Integrating Inverse Reinforcement Learning and Direct Policy Search for Modeling Multipurpose Water Reservoir Systems

Matteo Giuliani and Andrea Castelletti

***Abstract*—System identification and optimal control have always contributed to water resources systems planning and management. Although water control problems are commonly formulated as multi-objective Markov Decision Processes, accurately modeling reservoir systems controlled by human operators remains challenging due to the absence of a formal definition of the objective function guiding their behavior. In this letter, we introduce a mixed Reinforcement Learning approach to model the dynamics of multipurpose reservoir systems. Specifically, our method first uses Inverse Reinforcement Learning to extract the tradeoff among competing objectives from historical observations of the reservoir system dynamics. The identified objective function is then used in the formulation of an optimal control problem returning a closed-loop policy which allows the simulation of the observed dynamics of the reservoir system. We demonstrate the potential of the proposed method in a real-world application involving the multipurpose regulation of Lake Como in northern Italy. Results show that our approach effectively infers the tradeoff between flood control and water supply adopted in the observed system's operation, and yields a control policy that closely approximates the observed system dynamics.**

***Index Terms*—Inverse reinforcement learning, direct policy search, machine learning, human-in-the-loop control, control applications.**

I. Introduction

SYSTEM identification and optimal control have contributed to the design of efficient and sustainable water system operations since the 1955 Harvard Water Program [1]. Yet, they remain a very active research field [2], [3] as most water systems face several real-world challenges, making the use of optimal control tools particularly complex, as discussed in several review articles [4], [5], [6]. These challenges include the increasing variability in hydroclimatic conditions associated with climate change [7] as well as growing demands from multiple, often competing, sectors driven by population growth and socio-economic development [8].

Traditionally, water control problems are formulated as multi-objective Markov Decision Processes (MOMDP, see [9]), where the state variables represent the storage of water reservoirs and/or tanks and the water level in channels, while the control decisions include all actions for the actuators such as gates or pumps. The system then moves through a generally non-linear transition into a new state, influenced by the control decisions and the stochastic disturbances (e.g., inflows, rainfall, water demands), producing immediate rewards or costs after the transition (e.g., hydropower production, flood damages, water supply deficit).

A peculiar characteristic of reservoir systems with respect to other water systems - e.g., urban water networks - is the presence of a human operator in charge of implementing the control action [10]. Although reservoir operators might have access to a decision support system that recommends optimal operational decisions, they often prefer to make decisions largely based on their personal experience [11]. This tendency often generates a discrepancy between observational data and model-based experiments that simulate optimal control decisions. While some studies attribute this discrepancy to the non-rationality of human behaviors [12], we argue that the main challenge is the lack of a formal definition of the objective function driving the observed behavior of a human operator. For example, most existing global hydrologic models represent reservoir systems distinguishing only between irrigation and non-irrigation reservoirs, where the release of the former is estimated as a function of their corresponding water demand, while the release of the latter is their long-term mean inflow [13]. Recently, [14] showed how the accuracy of these global models improves when the reservoir dynamics is simulated using optimal control policies that are specific to the primary purpose of each reservoir. Yet, multipurpose reservoir systems, which represent a large share of existing water reservoirs (e.g., 40% of existing hydropower dams serve multiple demands), add another challenge to the modeling task as these systems also require defining a specific tradeoff determining the relative importance assigned to the different operating objectives, which is generally unknown.

In this letter, we contribute a mixed Reinforcement Learning (RL) method to identify a simulation model able to reproduce

Manuscript received 8 March 2024; revised 17 May 2024; accepted 29 May 2024. Date of publication 10 June 2024; date of current version 28 June 2024. Recommended by Senior Editor P. Tesi. *(Corresponding author: Matteo Giuliani.)*

The authors are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milan, Italy (e-mail: matteo.giuliani@polimi.it; andrea.castelletti@polimi.it).

Digital Object Identifier 10.1109/LCSYS.2024.3411831

the observed dynamics of existing multipurpose reservoir systems. Specifically, we first use Inverse Reinforcement Learning [15] to extract from the observed trajectory of reservoir levels and release the tradeoff underlying the decisions made by a human operator in the form of a weight vector balancing the competing objectives; then, we design via multi-objective Reinforcement Learning [16] an optimal policy targeting the identified tradeoff to enable the simulation of the observed reservoir dynamics. IRL falls within the broader category of Imitation Learning algorithms [17], aimed at learning from demonstrations. While IRL searches for a reward function, other methods, such as Behavioral Cloning [18], focus on generating an imitation policy obtained through supervised learning. However, policies derived from Behavioral Cloning are usually not adaptable to diverse environments. In contrast, the reward function produced by an IRL algorithm encapsulates the overarching intentions of an expert and remains applicable even amidst changes in the environment's dynamics, thus making this approach particularly attractive for modeling environmental systems that are exposed to non-stationary evolving drivers such as climate change. Additionally, this reward function can facilitate forward RL within the original environment, and the resulting solutions can be utilized in simulation models. Recently, [19] showed the IRL potential in a few real-world applications, including the identification of the tradeoff driving the multipurpose operation of a water reservoir but the use of this information in a forward RL problem has not been tested yet.

In the second phase of our analysis, we use the results of the IRL problem to formulate the objective function of an optimal control problem, whose resolution yields a closed-loop policy which allows the simulation of the observed dynamics of the reservoir system. Direct policy search (DPS, [20]) has recently emerged as one of the most popular multi-objective RL methods for solving complex MOMDP problems given its applicability to diverse tasks, scalability to high-dimensional state space, flexibility in using exogenous information via data-driven controller tuning approaches, and potential for broadening the scope and complexity of resolvable objectives [21]. The policy optimized via DPS is finally simulated to evaluate its accuracy in reproducing the observed data.

We illustrate our approach by applying it to the real-world case study of Lake Como, a multipurpose regulated lake located in Northern Italy. The water stored in the lake is primarily used downstream to irrigate four agricultural districts. The southwestern portion of Lake Como forms a dead end, posing flood risks to the city of Como. However, these flooding events can be mitigated by regulating the lake to minimize the occurrence of high water levels.

II. Methods

A. Problem Formulation

Control problems related to water reservoir systems involve making sequential decisions, denoted as u_t, regarding the volume of water to release at specific time intervals. These decisions are based on the current conditions of the system, described by the state variable x_t, which typically describes the reservoir storage. The system's state is subsequently modified through a stochastic transition function, influenced by a vector of stochastic external drivers denoted as ε_{t+1}, which may represent variables like reservoir inflows, precipitation, and evaporation losses. These systems can be effectively modeled as multi-objective Markov Decision Processes as $x_{t+1} = f_t(x_t, u_t, \varepsilon_{t+1})$ with $t = 0, 1, \ldots, h-1$, assuming the state is observable and the stochastic disturbance can be described by a probability density function (i.e., $\varepsilon_{t+1} \sim \phi_{t+1}$).

We model the human operator acting in the MOMDP using a parametric differentiable policy π_θ, where Θ is the policy parameter space. The execution of a policy π_θ in an MOMDP generates the trajectory τ, which is a sequence of state-control pairs over the time horizon $[0, H]$. This trajectory allows evaluating the expected performance of π_θ for a given cost function as

$$J(\pi_\theta) = E_\varepsilon\left[\sum_{t=0}^{H-1} g(x_t, u_t, \varepsilon_{t+1}) + G(x_H)\right] \tag{1}$$

where $g(\cdot)$ is the immediate cost function associated with the time transition from t to $t+1$ and $G(x_H)$ is a penalty function over the final state. The optimal control policy π_θ^* is then the one maximizing the expected performance $J(\pi_\theta)$, i.e.,

$$\pi_\theta^* = \arg\min_{\pi_\theta} J(\pi_\theta) \quad with \quad \theta \in \Theta \tag{2}$$

When the operations of the reservoir system influence various interests (e.g., hydropower production, water supply, environmental protection), Problem (2) is formulated including a q-dimensional objective function vector $\mathbf{J} = [J_1, \ldots, J_q]$. The multi-objective problem does not yield a single solution that minimizes all q objectives, as such a solution generally does not exist. It rather returns a set of Pareto optimal solutions $\mathcal{P}^*$.

Definition 1: Policy π dominates policy π', denoted by $\pi \prec \pi'$, if: $\forall i \in \{1, \ldots, q\}, J_i(\pi) \leq J_i(\pi') \wedge \exists i \in \{1, \ldots, q\}, J_i(\pi) < J_i(\pi')$.

Definition 2: If there is no policy π' such that $\pi' \prec \pi$, the policy π is Pareto optimal.

The traditional approach to solve such multi-objective problem is to reformulate it as a series of single-objective problems by combining the q objectives through a scalarization function $\Gamma : \mathbb{R}^q \to \mathbb{R}$, such as a convex combination of the objectives $\zeta(\pi_\theta, \omega) = \mathbf{J}(\pi_\theta)\omega$ using a weight vector $\omega \in \mathbb{R}^q_{\geq 0}$ and $\|\omega\|_1 = 1$.

B. Gradient-Based Inverse Reinforcement Learning

In the IRL problem, we assume that there exists an expert's policy π_θ^E which represents the policy describing the human operator's decisions who behaves optimally with respect to some unknown cost functions. Our goal becomes finding a scalarized cost function ζ^E such that π_θ^E is an optimal policy. When this optimality condition holds, the policy gradient $\nabla_\theta \zeta(\pi_\theta^E, \omega^E) = \nabla_\theta \mathbf{J}(\pi_\theta^E)\omega^E$ must vanish and the weight vector ω^E associated to the cost function minimized by the expert belongs to the null space of the Jacobian matrix, i.e.,

$$\begin{aligned} &if \quad \pi_\theta^E = \arg\min_{\pi_\theta} \zeta(\pi_\theta, \omega^E) \\ &then \quad \omega^E \in null(\nabla_\theta \mathbf{J}(\pi_\theta^E)) \end{aligned} \tag{3}$$

where $\nabla_\theta \mathbf{J}(\pi_\theta^E)$ is the Jacobian matrix defined as $\nabla_\theta \mathbf{J}(\pi_\theta^E) = [\nabla_\theta J_1(\pi_\theta^E)| \dots |\nabla_\theta J_q(\pi_\theta^E)]$.

The computation of the Jacobian matrix requires a parametric representation of the expert's policy that can be obtained via Behavioral Cloning. The expert policy parameters can be estimated from a dataset of observed trajectories $D = \{\tau_i\}_i^n$ using a Maximum Likelihood procedure. Then, given this policy parameterization, we can obtain an unbiased estimate of the Jacobian matrix using sample-based estimators for conventional policy gradient methods (e.g., G(PO)MDP [22]). The approximations introduced for the Jacobian estimation prevent the use of eq. (3) because the estimated Jacobian might result in full rank, although the true Jacobian has a rank smaller than q, leading to a zero-dimensional null space.

The Σ-Gradient inverse reinforcement learning (Σ-GIRL, [23]) addresses this issue by looking at the Jacobian estimate $\widehat{\nabla}_\theta \mathbf{J}(\pi_\theta)$ as a noisy version of the true Jacobian matrix, modeling it as a Gaussian random matrix $\mathcal{N}(\mathcal{M}, \frac{1}{n}\Sigma)$. Using the Gaussian likelihood model, we reformulate the IRL problem as the problem of finding the weights ω and the new Jacobian $\mathcal{M}$ that jointly maximize the likelihood of the estimated Jacobian. The IRL problem is then formulated as follows:

$$\omega^* = \underset{\substack{\omega\in\mathbb{R}^q_{\geq 0} \\ \|\omega\|_1=1}}{\arg\min} \left\| \widehat{\nabla}_\theta \mathbf{J}(\pi_\theta)\omega \right\|^2_{\left[(\omega\otimes\mathbf{I}_q)^T \Sigma (\omega\otimes\mathbf{I}_q)\right]^{-1}} \tag{4}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{I}_q$ is the identity matrix of order q. More details about Σ-GIRL are available in [23].

C. Direct Policy Search

Direct Policy Search represents an approximate dynamic programming method that diverges from the conventional Dynamic Programming policy design approach [24]. Rather than seeking the value function within the objective space, DPS directly searches for optimal control policies within an infinite-dimensional space of parameterized functions. This method offers the advantage of converting the challenge of designing functional control policies into a non-linear optimization problem. Finding an optimal parametric policy π_θ^* is equivalent to discovering the corresponding optimal policy parameters θ^*, where $\theta \in \Theta$. Various DPS strategies have emerged in the literature (refer to [20] for a comprehensive review and associated references). Their effectiveness, namely their ability to reveal high-quality solutions, is heavily based on two main factors: the choice of parameterization for the control policy [25], and the efficiency of the optimization algorithm used to explore optimal control policy parameters [26]. Specifically, we use the evolutionary multi-objective direct policy search (EMODPS, [25]) which combines DPS, nonlinear approximating networks, and multi-objective evolutionary algorithms to design Pareto-approximate closed-loop operating policies for multipurpose water reservoir systems.

In our EMODPS formulation, we use Gaussian radial basis functions (RBFs) to parameterize the control policy, as they are capable of representing policies for a large class of MOMDPs [27]. In this formulation, the control u_t is defined as:

$$u_t = \alpha + \sum_{i=1}^{A} \left[w_i \exp\left(- \sum_{j=1}^{B} \frac{((x_t)_j - c_{j,i})^2}{b_{j,i}^2} \right) \right] \tag{5}$$

where A is the number of RBFs, B the number of state variables, $\mathbf{c}_i$ and $\mathbf{b}_i$ the B-dimensional center and radius vectors of the i-th RBF, respectively, and w_i the weight of the i-th RBF which are nonnegative. The total number of policy parameters is equal to $A(2B + n_u) + 1$.

Searching the parameters of these nonlinear approximating networks entails navigating through high-dimensional spaces that map to noisy and multimodal objective function values. MOEAs offer an efficient means for solving this search. Although evolutionary strategies do not ensure optimal solutions, they present a promising alternative to gradient-based methods because evolving a set of candidate solutions based on their ranking has been shown to handle performance uncertainties more effectively than methods that rely on estimating absolute performance or performance gradients [28]. In particular, we solved Problem (2) using the Borg MOEA [29], which has been shown to be highly robust in solving multi-objective optimal control problems [26]. This algorithm incorporates epsilon-dominance archiving [30], adaptive population sizing [31], a steady-state algorithm structure [32], and multiple variation operators. The probability of selecting these operators changes adaptively during the search, reflecting recent successes in generating new non-dominated solutions. These characteristics address the challenge of tuning the algorithm parameters to match the unique fitness landscape of the problem. More details about EMODPS are available in [21], [25].

III. Real World Application

Lake Como, located in Northern Italy, is a regulated lake in the Adda River basin. The lake has a surface area of 145 km^2, an active storage capacity of 246.5 Mm^3, and its catchment covers an area of 4552 km^2 characterized by a subalpine hydrological regime featuring dry periods in winter and summer, and flow peaks in late spring and autumn fed by snowmelt and rainfall, respectively.

The regulation of Lake Como serves two primary and competing objectives: it provides flood protection to the city center of Como and it supplies downstream water to support the 1400 km^2 irrigation district and nine run-of-river power plants. In addition, local authorities recently started to consider the regulation as a means for preventing extremely low lake levels that negatively impact on navigation, tourism, and lake ecosystems. However, how much this new target has impacted the historical regulation is still unclear.

A. Model and Operating Objectives

The system is modeled as a discrete-time, periodic, nonlinear, stochastic MOMDP with the following features: a continuous state variable x_t representing the water volumes stored in the lake (m^3); a continuous control variable u_t representing the daily release decision (m^3/s); and a discrete-time,

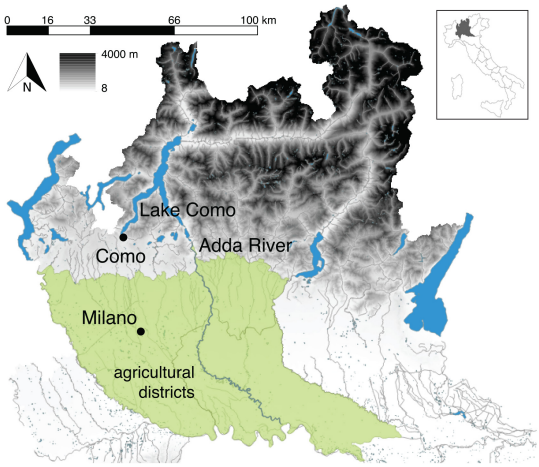

Fig. 1. Map of the Lake Como basin.

nonlinear state-transition function describing the mass balance equation of the lake water storage affected by a stochastic disturbances ε_{t+1} representing the net inflow (i.e., inflow minus evaporation losses) to the lake (m^3/s), i.e.,

$$x_{t+1} = x_t - R(x_t, u_t, \varepsilon_{t+1}) + \varepsilon_{t+1} \tag{6}$$

The adopted time step is 1 day, and the system is periodic with a period T = 365 days. The nonlinear dynamics of the system are due to the release function $R(\cdot)$, which determines the actual release from the lake as a function of the control u_t. This release generally coincides with u_t corrected, when necessary, to respect physical and legal constraints specifying the minimum and maximum volume that can be released over the time interval $[t, t+1)$ by keeping all the dam's gates completely closed and completely open, respectively. The Adda River is described by a plug-flow model, which simulates the routing of the lake releases to the intake of the irrigation canals.

The operating objectives involved in the Lake Como regulation and computed over the time horizon $[0, H]$ are the following:

- flood control: daily average excess of lake level above the flooding threshold $\bar{h} = 1.1m$:

$$J_F = \frac{1}{H}\sum_{t=0}^{H-1} \max\left(0, h_{t+1} - \bar{h}\right) \tag{7}$$

- water supply: daily average deficit between the lake release and the cyclostationary daily water demand of downstream users w_t, subject to the minimum environmental flow constraint $q_{MEF} = 22$ m^3/s to preserve adequate environmental conditions in the river downstream the abstraction point of the irrigation canals:

$$J_D = \frac{1}{H}\sum_{t=0}^{H-1} (\max(0, w_\tau - \max(0, r_{t+1} - q_{MEF}))) \tag{8}$$

- low-level control: daily average lake level scarcity compared to the low-level threshold $\underline{h} = -0.2m$:

$$J_L = -\frac{1}{H}\sum_{t=0}^{H-1} \min\left(0, h_{t+1} - \underline{h}\right) \tag{9}$$

B. Computational Experiments

Observational data on lake level, net inflow, and release were provided by the regulating authority (i.e., Consorzio dell'Adda) and are available from 1946 at a daily resolution. The calculation of the total net inflow is derived from the inversion of the mass balance eq. (6) to consider various tributaries and evaporation losses. The time period considered for the analysis is January 1st, 2000 to December 31st, 2019. During this period the lake was regulated by a single human operator.

The resolution of the IRL problem in eq. (4) returns the 3-dimensional weight vector $\omega^* = [\omega_F, \omega_D, \omega_L]$ that balances the competing objectives of flood control, water supply, and low-level control associated with the regulation of Lake Como. The RBF policies π_θ in eq. (5) are parameterized using 5 Gaussian bases that return the control decision u_t as a function of a 3-dimensional state vector $\mathbf{x}_t = [sin(2\pi t/365), cos(2\pi t/365), h_t]$, where the first two elements account for the time dependence and cyclostationarity of the system and, consequently, of the control policy. The total number of parameters of the control policy is equal to 36. The *IRL Policy* is then designed by solving Problem (2) using a scalarized objective that aggregates the vector of operating objectives $\mathbf{J} = [J_F, J_D, J_L]$ using the weights ω^* returned by Σ-GIRL.

To demonstrate the value of our RL approach, we compared the *IRL Policy* against four benchmarks:

- two state-of-the-art rules that are adopted in global hydrologic models as in [13], i.e., *HanasakiRule-I*, which considers Lake Como as an irrigation reservoir and sets the release equal to the downstream water demand, and *HanasakiRule-F*, which considers the lake as a non-irrigation reservoir and its release equal to the long-term mean inflow;
- two optimal control policies designed considering separately the Lake Como water supply (J_D *Policy*) and flood control (J_F *Policy*) objectives. These somehow maintain the same principles of the rules proposed by [13], while being optimized for the specific characteristics of the considered case study.

Since the Borg MOEA has been demonstrated to be relatively insensitive to the choice of parameters, we use the default algorithm parameterization suggested by [29]. Each optimization was run for 500,000 function evaluations. To improve solution diversity and avoid dependence on randomness, the solution set from each formulation is the result of 5 random optimization trials. In total, the EMODPS optimization runs 2.5 million simulations and requires approximately 68 hours on an Intel 440FX - 82441FX PMC [Natoma] with twp 2.0 GHz CPUs and 8 GB Ram. The final set of Pareto-optimal policies is defined as the set of non-dominated solutions from the results of all the optimization trials.

IV. Results

The human operator of Lake Como has a set of preferences that are unknown. We started our analysis by estimating the historical balance of the three competing objectives using the

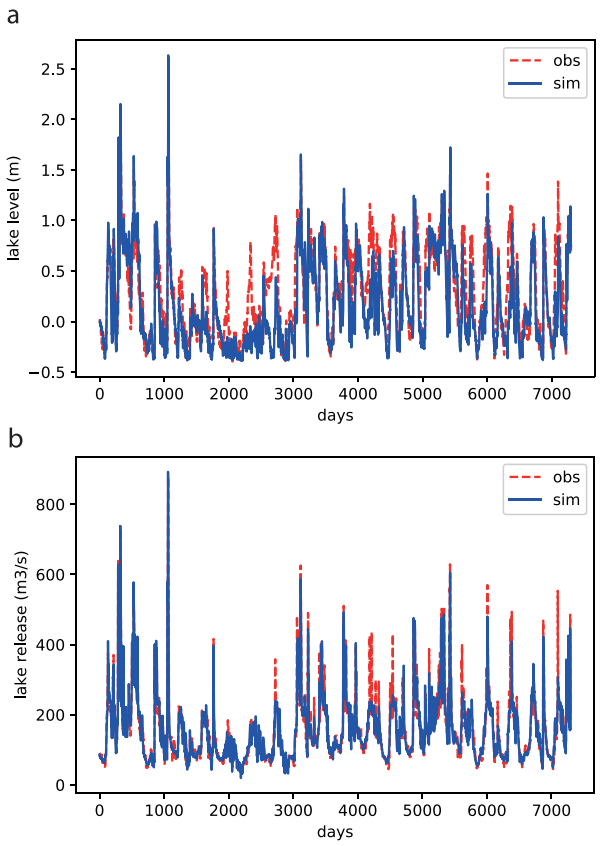

Fig. 2. Trajectories of observed and simulated lake levels (panel a) and release (panel b) under the IRL control policy.

Σ-GIRL algorithm (see Section II-B). The resulting weight vector $\omega^* = [0.84, 0.16, 0.00]$ assigns a higher weight to flood control than water supply interests, suggesting that flood control is relatively more important than water supply (or that it is easier to control floods than to minimize the water supply deficit). This result is consistent with previous studies [33] that found the performance of the historical regulation to be close to the portion of the Pareto front that favors flood control while accepting high water supply deficits. The low-level control is instead assigned a null weight and is therefore not considered in the next steps of the analysis. This result confirms how, historically, this interest was marginally influencing the lake regulation, becoming more relevant only in recent years as a consequence of severe drought events such as the summer and fall of 2022 (not included in our analysis).

In the second step of our analysis, we use the weights determined by Σ-GIRL in the design of a control policy optimized to target the identified tradeoff. The simulated lake level and release trajectories under such *IRL Policy* (Figure 2) provide a good approximation of the historical observations, capturing both seasonal patterns and many anomalous events such as the high lake levels in the first three years of the simulation horizon or the low releases during the 2005-2006 drought period (i.e., days 2190-2920). The accuracy of the *IRL Policy* in reproducing the historical dynamics of Lake Como is confirmed by considering the coefficient of determination (R^2) between the simulated and observed trajectories (Table I). Results show that the R^2 values of the *IRL Policy* outperform

TABLE I
ACCURACY OF DIFFERENT CONTROL POLICIES IN REPRODUCING THE HISTORICAL DYNAMICS OF LAKE COMO

	R^2 level	R^2 release
IRL Policy	0.59	0.90
HanasakiRule-I	0.14	0.83
HanasakiRule-F	-0.22	0.80
J_F *Policy*	0.42	0.87
J_D *Policy*	0.30	0.84

TABLE II
ACCURACY OF DIFFERENT CONTROL POLICIES IN REPRODUCING THE PERFORMANCE OF THE HISTORICAL REGULATION OF LAKE COMO

	J_F (cm)	J_D (m^3/s)
history	0.53	39.57
IRL Policy	0.59	37.69
HanasakiRule-I	0.87	37.26
HanasakiRule-F	0.86	39.14
J_F *Policy*	0.50	39.67
J_D *Policy*	0.74	34.99

all benchmarks, registering average relative improvements equal to 345% on lake levels and 10% on lake releases with respect to the two Hanasaki rules, and 68% on lake levels and 5% on lake releases when compared to the J_F and J_D *Policies*. These improvements are consistently larger in terms of the accuracy of the lake level simulation than the lake release one as reproducing the level dynamics is a more challenging modeling task, as confirmed by the lower values of R^2 than those referring to the release trajectory. This difficulty can be motivated by the direct impact of the stochastic disturbance's variability (i.e., inflow) on the level dynamics.

Lastly, we contrast the observed performance of the historical system's regulation against the simulated values of J_F and J_D objectives. Table II shows that the two Hanasaki rules and the J_D *Policy* do not accurately reproduce the historical performance, with large errors especially in terms of J_F. The simulated performance of the *IRL* and J_F *Policies* is instead closer to the historical one as both policies have similar weight vectors that assign more importance to flood control over water supply. Interestingly, the policy that most accurately simulates the historical values of the operating objectives is the J_F *Policy* albeit it is not the most accurate one in simulating the observed trajectories of lake level and release. This finding suggests the existence of solutions that attain the same (or similar) level of performance with a variety of diverse trajectories.

V. CONCLUSION

In this letter, we contribute a mixed method combining Inverse Reinforcement Learning and Direct Policy Search for modeling multipurpose water reservoir systems. The multipurpose regulation of Lake Como is used as a real-world application to demonstrate the potential of our approach. The numerical results show that the Σ-GIRL algorithm allows the inference of the tradeoff underlying the operation of the observed system, with a relative preference for flood control against the water supply that is consistent with previous research in the same case study. Moreover, the simulation of the control policy optimized targeting the identified tradeoff provides a good approximation of the observed lake dynamics.

Ongoing research activities are focused on testing the potential of the proposed methodology in other reservoir systems with different, possibly more operating objectives. Moreover, we will explore the scalability of the method to multi-reservoir systems, where the unknown objective function can differ across operators of the same system depending on the implemented level of coordination.

ACKNOWLEDGMENT

The authors would like to thank Consorzio dell'Adda for providing the data used in this letter.

REFERENCES

[1] A. Maass, M. Hufschmidt, R. Dorfman, H. Thomas Jr., S. Marglin, and G. Fair, *Design of Water–Resource Systems: New Techniques for Relating Economic Objectives, Engineering Analysis, and Governmental Planning*. Cambridge, MA, USA: Harvard Univ. Press, 1962.

[2] M. Zaniolo, M. Giuliani, and A. Castelletti, "Neuro-evolutionary direct policy search for multiobjective optimal control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5926–5938, Oct. 2022.

[3] P. Segovia, V. Puig, and E. Duviella, "A multilayer control strategy for the calais canal," *IEEE Trans. Control Syst. Technol.*, vol. 32, no. 2, pp. 311–325, Mar. 2024.

[4] A. Castelletti, F. Pianosi, and R. Soncini-Sessa, "Water reservoir control under economic, social and environmental constraints," *Automatica*, vol. 44, no. 6, pp. 1595–1607, 2008.

[5] L. García, J. Barreiro-Gomez, E. Escobar, D. Téllez, N. Quijano, and C. Ocampo-Martínez, "Modeling and real-time control of urban drainage systems: A review," *Adv. Water Resources*, vol. 85, pp. 120–132, Nov. 2015.

[6] A. Castelletti et al., "Model predictive control of water resources systems: A review and research agenda," *Annu. Rev. Control*, vol. 55, pp. 442–465, Apr. 2023.

[7] S. Stevenson et al., "Twenty-first century hydroclimate: A continually changing baseline, with more frequent extremes," *Proc. Nat. Acad. Sci.*, vol. 119, no. 12, 2022, Art. no. e2108124119.

[8] D. Billington and D. Jackson, *Big Dams of the New Deal Era: A Confluence of Engineering and Politics*. Norman, OK, USA: Univ. Oklahoma Press, 2017.

[9] D. J. White, "Multi-objective infinite-horizon discounted Markov decision processes," *J. Math. Anal. Optim.*, vol. 89, no. 2, pp. 639–647, 1982.

[10] M. Giuliani, J. Lamontagne, P. Reed, and A. Castelletti, "A state-of-the-art review of optimal reservoir control for managing conflicting demands in a changing world," *Water Resources Res.*, vol. 57, Dec. 2021, Art. no. e2021WR029927.

[11] B. Dobson, T. Wagener, and F. Pianosi, "An argument-driven classification and comparison of reservoir operation optimization methods," *Adv. Water Resources*, vol. 128, pp. 74–86, Jun. 2019.

[12] J. Yoon et al., "A typology for characterizing human action in multisector dynamics models," *Earth's Future*, vol. 10, no. 8, 2022, Art. no. e2021EF002641.

[13] N. Hanasaki, S. Kanae, and T. Oki, "A reservoir operation scheme for global river routing models," *J. Hydrol.*, vol. 327, nos. 1–2, pp. 22–41, 2006.

[14] G. W. Abeshu et al., "Enhancing the representation of water management in global hydrological models," *Geoscientific Model Develop. Discuss.*, vol. 16, pp. 5449–5472, Feb. 2023. [Online]. Available: https://doi.org/10.5194/gmd-16-5449-2023

[15] S. Adams, T. Cody, and P. A. Beling, "A survey of inverse reinforcement learning," *Artif. Intell. Rev.*, vol. 55, no. 6, pp. 4307–4346, 2022.

[16] C. F. Hayes et al., "A practical guide to multi-objective reinforcement learning and planning," *Auton. Agents Multi-Agent Syst.*, vol. 36, no. 1, p. 26, 2022.

[17] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *Found. Trends® Robot.*, vol. 7, nos. 1–2, pp. 1–179, 2018.

[18] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[19] A. Likmeta, A. Metelli, G. Ramponi, A. Tirinzoni, M. Giuliani, and M. Restelli, "Dealing with multiple experts and non-stationarity in inverse reinforcement learning: An application to real-life problems," *Mach. Learn.*, vol. 110, pp. 2541–2576, Sep. 2021.

[20] M. Deisenroth, G. Neumann, and J. Peters, "A survey on policy search for robotics," *Found. Trends Robot.*, vol. 2, pp. 1–142, Aug. 2013.

[21] M. Giuliani, J. D. Quinn, J. D. Herman, A. Castelletti, and P. M. Reed, "Scalable multiobjective control for large-scale water resources systems under uncertainty," *IEEE Trans. Control Syst. Technol.*, vol. 26, no. 4, pp. 1492–1499, Jul. 2018.

[22] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *J. Artif. Intell. Res.*, vol. 15, pp. 319–350, Nov. 2001.

[23] G. Ramponi, A. Likmeta, A. M. Metelli, A. Tirinzoni, and M. Restelli, "Truly batch model-free inverse reinforcement learning about multiple intentions," in *Proc. 23nd Int. Conf. Artif. Intell. Stat.*, 2020, pp. 2359–2369.

[24] R. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[25] M. Giuliani, A. Castelletti, F. Pianosi, E. Mason, and P. Reed, "Curses, tradeoffs, and scalable management: Advancing evolutionary multiobjective direct policy search to improve water reservoir operations," *J. Water Resources Plan. Manag.*, vol. 142, no. 2, 2016, Art. no. 4015050.

[26] J. Zatarain-Salazar, P. Reed, J. Herman, M. Giuliani, and A. C. lletti, "A diagnostic assessment of evolutionary algorithms for multi-objective surface water reservoir control," *Adv. Water Resources*, vol. 92, pp. 172–185, Jun. 2016.

[27] L. Busoniu, D. Ernst, B. De Schutter, and R. Babuska, "Cross–entropy optimization of control policies with adaptive basis functions," *IEEE Trans. Syst., Man Cybern. Part B, Cybern.*, vol. 41, no. 1, pp. 196–209, Feb. 2011.

[28] R. Busa-Fekete, B. Szörényi, P. Weng, W. Cheng, and E. Hüllermeier, "Preference-based reinforcement learning: Evolutionary direct policy search using a preference-based racing algorithm," *Mach. Learn.*, vol. 97, no. 3, pp. 327–351, 2014.

[29] D. Hadka and P. Reed, "Borg: An auto–adaptive many–objective evolutionary computing framework," *Evol. Comput.*, vol. 21, no. 2, pp. 231–259, May 2013.

[30] M. Laumanns, L. Thiele, K. Deb, and E. Zitzler, "Combining convergence and diversity in evolutionary multiobjective optimization," *Evol. Comput.*, vol. 10, no. 3, pp. 263–282, Sep. 2002.

[31] J. B. Kollat and P. M. Reed, "A computational scaling analysis of multiobjective evolutionary algorithms in long-term groundwater monitoring applications," *Adv. Water Resources*, vol. 30, no. 3, pp. 335–353, 2007.

[32] K. Deb, M. Mohan, and S. Mishra, "Evaluating the epislon-domination based multiobjective evolutionary algorithm for a quick computation of Pareto-optimal solutions," *Evol. Comput. J.*, vol. 13, no. 4, pp. 501–525, 2005.

[33] M. Giuliani, Y. Li, A. Castelletti, and C. Gandolfi, "A coupled human-natural systems analysis of irrigated agriculture under changing climate," *Water Resources Res.*, vol. 52, no. 9, pp. 6928–6947, 2016.

Open Access funding provided by 'Politecnico di Milano' within the CRUI CARE Agreement