Tarvainen, Sami; Hirvijoki, Eero; Laukkanen, Timo

Evaluation of reinforcement learning and model predictive control for apartment heating with heat pump and water storage tank

Published in:
Journal of Energy Storage

DOI:
10.1016/j.est.2026.120509

Published: 30/03/2026

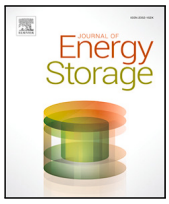
Document Version
Publisher's PDF, also known as Version of record

Journal of Energy Storage 152 (2026) 120509

Contents lists available at ScienceDirect

Journal of Energy Storage

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



Research papers

Evaluation of reinforcement learning and model predictive control for apartment heating with heat pump and water storage tank

Sami Tarvainen *, Eero Hirvijoki, Timo Laukkanen

Research Group of Energy Conversion and Systems, Department of Energy and Mechanical Engineering, Aalto University, PO Box 11000, FI-00076 AALTO, Finland

ARTICLE INFO

Dataset link: https://version.aalto.fi/gitlab/ehirvijo/RL-MPC-heating-control-evaluation

Keywords:
Reinforcement learning
Model predictive control
Proximal policy optimisation
Apartment space heating
Heat pump
Heat storage

ABSTRACT

Optimising apartment heating systems is becoming increasingly crucial due to growing use of heat pumps and heat storages. This study compares reinforcement learning (RL) and model predictive control (MPC) to optimise and control a simulated apartment heating system. We explore various RL designs, including binary and continuous action spaces, different reward functions, and three storage sizes. We find that MPC outperforms RL when the optimisation period aligns with the actual problem. This is shown in the lower electricity costs of MPC compared to RL with small and base storage sizes. However, the two methods yielded similar electricity costs with the large storage size. We also find that the RL design significantly affects its performance and robustness. The RL model with binary actions and a reward function promoting the active use of storage and profit maximisation outperforms other RL configurations. In turn, a reward function representing the actual problem of cost minimisation was found to be ineffective for agent training. Future studies that compare the two methods for the optimisation of heating systems with long-term storage could focus on extending the prediction horizon or enhancing the terminal cost term in MPC.

1. Introduction

Heat pumps are considered one of the key technologies for reducing the annual global greenhouse gas emissions of 4.2 GtCO2eq related to space and water heating [1], particularly when powered with low-carbon electricity sources. However, a key issue relates to the availability of affordable low-carbon electricity, as the production of wind or solar power may not coincide with the demand for heat. This issue can be addressed with an energy storage, such as a buffering water tank or an electric battery. The control and optimisation of such systems must consider certain constraints, such as energy conservation and occupant comfort [2], while trying to minimise, for example, energy consumption or costs.

There are at least three types of methods for the control of space heating: rule-based control (RBC), model predictive control (MPC), and reinforcement learning (RL). RBC is a conventional approach that keeps the system within defined limits based on predetermined rules [2]. Although RBC algorithms can be designed to achieve cost or energy savings (see e.g. [3]), these are not typically well suited for such tasks due to the reliance on heuristic rules. Increasing research has been focused on more advanced control methods, such as MPC and RL, to achieve smarter control of heating and other building control systems.

1.1. Existing literature

MPC methods have been widely studied in heating applications due to the methods' ability to reliably achieve multiple goals, such as thermal comfort and cost savings [4]. These methods use a model of the actual system that is optimised for a given time period, making it an effective approach for optimisation tasks. MPC employs a sliding window approach in which the optimisation results for the first time step are used to update the control actions of the system, after which the procedure is repeated for the next time step [5]. Many studies have shown MPC methods to achieve cost savings compared to RBC in simulation environments. Weeratunge et al. [6] and D'Ettorre et al. [7] reported that MPC can reduce electricity costs by about 8%, while Gholamibozanjani et al. [8] found savings of 12 to 57% compared to an RBC-based approach. MPC has also been applied in real-world cases to reduce energy consumption and costs (see e.g. [9,10]). The challenges of applying MPC methods to actual systems can include high computational costs, especially in accurate models with nonlinearities [11], and the costs of developing the model and installing the necessary hardware [9].

* Corresponding author.
E-mail address: sami.tarvainen@aalto.fi (S. Tarvainen).

https://doi.org/10.1016/j.est.2026.120509
Received 1 July 2025; Received in revised form 19 December 2025; Accepted 10 January 2026
Available online 24 January 2026

RL has also gained popularity in the scientific literature of building systems control, as it can handle complex control tasks without necessarily requiring accurate modelling of the actual system [2]. Additional advantages include possibility to handle uncertainties and solve long-term problems [11]. Instead of relying on a model of the system, model-free RL algorithms receive observations of the environment and take actions based on the observation and the current policy, i.e., a mapping between observed states and taken actions. The policy is improved by allowing the agent to interact directly with its environment and collecting rewards through sequences of interactions. RL algorithms have been shown to result in lower operational costs than RBC through peak shaving and lower electricity consumption [12] as well as by taking advantage of low electricity prices [13]. These simulation-based studies suggest that RL can cut operational costs by 22%–35% compared to RBC [12,13]. However, the study by Wang et al. [14] suggests that MPC methods can still outperform RL methods in controlling heating, ventilation, and air conditioning (HVAC) systems. They compared the performance of MPC and three different RL algorithms in the control of a simulated HVAC system. They found that the MPC model achieved 12%–14% lower electricity costs than the RL methods, but the RL could achieve better thermal comfort in some cases. It should also be noted that RL algorithms have important drawbacks that have limited their study and application mainly to simulation environments [2]. These include, for example, long training times and the possibility for RL controllers to drive the system to an undesired state [2]. RL robustness can be improved by adding rule-based backup controls [15] or by integrating RL with MPC methods [11].

1.2. Contribution

We contribute to the comparison of MPC and RL methods within the context of heating system control by evaluating their performance in optimising an apartment heating system that includes a heat storage. A heat storage increases the timespan of the optimisation problem compared to systems where temporal flexibility is mainly provided by the building's thermal mass (see [14]). Additionally, previous literature has mainly applied discrete [12] or continuous action spaces [13] for RL algorithms separately. We extend on these studies by comparing the relative efficacy of continuous vs. binary actions. Finally, the reward functions used in the previous literature focus on costs or energy consumption [12–14]. We further extend the literature by introducing a new reward function focusing on profit maximisation and active use of heat storage instead of cost minimisation. Similarly to Wang et al. [14], we find that MPC outperforms RL when the available heat storage capacity is comparable to the optimisation window of the MPC model. Further, we find that binary actions outperform continuous action spaces and that the newly introduced reward function enables more effective learning from the environment, which ultimately decreases electricity costs and improves RL robustness compared to a cost-based reward function.

The article proceeds with a review of the material and methods in Section 2, followed by the presentation of the results in Section 3. Section 4 discusses the results in the context of the existing literature, presents key limitations, and provides suggestions for future studies. Finally, Section 5 presents the conclusions of the study.

2. Material and methods

2.1. Study design

In this study, we simulate an apartment space heating system that includes a buffer water tank as heat storage, a heat pump, and a space heating demand (see Section 2.3). The simulated space heating system is optimised by the MPC model (see Section 2.4) and by the RL agent (see Section 2.5). The goal of optimisation is to meet the hourly demand for space heating with minimal electricity costs. Furthermore, we assume that the thermal comfort is ensured every hour by backup controls, which we have implemented in the simulation model.

The MPC and RL models receive known information about the current state of the system and future information up to 24 h ahead. The length of the observation window was selected based on the availability of day-ahead prices, which are always available at least 24 h ahead. Future information also includes forecast heat demand and coefficient of performance (COP) values, which are based on forecast outdoor temperatures.

The optimised heat supply of the heat pump by the MPC/RL model is given as input to the simulation model. Based on the input and specific constraints, the simulation model calculates the accurate heat supply and electric power of the heat pump, as well as the change in storage, after which it returns the state of charge (SOC) and additional information to the MPC/RL model. This cycle is repeated every hour for the first 8760 h for the year 2024, which is used to evaluate RL and MPC performance. Data for years 2020–2023 are used as training data for the RL agent (see Section 2.5), and the data for 2023 are used to tune the MPC model (see Section 2.4). RL training, MPC tuning, and model evaluation are performed using the same simulation model.

The performance of MPC and RL is compared with different RL designs. We study both binary ("BiAct") and continuous ("ContAct") action spaces, as well as two different reward functions ("RewCost" and "RewProfit") for the RL agent. Both the action space and the reward function can influence the performance of the RL algorithms. In reinforcement learning, the policy is updated based on the received rewards, and therefore the reward function is crucial for effective learning from the environment. In turn, the selection of an appropriate action space is not always straightforward. Although continuous action spaces can give granularity to system control, which is often desirable in real-world systems, their use can have drawbacks compared to discrete actions. With continuous actions, the output value must be computed from an infinite set of actions, which makes solving these RL problems inherently difficult [16].

The best RL design is chosen to compare how MPC and RL perform with different storage sizes. The motivation for using different storage sizes is that the used MPC model only optimises for 24 h ahead, whereas the RL model can potentially detect long-term trends based on the historical data, even though it only has information about the next 24 h. The base case ("BaseSto") has a useful thermal storage capacity of 3600 kWh, which is compared to storages with a 1200 kWh ("SmallSto") and a 8400 kWh ("LargeSto") useful capacity. Assuming a 50 kW heat demand, a consumption typical in a modelled winter day (see Section 2.2), the useful capacities would correspond to storage sizes of one, three, and seven days in the SmallSto, BaseSto, and LargeSto scenarios, respectively.

2.2. Data description

Exogenous data include outdoor temperatures from Kaisaniemi, Finland [17], Finland's day-ahead electricity prices [18], and artificially created space heating demand derived from district heating consumption in Helsinki [19]. All data is on an hourly basis for the years 2020–2024. For space heating, it is assumed that heating demand is zero with outdoor temperatures of $T_t^{\mathrm{outd}} = T^{\mathrm{href}} = 17°\mathrm{C}$ and above. It is further assumed that the heating demand at the design temperature of $T^{\mathrm{des}} = -26°\mathrm{C}$ [20] is 100 kW. A maximum heating capacity of 290 kW corresponds to an apartment building of 120 apartments for buildings constructed after 2005 [21], suggesting that the maximum heating power of 100 kW would correspond approximately to a building with 40 apartments.

The hourly space heating demand in Helsinki's district heating ($\dot{Q}_t^{\mathrm{space}}$) is first approximated for each hour. District heating data also includes water heating that is extracted from the data. The water heating demand is approximated by taking the average consumption for hours with outdoor temperatures of 17°C and above. The approximate

district water heating demand of $\dot{Q}^{\text{water}} = 235$ MW is then deducted from the total hourly district heating demand ($\dot{Q}_t^{\text{dh}}$) to get the hourly space heating demand of district heating, as described by Eq. (1).

For the observation period, the minimum outdoor temperature is −21.1°C with an approximate district space heating demand of $\dot{Q}^{\text{spaceMax}} = 2187$ MW. Based on the assumed maximum consumption of 100 kW at the design temperature, space heating demand for the apartment with an outdoor temperature of −21.1°C is $\dot{Q}^{\text{apartMax}} = 88.6$ kW. The apartment space heating demand for all other hours is calculated using the maximum space heating demand for the apartment and district heating and the hourly space heating demand of district heating as shown in Eq. (2), where $\Delta t = 1$ h. Fig. 1 shows the resulting demand for apartment space heating together with day-ahead prices and outdoor temperatures.

$$\dot{Q}_t^{\text{space}} = \dot{Q}_t^{\text{dh}} - \dot{Q}^{\text{water}} \tag{1}$$

$$Q_t^{\text{demand}} = \max(\dot{Q}_t^{\text{space}} \frac{\dot{Q}^{\text{apartMax}}}{\dot{Q}^{\text{spaceMax}}}, 0\ \text{kW}) \times \Delta t \tag{2}$$

Although the space heating demand is strongly dependent on outdoor temperature, there always remains forecast error between forecast and actual heat consumption. Thus, we introduce two types of forecast errors: (1) forecast error in outdoor temperatures and (2) unexplained variance. For the first type of error, we create a synthetic forecast of outdoor temperatures (T_t^{outdFor}) from the observed temperature data assuming a Gaussian distribution with a standard deviation of 2°C. For comparison, recently published artificial intelligence based Aurora 0.1° can provide surface temperature predictions with root mean squared errors of approximately 2°C with a one-day lead time [22]. The forecast temperatures are then used to derive the forecast heat demand for each hour (Q_t^{demFor}), as shown in Eq. (3). Outdoor temperature forecasts are also used to forecast COP values, so this error is also a source of unpredictability in heat pump performance.

$$Q_t^{\text{demFor}} = \max(100 \times \frac{T^{\text{href}} - T_t^{\text{outdFor}}}{T^{\text{href}} - T^{\text{des}}}, 0)\ \text{kWh} \tag{3}$$

The second type of forecast error in heat demand results from deriving the space heating demand from actual district heating data. This error component represents unmodelled and unknown factors, such as solar irradiance and occupant behaviour.

2.3. Simulation of the apartment heating system

The simulated apartment heating system comprises an air-to-water heat pump with an electric power capacity of 100 kW, a cylindrical water tank as heat storage, a radiator network, and a radiator as a heat sink. Fig. 2 illustrates the simulated heating system. Further details on the simulation are available in Section S1.2 of the Supplementary materials.

The COP values are based on a separate transcritical carbon dioxide (CO2) heat pump model with an internal heat exchanger (IHX). Precise COP values are calculated with the heat pump model as a function of the outside temperature and the water inlet temperature in the condenser, which is assumed to be equal to the water temperature at the bottom of the tank ($T_t^{\text{hpIn}} = T_t^{\text{bottom}}$). Table 1 outlines the key assumptions for the heat pump, including the assumptions used to calculate the precise COP values, while Fig. 3 illustrates these calculated values. The COP for each hour is derived through linear interpolation from the precise values.

The supply and return temperatures of the radiator network are assumed to depend linearly on the outdoor temperature. We note that actual systems can exhibit non-linearities in the supply temperature. However, a linear approximation is used as this is not expected to materially impact the comparability of MPC and RL methods in this study. The supply and return temperature are assumed to be 45°C and 30°C at the design temperature ($T^{\text{des}} = -26$°C), respectively. Both the supply and return temperatures are assumed to be equal to 20°C when the outdoor temperature is 20°C. The water outflow from the water tank is assumed to be mixed with the water returning from the radiators to ensure the appropriate supply temperature to the radiator network.

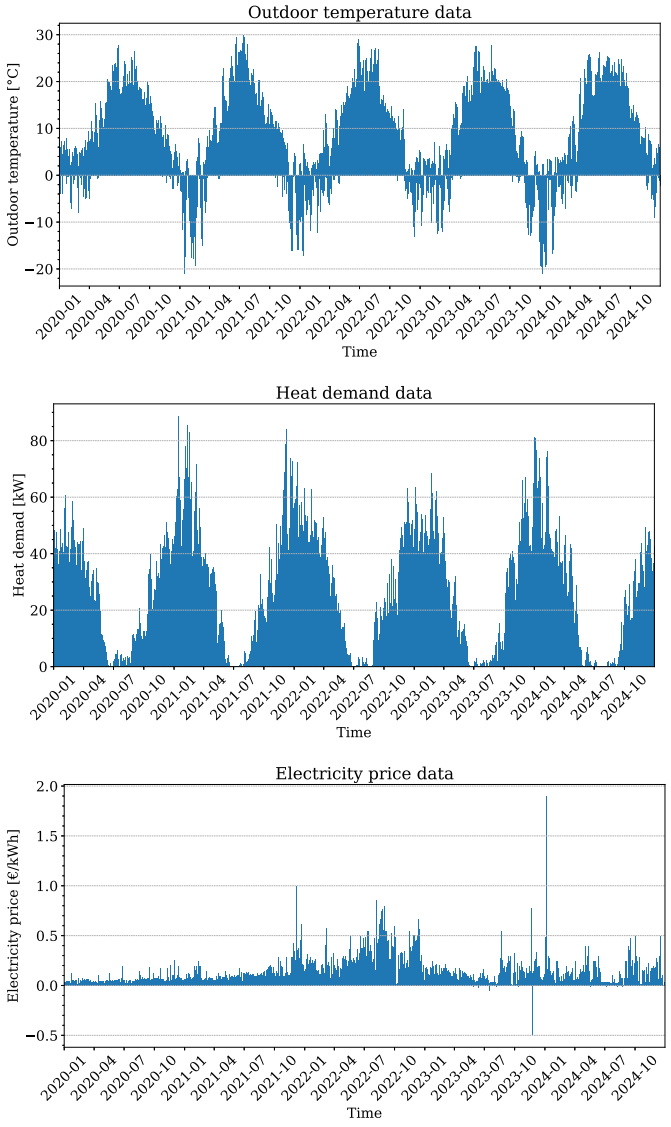


Fig. 1. Used data for hourly outdoor temperature [17], space heating demand [19], and day-ahead electricity prices [18].

Table 1
Key assumptions for the heat pump. Note: Evaporation and condenser temperatures refer to temperature of CO2. Discharge and temperature difference assumptions: $T^{\text{hpDis}} = 90$°C, $\Delta T = 5$°C.

Variable	Value	Unit
Capacity	100	kW_{el}
Evaporation temp.	$T_t^{\text{outd}} - \Delta T$	°C
Condenser inlet temp.	$\geq T^{\text{hpDis}} + \Delta T$	°C
Condenser outlet temp.	$T_t^{\text{bottom}} + \Delta T$	°C
Compressor isentr. effic.	0.80	–
Maximum discharge press.	130	bar

Table 2 presents the key assumptions for heat storage. The storage has a nominal capacity of 1600–11 200 kWh depending on the scenario and is assumed to be perfectly insulated. The minimum and maximum state of charge (SOC^{min} and SOC^{max}) limit the useful capacity. SOC^{min} ensures that heat storage can always deliver water to the

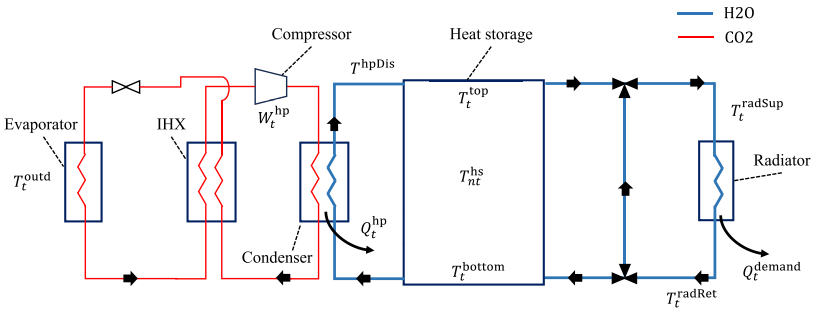


Fig. 2. Illustration of simulated apartment heating system.

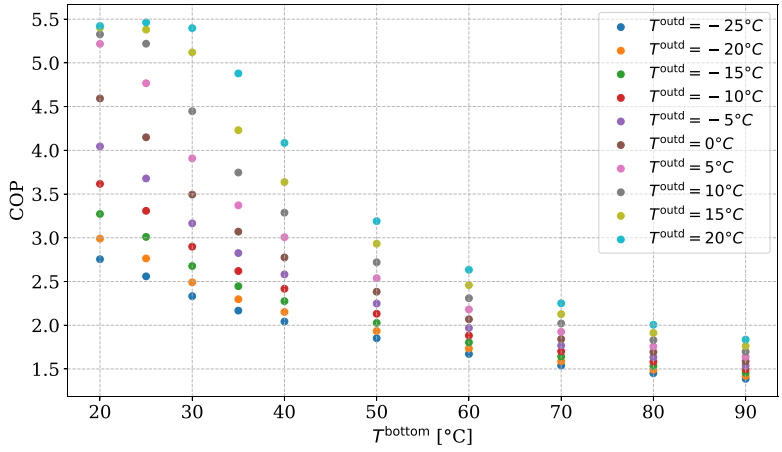


Fig. 3. Precise COP values used for the interpolation of hourly COP.

radiator network at the required temperature and SOC^{max} is assumed to avoid too high storage temperatures. The state of charge at each hour is defined by the mean (T_t^{mean}), minimum (T^{min}), and maximum temperature (T^{max}), as shown in Eq. (4).

$$SOC_t = (T_t^{mean} - T^{min})/(T^{max} - T^{min}), \quad (4)$$

where $T^{max} = 90°C$ and $T^{min} = 20°C$.

The storage is divided into differential fluid elements, each with a height of 1 mm. The temperature profile, energy flows, and electricity consumption are updated every hour. One simulation cycle includes the following stages: (1) optimised heat supply by the heat pump is received from the MPC or RL model, (2) the actual COP value is calculated, (3) the heat pump's thermal output is controlled to ensure that its power capacity is not exceeded and so that the storage remains within the defined limits, (4) the change in storage energy level is calculated from the heat pump supply and the demand for space heating, (5) the shift in the storage temperature profile is calculated, (6) the change in storage energy level and heat pump supply are adjusted to match to the energy transfer of discretised fluid elements, (7) the electricity consumption of the heat pump is calculated using the resulting heat pump supply and COP values, (8) convective mixing is calculated at the bottom of the tank if $T_t^{radRet} > T_t^{bottom}$, (9) heat diffusion is solved within the water tank applying the implicit Euler method in a one-dimensional case [23].

After the simulation is updated, the new state of charge is returned to the RL or MPC model. The simulated electricity consumption of the heat pump is used to calculate the actual electricity costs, which is used to compare the performance of RL and MPC.

2.4. Model Predictive Control (MPC)

The MPC model uses mixed-integer quadratically constrained programming (MIQCP) to represent the simulated apartment heating system approximately and optimises the system for a 24-hour window. At the beginning of each hour, the model receives information about the current state of the system and the electricity prices, forecast heat demand, and temperature-adjusted piecewise-linear functions of the COP values for the observation window. After finding an optimal solution, the heat pump supply for the first optimised hour is given as input for the simulation model, and the initial state of charge for the next optimisation window is updated based on the simulation output. This cycle is repeated hour by hour. Gurobi version 12.0.2 [25] for Python is used as the optimiser. Gurobi solves the model as a non-convex MIQCP model.

Eq. (5) shows the objective function of the MPC model with optimised variables highlighted in bold. The first term represents the optimised electricity costs, determined by the optimised electricity consumption of the heat pump ($\boldsymbol{W}_t^{hp}$) and the day-ahead prices (p_t^{el}). The second term aims to ensure stable long-term operation by penalising deviations from the designed state of charge $SOC^{des} = 0.45$. The parameter ω was tuned separately for each storage size based on the year 2023. Values of 0.1 to 1 were tested with 0.1 intervals, and the best (SmallSto: 0.7, BasteSto: 0.7, LargeSto: 0.6) were selected for the evaluation of the MPC model for the year 2024. In addition, since the magnitude of the first term is heavily dependent on the demand for

Table 2
Heat storage assumptions [24].

Variable	Value	Unit
Storage capacity	1600–11 200	kWh
Height-to-diameter	1/1	–
SOC^{min}	0.20	–
SOC^{max}	0.95	–
Specific heat capacity	4.18	kJ/kg/K
Water density	986	kg/m^3
Thermal diffusivity	0.157×10^{-6}	m^2/s

heat, we weight the second term with the forecast demand. Otherwise the model might overly emphasise short-term cost optimisation during winter or SOC balancing during warmer periods.

$$\min \sum_{t=1}^{24} (\boldsymbol{W}_t^{\mathrm{hp}}\, p_t^{\mathrm{el}}) + \omega(\boldsymbol{SOC}_{t=24} - SOC^{\mathrm{des}})^2 \sum_{t=1}^{24} Q_t^{\mathrm{demFor}} \tag{5}$$

A constraint is given to ensure a balance between heat supply and demand. The balance constraint comprises hourly heat pump supply ($\boldsymbol{Q}_t^{\mathrm{hp}}$), forecast heat demand (Q_t^{demFor}), and the change in storage level ($\boldsymbol{\Delta Q}_t^{\mathrm{hs}}$), as illustrated in Eq. (6).

$$\boldsymbol{Q}_t^{\mathrm{hp}} = Q_t^{\mathrm{demFor}} + \boldsymbol{\Delta Q}_t^{\mathrm{hs}} \tag{6}$$

2.4.1. Heat storage constraints

A maximum and a minimum mean temperature are assumed for heat storage, which determine the storage capacity together with the storage dimensions, as shown in Eq. (7). The maximum temperature is equal to the discharge temperature of the heat pump, and the minimum temperature is assumed to be $T^{\min} = 25°\mathrm{C}$, corresponding to the average return temperature of the radiator network. The hourly storage level ($\boldsymbol{Q}_t^{\mathrm{hs}}$) cannot exceed the capacity, as shown in Eq. (8).

$$Q^{\mathrm{hsCap}} = \pi r^2 H \rho c_p (T^{\max} - T^{\min}) \tag{7}$$

$$\boldsymbol{Q}_t^{\mathrm{hs}} \le Q^{\mathrm{hsCap}} \tag{8}$$

An optimised variable is created for the state of charge ($\boldsymbol{SOC}_t$), which affects the COP of the heat pump as described in Section 2.4.2. The state of charge is calculated as shown in Eq. (9). Furthermore, minimum ($SOC^{\min}$) and maximum ($SOC^{\max}$) charge levels are assumed to limit the useful storage capacity, as shown in Eqs. (10) and (11).

$$\boldsymbol{SOC}_t = \boldsymbol{Q}_t^{\mathrm{hs}} / Q^{\mathrm{hsCap}} \tag{9}$$

$$\boldsymbol{SOC}_t \le SOC^{\max} \tag{10}$$

$$\boldsymbol{SOC}_t \ge SOC^{\min} \tag{11}$$

The change in storage level ($\boldsymbol{\Delta Q}_t^{\mathrm{hs}}$) is determined by the charge ($\boldsymbol{Q}_t^{\mathrm{cha}}$) and discharge ($\boldsymbol{Q}_t^{\mathrm{dis}}$) of storage, as depicted by Eq. (12). The hourly storage at time $t+1$ is determined by the storage level in the last time step, the change in storage, and the heat losses ($\boldsymbol{Q}_t^{\mathrm{loss}}$), as shown in Eq. (13). However, the heat losses are assumed to be negligible ($c = 0$) for the modelled storage (see Eq. (14)).

$$\boldsymbol{\Delta Q}_t^{\mathrm{hs}} = \boldsymbol{Q}_t^{\mathrm{cha}} - \boldsymbol{Q}_t^{\mathrm{dis}} \tag{12}$$

$$\boldsymbol{Q}_{t+1}^{\mathrm{hs}} = \boldsymbol{Q}_t^{\mathrm{hs}} + \boldsymbol{\Delta Q}_t^{\mathrm{hs}} - \boldsymbol{Q}_t^{\mathrm{loss}} \tag{13}$$

$$\boldsymbol{Q}_{t+1}^{\mathrm{loss}} = c\ \boldsymbol{Q}_t^{\mathrm{hs}} \tag{14}$$

An initial storage level ($\boldsymbol{Q}_{t=0}^{\mathrm{hs}}$) is set based on the state of charge at the end of the last hour, as shown in Eq. (15). For the first hour of the year, the initial state of charge is set as $SOC^{\mathrm{init}} = 0.5$ and for all other hours the initial storage level is set based on the SOC level after simulation, as described in Section S1.2.2 of the Supplementary material.

$$\boldsymbol{Q}_{t=0}^{\mathrm{hs}} = SOC^{\mathrm{init}} Q^{\mathrm{hsCap}} \tag{15}$$

2.4.2. Heat pump constraints

Constraints for the heat pump are presented next. The electric power of the heat pump cannot exceed its capacity, as shown in Eq. (16). The hourly heat supply is determined by the optimised work, or electricity consumption ($\boldsymbol{W}_t^{\mathrm{hp}}$), and coefficient of performance ($\boldsymbol{COP}_t$), giving us the quadratic constraint shown in Eq. (17).

$$\boldsymbol{W}_t^{\mathrm{hp}} \le W^{\mathrm{hpCap}} \tag{16}$$

$$\boldsymbol{Q}_t^{\mathrm{hp}} = \boldsymbol{W}_t^{\mathrm{hp}} \boldsymbol{COP}_t \tag{17}$$

The simulation model assumes a nonlinear relationship between the buffer tank's bottom temperature and COP value (see Fig. 3). The MPC model assumes a simplified representation where the temperature profile is not explicitly modelled, but the bottom temperature is approximated as a function of the SOC. Thus, we impose a piecewise-linear constraint between $\boldsymbol{COP}_t$ and $\boldsymbol{SOC}_t$, as shown in Eq. (18). Gurobi transforms a model with piecewise-linear constraints into a mixed-integer program [25]. We note that the piecewise-linear constraint could probably be replaced by a smooth function, which would eliminate the need for mixed-integer formulation and potentially speed up computational time. However, the recorded solving times suggest that the chosen approach is also deployable in practice (see Section 3.3).

$$\boldsymbol{COP}_t = f_t(\boldsymbol{SOC}_t) \tag{18}$$

The function f_t (see Eq. (S44)-(S45)) is defined by a set of points (SOC_j, COP_{jt}), where SOC_j replaces the T_j^{bottom} values used for the calculation of accurate COP in the simulation model. The points COP_{jt} are defined for each hour using predicted outdoor temperatures. Further details are available in Section S1.3 of the Supplementary materials.

2.5. Reinforcement Learning (RL) model

Proximal Policy Optimisation (PPO) was chosen as the reinforcement learning method to optimise the simulated apartment heating system. PPO belongs to the field of deep reinforcement learning, and is a family of policy gradient methods with a clipped surrogate objective [26]. Clipping of the surrogate objective penalises too large policy updates, which results in better stability and reliability [26]. PPO can handle both discrete and continuous actions spaces, making it well suitable to a wide range of optimisation problems. Please refer to the article by Schulman et al. [26] for further information on PPO.

A custom RL environment was created using Gymnasium, a fork of OpenAI's Gym library [27]. The custom Gymnasium environment was integrated with the apartment heating system simulation model to allow direct interaction between the agent and the simulation model. Stable Baselines3 [28] was used for the implementation of the PPO algorithm.

2.5.1. Observation space and environment

The simulation model is the environment in which the agent takes actions based on the policy and available observations. Unlike MPC, the RL model does not require any approximate model of the simulation model as it can learn to optimise the system by directly interacting with it. However, the RL agent requires enough information about the environment to effectively optimise the system. The information provided to the agent includes the state of charge and six evenly spaced storage temperature values from the bottom to the top of the tank ($T_{n,t-1}^{\mathrm{hs}}$) at the end of the last hour, the forecast COP value for the next hour (COP_t^{for}) which is based on the bottom temperature of the tank and the forecast outdoor temperature, as well as the demand forecast and electricity prices for the next 24 h.

The motivation for providing the agent with additional information about the tank's temperature profile is to inform the agent about future COP values beyond the next hour. The COP value of the heat pump is determined by the outdoor temperature and the bottom temperature of the heat storage, so the agent could estimate the future COP value with the information of tank's temperature profile and outdoor temperature forecasts. Please note that we do not explicitly provide forecast outdoor temperatures as observations because the same information is fully captured in forecast demand.

All observations are normalised and clipped between 0 and 1 based on assumed minimum and maximum values, as shown in Eq. (19). The assumptions for the minimum and maximum values are tabulated in Table 3.

$$Obs_k^{\mathrm{norm}} = \max\left[\min\left(\frac{Obs_k - Obs_k^{\min}}{Obs_k^{\max} - Obs_k^{\min}}, 1\right), 0\right] \tag{19}$$

Table 3
Assumptions for the normalisation of RL observations.

Variable	$Obs_k^{\min}$	$Obs_k^{\max}$	Unit
SOC_{t-1}	0	1	–
$T_{n,t-1}^{\mathrm{hs}}$	20	90	°C
COP_t^{for}	1	6	–
$Q_{t\ldots t+23}^{\mathrm{demFor}}$	0	100	kWh
$p_{t\ldots t+23}^{\mathrm{el}}$	−0.01	0.2	€/kWh

2.5.2. Action space and reward function

The RL agent controls the heat pump with continuous or binary actions. In ContAct models, the agent takes a continuous action $\boldsymbol{a}_t \in [-1, 1]$ which is converted into electricity consumption of the heat pump, as shown in Eq. (20). These agents are then compared with the BiAct models, which have binary actions $\boldsymbol{a}_t \in \{0, 1\}$, i.e., on/off control over the heat pump (see Eq. (21)). The power consumption of the heat pump is then converted into heat supply (see Eq. (22)), which is given as input to the simulation model.

$$\boldsymbol{W}_t^{\mathrm{hp}} = (0.5\boldsymbol{a}_t + 0.5)W^{\mathrm{hpCap}} \quad \text{(ContAct)} \tag{20}$$

$$\boldsymbol{W}_t^{\mathrm{hp}} = \boldsymbol{a}_t W^{\mathrm{hpCap}} \quad \text{(BiAct)} \tag{21}$$

$$\boldsymbol{Q}_t^{\mathrm{hp}} = COP_t^{\mathrm{for}} \boldsymbol{W}_t^{\mathrm{hp}} \tag{22}$$

Two reward functions are used with the RL modelling. In the RewCost models, the reward for every step is the negative of the simulated electricity cost for hour t (see Eq. (23)). Please note that the simulated electricity consumption (W_t^{hp}) can deviate from the agent's output $\boldsymbol{W}_t^{\mathrm{hp}}$ due to the simulation controls as described in Section 2.3. Electricity costs were also used in the reward function by Schmitz et al. [13], but they also included a reward based on the state of charge. Due to the separate control of heat output in the simulation model, we do not include a reward for keeping the state of charge within the boundaries.

The second reward function (RewProfit) is designed from the perspective of heat storage. With positive electricity prices, the agent receives a positive reward when discharging the storage and a negative reward when charging the storage, as described by the first term in Eq. (24). The second term gives a positive reward both for charging and for discharging. The multiplier 0.01 for the second term is used to ensure that the rewards from the first and second terms have the same order of magnitude. This reward function is designed to encourage the agent to take advantage of the differences in electricity prices and COP values. By doing so, the agent is incentivised for the maximisation of 'profits' by the strategic usage of storage, rather than solely focusing on minimising electricity costs.

$$r_t = -W_t^{\mathrm{hp}}\, p_t^{\mathrm{el}} \quad \text{(RewCost)} \tag{23}$$

$$r_t = -\frac{p_t^{\mathrm{el}}}{COP_t}\Delta Q_t^{\mathrm{hs}} + 0.01|\Delta Q_t^{\mathrm{hs}}| \quad \text{(RewProfit)} \tag{24}$$

2.6. Training setup

Each agent is trained for a total of a million steps by interacting with the apartment simulation model. Data for years 2020–2023 are used for the training process. The training process applies the same principles as used for the model evaluation for the year 2024. For example, the agent receives forecast information about COP values and heat demand, but the simulation model is updated based on actual values. Following Schmitz et al. [13], the number of steps between each policy update is set as one year (8760 steps), and the learning rate is set to decrease linearly from 0.001 to 0 during the learning process. The mini batch size is set as 365 h, or approximately 15 days. Otherwise, the default parameters of PPO in Stable Baselines3 [28] are used. Due to the stochastic nature of the learning process, five agents are trained for each RL design with a different seed for the pseudo-random generator.

3. Results

The results for MPC and different RL algorithms are reviewed first for the base storage size in Section 3.1. The results with different storage sizes are further compared with different storage sizes with the best performing RL algorithm in Section 3.2. Additionally, Section 3.3 presents computational times of the different models.

3.1. RL and MPC results with base storage size

Fig. 4 shows the total cost of electricity for the test period, that is, for the year 2024. The figure shows that MPC outperforms all RL algorithms in the BaseSto scenario. The better performance of MPC can be attributed to three key factors: (1) The MPC model can reliably find an optimal solution to the simplified model, whereas the actions taken by RL agents are by design approximate; (2) The MPC model represents the simulated system accurately enough; and (3) The RL agents do not gain significant enough advantage by optimising beyond the observation window with this storage size. However, the outperformance of MPC fades with the large storage size, as discussed in Section 3.2.

We find significant differences between the RL designs in their robustness and performance. The BiAct-RewProfit RL model outperforms other RL designs in terms of electricity costs with mean electricity costs 19.4% above the MPC model. The BiAct-RewProfit is also the most robust design with the lowest variation in costs between agents trained with different seeds, as shown by the error bars in Fig. 4. The mean electricity costs of RL models ContAct-RewProfit, BiAct-RewCost, and ContAct-RewCost exceed the costs of MPC model by 46.8%, 53.0%, and 66.4%, respectively. RL agents with the RewCost reward function have the highest electricity costs on average and the largest variation in performance.

Differences in RL performance can be attributed to their behaviour. Agents with the RewCost reward function use the heat pump too rarely or with too low power in the case of continuous actions. The cautious use of heat pump can be seen in the upregulation in simulation model backup controls, as shown in Table 4. Upregulation is on average 4.4 kW and 6.4 kW with the ContAct-RewCost and BiAct-RewCost designs, respectively. In contrast, the average upregulation is only 0.7–0.8 kW with the ContAct-RewProfit and BiAct-RewProfit designs. Upregulation controls increase electricity consumption during the hours of high electricity prices, which increases total costs. The worse performance of the agents with the RewCost reward function is likely due to the fact that the agent can only receive negative rewards from the use of the heat pump. Therefore, the agent is biased to use the heat pump with caution. On the other hand, the results and agent actions indicate that the best performance of the BiAct-RewProfit design is explained by the combination of a good reward function, which encourages for the utilisation of storage and profit seeking behaviour, as well as the binary actions that force the agent to take decisive actions.

Fig. 5 further illustrates the differences by showing the energy supply and consumption in 1–7 December 2024 for the MPC, BiAct-RewProfit, and ContAct-RewCost models. Graphs for the other models are available in Section S2.2 of the Supplementary material. Only one agent (the agent with the mid-performing seed) is displayed for the RL cases. The figure shows that the use of heat pump for MPC and the BiAct-RewProfit agent coincide quite closely in time. The figure illustrates how the BiAct-RewProfit agents learn to take advantage of the low electricity prices with a limited need for upregulation from the simulation model controls. In turn, the ContAct-RewCost agent uses the heat pump with lower power compared to the MPC model, and the agent's use of the heat pump is more spread out in time. The overly cautious behaviour of the ContAct-RewCost agents is evident in their low-power use of the heat pump and the upregulation by simulation controls, which ultimately results in higher costs.

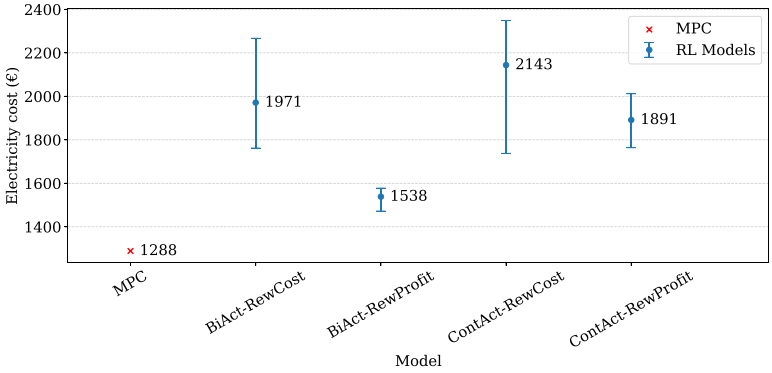


Fig. 4. Electricity cost for the BaseSto scenario for the test period (year 2024). Note: The bars for RL models represent the range of values across different seeds, while the dot indicates the mean value.

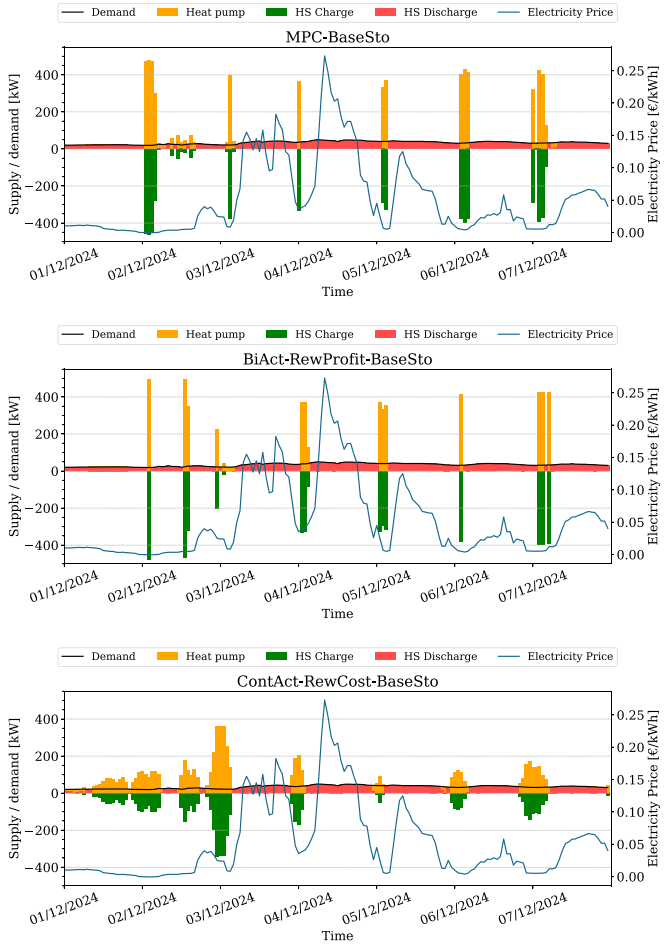


Fig. 5. Energy supply and consumption in the first week of Dec-2024 for the BaseSto scenario with MPC as well as BiAct-RewProfit and ContAct-RewCost agents.

3.2. The effect of storage size

The best performing RL design, that is, BiAct-RewProfit, is further benchmarked against the MPC model with small and large storage sizes. Fig. 6 shows the total electricity costs for these scenarios. The average electricity cost increases from the BaseSto to SmallSto scenario both for the MPC and BiAct-RewProfit models. This is an expected result, as the larger storage size provides more opportunities for the utilisation of low electricity prices.

Table 4
Average upregulation of the heat pump by simulation model backup controls for the test period (year 2024). The average of all seeds is shown for the RL agents.

Model	Average upregulation [kW]		
	SmallSto	BaseSto	LargeSto
MPC	0.16	0.37	0.58
BiAct-RewProfit	0.50	0.75	1.00
ContAct-RewProfit	*n/a*	0.66	*n/a*
BiAct-RewCost	*n/a*	6.36	*n/a*
ContAct-RewCost	*n/a*	4.35	*n/a*

The gap between the electricity costs of the MPC and the BiAct-RewProfit agents widens when a small storage size is assumed. The BiAct-RewProfit agents operate on average with 27.1% higher electricity costs compared to the MPC model, while in the BaseSto scenario the agents' costs were only 19.4% higher. This is likely explained by the fact that the optimisation window of 24 h used in the MPC model closely matches the capacity of the small storage. For example, the useful capacity of the SmallSto scenario (1200 kWh) is discharged within 23 h with the average heat demand of January 2024 (52.5 kW).

Interestingly, the average electricity costs of the MPC and BiAct-RewProfit models are practically on par with the large storage size. This suggests that the advantage of the RL model in inferring beyond the observation window balances its less accurate representation of the system. However, the electricity costs of the MPC model also decrease by 14.9% from the BaseSto to the LargeSto scenario. This is because the changes in SOC are smaller with a larger storage, allowing MPC to restrict the use of the heat pump during relatively high electricity prices without compromising long-term performance.

Fig. 7 illustrates the actions of BiAct-RewProfit agents with small and large storage sizes. In the SmallSto scenario, the agent uses the heat pump regularly throughout the week with a one-day cycle time. In contrast, the agent goes for 46 h without activating the heat pump in 3–5 December in the LargeSto scenario, showing its ability to infer beyond the observation window.

Table 5 presents the average COP values by storage size and model. The MPC model shows a 1%–3% higher average COP with both base and large storage sizes compared to the RL models. However, with the small storage size, the MPC model operates with a lower COP. In this case, the model seems to trade off heat pump performance to take advantage of lower electricity prices and utilise a greater share of storage capacity.

3.3. Computation time

Table 6 presents the model runtimes to inform about the deployability of the different methods. The modelling was done using a HP

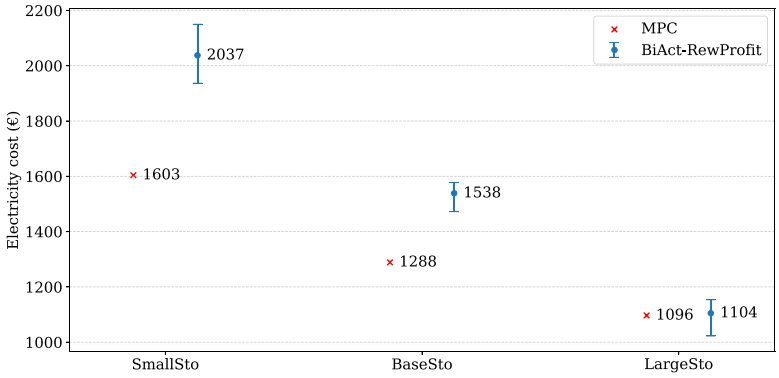


Fig. 6. Electricity cost for MPC and the BiAct-RewProfit models for the test period (year 2024) by storage scenario. Note: The bars for RL models represent the range of values across different seeds, while the dot indicates the mean value.

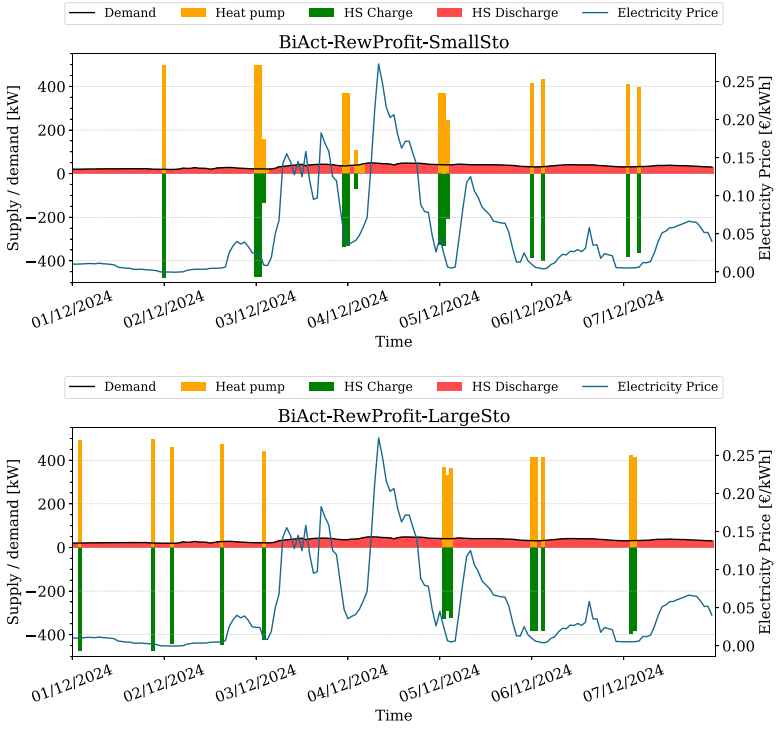


Fig. 7. Energy supply and consumption in the first week of Dec-2024 for BiAct-RewProfit with small and large storage size.

Table 5
Average COP values for the test period (year 2024). The average of all seeds is shown for the RL agents.

Model	Average COP [–]		
	SmallSto	BaseSto	LargeSto
MPC	3.99	4.09	4.12
BiAct-RewProfit	4.02	4.01	4.06
ContAct-RewProfit	*n/a*	3.97	*n/a*
BiAct-RewCost	*n/a*	4.05	*n/a*
ContAct-RewCost	*n/a*	4.03	*n/a*

Elitebook 840 14 inch G10 Notebook PC with the 13th Gen Intel® Core™ i5-1335U (1300 MHz) processor with 10 cores and 12 logical processors, and with a total RAM of 16 GB.

The computational times for simulation, MPC solving, RL training, and RL inference suggest that these models could be deployed in practice. No significant numerical problems were detected with the MPC model, as the maximum solving time was seven seconds. However, a more detailed MPC model with a longer optimisation period could incur a higher computational cost. We also note that the reported RL training times are based on a simulation environment. The large number of training steps required for RL training would make training in an actual system impractical. This means that a simulation model

Table 6
Runtimes for the MPC, RL, and the simulation environment. The simulation times were recorded with the MPC modelling.

Variable	Min	Median	Max	Unit
Simulation	<0.01	0.01	0.21	s
MPC solving	<0.01	0.01	7.3	s
RL training	101	145	249	min
RL inference	<0.01	<0.01	0.02	s

might have to be developed of the actual system if the used RL method was deployed in practice.

4. Discussion

This section discusses the results in the context of existing literature and the key limitations of the study. Then, suggestions for future studies are presented based on the limitations and results.

4.1. Results in the context of existing literature

Based on the results, MPC outperforms all RL designs with the base storage size. The electricity costs of the RL agents exceed the MPC costs on average by 19.4–66.4% in the BaseSto scenario, depending on the RL design. The best performing RL design is BiAct-RewProfit, which uses binary actions spaces and a reward function promoting for the active use of storage and profit maximisation. This RL design outperforms the others in terms of average electricity costs and robustness. The use of continuous action spaces and the RewCost reward function, which only considers electricity cost, is found to result in overly cautious use of the heat pump and storage. The fact that a reward function representing the real problem of cost minimisation resulted in poor performance of the agent is somewhat worrisome. Additional reward terms that do not represent the actual objective could result in unexpected behaviour, and the need to design reward functions can increase the costs of developing RL algorithms.

The BiAct-RewProfit design was further studied with small and large storage sizes. The gap between MPC and RL costs was found to grow with smaller storage size. This is a result of the fact that the MPC model can represent the optimisation problem more accurately than the RL model for the length of the observation window. The smaller storage size yielded a closer alignment between the optimisation problem and the MPC model because of shorter storage cycle times, giving a greater edge to the MPC approach. However, when the RL and MPC models were applied to large storage, the benefits observed with small storage were reversed. In this case, the average electricity costs of the MPC and RL models were approximately the same. The RL model could gain a greater advantage with the large storage size because of its ability to infer optimal policies beyond the observation window, which counterbalanced the less accurate representation of the system.

The results suggest that MPC models can provide heat energy with lower costs than RL algorithms when combined with short-term heat storage (1–3 days) and a heat pump. This is important contribution to studies that have found RL methods to outperform RBC methods in heating systems including a heat storage [12,13]. Our results are similar to those of Wang et al. [14] who found that MPC outperforms RL methods in the control of HVAC systems. The results indicate that MPC outperforms RL in apartment heating systems when the available heat storage capacity is comparable to the optimisation window of the MPC model. However, the RL and MPC models resulted in similar costs when the actual problem of cost minimisation clearly exceeded the observation horizon in the case of large storage. It should be noted that possibilities remain for improving the performance of MPC in long-term optimisation by extending the observation horizon with predicted values or by introducing a better terminal cost function, but this work is left for future studies.

4.2. Limitations

The study has limitations that should be considered when interpreting the results. The first key limitation is that the demand for space heating is derived from publicly available district heating data [19]. This also includes domestic water heating, which could not be accurately excluded from the raw data. This can lead to higher variability in space heating demand than would actually be the case for an apartment building.

Second, the modelling was done using a simulated heating system which has many simplifying assumptions. For example, the study does not consider heat losses or mass flows in the heat pump (which could affect compressor efficiency) and the space heating system. Thus, there would likely be larger forecast errors and modelling inaccuracies between the control methods and the system if these were applied in real-world. We also note that the training of an RL agent would be impractical in an actual system, which could necessitate the development of a simulation model. This requirement could negate one of the key benefits of model-free methods, which are intended to be applicable without an explicit system model.

Third, we note that the RL model receives more information about the state of the system than the MPC model by receiving information about the heat tank's temperature profile. Although the results suggest that the MPC model can represent the simulated system more accurately, this approach could give some edge to the RL model. In the case of the MPC model, we assumed a piecewise-linear function between the SOC and COP values, and thus did not utilise information about the temperature profile. This simplification speeds up computation and simplifies system representation. In turn, feeding more information to the RL algorithm does not significantly affect the observed inference time or complicate engineering related to the model development as RL is by design model-free, and hence a larger input space was used in the case of RL. However, a larger input space can increase training time or necessitate installation of additional hardware (e.g. temperature sensors) which could hinder broadening of the input space in practice.

Fourth, the results of the RL model could be affected by hyperparameters that were not specifically tuned for this case. This is somewhat counteracted by the fact that the proximal policy optimisation was designed to be relatively robust and general, that is, the same hyperparameters should be applicable to a variety of problems [26].

Finally, only 24 h of forward looking data were provided for the MPC and RL models with all modelled storage sizes. Using a longer data horizon with predicted data could improve the performance of both the MPC and the RL models. In this study, the data window used could give an advantage to the RL models with the base and large storage sizes, since the RL models can learn long-term trends based on historical data, but the MPC optimisation is only based on the next 24 h with a terminal cost term only for the state of charge. It should be noted that RL can be combined with MPC to potentially develop policies that outperform the standalone approaches of MPC and RL [11,29]. However, exploring this combined approach is outside the scope of this study, as we have focused on comparing the performance of RL and MPC as independent methods.

4.3. Future studies

Future research could study how RL and MPC compare with longer optimisation periods, such as in seasonal energy storage. These studies could incorporate a longer prediction horizon or put more effort into shaping the terminal cost function of the MPC model. Forecast errors would also become a larger issue in long-term forecasting, which could be considered in future studies. Furthermore, research on more detailed simulated systems or physical systems would be useful for comparing the alternative control methods. The MPC and RL algorithms used in this study could prove impractical in detailed systems due to long solver runtimes or long RL training time, which could require novel approaches to solving such problems.

5. Conclusion

This article studied the use of reinforcement learning and model predictive control to optimise the heating system of an apartment building. The simulated heating system comprises an air-to-water heat pump, a buffering water tank as heat storage, and a heat sink for space heating. The results showed that the best RL model had, on average, 19.4% higher electricity costs than the MPC model with the base storage size. Binary action spaces coupled with a reward function that promotes active storage use and profit maximisation were found to significantly improve the RL performance in terms of average electricity costs and robustness, i.e., reduced variation in results.

The best performing RL model was further tested against MPC with small and large storage sizes. MPC outperformed RL also with the small storage size, as the average costs of RL exceeded those of MPC by 27.1%. However, with the large storage size, the RL agents achieved similar costs to those of the MPC approach. This was explained by the fact that the applied MPC model optimised electricity costs over a 24-hour window, whereas the RL agents could infer good policies beyond the 24-hour period.

These results suggest that MPC methods can outperform RL algorithms in optimisation of heating systems, given that the optimisation period of the actual problem and the MPC model are aligned and the MPC model represents the actual problem with the necessary accuracy. Future studies could use a longer prediction window or consider alternative terminal cost functions in MPC to move from finite to infinite horizon optimisation. These improvements could provide better insights into the relative performance of MPC and RL methods in optimising and controlling heating systems with long-term storage.

Glossary

Abbreviations	
CO2	Carbon dioxide
COP	Coefficient of performance
HVAC	Heating, ventilation, and air conditioning
IHX	Internal heat exchanger
MIQCP	Mixed-integer quadratically constrained program
MPC	Model predictive control
PPO	Proximal policy optimisation
RBC	Rule-based control
RL	Reinforcement learning
SOC	State of charge

Variables and parameters			
$\boldsymbol{a}_t$	Agent action [–]	$\dot{Q}_t^{\text{spaceMax}}$	Max. district space heating demand [kW]
c	Coefficient for heat loss [–]	$\dot{Q}_t^{\text{water}}$	District water heating demand [kW]
c_p	Specific heat capacity [kJ/kg/K]	r_t	RL reward [–]
COP_t	Modelled COP [–]	SOC^{init}	Initial SOC [–]
ΔQ_t^{hs}	Change in storage [kWh]	SOC^{max}	Maximum SOC [–]
ΔT	Temperature difference [°C]	SOC^{min}	Minimum SOC [–]
Δt	Time step [h]	SOC_t	Modelled SOC [–]
H	Storage height [m]	T^{bottom}	Temperature at bottom of tank [°C]

Variables and parameters			
Obs_k	RL observation k	T^{des}	Design temperature [°C]
Obs_k^{max}	Max. for RL normalisation	T^{hpDis}	Heat pump's discharge temperature [°C]
Obs_k^{min}	Min. for RL normalisation	T_t^{hpIn}	Heat pump's inlet temp. [°C]
Obs_k^{norm}	Normalised observation	T^{href}	Reference temperature for heating cut-off [°C]
p_t^{el}	Electricity price [€/kWh]	T_{nt}^{hs}	Storage temperature at point n [°C]
$\dot{Q}^{\text{apartMax}}$	Max. apartment building heat demand [kW]	T^{max}	Max. storage temp. [°C]
$\boldsymbol{Q}_t^{\text{cha}}$	Charge of storage [kWh]	T_t^{mean}	Mean storage temp. [°C]
Q_t^{demand}	Modelled heat demand [kWh]	T^{min}	Min. storage temp. [°C]
Q_t^{demFor}	Demand forecast [kWh]	T_t^{outd}	Outdoor temperature [°C]
$\dot{Q}_t^{\text{dh}}$	District heating demand [kW]	T_t^{outdFor}	Forecast outdoor temperature [°C]
$\boldsymbol{Q}_t^{\text{dis}}$	Discharge of storage [kWh]	W_t^{hp}	Heat pump work [kWh]
$\boldsymbol{Q}_t^{\text{hp}}$	Heat pump supply [kWh]	W^{hpCap}	Heat pump capacity [kW]
$\boldsymbol{Q}_t^{\text{hs}}$	Heat storage charge level [kWh]		
Q^{hsCap}	Heat storage capacity [kWh]	*Greek symbols*	
$\boldsymbol{Q}_t^{\text{loss}}$	Heat loss of heat storage [kWh]	ρ	Water density [kg/m^3]
$\dot{Q}_t^{\text{space}}$	District space heating demand [kW]	ω	Weight parameter in MPC objective [–]

CRediT authorship contribution statement

Sami Tarvainen: Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Eero Hirvijoki:** Writing – review & editing, Software, Project administration, Methodology, Conceptualization. **Timo Laukkanen:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Funding sources

This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant number 353299.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.est.2026.120509.

Data availability

The research data, including the source code, the simulation results and the results figures, are available in the following repository https://version.aalto.fi/gitlab/ehirvijo/RL-MPC-heating-control-evaluation.

References

[1] IEA, CO2 emissions from buildings space and water heating by fuel in the net zero scenario, 2010–2030, 2023, URL https://www.iea.org/data-and-statistics/charts/co2-emissions-from-buildings-space-and-water-heating-by-fuel-in-the-net-zero-scenario-2010-2030.

[2] Z. Wang, T. Hong, Reinforcement learning for building controls: The opportunities and challenges, Appl. Energy 269 (2020) 115036, http://dx.doi.org/10.1016/J.APENERGY.2020.115036.

[3] B. Alimohammadisagvand, J. Jokisalo, S. Kilpeläinen, M. Ali, K. Sirén, Cost-optimal thermal energy storage system for a residential building with heat pump heating and demand response control, Appl. Energy 174 (2016) 275–287, http://dx.doi.org/10.1016/J.APENERGY.2016.04.013.

[4] D. Kim, J. Lee, S. Do, P.J. Mago, K.H. Lee, H. Cho, Energy modeling and model predictive control for HVAC in buildings: A review of current research trends, Energies 15 (2022) 7231, http://dx.doi.org/10.3390/EN15197231, URL https://www.mdpi.com/1996-1073/15/19/7231.

[5] A. Afram, F. Janabi-Sharifi, Theory and applications of HVAC control systems – a review of model predictive control (MPC), Build Environ. 72 (2014) 343–345, http://dx.doi.org/10.1016/J.BUILDENV.2013.11.016.

[6] H. Weeratunge, G. Narsilio, J. de Hoog, S. Dunstall, S. Halgamuge, Model predictive control for a solar assisted ground source heat pump system, Energy 152 (2018) 974–984, http://dx.doi.org/10.1016/J.ENERGY.2018.03.079.

[7] F. D'Ettorre, P. Conti, E. Schito, D. Testi, Model predictive control of a hybrid heat pump system and impact of the prediction horizon on cost-saving potential and optimal storage capacity, Appl. Therm. Eng. 148 (2019) 524–535, http://dx.doi.org/10.1016/J.APPLTHERMALENG.2018.11.063.

[8] G. Gholamibozanjani, J. Tarragona, A. de Gracia, C. Fernãndez, L.F. Cabeza, M.M. Farid, Model predictive control strategy applied to different types of building for space heating, Appl. Energy 231 (2018) 959–971, http://dx.doi.org/10.1016/J.APENERGY.2018.09.181.

[9] D. Sturzenegger, D. Gyalistras, M. Morari, R.S. Smith, Model predictive climate control of a swiss office building: Implementation, results, and cost-benefit analysis, IEEE Trans. Control. Syst. Technol. 24 (2016) 1–12, http://dx.doi.org/10.1109/TCST.2015.2415411.

[10] M. Castilla, J.D. Ãlvarez, J.E. Normey-Rico, F. Rodríguez, Thermal comfort control using a non-linear MPC strategy: A real case of study in a bioclimatic building, J. Process Control 24 (2014) 703–713, http://dx.doi.org/10.1016/J.JPROCONT.2013.08.009.

[11] W. Cai, S. Sawant, D. Reinhardt, S. Rastegarpour, S. Gros, A learning-based model predictive control strategy for home energy management systems, IEEE Access 11 (2023) 145264–145280, http://dx.doi.org/10.1109/ACCESS.2023.3346324.

[12] G. Han, H.J. Joo, H.W. Lim, Y.S. An, W.J. Lee, K.H. Lee, Data-driven heat pump operation strategy using rainbow deep reinforcement learning for significant reduction of electricity cost, Energy 270 (2023) 126913, http://dx.doi.org/10.1016/J.ENERGY.2023.126913.

[13] S. Schmitz, K. Brucke, P. Kasturi, E. Ansari, P. Klement, Forecast-based and data-driven reinforcement learning for residential heat pump operation, Appl. Energy 371 (2024) 123688, http://dx.doi.org/10.1016/J.APENERGY.2024.123688.

[14] D. Wang, W. Zheng, Z. Wang, Y. Wang, X. Pang, W. Wang, Comparison of reinforcement learning and model predictive control for building energy system optimization, Appl. Therm. Eng. 228 (2023) 120430, http://dx.doi.org/10.1016/J.APPLTHERMALENG.2023.120430.

[15] O.D. Somer, A. Soares, K. Vanthournout, F. Spiessens, T. Kuijpers, K. Vossen, Using reinforcement learning for demand response of domestic hot water buffers: A real-life demonstration, in: 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe, ISGT-Europe 2017 - Proceedings, vol. 2018-January (2017) 1–7, http://dx.doi.org/10.1109/ISGTEUROPE.2017.8260152.

[16] L.L. Custode, G. Iacca, A co-evolutionary approach to interpretable reinforcement learning in environments with continuous action spaces, in: 2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2021, http://dx.doi.org/10.1109/SSCI50451.2021.9660048.

[17] Finnish Meteorological Institute, Havaintojen lataus [dataset], 2024, URL https://www.ilmatieteenlaitos.fi/havaintojen-lataus. (Accessed 12 December 2024).

[18] ENTSO-E, ENTSO-E transparency platform, 2025, URL https://transparency.entsoe.eu/. (Accessed 30 April 2025).

[19] Helen, Open data [dataset], 2025, URL https://www.helen.fi/en/about-us/helen/about-us/open-data. (Accessed 20 March 2025).

[20] Ministry of the Environment (Ympäristöministeriö), 1010/2017 ympäristöministeriön asetus uuden rakennuksen energiatehokkuudesta, 2017, URL https://www.finlex.fi/en/legislation/collection/2017/1010.

[21] Finnish Real Estate Management Federation (Isännöintiliitto), Taloyhtiön Energia-ja Ilmastoasiat Isännöinnin Opas, Tech. rep., 2024, URL https://www.isannointiliitto.fi/wp-content/uploads/2024/10/taloyhtion_energia_ja_ilmastoasiat_isannoinnin_opas_10_2024.pdf.

[22] C. Bodnar, W.P. Bruinsma, A. Lucic, M. Stanley, A. Allen, J. Brandstetter, et al., A foundation model for the earth system, Nature (2025) 1180–1187, http://dx.doi.org/10.1038/s41586-025-09005-y.

[23] P.K. Kundu, I.M. Cohen, Fluid Mechanics, fourth ed., Elsevier, 2008, pp. 413–415.

[24] Verein Deutscher Ingenieure, VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen (GVC), VDI Heat Atlas, second ed., Springer Berlin Heidelberg, Dusseldorf, 2010, http://dx.doi.org/10.1007/978-3-540-77877-6,

[25] LLC Gurobi Optimization, Gurobi optimizer reference manual, 2025, Version 12.0. URL https://www.gurobi.com.

[26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017, http://dx.doi.org/10.48550/arXiv.1707.06347.

[27] M. Towers, A. Kwiatkowski, J. Terry, J.U. Balis, G.D. Cola, T. Deleu, et al., Gymnasium: A standard interface for reinforcement learning environments, 2024, URL https://arxiv.org/abs/2407.17032v3.

[28] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, N. Dormann, Stable Baselines3 [software], 2019, URL https://github.com/DLR-RM/stable-baselines3.

[29] W. Cai, A.B. Kordabad, S. Gros, Energy management in residential microgrid using model predictive control-based reinforcement learning and Shapley value, Eng. Appl. Artif. Intell. 119 (2023) 105793, http://dx.doi.org/10.1016/J.ENGAPPAI.2022.105793.