CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Control adaptativo basado en aprendizaje por refuerzo para sistemas no lineales complejos: aplicación a un CSTR

Antonio Martínez-Roa[a,*], Juan D. Gil[a], Igor M.L. Pataro[a], Antonio Del Rio Chanona[b], José L. Guzmán[a], Manuel Berenguel[a]

[a] *Departamento de Informática, CIESOL-ceiA3, Universidad de Almería, Ctra. Sacramento s/n, 04120, Almería, España.*
[b] *Sargent Centre for Process Systems Engineering, Imperial College London, SW7 2AZ, UK.*

To cite this article: Martínez-Roa, Antonio, Gil, Juan D., Pataro, Igor M. L., Del Rio Chanona, Antonio, Guzmán, José L., Berenguel, Manuel. 2025. Supervision and control system for a solar collector field with reflectors. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13801

Resumen

Este trabajo propone una estrategia de control adaptativo para la regulación de la concentración en un reactor continuo de tanque agitado (CSTR, por sus siglas en inglés *Continuous Stirred Tank Reactor*), en la que un agente de aprendizaje por refuerzo ajusta en línea las ganancias de un controlador PID con el objetivo de imponer una dinámica deseada en bucle cerrado. La metodología combina la interpretabilidad y robustez de los controladores clásicos tipo PID con la capacidad adaptativa del aprendizaje por refuerzo, mediante una función de recompensa definida a partir del error entre la respuesta del sistema y el comportamiento dinámico de referencia. Para ello, se emplea un agente basado en el algoritmo *Deep Deterministic Policy Gradient* (DDPG). Los resultados obtenidos en simulación muestran que la estrategia propuesta permite mantener una respuesta dinámica más consistente frente a variaciones del punto de operación, superando el desempeño de enfoques clásicos de control adaptativo, como la planificación de ganancias.

Palabras clave: Control de procesos, Control adaptativo por modelo de referencia, Aprendizaje automático, Planificación de ganancias

Reinforcement Learning-Based Adaptive Control for Complex Nonlinear Systems: Application to a CSTR

Abstract

This work proposes an adaptive control strategy for concentration regulation in a Continuous Stirred Tank Reactor (CSTR), in which a reinforcement learning agent adjusts the gains of a PID controller online in order to impose a desired closed-loop dynamic behavior. The proposed methodology combines the interpretability and robustness of classical PID-type controllers with the adaptive capabilities of reinforcement learning through a reward function defined from the error between the system response and a reference dynamic behavior. To this end, an agent based on the Deep Deterministic Policy Gradient (DDPG) algorithm is employed. Simulation results show that the proposed strategy is able to maintain a more consistent dynamic response under variations in the operating point, outperforming classical adaptive control approaches such as gain scheduling.

Keywords: Process control, Model reference adaptive control, Machine Learning, Gain scheduling

1. Introducción

El control de sistemas no lineales y fuertemente acoplados en ingeniería de procesos continúa siendo un desafío relevante (Bequette, 1991), especialmente en aplicaciones con dinámicas complejas como el reactor continuo de tanque agitado (CSTR, *Continuous Stirred Tank Reactor*). Este tipo de sistemas puede presentar múltiples estados estacionarios, fuertes no linealidades y sensibilidad a perturbaciones, lo que li-

*Autor para correspondencia: amr037@ual.es

mita la efectividad de enfoques de control lineal diseñados alrededor de un único punto de operación (Seborg et al., 2016).

Entre las estrategias más utilizadas en la industria, los controladores PID destacan por su simplicidad y amplia adopción. Sin embargo, su desempeño se degrada cuando el sistema opera lejos del punto de sintonía, situación habitual en procesos no lineales como el CSTR, lo que puede derivar en respuestas transitorias deficientes e incluso inestabilidad (Åström and Hägglund, 2006; Seborg et al., 2016).

En este contexto, el aprendizaje por refuerzo (RL, *Reinforcement Learning*) ha emergido como una alternativa para el control de sistemas dinámicos complejos, al permitir el aprendizaje de políticas de decisión a partir de la interacción con el entorno o datos históricos (Bloor et al., 2025b; Lillicrap et al., 2020). Basado en el formalismo de los procesos de decisión de Markov (MDP, *Markov Decision Process*), RL ha sido extendido mediante técnicas de aprendizaje profundo para abordar espacios de acción continuos, como los presentes en procesos industriales (Bertsekas, 2025). No obstante, su aplicación en sistemas físicos está limitada por la ausencia de garantías de estabilidad, interpretabilidad y seguridad (García and Fernández, 2015; Bloor et al., 2025b).

En particular, en el campo de ingeniería de procesos, el uso de RL ha crecido en aplicaciones de control, optimización y planificación, especialmente en procesos continuos y por lotes, destacando su potencial pero también la necesidad de incorporar conocimiento previo y estructuras de control más seguras (Gil et al., 2026; Monteiro and Kontoravdi, 2024; Parambil and Vasantha, 2026).

El presente trabajo toma como referencia la propuesta de (Bloor et al., 2025a), donde se hace uso de una estrategia de RL para ajustar las ganancias de un controlador de tipo PID. Sobre esta base, la principal contribución de este trabajo consiste en extender dicho enfoque mediante la integración de un esquema actor-crítico con una arquitectura de control adaptativo por modelo de referencia (MRAC, *Model Reference Adaptive Control*). De este modo, se propone una estrategia híbrida en la que un agente de RL ajusta en línea los parámetros de un controlador clásico, preservando su estructura e interpretabilidad (Luo et al., 2017; Sachio et al., 2021).

En la estrategia propuesta, a diferencia de enfoques convencionales basados en la minimización directa del error de seguimiento, la señal de recompensa se define respecto a un modelo de referencia que impone una dinámica deseada en lazo cerrado. Este planteamiento permite guiar el aprendizaje hacia un comportamiento dinámico predefinido, favoreciendo una respuesta homogénea ante cambios en el punto de operación. En consecuencia, el agente actúa como un mecanismo adaptativo dependiente del estado que ajusta las ganancias del controlador para seguir la dinámica del modelo de referencia, restringiendo además el espacio de acción a parámetros físicamente interpretables, lo que facilita la incorporación de límites operativos y consideraciones de estabilidad.

El documento se organiza de la siguiente manera: la sección 2 presenta los materiales y métodos. La sección 3 describe la metodología propuesta y su formulación. La sección 4 muestra los resultados obtenidos y la comparación con otras estrategias de control adaptativo. Finalmente, la sección 5 recoge las conclusiones y líneas futuras de trabajo.

2. Materiales y métodos

2.1. Reactor continuo de tanque agitado

Un sistema CSTR (véase Fig. 1) puede modelarse mediante balances de materia y energía bajo el supuesto de mezcla perfecta. Para simplificar el modelo, se asume que la temperatura de la chaqueta es una variable directamente manipulable, por lo que no es necesario considerar un balance de energía en la chaqueta. Bajo estas hipótesis, los balances que describen el comportamiento del sistema son:

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{Af} - C_A) - k_0 e^{-\frac{E}{RT}} C_A, \tag{1}$$

$$\frac{dT}{dt} = \frac{F}{V}(T_f - T) - \frac{\Delta H}{\rho C_p} k_0 e^{-\frac{E}{RT}} C_A - \frac{UA}{\rho C_p V}(T - T_j), \tag{2}$$

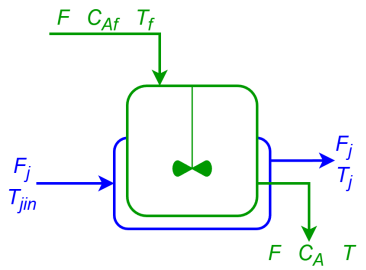

Figura 1: Esquema CSTR.

donde C_A es la concentración del componente A en el reactor, C_{Af} la concentración en la corriente de alimentación, F el caudal volumétrico, V el volumen del reactor, k_0 y E los parámetros de Arrhenius, R la constante universal de los gases, T la temperatura del reactor, T_f la temperatura de la corriente de alimentación, ΔH el calor de reacción, ρ la densidad del fluido, C_p la capacidad calorífica, U el coeficiente global de transferencia de calor, A el área de intercambio térmico y T_j la temperatura de la chaqueta. Los valores numéricos de estos parámetros puede ser consultados en Koiš et al. (2026).

De entre las posibles configuraciones para el estudio del control del sistema, se ha seleccionado la más simple, en la cual la temperatura de la chaqueta actúa como variable manipulada para regular la concentración del reactor. Asimismo, las perturbaciones en la temperatura y concentración de la corriente de alimentación se consideran constantes.

2.2. Aprendizaje por refuerzo

El aprendizaje por refuerzo es una rama del aprendizaje automático que se basa en la interacción entre un agente, un entorno y un objetivo. Esta interacción puede formularse mediante un MDP, definido formalmente como la tupla (s, a, p, r, γ), donde s es el conjunto de estados del sistema, a el conjunto de acciones posibles, p la función de probabilidad de transición de estados, r la función de recompensa, y $\gamma \in [0, 1]$ el factor de descuento que pondera la importancia de recompensas futuras.

En muchos entornos reales el agente no tiene acceso completo al estado del sistema, sino únicamente a observaciones parciales. En estos casos, el problema se modela como un MDP parcialmente observable (POMDP), en el que el agente recibe observaciones o relacionadas con el estado real del sistema. El objetivo del agente es aprender una política que

maximice la recompensa acumulada mediante interacción secuencial con el entorno.

Los algoritmos de aprendizaje por refuerzo pueden clasificarse como *on-policy* y *off-policy*, dependiendo de si el agente aprende a partir de las acciones generadas por su política actual o a partir de datos obtenidos con una política diferente. En *on-policy*, el agente aprende de la política que ejecuta, mientras que en *off-policy* puede aprender de datos generados por otras políticas, permitiendo reutilización de experiencia. En función de cómo se recopilan los datos, los métodos de RL también pueden clasificarse en aprendizaje en línea (*online*) y fuera de línea (*offline*). En el aprendizaje *online* el agente interactúa directamente con el entorno, mientras que en *offline* aprende a partir de datos previamente registrados.

Entre los diferentes algoritmos de RL, una de las arquitecturas más utilizadas es la de actor-crítico, que combina las ventajas de los métodos basados en políticas y en funciones de valor. En esta estructura, el actor define la política de control y el crítico estima el valor de las acciones para guiar la actualización de dicha política.

2.3. Algoritmo de Gradiente de Política Determinista Profunda

El algoritmo Gradiente de Política Determinista Profunda (DDPG, por sus siglas en inglés *Deep Deterministic Policy Gradient*) (Rajasekhar et al., 2025) es un método actor-crítico *off-policy* diseñado para problemas con espacios de acción continuos. Se basa en un actor determinista $\pi(o;\theta)$ y un crítico $Q(o,a;\Phi)$ que estima el valor de acción, siendo θ y Φ los pesos de sus respectivas redes.

En un esquema de entrenamiento en línea, las transiciones (o_t, a_t, r_t, o_{t+1}) se almacenan en un buffer de experiencias, de tamaño M, y se muestrean aleatoriamente para estabilizar el aprendizaje.

El crítico se entrena minimizando el error cuadrático medio respecto a un valor objetivo definido como:

$$y_i = r_i + \gamma Q_T(o_{i+1}, \pi_T(o_{i+1};\theta_T);\Phi_T), \tag{3}$$

$$L(\Phi) = \frac{1}{M}\sum_{i=1}^{M}\left(y_i - Q(o_i,a_i;\Phi)\right)^2, \tag{4}$$

donde $\gamma \in [0,1]$ es el factor de descuento. Q_T y π_T son redes objetivo (copias retardadas) utilizadas para mejorar la estabilidad del entrenamiento. El subíndice T denota la pertenencia de de estos parámetros a las redes objetivo.

La actualización de dichas redes objetivo se realiza mediante una interpolación suave de parámetros:

$$\Phi_T \leftarrow \tau\Phi + (1-\tau)\Phi_T, \quad \theta_T \leftarrow \tau\theta + (1-\tau)\theta_T, \tag{5}$$

donde $\tau \ll 1$ controla la velocidad de actualización.

La política del actor se optimiza maximizando la recompensa acumulada esperada mediante el gradiente de política determinista:

$$\nabla_\theta J \approx \frac{1}{M}\sum_{i=1}^{M}\nabla_a Q(o_i,a;\Phi)\Big|a=\pi(o_i)\nabla\theta\pi(o_i;\theta). \tag{6}$$

DDPG permite un aprendizaje estable en entornos continuos gracias al uso de redes objetivo y reutilización de experiencias.

2.4. Controladores de tipo PID

Como base para la estrategia de control propuesta, se empleó un controlador PI en estructura paralela para la regulación de la concentración del reactor mediante la manipulación de la temperatura de la chaqueta. La expresión del controlador en el dominio de Laplace se define como (Åström and Hägglund, 2006):

$$C(s) = K_p + \frac{K_i}{s}, \tag{7}$$

donde K_p representa la ganancia proporcional y K_i la ganancia integral. Nótese que en este trabajo se utiliza un controlador PI por simplicidad, pero los resultados expuestos son igualmente aplicables a controladores con parte derivativa.

2.5. Índices de desempeño

Con el objetivo de cuantificar la mejora de la metodología propuesta, se emplean dos índices de desempeño: la integral del error absoluto y el esfuerzo acumulado de la señal de control. La integral del error absoluto (IAE, por sus siglas en inglés *Integral of Absolute Error*) mide la desviación acumulada entre la referencia deseada y la variable del proceso. En este trabajo, este índice se evalúa tanto respecto a la referencia, IAE_{SP}, como respecto a la dinámica impuesta por el modelo de referencia, IAE_{MR}.

Por otro lado, el esfuerzo acumulado de la señal de control (CCE, por sus siglas en inglés *Cumulative Control Effort*) cuantifica la variación total de la acción de control a lo largo del tiempo. Ambos índices se definen como (siendo T el tiempo total de simulación):

$$IAE = \int_0^T |e_t|\,dt, \quad CCE = \int_0^T |\Delta u_t|\,dt, \tag{8}$$

donde e_t representa el error de seguimiento en el instante t, definido como la diferencia entre la señal de referencia (referencia externa o modelo de referencia y la salida del sistema, mientras que Δu_t corresponde a la variación de la acción de control entre dos instantes consecutivos.

3. Control PID adaptativo por modelo de referencia mediante aprendizaje por refuerzo

3.1. Descripción de la metodología propuesta

El objetivo de este trabajo es desarrollar una estrategia de control adaptativo para sistemas no lineales, capaz de mantener un comportamiento dinámico similar y homogéneo bajo distintas condiciones de operación. Para ello, se propone un enfoque denominado Control PID adaptativo por modelo de referencia mediante aprendizaje por refuerzo (RL-MRAC-PID, por sus siglas en inglés *RL-Based Model Reference Adaptative PID Control*), en el que un agente de aprendizaje por refuerzo ajusta en línea los parámetros de un controlador PID con el fin de que la salida del sistema siga la dinámica impuesta por un modelo de referencia previamente definido, el cual, caracteriza la respuesta deseada en bucle cerrado. A diferencia de los enfoques tradicionales de control adaptativo o de aprendizaje por refuerzo puro, la metodología propuesta combina la interpretabilidad y robustez de los controladores

clásicos de tipo PID, con la capacidad de adaptación y aprendizaje del RL, utilizando una función de recompensa basada en el error de seguimiento respecto al modelo de referencia, el cual refleja la dinámica deseada en bucle cerrado. De este modo, se busca garantizar una respuesta homogénea del sistema, independientemente del punto de operación, al tiempo que se preservan criterios de estabilidad y restricciones físicas del proceso.

3.2. Formulación del POMDP para el RL-MRAC-PID

3.2.1. Espacio de observaciones

El espacio de observaciones se define como el conjunto de variables accesibles al agente que describen el comportamiento del sistema en cada instante de tiempo. Dado que en aplicaciones reales no se dispone de acceso completo al estado interno del proceso, el agente basa sus decisiones en medidas disponibles y variables derivadas relevantes para el control.

En este trabajo, el vector de observaciones en el instante de tiempo t se define como:

$$o_t = [C_t, T_{j,t}, SP_t, e_t, e_{MR,t}, K_{p,t}, K_{i,t}], \tag{9}$$

donde C_t representa la concentración del reactor, $T_{j,t}$ la temperatura de la chaqueta, y r_t la referencia del sistema. El término $e_t = r_t - C_t$ corresponde al error de seguimiento respecto a la referencia, mientras que $e_{MR,t} = C_t - C_{MR,t}$ representa el error respecto al modelo de referencia. Además, se incluyen las acciones aplicadas en el instante anterior, representadas por las ganancias del controlador $K_{p,t}$ y $K_{i,t}$, con el objetivo de proporcionar al agente información de contexto sobre la dinámica reciente del sistema y mejorar la estabilidad del aprendizaje.

La inclusión de estas variables permite capturar tanto el estado actual del proceso como su evolución temporal, facilitando la toma de decisiones del agente en el ajuste de los parámetros del controlador. En particular, la incorporación explícita de la referencia, junto con el error respecto al modelo de referencia, permite guiar el aprendizaje hacia el seguimiento de una dinámica deseada, en línea con la filosofía de control del MRAC.

3.2.2. Espacio de acciones

El espacio de acciones se define como el conjunto de variables que el agente puede modificar para influir en el comportamiento del sistema. En este trabajo, las acciones corresponden a las componentes proporcional e integral de un controlador PID en su forma paralela, es decir, las ganancias K_p y K_i.

De este modo, en cada instante de tiempo el agente selecciona una acción de la forma:

$$a_t = [K_{p,t}, K_{i,t}], \tag{10}$$

lo que permite interpretar el aprendizaje como un mecanismo de sintonía adaptativa en línea del controlador PID. Este enfoque aproxima el comportamiento del agente de RL al de las estrategias clásicas de control adaptativo, donde los parámetros del controlador se ajustan en función del estado del sistema.

Con el fin de garantizar la viabilidad física y la estabilidad del proceso, las acciones se encuentran acotadas dentro de un rango predefinido, en particular, y debido a la forma de implementar el PID (ver sección 2.4) y a la dinámica del sistema CSTR, las ganancias deben restringirse a valores negativos, lo cual se tiene en cuenta en la definición del espacio de acciones.

3.2.3. Recompensa

Dado que el objetivo del problema es el seguimiento de una dinámica deseada en bucle cerrado, se define una función de recompensa basada en el error entre la salida del sistema y la salida del modelo de referencia. De este modo, el aprendizaje del agente se orienta a reproducir una respuesta dinámica previamente especificada para todos los puntos de operación del sistema.

En particular, la función de recompensa se define como una combinación ponderada del error cuadrático instantáneo y la integral del error, con el fin de penalizar tanto desviaciones transitorias como errores en estado estacionario:

$$r_t = -\left(\alpha\, e_{MR,t}^2 + \beta \int_0^t |e_{MR,t}|\, dt\right), \tag{11}$$

donde $\alpha = 100$ y $\beta = 0{,}01$ son los pesos asociados al error cuadrático y al término integral, respectivamente.

Esta formulación permite guiar el aprendizaje hacia el seguimiento de la dinámica deseada, penalizando tanto desviaciones rápidas como errores persistentes, en línea con los objetivos del control adaptativo por modelo de referencia.

3.3. Diagrama de bloques resultante para el entrenamiento

En la Figura 2 se muestra el diagrama de bloques correspondiente a la fase de entrenamiento, donde se integran los distintos componentes del sistema descritos anteriormente. Como se observa, el agente de RL opera dentro de un esquema de control adaptativo en bucle cerrado, ajustando en línea los parámetros del controlador en función de las observaciones del sistema.

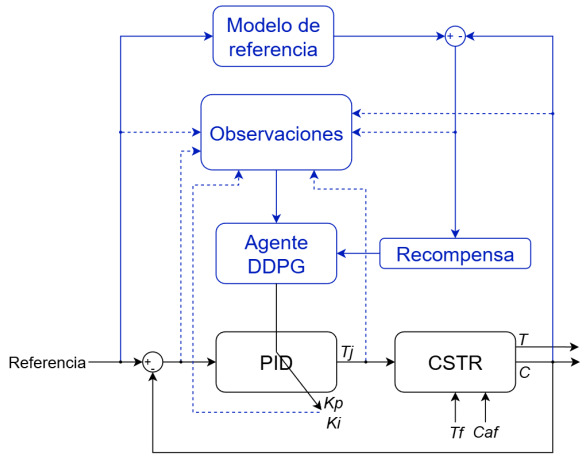

Figura 2: Esquema para el entrenamiento del agente RL.

El comportamiento del agente está guiado por la función de recompensa, la cual incorpora la dinámica del modelo de referencia, permitiendo orientar el proceso de aprendizaje hacia la obtención de una respuesta homogénea y deseada en diferentes condiciones de operación. Finalmente, y para garantizar tanto la seguridad operativa como la eficiencia del entrenamiento, se han implementado cuatro criterios de parada,

dos relacionados con las salidas del CSTR (concentración y temperatura) y otros dos relacionados con los parámetros del controlador (K_p y K_i).

4. Resultados

En la simulaciones presentadas en esta sección, se ha considerado una constante de tiempo de 150 minutos para el modelo de referencia de la estructura RL-MRAC-PID.

4.1. *Entrenamiento en línea del agente*

Con el objetivo de evaluar la capacidad de adaptación de la metodología propuesta, se entrenaron diferentes agentes empleando el esquema RL-MRAC-PID descrito anteriormente. Durante el entrenamiento se consideraron distintas estructuras de la red neuronal utilizada en el actor y se varió el número de neuronas de ambas redes (128 neuronas en el Agente 1, 32 en los Agentes 2 y 3, y 8 en el Agente 4). Las distintas configuraciones pueden observarse de manera ilustrativa en la Figura 3, y fueron seleccionadas mediante un proceso de prueba y error, realizando un análisis de sensibilidad sobre la estructura de las redes y el número de neuronas. El tiempo de entrenamiento fue de entre 2 y 2.5 horas por agente.

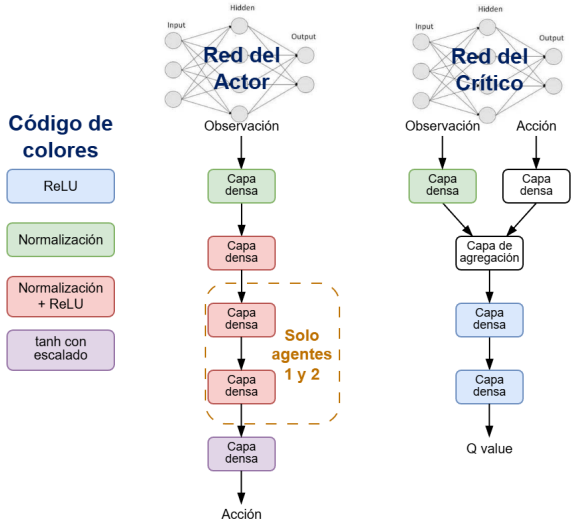

Figura 3: Redes utilizadas para el actor y el crítico.

La Figura 4 muestra la respuesta temporal obtenida para los distintos agentes entrenados (C_{ai} con $i \in [1,4]$ para los 4 agentes entrenados), observándose diferencias en el comportamiento dinámico y en la capacidad de seguimiento de la dinámica deseada. Asimismo, la Figura 5 presenta la evolución temporal de las ganancias K_p y K_i, evidenciando el carácter adaptativo del enfoque propuesto.

Con el fin de cuantificar el desempeño de cada agente, la Tabla 1 presenta los índices de evaluación definidos previamente. Los resultados muestran que el Agente 1 obtiene el mejor compromiso entre seguimiento de la dinámica deseada IAE_{MR} y esfuerzo de control CCE. Por otro lado, el Agente 3 consigue el menor error respecto a la referencia (IAE_{SP}), aunque a costa de un incremento significativo del esfuerzo de control.

Tabla 1: Comparación de desempeño entre agentes RL

Agente	*Reward*	IAE_{MR}	IAE_{SP}	CCE
1	-2699.6	32.39	101.8	312.4
2	-4119.0	48.59	106.7	313.4
3	-4450.3	49.93	80.4	336.7
4	-7369.8	80.7	113	333.7

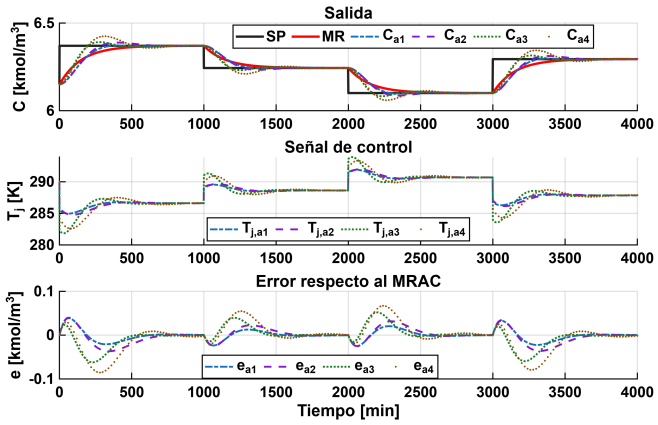

Figura 4: Comparativa de las salidas entre agentes RL.

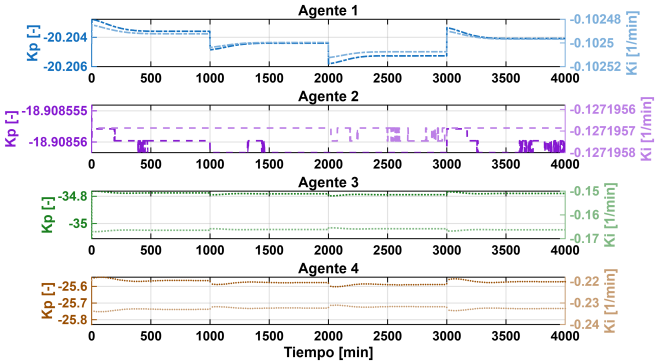

Figura 5: Parámetros del controlador para los diferentes agentes.

4.2. *Comparativa con dos estrategias de control adaptativo*

Finalmente se evaluó el desempeño de la metodología propuesta comparando el controlador RL-MRAC-PID, obtenido con el agente 1, con dos estrategias convencionales de control adaptativo: planificación de ganancias clásico (GS, por sus siglas en inglés *Gain Scheduling*), y ajuste basado en polinomios (POL en las gráficas). El enfoque GS se ha implementado mediante el diseño de controladores locales en torno a cuatro puntos de operación, obtenidos a partir de la linealización del sistema en cada uno de ellos. A partir de los valores de K_p y K_i calculados en estos puntos, el controlador basado en polinomios se ha obtenido mediante un ajuste polinomial de dichos parámetros. En ambos casos se usa una especificación de constante de tiempo de bucle cerrado de 150 minutos.

Las Figuras 6 y 7 muestran, respectivamente, la evolución temporal de la concentración, la señal de control y el error respecto al modelo de referencia, así como la evolución temporal de las ganancias del controlador para las distintas estrategias consideradas. En general, el enfoque basado en RL presenta

una respuesta más cercana a la dinámica deseada y una menor desviación durante los transitorios (ver IAE_{MR} en Tabla 2).

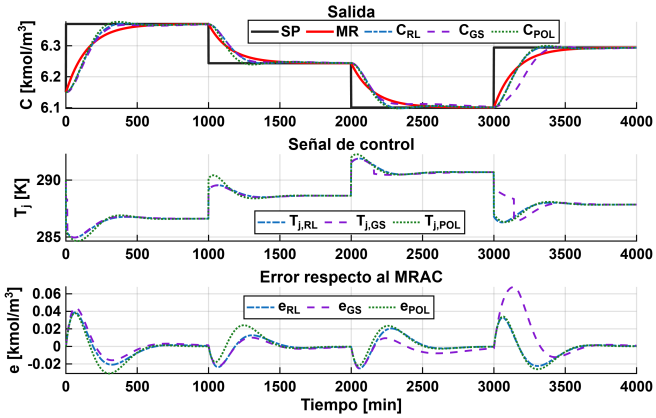

Figura 6: Comparativa de las salidas entre las diferentes técnicas de control.

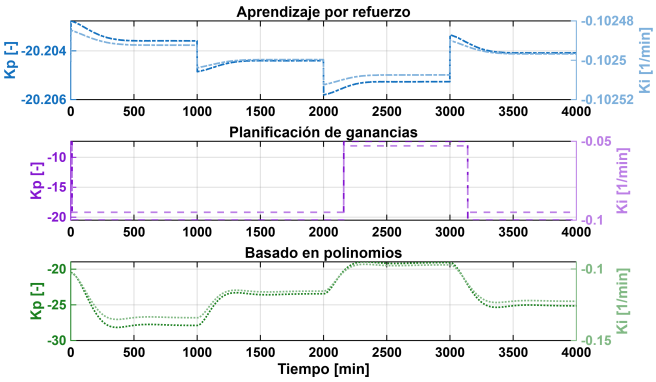

Figura 7: Parámetros del controlador para las diferentes técnicas de control.

La Tabla 2 resume los índices de desempeño obtenidos. El controlador basado en RL reduce el error de seguimiento de la dinámica deseada en aproximadamente un 15 % respecto a la estrategia de planificación de ganancias y en un 10 % respecto al método basado en polinomios. En términos de esfuerzo de control, las tres metodologías presentan valores de CCE similares, con diferencias inferiores al 1 %.

Tabla 2: Comparación de desempeño entre estrategias de control

Estrategia	*Reward*	IAE_{MR}	IAE_{SP}	*CCE*
RL-MRAC-PID	-2127.5	32.39	101.8	312.4
GS	-2334.8	38.42	124.2	312.1
Polinomial	-2417.8	35.95	96.5	316.2

En resumen, los resultados muestran que el enfoque propuesto proporciona una respuesta más uniforme en distintos puntos de operación, manteniendo un esfuerzo de control comparable al de las estrategias adaptativas convencionales.

5. Conclusiones y trabajos futuros

En este trabajo se ha desarrollado una estrategia de control adaptativo basada en aprendizaje por refuerzo y modelo de referencia para el control de un CSTR. La metodología propuesta, denominada RL-MRAC-PID, emplea un agente de aprendizaje por refuerzo para ajustar en línea las ganancias de un controlador PI con el fin de imponer una dinámica deseada en bucle cerrado. Los resultados obtenidos muestran que la estrategia propuesta mantiene un comportamiento dinámico más uniforme entre distintos puntos de operación que los enfoques adaptativos convencionales, mejorando el índice IAE entre un 10 % y un 15 % sin incrementar el esfuerzo de control.

Como líneas futuras de investigación, se propone la incorporación de una acción feedforward para mejorar el rechazo de perturbaciones, así como la validación experimental de la metodología en una planta real.

Agradecimientos

Este trabajo ha sido financiado por la Consejería de Universidad, Investigación e Innovación de la Junta de Andalucía a través del proyecto DGP_PIDI_2024_02211, cofinanciado por el Fondo Europeo de Desarrollo Regional (ERDF/-FEDER) en el marco del Programa FEDER Andalucía 2021-2027, y es parte del proyecto de I+D+i PID2023-150739OB-I00 financiado por MCIN/AEI/10.13039/5011000 11033/ y "FEDER Una manera de hacer Europa". Además, cuenta con financiación del Plan Propio de Investigación y Transferencia de la Universidad de Almería.

Referencias

Åström, K. J., Hägglund, T., 2006. Advanced PID control. IEEE Control Systems 26 (1), 98–101.

Bequette, B. W., 1991. Nonlinear control of chemical processes: A review. Industrial & Engineering Chemistry Research 30 (7), 1391–1413.

Bertsekas, D. P., 2025. Neuro-dynamic programming. In: Encyclopedia of optimization. Springer, pp. 1–6.

Bloor, M., Ahmed, A., Kotecha, N., Mercangoz, M., Tsay, C., del Rio-Chanona, E. A., 2025a. Control-informed reinforcement learning for chemical processes. Industrial & Engineering Chemistry Research 64 (9), 4966–4978.

Bloor, M., Mowbray, M., del Rio Chanona, E. A., Tsay, C., 2025b. A survey and tutorial of reinforcement learning methods in process systems engineering. Computers & Chemical Engineering, 109515.

García, J., Fernández, F., 2015. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research 16 (1), 1437–1480.

Gil, J. D., Chanona, E. A. D. R., Guzmán, J. L., Berenguel, M., 2026. Reinforcement learning meets bioprocess control through behavior cloning: Real-world deployment in an industrial photobioreactor. Engineering Applications of Artificial Intelligence 164, 113326.

Koiš, R., Pataro, I. M. L., Gil, J. D., Záková, K., Guzmán, J. L., Berenguel, M., 2026. A web-based virtual laboratory for PID control of CSTR processes. In: Proceedings of the 23rd IFAC World Congress. Accepted for publication / In press.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N. M. O., Erez, T., Tassa, Y., Silver, D., Wierstra, D. P., Sep. 15 2020. Continuous control with deep reinforcement learning. US Patent 10,776,692.

Luo, B., Liu, D., Wu, H.-N., 2017. Adaptive constrained optimal control design for data-based nonlinear discrete-time systems with critic-only structure. IEEE Transactions on Neural Networks and Learning Systems 29 (6), 2099–2111.

Monteiro, M., Kontoravdi, C., 2024. Bioprocess control: A shift in methodology towards reinforcement learning. In: Computer Aided Chemical Engineering. Vol. 53. Elsevier, pp. 2851–2856.

Parambil, S. K., Vasantha, D. R., 2026. Advanced temperature control of a continuous ethanol fermentation bioreactor using deep reinforcement learning. The Canadian Journal of Chemical Engineering.

Rajasekhar, N., Radhakrishnan, T., Samsudeen, N., 2025. Exploring reinforcement learning in process control: a comprehensive survey. International Journal of Systems Science, 1–30.

Sachio, S., del Rio-Chanona, E. A., Petsagkourakis, P., 2021. Simultaneous process design and control optimization using reinforcement learning. IFAC-PapersOnLine 54 (3), 510–515.

Seborg, D. E., Edgar, T. F., Mellichamp, D. A., Doyle III, F. J., 2016. Process Dynamics and Control. John Wiley & Sons.