Preprints, 4th IFAC Conference on
Advances in Proportional-Integral-Derivative Control
Almería, Spain | June 12-14, 2024

A Comparison of Classical and Reinforcement Learning-based Tuning Techniques for PI controllers ⋆

Abad-Alcaraz, V. *,** **Castilla, M.** *,** **Álvarez, J.D.** *,**

* *CIESOL, Solar Energy Research Centre, University of Almería-ceiA3, Ctra. Sacramento s/n, La Cañada de San Urbano, Almería 04120, Spain.*
** *Department of Informatics, University of Almería- ceiA3, Ctra. Sacramento s/n, La Cañada de San Urbano, Almería 04120, Spain, (e-mail: vabadalcaraz@ual.es, mcastilla@ual.es, jhervas@ual.es)*

Abstract:
This study compares two tuning techniques for Proportional-Integral (PI) controllers. The first strategy uses the Pole-Zero Cancellation method, a well-established technique in the field of dynamic systems control. The second strategy introduces an innovative approach by using a reinforcement learning-based technique to adaptively tune a PI controller. To compare these tuning methodologies, a Heating, Ventilation and Air Conditioning (HVAC) control system was selected as a case study to guarantee users' thermal comfort. In both cases, the HVAC process was modelled as a first-order system without time delay. In addition, the performance of the proposed controllers analysed by evaluating temperature set-point tracking and disturbance rejection caused by the occupancy of people in the room. The results demonstrate that classical methods are efficient and quick to implement, while the use of RL also enables optimization of energy consumption and reduction of operating costs.

Keywords: PID control, Reinforcement learning, HVAC systems, Energy efficiency

1. INTRODUCTION

The Proportional-Integral-Derivative (PID) controller is recognised for its versatility and applicability in a wide range of industrial environments, see McMillan (2012). Its ability to reject unforeseen disturbances and unknown dynamics makes it an optimal choice in many scenarios. PID control is widely used in key sectors, including process industries, robotics, manufacturing, power electronics and biomedical engineering. The widespread adoption of PID controller has motivated continuous research and development in order to optimise its performance, maintaining a constant search for innovative tools and methods to further refine PID configuration and tuning as is emphasised in Dubey et al. (2022).

Classical tuning techniques for PID controllers are typically divided into two classes: frequency-domain and time-domain approaches. Nevertheless, it is important to note that many of these conventional methodologies fail to achieve an accurate tuning. Concretely, some of them require manual intervention, while others, rely on rigid preset configurations, such as the Ziegler-Nichols tuning method (Muresan and De Keyser (2022)). Therefore, a precise tuning process allows preventing unwanted phenomena such as overshoot and oscillations in the closed-loop system response. To address this issue, new techniques are being explored, including those that make use of Artificial Intelligence (AI). These strategies employ continuous and dynamic approaches using tools such as neural networks (Günther et al. (2020)), genetic algorithms, fuzzy logic, and optimisation algorithms (Ali et al. (2021)). The main objective of applying AI-based tuning methods is to automatically and precisely adjust the characteristic parameters of a PID controller in order to achieve better results in comparison to classical tuning techniques.

In this paper, a comparison between classical and Reinforcement Learning (RL)-based tuning techniques for first-order systems without time delay has been performed. For this purpose, the control of a Heating, Ventilation and Air Conditioning (HVAC) system to ensure healthy and comfortable environments for people has been selected as a case of study. HVAC systems rely on key equipment such as chillers and boilers, and thus, they are responsible for most of the energy consumed by buildings. Several studies, as López-Alonso et al. (2018), have concluded that the development of appropriate control architectures for HVAC systems is essential to maintain an optimal indoor air temperature and airflow conditions guaranteeing users' comfort and minimising both carbon dioxide emissions and energy consumption.

The use of classical tuning approaches, such as the Pole-Zero cancellation method, for Proportional-Integral (PI) controllers to manage HVAC systems have been used

⋆ This work is part of the I+D+i TED2021-131655B-I00 research project funded by AEI/10.13039/501100011033/ and "Unión Europea NextGenerationEU"

in an effective way, see Castilla et al. (2011). However, although PID control is widely accepted for its simplicity, techniques, such as reinforcement learning, that do not assume a single linear model as system dynamics, but, they can be extrapolated to more complex problems are being investigated, as it is shown in Shuprajhaa et al. (2022). These strategies allows including within the control decision users' preferences (Lei et al. (2022)), energy efficiency (Yu et al. (2019)) or fault detection (Matetić et al. (2023)).

The document is organised as follows: Section 2 includes a description of the case of study used in this work. In Section 3 the proposed control architecture and the used tuning techniques are widely explained. Section 4 is devoted to the obtained results. Finally, Section 5 summarises the main conclusions.

2. CASE OF STUDY: THE CIESOL BUILDING

The CIESOL building (http://www.ciesol.es) is a solar energy research centre located within the campus of the University of Almería (southeastern Spain). This research centre was built following bioclimatic architecture criteria and it also counts with some active strategies, such as a HVAC system based on solar cooling, see Pasamontes et al. (2009). In particular, this HVAC system uses a solar collector field, a hot water storage system, a boiler, and an absorption machine with its cooling tower to provide both, hot and chilled water, to the fancoils units located in all the rooms of the building. Furthermore, the CIESOL research centre has a wide network of sensors distributed throughout the building.

More in detail, the fancoil unit available at each room of the CIESOL building, see Fig. 1, can be categorised as a Multiple Inputs, Single Outputs (MISO) system. Specifically, it is possible to regulate the temperature of the air introduced into the room by controlling the amount of water which flows through the fancoil unit (q_W) and the fan velocity (V_{fan}). However, in this paper, it has been considered that the only control variable available is V_{Fan} maintaining the water flow at the maximum possible value.

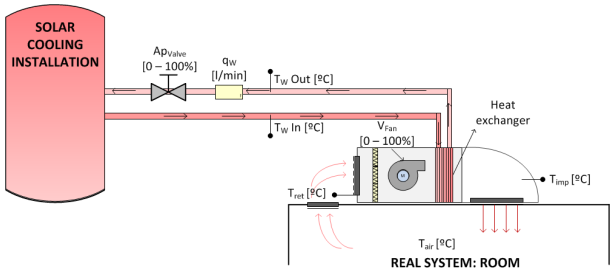

Fig. 1. Fancoil unit. Source: Castilla et al. (2014)

Therefore, a nominal linearized model at a typical operation point has been obtained by means of classical identification techniques. This model represents the indoor air temperature dynamics ($°C$) as a function of fan velocity (%). Equation 1 shows the continuous-time first-order without time delay transfer function in Laplace variable s for winter operating mode:

$$G(s) = \frac{Y_T(s)}{U(s)} = \frac{k}{\tau s + 1}; \quad \{ k = 0.0755,\ \tau = 44.17 \quad (1)$$

where k represents the static gain expressed in ($°C/\%$) and τ is the time constant in minutes.

In addition, it should be noted that the proposed case of study is also subjected to disturbances. Concretely, it has been considered that the main source of disturbances is people. The movement of people among the different rooms of the building can affect thermal comfort, as the presence of human beings and their activities impact into the indoor ambient temperature.

3. CONTROL ARCHITECTURE

PID control is one of the strategies most widely used in the industry. Tuning the parameters of a PID controller triggers effective control action adapted to the specific needs of the process. The characteristic equation of PID controller is defined as it is shown in equation (2):

$$u(t) = K_p \cdot e(t) + K_i \cdot \int_0^t e(\tau)\, d\tau + K_d \cdot \frac{de(t)}{dt} \quad (2)$$

where $u(t)$ symbolises the control signal at a given time t, $e(t)$ represents the error between the desired value of a process variable and its measured value at a certain time. Finally, K_p, K_i and K_d are the proportional, integral and derivative gains respectively.

However, as the *Derivative* term of a PID controller may be affected by system fluctuations or noise, it has been decided to use a Proportional-Integral (PI) controller. More in detail, in this work, a comparison of different tuning techniques for PI controllers will be performed. Therefore, the main control objective is to manage the Fancoil unit described in Section 2 to guarantee an appropriate indoor air temperature for the users of a building, see Fig. 2. To do that, different tuning techniques have been used to obtain K_p and K_i parameters of a PI controller. They can be classified into classical tuning techniques and RL-based ones.

3.1 Classical tuning techniques

As it was commented before, the selected case of study has been modelled as a first-order system without time delay. Therefore, the Pole-Zero Cancellation method has been chosen as classical tuning technique, see Åström and Hägglund (1995). This method allows the design of a PI controller ensuring the stability of the system by strategically cancelling poles and zeros in the closed-loop transfer function. Besides to mitigating oscillations, it is also possible to establish the closed-loop time response during the tuning process.

In particular, two PI controllers tuned using the Pole-Zero cancellation method have been designed. First, a conservative approach with a closed-loop time constant (τ_{bc}) equals to 0.9 times the open-loop time constant (τ), that is, a proportional gain of $K_p = 18.92\ \%/°C$ and an integral time equals to $T_i = 44.17\ min$. Second, an aggressive approach using $\tau_{bc} = 0.7\tau$, that is, the proportional gain and the integral time parameters have been defined as $K_p = 14.62\ \%/°C$ and $T_i = 44.17\ min$ respectively.

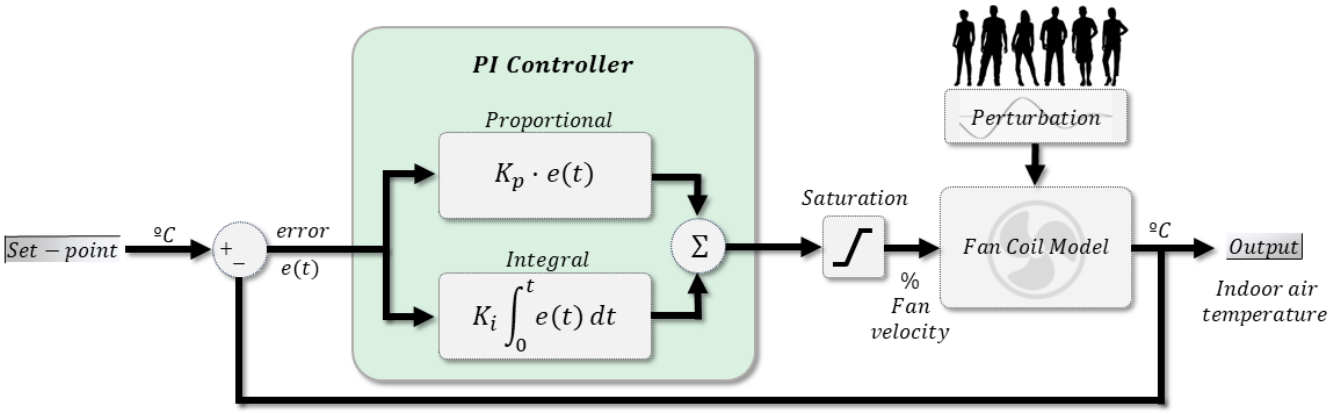

Fig. 2. Proposed control architecture

3.2 Reinforcement Learning-based tuning techniques

RL is a field within Machine Learning (ML) that is inspired by how living organisms learn by interacting with the environment around them. RL stars with an initial observation, from which, the agent makes a decision and applies an action to the environment. Subsequently, the environment changes its current state and the agent obtains a new state together with its associated reward. The purpose of RL is to decide the action that maximises the obtained reward. The RL-based architecture used to tune of a PI controller can be observed in Fig. 3. Therefore, the main components of a RL-based tuning approach are: observations, environment, agent, reward and policy.

Observations. They gather information about the environment current state that will be provided to the agent. In the proposed architecture, the PI controller tuning parameters (K_p and K_i) and the disturbance caused by the occupancy of people in the room have been selected as observations to be calculated from the error signal.

Environment. It represents the system with which the agent interacts. In this paper, the environment represents the case of study presented in Section 2.

Agent. It is a key component in the RL framework which is responsible for making decisions and taking actions. Therefore, an agent can be defined as an intelligent entity that interacts with the environment, make decisions to maximise rewards and learns from experience. An agent can be divided into the following elements: the learning algorithm and the policy. The first one, is the method used by the agent to optimise its policy. On the other hand, the policy is the function used to select actions based on the observations and the environment. In literature, it is possible to find different types of agents for continuous action space, for example, Deep Determinist Policy Gradient (DDPG), Twin-Delayed Deep Deterministic Policy Gradient (TD3) or Soft Actor-Critic (SAC).

In this work, a TD3 agent has been selected. It is an actor-critic RL agent that looks for an optimal policy able to maximise a cumulative reward, see MathWorks (2024). More in detail, the actor uses a feed-forward neural network to map observations from the environment to specific actions. Besides, the critic uses two neural networks to approximate Q-value functions and evaluate the actions taken in a specific state. These neural networks consist of fully connected layers able to process observations and actor actions. The Rectified Linear Unit (ReLU) activation function is used in these layers to introduce a non-linearity in order to improve model representability.

Reward. It is a function which provides a number as a function of the observations, the current state of the environment and the action decided by the agent. Hence, a reward function allows the learning algorithm to recognise when its policy is improving and ultimately converging on the desired outcome.

In the RL paradigm, flexibility in the definition of reward functions is critical. For instance, the reward function can evaluate set-point tracking or also evaluate the stability of the PI controller and the rejection of disturbances. However, this freedom can also lead to situations where the reward is scarce. To evaluate the flexibility and performance of reward functions, in this paper, three different reward functions have been proposed:

The band-limiting strategy. This reward function restricts the acceptable indoor air temperature within the range $\pm 1\ ^{\circ}C$ around the reference temperature, see equation 3. This approach aims to guide the agent towards solutions within this specific range, avoiding sparse training and focusing on reaching the desired indoor air temperature.

$$f(x) = \begin{cases} R = R - C1 \cdot (T - (Rf - 1))^2 & if\ \ T < Rf - 1 \\ R = R - C1 \cdot (T - (Rf + 1))^2 & if\ \ T > Rf + 1 \\ R = R + C2 & if \quad T = Rf \\ R = R + C2 \cdot min(1, (1/error)^2) & \\ \qquad\qquad if \quad (Rf - 1) \leq\ T\ \geq (Rf + 1) & \end{cases} \tag{3}$$

where R is the reward, T is the current temperature in $(^{\circ}C)$, Rf the imposed reference temperature expressed in $(^{\circ}C)$, $error$ is the difference between the desired and the current temperatures in $(^{\circ}C)$. Finally, $C1$ and $C2$ are the penalty and reward coefficients, with values of 0.01 and 0.005, respectively.

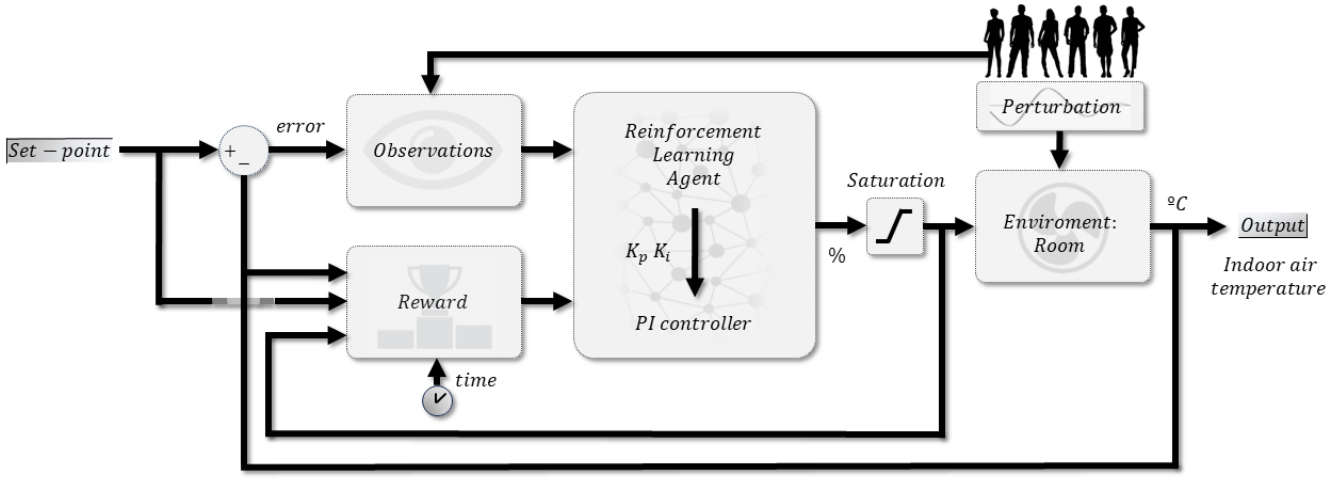

Fig. 3. Reinforcement Learning based architecture used to tune a PI controller

The disadvantage of this function is that the agent may choose to maintain temperatures slightly above or below the desired temperature. Various attempts have been made to mitigate this effect, such as adjusting the coefficients C1 and C2, modifying the training hyperparameters, or increasing the training duration. However, the agent prioritises obtaining an acceptable immediate reward within the imposed constraints, rather than achieving the best possible reward. This situation also has an impact on the ability of the agent to efficiently reject perturbations.

Penalties based on control signals and closed-loop time response. This approach intends to raise the controller's awareness about two key issues: the adverse effects of selecting high control signals, and the importance of the time taken to reach the set-point. Besides, it is essential to penalise the set-point tracking error as a percentage basis, particularly when dealing with very low references, as the system may underestimate its significance in achieving the set-point.

The proposed reward function incorporates a time variable, t, and adjusts a coefficient $C3$ to prevent excessively high reward values that could cause the agent to overlook minimal improvements. As for the control signals, they will only start to negatively affect the agent when the tracking error of the set-point is less than 40% as it indicates to the agent that it is approaching the target and a control signal should decrease.

These adjustments aim to provide to the agent a more balanced and accurate understanding of the impact of selecting high control signals and the value of time along the learning process. Taking into account that $C3$ is 0.001, this reward function can be defined as it is shown in (4) and (5):

$$rate = \frac{error \cdot 100}{(Rf - T_o)} \tag{4}$$

$$f(x) = \begin{cases} R = -C3 \cdot |rate| \cdot t & if\ rate \geq 40 \\ R = -C3 \cdot |rate| \cdot t \cdot u & if\ rate < 40 \end{cases} \tag{5}$$

where T_o is the initial indoor air temperature in ($^\circ C$).

This strategy allows the agent to follow a temperature set point very quickly and to reject any disturbances. The use of the control signal restricts its action when the system is near the imposed temperature reference. However, we observed an excessive constraint that limited the agent's freedom and could lead to significant stability problems during training.

Stability penalties. System stability is a priority to prevent abrupt oscillations and to mitigate the effects of disturbances. This approach is developed considering that the system should approach the desired setpoint from the first 40 seconds. Up to this point, the only relevant objective to consider in the reward function is the difference between the desired setpoint and the actual temperature. However, after 40 seconds, the difference between the current error and the error at the previous instant is also rewarded, which means that the agent must avoid sudden changes in temperature.

Another aspect that has been refined through experimentation is the definition of a narrow margin of error in percentage terms. In this sense, the maximum error allowed, without considering the penalty for lack of stability, is set at 2%. This strategy ensures that the agent does not exclude small improvements in the reward

The function defined in (6) and (7) allows the system to make less abrupt decisions, even when disturbances are introduced. The stability penalty is limited by the empirically derived coefficient C4, which has a value of 10.

$$rate = \frac{error \cdot 2}{(Rf - T_o)} \tag{6}$$

$$f(x) = \begin{cases} rate = 2 & if\ \ rate < 0 \\ rate = rate + C4 \cdot |error - error_{-1}| & if\ time \geq 40 \\ R = -|rate| & In\ all\ cases \end{cases} \tag{7}$$

where $error_{-1}$ is the difference between the desired temperature and the temperature in the previous step in ($^\circ C$).

4. RESULTS AND DISCUSSION

In this section, the performance of the proposed PI controllers will be evaluated using the case of study proposed in Section 2. To this end, the initial temperature of the room (T_o) has been established to $18°C$ and the reference temperature (Rf) to $22°C$.Moreover, it has been assumed that during the simulation tests the room was occupied on the basis of the occupancy profile given by the Fig. 4.

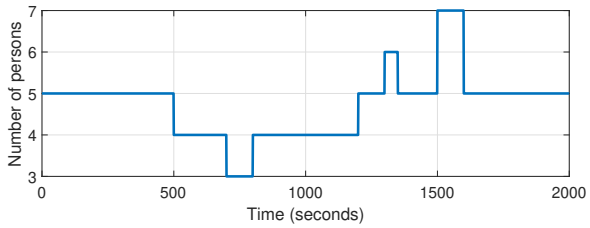

Fig. 4. Occupancy disturbance during simulation tests

First, a detailed analysis of the performance of PI controllers tuned using RL has been carried out. The main objective of this analysis is to decide which reward function provides better results as a function of set-point tracking error and stability. The obtained results are shown in Fig. 5. In particular, the upper graph shows the indoor air temperature and the bottom graph depicts the control signal, that is, the fan velocity in (%).

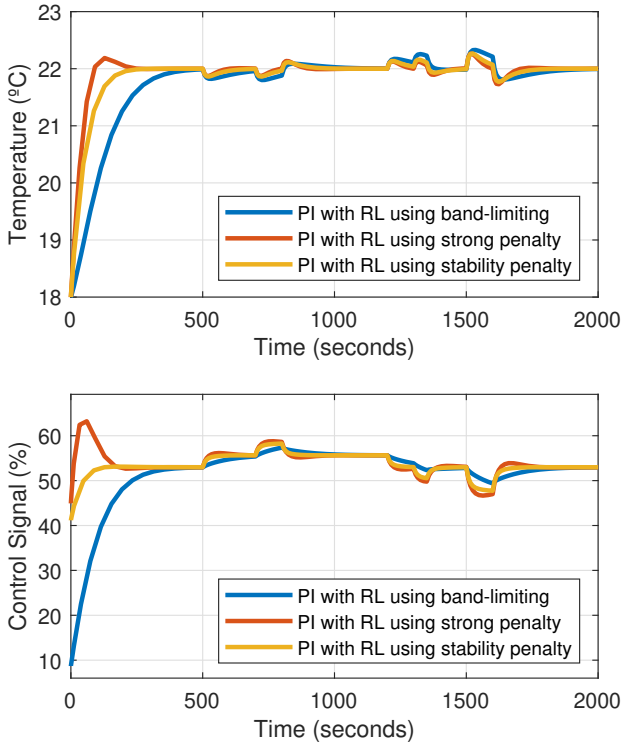

Fig. 5. PI controllers tuned with RL. Evaluation of reward functions

As it can be observed, the band-limiting reward function approach (blue line) is insufficient as it can be interpreted that it is not necessary to reach the desired set-point in order to obtain an acceptable reward. During the training stage, this leads to a high settling time or, in other case, to a agresive response characterised by continuous oscillations around the set-point. On the other hand, the second approach, which estimates penalties based on the closed-loop time response and the selected control signals (red line) shows an overly aggressive behaviour. This approach tends to seek greater rewards by reaching the set-point as quickly as possible, taking a considerable risk on the control signal during the first moments of operation, when the penalty time is minimal. Finally, as it was commented before, the third reward function was designed taking into consideration the limitations of the previous reward functions. Figure 5 shows how this approach provides the best results (yellow line). The stable strategy reward function is able to reach the set-point quickly using a smoother control signal despite the disturbances caused by people.

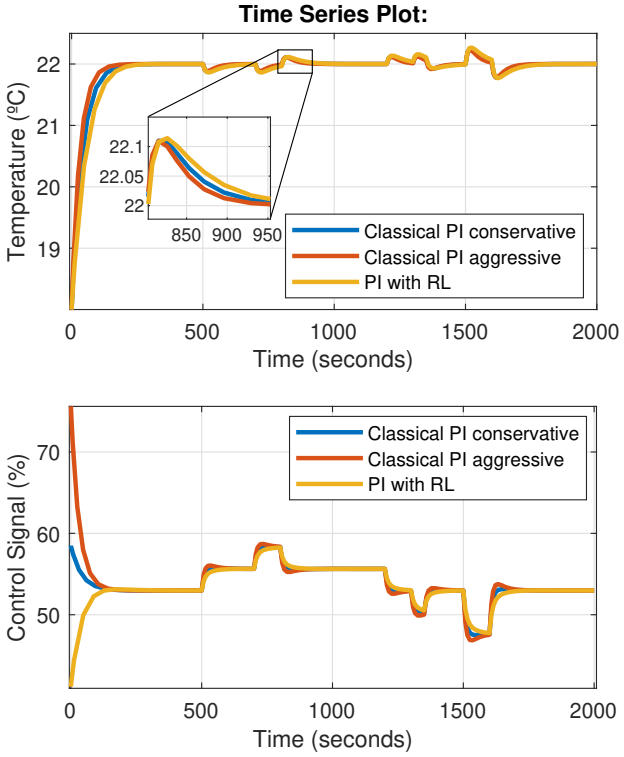

Fig. 6. Comparison of classical and RL tuned PI controllers

Therefore, the PI controller tuned by RL and that makes use of the stable penalties reward function has been selected. The obtained controller parameters are $K_p = 10.28\%/°C$ and $T_i = 0.02\ min$. The non-aggressive nature of the reward indicates a preference for stability over speed. This approach is considered effective as it encourages a non-aggressive behaviour in the control signal, guarantees to reach the reference, and enables swift rejection of external disturbances. Subsequently, it will be compared with the PI controllers tuned using classical approaches, that is, by using the Pole-Zero Cancellation method.

Figure 6 shows a comparison among the PI controllers tuned by both classical and RL-based methods. In particular, the upper graph shows the evolution of indoor air temperature and the bottom graph represents the Fan velocity. From these graphs, it can be inferred that the evolution of indoor air temperature temperature obtained for the three PI controllers are closely matched. However,

significant differences related to the control signal can be observed. PI controllers tuned using classical methods provides more aggressive control signals compared to the other PI controller at the cost of minimally decrease the time they take to reach the set-point. Therefore, in terms of efficiency, the PI controller tuned using RL is more favourable because the control signal will gradually increase until the set-point temperature is reached avoiding abrupt changes. In addition, all the approaches show an appropriate response to disturbances rejection.

Table 1. Performance evaluation of control systems with IAE, ISE, ITAE and IAVU methods

	IAE	ISE	ITAE	IAVU
Classical PI conservative	234.0	356.2	90689.5	24.0
Classical PI aggressive	184.4	275.9	72983.5	41.2
PI with RL	295.4	482.6	112206.6	28.2

Finally, an analysis of the obtained results based on some performance indexes such as the Integral Absolute Error (IAE), the Integral Square Error (ISE), the Integral Time-weighted Absolute Error (ITAE) and the Integral of the Absolute Value for the Variation of the Control Signal (IAVU) has been performed. For the three indexes related to set-point tracking errors, the best results are provided by the classical PI aggressive and the worst ones by the PI controller tuned using RL. Nevertheless, for the index related to control effort, worst results are provided by the classical PI aggressive. This is consistent with the graphical results.

Furthermore, a comparative analysis was conducted based on the energy consumption of each control approach, taking into account the average consumption of the fan-coil at each time interval. The results indicate that the RL-tuned PI controller achieves a 1% energy saving compared to the most energy-consuming approach, namely the aggressive classical PI controller. It demonstrates the effectiveness of using reinforcement learning techniques in control systems to optimize energy consumption and decrease operating expenses. This effectiveness can be improved by modifying the reward function to include energy efficiency.

5. CONCLUSIONS

In this work, a comparison of tuning methodologies for PI controllers has been performed. More in detail, the Pole-Zero cancellation method has been compared to reinforcement learning technique. In addition, a real case of study, the control of a HVAC system to ensure users' thermal welfare has been defined in order to evaluate the goodness of the proposed PI controllers. The obtained results show that PI controllers tuned using classical methods are highly effective and can be implemented quickly, particularly in applications where stability and speed of response are important. However, for more complex and multivariable models, reinforcement learning may be a better alternative, despite its lengthy training process.

As future works, methods to improve the learnability and stability of reinforcement learning will be explored. Additionally, this technique will be applied to more complex environments, particularly those where tuning a controller is challenging by using classical methods.

REFERENCES

Ali, M., Firdaus, A.A., Arof, H., Nurohmah, H., Suyono, H., Putra, D.F.U., and Muslim, M.A. (2021). The comparison of dual axis photovoltaic tracking system using artificial intelligence techniques. *IAES Int. J. Artif. Intell*, 10(4), 901.

Åström, K.J. and Hägglund, T. (1995). Pid controllers: Theory, design, and tuning.

Castilla, M., Álvarez, J., Berenguel, M., Rodríguez, F., Guzmán, J., and Pérez, M. (2011). A comparison of thermal comfort predictive control strategies. *Energy and buildings*, 43(10), 2737–2746.

Castilla, M., Álvarez, J., Normey-Rico, J., and Rodríguez, F. (2014). Thermal comfort control using a non-linear mpc strategy: A real case of study in a bioclimatic building. *Journal of Process Control*, 24(6), 703–713.

Dubey, V., Goud, H., and Sharma, P.C. (2022). Role of pid control techniques in process control system: a review. *Data Engineering for Smart Systems: Proceedings of SSIC 2021*, 659–670.

Günther, J., Reichensdörfer, E., Pilarski, P.M., and Diepold, K. (2020). Interpretable pid parameter tuning for control engineering using general dynamic neural networks: An extensive comparison. *Plos One*, 15(12), e0243320.

Lei, Y., Zhan, S., Ono, E., Peng, Y., Zhang, Z., Hasama, T., and Chong, A. (2022). A practical deep reinforcement learning framework for multivariate occupant-centric control in buildings. *Applied Energy*, 324, 119742.

López-Alonso, M., Álvarez, J.D., Guzmán, J.L., and Berenguel, M. (2018). Nonlinear control of a fan-coil operation. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, 000213–000218. IEEE.

Matetić, I., Štajduhar, I., Wolf, I., and Ljubic, S. (2023). Improving the efficiency of fan coil units in hotel buildings through deep-learning-based fault detection. *Sensors*, 23(15), 6717.

MathWorks (2024). Twin-Delayed Deep Deterministic (TD3) Policy Gradient Agents. https://es.mathworks.com/help/reinforcement-learning/ug/td3-agents.html Last accessed. 9th January 2024.

McMillan, G.K. (2012). Industrial applications of pid control. *PID control in the third millennium: Lessons learned and new approaches*, 415–461.

Muresan, C.I. and De Keyser, R. (2022). Revisiting ziegler–nichols. a fractional order approach. *ISA transactions*, 129, 287–296.

Pasamontes, M., Álvarez, J.D., Guzmán, J.L., and Berenguel, M. (2009). Hybrid modeling of a solar cooling system. *IFAC Proceedings Volumes*, 42(17), 26–31.

Shuprajhaa, T., Sujit, S.K., and Srinivasan, K. (2022). Reinforcement learning based adaptive pid controller design for control of linear/nonlinear unstable processes. *Applied Soft Computing*, 128, 109450.

Yu, L., Xie, W., Xie, D., Zou, Y., Zhang, D., Sun, Z., Zhang, L., Zhang, Y., and Jiang, T. (2019). Deep reinforcement learning for smart home energy management. *IEEE Internet of Things Journal*, 7(4), 2751–2762.