Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 50-1 (2017) 3183–3188

Scenario-based fitted Q-iteration for adaptive control of water reservoir systems under uncertainty

Federica Bertoni * **Matteo Giuliani** * **Andrea Castelletti** *,**

* *Department of Electronics, Information, and Bioengineering,*
Politecnico di Milano, Milan, Italy (e-mail: name.surname@polimi.it)
** *Institute of Environmental Engineering, ETH, Zurich, Switzerland.*

Abstract: This study presents a novel approach called scenario-based Fitted Q-Iteration (sFQI) for controlling water reservoir systems under climate uncertainty. In these problems, robust control frameworks, based on worst-case realization, are usually adopted. Yet, these might be overly conservative. In this paper, we use sFQI to design adaptive control policies by enlarging the state space to include the space of the uncertain system's parameters. This allows obtaining a control policy for any scenario in the uncertainty set with a single learning process. The method is demonstrated on a simplified model of the Lake Como system, a regulated lake operated for ensuring reliable water supply to downstream users. Numerical results show that the sFQI algorithm successfully identifies adaptive solutions to operate the system under different inflow scenarios, which outperform the control policy designed under historical conditions. Moreover, the sFQI policy generalizes over inflow scenarios not directly experienced during the policy design, thus alleviating the risk of mis-adaptation, namely the design of a solution fully adapted to a scenario that is different from the one that will actually realize.

Keywords: modelling and control under change; adaptive control; reinforcement learning; climate change; environmental engineering

1. INTRODUCTION

The use of mathematical models to support planning and management of water resources systems is rapidly expanding over recent years due to advances in scientific knowledge of the natural processes, efficiency of the optimization techniques, and availability of computational resources (Washington et al., 2009). Most of the operation problems involving water resources systems with mid-term and long-term control objectives can be formulated as Markov decision processes (MDPs, see White (1982)) and solved via Dynamic Programming or Reinforcement Learning (Powell, 2007; Busoniu et al., 2010). For example, water reservoir operations are a sequence of release decisions, made at discrete time instants, over a system affected by stochastic disturbances (i.e., inflows). Yet, the optimal operations of these systems represent a wide and challenging application domain for optimal control methodologies due to non-linearities in the model and the objective functions, high dimensional state-control space, and strong uncertainties in the inputs due to the variability in the hydrological regimes (Castelletti et al., 2008).

This increasing uncertainty associated to the hydrological regimes, further enhanced by the impacts of climate change, violates the stationarity assumption generally used for describing the inflow processes, where the statistical characteristics of future inflows are considered equivalent to those observed in the historical data. This assumption of a stationary climate is unlikely to be valid in the future (Milly et al., 2008). In this case, the probability density function used for modeling the stochastic disturbances becomes a deeply uncertain parameter (Kwakkel et al., 2016), which cannot be described via any stochastic model, but rather in a deterministic and set-membership based fashion (Giuliani and Castelletti, 2016). These problems call for robust and adaptive solutions able to withstand deviations from the conditions for which they were designed (Herman et al., 2015). Many studies adopt a robust MDP framework (e.g., Nilim and El Ghaoui, 2005; Iyengar, 2005) where, assuming the uncertain parameters (i.e., inflow PDF) fall within a given uncertainty set, they search a control policy that performs the best under the worst realization of the parameters (Ben-Tal et al., 2009). These robust solutions, however, can be overly conservative since they are based on worst-case realizations (Lim et al., 2013).

In this paper, we contribute a novel method for designing optimal, adaptive policies for controlling water reservoir systems under climate-related uncertainty. In particular, we propose an extension of the Fitted Q-Iteration (FQI) algorithm, a batch-mode reinforcement learning (RL) algorithm that combines RL concepts of off-line learning and functional approximation of the value function (Ernst et al., 2005; Castelletti et al., 2010). The proposed method, called scenario-based FQI (sFQI), extends the continuous approximation of the action-value function, originally performed by FQI over the state-control space, to the space of the uncertain parameters. As a result, sFQI embeds the

★ The work has been partially funded by the DAFNE project funded by the European Commission under the Horizon 2020 framework programme (grant n. 690268).

2405-8963 © 2017, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2017.08.340

set-membership uncertainty of the future inflow scenarios in the action-value function and is able to approximate, with a single learning process, the optimal control policy associated to any parameter (i.e., inflow scenario) included in the uncertainty set.

The proposed method is demonstrated on a simplified model of the Lake Como (Italy), a regulated lake in Northern Italy mainly operated for ensuring reliable water supply to downstream users (Giuliani et al., 2016c). Beside the historical inflow scenario, we consider four alternative, uncertain scenarios of inflows, which approximate the uncertain impacts of climate change (IPCC, 2013). We first test the impacts of the inflows' scenarios on the performance of the control policy designed over historical conditions. Then, we analyze the potential of the proposed sFQI algorithm for designing solutions that perform well under all the scenarios. The performance of the sFQI policies will be finally contrasted with the upper bound of system performance, represented by *fully adapted* policies (i.e., solutions evaluated over the same inflow scenario used in the policy design), as well as with the performance obtained by *mis-adapted* policies (i.e., solutions evaluated over an inflow scenario that is different from the one used in the policy design).

The paper is organized as follows. In the next section, the methodological aspects of the proposed approach are presented, while Section 3 provides a short description of the case study application. Numerical results are reported in Section 4, while final remarks and issues for further research are discussed in the last section.

2. METHODOLOGY

2.1 Markov Decision Processes

Water reservoir operation problems generally require to take sequential decisions u_t at discrete time instants ($t = 1, 2, \dots$) on the basis of the current system conditions described by the state vector x_t (e.g., reservoir storage). The control decisions are determined by a feedback control policy $u_t = \pi(x_t)$ and alter the state of the system according to a transition function $P_w(x_{t+1}|x_t, u_t)$ affected by a vector of stochastic external drivers $q_{t+1} \sim \phi_t(w)$ (e.g., reservoir inflows) described by a probability density function, which depends on the considered climate scenario $w \in \Xi$. Such system can be modeled as a discrete-time, non-linear, stochastic Markov Decision Process defined as a tuple

$$< \mathcal{X}, \mathcal{U}, \mathcal{P}_w, \mathcal{R}, \gamma >$$

where $\mathcal{X} \subset \mathbb{R}^{n_x}$ is the continuous state space, $\mathcal{U} \subset \mathbb{R}^{n_u}$ is the continuous control space, $\mathcal{P}_w(x_{t+1}|x_t, u_t)$ is the transition model defining the transition density between state x_t and x_{t+1} under control u_t for a specific scenario $w \in \Xi$, $\mathcal{R}(x_t, u_t, x_{t+1})$ is a reward function that specifies the instantaneous reward when state x_{t+1} is reached from state x_t by taking action u_t, $\gamma \in [0, 1]$ is a discount factor. The policy is characterized by a density distribution $\pi(u_t|x_t)$ that specifies the probability of taking action u_t in state x_t (Castelletti et al., 2011).

Solving an MDP means to find a policy that maximizes the action-value function in each state:

$$\pi^* = \arg \max_{u_t \in \mathcal{U}} Q^*(x_t, u_t) \tag{1}$$

where the optimal action-value function Q^* is the solution of the following equation

$$Q^*(x_t, u_t) = \int_{\mathcal{X}} [\mathcal{R}(x_t, u_t, x_{t+1}) + \gamma \max_{u_{t+1} \in \mathcal{U}} Q^*(x_{t+1}, u_{t+1})] P_w(dx_{t+1}|x_t, u_t) \tag{2}$$

2.2 Fitted Q-iteration

Fitted Q-iteration (FQI) is a batch-mode reinforcement learning (RL) algorithm that estimates an approximation of $Q^*(x_t, u_t)$ from experience samples, collected either from the system observations or via model simulations, which constitute a finite sample dataset $\mathcal{F}$ defined as

$$\mathcal{F} = \{< x_t^i, u_t^i, x_{t+1}^i, r_{t+1}^i >, \quad i = 1, \dots, N\} \tag{3}$$

where $r_{t+1} = \mathcal{R}(x_t, u_t, x_{t+1})$ and N is the cardinality of $\mathcal{F}$. Each tuple is a sample of the one-step transition dynamics of the system, regardless the way it is generated, whether from one single trajectory of the system (e.g., the historical one) or from several, independently generated, one-step or multi-step simulations of the system dynamics.

In particular, the fitted Q-iteration (FQI) algorithm proposed by Ernst et al. (2005) and derived from fitted values iteration works (Ormoneit and Sen, 2002), combines the RL idea of learning from experience with the concept of continuous approximation of the value function (Gordon, 1995). The advantage of this approach is twofold: first, a continuous mapping of the state-action pair into the action-value function should permit the same level of accuracy as a look-up table representation based on an extremely dense grid, but using a significantly coarser grid; second, the learning process is performed offline, without the need for a direct interaction with the real system. FQI has been recently applied in different research fields, such as robotics (Bonarini et al., 2008; Riedmiller et al., 2009), control theory application for active suspensions (Tognetti et al., 2009) and energy systems (Ernst et al., 2009), biology and medicine (Pineau et al., 2009; Zhao et al., 2011), environmental management (Castelletti et al., 2010; Pianosi et al., 2013; Giuliani et al., 2014).

Given the dataset $\mathcal{F}$, FQI reformulates Problem (1) as a sequence of regression problems, producing a sequence of $\hat{Q}_h$-functions which approximate the optimal action-value function $Q^*(x_t, u_t)$ defined in eq. (2) by iteratively extending the optimization horizon h. In the first iteration ($h = 1$), the algorithm produces an approximation of the expected immediate reward $\hat{Q}_1(x_t, u_t) = \mathrm{E}[\mathcal{R}(x_t, u_t, x_{t+1})]$ for each tuple, i.e. it performs a regression as $(x_t^i, u_t^i) \rightarrow r_{t+1}^i$. Based on this approximation, the second iteration extends the optimization horizon h by estimating the function $\hat{Q}_2(x_t, u_t)$ through a regression performed on the following training set:

$$\mathcal{TS}_2 = \left\{ \left[(x_t^i, u_t^i) \rightarrow r_{t+1}^i + \max_{u_{t+1}} \hat{Q}_1(x_{t+1}^i, u_{t+1}) \right] \right\} \tag{4}$$

By proceeding in the same way, at the h-th iteration it is possible to compute an approximation of the optimal action-value function Q_h^* at horizon h (Castelletti et al., 2011). The procedure iterates until the Q-function converges or a maximum number of iterations is reached (see

Algorithm 1 sFQI Algorithm

```
**Inputs**: a learning set of tuples $\mathcal{F}_w$ and a regression algorithm
**Initialization**:
Set $h = 0$
Set $\hat{Q}_0(\cdot) = 0$ over the whole enlarged state-action space $(X \times \Xi) \times U$
**Iterations**: repeat until stopping conditions are met
- $h \leftarrow h + 1$
- build the training set
$\mathcal{TS} = \{(IN^i, OUT^i), i = 1, \ldots, N_w\}$
where
$IN^i = (x_t^i, w^i, u_t^i)$
$OUT^i = r_{t+1}^i + \gamma \max_{u_{t+1} \in U} \hat{Q}_{h-1}(x_{t+1}^i, w^i, u_{t+1}^i)$
- Run the regression algorithm on $\mathcal{TS}$ to get $\hat{Q}_h(\cdot)$
```

Ernst et al. (2005) for a discussion about the stopping condition and the convergence properties of the algorithm).

2.3 Scenario-based FQI

In this paper, we propose to extend the FQI algorithm to solve MDPs under set-membership uncertainty. The idea of the scenario-based fitted Q-iteration (sFQI) is to enlarge the state space $\mathcal{X}$ to include the uncertain parameter $w \in \Xi$ to account for the uncertainty in the disturbance vector's PDF. The original sample dataset $\mathcal{F}$ is hence modified to construct a new dataset $\mathcal{F}_w$, defined as:

$$\mathcal{F}_w = \{< x_t^i, w^i, u_t^i, x_{t+1}^i, w^i, r_{t+1}^i >, \quad i = 1, \ldots, N_w\} \tag{5}$$

where $N_w = N \cdot n_w$ is the number of tuples in the sFQI dataset, which is larger than N because sFQI operates in an enlarged state space, and n_w is the number of uncertain parameters w sampled in the algorithm learning set. Given $\mathcal{F}_w$, sFQI estimates an approximation of the optimal action-value function $Q^*(x_t, w, u_t)$ that generalizes over the uncertain parameter space Ξ. A tabular version of the algorithm is given in Algorithm 1.

As a consequence, sFQI allows learning a continuous approximation over Ξ of the optimal action-value function, and hence deriving an adaptive control policy π^* conditioned on the uncertain scenarios $w \in \Xi$. Note that the continuous approximation over Ξ allows the sFQI policy to deal with scenarios $w' \in \Xi$ not directly experienced during the training process. Indeed, like the continuous FQI mapping of state-action pairs into $Q^*(x_t, u_t)$ allows attaining the same level of accuracy as a look-up table representation based on an extremely dense grid, but using a definitely coarser grid for the state-action space, the continuous sFQI mapping of state-scenario-action into $Q^*(x_t, w, u_t)$ can be performed on a limited number of scenarios w, while ensuring the same level of accuracy in scenarios $w' \neq w$.

3. CASE STUDY DESCRIPTION

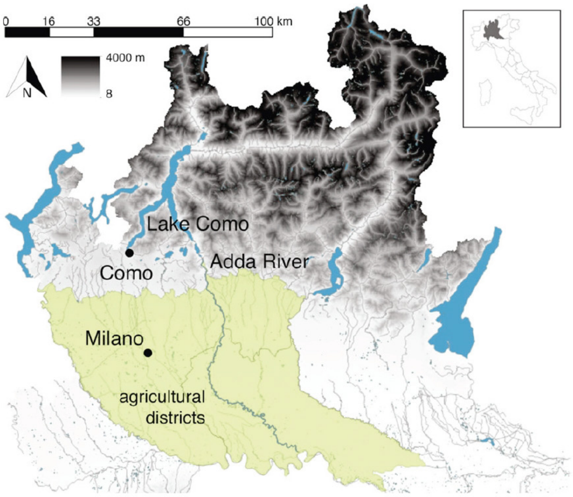

Fig. 1. Map of the Lake Como system.

The proposed scenario-based FQI approach is demonstrated on a simplified model of the Lake Como system, a regulated lake in Northern Italy (Figure 1). Lake Como is characterized by an active storage capacity of 254 Mm^3 and a mean inflow rate of 160 m^3/s. It is fed by a 4,552 km^2 catchment, which is located at an average elevation of 1500 m in the central part of the Alps and is characterized by the typical mixed snow-rain Alpine hydrological regime (Denaro et al., 2017). Lake Como inflow and effluent is the Adda River, a tributary of the Po River and the fourth longest Italian river. The lake release feeds eight run-of-river hydroelectric power plants and supports four agricultural districts with a total surface of 1,400 km^2. Major crops are cereals, especially maize, along with temporary grasslands for livestock (Giuliani et al., 2016b).

The Lake Como system is modeled as a discrete-time, stationary, non-linear, stochastic MDP. The state variable x_t is the reservoir storage and the control u_t is the release decision. The system is affected by a stochastic disturbance q_{t+1} representing the inflow to the reservoir in the time interval $[t, t+1)$. In the adopted notation, the time subscript of a variable represents the instant at which its value is deterministically known. The state x_t is observed at time t, whereas the disturbances vector has subscript $t+1$ denoting the realization of the stochastic process in the time interval $[t, t+1)$. The transition function is defined by the mass-balance equation of the lake storage, i.e.

$$x_{t+1} = x_t + q_{t+1} - q_{t+1}^R \tag{6}$$

where the release q_{t+1}^R depends on the control u_t constrained within a certain zone of operation discretion by the maximum and minimum feasible release functions, which mathematically embody some physical and normative constraints according to the current water level of the reservoir.

The daily operations of the lake aims at minimizing the vulnerability of the water supply to the downstream users. According to previous works (Giuliani et al., 2016b), this cost function $r_{t+1}(\cdot)$ is formulated as the quadratic daily deficit with respect to a fixed water demand downstream d=370 m^3/s, i.e.

$$r_{t+1} = \left(\max\left((d - q_{t+1}^R), 0\right)\right)^2 \tag{7}$$

where the quadratic formulation penalizes severe deficits in a single time step, while allowing for more frequent, small shortages (Hashimoto et al., 1982).

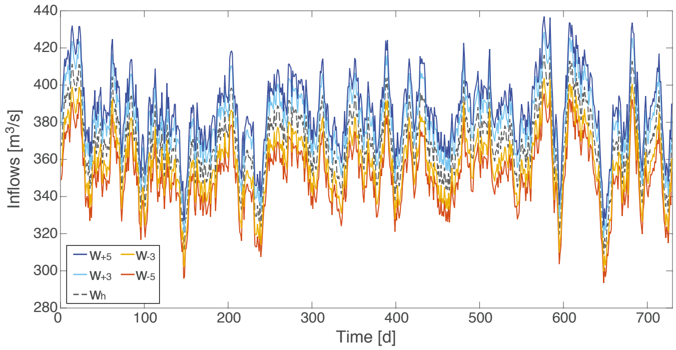

Fig. 2. Comparison of the historical inflow trajectory with the four perturbed scenarios.

4. APPLICATION RESULTS

4.1 Experiment Setting

In this work, we consider four scenarios of inflows, approximating the impacts of climate change. These scenarios, displayed in Figure 2, present daily inflow values for 730 days and are obtained via perturbation of the historical inflow w_h by reducing or increasing the historical average by 3% and 5% (Culley et al., 2016).

To demonstrate the potential of the proposed scenario-based FQI in controlling water reservoir systems under uncertainty, we performed the following experiments:

- *Historical policy* (π_h): the optimal control policy is determined via FQI using a sample dataset $\mathcal{F}$ generated through 100 model simulations with pseudo-random controls and historical inflows. The performance of this policy will be then evaluated against both the historical inflows (i.e., the same used in the policy design) and the four perturbed scenarios of inflows.
- *Fully Adapted policy* (π_{fa}): the optimal control policy is determined via FQI using a sample dataset $\mathcal{F}$ generated through 100 model simulations with pseudo-random controls and one perturbed scenario of inflows. This policy represents the upper-bound of the system performance under climate change, assuming the system operator recognizes the change and optimally re-operates the system based on the new hydrological conditions. In particular, we consider two fully adapted policies, namely p_{fa-5} and p_{fa+5}, which are designed under scenarios w_{-5} and w_{+5}, respectively.
- *sFQI policy* (π_{sFQI}): the optimal control policy is determined via sFQI using a sample dataset $\mathcal{F}_w$ generated through 100 model simulations with pseudo-random controls and three scenarios of inflows, including w_h, w_{+5}, and w_{-5}.

We tested several parameter settings for FQI and sFQI and stable solutions have been reached when the parameters of both algorithms are set as follows:

- number of trees $M = 300$;
- number of random cut directions $K = 2$ for FQI and $K = 3$ for sFQI;
- minimum number of points per leaf $n_{min} = 25$;
- number of iterations for approximating the Q-function $h = 40$.

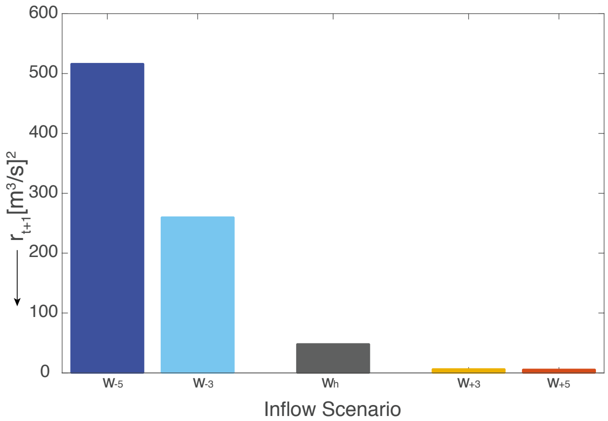

Fig. 3. Performance of policy π_h evaluated over different inflow scenarios.

4.2 Numerical results

Figure 3 illustrates the impacts of the different inflow scenarios on the performance of the historical policy π_h. When evaluated over the historical inflow scenario w_h, π_h attains a performance equal to 48.15 $[\mathrm{m}^3/\mathrm{s}]^2$, corresponding to a daily volume of 5.99e+05 m^3. This latter degrades to 259.94 $[\mathrm{m}^3/\mathrm{s}]^2$ (1.39e+06 m^3/d) and to 516.32 $[\mathrm{m}^3/\mathrm{s}]^2$ (1.96e+06 m^3/d) if the same policy is evaluated with no adaptation under scenarios w_{-3} and w_{-5}, respectively. On the contrary, the performance of π_h is equal to 6.32 $[\mathrm{m}^3/\mathrm{s}]^2$ (2.17 m^3/d) and 5.61 $[\mathrm{m}^3/\mathrm{s}]^2$ (2.05e+05 m^3/d) when evaluated under scenarios w_{+3} and w_{+5}, respectively. The same trends apply also to the performance of the fully adapted and the sFQI policies, as reported in Table 1: decreasing inflows always induce a worsening of the performance, while increasing inflows reduces the water supply deficit. These results demonstrate that the control policy performance is sensitive to the climate scenario that will realize. Moreover, the uncertainty in the scenarios is transferred and amplified when evaluated in terms of policy performance (Giuliani et al., 2016a).

Table 1. Performance of the four considered policies (rows) evaluated over the four perturbed inflow scenarios (columns).

	w_{-5}	w_{-3}	w_{+3}	w_{+5}
π_h	516.32	259.94	6.32	5.61
π_{fa-5}	500.91	256.05	9.58	7.98
π_{fa+5}	514.31	258.29	6.32	5.60
π_{sFQI}	505.45	249.50	6.32	5.60

The results in Table 1 suggest that the most challenging scenario is w_{-5} (first column). In fact, all the policies attain their worst performance under this scenario. However, despite the challenging conditions, adapting the operations of the lake allows a partial mitigation of these adverse climate change impacts. In fact, the performance of π_h is equal to 516 $[\mathrm{m}^3/\mathrm{s}]^2$, while policy π_{fa-5}, which is fully adapted to this scenario, reduces the water deficit by 3% (Figure 4). It is worth noting that such improvement requires adapting to the actual scenario that will realize in the future. Since the inflow scenarios are uncertain, such adaptation might be difficult. We therefore need to consider the risk of mis-adaptation, namely the design of a solution fully adapted to an inflow scenario that is different

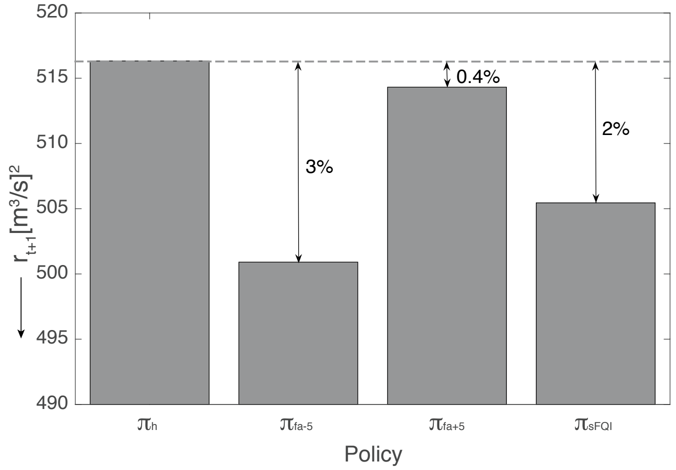

Fig. 4. Performance of the four policies in Table 1 evaluated over scenario w_{-5}.

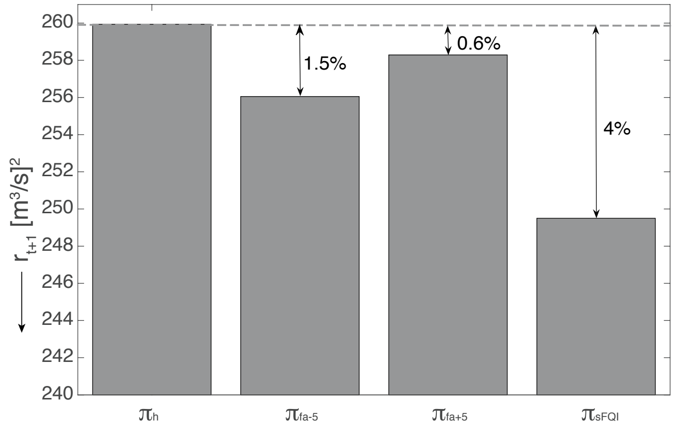

Fig. 5. Performance of the four policies in Table 1 evaluated over scenario w_{-3}.

from the one that will actually realize. Figure 4 shows that the mis-adapted policy π_{fa+5}, designed assuming scenario w_{+5} and evaluated under scenario w_{-5} is almost equivalent to the historical policy. The performance of π_{sFQI}, instead, improves the historical one by 2%. Moreover, without assuming the realization of scenario w_{-5}, it attains a performance that is close to the upper-bound represented by π_{fa-5}. Beside comparing the performance of the different policies under scenario w_{-5}, it is interesting to focus also on scenario w_{-3} (see Table 1, second column). This latter represents a scenario with the same overall trend than w_{-5} (i.e., a reduction of inflows) which is not used in the design of any policy reported in Table 1. Scenario w_{-3} hence allows testing the effectiveness of sFQI in learning a continuous approximation over the entire space of scenarios. Figure 5 shows that, again, π_h and π_{fa+5} attain the worst performance (i.e., highest water deficit). Under this scenario, the performance improvement obtained with policy π_{fa-5} is reduced by 1.5%. The solution designed via sFQI, instead, is able to generalize over this scenario and produces a 4% improvement in the control policy performance. This feature of sFQI makes it extremely promising for handling the deep uncertainty associated to the future climate.

Finally, the performance of all the policies improve under increasing inflows (i.e., w_{+3} and w_{+5}) with respect to their evaluation under the historical inflows (see Table 1, third and fourth columns). This is due to the increased water availability in the system, which simplifies the operations of the lake and allows reducing the water deficit. It is still interesting to observe that a mis-adapted policy, which in these scenarios is π_{fa-5}, attains the worst performance in both scenarios as it was designed assuming a decrease of inflows. On the contrary, the policy designed via sFQI outperforms π_h and is equivalent to π_{fa+5}.

5. CONCLUSION

The paper presents a novel method, called scenario-based Fitted Q-Iteration, to control water reservoir systems under climate change, which are modeled as MDPs under set-membership uncertainty. This method represents an extension of the Fitted Q-Iteration algorithm to learn a continuous approximation of the action-value function over the space of the uncertain parameters (i.e., the uncertain climate scenarios). As a result, sFQI embeds the set-membership uncertainty of the inflow scenarios and is able to approximate with a single learning process the optimal control policy for any scenario in the uncertainty set. The method is demonstrated on a simplified model of the Lake Como (Italy) under historical and perturbed inflow scenarios, where decreasing inflows significantly challenge the ability of the system of ensuring a reliable water supply to the downstream users.

Results show that the sFQI algorithm performs satisfactorily compared to the original FQI, regardless of the inflow scenario used for the evaluation of the policy. Since the sFQI policy is continuously approximated over the scenarios' space, it is also able to deal with realizations that were not directly experienced during the learning process. This feature is extremely valuable for designing adaptive control policies for managing water reservoir systems under climate-related uncertainty, where, generally, a solution fully adapted to a specific scenario suffers a degradation of performance when evaluated on a different scenario.

Future research efforts will focus on testing the sensitivity of sFQI to increasing levels of uncertainty by enlarging the range of variability of the inflows' scenarios as well as by including other co-varying factors affecting water reservoir systems (e.g., energy price, water demand). Moreover, we will assess the performance of sFQI in multi-objective control problems and we will validate its potential in real-world applications.

REFERENCES

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust optimization.* Princeton University Press.

Bonarini, A., Caccia, C., Lazaric, A., and Restelli, M. (2008). Batch Reinforcement Learning for controlling a Mobile Wheeled Pendulum robot. In *Artificial Intelligence in Theory and Practice II*, volume 276, 151–160.

Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement Learning and Dynamic Programming Using Function Approximators.* CRC Press, New York.

Castelletti, A., Galelli, S., Restelli, M., and Soncini-Sessa, R. (2010). Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research*, 46(W09507).

Castelletti, A., Pianosi, F., and Restelli, M. (2011). Multi-objective fitted Q-iteration: Pareto frontier approximation in one single run. In *Networking, Sensing and Control (ICNSC), 2011 IEEE International Conference on*, 260–265. Delft, NL.

Castelletti, A., Pianosi, F., and Soncini-Sessa, R. (2008). Water reservoir control under economic, social and environmental constraints. *Automatica*, 44(6), 1595–1607.

Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H., Giuliani, M., and Castelletti, A. (2016). A bottom-up approach to identifying the maximum operational adaptive capacity of water resource systems to a changing climate. *Water Resources Research.* doi: 10.1002/2015WR018253.

Denaro, S., Anghileri, D., Giuliani, M., and Castelletti, A. (2017). Informing the operations of water reservoirs over multiple temporal scales by direct use of hydro-meteorological data. *Advances in Water Resources*, 103, 51–63.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 503–556.

Ernst, D., Glavic, M., Capitanescu, F., and Wehenkel, L. (2009). Reinforcement Learning Versus Model Predictive Control: A Comparison on a Power System Problem. *IEEE Transactions on Systems, Man, and Cybernetics Part B-Cybernetics*, 39(2), 517–529.

Giuliani, M., Anghileri, D., Vu, P., Castelletti, A., and Soncini-Sessa, R. (2016a). Large storage operations under climate change: expanding uncertainties and evolving tradeoffs. *Environmental Research Letters*, 11(3). doi:10.1088/1748-9326/11/3/035009.

Giuliani, M. and Castelletti, A. (2016). Is robustness really robust? How different definitions of robustness impact decision-making under climate change. *Climatic Change*, 135, 409–424. doi:10.1007/s10584-015-1586-9.

Giuliani, M., Galelli, S., and Soncini-Sessa, R. (2014). A dimensionality reduction approach for many-objective Markov Decision Processes: Application to a water reservoir operation problem. *Environmental Modelling & Software*, 57, 101–114.

Giuliani, M., Li, Y., Castelletti, A., and Gandolfi, C. (2016b). A coupled human-natural systems analysis of irrigated agriculture under changing climate. *Water Resources Research*, 52(9), 6928–6947. doi: 10.1002/2016WR019363.

Giuliani, M., Li, Y., Cominola, A., Denaro, S., Mason, E., and Castelletti, A. (2016c). A matlab toolbox for designing multi-objective optimal operations of water reservoir systems. *Environmental Modelling & Software*, 85, 293–298.

Gordon, G. (1995). Online fitted reinforcement learning. In *Proceedings of the Workshop on Value Function Approximation at the 12th International Conference on Machine Learning, July 9*. Tahoe City, CA.

Hashimoto, T., Stedinger, J., and Loucks, D. (1982). Reliability, resilience, and vulnerability criteria for water resource system performance evaluation. *Water Resources Research*, 18(1), 14–20. doi:10.1029/WR018i001p00014.

Herman, J.D., Reed, P.M., Zeff, H.B., and Characklis, G.W. (2015). How Should Robustness Be Defined for Water Systems Planning under Change? *Journal of Water Resources Planning and Management.* doi: 10.1061/(ASCE)WR.1943-5452.0000509.

IPCC (2013). Climate change 2013: The physical science basis. Technical report, Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC).

Iyengar, G. (2005). Robust dynamic programming. *Mathematics of Operations Research*, 30(2), 257–280.

Kwakkel, J.H., Walker, W.E., and Haasnoot, M. (2016). Coping with the wickedness of public policy problems: approaches for decision making under deep uncertainty. *Journal of Water Resources Planning and Management*, 142(3), 01816001.

Lim, S., Xu, H., and Mannor, S. (2013). Reinforcement Learning in Robust Markov Decision Processes. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, 701–709. Curran Associates, Inc.

Milly, P., Betancourt, J., Falkenmark, M., Hirsch, R., Kundzewicz, Z., Lettenmaier, D., and Stouffer, R. (2008). Stationarity is dead. *Science*, 319, 573–574.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5), 780–798.

Ormoneit, D. and Sen, S. (2002). Kernel-based reinforcement learning. *Machine Learning*, 49(2-3), 161–178.

Pianosi, F., Castelletti, A., and Restelli, M. (2013). Tree-based fitted Q-iteration for multi-objective Markov decision processes in water resource management. *Journal of Hydroinformatics*, 15(2), 258–270.

Pineau, J., Guez, A., Vincent, R., Panuccio, G., and Avoli, M. (2009). Treating epilepsy via adaptive neurostimulation: a reinforcement learning approach. *International Journal of Neural Systems*, 19(4), 227–240.

Powell, W. (2007). *Approximate Dynamic Programming: Solving the curses of dimensionality.* Wiley, NJ.

Riedmiller, M., Gabel, T., Hafner, R., and Lange, S. (2009). Reinforcement learning for robot soccer. *Autonomous Robots*, 27(1), 55–73.

Tognetti, S., Savaresi, S., Spelta, C., and Restelli, M. (2009). Batch Reinforcement Learning for Semi-Active Suspension Control. In *2009 IEEE Control Applications CCA & Intelligent Control (ISIC), VOLS. 1-3*, IEEE International Conference on Control Applications, 582–587. St Petersburg, RUS.

Washington, W., Buja, L., and Craig, A. (2009). The computational future for climate and earth system models: on the path to petaflop and beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890), 833–846.

White, D. (1982). Multi-objective infinite-horizon discounted markov decision processes. *Journal of Mathematical Analysis and Optimization*, 89(2), 639–647. doi: 10.1016/0022-247X(82)90122-6.

Zhao, Y., Zeng, D., Socinski, M., and Kosorok, M. (2011). Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer. *Biometrics*, 67(4), 1422–1433.