A Comparative Analysis of Proactive Flood Control Policies Using Reinforcement Learning

Eyrin Kim
Stanford University
eyrinkim@stanford.edu

Madison Ho
Stanford University
mlho@stanford.edu

Holland Ferguson
Stanford University
hferg@stanford.edu

Abstract

Dam operators face a persistent trade-off between flood control and water supply, a challenge traditionally managed with static, heuristic-based rule curves. While Reinforcement Learning (RL) has been proposed as a tool for optimizing these operations, its potential as a framework for comparatively analyzing different policy philosophies remains underexplored. This paper evaluates U.S. dam operating strategies by formulating the control problem as a finite Markov Decision Process (MDP). We train and compare three distinct tabular Q-learning agents based on historical data from Folsom Dam, CA. Each agent is trained under a unique reward structure representing a different operational philosophy: (1) a risk-averse policy driven by penalties for extreme states, (2) a rule-curve-following policy incentivized to meet seasonal storage targets, and (3) a myopic, reactive policy with a low discount factor. We evaluate these learned policies against a rule-based heuristic baseline derived from USACE Water Control principles. Our framework quantifies the performance trade-offs inherent in each strategy, revealing how differences in reward structure and long-term planning manifest as measurable impacts on flood mitigation and water supply reliability. This work demonstrates the utility of RL as both an optimization tool and a method for evaluating the implicit logic of water management strategies.

1. Introduction

The management of large-scale reservoir systems is a task of critical national importance, directly impacting public safety, agricultural productivity, and ecological stability. Dam operators are charged with navigating a fundamental trade-off: releasing water to create storage capacity for flood mitigation versus retaining water to ensure supply for agricultural, municipal, and environmental needs. This decision-making process is complicated by deep hydrological uncertainty, driven by climate variability and increasingly extreme weather events [6].

Historically, federal agencies like the U.S. Army Corps of Engineers (USACE) have relied on static “rule curves” detailed in Water Control Manuals to guide operations. These heuristics dictate target reservoir levels based on the time of year, derived from long-term historical averages [1]. While robust, such static policies are inherently sub-optimal, as they cannot adapt to near-term forecasts or anomalous conditions, potentially sacrificing either safety or efficiency. In response, modern paradigms such as Risk-Informed Decision Making (RIDM) have emerged, advocating for dynamic, probabilistic approaches to operational planning [2].

While the application of Reinforcement Learning (RL) to optimize reservoir control has been demonstrated in prior work [3, 4], a significant gap remains in using RL as a tool to systematically compare the underlying philosophies of these divergent regulatory frameworks. The core logic of a traditional rule curve, a risk-informed strategy, or a purely reactive approach can be explicitly encoded within an RL agent’s reward structure and discount factor. This allows for a principled, quantitative comparison of how these philosophies perform under identical simulated conditions.

This paper introduces a comparative RL framework to analyze and quantify the performance of different dam operation strategies. We use the Folsom Dam system, a site prone to sudden atmospheric river events [6], as a representative case study, and address the following research questions:

1. How does the performance of RL-derived policies compare to a static baseline designed from USACE operational principles [7]?
2. How do distinct reward structures alter operational policies and their resulting safety-supply trade-offs?
3. Can this framework quantify the value of long-term planning?

Through our investigation we aim to provide a data-driven analysis of the strengths and weaknesses inherent in different strategic approaches to water management.

2. Related Work

The optimization of reservoir operations is supported by significant literature. Classical approaches have frequently employed mathematical programming techniques. However, traditional USACE and Bureau of Reclamation operations often still rely on static rules based on historical averages [1], despite the growing availability of advanced risk assessment frameworks [2].

More recently, the field has turned to Reinforcement Learning (RL), which is well-suited to the sequential decision-making nature of dam control [8]. Early applications demonstrated the feasibility of using tabular Q-learning on discretized state spaces to derive effective policies for single reservoirs [5]. As the field has advanced, so have the methods. Deep Reinforcement Learning (DRL) approaches, utilizing neural networks as function approximators, have enabled the use of continuous state and action spaces, capturing system dynamics with higher fidelity. For instance, Xu et. al. [3] successfully applied DRL to cascaded hydropower reservoirs considering inflow forecasts, while Tabas et. al. [4] demonstrated the efficacy of policy gradient methods in balancing competing reservoir objectives.

While this body of work has established RL as a powerful optimization tool for finding high-performing control policies, our research takes a different approach. Prior work has largely focused on identifying a single, superior policy. In contrast, our work leverages RL as a comparative analytical framework. By training agents to optimize for goals analogous to “follow the rule curve” or “react only to immediate threats,” we can directly quantify the emergent behaviors and performance trade-offs of various strategies.

3. Methodology

We frame reservoir operation as a sequential decision-making problem and evaluate alternative operational philosophies within a unified reinforcement learning (RL) framework. Our approach is to (1) formulate the control process as a finite Markov Decision Process (MDP), (2) generate a realistic, data-driven simulation environment from historical Folsom Dam records, and (3) train tabular Q-learning agents whose reward structures encode different management strategies (risk-averse, rule-curve-following, and reactive). A USACE-inspired heuristic policy serves as a real world baseline.

All agents interact with the same environment and differ only in their reward functions and discount factors. This isolates the effect of operational philosophy from other modeling choices. Policies are trained via tabular Q-learning and subsequently evaluated under a standardized risk-averse metric to ensure fair comparison. Full details of the MDP formulation, state space, transition model, and agent configurations are provided in Section 4.

4. Data

Our environment is constructed using two years of hourly operational data for Folsom Dam retrieved via the CWMS (Corps Water Management System) Python API. Using CWMS, we queried high-resolution time series for (1) reservoir storage (acre-feet), (2) reservoir inflow (cfs), (3) reservoir outflow/release (cfs), and (4) forebay elevation, yielding more than 17,000 observations per variable. The raw dataset includes occasional gaps and telemetry anomalies typical of hydrologic records; our preprocessing script handles timestamp alignment, conversion of CWMS interval identifiers, unit normalization, and forward-filling or removal of missing or nonphysical values. After cleaning, the hourly observations are aggregated to daily means to produce a consistent operational record suitable for RL training. Reservoir storage values are then discretized into ten quantile bins, while daily inflows are mapped into four categorical regimes (low, normal, high, flood-watch) to capture the skewed, atmospheric-river-driven inflow dynamics characteristic of the American River watershed.

In order to model hydrologic uncertainty in the RL environment, we compute month-dependent inflow transition probabilities directly from the historical data. These transitions, combined with the discretized level dynamics, allow each simulated episode to reflect realistic seasonal behavior without requiring explicit physical modeling. The resulting dataset was used for training and evaluating the Q-learning agents.

4.1. Environment and Setup

To analyze dam operating policies, we formulate the control problem as a finite Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. This formulation allows for the use of tabular Q-learning [8], a model-free RL algorithm that can learn an optimal action-value function, $Q^*(s, a)$, for a finite state and action space.

4.2. MDP Formulation

- **State Space ($\mathcal{S}$):** The state $s_t \in \mathcal{S}$ is a discrete representation of the system at time t. It is a tuple comprising: (1) reservoir level, discretized into 10 bins; (2) month of the year, from 0 (January) to 11 (December); and (3) inflow regime, categorized into 4 bins (low, normal, high, flood watch) based on historical daily inflow quantiles. This results in a state space of $10 \times 12 \times 4 = 480$ unique states.

- **Action Space ($\mathcal{A}$):** The agent can select one of four discrete release actions $a_t \in \mathcal{A}$: {0: Hold, 1: Low

Release, 2: Medium Release, 3: High Release}, corresponding to operational decisions found in Water Control Manuals.

- **Transition Function ($\mathcal{P}$):** The dynamics of reservoir level are modeled as a discrete, deterministic function of inflow and outflow actions. Hydrological uncertainty is captured by a stochastic transition model for the inflow regime, $P(\text{inflow}_{t+1}|\text{inflow}_t, \text{month}_t)$, estimated from historical Folsom Dam data [5]. This captures seasonal patterns without requiring explicit weather forecasting.
- **Reward Function ($\mathcal{R}$):** The reward function, $R(s, a, s')$, is the primary experimental variable used to encode different policy philosophies, detailed in the following section.

4.3. Policy Agents

We design and evaluate four distinct policy agents: one rule-based heuristic and three RL agents. Each RL agent is trained using tabular Q-learning with $\alpha = 0.1$, ϵ decaying from 1.0 to 0.05 over 4000 episodes, for a total of 5000 training episodes.

USACE Heuristic Baseline. We implement a rule-based policy derived from USACE Water Control Manual principles [9]. This heuristic encodes: (1) a seasonal pool guide with monthly target levels, lower during flood season (October to April) to reserve capacity, higher during conservation season (May to September) for water supply; (2) zone-based operations that adjust releases based on deviation from seasonal targets; (3) emergency protocols for critical pool levels; and (4) inflow-responsive logic that increases releases during high-inflow conditions.

Policy A: Standard RL (Risk-Averse). This agent uses a fixed-band reward structure designed to penalize unsafe operating conditions. The reward function assigns $+10$ for levels 3 to 7 (safe operating range), -5 for levels 1, 2, or 8 (marginal zones), -25 for level 0 (empty reservoir), and -100 for level 9 (overflow risk). This structure encodes risk aversion by heavily penalizing extreme states while rewarding operation within safe bounds. We use $\gamma = 0.99$ to encourage long-term planning.

Policy B: Rule-Curve RL (Seasonal Targets). This agent is trained with a reward based on distance from monthly storage targets, mimicking traditional rule-curve logic. The reward is $+15$ when at the seasonal target, $+8$ when within one bin, 0 when within two bins, and increasingly negative beyond. Critical penalties (-100 for overflow, -25 for empty) are preserved. We use $\gamma = 0.99$.

Table 1. Policy Performance Comparison (1000 episodes)

Policy	Mean Reward	Std Dev
Policy A (Standard RL)	3476.4	70.4
Policy B (Rule-Curve RL)	3221.6	131.3
Policy C (Reactive RL)	3641.3	18.3
USACE Heuristic	3199.0	85.3

Policy C: Reactive RL (Myopic). This agent tests the value of long-term planning by using a lower discount factor ($\gamma = 0.80$) and a simplified reward that peaks at level 5: $R = 10 - 3 \cdot |level - 5|$. Notably, level 5 sits at the center of Policy A's safe operating range (levels 3 to 7), making this a "stay in the middle" strategy. The low γ encodes a short-term focus, reacting to current conditions without seasonal awareness.

5. Experiments and Results

5.1. Experimental Setup

Training (5000 episodes, 3 min on A100) and evaluation (1000 Monte Carlo episodes) were conducted in Python. Each episode simulates 365 days. For fair comparison, all policies are evaluated using Policy A's reward function.

5.2. Training Performance

Figure 1 shows representative learning curves for the three RL agents. All agents demonstrate convergence, with Policy A and Policy C achieving stable performance by episode 2000, while Policy B (Rule-Curve) exhibits slightly more variance due to its distance-based reward structure that creates a more complex optimization landscape.

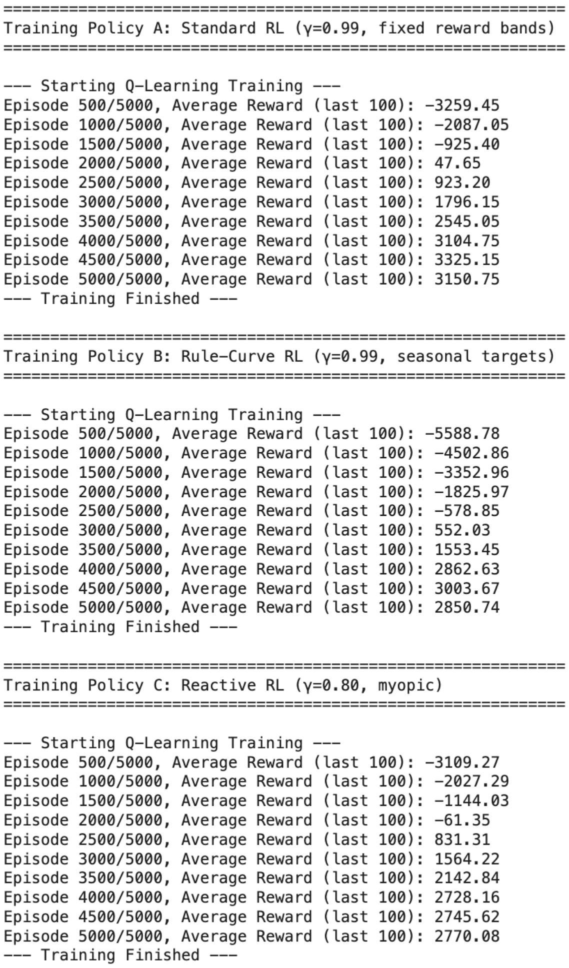
5.3. Policy Comparison

Figure 2 presents the reward distribution across all four policies evaluated over 1000 episodes. Table 1 summarizes the quantitative performance metrics.

Key Finding 1: RL outperforms the heuristic. All three RL policies exceed the USACE baseline (3199.0), demonstrating RL can improve upon expert-designed rules.

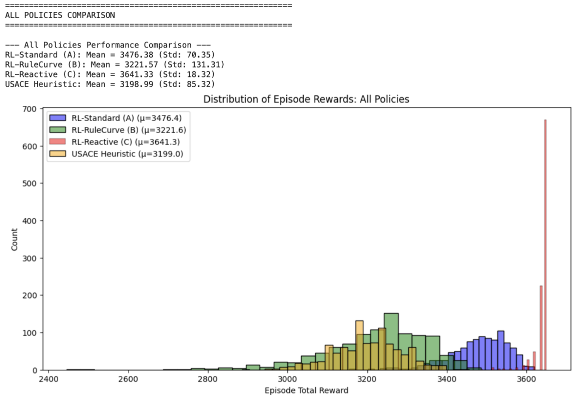
Key Finding 2: Simplicity wins. Policy C achieves the highest mean (3641.3) with lowest variance (18.3). Its "stay at level 5" objective aligns with the safe operating range (3 to 7), yielding near-optimal performance.

Key Finding 3: Rule-curve goals misalign with safety. Policy B scores lowest (3221.6), suggesting seasonal target adherence conflicts with risk-averse objectives.



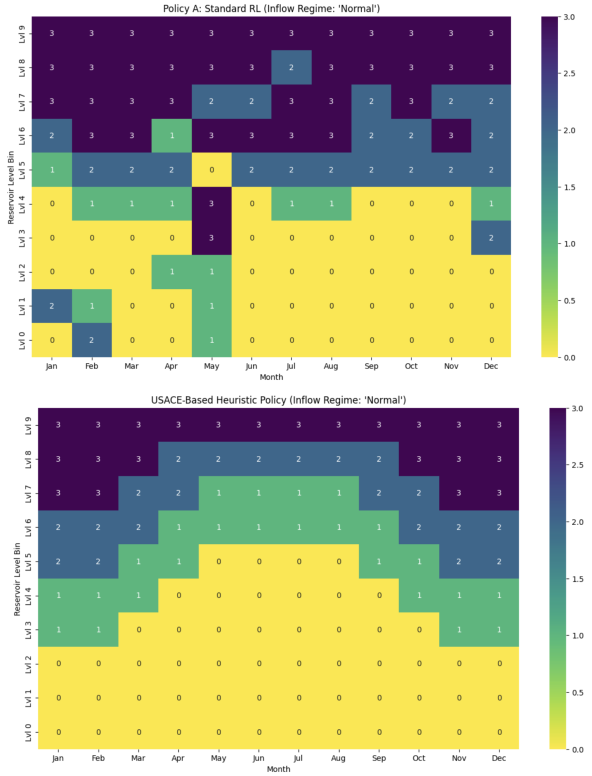
```
============================================================
Training Policy A: Standard RL (γ=0.99, fixed reward bands)
============================================================

--- Starting Q-Learning Training ---
Episode 500/5000, Average Reward (last 100): -3259.45
Episode 1000/5000, Average Reward (last 100): -2087.05
Episode 1500/5000, Average Reward (last 100): -925.40
Episode 2000/5000, Average Reward (last 100): 47.65
Episode 2500/5000, Average Reward (last 100): 923.20
Episode 3000/5000, Average Reward (last 100): 1796.15
Episode 3500/5000, Average Reward (last 100): 2545.05
Episode 4000/5000, Average Reward (last 100): 3104.75
Episode 4500/5000, Average Reward (last 100): 3325.15
Episode 5000/5000, Average Reward (last 100): 3150.75
--- Training Finished ---

============================================================
Training Policy B: Rule-Curve RL (γ=0.99, seasonal targets)
============================================================

--- Starting Q-Learning Training ---
Episode 500/5000, Average Reward (last 100): -5588.78
Episode 1000/5000, Average Reward (last 100): -4502.86
Episode 1500/5000, Average Reward (last 100): -3352.96
Episode 2000/5000, Average Reward (last 100): -1825.97
Episode 2500/5000, Average Reward (last 100): -578.85
Episode 3000/5000, Average Reward (last 100): 552.03
Episode 3500/5000, Average Reward (last 100): 1553.45
Episode 4000/5000, Average Reward (last 100): 2862.63
Episode 4500/5000, Average Reward (last 100): 3003.67
Episode 5000/5000, Average Reward (last 100): 2850.74
--- Training Finished ---

============================================================
Training Policy C: Reactive RL (γ=0.80, myopic)
============================================================

--- Starting Q-Learning Training ---
Episode 500/5000, Average Reward (last 100): -3109.27
Episode 1000/5000, Average Reward (last 100): -2027.29
Episode 1500/5000, Average Reward (last 100): -1144.03
Episode 2000/5000, Average Reward (last 100): -61.35
Episode 2500/5000, Average Reward (last 100): 831.31
Episode 3000/5000, Average Reward (last 100): 1564.22
Episode 3500/5000, Average Reward (last 100): 2142.84
Episode 4000/5000, Average Reward (last 100): 2728.16
Episode 4500/5000, Average Reward (last 100): 2745.62
Episode 5000/5000, Average Reward (last 100): 2770.08
--- Training Finished ---
```

Figure 1. Training curves showing episode reward over 5000 episodes. The red line indicates a 100-episode moving average.



```
============================================================
ALL POLICIES COMPARISON
============================================================

--- All Policies Performance Comparison ---
RL-Standard (A): Mean = 3476.38 (Std: 70.35)
RL-RuleCurve (B): Mean = 3221.57 (Std: 131.31)
RL-Reactive (C): Mean = 3641.33 (Std: 18.32)
USACE Heuristic: Mean = 3198.99 (Std: 85.32)
```

Figure 2. Distribution of episode rewards across all policies. Higher rewards indicate better performance under the standard (risk-adverse) evaluation metric.



Policy A: Standard RL (Inflow Regime: 'Normal')

Reservoir Level Bin	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Lvl 9	3	3	3	3	3	3	3	3	3	3	3	3
Lvl 8	3	3	3	3	3	3	2	3	3	3	3	3
Lvl 7	3	3	3	3	2	2	3	3	2	3	2	2
Lvl 6	2	3	3	1	3	3	3	3	2	2	3	2
Lvl 5	1	2	2	2	0	2	2	2	2	2	2	2
Lvl 4	0	1	1	1	3	0	1	1	0	0	0	1
Lvl 3	0	0	0	0	3	0	0	0	0	0	0	2
Lvl 2	0	0	0	1	1	0	0	0	0	0	0	0
Lvl 1	2	1	0	0	1	0	0	0	0	0	0	0
Lvl 0	0	2	0	0	1	0	0	0	0	0	0	0

USACE-Based Heuristic Policy (Inflow Regime: 'Normal')

Reservoir Level Bin	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Lvl 9	3	3	3	3	3	3	3	3	3	3	3	3
Lvl 8	3	3	3	2	2	2	2	2	2	3	3	3
Lvl 7	3	3	2	2	1	1	1	1	2	2	3	3
Lvl 6	2	2	2	1	1	1	1	1	1	2	2	2
Lvl 5	2	2	1	1	0	0	0	0	1	1	2	2
Lvl 4	1	1	1	0	0	0	0	0	0	1	1	1
Lvl 3	1	1	0	0	0	0	0	0	0	0	1	1
Lvl 2	0	0	0	0	0	0	0	0	0	0	0	0
Lvl 1	0	0	0	0	0	0	0	0	0	0	0	0
Lvl 0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3. Policy heatmaps under normal inflow: Policy A (top) vs USACE Heuristic (bottom). Values indicate action (0=Hold, 3=High Release).

5.4. Policy Behavior Analysis

To understand the learned behaviors, we examine policy heatmaps showing action selection across reservoir levels (y-axis) and months (x-axis) for different inflow conditions. Figure 3 compares Policy A (Standard RL) with the USACE Heuristic under normal inflow conditions.

The heatmaps reveal distinct operational philosophies. Policy A shows more aggressive releasing at high levels (actions 2 to 3 at levels 7+) regardless of season, reflecting its focus on avoiding the overflow penalty. The USACE heuristic exhibits clearer seasonal patterns, with lower releases during conservation season (May to September) to maintain water supply. Additional heatmaps for all policies under both normal and flood-watch conditions are provided in Appendix A.

5.5. Discussion

Our results reveal several important insights about dam operation strategies:

The value of simplicity. Policy C's success challenges the assumption that complex, seasonally-aware policies are necessary for effective dam management. Under the risk-adverse evaluation framework, a simple "stay in the middle" strategy proves highly effective. This has practical implications: simpler policies may be easier to implement, explain to stakeholders, and verify for safety compliance.

Misalignment between rule-curve goals and safety objectives. Policy B's underperformance highlights a potential tension between traditional rule-curve operations and modern risk-informed objectives. When evaluated on flood/drought prevention (extreme state avoidance), policies optimized for seasonal target adherence may make suboptimal trade-offs.

Variance as a risk metric. Policy C's low variance suggests consistent, predictable operations, a desirable property for critical infrastructure. In contrast, the higher variance of Policy B and the heuristic indicates more variable outcomes, which may represent operational risk even if mean performance is acceptable.

Limitations. Our model simplifies several real-world complexities: continuous state/action spaces, multi-day forecasts, downstream constraints, and multi-reservoir coordination. The discretization into 480 states, while tractable for tabular methods, may miss important nuances. Future work could address these limitations using deep RL approaches.

6. Conclusion and Future Work

This paper presented a comparative RL framework for analyzing dam operation strategies. By encoding different operational philosophies (risk-averse, rule-curve-following, and reactive) into distinct reward structures, we quantified the performance trade-offs inherent in each approach. Our key findings are:

1. All RL policies outperformed the USACE heuristic baseline, demonstrating RL's potential for improving dam operations.
2. The reactive policy ($\gamma = 0.80$) achieved the best performance with the lowest variance, suggesting that simpler strategies can be highly effective.
3. Rule-curve-optimized policies underperformed when evaluated on risk-adverse criteria, revealing potential misalignment between traditional operational goals and safety objectives.

Future work could extend this framework to continuous state spaces using Deep RL, incorporate multi-day inflow forecasts, model downstream flood constraints, and apply the methodology to multi-reservoir systems. Additionally, evaluating policies under climate change scenarios with shifted inflow distributions would provide valuable insights for long-term infrastructure planning.

7. Contributions and Acknowledgments

Eyrin Kim built the environment, MDP, and the initial functional agent and heuristic. Madison Ho designed and implemented the reinforcement learning framework, including the three policy reward structures (risk-averse, rule-curve, and reactive), and the USACE-aligned heuristic baseline. Holland Ferguson handled the acquisition, cleaning, and processing of the data, set up the data pipeline, and defined fundamental aspects of the MDP. All contributed to the planning and direction of the research and final writeup.

References

[1] U.S. Army Corps of Engineers (USACE). (2014). Safety of Dams – Policy and Procedures, ER 1110-2-1156. Washington, D.C.: U.S. Army Corps of Engineers. 1, 2

[2] Federal Energy Regulatory Commission (FERC). (2016). Risk-Informed Decision Making (RIDM) Guidelines. Division of Dam Safety and Inspections. 1, 2

[3] Xu, W., Zhang, X., Peng, A., & Liang, Y. (2020). Deep reinforcement learning for cascaded hydropower reservoirs considering inflow forecasts. *Water Resources Management*, 34, 3003-3018. 1, 2

[4] Sadeghi Tabas, S., Samadi, S., et al. (2024). Fill-and-Spill: Deep reinforcement learning policy gradient methods for reservoir operation decision and control. *Journal of Water Resources Planning and Management*, 150(7). 1, 2

[5] Castelletti, A., Galelli, S., Restelli, M., & Soncini-Sessa, R. (2010). Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 46(9), W09507. 2, 3

[6] Public Policy Institute of California (PPIC). (n.d.). How February's atmospheric rivers affected California's water supply. Accessed Dec 2025. 1

[7] U.S. Army. (2025, March 31). Managing the Cumberland River: How the corps works to reduce flood risk. www.army.mil. 1

[8] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). Cambridge, MA: The MIT Press. 2

[9] U.S. Army Corps of Engineers (USACE). (2016). Preparation of Water Control Manuals, ER 1110-2-8156. Washington, D.C.: U.S. Army Corps of Engineers. 3

Appendix A: Policy Heatmaps

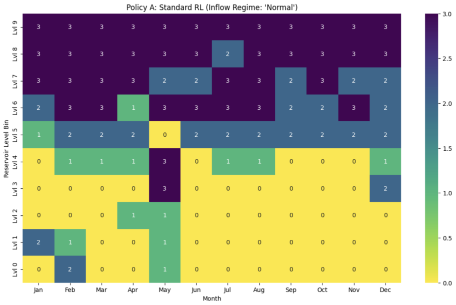


Figure 4. Policy A (Risk-Averse) under normal inflow.

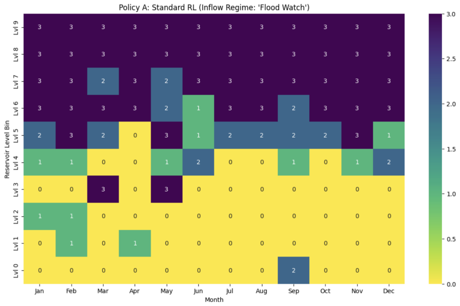


Figure 5. Policy A (Risk-Averse) under flood watch.

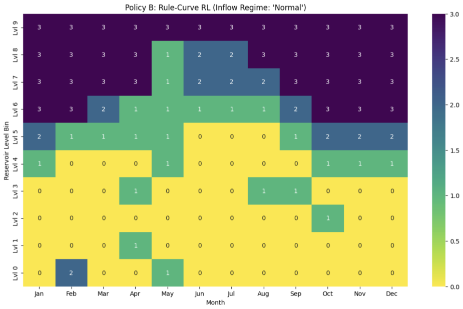


Figure 6. Policy B (Seasonal Targets) under normal inflow.

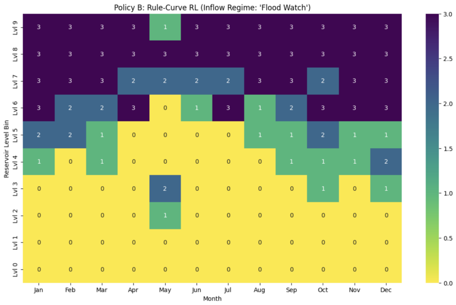


Figure 7. Policy B (Seasonal Targets) under flood watch.

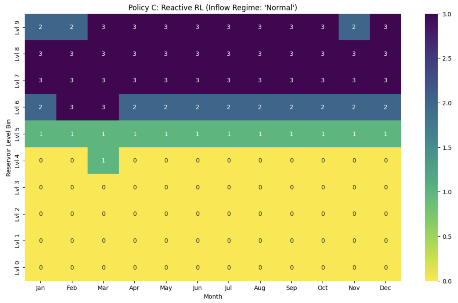


Figure 8. Policy C (Myopic) under normal inflow.

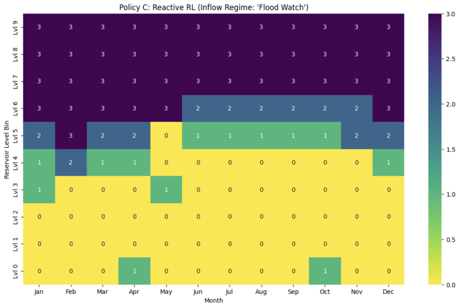


Figure 9. Policy C (Myopic) under flood watch.

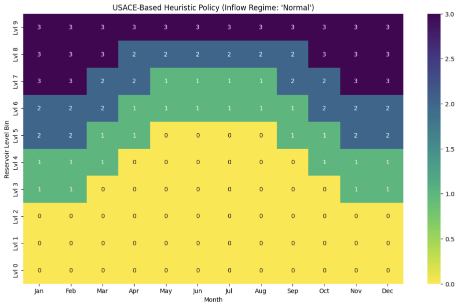


Figure 10. USACE Heuristic under normal inflow.

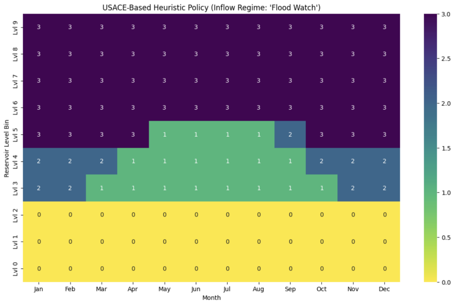


Figure 11. USACE Heuristic under flood watch.