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

Jornadas de Automática

Aprendizaje por refuerzo fuera de línea para el control de la temperatura de invernaderos mediante ventilación natural

Daniel Pérez-Sánchez[a,*], Juan D. Gil[a], Francisco García-Mañas[b], Manuel Berenguel[a], Francisco Rodríguez[a]

[a]*Dpto. de Informática, Universidad de Almería, CIESOL, ceiA3, Carretera Sacramento, s/n, 04120, Almería, España*
[b]*Dpto. de Ingeniería de Sistemas y Automática, Universidad de Sevilla, Camino de los Descubrimientos, s/n, 41092, Sevilla, España*

To cite this article: Pérez-Sánchez, D., Gil, J. D., García-Mañas, F., Berenguel, M., Rodríguez, F. 2026. Offline Reinforcement Learning for Greenhouse Temperature Control through Natural Ventilation. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13686

Resumen

El control climático de invernaderos constituye un problema complejo debido a su dinámica no lineal y a la influencia de fuertes perturbaciones externas. En este trabajo se propone una estrategia de control basada en Aprendizaje por Refuerzo (*Reinforcement Learning*, RL) entrenado en modo *offline* para regular la temperatura de un invernadero mediterráneo mediante la ventilación natural. Para ello, se emplea el algoritmo *Deep Deterministic Policy Gradient* (DDPG), entrenado a partir de experiencias obtenidas sobre un sistema real mediante un controlador proporcional integral (PI) considerado como sistema experto. Las políticas obtenidas se evalúan sobre datos no empleados para el aprendizaje. Los resultados muestran que el agente reproduce adecuadamente el comportamiento del controlador experto, generando señales de control suaves y coherentes con la dinámica del sistema, con un esfuerzo de control inferior al 7,4 % respecto al PI. Este trabajo se limita al entrenamiento y validación *offline* del agente a partir de datos reales del proceso, sin contemplar en esta etapa su despliegue sobre la planta.

Palabras clave: Control basado en datos, Sistemas de control inteligente, Control de sistemas ambientales, Automatización de invernaderos, Aprendizaje automático.

Offline Reinforcement Learning for Greenhouse Temperature Control through Natural Ventilation

Abstract

Greenhouse climate control is a complex problem due to its nonlinear dynamic and the influence of strong external disturbances. This work proposes a control strategy based on Reinforcement Learning (RL) trained in offline mode to regulate the temperature of a Mediterranean greenhouse through natural ventilation. To this end, the Deep Deterministic Policy Gradient (DDPG) algorithm is employed, trained using experience collected from a real system under the operation of a proportional-integral controller (PI) considered as an expert system. The resulting policies are evaluated using data not used for learning. The results show that the agent accurately replicates the behavior of the expert controller, generating smooth control signals consistent with the system dynamics, with a control effort reduction of 7.4 % compared to the PI. This work is limited to the offline training and validation of the agent using real process data, without considering its direct deployment on the plant at this stage.

Keywords: Data-driven control, Intelligent control system, Enviromental systems control, Greenhouse automation, Machine learning.

1. Introducción

El cultivo bajo invernadero constituye uno de los pilares fundamentales en la actividad económica del sureste español, principalmente en la provincia de Almería (Castro et al., 2019). Estas estructuras permiten crear un microclima controlable que favorece la producción intensiva durante gran parte

*Autor para correspondencia: daniel.perezz@ual.es

del año. Además, los invernaderos son una solución altamente eficiente para proteger al cultivo frente a condiciones meteorológicas adversas y mantener unas condiciones climáticas adecuadas, particularmente en lo relativo a temperatura del aire, humedad, radiación solar y concentración de CO_2 (Yusuf et al., 2025). Entre estas variables, destaca la temperatura del aire interior del invernadero, ya que influye directamente en el desarrollo fisiológico de las plantas.

Los invernaderos mediterráneos de tipo parral, que son los ampliamente extendidos en Almería, tienen como principal mecanismo de regulación térmica la ventilación natural. La apertura de las ventanas permite el intercambio convectivo del aire con el exterior, permitiendo evacuar el aire caliente acumulado en el interior, reduciendo así la temperatura interna. Sin embargo, la dinámica asociada a este proceso es altamente no lineal y está fuertemente condicionada por las perturbaciones ambientales como la radiación solar, la velocidad y dirección del viento, la temperatura o incluso el estado fisiológico del cultivo (Rodríguez et al., 2015).

En este contexto, el presente trabajo se enmarca en el enfoque del nexo agua-energía-alimentación (González et al., 2025), donde la gestión eficiente de los sistemas agrícolas requiere considerar de forma conjunta la productividad del cultivo y la minimización de los recursos consumidos (Rodríguez et al., 2024). En particular, el control climático en invernaderos mantiene una relación directa tanto con el rendimiento agronómico, como con el uso eficiente de recursos. En consecuencia, durante los últimos años se han desarrollado numerosas estrategias de control para la regulación de las condiciones internas y la eficiencia operativa del sistema (Li et al., 2025).

Se han propuesto distintas estrategias de control de bajo nivel para la regulación climática de invernaderos mediante ventilación natural. Entre ellas destacan soluciones de control multivariable para el control coordinado de temperatura y humedad relativa (García-Mañas et al., 2024), empleando estructuras PI que abordan el fuerte acoplamiento existente entre ambas variables climáticas. Asimismo, se han llevado a cabo otras estrategias clásicas de control adaptativo y esquemas *feedforward* aplicadas a la apertura de ventanas, con el objetivo de mejorar el rechazo a las perturbaciones meteorológicas (Montoya-Ríos et al., 2020). Por otro lado, el control predictivo basado en modelo (*Model Predictive Control*, MPC) también ha sido apliamente utilizado, empleando modelos físicos del sistema y utilizando la ventilación natural y otros actuadores como variables manipuladas (Svensen et al., 2024). A pesar de los buenos resultados obtenidos, estos enfoques suelen depender de modelos precisos, requiriendo de procesos de identificación y ajuste para mantener un rendimiento adecuado frente a las fuertes no linealidades, el acoplamiento de variables y la variabilidad estacional del microclima de un invernadero.

Bajo esta perspectiva, el Aprendizaje por Refuerzo (*Reinforcement Learning*, RL) permite el diseño de controladores climáticos basados en la interacción con el entorno, eliminando la necesidad de un modelo explícito del sistema. Esta formulación facilita la obtención de políticas de control capaces de adaptarse a dinámicas no lineales y a perturbaciones externas variables, lo que resulta especialmente relevante para la gestión del clima de un invernadero.

Una clasificación que se puede hacer de estos algoritmos es en *on-policy*, donde el agente basa su aprendizaje exclusivamente en los datos derivados de sus propias decisiones (su política) (Sachio et al., 2021), y en *off-policy*, donde el aprendizaje se realiza a partir de información generada con políticas independientes. Esta capacidad de los métodos *off-policy* para reutilizar experiencias previas resulta de vital importancia en la climatización de invernaderos. Llevar a cabo un entrenamiento exploratorio en tiempo real dentro de estas instalaciones es inviable, puesto que las desviaciones térmicas que se inducen pueden provocar un estrés en el cultivo, además de suponer un alto coste energético. De esta forma, se evita comprometer la producción agronómica, además de acelerar el proceso de entrenamiento del agente.

Con todo lo expuesto, este trabajo propone aplicar un algoritmo basado en aprendizaje por refuerzo *off-policy* para el control de la temperatura de un invernadero. En concreto se utiliza un agente *Deep Deterministic Policy Gradient* (DDPG), debido a su capacidad para trabajar con sistemas con variables continuas. En este caso, el agente aprende de los datos históricos generados previamente con un controlador PI clásico. Esta metodología de entrenamiento *offline* evita la interacción con el entorno real (Gil et al., 2026). Para comprobar la fiabilidad y flexibilidad del entrenamiento, se reservaron datos para validar las políticas obtenidas. En las pruebas de validación, se realizó una comparación entre la señal de control del controlador experto y la devuelta por el agente. Cabe destacar que el alcance de este trabajo se limita al entrenamiento y validación *offline* del agente a partir de datos reales del proceso, sin contemplar en esta etapa su despliegue directo sobre la planta.

2. Materiales y métodos

2.1. Descripción del invernadero

Para este trabajo, se ha usado como referencia un invernadero de tipo parral de 1900 m^2, situado en las instalaciones de Agroconnect [1] (Sánchez-Molina et al., 2025), junto a la Universidad de Almería (Figura 1). Cuenta con numerosos actuadores, entre los que se encuentra una instalación de frío y calor, sistemas de humidificación y deshumidificación y un sistema de inyección de CO_2.

Figura 1: Invernadero de Agroconnect.

El actuador considerado en este estudio es el sistema de ventilación natural del invernadero. Está compuesto por seis

[1] https://agroconnect.es/

ventanas cenitales distribuidas longitudinalmente en la cubierta y siete ventanas laterales ubicadas en el perímetro de la estructura. Cada una de estas aperturas es accionada mediante un motor trifásico de corriente alterna de 0,11 kW.

Para la recopilación de los datos, la instalación cuenta con una instrumentación orientada a la supervisión continua del clima interior y de las principales perturbaciones externas que afectan a su dinámica térmica. El sistema de adquisición registra, cada 30 segundos, variables interiores como la temperatura del aire y la humedad relativa. También se miden variables meteorológicas exteriores, entre ellas la radiación solar, la temperatura ambiente, la humedad exterior, la lluvia y la velocidad y dirección del viento.

2.2. Descripción de RL

Desde la perspectiva de la teoría de control, los algoritmos RL abordan problemas de toma de decisiones mediante la formulación del entorno como un Proceso de Decisión de Markov (*Markov Decision Process*, MDP). Este marco se define como una quíntupla $\mathcal{S}, \mathcal{A}, P, R, \gamma$, donde $\mathcal{S}$ es el espacio de estados del sistema, $\mathcal{A}$ el conjunto de acciones, $P : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, 1]$ la función de probabilidad de transición de estados, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ la función de recompensa y γ el factor de descuento.

Sin embargo, en entornos como los invernaderos, es necesario realizar una reformulación del MDP. Esto es debido a que es bastante complejo conocer el estado completo del sistema, derivado de las limitaciones del sistema de medida. Esto lleva a una formulación más realista basada en un entorno parcialmente observable MDP (POMDP), donde el agente construye su política de control a partir de un conjunto de observaciones parciales, O, obtenidas del sistema (Figura 2).

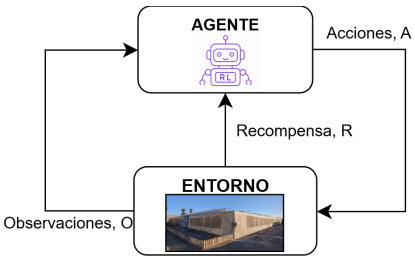

Figura 2: Flujo de trabajo simplificado de un sistema de RL.

Para operar en este tipo de entornos, resulta especialmente adecuada la arquitectura *actor-crítico*. En esta estructura, el *actor* implementa una política determinista definida por la función $\pi(o_k; \boldsymbol{\theta})$, la cual, a partir de una observación o_k, genera las acciones. De forma complementaria, el *crítico* evalúa dichas acciones mediante una función de valor de acción $Q(o_k, a_k; \boldsymbol{\Phi})$, que estima el rendimiento esperado asociado a cada par observación-acción. Durante el proceso de aprendizaje, esta evaluación actúa como señal de realimentación para ajustar la política del actor. En este contexto, los vectores de parámetros $\boldsymbol{\theta}$ y $\boldsymbol{\Phi}$ corresponden a los pesos de los modelos que implementan el actor y el crítico, respectivamente, y se optimizan de manera iterativa para mejorar la política de decisión como su evaluación.

2.3. Descripción del algoritmo DDPG

En el marco *actor-crítico*, el algoritmo DDPG está diseñado para sistemas con espacios de acción continuos, habituales en la mayoría de aplicaciones industriales (Rajasekhar et al., 2025). Este algoritmo emplea redes neuronales para aproximar tanto la política determinista del actor como la función de valor del crítico.

El proceso de entrenamiento fuera de línea (*offline*) se realiza a partir de un conjunto de datos históricos del proceso, los cuales contienen información de las observaciones, acciones y recompensas para cada instante de tiempo muestreado. Estas experiencias se almacenan en un *buffer*, del cual, durante el aprendizaje del agente, se extraen mini-lotes de tamaño M de forma aleatoria y así, actualizar los pesos de las redes del actor y del crítico.

El entrenamiento del crítico se formula como un problema de regresión supervisada, minimizando el error entre la estimación actual de la función de valor y un valor objetivo y_i, definido como:

$$y_i = r_i + \gamma Q_t(o_i, \pi_t(o_i; \boldsymbol{\theta}_t); \boldsymbol{\Phi}_t). \tag{1}$$

Este valor objetivo combina la recompensa inmediata r_i con una estimación de la recompensa futura descontada. Dicha estimación se obtiene mediante redes objetivo (*target networks*), que son copias de las redes principales cuyos parámetros $\boldsymbol{\theta}_t$ y $\boldsymbol{\Phi}_t$ se actualizan de forma gradual. Este desacoplamiento introduce un efecto de suavizado que reduce las oscilaciones y contribuye a la estabilidad del aprendizaje (Lillicrap et al., 2015).

La función de pérdida del crítico se define como:

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

Por su parte, el actor se entrena con el objetivo de maximizar la recompensa acumulada esperada. Para ello, se emplea el gradiente de la política, estimado a partir de la información proporcionada por el crítico:

$$\nabla_{\boldsymbol{\theta}} J \approx \frac{1}{M} \sum_{i=1}^{M} \mathbf{G}_{a_i} \mathbf{G}_{\pi_i}, \tag{3}$$

donde $\mathbf{G}_{a_i}$ representa el gradiente de la función de valor respecto a la acción, evaluado por el crítico, y $\mathbf{G}_{\pi_i}$ corresponde al gradiente de la política respecto a sus propios parámetros. Este mecanismo permite ajustar iterativamente la política en la dirección que incrementa el valor esperado de las acciones generadas.

3. Algoritmo de control RL *offline* propuesto para entornos en agricultura

3.1. Descripción del POMDP

Como se ha comentado anteriormente, este problema de control se modela como un POMDP. En este, el agente toma decisiones a partir de información parcial del entorno:

- Observaciones. El vector de observaciones o_k está compuesto por las variables del invernadero que contienen información relevante sobre su estado térmico. Estas incluyen tanto variables climáticas internas como externas, así como señales asociadas al estado del sistema.

- Acciones. El vector de acciones a_k corresponde a las variables manipuladas del sistema. Aunque el sistema cuenta con numerosos actuadores, en este caso solo se considera como acción principal el grado de apertura del sistema de ventilación:

$$a_k = u_k \in [u_{\text{mín}}, u_{\text{máx}}], \tag{4}$$

  donde u_k representa el porcentaje de apertura de las ventanas del invernadero.

- Función de recompensa. La función de recompensa busca un compromiso entre el seguimiento de la referencia térmica y la suavidad de la acción de control. Se plantea una formulación basada en dos términos: un término de error de seguimiento y un término de penalización del esfuerzo de control:

$$r_k = -log(e_k^2 + \varepsilon) - \lambda \Delta u_k^2, \tag{5}$$

  donde $e_k = T_{ref} - T_k$ es el error en el tiempo de muestreo k entre la temperatura de referencia T_{ref} y el valor actual T_k. Además, existe un parámetro ϵ para ajustar el máximo del término logarítmico y un parámetro λ para ajustar la penalización de la variación de la señal de control.

  El uso de una función logarítmica en el término de error permite amplificar la sensibilidad del agente en las proximidades de la referencia, donde se concentran la mayoría de las muestras debido al uso de datos procedentes de un controlador PI. De este modo, se mejora la calidad de la señal de gradiente durante el entrenamiento, favoreciendo el aprendizaje fino en condiciones cercanas al régimen estacionario. Por otro lado, el término cuadrático asociado a la variación de la acción introduce un efecto de regularización, evitando cambios bruscos en la señal de control y garantizando un comportamiento más suave y físicamente realizable.

3.2. *Algoritmo de control*

A partir de la formulación del problema como un POMDP, se propone un esquema de control basado en aprendizaje por refuerzo de forma *offline*, en el cual la política de control se obtiene a partir de datos históricos del sistema. El procedimiento que se ha seguido para desarrollar el controlador se estructura en tres etapas principales:

1. **Preparación del conjunto de datos.** Antes de comenzar con el entrenamiento del agente, se define el conjunto de datos en forma de tuplas del tipo:

$$\mathcal{D} = \{(o_i, a_i, r_i, o_{i+1})\}_{i=1}^{N}, \tag{6}$$

   cuyo vector de observaciones o_i está formado por T_{ext}, H_{ext}, R_{ext}, VV_{ext} y DV_{ext} que representan las perturbaciones exteriores; T_{int} que incluye el valor actual y tres regresores para capturar la dinámica térmica; T_{suelo} es la temperatura del suelo; T_{SP} es la referencia de temperatura interior; además del error de seguimiento y su integral acumulada. El uso de los errores proporciona un conocimiento de la desviación con respecto al objetivo, así como su acumulación temporal.

2. **Entrenamiento *offline* del agente.** El agente basado en DDPG se entrena utilizando el conjunto $\mathcal{D}$, aproximando mediante redes neuronales tanto la política del actor como la función de valor del crítico. Durante este proceso: (i) se inicializan aleatoriamente los parámetros de las redes, (ii) se extraen mini-lotes aleatorios del conjunto de datos, y (iii) se actualizan iterativamente los parámetros del actor y del crítico aplicando el procedimiento indicado en la sección 2.3.
   Este procedimiento permite al agente aprender la dinámica del sistema sin necesidad de interacción directa con la planta.
3. **Implementación del controlador.** Una vez finalizado el entrenamiento, la red del actor se emplea como el controlador en tiempo real. En cada instante de muestreo k, el controlador recibe el vector de observaciones o_k y genera la acción correspondiente:

$$u_k = \pi(o_k; \boldsymbol{\theta}). \tag{7}$$

   La Figura 3 muestra el esquema de entrenamiento.

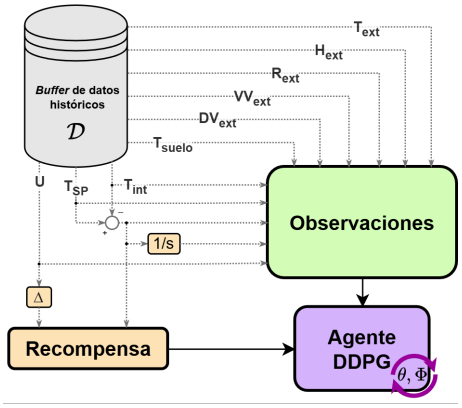

Figura 3: Esquema de entrenamiento del RL.

4. Resultados

4.1. *Extracción de experiencias expertas del sistema real*

El conjunto de datos que se han empleado para el entrenamiento *offline* del agente proviene de una campaña experimental sobre el invernadero bajo condiciones normales de operación. Durante dicha campaña, el control climático fue llevado a cabo mediante un controlador PI, cuya respuesta se consideró como comportamiento experto del sistema. Asimismo, se emplearon distintas referencias de temperatura interior según las condiciones de cada día, lo que incrementa la variabilidad de los datos y supone un reto adicional para el RL en términos de generalización.

La Figura 4 muestra el esquema de control utilizado durante la adquisición de datos, con un tiempo de muestreo para la actuación de un minuto. El controlador presenta una configuración en serie sin término derivativo, donde la ganancia proporcional se recalcula en cada iteración en función de dos perturbaciones que influyen en el directamente en el flujo de ventilación a través de las ventanas: la temperatura exterior (T_{ext}) y la velocidad del viento (VV_{ext}) (Rodríguez et al., 2015). La ganancia se limitó en un rango para evitar comportamientos agresivos, siendo dicho rango $K_p \in [-25, -5]$ %/ºC.

Adicionalmente, la acción integral se configuró con un tiempo integral de $T_i = 660$ s. Este valor se justifica tanto por la inercia térmica característica del invernadero como por la necesidad de que la acción integral elimine errores estacionarios sin introducir oscilaciones significativas en la señal de control.

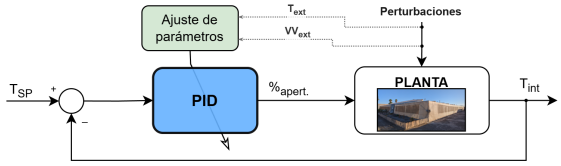

Figura 4: Esquema de control PI para invernaderos.

Los datos que han sido utilizados corresponden a los días 9, 10, 21 y 22 de marzo y 13 y 16 de abril de 2026. Se muestran en la Figura 5.

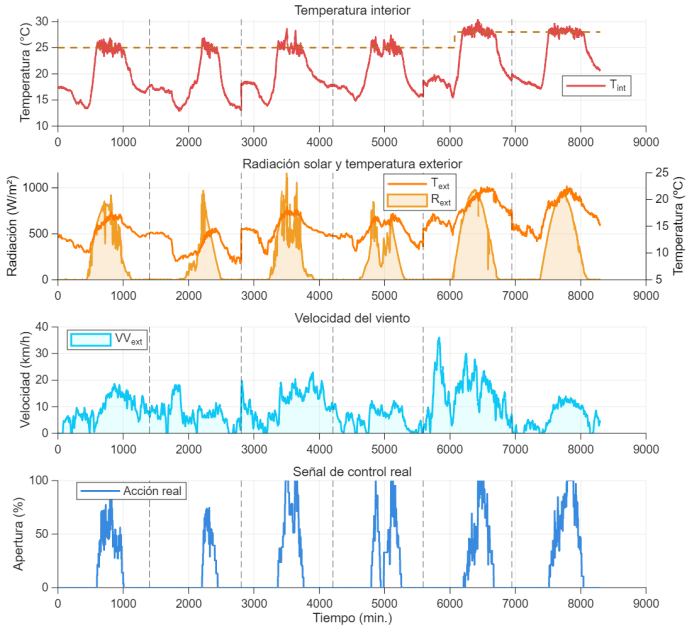

Figura 5: Datos recogidos con el controlador PI.

4.2. *Implementación y entrenamiento del agente*

Para la configuración del agente, los datos de entrada son normalizados para garantizar que todas las variables operen en rangos comparables. Los datos se han escalado al intervalo [0,1, 0,9] en lugar de [0, 1] con el objetivo de evitar valores nulos en las entradas, ya que estos pueden provocar saturación o inactividad de ciertas neuronas, especialmente en funciones de activación que responde de forma no lineal cerca de cero.

En cuanto a la función de recompensa, se han seleccionado unos valores específicos para ε y λ, fijándose $\varepsilon = 10^{-3}$ y $\lambda = 0{,}015$. Estos valores se han ajustado buscando un compromiso. En particular, el valor de ε proporciona una escala suficientemente pequeña para no distorsionar la magnitud del error, pero evitando problemas numéricos cuando este tiende a cero. Por otro lado, el valor de λ se ha fijado como una penalización moderada sobre la variación de la señal de control, de forma que no domina la función de recompensa pero sí es suficiente para suavizar la señal de control y evitar señales agresivas.

La red del actor está compuesta por una serie de capas completamente conectadas con funciones de activación ReLU en las capas ocultas y una función hiperbólica a la salida. Las dos primeras capas ocultas contienen 64 y 32 neuronas, respectivamente. Luego, se incorpora una capa de escalado para adaptar el rango de salida de la red al rango físico permitido para la acción de control del sistema. En cuanto a la red del crítico, esta recibe las entradas tanto de las observaciones del entorno como la acción aplicada. Ambas señales son procesadas inicialmente de forma independiente mediante capas completamente conectadas de 64 neuronas y posteriormente conectadas en una única representación interna. A partir de esta combinación, la información atraviesa una nueva capa oculta de 64 neuronas con activación ReLU antes de obtener la estimación final del valor $Q(o_k, a_k)$, que representa la calidad de realizar una determinada acción dadas unas observaciones concretas.

La selección de esta arquitectura y del número de neuronas se ha realizado buscando que las redes no fueran muy complejas pero que, a su vez, tuviesen capacidad de representación. Se han empleado capas de tamaño moderado (64 y 32 neuronas) con el objetivo de proporcionar suficiente capacidad para modelar relaciones no lineales presentes en la dinámica térmica del invernadero, evitando al mismo tiempo arquitecturas excesivamente profundas que aumentasen el coste computacional o que potencialmente causen problemas de sobreajuste. La configuración final se ha ajustado empíricamente a partir de varios entrenamientos, seleccionando la estructura que ofrecía un comportamiento más estable y un mejor equilibrio entre rendimiento y tiempo de entrenamiento.

Todo esto se ha implementado en el *software* de MATLAB (The MathWorks, EE.UU.). El entrenamiento *offline* se ha realizado con un tiempo de muestreo de un minuto, igual al muestreo usado en el PI. Además, el agente ha sido entrenado durante 4000 épocas, seleccionando finalmente el agente que presentó el mejor comportamiento de acuerdo con los requerimientos establecidos.

4.3. *Evaluación de las políticas obtenidas*

Durante el entrenamiento, los agentes se fueron almacenando cada 20 épocas. Posteriormente, se realizó una validación *offline* de cada uno de ellos utilizando días distintos a los empleados durante el entrenamiento. Para ello, se comparó la señal de control generada por cada agente con la señal real calculada por el controlador PI a partir de los datos registrados en el invernadero real.

La evaluación se abordó mediante un enfoque multicriterio. Por un lado, se utilizó la Raíz del Error Cuadrático Medio (*Root Mean Square Error*, RMSE) como medida de proximidad respecto al comportamiento del controlador PI, considerado como una política de referencia ya validada. En este sentido, el RMSE no se emplea con el objetivo de forzar una reproducción exacta de dicha señal, sino para cuantificar la coherencia global del agente respecto a una estrategia de control estable. Por otro lado, se empleó el criterio del Esfuerzo de Control (EC), con el fin de evaluar la suavidad de la acción generada por el agente. Finalmente, la selección del mejor agente se realizó combinando estas métricas (calculadas como se muestra en (8)), el valor de la función Q y una inspección visual de las señales, permitiendo valorar tanto el

seguimiento global del comportamiento de referencia como la calidad dinámica de la acción de control.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(u_{PI,i} - u_{agent,i})^2}{N}}, \quad \text{EC} = \sum_{i=1}^{N} |\Delta u_{agent,i}| \quad (8)$$

Los datos utilizados para validar corresponden a los días 24 de marzo y 17 de abril de 2026. Los resultados se observan en la Figura 6 y sus métricas en la Tabla 1.

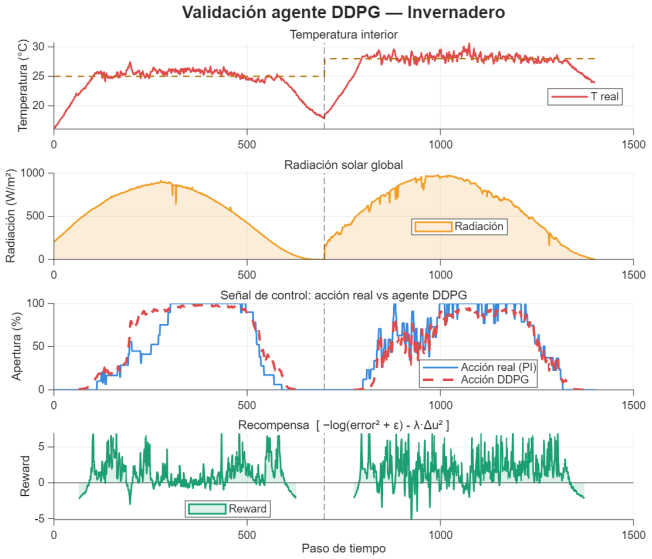

Figura 6: Comparativa PI vs. Agente DDPG.

Se muestra la comparación entre la acción de control real del PI y la señal generada por el agente DDPG durante la validación *offline*. El agente reproduce adecuadamente la tendencia global de la estrategia de control, adaptando la apertura de ventilación en función de las condiciones climáticas y de la temperatura interior, destacando especialmente su capacidad para mantener un comportamiento coherente ante cambios de consigna en la temperatura de referencia. Existen diferencias notables, sobre todo en el primer día donde, en torno al minuto 200, la temperatura interior sube repentinamente y el agente responde con una apertura abrupta, mantiéniendose en torno a esta posición. No obstante, el comportamiento del agente se mantiene coherente y no realiza una simple replicación de la señal de control. Además, la acción generada por el DDPG presenta una mayor suavidad temporal (reflejándose en el EC), reduciendo variaciones bruscas entre instantes consecutivos, traduciéndose en una señal más adecuada para actuadores reales.

Tabla 1: Métricas de validación

Controlador	RMSE	EC
PI	-	1750
DDPG	17.04	1620

5. Conclusiones y trabajos futuros

Los resultados evidencian la viabilidad del entrenamiento *offline* de un agente DDPG para el control de la temperatura utilizando la ventilación natural. El agente es capaz de reproducir una estrategia de control coherente con la política de referencia (controlador PI) y generar señales de control más suaves, lo que supone una reducción de la exigencia mecánica sobre los actuadores. Por tanto, el RL puede constituir una alternativa prometedora para el desarrollo de estrategias de control climático basadas en datos históricos de operación real.

Como trabajo futuro, se plantea la implementación y validación *online* del agente desarrollado sobre el invernadero real, permitiendo analizar la capacidad de adaptación del agente, pudiendo incluso realizar un ajuste fino de los parámetros diariamente. Además, se contempla la ampliación del conjunto de datos a distintas estaciones agrícolas para evaluar la generalización del agente bajo diversas condiciones.

Agradecimientos

Este trabajo es resultado del Proyecto PID2024-157228OB-I00 financiado por MICIU /AEI /10.13039/501100011033 / FEDER, UE. Daniel Pérez Sánchez cuenta con una ayuda 25458, financiada por la Junta de Andalucía /CUII y por el FSE+.

Referencias

Castro, A. J., López-Rodríguez, M. D., Giagnocavo, C., Gimenez, M., Céspedes, L., Calle, A. L., Gallardo, M., Pumares, P., Cabello, J., Rodríguez, E., Uclés, D., Parra, S., Casas, J., Rodríguez, F., Fernandez-Prados, J. S., Alba-Patiño, D., Expósito-Granados, M., Murillo-López, B., Vasquez, L. M., Valera, D. L., 2019. Six collective challenges for sustainability of almería greenhouse horticulture. International Journal of Environmental Research and Public Health 16 (21).

García-Mañas, F., Hägglund, T., Guzmán, J. L., Rodríguez, F., Berenguel, M., 2024. A practical solution for multivariable control of temperature and humidity in greenhouses. European Journal of Control 77, 100967.

Gil, J. D., Del Rio Chanona, E. A., 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.

González, R. A., Ramos-Teodoro, J., Rodríguez, F., Berenguel, M., 2025. Multi-resource scheduling of the water–energy–food nexus in agro-industrial environments. In: Ocampo-Martinez, C., Quijano, N. (Eds.), Energy Systems Integration for Multi-Energy Systems: From Operation to Planning in the Green Energy Context. Springer, pp. 179–204.

Li, K., Shi, J., Hu, C., Xue, W., 2025. The intelligentization process of agricultural greenhouse: A review of control strategies and modeling techniques. Agriculture 15 (20).

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

Montoya-Ríos, A. P., García-Mañas, F., Guzmán, J. L., Rodríguez, F., 2020. Simple tuning rules for feedforward compensators applied to greenhouse daytime temperature control using natural ventilation. Agronomy 10 (9).

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

Rodríguez, F., Berenguel, M., García-Mañas, F., Guzmán, J. L., Sánchez-Molina, J. A., 2024. A multilayer control architecture for greenhouse crop production in agro-industrial districts: Conceptual framework, prospects and challenges. Smart Agricultural Technology 9, 100657.

Rodríguez, F., Berenguel, M., Guzmán, J. L., Ramírez-Arias, A., 2015. Modeling and Control of Greenhouse Crop Growth. Springer.

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.

Svensen, J. L., Cheng, X., Boersma, S., Sun, C., 2024. Chance-constrained stochastic mpc of greenhouse production systems with parametric uncertainty. Computers and Electronics in Agriculture 217, 108578.

Sánchez-Molina, J. A., Gil, J. D., González, R. A., García-Mañas, F., Muñoz, M., 2025. Agroconnect research station: General description and control challenges. In: Maestre, J. M., Ocampo-Martinez, C. (Eds.), Control Systems Benchmarks. Springer, pp. 29–43.

Yusuf, A. G., Al-Yahya, F. A., Saleh, A. A., Abdel-Ghany, A. M., 2025. Optimizing greenhouse microclimate for plant pathology: challenges and cooling solutions for pathogen control in arid regions. Frontiers in Plant Science 16.