Reinforcement Learning for a Parabolic Trough Solar Collector

Marta Leal[1,2], Verónica Abad-Alcaraz[1,2], José Domingo Álvarez[1,2] and María del Mar Castilla[1,2]

Abstract— **Parabolic Trough Solar Collectors (PTCs) are a prevalent technology used to provide heat in industrial processes. The primary objective of the control system is to regulate the Heat Transfer Fluid (HTF) temperature to a target value, despite external disturbances. In this study, a Reinforcement Learning (RL) based control system is developed to ensure accurate tracking of the temperature reference at the PTC output, considering the disturbances introduced by solar irradiance. Furthermore, the performance of the proposed system is compared with classical control strategies such as Proportional-Integral-Derivative (PID) and feedforward control, aiming to enhance disturbance rejection and overall performance. The results demonstrate the potential of RL-based control systems for managing complex and non-linear systems.**

I. Introduction

The heat sector remains the largest end-use category, accounting for nearly 50 % of total final energy demand and approximately 40 % of CO_2 emissions linked to energy production in 2023. Between the year 2017 and 2023, global heat demand increased by 7 % (+14 EJ). However, modern renewable sources only covered a portion of this increase, resulting in a 5 % rise in heat-related CO_2 emissions, mostly driven by industrial processes. Looking forward, the use of renewable sources for heat is projected to increase by more than 50 % (an increase of over 15 EJ) between 2024 and 2030 [1]. Among renewable heat sources, solar thermal technologies have achieved the necessary maturity to be reliable and competitive across diverse applications, ranging from power generation and regional heating networks to self-sufficient heat and electricity for buildings, industries, and isolated communities.

Parabolic Trough Solar Collector (PTC) is one of the most predominant technologies used to provide heat to industrial processes [2]. These systems are typically employed for medium-range thermal processes, generally operating between 100 °C and 250 °C, though they can also reach temperatures up to 400 °C under certain conditions. PTCs concentrate direct solar radiation onto a receiver pipe located along their focal line. A Heat Transfer Fluid (HTF), such as synthetic oil or water, flows through the pipe, absorbing the concentrated energy to increase its enthalpy.

The primary objective of the control system in a distributed collector field is to regulate the average temperature of the HTF to a target value, despite external disturbances. Among these disturbances, irradiance significantly influences the performance of PTCs in solar energy systems. Conventional control strategies that rely on simplified global parameters often fail to address the distributed characteristics of collectors and the impact of irradiance on the dynamics of the system. Feedback control has been shown to be more effective than open-loop control, yet it is unable to entirely mitigate the adverse effects of disturbances on system output. To overcome this limitation, feedforward control is implemented to anticipate and counteract disturbances before they affect the process, and this method is now commonly integrated in distributed control systems and is widely used in basic control applications to enhance performance [3]. However, due to the stochastic nature of solar irradiance, feedforward controllers reduce their performance by counteracting this disturbance.

Reinforcement Learning (RL) allows for improved responsiveness of the control loop to dynamic disturbances such as irradiance fluctuations. Moreover, compensating the disturbance through flow rate adjustments modifies the system dynamics, which are strongly nonlinear. This nonlinear behaviour can be managed more effectively with RL-based controllers compared to traditional Proportional-Integral-Derivative (PID) controllers, offering a more adaptive and robust approach to disturbance rejection.

Solar energy systems are typically costly relative to the energy they generate. Therefore, enhancing their performance through the implementation of advanced control techniques could contribute to positioning them as a practical alternative to traditional energy sources [4]. RL is a subfield of Machine Learning (ML) inspired by the way organisms learn through interactions with their environment. The main objective of RL is to identify actions that maximize the cumulative reward. In the context of solar thermal systems, various RL algorithms have been explored. For example, a recent study [5] proposed a model-free deep RL approach using the Soft Actor-Critic (SAC) algorithm to optimize heliostat aiming strategies. In another study [6], the challenges of controlling Solar Collector Fields (SCFs) using complex models were addressed by introducing an adaptive model-free controller based on the Q-Learning algorithm. Additionally, in [7], a RL-based control model is implemented for a solar thermal cooling system powered by linear Fresnel collectors, tailored to the specific case study presented in their research.

The main objective of this study is to develop a control

*This work has been funded by the National R+D+i Plan Project PID2021-126889OB-I00 of the Spanish Ministry of Science and Innovation and EIE funds, and TED2021-131655B-I00 research project funded by AEI/10.13039/501100011033 and by the "European Union NextGenerationEU".

[1] All authors are with CIESOL, Solar Energy Research Centre, University of Almería - ceiA3, Almería, Spain.

[2] All authors are with Department of Informatics, University of Almería - ceiA3, Ctra. Sacramento s/n, La Cañada de San Urbano, 04120 Almería, Spain. {marta.leal, vabadalcaraz, jhervas, mcastilla}@ual.es

system based on RL to ensure accurate tracking of the temperature reference at the output of a PTC, taking into account the disturbances introduced by irradiance. In addition, the performance of the proposed RL control system is compared with classical control strategies such as PID and feedforward control, aiming to enhance disturbance rejection and overall performance.

The structure of this paper is organised as follows. Section II introduces the non-linear model employed in the case study. Section III examines the different control strategies considered. Section IV presents the results obtained from applying the proposed approach. Lastly, Section V outlines the main conclusions and suggests possible avenues for future research.

II. Case of Study

The case study investigated in this work, involves a distributed solar collector field utilising PTC technology. A solar collector acts as a large-scale heat exchanger, and this system is prevalent in the process industry. Consequently, the expertise acquired in managing solar collector fields can be applied to a variety of standard industrial processes [8]. Fig. 1 shows the temperatures involved in the energy balance of the absorber tube of a parabolic trough collector.

Under typical assumptions and hypotheses, the distributed solar collector field can be described by a distributed parameter model based on temperature. The dynamics of the solar collector field are governed by a system of Partial Differential Equations (PDEs) representing the energy balance (1). Applying the law of energy conservation on an interval of length Δx and time interval dt, the following equation is obtained for the fluid that circulates through the pipe:

$$\begin{aligned} &A_i \rho C \frac{\partial T}{\partial t}(x,t) + \dot{q} \rho C \frac{\partial T}{\partial x}(x,t) \\ &= \pi D_i h_i \cdot (T_\omega(x,t) - T(x,t)) \end{aligned} \tag{1}$$

where $T(x,t)$ is the fluid temperature and $T_\omega(x,t)$ is the pipe wall temperature. To simplify the partial derivative with respect to length, the relationship $\frac{\partial T}{\partial x} \approx \frac{T(L,t)-T(0,t)}{L}$ can be applied, resulting in (2):

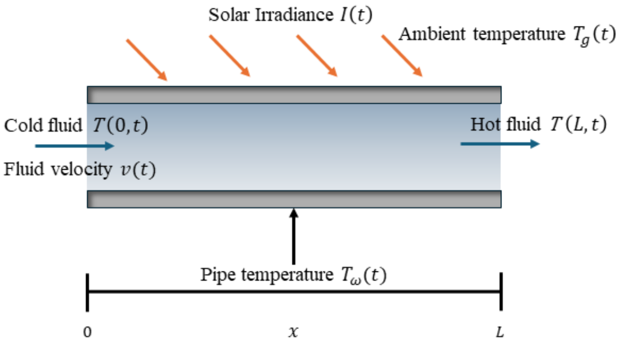


Fig. 1: Absorber tube diagram of a parabolic trough collector

$$\begin{aligned} A_i \rho C \frac{\partial T}{\partial t}(x,t) + \dot{q} \rho C \frac{T(L,t) - T(0,t)}{L} \\ = \pi D_i h_i \cdot (T_\omega(x,t) - T(x,t)) \end{aligned} \tag{2}$$

Similarly, an energy balance applied to the pipe wall of the collector yields is in (3) :

$$\begin{aligned} \rho_\omega C_\omega A_o \frac{\partial T_\omega}{\partial t}(x,t) = I \eta_o G \\ -\pi D_o h_o \left(T_\omega(x,t) - T_g(x,t)\right) \\ -\pi D_i h_i \left(T_\omega(x,t)) - T(x,t)\right) \end{aligned} \tag{3}$$

where $T_g(x,t)$ represents the ambient temperature.

Table I provides the definitions for the parameters used in (1), (2) and (3) is presented in Table I. Due to space limitations, this work does not include the entire mathematical development. However a detailed description of the plant along with a detailed presentation of the physical parameters can be found in [4].

The model is regarded as semi-physical, as it incorporates established knowledge of the system's dynamics. It adopts a hybrid approach that combines empirical observations with theoretical principles, featuring adjustable parameters that permit a physically meaningful interpretation. The simulations conducted in this study are based on the non-linear model of the collector set out in (2) and (3). The inputs to the non-linear model include the fluid temperature, ambient temperature, solar irradiance, and fluid velocity ($T(0,t), T_g(t), I(t)$ and $v(t)$ respectively). The output variable is the fluid temperature at the collector outlet ($T(L,t)$). This study concentrates on the dynamic relationship between solar irradiance and fluid velocity, assuming that the remaining disturbances remain approximately constant during the operation of the system.

TABLE I
MODEL PARAMETERS OF A SOLAR TUBE HEAT EXCHANGER

Parameter	**Description**	**Unit**
t	Time	s
x	Space	m
η_0	Optical efficiency of the collector	–
ρ	Fluid density	$\mathrm{kg/m^3}$
ρ_ω	Pipe density	$\mathrm{kg/m^3}$
A_i	Inner pipe area	$\mathrm{m^2}$
A_o	Outer pipe area	$\mathrm{m^2}$
C	Fluid specific heat capacity	$\mathrm{J/(kg^\circ C)}$
C_ω	Pipe material heat capacity	$\mathrm{J/(kg^\circ C)}$
D_i	Inner pipe diameter	m
D_o	Outer pipe diameter	m
G	Opening of the collector	m
h_i	In-pipe convective HTC	$\mathrm{W/(m^2 {}^\circ C)}$
h_o	Convective HTC outside pipeline	$\mathrm{W/(m^2 {}^\circ C)}$
$I(t)$	Solar irradiance	$\mathrm{W/m^2}$
L	Pipe length	m
$\dot{q}(t)$	Volumetric flow	$\mathrm{m^3/s}$
$T(x,t)$	Fluid temperature	$^\circ$C
$T_g(t)$	Ambient temperature	$^\circ$C
$T_\omega(x,t)$	Pipe temperature	$^\circ$C
$v(t)$	Fluid velocity	m/s

[a]HTC: Heat transfer coefficient.

III. CONTROL ARCHITECTURE

A. PID Control

This study compares three control architectures. The first is a Proportional-Integral (PI) controller tuned as described in this section.

The only control variable is the fluid velocity, $v(t)$, while the control parameter is the volumetric flow, $\dot{q}(t)$, obtained by multiplying $v(t)$ by the inner pipe area, A_i .

To design the PI controller, it is essential to obtain the transfer function that relates the volumetric flow, $\dot{q}(s)$, to the outlet temperature of the absorber tube, $T(L,s)$. A linear approximation of the non-linear model from (2) and (3) is obtained using Taylor series development, followed by a Laplace transform to determine the transfer functions relating $T(L,s)$ to $\dot{q}(s)$. The parameters of these transfer functions have been identified through reaction curve tests conducted in simulations of the linearised model and are presented in (4). A comprehensive analysis is provided in the referenced work, and the resulting transfer function is presented in [8].

$$\frac{T(L,s)}{\dot{q}(s)} = \frac{k_g}{\tau_g s + 1} \tag{4}$$

where k_g refers to the static gain in $\mathrm{m}^3/(\mathrm{s}°\mathrm{C})$, while τ_g is the time constant in s.

The tuning method selected, given that this relation is modelled as a first-order system without time delay, is the Pole-Zero Cancellation method, which is a classical tuning technique presented in [9] .

B. Feedforward Control

The second approach consists of integrating a feedforward component into the design of the PI controller. The proposed feedforward structure is a lead-lag configuration, as given by (5).

$$F_{ff}(s) = \frac{k_d}{k_g}\frac{\tau_g s + 1}{\tau_d s + 1} \tag{5}$$

where k_d refers to the static gain in $\mathrm{W}/(\mathrm{m}^2°\mathrm{C})$, while τ_d is the time constant in s obtained during the open-loop analysis and represents the transfer function that relates the solar irradiance to the outlet temperature.

This feedforward strategy is based on classical process control principles, as described in [10].

C. Reinforcement Learning

The basic functioning of RL can be described as follows: first, an observation is elicited and sent to the agent. The agent analyses the information, makes a decision and performs an action that affects the environment. Then, following the agent's action, the environment changes state, and the agent receives the new state along with a corresponding reward. The main objective is that the agent learns to maximise the reward received. Fig. 2 shows the control architecture created to manage the volumetric flow of the PTC. The main elements and their configuration are detailed below.

Observations. The main state variables of the environment that are relevant to the agent's decision are collected. In this implementation, five key elements are defined as observations:

- Error integral: Cumulative indicator that measures the sustained deviation between the collector output temperature and the reference value.
- Instantaneous error: Direct difference between the actual collector output temperature and the desired setpoint.
- Error derivative: The rate of change of the error over time, useful for anticipating trends in system dynamics.
- Collector outlet temperature: Variable directly related to collector performance and the main control objective.
- Incident solar radiation: Main disturbance of the system, its inclusion as an observed variable allows the agent to develop active disturbance rejection strategies, improving the stability and performance of the controller.

These observations are extracted and processed in real time to provide a comprehensive description of the dynamic state of the system.

Environment. This is the physical and dynamic system with which the RL agent interacts. In this instance, the non-linear parabolic solar collector outlined in Section II is used.

Agent. This framework is at the core of RL, and its purpose is to make decisions and execute actions in the environment to maximise cumulative reward [11]. In this study, an agent based on the Deep Deterministic Policy Gradient (DDPG) algorithm was implemented. This algorithm was selected for its ability to handle continuous action spaces and its relative simplicity in hyperparameter tuning. In addition, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm was investigated as a potential alternative, offering a valuable point of comparison.

The DDPG agent uses an actor-critic architecture: the actor generates actions based on observations, while the critic evaluates these actions using Q-value estimations. The critic network is updated via gradient descent to refine its Q-value predictions, improving the feedback it provides to the actor. In contrast, the TD3 algorithm introduces twin critics—identical to the one depicted— to mitigate overestimation bias and incorporates delayed policy updates to enhance stability. While both algorithms exhibit strengths, their performance can vary depending on the specific dynamics of the environment.

Reward. A reward function has been designed to guide the agent towards optimal behaviour, taking into account both the error with respect to the target and the system's constraints. From the perspective of safety and optimisation of training time, a temperature range between 100 °C and 240 °C has been established. This range is designed to prevent the system from being exposed to temperatures that exceed its operational limits and to ensure that it does not deviate from its expected behaviour. Should the temperature fall outside this range, the reward is reduced by $Bd = -300$ and the episode is terminated.

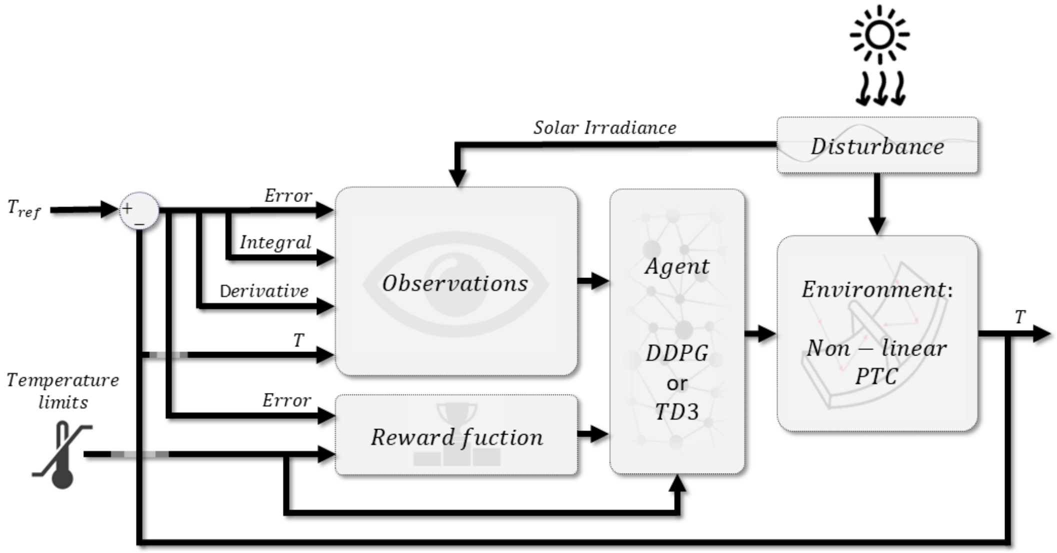


Fig. 2: RL based control architecture for PTC

Within the admissible temperature range, if the absolute error (the difference between the desired temperature and the collector temperature) is less than or equal to 0.15 °C, a reward of 10 is given. Should the absolute error exceed 0.15 °C, the reward is reduced to -1. Expressions (6) and (7) provide a suitable description of the reward:

$$\mathrm{RF}(t) = \begin{cases} 10 + Bd & \text{si } T_{\text{ref}}(t) - T(x,t) < 0.15 \\ -1 + Bd & \text{si } T_{\text{ref}}(t) - T(x,t) \geq 0.15 \end{cases} \tag{6}$$

$$Bd = \begin{cases} 0 & \text{si } T(x,t) \in [100, 240] \\ -300 & \text{si } T(x,t) \notin [100, 240] \end{cases} \tag{7}$$

IV. Results

After defining the control architecture, the implementation was carried out using the RL Designer Toolbox in MATLAB [12]. Two case studies are proposed to evaluate the performance of the control strategies. In the first case, a comparative analysis is conducted between two different types of RL agents: DDPG and TD3. In the second case, the performance of the DDPG based controller is compared with traditional control strategies, including a PI controller and a PI controller enhanced with classical feedforward compensation. To train the different agents, a sample time of 10 seconds for calculating the reward has been considered. The training process will simulate up to 3000 episodes to ensure sufficient exploration and convergence of the learning algorithm. Each episode will start with different initial state conditions, including variations in the initial values of T_{ref} and the disturbance $I(t)$. More specifically, the initial value of the desired fluid outlet temperature T_{ref} is within the range of 100 °C to 240 °C and the initial value of the irradiance $I(t)$ is within the range of $400\,\mathrm{W/m^2}$ to $800\,\mathrm{W/m^2}$. Disturbances such as the ambient temperature $T_g(t)$ or the initial fluid temperature $T(0,t)$ are kept constant at an operating point for this work. The stopping criterion for the agents' learning process is that the average reward of the last N episodes exceeds a predefined threshold or that the maximum number of episodes is reached. This criterion aims to train an agent that achieves good reward function values across N different cases.

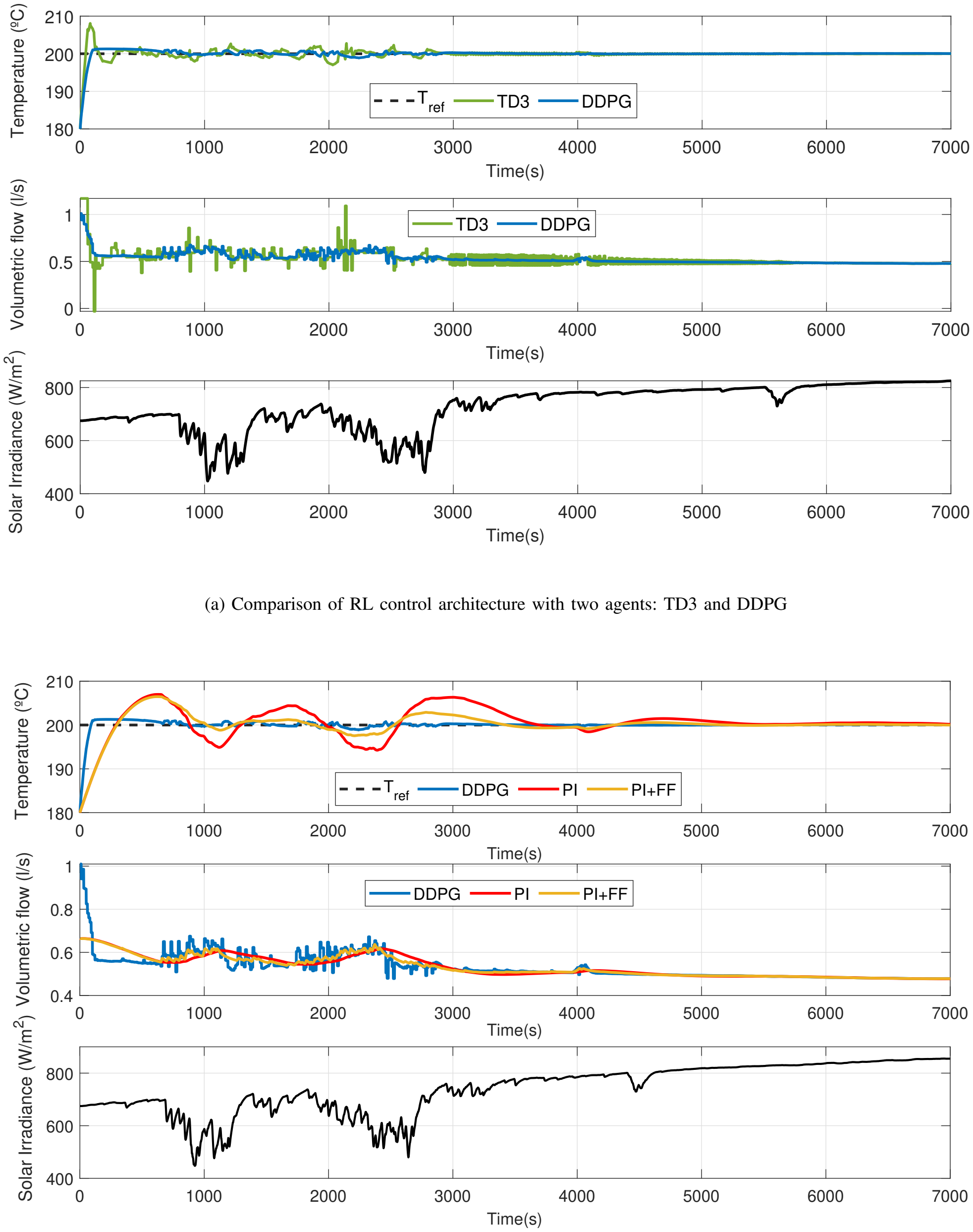
Among the most relevant hyperparameters, both agents used an exploration noise variance of 0.3 with a decay rate of 10^{-5}, promoting initial exploration and gradual exploitation. Additionally, TD3 incorporates target policy smoothing with the same variance (0.3) and decay rate as DDPG, a feature that enhances training stability.

To demonstrate the efficiency of RL, it will be compared with a classical PI controller tuned using pole-zero cancellation. An aggressive approach will be considered, using a closed-loop time constant $\tau_{bc} = 0.7\tau$ with a proportional gain of $k_p = \tau_i/(k_g \tau_{bc}) = 4.72 \times 10^{-6}\ \mathrm{m^3/(s°C)}$, and an integral time of $\tau_i = \tau = 215.16$ s. To validate the proposed cases through simulation, real operational data collected from a solar plant were used, ensuring realistic testing conditions.

Finally, to compare the performance of the proposed controllers in each case, the results were analyzed using several well-established performance criteria. Specifically, the following indices were employed: the Integral Absolute Error (IAE) $\left(\int |e(t)|\, dt\right)$, the Integral Square Error (ISE) $\left(\int e^2(t)\, dt\right)$ and the Integral Time-weighted Absolute Error (ITAE) $\left(\int |te(t)|\, dt\right)$. These indices evaluate controller performance based on the error between the reference and output, defined as $(e(t) = T_{ref}(t) - T(x,t))$. Additionally, the Control Effort Index (CEI) $\left(\int |\Delta u(t)|\, dt\right)$ was used to assess the control effort required to achieve the desired performance. The results of this analysis are summarized in Table II.

A. Case 1. Comparison of DDPG and TD3 Agents

The performance of two RL agents, DDPG and TD3, was compared to determine which algorithm is better suited



(a) Comparison of RL control architecture with two agents: TD3 and DDPG

(b) Comparison of DDPG, PID, and PID+Feedforward

Fig. 3: Simulation results of the RL control approach

TABLE II
PERFORMANCE CRITERIA OF DIFFERENT CONTROLLERS

Controller	**IAE**	**ISE**	**ITAE**	**CEI**
TD3	345.9	1093.4	2.42×10^5	32.70
DDPG	238.5	953.5	1.67×10^5	6.68
PID Controller	1680.3	9182.3	1.18×10^6	0.48
PID Controller + FF	960.5	5411.2	6.72×10^5	1.35

to the dynamics of the PTC system. Fig. 3a presents the simulation results for both agents. The graph above illustrates the temporal evolution of the output temperature in response to the two proposed control strategies: the controller using a DDPG agent (in blue) and the controller using a TD3 agent (in green). The middle graph illustrates the control signal, representing how the volumetric flow rate varies over time in order to regulate the output temperature at the reference value of $200\,^{\circ}\mathrm{C}$. The lower graph shows the irradiance profile for a selected day of operation, which ranges from $448\,\mathrm{W/m^2}$ to $855\,\mathrm{W/m^2}$.

The TD3 controller exhibited larger oscillations in the output temperature, particularly during periods of rapid changes in solar irradiance. This behaviour is attributed to the difficulty of the algorithm in adapting to the high-frequency dynamics of the system, as evidenced by the higher IAE and ISE values compared to the DDPG agent (see Table II). The DDPG controller demonstrated superior performance, maintaining the output temperature closer to the reference value with minimal oscillations. This is reflected in the lower values of IAE, ISE, and ITAE, indicating better tracking accuracy and disturbance rejection. Regarding the control action, when the TD3 agent is used, the control signal is more aggressive, and the control effort is significantly higher, as shown by the CEI index, which is five times greater for TD3. Therefore, for this specific case study, the use of the DDPG agent is more relevant, and it will be used for comparison in the next case.

B. Case 2. Comparison of DDPG with Tradicional Control Strategies

The performance of the DDPG-based RL controller was further compared with traditional control strategies, including PID control and PID with feedforward (FF) compensation. Fig. 3b illustrates the simulation results for these controllers.

The traditional PID controller (in red) showed significant deviations from the reference temperature, particularly during periods of high irradiance variability. This is evident in the large oscillations and slower response times observed in Fig. 3b. The high IAE, ISE, and ITAE values highlight the limitations of PID control in handling the non-linear dynamics and disturbances inherent in the PTC system. However, this case exhibits the least variation in the control signal, with a relatively low CEI index, indicating minimal control effort. The addition of a feedforward component (in yellow) improved the performance of the PID controller by reducing IAE, ISE, and ITAE. However, it negatively impacted the control action, requiring greater control effort, as evidenced by the higher CEI value. It is worth noting that the classical PID controller always has the advantage of immediate tuning, whereas RL controllers may require an undetermined amount of time to converge.

Comparing classical control techniques with DDPG (in blue), it can be observed that DDPG provides better reference tracking and disturbance rejection. However, this comes at the cost of a more aggressive behaviour, requiring a control effort approximately five times higher than the case with feedforward compensation. This is consistent with the response time of the DDPG agent being much faster than the time of the PID controller, see Fig. 3a.

The simulation results demonstrate the potential of RL-based control systems for managing complex, non-linear systems like parabolic trough solar collectors. The DDPG controller outperformed both the TD3 controller and traditional control strategies, particularly in scenarios with significant disturbances, such as fluctuations in solar irradiance. This suggests that RL-based controllers can provide a more adaptive and robust solution for solar thermal systems, enhancing their overall efficiency and reliability.

V. Conclusions

This study has demonstrated the effectiveness of RL-based control systems for managing the temperature regulation of PTCs. The proposed DDPG controller outperformed traditional control strategies, such as PID and PID with feedforward control, in terms of tracking accuracy and disturbance rejection. Specifically, the DDPG controller achieved significantly lower IAE, ISE, and ITAE values, highlighting its ability to handle the non-linear dynamics and external disturbances inherent in solar thermal systems.

The results underscore the potential of RL-based control systems to enhance the performance and reliability of solar thermal technologies, which are critical for meeting the growing demand for renewable heat in industrial processes. By leveraging the adaptive and learning capabilities of RL, the proposed controller provides a robust solution for maintaining stable operation under varying environmental conditions, such as fluctuations in solar irradiance.

In real-world implementations, it is important to consider that achieving an ideal theoretical control signal is rarely feasible due to the physical constraints and imperfections of real systems. Therefore, the dynamics and nonlinearities of the plant's actuators—such as pumps and valves—must be taken into account when translating the control signal computed by the RL agent into actual control actions.

Future research could explore the application of RL with other advanced control techniques, such as model predictive control (MPC). Additionally, further investigation into the scalability and real-time implementation of RL-based controllers in large-scale solar thermal plants would be valuable for advancing their practical adoption.

References

[1] International Energy Agency, *Renewables 2024*. Paris, France: IEA, 2024. [Online]. Available: https://www.iea.org/reports/renewables-2024. [Accessed: Jan. 30, 2025].

[2] P. Horta and Fraunhofer Inst. Solar Energy Syst., "Process Heat Collectors: State of the Art and Available Medium-Temperature Collectors," IEA SHC Task 49/IV—Deliverable A, vol. 1, 2015.

[3] J. L. Guzmán and T. Hägglund, "Simple tuning rules for feedforward compensators," *J. Process Control*, vol. 21, no. 1, pp. 92–102, Jan. 2011.

[4] E. F. Camacho, F. R. Rubio, M. Berenguel, and L. Valenzuela, "A survey on control schemes for distributed solar collector fields. Part I: Modeling and basic control approaches," *Solar Energy*, vol. 81, no. 10, pp. 1240–1251, Oct. 2007.

[5] J. A. Carballo *et al.*, "Reinforcement learning for heliostat aiming: Improving the performance of solar-tower plants," *Appl. Energy*, vol. 377, Art. no. 124574, 2025.

[6] I. M. L. Pataro *et al.*, "Optimal model-free adaptive control based on reinforcement Q-learning for solar thermal collector fields," *Eng. Appl. Artif. Intell.*, vol. 126, Art. no. 106785, 2023.

[7] J. J. Díaz and J. A. Fernández, "The potential of control models based on reinforcement learning in the operation of solar thermal cooling systems," *Processes*, vol. 10, no. 8, Art. no. 1649, 2022.

[8] J. Álvarez, L. Yebra, and M. Berenguel, "Adaptive repetitive control for resonance cancellation of a distributed solar collector field," *Int. J. Adapt. Control Signal Process.*, vol. 23, pp. 331–352, 2009.

[9] K. J. Åström and T. Hägglund, *Advanced PID Control*. Madrid, Spain: Pearson, 2009.

[10] B. A. Ogunnaike and W. H. Ray, *Process Dynamics, Modeling, and Control*, Oxford, UK: Oxford University Press, 1994, 1260 p.

[11] MathWorks, "Reinforcement Learning Agents." [Online]. Available: https://es.mathworks.com/help/reinforcement-learning/ug/create-agents-for-reinforcement-learning.html. [Accessed: Jan. 23, 2025].

[12] MathWorks, "Reinforcement Learning Toolbox." [Online]. Available: https://es.mathworks.com/help/reinforcement-learning.html. [Accessed: Jan. 30, 2025].