frontiers | Frontiers in Water

TYPE Original Research
PUBLISHED 01 October 2025
DOI 10.3389/frwa.2025.1649284

OPEN ACCESS

EDITED BY
Evan G. R. Davies,
University of Alberta, Canada

REVIEWED BY
Ying Zhu,
Xi'an University of Architecture and Technology, China
Paweł Tomczyk,
Wroclaw University of Environmental and Life Sciences, Poland

*CORRESPONDENCE
Angel Udias
✉ angel.udias@ec.europa.eu

RECEIVED 18 June 2025
ACCEPTED 28 August 2025
PUBLISHED 01 October 2025

CITATION
Campo Carrera JM and Udias A (2025) Deep Reinforcement Learning for complex hydropower management: evaluating Soft Actor-Critic with a learned system dynamics model. *Front. Water* 7:1649284.
doi: 10.3389/frwa.2025.1649284

Deep Reinforcement Learning for complex hydropower management: evaluating Soft Actor-Critic with a learned system dynamics model

José María Campo Carrera[1,2] and Angel Udias[3,4]*

[1]Universidad de Alcalá, Alcalá de Henares, Spain, [2]Corporación Eléctrica del Ecuador (CELEC EP), Unidad de Negocio Hidronación, Guayaquil, Ecuador, [3]European Commission Joint Research Centre, Ispra, Italy, [4]Rey Juan Carlos University, Móstoles, Spain

Introduction: Optimizing the operation of interconnected hydropower systems presents significant challenges due to complex non-linear dynamics, hydrological uncertainty, and the need to balance competing objectives like economic maximization and operational safety. Traditional optimization methods often struggle with these complexities, particularly for high-resolution intraday decision-making.

Methods: This paper proposes and evaluates a Deep Reinforcement Learning (DRL) framework, specifically utilizing the Soft Actor-Critic (SAC) algorithm, to optimize the hourly operation of the Baba hydropower facility and its strategic water transfers to the downstream Marcel Laniado De Wind (MLDW) system in Ecuador's Guayas basin. A key component of our approach is a custom Gymnasium simulation environment incorporating a validated internal dynamics model based on a pre-trained neural network. This learned model, developed using historical inflow data, accurately simulates the system's hydraulic and energy state transitions. The SAC agent was trained within this environment using synthetically generated data (KNN-resampled) to learn policies that maximize the combined economic revenue from Baba generation and the estimated downstream MLDW generation benefit, while adhering to stringent operational and safety constraints.

Results: Results demonstrate that the learned SAC policies significantly outperform historical operations, achieving up to a 9.43% increase in total accumulated economic gain over a decade-long validation period. Furthermore, the agent effectively learned to manage constraints, notably reducing peak uncontrolled spillway discharges by up to 9%.

Discussion: This study validates the effectiveness of SAC combined with a learned internal dynamics model as a robust, data-driven approach for optimizing complex, interconnected hydropower systems, offering a promising pathway toward more efficient and resilient water resource management.

KEYWORDS

water resources management, reservoir operation, Deep Reinforcement Learning, hydropower optimization, Soft Actor-Critic, Guayas River

1 Introduction

Hydropower serves as a cornerstone in the global transition toward sustainable energy portfolios, valued not only for its renewable generation capacity but also for its essential grid regulation services and integrated water resource management capabilities (Bautista et al., 2022). However, the optimal operation of

hydropower systems presents a significant challenge, necessitating a complex trade-off between maximizing economic revenue and adhering to a multitude of critical constraints. These include safeguarding dam structural integrity, respecting dynamic hydrological limits, ensuring operational safety, and satisfying prioritized downstream water demands for human consumption, irrigation, and ecological flows (Wu et al., 2024; Zarfl et al., 2015). Effectively navigating this trade-off is further complicated by the highly non-linear dynamics inherent in hydraulic and power generation processes and the significant uncertainties associated with hydrological inflows, especially at short timescales (Tabas and Samadi, 2024; Negm et al., 2024). These operational complexities are particularly amplified in large-scale, interconnected river systems involving multiple reservoirs and strategic water transfers, such as the Guayas basin central to this study, demanding increasingly sophisticated optimization approaches (Wu et al., 2024).

The optimization of short-term operations, specifically at the intraday (e.g., hourly) resolution, is becoming paramount (Ramos et al., 2019). This is driven by the increasing need for operational flexibility to integrate intermittent renewable sources, participate in volatile hourly electricity markets, and respond rapidly to hydro-meteorological events. Traditional optimization techniques, including linear programming (LP), non-linear programming (NLP), mixed-integer linear programming (MILP), and various forms of dynamic programming (DP/SDP), have been the mainstay for reservoir operation scheduling (Castelletti et al., 2010). However, applying these methods to complex, interconnected systems like the Baba-Daule-Peripa for high-resolution intraday decision-making reveals significant limitations (Castro-Freibott et al., 2025). These conventional approaches often struggle to adequately capture the system's inherent hydrological variability (including pronounced seasonality and extreme events like El Niño) and complex non-linear system dynamics (e.g., turbine efficiencies, hydraulic losses) (Hidalgo-Proaño, 2017; Ilbay-Yupa et al., 2019; Ghafoor et al., 2024). Furthermore, they frequently encounter the "curse of dimensionality" when dealing with multiple state variables and fine temporal discretization, often leading to prohibitive computational times that hinder their use for real-time or near-real-time control (Tabas and Samadi, 2024; De Mel et al., 2022; Wu et al., 2024). Such approaches may also necessitate substantial model simplifications or exhibit heavy reliance on the accuracy of short-term forecasts, which can be unreliable or unavailable (Ghobadi and Kang, 2023).

To overcome these limitations (Villeneuve et al., 2023), DRL has emerged as a powerful, data-driven paradigm for tackling complex sequential decision-making problems under uncertainty (Sutton and Barto, 2018; Negm et al., 2024). DRL agents learn optimal control policies through direct trial-and-error interactions with an environment (either real or simulated) (Ortega et al., 2024) bypassing the need for explicit, often simplified, system models and demonstrating the capacity to effectively handle non-linearities, stochasticity, and high-dimensional state-action spaces inherent in complex systems (Ortega et al., 2024). Within the DRL landscape, the SAC algorithm (Haarnoja et al., 2018, 2019) stands out as particularly well-suited for continuous control problems like hydropower operation (Tabas and Samadi, 2024; Riemer-Sørensen and Rosenlund, 2020). As an off-policy actor-critic method grounded in the maximum entropy framework, SAC offers notable stability and sample efficiency (Haarnoja et al., 2018; Raffin et al., 2021). Its unique principle of maximizing both expected reward and policy entropy intrinsically encourages robust exploration, reducing the risk of converging prematurely to suboptimal solutions in complex reward landscapes, while its off-policy nature enhances learning efficiency by effectively leveraging past experiences.

This paper proposes and evaluates a DRL framework for optimizing the hourly operation of the interconnected Baba-Daule-Peripa hydropower system in Ecuador's Guayas basin. We specifically employ the SAC algorithm (Haarnoja et al., 2019, 2018), a method well-suited for complex continuous control tasks. While alternative DRL algorithms like Proximal Policy Optimization (PPO) were considered, preliminary investigations indicated challenges in achieving stable convergence for this specific problem, further motivating the selection of SAC. A cornerstone of our methodology is the development and integration of a data-driven, internal NN-based dynamics model within a standard Gymnasium simulation environment (Towers et al., 2024). This learned model is trained on historical and synthetically generated data (using KNN resampling) to approximate the system's complex, non-linear hydraulic and energy state transitions (Hidalgo-Proaño, 2017; Ilbay-Yupa et al., 2019), allowing the SAC agent to learn effective policies without requiring explicit differential equations. The primary objective is to derive operational policies that maximize the combined economic revenue from the system while adhering to operational and safety constraints, implicitly respecting established water use priorities (Asamblea Constituyente del Ecuador, 2008; art. 318). Ultimately, this work validates the effectiveness of this DRL-based approach as a robust and scalable strategy for optimizing complex, interconnected hydropower systems (Tabas and Samadi, 2024; Wu et al., 2024).

2 Methodology

To achieve the optimization objectives outlined previously, we applied a approach to derive optimal hourly operational policies for the Baba hydropower facility. This data-driven methodology allows an artificial decision-maker, termed the RL Agent, to learn effective operational strategies through simulated trial-and-error interactions with the system (Sutton and Barto, 2018). We specifically implemented the SAC Algorithm (Haarnoja et al., 2018, 2019), a state-of-the-art, off-policy DRL technique adept at handling the continuous control variables (e.g., turbine power, gate adjustments) inherent in hydropower operations (Tabas and Samadi, 2024; Riemer-Sørensen and Rosenlund, 2020).

The RL Agent learns within a custom Simulation Environment developed according to the Gymnasium standard (Towers et al., 2024), see Figure 1. This environment simulates the Baba system's response to the agent's actions and provides feedback via a Reward Function. This function is designed to reflect the primary goal of maximizing economic revenue (considering both Baba generation

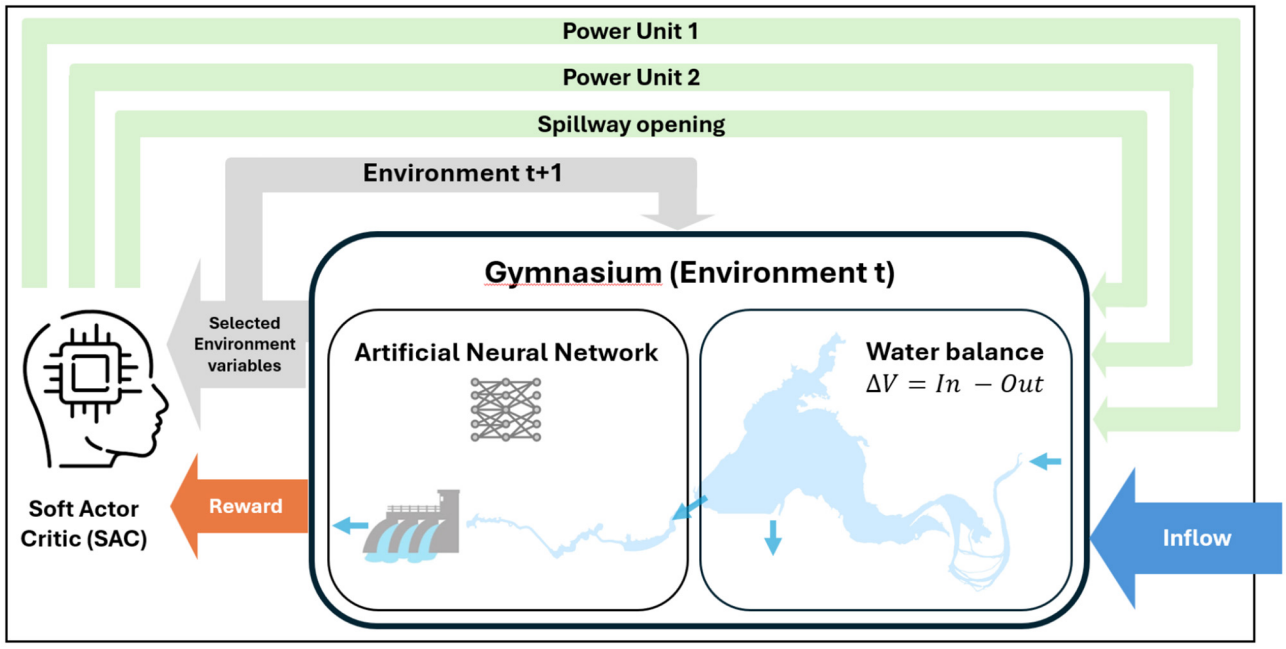

FIGURE 1
Schematic diagram of the implemented Deep Reinforcement Learning framework. The SAC Agent interacts with the Gymnasium-based Simulation Environment, receiving state information and rewards. The internal Dynamics Model predicts system responses. The reward signal guides the agent's policy learning during training.

and estimated downstream MLDW benefits) while incorporating penalties for violating critical operational and safety constraints. Central to the environment's realism is an NN-based internal Dynamics Model. This is a neural network pre-trained on extensive historical data to predict the non-linear hydraulic and energy state transitions resulting from the agent's actions and stochastic river inflows (Ortega et al., 2024). Through extensive interactions within this environment, guided by the reward signal, the SAC Agent develops an operational policy mapping observed system states to optimized actions, aiming to maximize cumulative rewards over the operational horizon.

2.1 The Guayas River basin

This study focuses on the Guayas River basin, the largest watershed draining into the Pacific Ocean from South America, located in western Ecuador and encompassing approximately 45,948 km^2. Originating in the Andean highlands and flowing through extensive coastal plains, the basin's hydrology is fundamental to the region's ecological and socio-economic stability. It provides the primary source of potable water for over eight million inhabitants and supports vast agricultural activities, irrigating more than 300,000 hectares of diverse crops. The basin's hydrological regime is characterized by a distinct unimodal rainfall pattern, with a pronounced wet season typically occurring between February and March, followed by a marked dry season from August to September. This seasonality leads to significant variations in river discharge. Furthermore, the basin is highly susceptible to interannual climate variability, particularly the influence of El Niño Southern Oscillation (ENSO) events, which can drastically alter precipitation patterns, leading to both severe droughts and extreme flood events (Hidalgo-Proaño, 2017; Ilbay-Yupa et al., 2019).

2.1.1 The Baba-Daule Peripa interconnected hydropower system

Within this basin, the study centers on a strategically important interconnected hydropower system managed by the public utility CELEC EP Hidronación (Campo-Carrera et al., 2025; Gelati et al., 2014; CELEC EP. n.d..). The system comprises two main facilities (Figure 2):

- **The Daule-Peripa Reservoir and Marcel Laniado de Wind (MLDW) Hydropower Plant:** Located on the Daule River, this is the region's primary storage facility (approx. 5,200 hm^3 capacity) (web Corporacion Electrica del Ecuador). It serves multiple functions: hydropower generation (3 x 71 MW turbines, avg. 1,033 GWh/year), potable water supply, irrigation, flood control, and downstream salinity management.
- **The Baba Reservoir and Hydropower Plant:** Situated upstream on a tributary, operational since 2013 with a smaller reservoir (approx. 70 hm^3) (web Corporacion Electrica del Ecuador). Its plant (2 x 21 MW turbines, avg. 154 GWh/year) strategically transfers water from the wetter Quevedo basin to Daule-Peripa, augmenting MLDW inflows, especially during dry periods. Key control components include, see Figure 3:

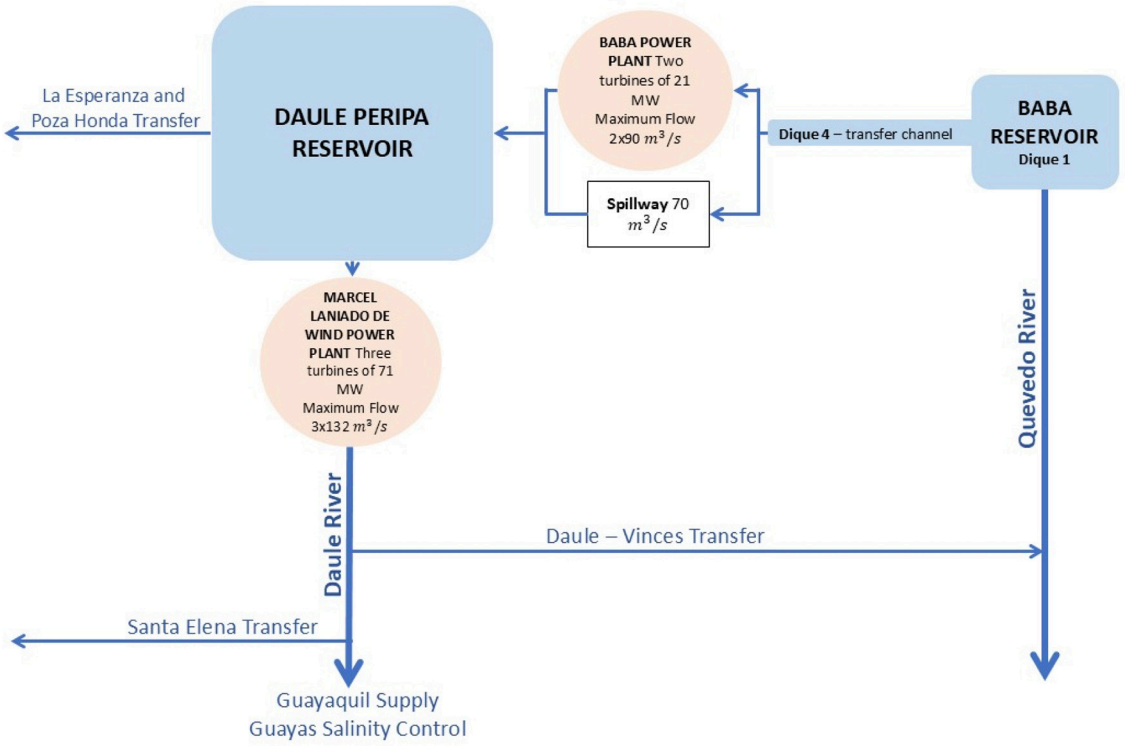

FIGURE 2
Schematic overview of the interconnected Baba and MLDW hydropower facilities within the Guayas River basin, Ecuador, indicating key components and flow paths.

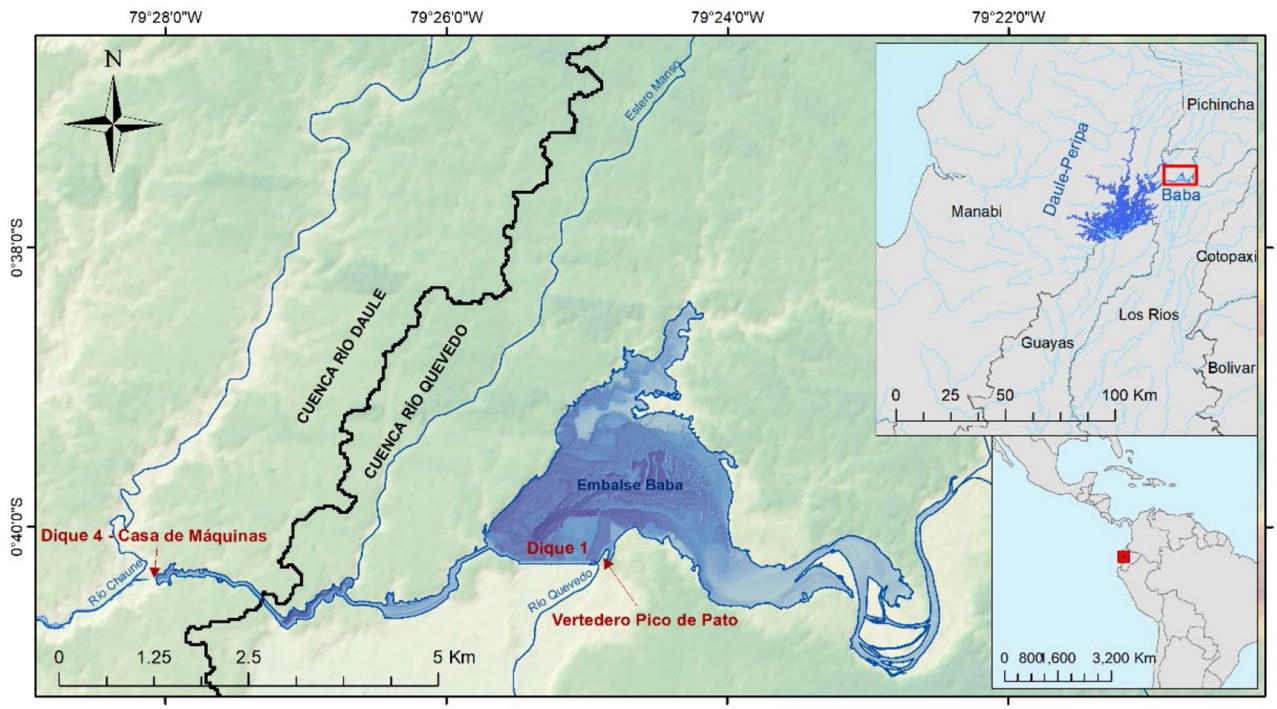

FIGURE 3
Detail of baba reservoir, dam 1, dam 4 (Powerhouse), conveyance channels, and Pico de Pato spillway.

- An uncontrolled **free spillway** *("Pico de Pato")* for dam safety overflows.
- A **gated spillway** ("Extravasor") allowing regulated water transfer (Nominal capacity, 70 m^3/s) toward Daule-Peripa, independent of turbine operation.
- The **powerhouse turbines**, whose discharge also contributes to the transfer while generating electricity.

The interconnection enables flexible water management but introduces significant operational complexity due to the dependencies between the facilities.

2.1.2 Operational challenges and optimization objectives

Operating the Baba hydropower plant optimally presents significant challenges stemming from:

- **Hydrological Uncertainty:** High seasonality and unpredictable interannual variations in Quevedo River inflows, exacerbated by ENSO events.
- **Complex Hydraulics:** Non-linear relationships govern turbine efficiency, head losses in the transfer system, and spillway discharge dynamics.
- **Interdependent System:** Baba's operation directly impacts water availability and generation potential at the downstream MLDW plant.
- **Multiple, Potentially Conflicting Objectives:** The need to maximize economic revenue from energy sales (subject to different energy tariffs at Baba and MLDW) must be balanced against maintaining dam safety (reservoir level limits), ensuring structural integrity and efficiency of turbines (operational range constraints, minimizing start/stops), and adhering to downstream flow requirements implicitly managed through the large Daule-Peripa storage.

Therefore, the primary objective of this study is to develop and evaluate an optimal hourly operational policy specifically for the **Baba facility** (controlling its two turbines and the *Extravasor* spillway) using a Deep Reinforcement Learning approach. The goal is to maximize the combined economic benefit derived from generation at Baba and the estimated downstream generation benefit at MLDW resulting from the transferred water, while strictly respecting all operational and safety constraints.

2.1.3 Data acquisition and preparation

Developing and validating the proposed methodology relied on comprehensive datasets:

- **Historical Operational and Hydrological Data:** Hourly operational records from January 1, 2015, to December 31, 2024, were obtained from CELEC EP Hidronación. These included reservoir water levels (at various points like Dique 1, Dique 4, see Figure 3), turbine discharges, generated power for each unit, and gate openings for the Extravasor. Crucially, as direct inflow measurements to Baba were unavailable, hourly inflows for this period were derived using a water balance calculation based on the observed changes in storage and measured outflows. This historical dataset was primarily used for: (a) training (70%) and validating (30%) the internal neural network simulation model (Section 2.4.2), and (b) establishing a baseline performance benchmark for comparison against the DRL agent's policies.
- **Extended Synthetic Flow Series for Agent Training:** Deep Reinforcement Learning agents typically require extensive interaction with the environment, often spanning longer periods than available historical records, to learn robust policies across a wide range of hydrological conditions. To address this, a long-term synthetic hourly inflow series for the Baba reservoir was generated, effectively covering the period 1950–2015 for training purposes. This series was constructed based on historical mean monthly data using the k-Nearest Neighbors (KNN) resampling technique (Lall and Sharma, 1996; Yates et al., 2003), which was selected after comparative analysis demonstrated its superior ability over simpler methods (like the Method of Fragments) to preserve key statistical properties (mean, variance, probability distribution) and temporal correlation structures of the historical flows observed post-2015 (see Supplementary Figures 1–3 for time-series comparisons, QQ-plots and Cumulative Distribution Functions). This high-fidelity synthetic series provided the necessary long-term hydrological variability for effective agent training.

2.2 Reinforcement learning approach

Reinforcement learning can be explained mathematically as Markov Decision Processes (MDPs, Bellman, 1957). An MDP is an extension of Markov chains that involves decision-making and actions taken by an agent to maximize cumulative rewards over time. Like Markov chains, MDPs are based on a fixed set of states, where each represents the current environment situation. With MDPs, the agent can take actions to influence state transitions and achieve specific goals. The agent's actions determine the probability of transitioning to different states. MDPs contain rewards associated with state transitions and actions. The agent's goal is to learn a strategy (policy) that maximizes the cumulative rewards achieved over time.

Applying this framework to our hydropower operation problem, the core components are defined as follows:

- **Agent:** The decision-maker (in our case, the SAC algorithm) that learns the operational policy.
- **Environment:** A simulation representing the system being controlled (the Baba reservoir, plant, and associated hydraulics).
- **State (s_t):** A vector representing the environment's condition at hourly time step t (e.g., reservoir levels, inflow, previous operational actions).
- **Action (a_t):** Decisions made by the agent at time step t (e.g., setting turbine power output, adjusting spillway gate opening).

- **Reward (r):** A scalar feedback signal from the environment indicating the immediate desirability of the action taken *a_t* in state *s_t*. in the current state *s_t*.
- **Policy (π):** The strategy learned by the agent, mapping states to actions, aiming to maximize the cumulative reward over time.

The agent learns through a cycle of observing state *s_t*, selecting action *a_t*, receiving reward *r_t*, observing the next state *s_{t+1}*, and updating its policy π.

2.3 Soft Actor-Critic algorithm

We selected the SAC algorithm (Haarnoja et al., 2019, 2018), a state-of-the-art DRL algorithm particularly well-suited for continuous control problems like hydropower operation. Its key advantages in this context include:

- **Continuous Action Space:** SAC naturally handles continuous actions (e.g., precise power levels in MW or gate openings in meters), allowing for finer operational control compared to algorithms restricted to discrete actions.
- **Entropy Maximization:** SAC optimizes for both expected return and policy entropy. This encourages structured exploration, mitigating the risk of premature convergence to suboptimal policies in the complex, potentially multi-modal reward landscape of hydropower operation.
- **Off-Policy Learning & Sample Efficiency:** SAC is an off-policy algorithm, meaning it can learn efficiently from past experiences stored in a replay buffer, even if those experiences were generated by a previous version of the policy. This improves data efficiency, crucial when environment interactions (simulations) can be computationally intensive.

SAC typically employs neural networks to approximate the policy (the "actor") and value functions (the "critics") that estimate the expected return of state-action pairs.

2.4 Simulation environment design

2.4.1 Environment design and interface

A custom simulation environment was developed following the Gymnasium standard (Brockman et al., 2016), providing a consistent interface for agent-environment interaction.

- **State Representation (s_t)**: The state vector provided to the agent at each hourly step comprised: current reservoir water levels (e.g., Dique 1, Dique 4), estimated inflow for the current hour, power generated by each turbine in the previous hour (MW), and the Extravasor gate opening in the previous hour (m).
- **Action Space (a_t):** The agent outputs a 3-dimensional continuous action vector: target power for Unit 1 (MW), target power for Unit 2 (MW), and target Extravasor opening (m). These actions are typically normalized [e.g., to (−1, 1) or (0, 1)] by the SAC algorithm and then scaled by the environment to the physical operational limits before application.
- **Environment Step Logic**: At each hour t, the environment: (1) receives action a_t from the agent, (2) scales the action to physical units, (3) uses the internal NN Dynamics Model (Section 1.4.2) to predict the resulting physical state transitions (flows, levels, power), (4) calculates the change in storage via water balance, (5) computes the reward r_t (Section 1.5), (6) determines the next state s {t+1}, and (7) returns (s_{t+1}, r_t, termination/truncation flags) to the agent.

To simulate the environment, a combination of classical water balance equations and discharge equations for outflow structures has been implemented, integrated with a multilayer neural network model.

2.4.2 NN-based dynamics model

To accurately simulate the complex, non-linear response of the Baba hydropower system within the environment, a dedicated internal dynamics model was developed using a pre-trained feedforward neural network (Multilayer Perceptron, MLP).

- **Rationale:** An NN approach captures the complex, non-linear relationships (e.g., turbine efficiency curves, head losses, spillway hydraulics) more effectively than simplified analytical models, learning these directly from historical data.
- **Architecture and Prediction:** The MLP takes relevant components of the current state *s_t* and the applied physical action *a_t* as input. It predicts key physical outcomes for the hour: resulting water levels (Dique 4, discharge point), total turbine flow, Extravasor flow, individual turbine flows, and total energy generated (MWh).
- **Training and Validation:** This NN model was trained and validated *offline* using the 2015–2024 historical dataset *before* integration into the RL environment. Supervised learning techniques were used to minimize prediction errors for the key outputs. The validation process confirmed the NN's ability to accurately reproduce observed system dynamics across various operating conditions, justifying its use as the core simulation engine. (Detailed validation metrics are presented in Section 3.1).

2.4.3 Water balance in the reservoir

In order to estimate the inflows into Dam 1, the water balance of the reservoir is defined as follows.:

$$Balance = Inflow - (Q_{ecological}(Dam\ 1\ Level) + Q_{Pico\ de\ Pato}(Dam\ 1\ Level) + Q_{turbines}(NN\ Model) + Q_{Extravasor}(NN\ Model))$$

where:

$Q_{ecological}$ and $Q_{Pico\ de\ Pato}$ are functions dependent on the Dam 1 level.

$Q_{turbines}$ and $Q_{Extravasor}$ are values estimated by the neural network model.

Infiltration and evaporation were not explicitly included in the water balance, as they are accounted for within the inflow estimates;

evaporation, in particular, was omitted due to its negligible impact, amounting to only 0.014% of the inflow.

2.5 Reward function design

The reward function translates the multi-objective operational goals into a single scalar value guiding the agent's learning. It was designed to balance economic revenue with operational constraints:

- **Economic Objective:** A positive reward component calculated as the sum of: (a) revenue from energy generated at Baba (MWh * Baba tariff), and (b) estimated revenue from potential energy generated at the downstream MLDW plant due to water transferred from Baba. This incentivizes both direct generation and beneficial water transfer.

$$Total\ Revenue(t) = Revenue\_Baba(t) + Estimated\_Revenue\_MLDW(t)$$

where:

$$Revenue\_Baba(t) = Energy\_Baba(t)[MWh] \times Tariff\_Baba[\frac{\$}{MWh}]$$

$$Estimated\_Revenue\_MLDW(t) = Transferred\_Vol(t)[m^3] \times Yield\_Factor[\frac{MWh}{m^3}] \times Tariff\_MLDW[\frac{\$}{MWh}]$$

- **Operational Constraints (Penalties):** Negative rewards (penalties) were applied to discourage violations of operational limits and unsafe conditions. These included penalties for:
  - Exceeding maximum reservoir levels (magnitude-scaled penalty).
  - Operating turbines outside efficient/safe power ranges (penalty, Figure 4).
  - Excessive turbine start/stop frequency (penalty per event).
  - Significant flow over the uncontrolled "Pico de Pato" spillway (penalty, potentially magnitude-scaled, indicating high levels or inefficient water use).

2.6 SAC agent training and implementation

- **Training Protocol:** The SAC agent was trained by interacting with the custom Gymnasium environment over millions of hourly time steps. The environment was driven by the long-term **synthetic** hourly inflow series (1950–2015) generated using the KNN method (Section 2.1.3). This ensured the agent experienced a wide spectrum of hydrological conditions, promoting the learning of a robust policy adaptable to different flow regimes.
- **Implementation Details:** The implementation utilized Python, employing the Stable-Baselines3 library (Raffin et al., 2021) for the SAC algorithm and PyTorch (Paszke et al., 2019) for the internal NN model. Key SAC hyperparameters (e.g., learning rates, batch size, network architecture for actor and critic, discount factor γ, entropy coefficient α) were configured based on preliminary testing and established practices (specific values detailed in Table 1 of the original document). The agent's policy and associated models were saved periodically during training.

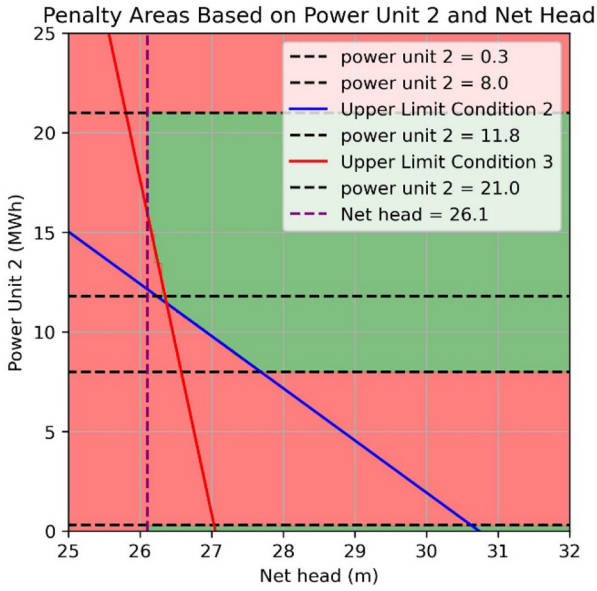

FIGURE 4
Penalty areas of the units, based on historical operation.

3 Results

This section presents the results obtained from validating the simulation components and evaluating the performance of the SAC agent trained to optimize the hourly operation of the Baba hydropower facility.

3.1 Validation of the NN-based system dynamics model

The fidelity of the internal neural network (NN) model, responsible for simulating the hydraulic and electrical dynamics within the RL environment (Section 2.4.2), was validated against historical operational data (January 2015—October 2022). The comparison between the NN model's predictions and the actual recorded measurements demonstrated high accuracy:

TABLE 1 Training parameters and architecture of the SAC models.

Model	Training steps	Network architecture	Learning rate	Batch size	Training frequency
SAC M01	6,100,000	(256, 512, 1024, 1024, 512, 256)	2.00E-04	4,096	10
SAC M02	5,400,000	(256, 512, 1024, 1024, 1024, 512, 256)	3.00E-04	8,192	10
SAC M03	6,750,000	(256, 512, 1024, 1024, 1024, 256)	2.00E-04	8,192	25
SAC M04	11,090,000	(256, 512, 1024, 1024, 1024, 256)	2.00E-04	8,192	10
SAC M05	17,000,000	(256, 512, 1024, 1024, 1024, 512, 256)	3.00E-04	8,192	10
SAC M06	5,500,000	(256, 512, 1024, 2048, 2048, 1024, 256)	3.00E-04	4,096	25
SAC M07	3,700,000	(256, 512, 1024, 1024, 512, 256)	2.00E-04	4,096	10

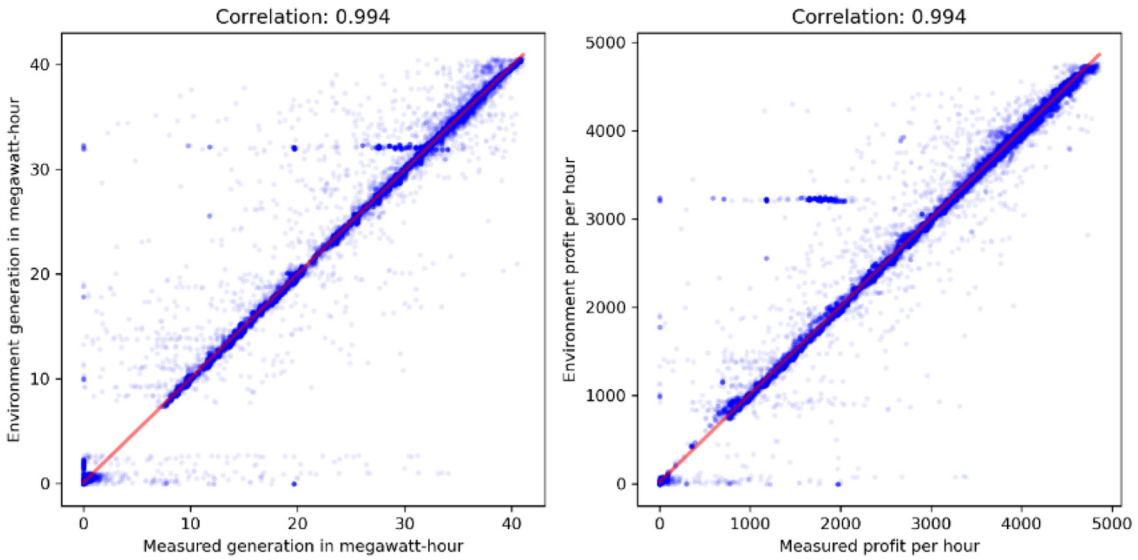

FIGURE 5
Comparison of results with historical actions: measured results vs. outcomes generated by the environment.

- **Energy Production:** A Pearson correlation coefficient of 0.994 was achieved between the simulated and historically recorded hourly energy generation at the Baba plant. This metric was chosen to specifically quantify the strength of the linear relationship, which is visually confirmed in scatter plots (Figures 5, 6) and is the primary assumption for this validation. The test is also considered robust to normality deviations in large datasets. The deviation in the total accumulated energy over the entire validation period was only 0.12%.
- **Economic Gain:** When considering the total economic gain (including estimated downstream benefits from transfer), the correlation between simulated and historical values was also 0.994, with a cumulative deviation of 0.89% over the period.

Scatter plots comparing simulated vs. measured values (Figure 5) visually confirm the strong linear relationship and minimal dispersion.

Further validation using specific hydraulic variables (discharge levels, dam levels, turbine flows—Figure 6) showed excellent agreement, with coefficients of determination (R^2) generally exceeding 0.99, and above 0.92 even for the most sensitive variables. Minor deviations observed are likely attributable to potential inaccuracies in manual historical data logging rather than systemic model flaws. These results confirm the NN model's capability to reliably reproduce the complex system dynamics, providing a high-fidelity simulation environment for RL agent training.

3.2 Validation of synthetic inflow generation

To ensure robust agent training across diverse hydrological conditions, the k-Nearest Neighbors (KNN) method was selected for generating the long-term synthetic hourly inflow series (1950–2015), as described in Section 2.1.3. The suitability of KNN was confirmed by comparing its statistical properties against both historical data (2015–2024) and an alternative generation method (Method of Fragments):

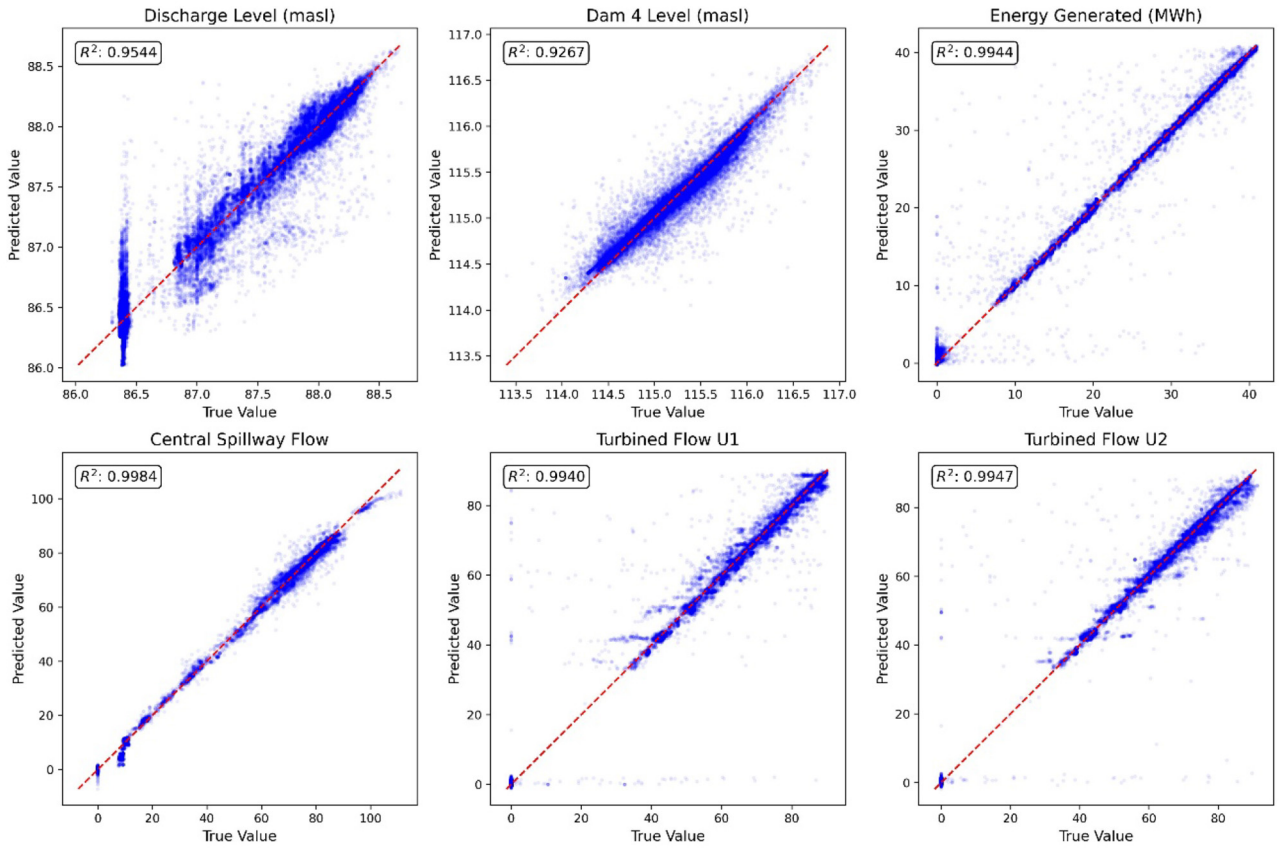

FIGURE 6
Internal model for simulating the behavior of diversion channels and dikes along with the Hydroelectric Power Plant. Comparison of environment-predicted results with historical measurements.

- **Statistical Fidelity:** KNN demonstrated superior performance in preserving key statistics. Compared to the Method of Fragments, KNN yielded significantly lower Root Mean Square Error (RMSE: 63.6 vs. 111.5 m^3/s) and Mean Absolute Percentage Error (MAPE: 29.4% vs. 76.2%) relative to the historical monthly means.
- **Distributional and Temporal Similarity:** The KNN method demonstrated superior performance in replicating the historical series compared to the Fragments method. For temporal similarity, KNN achieved a higher Pearson correlation (0.906 vs. 0.709) and a significantly lower Dynamic Time Warping (DTW) distance (84,198 vs. 253,055), indicating a better reproduction of temporal patterns. Furthermore, for distributional similarity, the Kolmogorov-Smirnov (KS) test confirmed that the KNN-generated distribution was closer to the historical data, yielding a smaller KS statistic (0.106 vs. 0.112; $p < 0.001$ for both).
- Time-series comparisons (Supplementary Figure 1), QQ-plots (Supplementary Figure 2) and empirical cumulative distribution functions (Supplementary Figure 3) visually corroborate these findings. The KNN method provided a statistically robust and temporally coherent long-term inflow series, deemed essential for effective DRL training.

3.3 DRL agent training and learned policy characteristics

Multiple SAC agents were trained using different NN architectures and hyperparameter configurations (detailed in Table 1) interacting with the validated simulation environment driven by the synthetic KNN inflow series.

- **Convergence:** Training runs typically converged toward stable policies, exhibiting characteristic learning curves where the cumulative reward per episode increased and stabilized over millions of simulation steps (Learning curves can be shown in Supplementary material if needed).
- **Policy Variations:** Despite converging to high-performance policies, different training configurations resulted in distinct operational strategies, particularly regarding reservoir level management and turbine usage (Table 2, Figure 7):
  - **Reservoir Level Management:** While most models maintained average reservoir levels close to the historical mean (approx. 115.68 m a.s.l.), the temporal patterns varied. Some models (e.g., SAC M05) tended to maintain slightly higher levels, potentially maximizing head for dry periods but increasing spill risk during floods.

TABLE 2 Results of the different SAC optimizations during the modeled historical validation period (2015–2024).

Model	Total gain 2015–2024 (USD)	Gain diff. vs. historical (%)	Transfer diff. vs. historical (%)	Spillway "Extravasor" diff. vs. historical (%)	Spillway "Pico de Pato" diff. vs. historical (%)	% Generation Baba Diff. vs. historical	% Generation MLDW diff. vs. historical	Average reservoir level (m.a.s.l.)	Daily unit startups
HISTORICAL	153,715,278	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	115.68	0.3599
SAC M01	166,134,973	8.08%	5.74%	13.95%	−26.27%	6.81%	9.66%	115.83	0.6884
SAC M02	165,278,110	7.52%	6.48%	27.26%	−28.22%	5.23%	10.37%	115.86	0.7614
SAC M03	168,214,685	9.43%	4.46%	−8.76%	−22.23%	10.21%	8.46%	115.42	0.4376
SAC M04	156,983,019	2.13%	1.06%	19.60%	−13.87%	−0.18%	4.99%	115.9	0.7608
SAC M05	160,882,230	4.66%	6.70%	53.55%	−29.06%	−0.15%	10.65%	115.93	1.0901
SAC M06	167,048,977	8.67%	11.10%	62.32%	−39.22%	3.76%	14.78%	115.46	0.8692
SAC M07	166,771,305	8.49%	8.32%	37.08%	−33.17%	5.48%	12.24%	115.83	0.7054

Others (e.g., SAC M06) operated at slightly lower levels, reducing spill probability but potentially foregoing some generation opportunities.

- **Turbine Operation**: A trade-off was observed between maximizing immediate economic gain and minimizing turbine wear-and-tear. Some configurations (e.g., SAC M05 achieved high gains but exhibited more frequent daily unit startups (Table 2). Others (e.g., SAC M03) achieved comparable or higher gains with significantly fewer startups, indicating a more stable operational regime.

- **Adaptability:** The emergence of multiple high-performing, yet distinct, policies highlights the flexibility of the DRL approach and its ability to find different balances between competing objectives based on subtle differences in training setup or inherent system trade-offs.

3.4 Performance evaluation against historical benchmark

The performance of the trained SAC policies was evaluated by simulating their operation over the historical period (January 2015—December 2024) using the actual derived historical inflows and comparing the outcomes against the actual historical operation record.

- **Economic Gain:** All evaluated SAC models demonstrated a significant improvement in total accumulated economic gain (Baba generation revenue + estimated MLDW transfer benefit) compared to the historical baseline (153.7 million USD). Annualized improvements ranged from approximately +2.13% to a maximum of +9.43% (Table 2). The top-performing models were SAC M03 (+9.43%), SAC M06 (+8.67%), and SAC M07 (+8.49%).
- **Constraint Adherence and Spill Management:** The reward function penalties effectively guided the agents to respect operational limits:
  - **Uncontrolled Spillway ("Pico de Pato"):** The specific penalties applied for high discharges via the Pico de Pato spillway resulted in all SAC models significantly reducing the peak discharge compared to the historical maximum (1,136.37 m^3/s). Reductions reached nearly 9% in the best case (SAC M05: 1,035.30 m^3/s) (Table 3). Analysis of the peak historical spill event (April 2022) showed that the trained models successfully maintained discharges below the observed historical peak, demonstrating effective flood mitigation behavior learned via the reward signal (Figure 8).
  - **Reservoir Levels**: As noted previously (Figure 7), average levels were maintained within a safe and operationally reasonable range comparable to historical practice, although with differing dynamic behaviors.

- **Water Balance Conservation:** Analysis of the difference between total inflows and outflows for each model over the

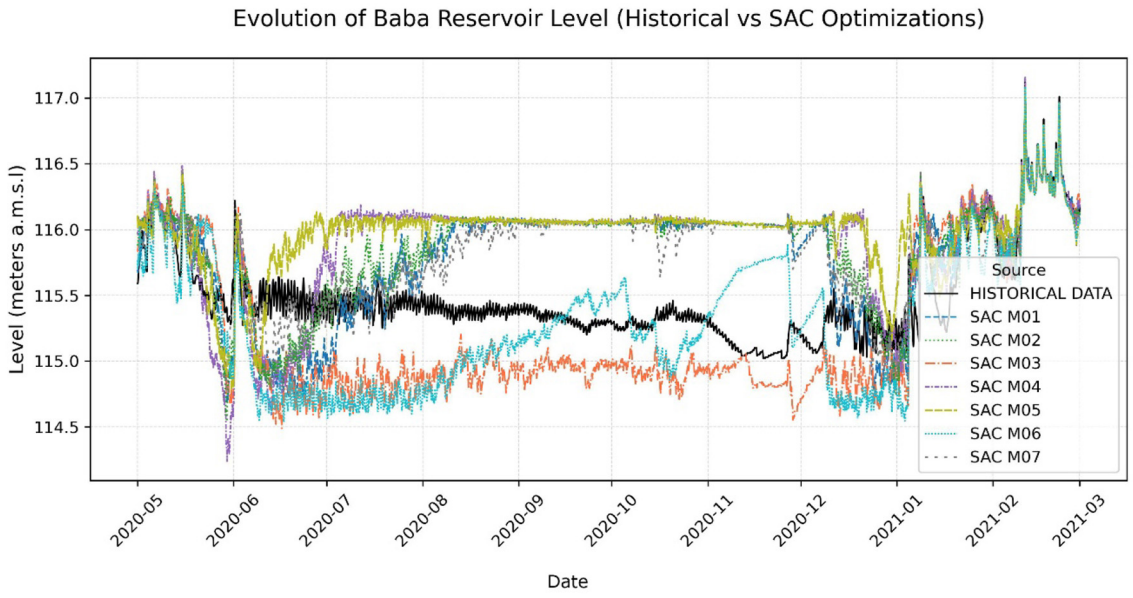

FIGURE 7
Baba reservoir level evolution: comparison of SAC optimization models and historical data.

TABLE 3 Percentage difference between inflows and outflows during the modeled historical validation period (2015–2024).

Model	% In-out diff. vs. historical	Maximun discharge at Pico de Pato 2015–2024 (m3/s)
HISTORICAL	0.00000%	1,136.37
SAC M01	0.00558%	1,044.62
SAC M02	0.00154%	1,040.05
SAC M03	−0.00950%	1,043.51
SAC M04	0.00391%	1,050.65
SAC M05	−0.00609%	1,035.3
SAC M06	−0.01530%	1,054.9
SAC M07	0.00290%	1,055.33

Maximum flows released to the river through the Pico de Pato spillway for different SAC optimizations and historical data.

validation period showed negligible deviations (ranging from +0.00558% to −0.01530% relative to historical flows, Table 3). This confirms that the simulation environment, including the NN model, rigorously conserves mass, ensuring that economic gains are not based on artificial water creation or loss.

4 Discussion

The results of this study demonstrate the significant potential of DRL for optimizing the complex, hourly operation of the Baba hydropower facility. The development of operational policies that yield up to a 9.43% increase in economic gain over the historical baseline is a core achievement, highlighting the capacity of DRL to identify and exploit complex efficiencies often missed by traditional methods or manual operation. The magnitude of this improvement is considerable and aligns with performance gains reported in other recent studies applying DRL to hydropower optimization (e.g., Wu et al., 2024; Tabas and Samadi, 2024). While direct numerical comparisons are challenging due to differences in reservoir characteristics and objectives, our results confirm that the proposed framework is a state-of-the-art approach, particularly in its ability to concurrently optimize economic revenue while enhancing operational safety, evidenced by the up to 9% reduction in peak uncontrolled spillway discharges.

The fidelity of these results is strongly supported by the rigorous validation of the underlying simulation framework. The integration of a pre-trained NN-based dynamics model within the Gymnasium environment proved highly effective, accurately reproducing the system's non-linear hydraulics and energy generation ($R^2 > 0.99$ for key variables, Figure 6) while strictly conserving mass balance (Table 3). Furthermore, the use of a statistically validated, long-term synthetic inflow series generated via the KNN method ensured that the agent was trained across a diverse and representative range of hydrological conditions, fostering the robustness of the learned policies.

Beyond the simulation framework, the choice of the SAC algorithm was instrumental. Extensive experimentation was conducted with the PPO algorithm, exploring various architectures and hyperparameter configurations, yet it failed to achieve satisfactory performance on this hydroelectric dispatch problem. In contrast, SAC's unique principle of maximizing both expected reward and policy entropy intrinsically encourages robust exploration. This, combined with its off-policy nature, allowed it

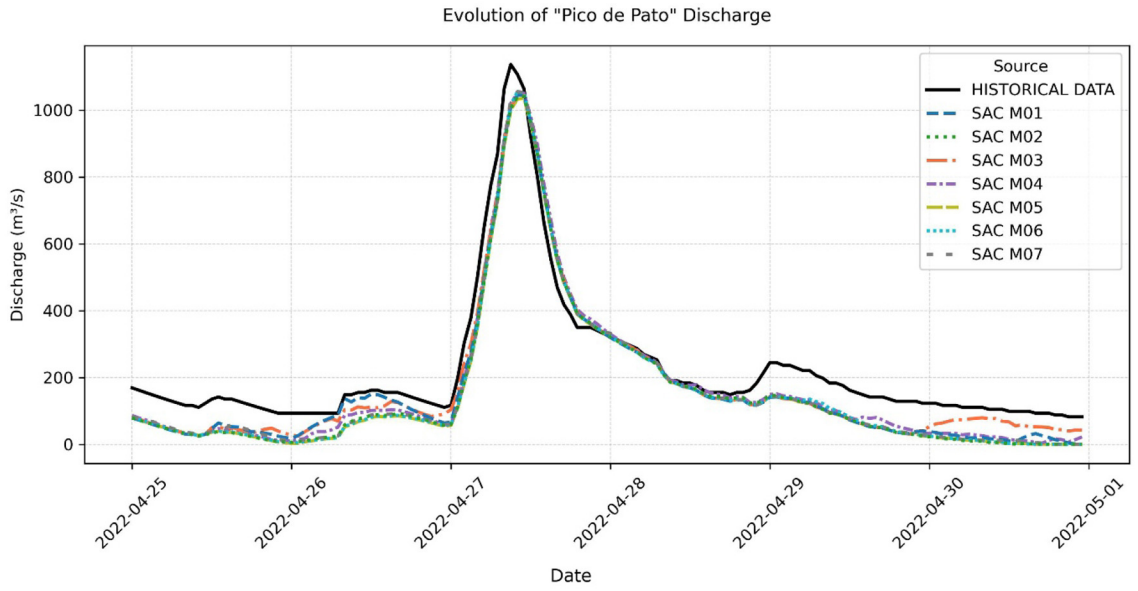

FIGURE 8
Evolution of the flow discharged through Pico de Pato between April 26 and 30, 2022, for different SAC optimizations and historical data.

to excel in a scenario requiring continuous, highly interdependent actions, confirming SAC as a more effective and stable option for the operation of the Baba power plant. This success led to the discovery of not a single optimal policy, but a suite of diverse, high-performing strategies. This presents a key advantage in the form of operational flexibility, offering human operators a choice of policies adaptable to varying real-time conditions or strategic priorities. However, it also reveals inherent trade-offs; for example, some policies achieved high gains at the cost of more frequent turbine startups, which could increase long-term maintenance costs, whereas others found a more stable operational regime with comparable gains.

Despite the promising results, certain limitations inherent to DRL applications warrant discussion. The performance is ultimately bound by the accuracy of the simulation model; while validated, discrepancies between the NN model and true system dynamics could arise under novel conditions. The representativeness of the training data also influences policy generalization. Moreover, further refinement of the simulation framework and rewards remains an open direction, since hydropower plants are inherently complex systems influenced by diverse and evolving factors that cannot be fully captured at once, but can be progressively incorporated. Furthermore, the design of the reward function involves subjective weighting of competing objectives, and alternative weighting schemes could lead to different optimal policies. Finally, the computational cost and expertise required for DRL training and hyperparameter tuning remain considerations for practical deployment. These limitations suggest clear directions for future work, including the exploration of online learning to close the model-reality gap, the use of multi-objective DRL to map Pareto-optimal policies, and the integration of climate forecasts to enhance policy robustness against future uncertainties.

5 Conclusions

This study successfully demonstrated that a Deep Reinforcement Learning framework using the Soft Actor-Critic algorithm can derive high-performing operational policies for a complex, real-world interconnected hydropower system. The combination of a high-fidelity, NN-based simulation model and an advanced DRL algorithm capable of efficient exploration in continuous control spaces proved key to this success.

The learned policies significantly outperformed historical operations, achieving up to a 9.43% increase in total economic gain over a decade-long validation period. Crucially, this economic optimization was achieved while respecting operational and safety constraints, most notably by reducing peak uncontrolled spillway discharges by up to 9%. The research also highlighted the discovery of a diverse set of viable policies, revealing practical trade-offs between maximizing immediate revenue and ensuring long-term operational stability.

Ultimately, this work provides a strong foundation and a methodological blueprint for leveraging DRL to develop more resilient, adaptive, and economically efficient hydropower management systems. It paves the way for a new generation of data-driven tools that can help operators navigate the increasing complexities of water and energy resource management.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data is property of Corporacion Electrica del Ecuador. It has allowed us to use them for this work under

conditions of confidentiality. Requests to access these datasets should be directed to informacion@celec.gob.ec.

Author contributions

JC: Writing – original draft, Visualization, Software, Validation, Methodology, Conceptualization, Formal analysis, Data curation. AU: Supervision, Writing – review & editing, Conceptualization.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work has been partially supported by grant "Matemáticas explicables para soluciones interdisciplinares de ciencia de datos, ref. PID2021-122640OB-I00, funded by the Spanish Ministry of Science and Innovation.

Acknowledgments

The authors express their gratitude to Corporación Eléctrica del Ecuador CELEC EP for their support. This work was partly developed in the context of the "Water Resilience Portfolio" at the Joint Research Centre (JRC) of the European Commission.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The author(s) gratefully acknowledge the use of generative AI tools in the preparation of this manuscript. In particular, Gemini 2.5 was used to improve language and readability. Subsequently, the author(s) reviewed and edited the content as necessary and take full responsibility for the final publication.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2025.1649284/full#supplementary-material

References

Asamblea Constituyente del Ecuador (2008). *Constitución de la República del Ecuador*. Quito: Tribunal Constitucional del Ecuador. Spanish. Registro oficial Nro 449, 79–93.

Bautista, E. L. V., Ângulo Guerrero, R. J., Farfán Bone, J. M., Verá Lozano, C. J., Arboleda Cheres, I. A., Orobio Arboleda, T. J., et al. (2022). Una revisión del suministro de energía renovable y las tecnologías de eficiencia energética. *Polo del Conocimiento* 7, 83. Spanish. Available online at: https://polodelconocimiento.com/ojs/index.php/es/article/view/3934/9143

Bellman, R. (1957). A Markovian decision process. *J. Math. Mech.* 6, 679–684. doi: 10.1512/iumj.1957.6.56038

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., et al. (2016). OpenAI Gym. *arXiv* [Preprint]. doi: 10.48550/arXiv.1606.01540

Campo-Carrera, J. M., Cedeño-Villarroel, M. A., Boada-Herrera, M., and Udias, A. (2025). Sistema de ayuda a la decisión para la gestión hidrológica del río Guayas. *Tecnol. Cienc. Agua* 16, 237–294. Spanish. doi: 10.24850/j-tyca-2025-01-06

Castelletti, A., Galelli, S., Restelli, M., and Soncini-Sessa, R. (2010). Tree-based reinforcement learning for optimal water reservoir operation. *Water Resour. Res.* 46:2009WR008898. doi: 10.1029/2009WR008898

Castro-Freibott, R., García-Sánchez, Á., Espiga-Fernández, F., and González-Santander de la Cruz, G. (2025). Deep Reinforcement Learning for intraday multireservoir hydropower management. *Mathematics* 13:151. doi: 10.3390/math13010151

De Mel, I., Klymenko, O. V., and Short, M. (2022). Balancing accuracy and complexity in optimisation models of distributed energy systems and microgrids with optimal power flow: a review. *Sustain. Energy Technol. Assess.* 52:102066. doi: 10.1016/j.seta.2022.102066

Gelati, E., Madsen, H., and Rosbjerg, D. (2014). Reservoir operation using El Niño forecasts—case study of Daule Peripa and Baba, Ecuador. *Hydrol. Sci. J.* 59, 1559–1581. doi: 10.1080/02626667.2013.831978

Ghafoor, J., Forio, M. A. E., Nolivos, I., Arias-Hidalgo, M., and Goethals, P. L. M. (2024). Model-based analysis of the impact of climate change on hydrology in the Guayas River basin (Ecuador). *J. Water Clim. Change* 15, 5021–5040. doi: 10.2166/wcc.2024.064

Ghobadi, F., and Kang, D. (2023). Application of machine learning in water resources management: a systematic literature review. *Water* 15:620. doi: 10.3390/w15040620

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). "Soft Actor-Critic: off-policy maximum entropy Deep Reinforcement Learning with a stochastic actor," in *International Conference on Machine Learning* (Stockholm, Sweden: PMLR), 1861–1870.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., et al. (2019). Soft Actor-Critic algorithms and applications. *arXiv* [Preprint]. doi: 10.48550/arXiv.1812.05905

Hidalgo-Proaño, M. (2017). Variabilidad climática interanual sobre el Ecuador asociada a ENOS. *Rev. Cienciamérica* 6, 42–47. Spanish. Available online at: https://cienciamerica.edu.ec/index.php/uti/article/view/82

Ilbay-Yupa, M., Zubieta Barragán, R., and Lavado-Casimiro, W. (2019). Regionalización de la precipitación, su agresividad y concentración en la cuenca del río Guayas, Ecuador. *LA GRANJA. Rev. Cienc. Vida* 30, 57–76. Spanish. doi: 10.17163/lgr.n30.2019.06

Lall, U., and Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.* 32, 679–693. doi: 10.1029/95WR02966

Negm, A., Ma, X., and Aggidis, G. (2024). Deep Reinforcement Learning challenges and opportunities for urban water systems. *Water Res.* 253:121145. doi: 10.1016/j.watres.2024.121145

Ortega, R., Carciumaru, D., and Cazares-Moreno, A. D. (2024). Reinforcement learning for watershed and aquifer management: a nationwide view in the country of Mexico with emphasis in Baja California Sur. *Front. Water* 6:1384595. doi: 10.3389/frwa.2024.1384595

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: reliable reinforcement learning implementations. *J. Mach. Learn. Res.* 22, 1–8. Available online at: https://jmlr.org/papers/volume22/20-1364/20-1364.pdf

Ramos, H. M., McNabola, A., López-Jiménez, P. A., and Pérez-Sánchez, M. (2019). Smart water management towards future water sustainable networks. *Water* 12:58. doi: 10.3390/w12010058

Riemer-Sørensen, S., and Rosenlund, G. H. (2020). "Deep Reinforcement Learning for long term hydropower production scheduling," in *2020 International Conference on Smart Energy Systems and Technologies (SEST)* (Istanbul, Turkey: IEEE), 1–6.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT press.

Tabas, S. S., and Samadi, V. (2024). Fill-and-Spill: Deep Reinforcement Learning policy gradient methods for reservoir operation decision and control. *J. Water Resour. Plan. Manag.* 150:04024022. doi: 10.1061/JWRMD5.WRENG-6089

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., et al. (2024). Gymnasium: A Standard Interface for Reinforcement Learning Environments. *arXiv* [Preprint]. doi: 10.48550/arXiv.2407.17032

Villeneuve, Y., Séguin, S., and Chehri, A. (2023). Ai-based scheduling models, optimization, and prediction for hydropower generation: opportunities, issues, and future directions. *Energies* 16:3335. doi: 10.3390/en16083335

Wu, R., Wang, R., Hao, J., Wu, Q., and Wang, P. (2024). Multiobjective multihydropower reservoir operation optimization with transformer-based Deep Reinforcement Learning. *J. Hydrol.* 632:130904. doi: 10.1016/j.jhydrol.2024.130904

Yates, D., Gangopadhyay, S., Rajagopalan, B., and Strzepek, K. (2003). A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resour. Res.* 39:2002WR001769. doi: 10.1029/2002WR001769

Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L., and Tockner, K. (2015). A global boom in hydropower dam construction. *Aquat. Sci.* 77, 161–170. doi: 10.1007/s00027-014-0377-0