Civil Engineering and Applied Solutions

Research Article

journal homepage: ceas.journals.umz.ac.ir

CEAS Civil Engineering and Applied Solutions

Optimizing Reservoir Operations with Reinforcement Learning: A Data-Driven Framework

Fariborz Masoumi [a], Mehdi Jorabloo [b*], Gholamreza Shobeyri [c]

[a] *Department of Civil Engineering, University of Mohaghegh Ardabili, Ardabil, Iran*
[b] *Department of Water Engineering, Islamic Azad University, Garmsar, Iran*
[c] *Faculty of Civil, Water & Environmental Engineering, Shahid Beheshti University, Tehran, Iran*

ARTICLE INFO

Keywords:
Reservoir operation
Reinforcement learning
Dez dam
Machine learning

Article history:
Received 28 July 2025
Accepted 20 August 2025
Available online 01 September 2025

ABSTRACT

Effective reservoir management demands adaptive, data-driven strategies to optimize storage and release decisions while balancing multiple, often competing, operational objectives. This study investigates the application of Q-Learning, a model-free reinforcement learning (RL) algorithm, for optimizing reservoir releases under dynamic and uncertain hydrological conditions. Unlike conventional rule-based or offline optimization methods, the proposed RL approach continuously refines its release policy by learning from environmental feedback and observed states, enabling real-time adaptation without the need for a predefined system model. The framework is tested on the Dez Reservoir in Iran, a real-world case study characterized by significant inflow variability and seasonal water demand. Simulation results demonstrate that Q-Learning effectively manages operational complexity, maintaining storage within prescribed bounds and delivering release patterns closely aligned with demand. To benchmark performance, a simplified Ant Colony Optimization (ACO) model is implemented for comparison. While ACO shows moderate capability in deficit reduction, Q-Learning outperforms it in terms of constraint satisfaction and long-term feasibility. Findings highlight the strong potential of reinforcement learning to support intelligent, scalable, and robust decision-making in modern reservoir operation systems under uncertainty.

1. Introduction

1.1. Background and importance of reservoir operation

Water resources management is critical to ensuring a sustainable water supply, hydropower generation, flood mitigation, and ecological conservation. Effective reservoir operation entails complex decision-making to balance competing objectives under uncertain and often dynamic hydrological conditions. Traditional reservoir operation methods, typically based on rule curves or static optimization, frequently lack the flexibility required to respond in real-time to changing environmental and operational scenarios.

1.2. Reinforcement learning and its application to reservoir systems

Reinforcement Learning (RL), a branch of machine learning, enables autonomous agents to learn optimal actions through interactions with their environment, guided by feedback in the form of rewards or penalties. Unlike conventional rule-based or deterministic optimization methods, RL can dynamically refine its decision policies based on observed outcomes, making it well-suited for managing systems with uncertainty and evolving constraints.

* Corresponding author.
E-mail addresses: jorabloo.mehdi@gmail.com (M. Jorabloo).

https://doi.org/10.22080/ceas.2025.29738.1032
ISSN: 3092-7749/
How to cite this article: Masoumi, F., Jorabloo, M., Shobeyri, G. Optimizing Reservoir Operations with Reinforcement Learning: A Data-Driven Framework. Civil Engineering and Applied Solutions. 2025; 1(4): 56–67. doi: 10.22080/ceas.2025.29738.1032.

Among RL techniques, Q-Learning is particularly advantageous for reservoir operations due to its model-free architecture. By building a state-action value function (Q-table), the algorithm learns to adjust reservoir releases based on system observations such as storage levels, inflow rates, and demand forecasts. Its integration into reservoir operations aims to: Improve adaptability in the face of hydrological uncertainty; and enhance operational efficiency while respecting system constraints.

2. Literature review on reservoir operation

Water resource challenges are both complex and complicated, shaped by dynamic interactions among natural processes, socio-economic factors, and diverse stakeholder interests. The complex aspect arises from system uncertainty, feedback loops, and adaptive human and ecological behaviors, while the complicated dimension reflects the involvement of numerous variables, nonlinear relationships, and technical constraints [1]. Reservoir operation optimization, an inherently complicated issue, has evolved significantly over recent decades, mirroring the growing multidimensionality of water resource systems [2, 3].

A wide range of optimization models has been developed to address various water-related problems. These include applications in non-point source pollution control [4-6], decentralized water resources management [7], urban water sustainability [8], sewer network design [9], and reservoir operation under temperature variability [10]. In addition, water reallocation policies have been developed under inflow and demand uncertainty [11], while optimization has also been combined with social behavior models [12], the water-energy-food-greenhouse gas nexus [13], and Water Sensitive Urban Design [14]. These diverse applications highlight the critical role of optimization techniques as powerful tools for navigating trade-offs, improving resource allocation efficiency, and supporting integrated, resilient water management strategies under uncertainty.

A review of over 50 studies reveals a rich landscape of methodologies from traditional optimization techniques to hybrid metaheuristics and intelligent algorithms applied across diverse hydrological, ecological, and operational contexts. Early methods in this domain primarily relied on Traditional Models (TM), such as Linear Programming (LP), Nonlinear Programming (NLP), and Dynamic Programming (DP). These methods provided a rigorous analytical foundation for reservoir optimization but were constrained by computational inefficiencies and limitations in handling stochastic variables and system complexity. For instance, Li et al. [15] and Zhao et al. [16] applied improved DP to multi-reservoir systems, highlighting the method's effectiveness but also its scale limitations. Similarly, Zeng et al. [17] introduced enhancements to DP for parallel reservoir systems to address computational challenges.

The need for flexible and scalable solutions led to a shift toward Evolutionary Algorithms (EA) and Swarm Intelligence (SI) methods, collectively termed EA-SI. Algorithms such as Genetic Algorithms (GA) [18], Particle Swarm Optimization (PSO) [19, 20], and Differential Evolution (DE) [21] have been widely employed. These algorithms allow efficient exploration of nonlinear and non-convex search spaces and are particularly well-suited to multi-objective and multi-reservoir operation problems.

A major development has been the emergence of Metaheuristic Algorithms (MHA), often hybridized with AI or data-driven approaches. For example, Emami et al. [22] introduced a hybrid constrained coral reefs optimization algorithm integrated with machine learning to improve multi-reservoir operational efficiency. Similarly, Moeini and Babaei [23] combined support vector machines (SVM) with a PSO framework for managing reservoir operations under uncertain future conditions.

Among notable metaheuristics, the Imperialist Competitive Algorithm (ICA) stands out. Afshar et al. [4] applied ICA for optimizing piecewise linear operating rule curves in the Dez reservoir, Iran, demonstrating efficient convergence and robustness. The method used a parameterization–simulation–optimization loop and incorporated adaptive penalty functions for constraint handling, making it effective in addressing both water supply and hydropower objectives.

Other innovative algorithms include the Seagull Optimization Algorithm (SOA) used by Ehteram et al. [24], the Jaya Algorithm by Chong et al. [25] and Paliwal et al. [26], and the Charged System Search (CSS) applied by Latif et al. [27]. The Moth-Flame Optimization (MFO) algorithm [28], Sine Cosine Algorithm (SCA) [29], and Hybrid Bat-Swarm Algorithm [30] further enrich the methodological diversity in the field.

Multi-objective optimization approaches are particularly prevalent, addressing trade-offs between hydropower production, water supply, and ecological needs. Liu et al. [31] proposed a bi-objective NSGA-II-based model, while Feng et al. [32] and Meng et al. [33] developed multi-objective frameworks that optimize ecological and economic outcomes simultaneously. Zhang et al. [34]employed a multi-elite guide PSO for multi-reservoir operation, and Raso et al. [35] evaluated operation scenarios to reduce conflicts between hydropower generation and traditional water uses.

Several studies applied optimization to special configurations, such as cascade systems [36, 37], gate-controlled reservoirs [20], or systems integrating hydro-photovoltaic power [38]. Others have considered sedimentation and water-quality dynamics [39, 40], while Allawi et al. [41] explored the use of intelligent models for hydrological parameter forecasting and its influence on operational performance.

Another important advancement is the use of Penalty Functions (PF) in objective formulation. Many reviewed studies incorporate soft constraint enforcement through penalty-based formulations to maintain feasibility under varying hydrological conditions [42, 43].

Furthermore, hybrid and adaptive algorithms are being used to improve convergence and solution stability. For instance, Ahmadianfar et al. [21] applied an adaptive hybrid DE for hydropower scheduling, while Turgut et al. [44] developed a master-slave algorithm for optimizing release policies under real-time constraints.

Despite the methodological diversity and advances, most reviewed studies rely on offline optimization frameworks. This reveals an important research gap: the lack of real-time adaptability in conventional algorithms, which are often unable to adjust to dynamic inflow conditions or user demands without re-optimization.

To bridge this gap, Reinforcement Learning (RL) has recently been identified as a promising paradigm. Unlike traditional methods, RL enables model-free policy learning through continuous interaction with the environment, adapting operational rules dynamically as new data becomes available. Given the high dimensionality, uncertainty, and nonlinearity of reservoir systems, RL offers a viable extension or alternative to existing optimization frameworks.

In summary, the reviewed body of work demonstrates extensive innovation in reservoir operation modeling, spanning deterministic and stochastic frameworks, single and multi-objective formulations, and hybrid intelligent algorithms. As the field progresses, integrating real-time adaptive learning such as RL with the rigor of metaheuristic optimization presents a promising direction for robust, resilient water resource management.

3. Methodology

Reservoir operation constitutes a complex, multi-objective decision-making process that requires adaptive strategies to manage inflow variability, storage dynamics, and release policies. Conventional optimization approaches such as linear programming (LP), nonlinear programming (NLP), and dynamic programming (DP) have been widely employed. However, these methods typically rely on fixed mathematical models, limiting their adaptability under uncertain and evolving hydrological conditions.

Reinforcement Learning (RL) provides a data-driven framework for reservoir operation by enabling an agent to learn optimal decisions through repeated interactions with its environment. This adaptive trial-and-error mechanism allows policies to evolve in response to changing conditions. In this study, Q-Learning, a model-free RL algorithm, is used to dynamically optimize reservoir release policies, eliminating the need for predefined operating rules or model assumptions.

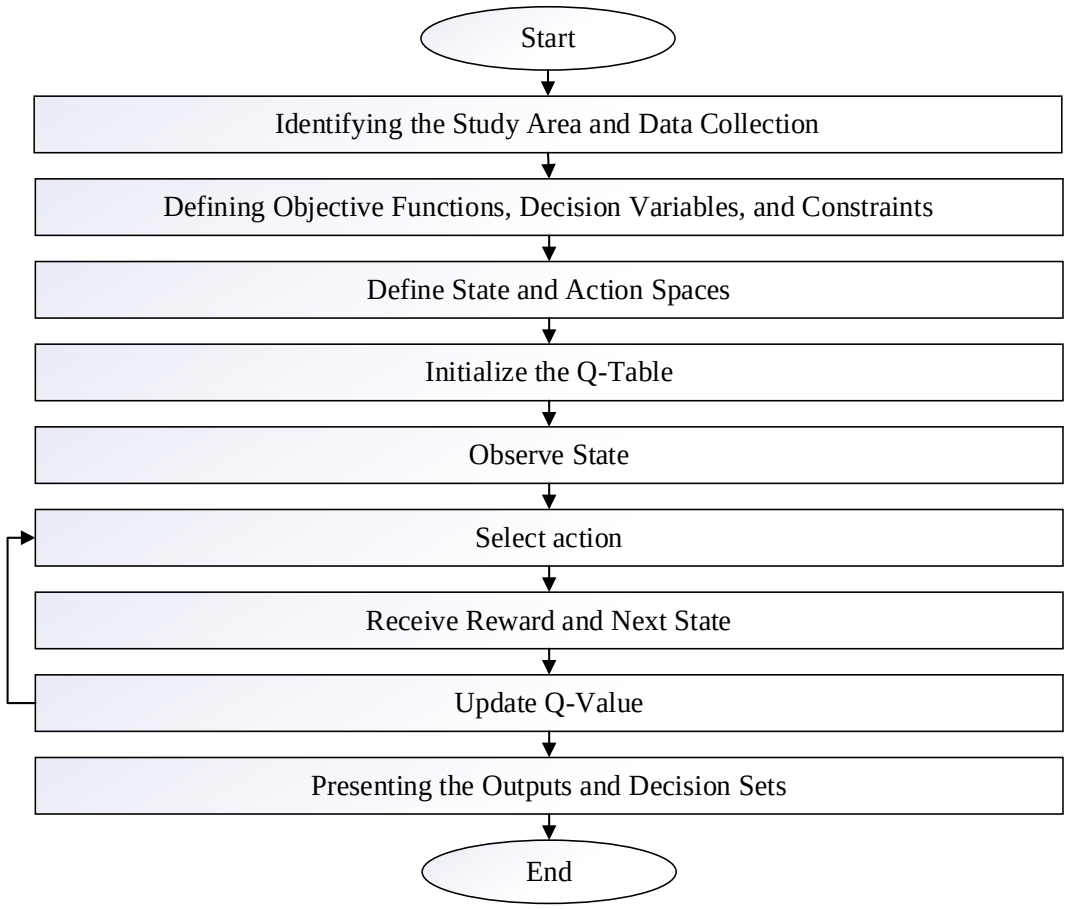

Fig. 1. The flowchart of the proposed methodology of RL application in reservoir operation.

In reinforcement learning, the system's state space represents the current storage level of the reservoir. In this study, the reservoir storage levels range from minimum (s_{min}) to maximum (s_{max}), with discrete intervals representing different states. The action space corresponds to the release decisions, constrained within a predefined range (r_{min} to r_{max}) to maintain operational limits.

State transitions are governed by the interplay between inflow variability, storage updates, and water demand at each time step. This dynamic formulation enables the RL agent to adaptively determine release decisions based on current conditions, allowing it to respond to fluctuations in system states and external inputs. The conceptual framework of the reinforcement learning approach is illustrated in Fig. 2, which summarizes the learning cycle, including environment observation, action execution, reward evaluation, and policy updates.

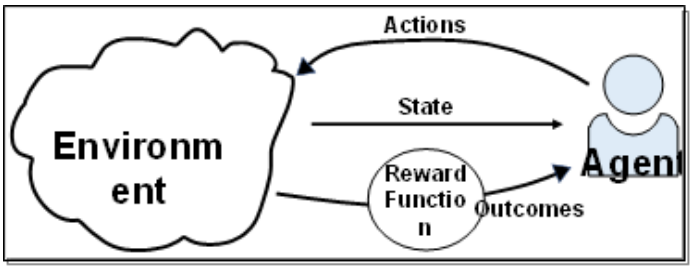

Fig. 2. The main concept of the RL [45].

Recent advancements in artificial intelligence (AI) and machine learning (ML) have introduced data-driven and adaptive strategies for reservoir management. Among these, reinforcement learning (RL) has gained significant attention due to its ability to learn optimal policies through trial-and-error interactions with the environment. In this study, we focus on Q-Learning, a model-free RL algorithm, to optimize reservoir release decisions while enhancing system efficiency and sustainability. Q-Learning is a value-based reinforcement learning method where an agent learns the optimal policy by estimating the expected cumulative reward of state-action pairs. The Q-Learning framework follows these core steps:

1. Initialization:
   - The agent begins with an initial reservoir storage value ($s_{initial}$).
   - A Q-table, representing the expected reward of each state-action pair, is initialized with zeros.
   - Historical or synthetic inflow data and demand sequences are predefined for training and evaluation purposes.
2. Action selection:
   - The agent selects reservoir release actions using an epsilon-greedy strategy, which balances exploration (randomized action selection) and exploitation (choosing the highest $Q-value$ action).
   - Initially, exploration dominates to allow the model to learn efficiently, but as training progresses, exploitation becomes more prominent to refine decision-making.
3. Reward function design:
   - The reward function incentivizes meeting demand while discouraging deviations and violations of system constraints. Neutral or positive rewards are assigned when releases closely match demand. Penalties are imposed when releases exceed or fall short of demand, or when storage or release constraints are violated.
4. Q-Value update (bellman equation):
   - The $Q-values$ are updated iteratively using the Bellman equation, which considers:
   - Immediate reward based on the release decision.
   - Future reward expectations, incorporating a discount factor (γ) to prioritize long-term outcomes.
   - Learning rate (α) to control the magnitude of updates.

   The Q-value update formula is given by:

$$Q_{new}(s_t, a_t) \leftarrow (1-\alpha) \times Q_{old}(s_t, a_t) + \alpha \times (r_t + \gamma \times max\, Q(s_{t+1}, a')) \quad (1)$$

where, t stands for the time step, $Q_{old}(s_t, a_t)$ is the Q-factor in the previous episode for action a_t and state s_t; $Q_{new}(s_t, a_t)$ is the Q-factor in the new episode for action a_t and state s_t; s_t is the state of the system (here, the reservoir condition) in the time step t, and a_t is the action in the time step t; s_{t+1} is the next state is determined based on the selected action a_t; the symbol α represents the learning rate, and it is between 0 and 1; it reflects how significantly the reward signal or r_t influences the update of the value function or Q; γ is the discount rate, it is between 0 and 1, it represents how a decision influences future possibilities and how the value function guides that decision-making process; $max\, Q(s_{t+1}, a')$ is the maximum Q-factor in the s_{t+1} mode for all possible actions, r_t is the value of the instantaneous reward received by the agent for the operation of a_t in s_t mode.

3.1. Mathematical formulation (single-reservoir, monthly time step)

In this section, the main formulation for the reservoir simulation model is presented. The main formulation is presented below.

$$S_{t+1} = S_t + I_t - R_t \quad (2)$$

where, S_{t+1}: Storage volume at the beginning of the month $t+1$. S_t: Storage volume at the beginning of the month t. I_t: Inflow volume into the reservoir during the month t. I_t: Inflow volume into the reservoir during the month t; R_t: Release volume from the reservoir during the month t.

3.2. Exploration vs exploitation mechanism

To promote learning, the model begins with a high probability of exploration, randomly selecting actions in early episodes. As learning progresses, this probability decreases, allowing the agent to increasingly exploit its learned policy. Periodic reintroduction of random actions helps avoid local optima and ensures broader policy robustness.

3.3. Training and convergence criteria

The model is trained over a fixed number of episodes (e.g., 500,000), each simulating reservoir operation across the defined time horizon. Training is considered converged when:

- $Q - values$ stabilize, indicating convergence towards an optimal policy.
- Exploration probability is sufficiently reduced, favoring exploitation.
- Reservoir release performance meets predefined operational benchmarks.

3.4. Implementation details

The RL algorithm is implemented in MATLAB, utilizing:

- Storage state discretization, defining reservoir states at incremental levels within operational bounds.
- Q-table structure, storing state-action values for iterative updates.
- Demand incorporation, ensuring releases are adjusted to meet water requirements dynamically.
- Performance tracking, analyzing learning progress and policy improvements through reward evolution.

3.5. Significance and practical implications

This RL-based approach offers multiple advantages over traditional reservoir operation methods: Dynamic Adaptation: The Q-Learning framework continuously adjusts decisions, improving release efficiency across varying hydrological conditions. Optimization Without Explicit Models: Unlike traditional optimization techniques, Q-Learning does not require explicit mathematical models, making it highly adaptable to data-driven environments. Efficient Water Resource Allocation: By refining release decisions, the methodology enhances water supply reliability, hydropower production efficiency, and environmental sustainability.

4. Case study

The Dez Reservoir, situated in southwestern Iran, is a critical infrastructure element for regional water supply, supporting agricultural irrigation, domestic consumption, and industrial use. Constructed on the Dez River, the reservoir plays a pivotal role in sustaining water availability across Khuzestan Province, where seasonal and climatic variability significantly influence demand. This case study examines the reservoir's hydrological features, operational parameters, and its strategic role in ensuring long-term water security for multiple sectors.

- Geographical and hydrological overview

Located in a semi-arid region, the Dez Reservoir receives its inflows predominantly from rainfall and snowmelt originating in the Zagros Mountains. With a total storage capacity exceeding 3 billion cubic meters, it ranks among the largest reservoirs in Iran. Inflow patterns exhibit substantial seasonal and interannual variability, shaped by precipitation and upstream hydrological dynamics. Effective management of these fluctuations is vital to avoid shortages and maintain a continuous supply for both agricultural and urban demands.

To regulate storage and distribution, the reservoir operates under a structured management framework, adjusting release rates based on agricultural demands and municipal consumption trends. Water allocation models are integrated into reservoir operations to ensure equitable distribution, particularly during dry periods when inflow rates decline.

- Water supply management and distribution

As a cornerstone of Iran’s water infrastructure, the Dez Reservoir supports extensive irrigation systems that serve large-scale agricultural operations. These networks rely on seasonal demand forecasts to guide water allocation, helping sustain crop yields while minimizing reservoir drawdown. In parallel, the reservoir provides essential potable water to urban and industrial sectors through municipal distribution systems. Downstream water treatment facilities ensure the supply of safe drinking water by incorporating filtration and purification processes. During drought events, reservoir managers employ adaptive release strategies to prioritize essential needs and conserve storage volumes.

Municipal water supply remains a key operational priority. The reservoir supports urban water distribution networks, delivering potable water to communities and industries. Water treatment facilities downstream process and distribute clean drinking water, integrating filtration and purification technologies to maintain quality standards.

During drought conditions, reservoir managers implement adaptive strategies, adjusting release patterns to conserve storage

while ensuring essential water needs are met.

For simulation purposes, the reservoir is characterized by a maximum storage capacity of 3,340 million cubic meters (MCM) and a minimum operational threshold of 830 MCM. The initial storage is set to 1,430 MCM, reflecting a mid-range condition. The model enforces a maximum monthly release limit of 1,000 MCM, with the flexibility to withhold releases entirely during low-flow conditions. Historical inflow data reveal considerable variability:

- 5-year average inflow: 5,303 MCM
- 20-year average: 5,990 MCM
- 40-year average: 5,951 MCM

This variability underscores the need for robust and flexible operational policies that can accommodate both short-term fluctuations and long-term hydrological trends.

5. Results and discussion

This section presents a comparative evaluation of simple Ant Colony Optimization (ACO) and Q-Learning algorithms applied to a single-reservoir operation problem over a 60-month planning horizon. Both models were implemented in MATLAB under consistent hydrological inputs, operational constraints, and demand profiles to ensure fairness and reproducibility. The objective was to minimize the total demand deficit while satisfying system constraints related to storage bounds and release capacities. Key performance indicators included cumulative fitness score, total unmet demand, number of constraint violations, and monthly operational trends. Graphical representations of release decisions, storage dynamics, and deficits offer qualitative insights into each algorithm's operational behavior and effectiveness.

5.1. Performance metrics comparison

The comparative analysis relies on system-wide performance indicators derived from each algorithm's output. Table 1 summarizes the key metrics, including fitness score, demand deficit, and constraint violations.

Table 1. Summary of performance metrics.

Metric	ACO	Q-Learning	Unit
Total fitness	752,879.33	4,322.03	Unitless (composite penalty score)
Total demand deficit	1,003.50	4,322.03	Million Cubic Meters (MCM)
Storage violations	32	0	Count
Release violations	0	0	Count

The total fitness score, which combines penalties from demand deficits and constraint violations, reveals a substantial performance gap between the two methods. Q-Learning achieved a significantly lower fitness score (4,322.03) compared to ACO (752,879.33), indicating a more feasible and operationally robust reservoir strategy. The low score for Q-Learning reflects its ability to maintain system constraints while balancing competing objectives.

Interestingly, the total demand deficit was higher under Q-Learning (4,322.03 MCM) than ACO (1,003.50 MCM). At first glance, this might suggest superior demand satisfaction by ACO. However, this interpretation overlooks the broader context: ACO violated the reservoir's storage constraints 32 times during the simulation, each breach contributing heavily to its overall penalty and inflated fitness score. These violations indicate instability in reservoir behavior, compromising the practical feasibility of the solution.

In contrast, Q-Learning maintained zero violations across all episodes, consistently operating within both storage and release limits. Although this came at the cost of some unmet demand, the agent prioritized long-term sustainability and operational reliability as core goals in real-world reservoir management.

Furthermore, both algorithms successfully avoided release violations, demonstrating that release bounds were effectively enforced. This reinforces the notion that the key differentiator between the two lies in how they handle storage dynamics and long-term feasibility, not just raw demand satisfaction.

5.2. Monthly release and storage trends

To evaluate the operational performance of both algorithms over time, monthly data on reservoir releases, storage levels, and unmet demand were analyzed across the 60-month simulation period. The results reveal clear distinctions in the temporal decision-making behaviors of Q-Learning and Ant Colony Optimization (ACO).

Fig. 3 provides a detailed visualization of the operational dynamics generated by the Ant Colony Optimization (ACO) algorithm over the 60-month simulation period. The top panel displays monthly release volumes alongside corresponding demand levels, illustrating that while ACO generally responds to fluctuations in demand, its release decisions are occasionally misaligned due to heuristic-driven exploration or constraint trade-offs.

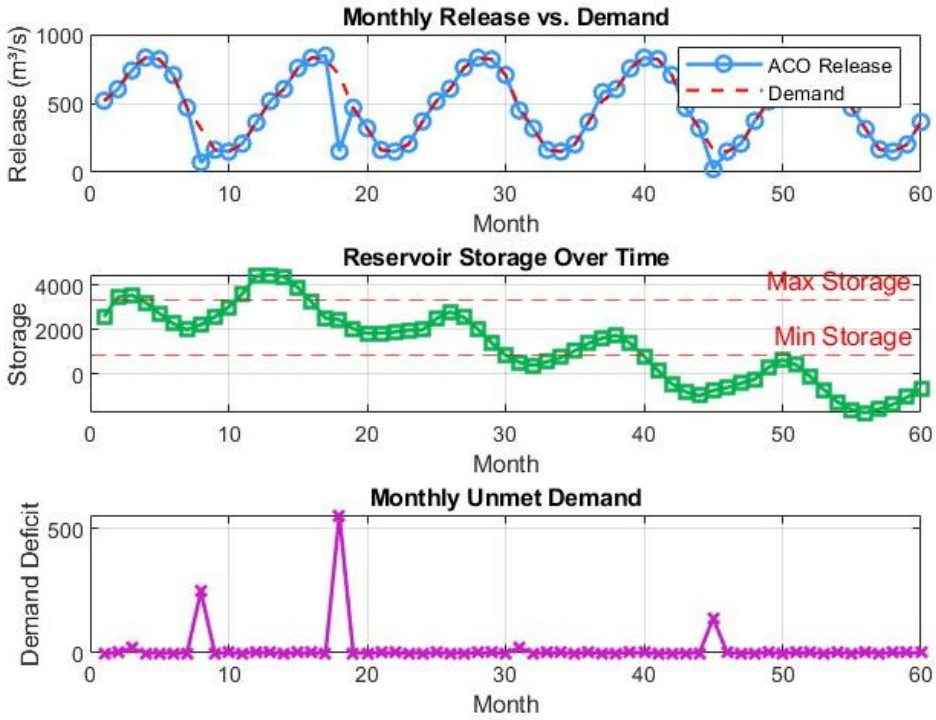

Fig. 3. ACO-based reservoir simulation: monthly dynamics of release, storage, and demand deficit over a five-year horizon.

The middle panel depicts reservoir storage levels, showing considerable variability with multiple exceedances of both the minimum and maximum storage thresholds, clear evidence of constraint violations. The bottom panel presents monthly unmet demand, characterized by intermittent spikes that highlight the system's difficulty in consistently satisfying water needs without breaching operational limits. Collectively, these plots reflect ACO's adaptive but volatile decision-making, which manages to meet demand in many months but often at the expense of violating storage constraints and compromising long-term system feasibility.

In the case of ACO, release decisions exhibited considerable volatility, characterized by frequent oscillations between over-releasing and overly conservative adjustments. This erratic behavior resulted in multiple instances of both storage overshoot and undershoot, ultimately leading to 32 documented violations of the reservoir's operational limits. Although such actions occasionally succeeded in minimizing short-term demand deficits, they undermined the overall stability of the reservoir system and incurred significant penalty costs due to constraint breaches.

In contrast, Q-Learning produced more stable and coherent release patterns. Leveraging an epsilon-greedy strategy and extensive training across 500,000 episodes, the agent progressively refined its policy to account for variations in inflow and evolving storage conditions. As a result, reservoir storage levels remained consistently within the defined operational bounds throughout the simulation period, demonstrating the model's capacity to manage hydrological risk effectively and comply with system constraints.

Fig. 4 highlights the effective performance of the Q-Learning algorithm in managing reservoir operations across a 60-month simulation period. In the top panel, the release trajectory produced by Q-Learning closely aligns with monthly water demand, demonstrating the model's capacity to learn and implement balanced release strategies. This responsiveness stems from the algorithm's reinforcement-based learning structure, which enables the agent to refine its decisions through continuous interaction with the environment.

The middle panel shows that reservoir storage levels remained consistently within operational limits throughout the simulation. This adherence to constraints reflects the model's ability to internalize system rules through repeated training and to maintain feasibility even under fluctuating inflow and demand conditions. The absence of storage violations throughout the entire time horizon is a key indicator of the model's reliability in real-world operational contexts.

In the bottom panel, monthly unmet demand levels remain moderate and well-distributed, suggesting that the model occasionally allowed small deficits in favor of preserving long-term operational stability. This behavior illustrates the agent's capacity to manage trade-offs intelligently, avoiding short-term overreactions and instead favoring sustained performance across the entire planning period.

It is important to note that the version of Ant Colony Optimization used in this study represents one of its simplest forms, implemented here primarily for baseline comparison. The strong performance of Q-Learning, especially in terms of constraint compliance and adaptive release behavior, underscores its potential as a robust and scalable approach for data-driven reservoir operation. Its ability to maintain operational stability while responding flexibly to environmental variability makes it a promising tool for future water resources management applications.

Fig. 5 presents a side-by-side comparison of reservoir operations generated by Q-Learning and a simplified implementation of Ant Colony Optimization (ACO) across three key dimensions: release behavior, storage dynamics, and demand deficit over the 60-month planning horizon. The release pattern produced by Q-Learning shows consistent alignment with monthly demand, highlighting the model's capacity to learn adaptive and balanced strategies that evolve in response to inflow and system state changes.

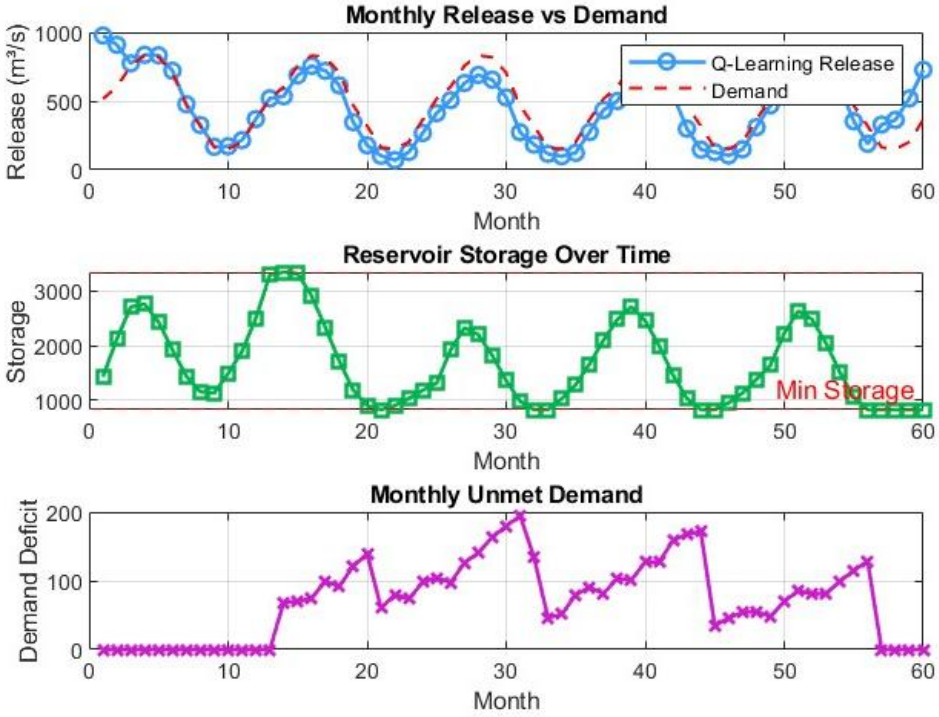

Fig. 4. Q-Learning-based reservoir operation: performance insights from release, storage, and deficit dynamics.

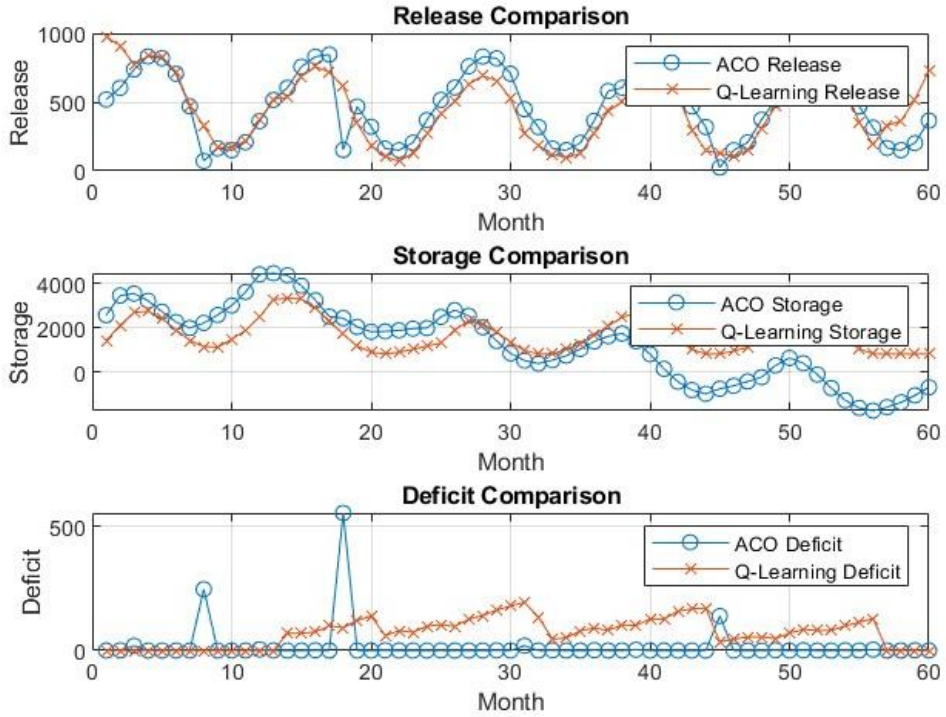

Fig. 5. Side-by-side algorithmic comparison of reservoir operations: ACO vs Q-Learning.

In terms of storage management, Q-Learning maintains reservoir levels well within operational boundaries throughout the simulation period. This reflects the algorithm's effective internalization of system constraints and its ability to avoid violations even under variable conditions. The storage profile demonstrates stability and discipline, indicating the agent's focus on long-term operational feasibility.

The deficit panel further illustrates Q-Learning's strategic behavior. While small deficits occur in certain months, they are

distributed evenly and remain within acceptable limits. This suggests the model occasionally allows controlled shortfalls to preserve broader system integrity over time, a hallmark of sustainable reservoir operation.

It should be noted that the ACO implementation used here serves as a basic benchmark to contrast learning-based decision-making with heuristic search methods. Despite its simplicity, the comparison helps to underscore the key advantage of Q-Learning: its ability to dynamically improve performance through experience, adapt to environmental variability, and maintain consistent compliance with operational rules. Overall, Fig. 5 reinforces Q-Learning's promise as a reliable, adaptive approach for managing reservoir systems under uncertainty.

5.3. Constraint handling and penalty impacts

Effective constraint management is a critical aspect of realistic reservoir operation modeling. In this study, both algorithms incorporated mechanisms to address storage and release constraints, but their strategies and outcomes varied significantly.

Q-Learning handled operational constraints through a built-in reward function that inherently discouraged violations. Over the course of training, the agent learned via repeated interactions and feedback to avoid actions that would lead to infeasible storage or release values. This internalization of system rules resulted in zero storage or release violations during the entire simulation, indicating the algorithm's strength in achieving not only high performance but also operational reliability.

The Ant Colony Optimization approach used here employed a basic penalty structure, where violations of constraints were accounted for by adding penalty terms to the fitness score. While this approach provided a mechanism to discourage infeasible actions, the version of ACO implemented in this study was intentionally kept simple for baseline comparison. As such, the algorithm occasionally prioritized short-term improvements in demand satisfaction at the expense of constraint adherence, leading to elevated fitness penalties due to storage violations.

This contrast highlights one of the key advantages of Q-Learning: its ability to evolve policies that balance immediate objectives with long-term feasibility. Rather than relying solely on external penalties, Q-Learning gradually developed an operational strategy that inherently respected system constraints while maintaining stable and adaptive behavior across varying hydrological conditions.

5.4. Computational efficiency and scalability

In addition to performance and feasibility, computational efficiency and scalability are important considerations in selecting an appropriate reservoir operation algorithm, particularly for real-time or large-scale applications.

The Q-Learning model, though trained over a large number of episodes (500,000 in this study), benefited from a lightweight and structured learning architecture. Its use of matrix-based Q-table representations allowed for efficient updates and rapid convergence tracking. Once training was completed, the learned policy could be executed instantaneously for decision-making, well-suited for real-time operational deployment. Moreover, the model's independence from explicit hydrological equations enhances its adaptability to different system configurations, suggesting strong potential for scalability to more complex, multi-reservoir systems or longer planning horizons.

The Ant Colony Optimization (ACO) approach, implemented here in a basic form for benchmarking purposes, followed a heuristic search process that involved multiple agents (ants) exploring the solution space. This method required repeated pheromone updates and solution evaluations across iterations, which increased computational overhead, particularly as the number of decision variables (e.g., months or release levels) expanded. While more advanced variants of ACO can reduce this burden through parallelization and parameter tuning, the simplified version applied in this study was not optimized for efficiency or scale.

Overall, Q-Learning demonstrated greater potential for practical application in terms of computational performance and scalability. Its structured learning process, adaptability, and low execution cost once trained position it as a promising tool for modern, data-driven water resources management systems.

5.5. Final verdict and implications

The comparative results of this study highlight the strong potential of Q-Learning as a reliable and adaptive tool for reservoir operation. Across all evaluated dimensions, constraint satisfaction, operational stability, computational efficiency, and scalability, Q-Learning consistently delivered feasible and well-balanced release strategies. Its ability to learn from interaction with the environment and evolve robust policies over time makes it especially well-suited for dynamic and uncertain hydrological systems.

The Q-Learning model achieved complete compliance with operational constraints, maintained storage levels within prescribed bounds, and produced release patterns closely aligned with demand. Although moderate unmet demand was observed in some months, these shortfalls were strategically distributed and acceptable within the context of maintaining long-term system stability. Once trained, the algorithm was computationally efficient and highly responsive, demonstrating potential for real-time deployment and scalability to more complex systems, such as multi-reservoir networks.

The Ant Colony Optimization (ACO) algorithm used in this study was implemented in its simplest form to serve as a heuristic-based baseline for comparison. While it showed some ability to reduce short-term demand deficits, it lacked the adaptive learning capabilities needed to consistently balance performance with feasibility. More advanced versions of ACO may perform better under different configurations; however, the results here emphasize the added value of incorporating reinforcement learning into water

resources planning frameworks.

In conclusion, this study reinforces the role of reinforcement learning, particularly Q-Learning, as a promising paradigm for intelligent reservoir management. Its model-free nature, adaptive behavior, and real-time decision-making capability position it as a valuable alternative or complement to traditional and metaheuristic optimization techniques. As climate variability and increasing demand continue to challenge water systems, data-driven and self-improving models like Q-Learning are likely to play a central role in the next generation of sustainable water resources management.

6. Conclusions

This study demonstrated the effectiveness of reinforcement learning, specifically Q-Learning, for optimizing reservoir operations under uncertain and dynamic hydrological conditions. By learning optimal release strategies through iterative interactions with the system, the Q-Learning agent was able to maintain storage and release decisions within prescribed constraints while achieving a high level of adaptability to variable inflows and demand scenarios. Compared to a simplified implementation of Ant Colony Optimization (ACO), the Q-Learning model exhibited superior performance in operational feasibility, stability, and scalability. Although ACO occasionally yielded lower demand deficits, it did so at the cost of constraint violations and operational instability. These results underscore the potential of model-free reinforcement learning approaches in supporting intelligent, data-driven water resources management. Future research may extend this framework to multi-reservoir systems, integrate climate uncertainty, and hybridize RL models with physical simulation tools to enhance realism and generalizability.

Statements & Declarations

Author contributions

Fariborz Masoumi: Conceptualization, Methodology, Formal analysis, Resources, Original Draft, Writing.

Mehdi Jorabloo: Conceptualization, Methodology, Formal analysis, Resources, Original Draft, Writing.

Gholamreza Shobeyri: Conceptualization, Methodology, Formal analysis, Resources, Original Draft, Writing.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declarations

The authors declare no conflict of interes.

References

[1] Emami-Skardi, M. J., Shobeyri, G., Kerachian, R. Analysis of Stakeholder Relationships and Conflicts Using the Conflict Tree Approach. Iran-Water Resources Research, 2023; 18: 57–74. doi:20.1001.1.17352347.1401.18.4.4.8.

[2] Afshar, A., Emami Skardi, M. J., Jerani, F. Pond designing optimization using Multi-Objective ant colony algorithm and swat model. Journal of Environmental Science and Technology, 2015; 16 (Special issue): 121–132.

[3] Masoumi, F., Masoumzadeh, S., Zafari, N., Emami-Skardi, M. J. Optimal operation of single and multi-reservoir systems via hybrid shuffled grey wolf optimization algorithm (SGWO). Water Supply, 2021; 22: 1663–1675. doi:10.2166/ws.2021.326.

[4] Afshar, A., Emami Skardi, M. J., Masoumi, F. Optimizing water supply and hydropower reservoir operation rule curves: An imperialist competitive algorithm approach. Engineering Optimization, 2015; 47: 1208–1225. doi:10.1080/0305215X.2014.958732.

[5] Emami Skardi, M. J., Afshar, A., Saadatpour, M., Sandoval Solis, S. Hybrid ACO–ANN-Based Multi-objective Simulation–Optimization Model for Pollutant Load Control at Basin Scale. Environmental Modeling & Assessment, 2015; 20: 29–39. doi:10.1007/s10666-014-9413-7.

[6] Skardi, M. J. E., Afshar, A., Solis, S. S. Simulation-optimization model for non-point source pollution management in watersheds: Application of cooperative game theory. KSCE Journal of Civil Engineering, 2013; 17: 1232–1240. doi:10.1007/s12205-013-0077-7.

[7] Moradikian, S., Emami-Skardi, M. J., Kerachian, R. A distributed constraint multi-agent model for water and reclaimed wastewater allocation in urban areas: Application of a modified ADOPT algorithm. Journal of Environmental Management, 2022; 317: 115446. doi:10.1016/j.jenvman.2022.115446.

[8] Emamjomehzadeh, O., Kerachian, R., Emami-Skardi, M. J., Momeni, M. Combining urban metabolism and reinforcement learning concepts for sustainable water resources management: A nexus approach. Journal of Environmental Management, 2023; 329: 117046. doi:10.1016/j.jenvman.2022.117046.

[9] Masoumi, F., Masoumzadeh, S., Zafari, N., Javad Emami-Skardi, M. Optimum sanitary sewer network design using shuffled gray wolf optimizer. Journal of Pipeline Systems Engineering and Practice, 2021; 12: 04021055. doi:10.1061/(ASCE(PS.1949-1204.0000597.

[10] Tabari, M. M. R., Azadani, M. N., Kamgar, R. Development of operation multi-objective model of dam reservoir under conditions of temperature variation and loading using NSGA-II and DANN models: a case study of Karaj/Amir Kabir dam. Soft Computing, 2020; 24: 12469–12499. doi:10.1007/s00500-020-04686-1.

[11] Tabari, M. M. R., Safari, R. Development of water re-allocation policy under uncertainty conditions in the inflow to reservoir and demands parameters: a case study of Karaj AmirKabir dam. Soft Computing, 2023; 27: 6521–6547. doi:10.1007/s00500-023-07885-8.

[12] Eyni, A., Skardi, M. J. E., Kerachian, R. A regret-based behavioral model for shared water resources management: Application of the correlated equilibrium concept. Science of The Total Environment, 2021; 759: 143892. doi:10.1016/j.scitotenv.2020.143892.

[13] Emamjomehzadeh, O., Omidi, F., Kerachian, R., Emami-Skardi, M. J., Momeni, M. Water-energy-food-greenhouse gases nexus management in urban environments: A robust multi-agent decision-support system. Sustainable Cities and Society, 2024; 113: 105676. doi:10.1016/j.scs.2024.105676.

[14] Sharifian, H., Emami-Skardi, M. J., Behzadfar, M., Faizi, M. Water sensitive urban design (WSUD) approach for mitigating groundwater depletion in urban geography; through the lens of stakeholder and social network analysis. Water Supply, 2022; 22: 5833–5852. doi:10.2166/ws.2022.206.

[15] Li, C., Zhou, J., Ouyang, S., Ding, X., Chen, L. Improved decomposition–coordination and discrete differential dynamic programming for optimization of large-scale hydropower system. Energy Conversion and Management, 2014; 84: 363–373. doi:10.1016/j.enconman.2014.04.065.

[16] Zhao, T., Cai, X., Lei, X., Wang, H. Improved dynamic programming for reservoir operation optimization with a concave objective function. Journal of Water Resources Planning and Management, 2012; 138: 590–596. doi:10.1061/(ASCE)WR.1943-5452.0000205.

[17] Zeng, X., Hu, T., Cai, X., Zhou, Y., Wang, X. Improved dynamic programming for parallel reservoir system operation optimization. Advances in Water Resources, 2019; 131: 103373. doi:10.1016/j.advwatres.2019.07.003.

[18] Tegegne, G., Kim, Y.-O. Representing inflow uncertainty for the development of monthly reservoir operations using genetic algorithms. Journal of Hydrology, 2020; 586: 124876. doi:10.1016/j.jhydrol.2020.124876.

[19] Al-Aqeeli, Y. H., Mahmood Agha, O. M. A. Optimal Operation of Multi-reservoir System for Hydropower Production Using Particle Swarm Optimization Algorithm. Water Resources Management, 2020; 34: 3099–3112. doi:10.1007/s11269-020-02583-8.

[20] Kim, Y.-G., Sun, B.-Q., Kim, P., Jo, M.-B., Ri, T.-H., Pak, G.-H. A study on optimal operation of gate-controlled reservoir system for flood control based on PSO algorithm combined with rearrangement method of partial solution groups. Journal of Hydrology, 2021; 593: 125783. doi:10.1016/j.jhydrol.2020.125783.

[21] Ahmadianfar, I., Kheyrandish, A., Jamei, M., Gharabaghi, B. Optimizing operating rules for multi-reservoir hydropower generation systems: An adaptive hybrid differential evolution algorithm. Renewable Energy, 2021; 167: 774–790. doi:10.1016/j.renene.2020.11.152.

[22] Emami, M., Nazif, S., Mousavi, S.-F., Karami, H., Daccache, A. A hybrid constrained coral reefs optimization algorithm with machine learning for optimizing multi-reservoir systems operation. Journal of Environmental Management, 2021; 286: 112250. doi:10.1016/j.jenvman.2021.112250.

[23] Moeini, R., Babaei, M. Hybrid SVM-CIPSO methods for optimal operation of reservoir considering unknown future condition. Applied Soft Computing, 2020; 95: 106572. doi:10.1016/j.asoc.2020.106572.

[24] Ehteram, M., Banadkooki, F. B., Fai, C. M., Moslemzadeh, M., Sapitang, M., Ahmed, A. N., Irwan, D., El-Shafie, A. Optimal operation of multi-reservoir systems for increasing power generation using a seagull optimization algorithm and heading policy. Energy Reports, 2021; 7: 3703–3725. doi:10.1016/j.egyr.2021.06.008.

[25] Chong, K. L., Lai, S. H., Ahmed, A. N., Wan Jaafar, W. Z., El-Shafie, A. Optimization of hydropower reservoir operation based on hedging policy using Jaya algorithm. Applied Soft Computing, 2021; 106: 107325. doi:10.1016/j.asoc.2021.107325.

[26] Paliwal, V., Ghare, A. D., Mirajkar, A. B., Bokde, N. D., Feijoo Lorenzo, A. E. Computer modeling for the operation optimization of mula reservoir, Upper Godavari Basin, India, Using the Jaya Algorithm. Sustainability, 2019; 12: 84. doi:10.3390/su12010084.

[27] Latif, S. D., Marhain, S., Hossain, M. S., Ahmed, A. N., Sherif, M., Sefelnasr, A., El-Shafie, A. Optimizing the operation release policy using charged system search algorithm: a case study of Klang Gates Dam, Malaysia. Sustainability, 2021; 13: 5900. doi:10.3390/su13115900.

[28] Zhang, Z., Qin, H., Yao, L., Liu, Y., Jiang, Z., Feng, Z., Ouyang, S. Improved Multi-objective Moth-flame Optimization Algorithm based on R-domination for cascade reservoirs operation. Journal of Hydrology, 2020; 581: 124431. doi:10.1016/j.jhydrol.2019.124431.

[29] Feng, Z.-k., Liu, S., Niu, W.-j., Li, B.-j., Wang, W.-c., Luo, B., Miao, S.-m. A modified sine cosine algorithm for accurate global optimization of numerical functions and multiple hydropower reservoirs operation. Knowledge-Based Systems, 2020; 208: 106461. doi:10.1016/j.knosys.2020.106461.

[30] Yaseen, Z. M., Allawi, M. F., Karami, H., Ehteram, M., Farzin, S., Ahmed, A. N., Koting, S. B., Mohd, N. S., Jaafar, W. Z. B., Afan, H. A., El-Shafie, A. A hybrid bat–swarm algorithm for optimizing dam and reservoir operation. Neural Computing and Applications, 2019; 31: 8807–8821. doi:10.1007/s00521-018-3952-9.

[31] Liu, D., Huang, Q., Yang, Y., Liu, D., Wei, X. Bi-objective algorithm based on NSGA-II framework to optimize reservoirs operation. Journal of Hydrology, 2020; 585: 124830. doi:10.1016/j.jhydrol.2020.124830.

[32] Feng, Z.-k., Niu, W.-j., Zhang, R., Wang, S., Cheng, C.-t. Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization. Journal of Hydrology, 2019; 576: 229–238. doi:10.1016/j.jhydrol.2019.06.045.

[33] Meng, X., Chang, J., Wang, X., Wang, Y. Multi-objective hydropower station operation using an improved cuckoo search algorithm. Energy, 2019; 168: 425–439. doi:10.1016/j.energy.2018.11.096.

[34] Zhang, R., Zhou, J., Ouyang, S., Wang, X., Zhang, H. Optimal operation of multi-reservoir system by multi-elite guide particle swarm optimization. International Journal of Electrical Power & Energy Systems, 2013; 48: 58–68. doi:10.1016/j.ijepes.2012.11.031.

[35] Raso, L., Bader, J.-C., Weijs, S. Reservoir operation optimized for hydropower production reduces conflict with traditional water uses in the Senegal River. Journal of Water Resources Planning and Management, 2020; 146: 05020003. doi:10.1061/(ASCE)WR.1943-5452.0001076.

[36] Luo, J., Qi, Y., Xie, J., Zhang, X. A hybrid multi-objective PSO–EDA algorithm for reservoir flood control operation. Applied Soft Computing, 2015; 34: 526–538. doi:10.1016/j.asoc.2015.05.036.

[37] Niu, W.-j., Feng, Z.-k., Liu, S. Multi-strategy gravitational search algorithm for constrained global optimization in coordinative operation of multiple hydropower reservoirs and solar photovoltaic power plants. Applied Soft Computing, 2021; 107: 107315. doi:10.1016/j.asoc.2021.107315.

[38] Li, F.-F., Qiu, J. Multi-objective optimization for integrated hydro–photovoltaic power system. Applied Energy, 2016; 167: 377–384. doi:10.1016/j.apenergy.2015.09.018.

[39] Bai, T., Wei, J., Chang, F.-J., Yang, W., Huang, Q. Optimize multi-objective transformation rules of water-sediment regulation for cascade reservoirs in the Upper Yellow River of China. Journal of Hydrology, 2019; 577: 123987. doi:10.1016/j.jhydrol.2019.123987.

[40] Kurek, W., Ostfeld, A. Multi-objective optimization of water quality, pumps operation, and storage sizing of water distribution systems. Journal of Environmental Management, 2013; 115: 189–197. doi:10.1016/j.jenvman.2012.11.030.

[41] Allawi, M. F., Jaafar, O., Mohamad Hamzah, F., Koting, S. B., Mohd, N. S. B., El-Shafie, A. Forecasting hydrological parameters for reservoir system utilizing artificial intelligent models and exploring their influence on operation performance. Knowledge-Based Systems, 2019; 163: 907–926. doi:10.1016/j.knosys.2018.10.013.

[42] Asadieh, B., Afshar, A. Optimization of water-supply and hydropower reservoir operation using the charged system search algorithm. Hydrology, 2019; 6: 5. doi:10.3390/hydrology6010005.

[43] He, Y., Xu, Q., Yang, S., Liao, L. Reservoir flood control operation based on chaotic particle swarm optimization algorithm. Applied Mathematical Modelling, 2014; 38: 4480–4492. doi:10.1016/j.apm.2014.02.030.

[44] Turgut, M. S., Turgut, O. E., Afan, H. A., El-Shafie, A. A novel Master–Slave optimization algorithm for generating an optimal release policy in case of reservoir operation. Journal of Hydrology, 2019; 577: 123959. doi:10.1016/j.jhydrol.2019.123959.

[45] Skardi, M. J. E., Kerachian, R., Abdolhay, A. Water and treated wastewater allocation in urban areas considering social attachments. Journal of Hydrology, 2020; 585: 124757. doi:10.1016/j.jhydrol.2020.124757.