EuroSun 2026

STATE-AWARE DEEP REINFORCEMENT LEARNING FOR REAL-TIME RECONFIGURATION OF DIRECTLY COUPLED PV–PEM SYSTEMS

Direct coupling of photovoltaic (PV) arrays with Proton Exchange Membrane (PEM) electrolyzers eliminates DC–DC conversion losses but suffers from frequent electrical mismatch due to the non-linear and weather-dependent nature of PV generation. Recent literature solutions rely on computationally intensive multi-objective optimization and repeated I–V curve intersection calculations, which are unsuitable for real-time switching control. This work proposes a Deep Reinforcement Learning (DRL) framework based on Proximal Policy Optimization (PPO) that reformulates PV–PEM matching as a discrete dynamic control problem. The RL agent directly controls MOSFET switching to reconfigure the number of active PV strings, learning an optimal policy through a physics-informed reward function that embeds efficiency and switching fatigue. By shifting computational complexity to offline training, the proposed approach enables real-time, sub-second PV array reconfiguration while maintaining physical feasibility and high solar-to-hydrogen efficiency.

Keywords: Deep reinforcement learning, PV–PEM coupling, PV arrays reconfiguration, MOSFETs control, Proximal Policy Optimization,

Field of application for the findings: Real-time control of PV–electrolyzer systems, power-to-hydrogen applications, and AI-assisted energy system operation.

Target audience: Researchers and practitioners in renewable energy systems, hydrogen technologies, and machine learning–based control, as well as system integrators developing real-time PV–electrolyzer solutions.

1. Introduction

Direct coupling of PV arrays and PEM electrolyzers is a highly attractive architecture. As highlighted by researchers such as Mas et al. (Mas et al., 2022) and Coppitters et al. (Coppitters et al., 2019), it eliminates the need for intermediate DC-DC converters, which subsequently reduces system complexity, capital costs, and transmission losses. Furthermore, as noted by Di Caro and Vitale (Di Caro & Vitale, 2024), eliminating the converter avoids the introduction of ripple currents, a phenomenon known to deteriorate the PEM membrane and reduce system reliability. However, the fundamental drawback of direct coupling, addressed extensively by Paul and Maroufmashat et al. (Maroufmashat et al., 2014), is the frequent electrical mismatch between the non-linear I-V characteristics of the PV array and the electrolyzer's load curve. Because PV generation is highly weather-dependent, variations in solar irradiance and temperature cause the system to frequently operate away from the Maximum Power Point (MPP), resulting in significant energy transfer losses.

To mitigate this mismatch and optimize system performance, researchers have traditionally focused on capacity sizing and static configuration optimization, such as defining the optimal number of series and parallel cells. As reviewed by Soyturk et al. (Soyturk et al., 2024), these conventional solutions predominantly rely on multi-objective optimization (MOO) techniques and metaheuristic algorithms. For instance, Mas et al. (Mas et al., 2022) utilized Genetic Algorithms (GA) for sizing optimization, Sayedin et al. applied Particle Swarm Optimization (PSO), and Maroufmashat et al. (Maroufmashat et al., 2014) employed the Imperialist Competitive Algorithm (ICA). To find optimal operating points, these models iteratively calculate the intersection points of the PV and PEM I-V curves. However, these traditional optimization methods are highly computationally intensive. As detailed by Mas et al. (Mas et al., 2022), algorithms like GA require evaluating large population sizes across multiple stalled generations, demanding immense computational resources and execution time. While these algorithms are effective for offline system planning and capacity configuration, their heavy computational burden makes them fundamentally unsuitable for real-time switching control.

To improve direct coupling efficiency under varying weather conditions, recent advancements have proposed dynamically reconfiguring the PV source itself. A notable approach by Di Caro and Vitale (Di Caro

& Vitale, 2024) introduces a direct coupling technique enhanced by a reconfigurable PV source, which utilizes power MOSFETs as static switches to actively alter the parallel connections of PV arrays depending on real-time solar radiation. This dynamic reconfiguration successfully shifts the I-V curve to maintain optimal matching without relying on a DC-DC converter. However, current implementations of this technique rely on basic threshold comparators. These static, rule-based switches lack the analytical depth and continuous adaptability required for truly optimal real-time switching.

Concurrently, Artificial Intelligence (AI) and Deep Learning (DL) have been introduced into solar-hydrogen systems to handle highly complex and non-linear data. Yet, as demonstrated in the work of Mert (Mert, 2021), the application of DL and Artificial Neural Networks (ANNs) in this domain has been largely restricted to predicting and forecasting solar irradiance or estimating hydrogen production yields. The use of advanced machine learning techniques for the real-time dynamic control and configuration switching of PV-PEM systems remains an unexplored frontier.

Based on the literature, a critical research gap is evident. While dynamic PV reconfiguration via MOSFETs, as proposed by Di Caro and Vitale (Di Caro & Vitale, 2024), is a proven physical solution for mitigating PV-PEM mismatch, current control mechanisms rely on simplistic, static thresholds. Conversely, accurate I-V matching relies on complex MOO algorithms (like the GA and PSO methods used by Mas et al. (Mas et al., 2022) and Maroufmashat et al. (Maroufmashat et al., 2014) that are far too computationally intensive for real-time, sub-second control.

Our proposed DRL framework, based on PPO, directly bridges this gap. By reformulating PV-PEM matching as a discrete dynamic control problem, we leverage the physical advantages of MOSFET-driven PV string reconfiguration while bypassing the limitations of simplified threshold comparators. Most importantly, by shifting the heavy computational complexity to an offline training phase, our RL agent avoids the real-time computational bottlenecks of repeated I-V curve intersection calculations seen in traditional GA and PSO methods. This enables a novel, physics-informed solution that achieves real-time, sub-second PV array reconfiguration to maintain physical feasibility.

2. Methodology

The system directly couples a PV array to a PEM electrolyzer without a DC–DC converter. Optimal performance requires modulating parallel PV strings to align the array's I–V curve with the electrolyzer's semi-empirical polarization curve. The baseline Multi-Objective Optimization (MOO) approach is computationally prohibitive for real-time hardware control, as it requires recursive solving of transcendental diode equations for every MOSFET state. Additionally, this steady-state method ignores temporal factors such as hardware fatigue from excessive switching and hydrogen storage limits. Consequently, the control problem is reframed as a Markov Decision Process (MDP) where a PPO agent learns a dynamic policy to map environmental and temporal states directly to hardware actions.

The RL agent operates on a normalized, seven-dimensional state vector ($[0, 1]$) comprising irradiance (G), temperature (T), previous electrical states (V, I), current string configuration, hydrogen storage state of charge (SoC), and the hour of the day. This provides the temporal awareness necessary to anticipate operational constraints. The action space is a discrete set of switching commands determining the number of active parallel PV strings. The reward function incentivizes hydrogen production and solar-to-hydrogen efficiency while penalizing electrical mismatch (voltage gap). To ensure long-term feasibility, penalties are applied for excessive switching to limit MOSFET fatigue and for violating storage capacities or physical PEM current limits. If an action leads to a physically infeasible state where curves do not intersect, hydrogen production is set to zero, and a severe penalty is applied.

The agent is trained offline using historical weather datasets, where the environment simulates high-fidelity I–V curve physics to calculate rewards. Through this process, the PPO agent learns the non-linear relationships between environmental fluctuations and optimal topology. During deployment, the heavy physics calculations are bypassed and the trained actor network executes a single forward pass to output the optimal configuration instantly. This transitions the system from a burdensome recursive solver to a lightweight, predictive, and real-time AI controller.

3. Preliminary Results and Discussion

The performance of the DRL framework was evaluated through an intensive training phase and a comparative analysis against the baseline MOO approach. Figure 1 illustrates the training progression of the PPO agent over 300 episodes. The "Episode Reward" curve demonstrates a rapid and stable convergence, transitioning from negative values to a sustained plateau after approximately 100 episodes. This indicates that the agent successfully learned to balance the competing objectives of maximizing hydrogen production and solar-to-hydrogen efficiency while minimizing electrical mismatch and switching fatigue. The "Total H_2/episode" plot shows a corresponding increase, stabilizing around 325 kg (-12.93 % compared to MOO), while the "Avg StH" stabilizes at 15.60 % (1.66 % compared to MOO). The convergence of these metrics suggests that the physics-informed reward function effectively guided the agent toward a policy that prioritizes long-term efficiency over short-term production spikes.

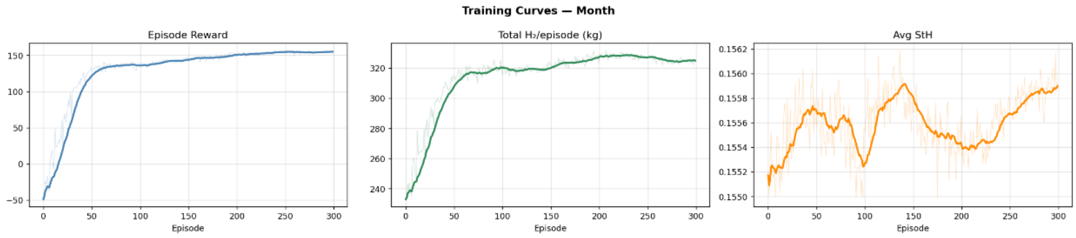

Figure 1: Training Curves

The real-time operational behavior of the trained agent is showcased in Figure 2, which compares the PV string reconfiguration choices of the RL agent against the MOO baseline over a full-month simulation (744 hours). The plot reveals that the RL agent (indicated by the red dashed line) closely tracks the behavior of the computationally intensive MOO solver (blue solid line), particularly during peak solar hours. However, a key distinction is observed in the RL agent's proactive management of switching transitions. While the MOO approach solves for a theoretical steady-state optimum at every discrete time step, the RL agent frequently adopts more conservative switching patterns to avoid excessive MOSFET fatigue, a behavior incentivized by the switching penalty in the reward function.

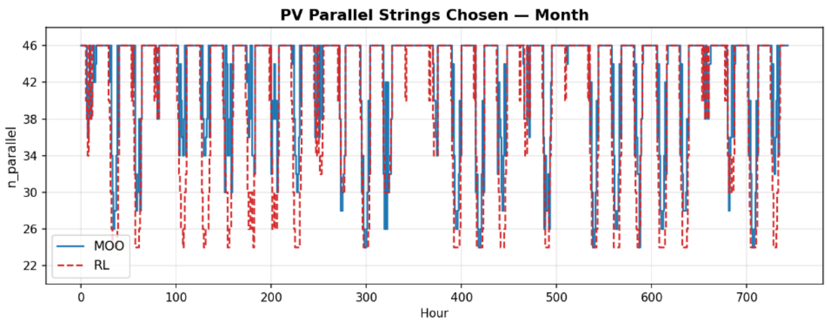

Figure 2: Behavior comparison

The RL agent's "state-aware" capability is particularly evident during low-irradiance transitions, where it maintains the PV voltage within the PEM operational window. Unlike the MOO solver, which reacts to infinitesimal irradiance changes, the RL policy employs temporal smoothing. This reduces unnecessary switching and enhances hardware durability while achieving near-optimal efficiency without the computational overhead of iterative curve-fitting, proving its viability for sub-second PV–PEM control. Furthermore, it is anticipated that performance metrics will continue to improve as the implementation is refined and the agent undergoes more extensive training iterations across broader climatic datasets.

Future research will expand this framework toward predictive control using weather forecasts and hardware degradation modeling via a "State of Health" variable. Additionally, we plan to utilize Explainable AI (XAI) to extract simplified mathematical rules from the neural network's policy. This will facilitate the deployment of optimized, physics-informed control strategies in industrial hardware without the need for complex AI infrastructure.

4. References

Coppitters, D., De Paepe, W., & Contino, F. (2019). Surrogate-assisted robust design optimization and global sensitivity analysis of a directly coupled photovoltaic-electrolyzer system under techno-economic uncertainty. *Applied Energy*, *248*, 310–320. https://doi.org/10.1016/j.apenergy.2019.04.101

Di Caro, A., & Vitale, G. (2024). Direct-Coupled Improvement of a Solar-Powered Proton Exchange Membrane Electrolyzer by a Reconfigurable Source. *Clean Technologies*, *6*(3), 1203–1228. https://doi.org/10.3390/cleantechnol6030059

Maroufmashat, A., Sayedin, F., & Khavas, S. S. (2014). An imperialist competitive algorithm approach for multi-objective optimization of direct coupling photovoltaic-electrolyzer systems. *International Journal of Hydrogen Energy*, *39*(33), 18743–18757. https://doi.org/10.1016/j.ijhydene.2014.08.125

Mas, R., Berastain, A., Antoniou, A., Angeles, L., Valencia, S., & Celis, C. (2022). Genetic algorithms-based size optimization of directly and indirectly coupled photovoltaic-electrolyzer systems. *Energy Conversion and Management*, *270*. https://doi.org/10.1016/j.enconman.2022.116213

Mert, İ. (2021). Agnostic deep neural network approach to the estimation of hydrogen production for solar-powered systems. *International Journal of Hydrogen Energy*, *46*(9), 6272–6285. https://doi.org/10.1016/j.ijhydene.2020.11.161

Soyturk, G., Cetinkaya, S. A., Aslani Yekta, M., Kheiri Joghan, M. M., Mohebi, H., Kizilkan, O., Ghandehariun, A. M., Colpan, C. O., Acar, C., & Ghandehariun, S. (2024). Dynamic analysis and multi-objective optimization of solar and hydrogen energy-based systems for residential applications: A review. *International Journal of Hydrogen Energy*, *75*, 662–689. https://doi.org/10.1016/j.ijhydene.2024.05.095