Jornadas de Automática

Refrigeración combinada en CSP: aprendizaje por refuerzo en una optimización multi-etapa

Juan Miguel Serrano[a], Juan Diego Gil[b], Patricia Palenzuela[c], Lidia Roca[c,*]

[a]*Centro Nacional de Energías Renovables (CENER), Ciudad de la Innovación, nº 7, 31621 Navarra, España.*
[b]*Departamento de Informática, CIESOL-ceiA3, Universidad de Almería, Ctra. Sacramento s/n, 04120, Almería, España.*
[c]*CIEMAT-Plataforma Solar de Almería-CIESOL, Ctra. de Senés s/n, Tabernas, 04200, Almería, España.*

To cite this article: Serrano, J.M, Gil, J.D., Palenzuela, P., Roca, L. 2026. [Refrigeración combinada en CSP: aprendizaje por refuerzo en una optimización multi-etapa]. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13808

Resumen

Con el fin de mejorar la sostenibilidad y la viabilidad económica de la energía solar térmica de concentración, se propone un sistema de refrigeración del ciclo de potencia que combina refrigeración húmeda y seca. Su eficacia se basa en aplicar estrategias de operación óptimas que equilibren los consumos de agua y electricidad. Para abordar este reto, se propone una optimización en dos etapas. En la primera, se resuelve el problema multi-objetivo para generar frentes de Pareto en el horizonte de predicción. En la segunda, estos frentes se recorren mediante un algoritmo de búsqueda que minimiza el coste económico acumulado de operación. Este trabajo se centra en la primera etapa, comparando dos metodologías para obtener los frentes de Pareto: i) una búsqueda exhaustiva, y ii) un agente de aprendizaje por refuerzo. Los resultados indican que este último método presenta un tiempo de cálculo favorable en este tipo de plantas, dada su prolongada vida útil, alcanzándose una ventaja computacional a partir del 1.46 % de dicha vida útil sin detrimento de la calidad de la solución.

Palabras clave: Optimización de proceso, Aprendizaje por refuerzo, Gestión de recursos, Sostenibilidad, Control de recursos energéticos renovables

Combined cooling in CSP: Reinforcement learning applied to a two-stage optimization

Abstract

To improve the sustainability and economic viability of concentrated solar thermal energy systems, this study proposes a cooling system that combines wet and dry cooling technologies. The effectiveness of this novel cooling solution relies on optimal operation strategies that balance water and electricity use. To address this challenge, a two-stage optimization framework is proposed. In the first stage, a sequence of multi-step optimization problems is solved to generate Pareto fronts over the prediction horizon. In the second stage, these fronts are traversed through a path-search algorithm that minimizes the operating cumulative cooling cost. This work focuses on the first stage by comparing two methodologies for obtaining Pareto fronts: (i) an exhaustive search across the entire operating range and (ii) a reinforcement learning agent. The results show that the latter significantly reduces computational time, which is particularly advantageous given the long operational lifetime of such plants, achieving a computational advantage once 1.46 % of the plant's useful life has elapsed, while maintaining a comparable solution quality.

Keywords: Process optimization, Reinforcement learning, Resource management, Sustainability, Control of renewable energy resources

*Autor para correspondencia: lidia.roca@psa.es

1. Introducción

El aumento de la demanda mundial de electricidad, impulsado por la electrificación, la digitalización y el crecimiento económico, está generando importantes desafíos para los sistemas energéticos. La expansión de la energía solar fotovoltaica ha acentuado el fenómeno de la "curva del pato", con excedentes de generación al mediodía y déficit en el atardecer, lo que dificulta la estabilidad de la red y pone de relieve la necesidad de tecnologías renovables gestionables.

A esta problemática se suma la creciente incertidumbre geopolítica, que ha evidenciado la vulnerabilidad de los sistemas dependientes de combustibles fósiles frente a interrupciones en el suministro y la volatilidad de los precios. En este contexto, las energías renovables se consolidan como una solución tanto climática como estratégica para reforzar la seguridad y la soberanía energética.

Entre ellas, la energía solar térmica de concentración (CSP por sus siglas en inglés, *Concentrated Solar Power*) destaca por su capacidad de almacenamiento térmico y generación gestionable, permitiendo producir electricidad incluso sin radiación solar y contribuyendo a la estabilidad del sistema eléctrico, tal y como se destaca en el reciente "Decálogo para impulsar la Energía Termosolar en España"(Protermosolar, 2026). España es líder mundial en esta tecnología, con una capacidad instalada de 2.3 MW.

No obstante, el desarrollo de nuevas plantas CSP depende de varios factores, entre ellos la disponibilidad de alta irradiancia solar, el terreno y el acceso al agua. Este último constituye una limitación crítica en regiones áridas y semiáridas, donde las plantas requieren importantes volúmenes de agua, principalmente para la refrigeración de los ciclos de potencia basados en tecnología Rankine.

En estos casos, la eficiencia del ciclo depende en gran parte de la temperatura de condensación del vapor, que está condicionada por la tecnología de refrigeración empleada. Las torres de enfriamiento húmedo (WCT por sus siglas en inglés, *Wet Cooling Tower*) permiten alcanzar temperaturas de condensación más bajas y, por tanto, una mayor producción eléctrica, pero requieren un elevado consumo de agua para compensar las pérdidas asociadas a la evaporación, arrastre y purgas (entre 1.8 y 4 L/kWh, (Aseri et al., 2022)). Por el contrario, los sistemas de refrigeración seca (DC por sus siglas en inglés, *Dry Cooler*) eliminan prácticamente el uso de agua al emplear el aire ambiente como medio de disipación térmica, aunque presentan una menor eficiencia energética, especialmente en condiciones de altas temperaturas ambiente, y demandan un mayor consumo de energía auxiliar para el funcionamiento de los ventiladores.

Una tercera alternativa son los sistemas combinados, que integran ambas tecnologías (WCT y DC) y son, por tanto, una solución de compromiso entre las ventajas y limitaciones de sistemas solo WCT o solo DC. Una diferencia importante de este tipo de sistemas con respecto a los tradicionales es que introducen flexibilidad en la operación: una misma carga térmica puede ser refrigerada con más de una estrategia de operación. La elección depende del criterio de consumo que se desee priorizar: minimizar el consumo de agua mediante un mayor uso del sistema seco (DC), o minimizar la penalización sobre la producción eléctrica mediante el mayor uso de WCT.

Esto supone que la aplicación de estrategias de optimización para la operación de estos sistemas sea un paso fundamental para aprovechar todo su potencial y superar en rendimiento a las alternativas aisladas. La única propuesta en la literatura que propone una estrategia de optimización para este tipo de sistemas se describe en (Serrano Rodríguez, 2026). Se trata de una estrategia de optimización con horizonte decreciente multi-etapa, donde en la primera etapa, se resuelven problemas multi-objetivo —uno por paso en el horizonte— para generar frentes de Pareto, mientras que en la segunda etapa, estos frentes se recorren mediante un algoritmo de búsqueda heurística que minimiza el coste económico de operación acumulado al final del horizonte. Este trabajo se centra en la mejora computacional de la primera etapa mediante la comparación de dos metodologías para la obtención de frentes de Pareto: (i) una búsqueda exhaustiva sobre todo el rango de operación (tal y como se plantea en (Serrano et al., 2026)) y (ii) un agente basado en aprendizaje por refuerzo.

2. Descripción del sistema

La planta piloto de refrigeración combinada de la Plataforma Solar de Almería (ver Figuras 1 y 2) es una instalación única que combina una torre de refrigeración húmeda (WCT) y un aeroenfriador (DC) dentro de una configuración hidráulica flexible, lo que permite el estudio y la validación de diversas configuraciones de refrigeración, así como estrategias de control y optimización.

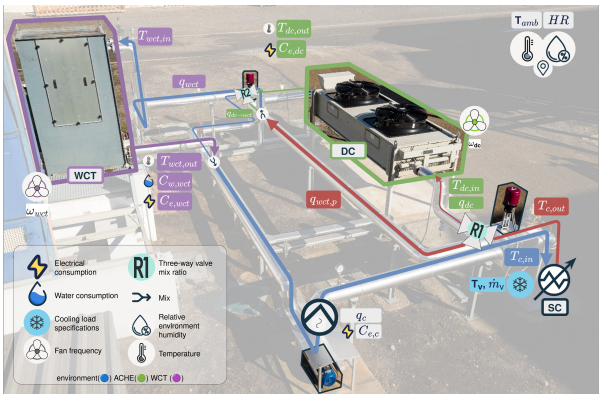

Figura 1: Imagen de la planta piloto de refrigeración combinada en las instalaciones de la Plataforma Solar de Almería

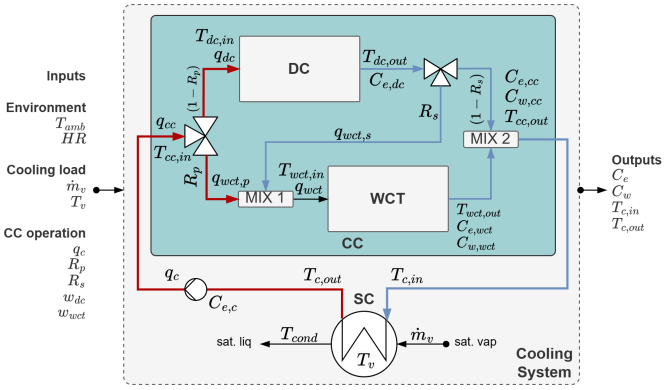

Figura 2: Esquema del circuito de refrigeración

En el circuito de refrigeración, el agua circula a través del haz de tubos de un condensador de superficie (SC, Figura 2), que puede refrigerarse mediante WCT y/o DC, ambos con una potencia térmica nominal de 204 kW_{th}. El funcionamiento del sistema puede tener distintas configuraciones hidráulicas según las posiciones de las válvulas R_p y R_s (serie, paralelo o combinaciones de estas).

2.1. Modelado del sistema de refrigeración combinado

En trabajos anteriores se desarrolló un modelo físico del sistema completo, que fue posteriormente calibrado y validado utilizando datos reales de la planta piloto (Serrano et al., 2026). Con el objetivo de reducir el coste computacional del modelo global, en este trabajo se han sustituido los componentes WCT y DC por modelos basados en datos obtenidos a partir de simulaciones de los modelos de primeros principios. En concreto, se han utilizado modelos de regresión por procesos gaussianos (GPR) (Rasmussen and Williams, 2006), obteniendo un MAPE (Error Porcentual Absoluto Medio) de 1.32 %, 1.35 % y 4.52 % en la validación de las temperaturas de salida de WCT, DC y en el consumo de agua en WCT, respectivamente. Los consumos eléctricos de cada componente son funciones polinómicas que fueron ajustadas a partir de datos experimentales.

2.2. El problema de optimización

El problema completo es un problema de optimización con horizonte decreciente donde se intenta minimizar el coste acumulado de refrigeración ($J^{(n)}$), caracterizado de forma general en la Ecuación 1.

$$\begin{aligned}\min_{\mathbf{x},\mathbf{e}} J^{(n)} &= J_e^{(n)} + J_w^{(n)} = \\ f(&\underbrace{\underbrace{\mathbf{T_{amb}}, \mathbf{HR}}_{\text{Ambiente}}, \underbrace{\dot{\mathbf{Q}}, \mathbf{T_v}}_{\text{Carga térm.}}}_{\text{Entorno, } \mathbf{e}}, \underbrace{V_{avail,0}}_{\text{Agua disp.}}, \underbrace{\mathbf{P_e}, \mathbf{P_{w,s1}}, \mathbf{P_{w,s2}}}_{\text{Económico}}, \\ &\underbrace{\mathbf{q_c}, \underbrace{\mathbf{R_p}, \mathbf{R_s}}_{\text{Conf. Hidráulica}}, \underbrace{\boldsymbol{\omega}_{dc}, \boldsymbol{\omega}_{wct}}_{\text{Refrig. Comp.}}}_{\text{Variables de decisión, } \mathbf{x}}).\end{aligned} \tag{1}$$

Este problema se compone de la suma de los costes acumulados asociados al consumo de agua ($J_w^{(n)}$) y eléctrico ($J_e^{(n)}$). Estos costes dependen a su vez de su precio asociado en cada paso (P_x). Para el recurso agua se considera que existen dos fuentes; una más económica ($P_{w,s1}$) pero de disponibilidad limitada y fijada al inicio del horizonte ($V_{avail,0}$), y otra ilimitada pero con coste asociado significativamente más elevado ($P_{w,s2}$).

Como se ha mencionado en la introducción, la resolución del problema se basa en la descomposición del problema definido en la Ecuación 1 en dos etapas. En la primera etapa, se resuelve el problema multi-objetivo para cada paso del horizonte de predicción (cada hora en un horizonte decreciente de 24 horas inicialmente), generando un número de frentes de Pareto. En la segunda etapa se formula un problema de optimización para decidir los puntos de operación en la secuencia de frentes de Paretos creados. Este trabajo se centra en la primera etapa, la resolución de problemas multi-objetivo por paso del horizonte. Este se puede definir formalmente para un paso particular en el horizonte como se muestra en la Ecuación 2.

$$\min_{\mathbf{x},\mathbf{e}} C_e, C_w = f(\underbrace{\underbrace{T_{amb}, HR}_{\text{Ambiente}}, \underbrace{\dot{Q}, T_v}_{\text{Carga térm.}}}_{\text{Entorno, } \mathbf{e}}, q_c, \underbrace{\underbrace{R_p, R_s}_{\text{Conf. Hidráulica}}, \underbrace{\omega_{dc}, \omega_{wct}}_{\text{Refrig. Comp.}}}_{\text{Variables de decisión, } \mathbf{x}}), \tag{2}$$

donde C_e y C_w son los consumos de electricidad y agua, respectivamente. El conjunto de variables de decisión mostrado en la Ecuación 2 está formado por variables operacionales del sistema, como son el caudal que circula por el circuito de refrigeración (q_c), la configuración hidráulica del sistema combinado (R_p y R_s) y la velocidad de la ventilación forzada de cada uno de los componentes ($\omega_{dc}, \omega_{wct}$). En cuanto al entorno (e), dada una planta particular situada en una ubicación determinada, queda definido un espacio $\mathbb{R}^4$ de todos los posibles escenarios del entorno. Este entorno viene caracterizado por las condiciones ambientales de temperatura y humedad relativa (T_{amb} y HR), así como la potencia térmica y la temperatura del vapor a refrigerar ($\dot{Q}$ y T_v).

Una explicación más detallada del problema y la justificación de la estrategia adoptada para su resolución se puede encontrar en (Serrano Rodríguez, 2026).

3. Métodos para la generación de frentes de Pareto

3.1. Búsqueda global

La búsqueda global (GS, *grid search*) es un método de optimización basado en la evaluación exhaustiva del espacio de decisión mediante una discretización previa de las variables de diseño. Para ello, se define un conjunto finito de valores posibles para cada variable y se evalúan todas las combinaciones resultantes, obteniendo así una representación completa del espacio de soluciones discretizado. Este enfoque permite identificar de forma sistemática las soluciones no dominadas y construir un frente de Pareto de referencia, siempre que la discretización sea lo suficientemente fina. Aunque se trata de un método conceptualmente simple y altamente paralelizable, su coste computacional crece exponencialmente con el número de variables y el nivel de discretización, lo que limita su aplicabilidad en problemas de alta dimensión.

3.2. Aprendizaje por refuerzo

El aprendizaje por refuerzo (RL, por sus siglas en inglés, *Reinforcement Learning*) es una rama del aprendizaje automático en la que un agente o controlador aprende a seleccionar acciones de control mediante la interacción con el sistema, con el objetivo de maximizar una función de recompensa que evalúa su desempeño. Una descripción detallada de sus fundamentos puede encontrarse en (Sutton and Barto, 2018). Entre los algoritmos de RL de última generación destaca Soft Actor-Critic (SAC), un método actor-crítico del tipo *off-policy* basado en el principio de máxima entropía, que combina estabilidad en el aprendizaje con una exploración eficiente del espacio de estados y acciones (Haarnoja et al., 2018).

Tal y como se muestra en la Figura 3 la arquitectura actor-crítico se compone de dos elementos: un actor, responsable de tomar decisiones sobre las acciones, y un crítico, que estima el

retorno esperado o valor asociado a dicha acción, proporcionando una señal de aprendizaje que permite mejorar la política del actor. Los parámetros θ y ϕ denotan los pesos para las redes del actor y crítico, respectivamente, y se actualizan durante el entrenamiento para optimizar el proceso (Gil et al., 2026).

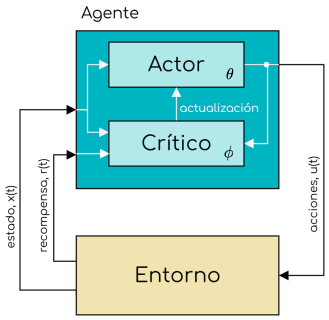

Figura 3: Diagrama del agente y su entorno.

Cuando el agente ejecuta una acción de control,$u(t)$, modifica el estado del entorno, $x(t)$, generando una recompensa, $r(t)$, asociada a dicha acción. A partir de esta información, el agente actualiza su conocimiento y determina las acciones que deberá ejecutar en situaciones futuras. Las acciones que producen recompensas más elevadas son consideradas más favorables y tienen una mayor probabilidad de ser seleccionadas por el agente durante el entrenamiento.

La Figura 3 muestra de manera general los componentes de un algoritmo de aprendizaje por refuerzo utilizando el método actor-crítico. Para el caso del sistema de refrigeración combinada analizado en este trabajo, el entorno y las variables involucradas se describen a continuación.

- Observaciones o estados, x(t). El vector de observación está formado por variables que pueden medirse directamente en el sistema real. Estas son: la temperatura ambiente, T_{amb}, humedad relativa, HR, flujo másico de vapor, $\dot{m}_v$, y temperatura del vapor, T_v.

- Acciones, u(t). El espacio de acción en esta formulación viene definido por aquellas variables que definen la configuración hidráulica y la potencia de enfriamiento. Estas variables son: caudal de agua que circula por el condensador, q_c, posición de válvulas de reparto paralelo y serie, R_p y R_s, y velocidad de los ventiladores de DC y WCT, ω_{dc} y ω_{wct}.

- Recompensa, $r(t)$. Cuando el entorno devuelve una solución factible, la recompensa se calcula como:

$$r(t) = -(\alpha \cdot C_{e,norm} + (1-\alpha) \cdot C_{w,norm}) \in [-1, 0], \quad (3)$$

  donde el parámetro α permite ajustar la política entre los extremos y la normalización es min-max sobre los rangos experimentales, acotando $r(t)$ a [-1, 0]:

$$\begin{aligned} C_{e,\text{norm}} &= \frac{C_e - C_{e,\text{mín}}}{C_{e,\text{máx}} - C_{e,\text{mín}}}, \\ C_{w,\text{norm}} &= \frac{C_w}{C_{w,\text{máx}}}. \end{aligned} \quad (4)$$

Las acciones que se detectan como inválidas (por no cumplir con las restricciones impuestas) reciben una recompensa $r(t)$=-2.

Para obtener el frente de Pareto para unas condiciones dadas de contorno, es necesario realizar el entrenamiento del agente n_{rl} veces. Un valor más elevado de este parámetro supone una mejor cobertura del frente de Pareto.

3.3. *Criterios para comparativa de métodos*

Para comparar las dos alternativas previamente mencionadas, se utilizan dos criterios de evaluación: la calidad de los frentes de Pareto generados y el presupuesto computacional.

En términos de calidad, se asume que la búsqueda global, dado un nivel de discretización suficientemente elevado, produce frentes de Pareto muy próximos al frente de Pareto ideal. Por ello, la primera comparación se realiza en función de la distancia entre el frente de Pareto generado por la alternativa basada en RL y este frente de referencia.

La calidad del frente de Pareto aproximado se evalúa mediante la métrica *Generational Distance* (GD), utilizada para cuantificar la convergencia respecto a un frente de referencia. Esta métrica se obtiene como la distancia media (en norma p, usualmente $p = 2$) entre cada solución del frente aproximado y su punto más cercano en el frente de referencia:

$$GD_{\text{norm}} = \left(\frac{1}{n_{rl}} \sum_{i=1}^{n_{rl}} \tilde{d}_i^{\,p} \right)^{1/p}, \quad (5)$$

donde n_{rl} es el número de soluciones del frente aproximado (el número de agentes entrenados), y $\tilde{d}_i$ representa la distancia mínima desde la solución i al frente de referencia en el espacio de objetivos.

Dado que los objetivos presentan unidades distintas, se aplica una normalización previa basada en los valores mínimo y máximo del frente de referencia. Finalmente, se introduce una transformación acotada definida como:

$$S_{\text{GD}} = \frac{1}{1 + GD_{\text{norm}}}, \quad (6)$$

que asigna valores cercanos a 1 a mejores aproximaciones del frente de Pareto.

Como nota, se emplea la versión normalizada de GD en lugar del hipervolumen, ya que en este estudio solo se han entrenado un número reducido de agentes como demostración de concepto, lo que implica una cobertura limitada del frente de Pareto. En este contexto, el hipervolumen puede penalizar excesivamente la falta de exploración del espacio objetivo, incluso cuando las soluciones obtenidas se encuentran próximas al frente de referencia. Por el contrario, el GD normalizado se centra en la proximidad media de las soluciones al frente de referencia, siendo más adecuado para comparar la calidad de aproximación en escenarios con un número reducido de políticas evaluadas.

En cuanto al presupuesto computacional, para cada alternativa se mide el tiempo de computación acumulado definido como $t_{ac} = t_0 + n_{eval} \cdot t_{eval}$, donde t_0, t_{eval} y n_{eval} son:

- **Tiempo de preparación o puesta en marcha** (t_0). Se define como el tiempo computacional necesario para

disponer del método empleado en condiciones de resolver el problema de optimización. En el caso de RL, este tiempo corresponde al entrenamiento. En la búsqueda global, este tiempo se considera nulo.

- **Tiempo por evaluación** (t_{eval}). Se define como el tiempo empleado por cada alternativa en resolver una instancia del problema, tomando como entrada un entorno y generando como resultado un frente de Pareto. En el caso de RL, aunque la aproximación del frente de Pareto requiere la evaluación de múltiples agentes, el tiempo total es equivalente al de una única evaluación, ya que estas se realizan en paralelo. En el caso de la búsqueda global, corresponde al tiempo invertido (considerando el máximo grado de paralelización posible) en generar un único frente de Pareto.
- **Número de evaluaciones por día de operación** ($n_{\text{eval/día}}$). Dada la estrategia descrita en la Sección 2.2 de horizonte decreciente, el número de evaluaciones por día operativo se puede calcular como:

$$n_{\text{evals/día}} = \sum_{k=n_{\text{horizon}}}^{1} k = \frac{n_{\text{horizon}}(n_{\text{horizon}}+1)}{2}, \quad (7)$$

donde n_{horizon} representa el número de pasos en el horizonte de predicción inicial. Aunque el valor de este parámetro es el mismo en ambos métodos, el tiempo acumulado obtenido difiere según el método empleado.

4. Resultados

Para comparar las dos alternativas propuestas, se analizan dos casos de estudio con distintas condiciones de entorno:

1. $T_{amb} = 30\,^{o}\text{C}$, $HR = 40\,\%$, $T_v = 42\,^{o}\text{C}$, $\dot{Q} = 200\,\text{kW}$
2. $T_{amb} = 30\,^{o}\text{C}$, $HR = 40\,\%$, $T_v = 42\,^{o}\text{C}$, $\dot{Q} = 120\,\text{kW}$

La Figura 4 muestra los puntos obtenidos con ambos métodos, siendo los valores de α empleados para entrenar al agente RL los siguientes: 0.25, 0.5, 0.75 y 0.99.

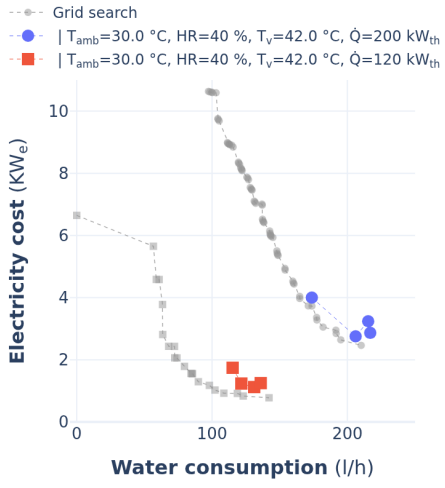

Figura 4: Frentes de Pareto obtenidos para dos escenarios (símbolos) representativos con búsqueda global (color gris) y aprendizaje por refuerzo (colores azul y rojo). Los círculos y cuadrados representan los casos de potencia térmica 200 kW_{th} y 120 kW_{th}, respectivamente.

Se puede apreciar cómo, en ambos casos de estudio, la búsqueda global genera frentes de Pareto de mayor calidad que el aprendizaje por refuerzo, el cual genera una aproximación adecuada. Sin embargo, aunque se ha intentado recorrer los frentes de Pareto mediante la variación del parámetro α, los puntos obtenidos se concentran en la región derecha. Al estar definiendo la recompensa como un valor mono-objetivo, no se puede garantizar la cobertura uniforme del frente de Pareto. Esto sugiere la necesidad de modificar la función de recompensa o, alternativamente, emplear estrategias de aprendizaje por refuerzo multi-objetivo.

4.1. Calidad

La Tabla 1 muestra los criterios de calidad y las variables de decisión obtenidas para cada caso de estudio y valor de α. Para la comparación de las variables de decisión, se consideran aquellos puntos cuyos valores de r son los más próximos.

Tabla 1: Calidad de frentes de Pareto generados por alternativa RL (S_{GD}) y comparativa de variables de decisión para los dos escenarios analizados.

	α/r	S_{GD}	Variables de decisión (GS - RL)				
			$q_c(m^3/h)$	R_p	R_s	$\omega_{dc}(\%)$	$\omega_{wct}(\%)$
1	0.25 / -0.4	0.93	14 - 19	0 - 0	1 - 1	50 - 49	54 - 39
	0.50 / -0.4		20 - 19	0 - 0	0.7 - 0.9	70 - 26	37 - 48
	0.75 / -0.3		16 - 17	0 - 0	0.9 - 0.9	50 - 26	48 - 51
	0.99 / -0.2		16 - 17	0 - 0	1 - 1	40 - 31	48 - 50
2	0.25 /-0.3	0.91	8 - 15	0 - 0	1 - 0.9	21 - 22	37 - 24
	0.50 /-0.2		16 -13	0 - 0	0 - 0.9	89 - 13	0 - 28
	0.75 /-0.1		10 - 13	0 - 0	1 - 1	50 - 19	23 - 27
	0.99 /-0.04		8 - 12	0 - 0	1 - 1	40 - 15	32 - 28

Se observa que se obtienen valores cercanos a 1 en S_{GD}, lo que indica que los valores obtenidos con RL se aproximan al ideal obtenido con GS. Los valores de las variables de decisión son, en la mayoría de los casos, comparables; no obstante, se demuestra que distintas combinaciones de estas variables pueden dar lugar a valores de r similares.

4.2. Tiempos

En la Figura 5 se muestra la evolución del tiempo acumulado de computación comparando la alternativa de búsqueda global (GS) y la alternativa propuesta en este trabajo basada en aprendizaje por refuerzo (RL).

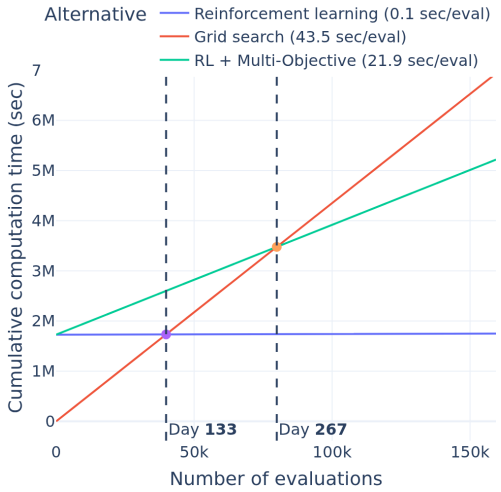

Figura 5: Evolución del tiempo de computación acumulado para las distintas estrategias.

Esta segunda se caracteriza por un tiempo computacional considerable y previo a realizar ninguna evaluación del problema de optimización: $t_{0,RL} \times n_{rl} = 48$ horas $\times$ 10 (asumiendo 10 puntos del Pareto bien repartidos). Sin embargo, su tiempo por evaluación es órdenes de magnitud inferior (0.1 segundos/evaluación) respecto al tiempo de GS (43.5 segundos/evaluación). Esto lleva a que inicialmente la ventaja computacional sea para el método de búsqueda global. Sin embargo, transcurridas 39850 evaluaciones se produce el cruce en coste computacional de ambos métodos, y a partir de ese punto la ventaja corresponde a RL.

Para evaluar la relevancia de este cruce, se calcula el tiempo de operación para que se produzca y se pondera con el tiempo de vida de la planta. Para una planta con tiempo de vida de 25 años, con evaluaciones de la estrategia realizadas cada hora, y para un horizonte de 24 horas, resulta en 300 evaluaciones diarias (Ecuación 7), o $2.74{\cdot}10^6$ evaluaciones a lo largo de la vida de la planta. En términos de tiempo de operación se produce el cruce transcurridos 133 días de operación. En términos relativos, esto supone que la ventaja computacional se alcanza al 1.46 % de la vida de la planta, lo que ampliamente justifica la inversión en el entrenamiento de los agentes.

Si el frente de Pareto obtenido mediante RL se refinara posteriormente con un algoritmo heurístico, habría que añadir su coste computacional. Suponiendo un tiempo de refinamiento equivalente a la mitad del tiempo de evaluación de la búsqueda global, la intersección se retrasaría hasta el 2.92 % de la vida útil de la planta (267 días, Figura 5), manteniéndose como una alternativa ventajosa.

5. Conclusiones

Este trabajo plantea el problema de optimización de operación de un sistema de refrigeración combinada para plantas CSP. En concreto, se comparan dos métodos para obtener los frente de Pareto necesarios en una optimización multi-etapa; búsqueda global y aprendizaje por refuerzo. Dado que la optimización en este tipo de sistemas se debe realizar en paralelo a la operación de la planta, cuya vida útil puede extenderse durante varias décadas, resulta razonable adoptar estrategias que concentren el coste computacional en las fases iniciales de la implementación. A cambio, estas estrategias permiten disponer de evaluaciones de bajo coste computacional durante la operación, facilitando su aplicación en tiempo real.

Además, el enfoque permite desacoplar las fases de entrenamiento y operación. De esta forma, se podría emplear un equipo especializado dotado de los recursos de hardware necesarios para desarrollar el agente en una planta dada, mientras que la evaluación en línea podría ser realizada por hardware de capacidad modesta, ubicado en la propia planta sin depender de servidores remotos.

Asimismo, el empleo de agentes previamente entrenados proporciona tiempos de evaluación predecibles, una característica especialmente relevante en aplicaciones industriales con requisitos operativos estrictos.

Por otro lado, esta mejora en la eficiencia computacional permite la extensión de la optimización a capas superiores, como estrategias de gestión del recurso agua mediante la planificación de su distribución, las cuales requieren una capa inferior de optimización de la operación rápida y fiable para que su implementación resulte viable en la práctica.

Entre las limitaciones encontradas, definir la recompensa como un solo escalar (resultado mono-objetivo) limita la exploración total del frente de Pareto, sugiriendo la necesidad de definir la recompensa como un vector de mejores soluciones o emplear directamente algoritmos de RL multi-objetivo. Además, sería interesante comparar los resultados con algoritmos multi-objetivo tradicionales, como la variante NSGA-II previamente empleada en (Serrano et al., 2024). Cabe destacar también que el agente basado en RL se entrena a partir de un modelo que representa el comportamiento del sistema de enfriamiento. Sin embargo, las características de este sistema pueden variar con el tiempo debido al envejecimiento y suciedad de los equipos. Estas variaciones pueden provocar una pérdida de rendimiento del agente, por lo que resulta recomendable monitorizar de forma continua el desempeño del controlador y contemplar estrategias de reentrenamiento o aprendizaje adaptativo que permitan mantener la eficacia de la política a lo largo de la vida útil de la planta.

Este trabajo constituye una primera aproximación a una solución alternativa para resolver de manera eficiente la primera etapa del problema de optimización propuesto. Como líneas futuras de investigación, sería de interés realizar una comparación más exhaustiva variando el nivel de discretización de la búsqueda global para encontrar un compromiso entre calidad de Pareto obtenido y tiempo por evaluación. Además, se puede optimizar el entrenamiento de los agentes haciendo uso de modelos y algoritmos alternativos, y optimizando los parámetros de entrenamiento.

Agradecimientos

Esta publicación es parte del proyecto I+D+i PID2021-126452OA-I00, financiado por MCIN/ AEI/ 10.13039/ 501100011033/ y “FEDER Una manera de hacer Europa”.

Referencias

Aseri, T. K., Sharma, C., Kandpal, T. C., 2022. Condenser cooling technologies for concentrating solar power plants: A review 24 (4), 4511–4565.

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.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning (ICML). pp. 1861–1870.

Protermosolar, Mar. 2026. Decálogo para impulsar la energía termosolar en España. Publicado el 18 de marzo de 2026.

Rasmussen, C. E., Williams, C. K. I., 2006. Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press.

Serrano, J. M., Gil, J. D., Bonilla, J., Palenzuela, P., Roca, L., 2024. Optimal operation of a combined cooling system. In: IFAC-PapersOnLine. Vol. 58. pp. 460–465.

Serrano, J. M., Palenzuela, P., Ruiz, J., Navarro, P., Muñoz-Cámara, J., Ortega-Delgado, B., Roca, L., Jan. 2026. Combined cooling for CSP plants: Modeling, experimental validation and optimization analysis. Energy Conversion and Management 348, 120752.

Serrano Rodríguez, J. M., 2026. Towards optimal resource management in solar thermal applications: CSP and desalination. Ph.D. thesis, University of Almería.

Sutton, R. S., Barto, A. G., 2018. Reinforcement Learning: An Introduction, 2nd Edition. MIT Press.