강화학습법을 이용한 유역통합 저수지군 운영

Basin-Wide Multi-Reservoir Operation Using Reinforcement Learning

이진희*·심명필**

Jin Hee Lee*·Myung Pil Shim**

Abstract

The analysis of large-scale water resources systems is often complicated by the presence of multiple reservoirs and diversions, the uncertainty of unregulated inflows and demands, and conflicting objectives. Reinforcement learning is presented herein as a new approach to solving the challenging problem of stochastic optimization of multi-reservoir systems. The Q-Learning method, one of the reinforcement learning algorithms, is used for generating integrated monthly operation rules for the Keum River basin in Korea. The Q-Learning model is evaluated by comparing with implicit stochastic dynamic programming and sampling stochastic dynamic programming approaches. Evaluation of the stochastic basin-wide operational models considered several options relating to the choice of hydrologic state and discount factors as well as various stochastic dynamic programming models. The performance of Q-Learning model outperforms the other models in handling of uncertainty of inflows.

Key words: Q-Learning, stochastic DP, Reinforcement Learning, mult-reservoir system

1. Introduction

As populations expand and economies develop, increasing competition for limited available water resources is occurring among both intrabasin and interbasin users. This has brought greater attention to integrated river basin management, requiring an extended scale of water management without losing model detail and accuracy. However, the analysis of large-scale water resources systems is often complicated by the presence of (1) multiple reservoirs and diversions, (2) the uncertainty of unregulated inflows and demands, and (3) conflicting objectives (e.g. flood control vs. conservation purpose). In particular, the uncertainty of inflows makes it impossible to precisely identify future impacts of current decision-making. As a result, the efficient operation of multiple reservoir systems is a difficult and challenging task for water resources managers and the need for incorporating uncertainties in the planning and operation of multiple reservoir systems is important and necessary.

A possible method for overcoming the computational challenge of stochastic optimization of multireservoir systems is reinforcement learning. The search space over the range of the possible releases is reduced in reinforcement learning so the algorithm is faster than SDP, yet it also finds better answers than the standard SDP since it is much easier to incorporate the stochastic nature of the inflows. That is to say, acquiring apriori knowledge of the stochastic structure of inflows in SDP is extremely difficult when complex spatial correlations exist among the system inflows. SDP approaches require access to a multivariate time series model to generate many synthetic inflow sequences to build when the available historical data are not

* 정회원·한국건설기술연구원 수자원연구부 박사후연구원·E-mail: kolnidre@kict.re.kr

** 정회원·인하대학교 환경토목공학부 토목공학과 교수·E-mail: shim@inha.ac.kr

sufficient. Contrast, reinforcement learning approaches do not require this process since they can find good models through a learning process regardless of the complex stochastic structure of the inflows.

2. Reinforcement Learning System

A reinforcement learning system consists of the agent, environment, and their interactions. The learner or decision maker is called the agent and everything except the agent is called the environment. The agent is connected to its environment via action, reward, and state. Figure 1 depicts the components of a reinforcement learning system and the agent-environment interaction.

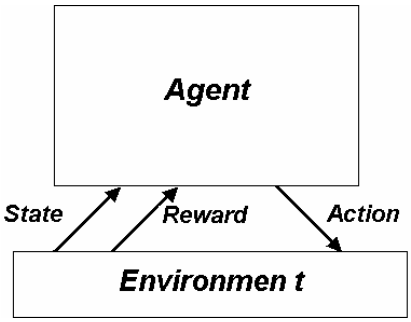

Figure 1. Reinforcement Learning System

Reinforcement learning generally consists of (1) a finite number of state $S=\{s_1, s_2, \cdots, s_n\}$ (2) a finite set of actions $A=\{a_1, a_2, \cdots, a_m\}$ available to an agent (3) a reward $r$given by the environment to the agent and (4) a state transition probability $P_{ss'}$ which determines the probability that the environment will make a transition to one state to another when the agent performs an action a.

The agent is supposed to find a policy $\pi_t(s,a): S\times A\rightarrow[0,1]$, mapping from the state to probabilities of selecting each possible action. A policy is denoted by $\pi_t(s,a)$ which is the probability of taking action a in state s. If the environment is stationary for simplicity, the probabilities of making state transitions or the immediate rewards do not change over time. As a result, the objective of the agent in stationary case is to determine a deterministic policy $\pi_t(s): S\rightarrow A$, mapping from the state to action.

Assuming as optimal policy is followed thereafter, the optimal action-value function $Q^*(s,a)$ can be defined as follows in terms of state-value function $V^*(s)$:

$$Q^*(s,a)=E(r_{t+1}+V^*(s_{t+1})|s_t=s, a_t=a) \tag{1}$$

Watkins (1989) develop an algorithm known as Q-learning, which learns optimal action-value function, $Q^*(s,a)$ is directly approximated from learned action-value function, $Q(s,a)$, regardless of the policy being followed. This algorithm has been proved early convergence (Sutton and Barto, 1998). The Q-learning algorithm performs the updates by following equation.

$$Q^*(s_t,a_t) := Q(s_t,a_t)+\alpha[r_t+\gamma\max_a Q(s_{t+1},a_{t+1})-Q(s_t,a_t)] \tag{2}$$

3. Case Study

The Keum River basin in Korea was chosen as a case study to demonstrate the applicability of the reinforcement learning algorithm. The priorities of water allocation in the sub basin are instream flows, domestic, industrial, agricultural water in the highest to the lowest. A deficit sharing policy was developed to allocate water deficits in accordance with the type of demand and level of use. This policy provides a dynamic approach to adjusting allocation of deficits considering the extent of demand satisfaction. According to the policy, all of the water demands are satisfied when there is sufficient water available. If there is a shortage in

water supply, it will first satisfy the first municipal demand block. Water continues to apply to the municipal water demand blocks until it reaches 100 percent supply level. The allocation process continues in a similar manner for the industrial water demands and then the agricultural water demands. The municipal and industrial demands used in the optimization are derived by averaging the demands in water years 2001 and 2002 to reflect the current situations of Keum River basin. The agricultural demands are the average of 19 years data from 1983 to 2002. The instreamflow for sub basin are the 95% exceedence percentile of 19-year historical flows for each sub basin and all other insreamflow requirements are provided by KOWACO.

3.1 Development of Model

In the study, the objectives are to minimize the water demand deficits and reservoir spills, and to maximize the hydropower generation. These multiple objectives are combined into a single objective function for the dynamic programming optimization using the weighting method. Equation (3) shows the immediate reward (return) with the weighting method during every time.

$$\sum_{i=1}^{N_{Generator}} w_1 P_i - \sum_{j}^{N_{Diversion}} \sum_{k=1}^{N_{Share}} w_{2jk} \left(\frac{100 \times (D_{jk} - U_{jk})}{D_{jk}} \right)^2 - \sum_{l=1}^{N_{Reservoir}} w_3 SP_l \tag{3}$$

where, w_1, w_2, w_3 are priority weighting factors for hydropower generation, diversions at the j^{th} diversion point and the k^{th} deficit sharing block, and reservoir spill, respectively; $N_{Generator}(=3)$ is the number of hydropower generators, $N_{Diversion}(=25)$ is the number of diversion points, $N_{Share}(=4)$ is number of deficit sharing blocks, and $N_{Reservoir}(=2)$ is number of reservoirs; P_i is the hydropower generation at generator i and SP_l is the spilled water from reservoir l; D_{jk} is demand at diversion point j and deficit sharing block k and U_{jk} is the actual diversion water at diversion point j and deficit sharing block k.

Unlike other reservoir optimization model there is no simplification of the simulation procedure representing the basin as precise as possible. In addition, the deficit sharing policy allocation along the river is based on the equation (4).

$$\arg\min_{U_{jk}} \sum_{j=1}^{N_{Diversion}} \sum_{k=1}^{N_{Share}} w_{2k} \left(\frac{100 \times (D_{jk} - U_{jk})}{D_{jk}} \right)^2 \tag{4}$$

$$\sum_{k=1}^{N_{Share}} U_{jk} \le \sum_{k=1}^{N_{Share}} D_{jk} \quad \text{for } j = 1,2,\cdots,N_{Diversion} \tag{5}$$

$$U_{jk+1} = 0 \qquad \text{if } U_{jk} < D_{jk} \quad \text{for } j = 1,2,\cdots,N_{Diversion} \text{ and } k = 1,2,\cdots,N_{Diversion} - 1 \tag{6}$$

3.2 Development of Operation Policy

Various stochastic dynamic programming approaches were applied to developing optimal coordinated operating rules for the two reservoir system of the Keum River basin. Three stochastic dynamic programming approaches were considered including implicit stochastic dynamic programming, SSDP, and reinforcement learning(Q-Learning). The conventional SDP is excluded due to the complexity of deriving transition probability of 12 sub basin inflows. Several options for defining the streamflow transition and discount factors in SSDP and reinforcement learning are tested. Figure 2 and 3 show the examples of operation policies derived.

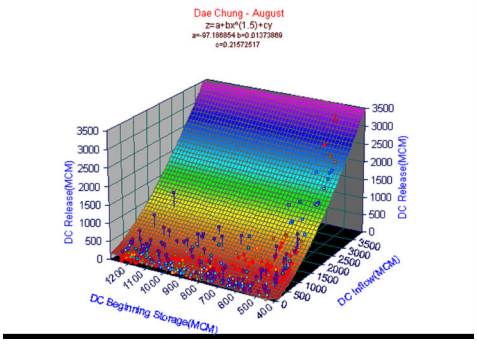

Figure 2. Implicit Stochastic DP

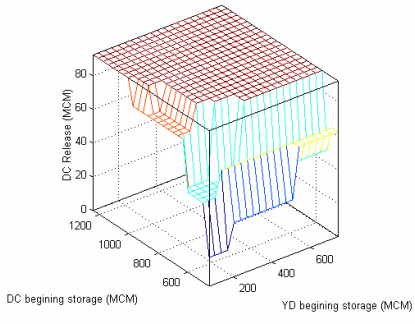

Figure 3. Reinforcement Learning

3.3 Simulation Analysis

Table 1. Simulation scenarios

Model	Hydrologic State	Discount Factor	Model Label
CSUDP	Inflow to the reservoir	NA	CSUDP
SSDP	NA	0.95	SSDP095
		0.80	SSDP08
Q-Learning	NA	0.95	Q095
		0.70	Q07
Q-Learning	Percentile	0.90	QP09
		0.80	QP08
Q-Learning	K-mean clustering	0.95	QK095
		0.90	QK09
Deterministic DP	NA	NA	OPT

Table 1 specifies all the scenarios evaluated in the study. The first column indicates the model option, such as CSUDP, SSDP, Q-Learning, and optimal policy. Hydrologic state in the second column specifies whether the model includes the basin-wide flow condition as a system state, whereas K-mean clustering and percentile approach are used to classify the hydrologic state. The options are only available for the Q-Learning model. Discount factors are applied to the SSDP and Q-Learning models, with value ranging from 0.7 to 0.95. Performances of release policies developed previously are evaluated using simulation analysis with the same performance measure.

The performances of the various operational rules were evaluated but some of the results are presented in this paper. Several explicit stochastic optimization models including Q095, QK095, and SSDP095 are compared with implicit stochastic optimization model (CSUDP) to investigate model performances. Operation of YongDam and DaeChung reservoirs are compared for the various models in Figure 3 and 4.

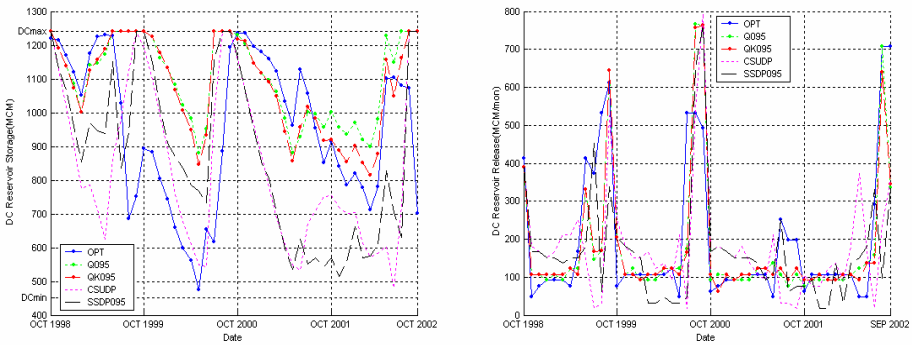

Figure 3. DaeChung reservoir storage change and release

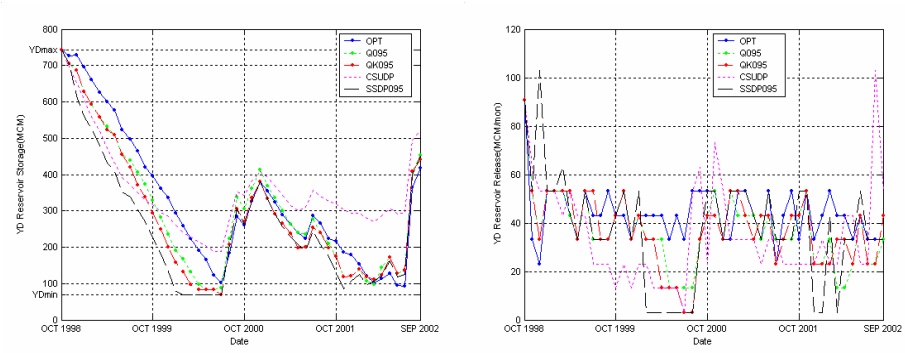

Figure 4. YongDam reservoir storage change and release

Figure 5 compares the monthly and total performance measures, showing that Q095 and OPT provide consistency in monthly performance measures, which the other models provide less performance measures during the low flow sequences. CSUDP has better performance in terms of large performance measure reduction than QK095 and SSDP095. The total performance measure indicates that the conditional or unconditional Q-Learning approaches outperform both CSUDP and SSDP. CSUDP and QK095 produced almost the same amount of total performance measures although their monthly performance measures are different.

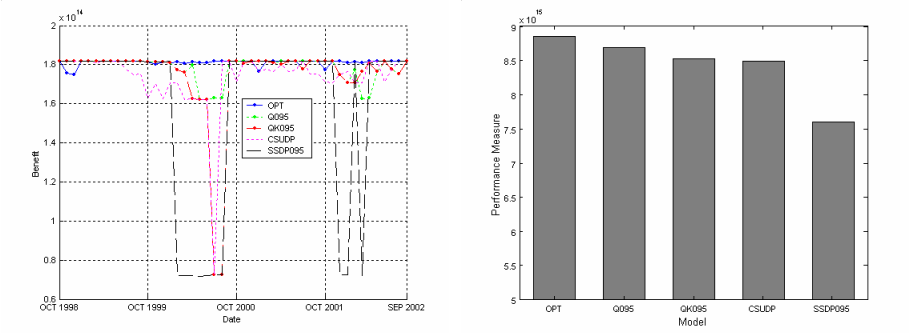

Figure 5. Monthly and total performance evaluation

5. Conclusion

The overall conclusions drawn by this study can be briefly summarized in the followings.

1) Large scale basin-wide reservoir operation rules were derived from Q-Learning models as well as various explicit stochastic dynamic models. The serial and cross correlations of streamflow were preserved by using historical streamflow data.
2) The primary advantage of the Q-Learning model is that predetermined transition probabilities of inflow and post inference procedure to derive the operational rules are not required.
3) The operating rules by Q-Learning models outperformed the other rules derived by SSDP and implicit stochastic optimization models.
4) A multiple linear or nonlinear regression model by implicit stochastic dynamic programming is well performed in basin-wide reservoir operation. However, the releases from the reservoirs are fluctuated according to the reservoir inflow condition.
5) Since SSDP was originally designed to use for the real time operation with a precise forecasting model it is not suitable for deriving long-term optimal operation rules.

References

1. Lee, Jin-Hee(2005), Basin-Wide Multi-Reservoir Operation Using Reinforcement Learning, PhD thesis, Colorado State University, Department of Civil Engineering.
2. Watkins, Christopher J.C.H.(1989), Learning from delayed rewards, PhD thesis, University of Cambridge, Psychology Department.
3. Sutton, R.S. and A.G. Barto.(1998). Reinforcement Learning:An Introduction. MIT Press, Cambridge, MA.