Journal of Hydrology 519 (2014) 732–742

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A game theory–reinforcement learning (GT–RL) method to develop optimal operation policies for multi-operator reservoir systems

Kaveh Madani [a,*], Milad Hooshyar [b]

[a] Centre for Environmental Policy, Imperial College London, London SW7 2AZ, UK
[b] Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form 28 July 2014
Accepted 28 July 2014
Available online 7 August 2014
This manuscript was handled by
Geoff Syme, Editor-in-Chief

Keywords:
Game theory
Reinforcement learning
Reservoir operation
Conflict resolution
Optimization
Evolutionary algorithm

SUMMARY

Reservoir systems with multiple operators can benefit from coordination of operation policies. To maximize the total benefit of these systems the literature has normally used the social planner's approach. Based on this approach operation decisions are optimized using a multi-objective optimization model with a compound system's objective. While the utility of the system can be increased this way, fair allocation of benefits among the operators remains challenging for the social planner who has to assign controversial weights to the system's beneficiaries and their objectives. Cooperative game theory provides an alternative framework for fair and efficient allocation of the incremental benefits of cooperation. To determine the fair and efficient utility shares of the beneficiaries, cooperative game theory solution methods consider the gains of each party in the status quo (non-cooperation) as well as what can be gained through the grand coalition (social planner's solution or full cooperation) and partial coalitions. Nevertheless, estimation of the benefits of different coalitions can be challenging in complex multi-beneficiary systems. Reinforcement learning can be used to address this challenge and determine the gains of the beneficiaries for different levels of cooperation, i.e., non-cooperation, partial cooperation, and full cooperation, providing the essential input for allocation based on cooperative game theory. This paper develops a game theory–reinforcement learning (GT–RL) method for determining the optimal operation policies in multi-operator multi-reservoir systems with respect to fairness and efficiency criteria. As the first step to underline the utility of the GT–RL method in solving complex multi-agent multi-reservoir problems without a need for developing compound objectives and weight assignment, the proposed method is applied to a hypothetical three-agent three-reservoir system.

1. Introduction

Economies of scale and negative externalities can potentially provide incentives for cooperation in multi-beneficiary resource management systems (Asgari et al., 2014). Multi-operator multi-reservoir systems provide a good example of multi-beneficiary resource management systems in which coordination of operation policies can create opportunities to increase the total benefits of the system. Such opportunities have encouraged a large group of researchers to use optimization models to develop operation policies that can maximize the total benefits of multi-reservoir systems (Labadie, 2004). Given the variety of multi-reservoir system operation problems and the difference in their complexity, a range of optimization methods have been applied to multi-reservoir problems, including linear programming (Willis et al., 1984), successive linear programming (Hiew, 1987), mixed-integer linear programming (Needham et al., 2000), successive quadratic programming (Tejada-Guibert and Stedinger, 1990), method of multipliers (Arnold et al., 1994), dynamic programming (Lee and Labadie, 2007), chance constrained programming (Loucks and Dorfman, 1975), stochastic linear programming (Thomas and Watermeyer, 1962), stochastic dynamic programming (Mousavi and karamouz, 2003), sampling stochastic dynamic programming (Kelman et al., 1990), network flow optimization (Lund and Ferreira, 1996) and reinforcement learning (Lee and Labadie, 2007; Mahootchi, 2009).

While the literature is rich in using optimization models to maximize the benefits of multi-reservoir systems, less attention has been paid to distribution of the incremental benefits of coordination in such systems. Cooperative game theory (CGT) can help with finding fair and efficient allocation of the incremental benefits of cooperation. Given the difference in notion of fairness, various

* Corresponding author.
E-mail addresses: k.madani@imperial.ac.uk (K. Madani), hooshyar.milad@knights.ucf.edu (M. Hooshyar).

http://dx.doi.org/10.1016/j.jhydrol.2014.07.061
0022-1694/

CGT methods propose different allocation solutions. Thus, stability analysis methods are employed to determine the most stable CGT allocation solution with the highest acceptance likelihood in each particular cooperative setting. CGT has been applied to different cost/benefit sharing problems in the water resources field (Parrachino et al., 2006; Madani, 2010) such as cost sharing of a joint enterprise (Nakayama and Suzuki, 1976), cost sharing among municipalities (Young et al., 1982), multi-objective management of a regional aquifer (Szidarovszky et al., 1984), water quality control by farmers (Ratner and Yaron, 1990), management of a regional water development project (Dinar, 2001), basin-wide cooperative water resources allocation (Wang et al., 2008), benefit and water allocation in inter-basin water transfer projects (Sadegh et al., 2010), hydropower licensing (Madani, 2011), sharing transboundary water systems (Teasley and McKinney, 2011; Imen et al., 2012), and cooperative management of shared aquifers (Madani and Dinar, 2012), among others.

CGT helps finding the share of each cooperating party from the overall gain, such that all parties remain cooperative. In order for the grand coalition (cooperation among all parties) to be stable, parties must not have any incentive to act non-cooperatively (in a singleton coalition) or form partial coalitions. This means that their gains from participation in the grand coalition must be greater than their possible gains through participating in partial or singleton coalitions. So, to determine the fair and efficient allocation of benefits, CGT considers the values (obtainable benefits) of all possible coalitions of players. Calculation of all coalition values can be challenging in problems with more than a few players, especially in problems with multiple time-steps (e.g., when players need to make monthly decisions with the objective of maximizing their gain over 20 years). This is because not only the players participating in a given coalition would act differently from the status quo, but also because the players who do not join the coalition would change their strategies in response to changed behavior of the other players. In such situations both groups of players (cooperating and non-cooperating) develop their best response strategies resulting in a new equilibrium for which the obtainable benefits of existing coalitions need to be determined. A considerable computational effort is required to calculate the obtainable benefits of all possible coalitions under their best response strategies in games with multiple players. Given the involved complexity, the water resources literature has either overlooked the significance of the possibility of partial coalitions and modifying the best response strategies, or made simplifying assumptions to determine the value of partial coalitions in complex multi-period problems with interacting agents.

To address the existing limitations in the literature this paper develops a new game theory–reinforcement learning (GT–RL) method for solving complex multi-period multi-beneficiary games with interacting agents. Reinforcement learning (Sutten and Barto, 2000) is a simulation-based optimization method which relies on interacting agents to find the optimal (or near-optimal) solution. RL is employed here to develop best response strategies under all possible coalitions in the game and to calculate the obtainable coalition benefits, providing the required information for CGT solution methods in order to determine fair and efficient allocations of the incremental benefits of cooperation. To show the applicability of the proposed method (Fig. 1), GT–RL is applied to a hypothetical numerical three-operator three-reservoir system in which parties can increase their gains from cooperative hydroelectricity production.

Generally, the GT–RL method (Fig. 1) has three major steps. First, the obtainable benefits under each possible coalition are obtained by applying the RL algorithm. Then, CGT solution methods are applied to find fair and efficient allocations of the incremental benefits of cooperation among the agents based on different notions of fairness. Finally, stability of each CGT allocation solution is examined using the stability analysis methods to find the allocation solution with the highest acceptance potential. Different components of the GT–RL methods are described in the following sections.

2. Cooperative game theory

CGT is concerned with allocating the incremental benefits of cooperation among the cooperating parties. So, CGT is only applicable when cooperation would generate extra benefits. A necessary condition for full cooperation is that each cooperating party's cooperative gain is higher than what he can achieve non-cooperatively or by participation in smaller coalitions. The interest is normally in the grand coalition involving all players, provided that the maximum incremental benefit is achievable when all parties cooperate.

The following three conditions form the core of the cooperative game, i.e., the set of all allocations that are potentially acceptable by the cooperating parties:

$$u_i^* \geqslant u_i \quad \forall i \in N \tag{1}$$

$$\sum_{i \in s} u_i^* \geqslant v(s) \quad \forall s \in S,\ S \subseteq N \tag{2}$$

$$\sum_{i \in N} u_i^* = v(N) \tag{3}$$

where u_i^* is the benefit share of the player i under cooperation, u_i is the player i's status quo (non-cooperative) gain, N is the number of players, $v(s)$ is the total gain of coalition s, and $v(N)$ is the total gain of the grand coalition involving all players.

The first equation sets the individual rationality condition, requiring that the cooperative gain of each player would be greater than his status quo (non-cooperative) gain. The second conditions sets the group rationality condition, requiring that the total of cooperative gains of any subset of players (subset of N) would be greater that their total gain if they form a coalition together. This condition does not leave any incentive for the players to leave the grand coalition in order to form partial coalitions. The third equation sets the efficiency condition, requiring the total gain under cooperation is fully allocated to the members of the grand coalition.

Based on different notions of fairness, researchers have proposed various solution methods to develop a unique 'fair' cooperative solution. In general, allocation solutions, which do not belong to the core have no practical value. While any solution belonging to the core is potentially acceptable by the cooperating parties, they might have different preferences over the core solutions. Researchers have shown more interest in using CGT solution methods that select the single allocation solution from the core.

Three of the most commonly used CGT solution methods for allocation of the cooperative gains are reviewed here.

2.1. Nash–Harsanyi bargaining solution

Harsanyi (1959) suggested the following optimization model for finding the fair cooperative solution for the n-player bargaining game over sharing the cooperation benefits as an extension to the 2-player bargaining solution proposed by Nash (1953).

$$\text{Max} \prod_{i=1}^{n} (u_i^* - u_i) \tag{4}$$

subject to the core conditions (Eqs. (1)–(3)).

This optimization model maximizes the product of the incremental gain of the players from cooperation. The optimization

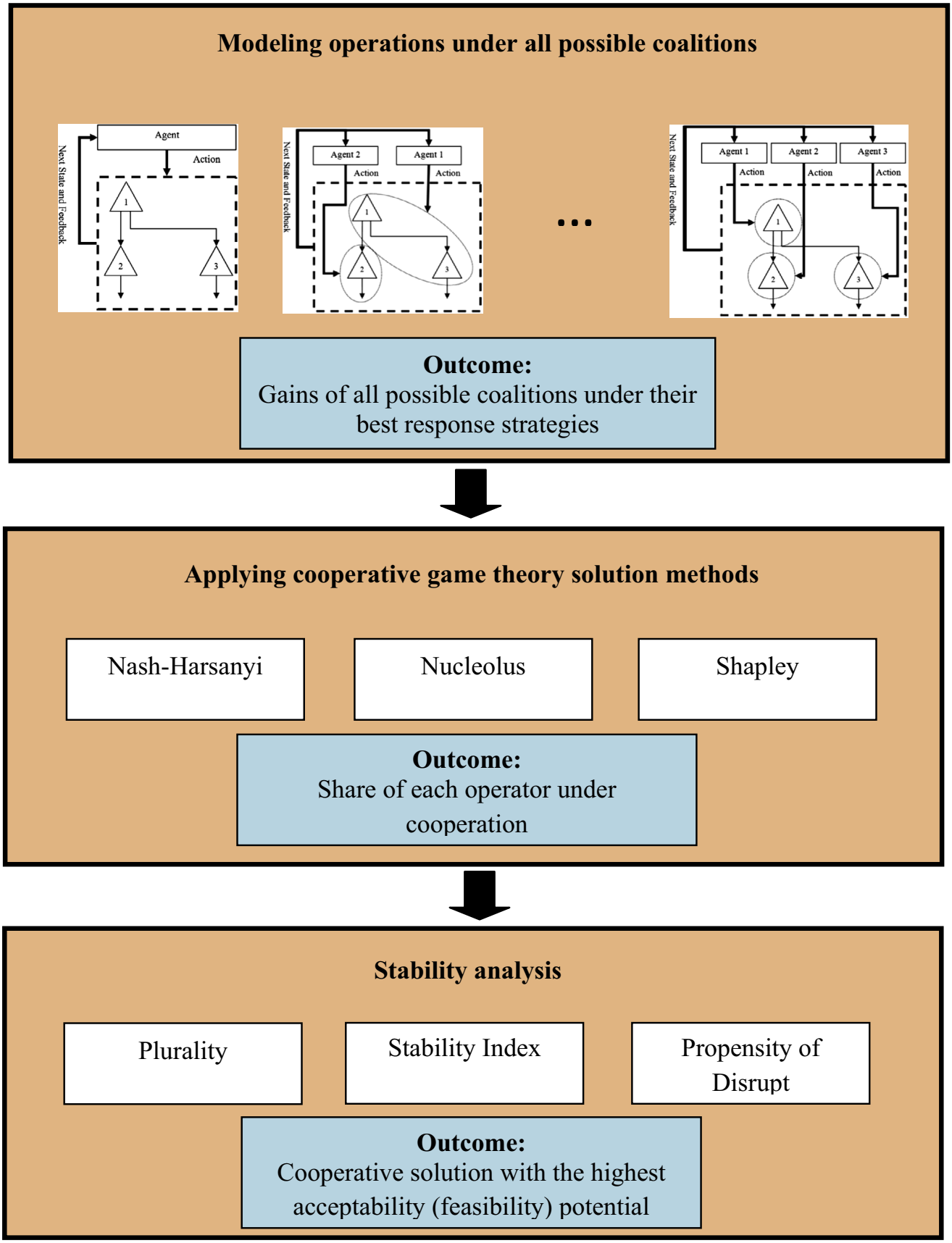

Fig. 1. The general procedure of developing stable cooperative reservoir operations solutions based on the RL–GT method.

constraints (Eqs. (1)–(3)) ensures that the optimization solution belongs to the core if the core is non-empty.

2.2. The Shapley value

Shapley (1953) proposed the following equation for calculation of the fair gain of the players from cooperation:

$$u_i^* = \sum_{\substack{S \subseteq N \\ i \in S}} \frac{(n-|s|)!(|s|-1)!}{n!}(v(s) - v(s-\{i\})) \quad (5)$$

where n is the total number of players, $|s|$ is the number of members in coalition s, and $v(s-\{i\})$ is the value of coalition s without member i.

Shapley's solution concept – Shapley value – defines a unique symmetric payoff vector. It determines the fair allocations based on the weighted average of the players' contributions to different coalitions and sequences, and assigns a zero gain to null players (players whose inclusion in the coalition does not result in any incremental benefit).

2.3. Nucleolus

Nucleolus (Schmeidler, 1969) is another commonly used cooperative game theoretic solution method. To find the fair allocation Nucleolus minimizes the dissatisfaction of the most dissatisfied coalition by imposing tax (ε) to all coalitions and keeping them in the core. The optimal level of tax and the allocation solution based on the Nucleolus concept can be determined using the following optimization model:

$$\text{Max } \varepsilon \quad (6)$$

subject to:

$$\varepsilon \leqslant \sum_{i \in s} u_i^* - v(s) \quad \forall s \in S,\ S \subseteq N \quad (7)$$

Eq. (3)

where $\sum_{i \in s} u_i^* - v(s)$ is the loss of players in coalition s if they leave the grand coalition.

The above optimization model has a unique solution belonging to the core if the core is not empty.

3. Reinforcement learning

Estimating the values of all possible coalitions ($v(s)$) is essential to successful application of CGT concepts (e.g., Nash–Harsanyi, Shapley and Nucleolus). As discussed earlier, calculation of coalition values can be challenging in complex multi-step multi-agent water problems with significant externalities, especially when non-cooperative parties have a chance to change their strategies in response to formation of partial coalitions by other players. Multi-agent reinforcement learning (RL) can be used to address this challenge and derive the best response strategies of all combinations of non-cooperative (staying out of coalitions) and cooperative (participating in coalitions) players.

Inspired by behaviorist psychology, RL (Sutten and Barto, 2000) is a simulation-based optimization method that utilizes interacting agents to find the optimal (or near-optimal) policy of an agent through interaction with an environment that includes all other active agents. Based on this evolutionary computational method, the learner (agent) is trained to take optimal (or near-optimal) actions through interaction with the environment. In contrast to supervised learning methods requiring example policies to be provided by an external supervisor (e.g., Artificial Neural Networks), the RL-based learning process is conducted through interaction with a dynamic environment and by analyzing the feedbacks from earlier decisions (Sutten and Barto, 2000).

RL comprises two main elements, the agent (learner) and the learning environment. Through a learning process, the agent finds out the set of optimal actions that can affect the environment. The agent has to be able to take any action included in the set of admissible actions, defined by the modeler (e.g., different values of water release from a reservoir) and to sense the feedbacks of his actions. Feedback is the only guide for the agent to improve its decisions. Environment is defined as the set of states which the agent might visit. The aim of learning is to find the optimal action in each state. In simple reservoir operations, for example, the agent can be the operator who makes release decisions. The set of discrete reservoir storage levels can be considered as the environment. The objective of learning in this case would be finding the best release strategy for a given level of storage with respect to the operation objective(s) (e.g., maximizing the hydropower generation profits, minimizing expected flood costs, minimizing water shortage costs, maximizing recreational benefits) and constraints (e.g., upper and lower storage/release levels, maximum ramping rate, maximum temperature).

Q-Learning (Watkins, 1989; Watkins and Dayan, 1992) is one of the commonly used RL techniques for finding the optimal policy (set of actions) for any finite Markov decision process. Q-Learning is a simulation-based version of Stochastic Dynamic Programming (SDP) (Gosavi, 2003). Based on this method, the value function of SDP, which is a value defined for every pair of action–state, is estimated using the Robbins–Monro algorithm (Robbins and Monro, 1951). The value function here has the same meaning as in SDP and includes two main components: immediate reward and accumulated reward. Because in RL the learning process is implemented by direct interaction with the environment, value functions have to be updated after each interaction using the following equation (Gosavi, 2003):

$$Q_{new}(s,a) = Q_{old}(s,a) + \alpha \times \left[r_{ss'}(a) + \gamma \max_{b \in A(s')} Q_{old}(s',b) - Q_{old}(s,a)\right] \quad (8)$$

where $Q(s,a)$ represents the accumulated reward of the agent once the agent takes action a in state s; the subscript old/new refers to before/after taking action a; s' is the next state; $r_{ss'}(a)$ is the immediate reward when action a is taken for transition from state s to state s'; $b \in A(s')$ is an action from the set of admissible actions (defined by the modeler) in state s' ($A(s')$); α is the learning parameter (i.e., the rate by which the Q-factors are updated) equal to $1/(NOV(s,a))$; $NOV(s,a)$ is the number of times that action–state pair (s,a) has been visited (tried) so far; and $\gamma < 1$ is the discount factor which depreciates the effect of agent's previous decisions on the accumulated reward.

Theoretically, RL will converge to an optimal policy if (Gosavi, 2003):

1. $\alpha = \frac{1}{NOV(s,a)} \quad \forall s,\ a$
2. $\gamma < 1$
3. $NOV(s,a) \to \infty \quad \forall s,\ a$

Since meeting the third condition is computationally intensive, RL convergence can be ensured by reasonable manual modification (tuning) of the solution algorithm variables/parameters to speed up the convergence process. Appropriate values of the tuning variables are normally determined through an iterative process involving exploratory runs by the modeler, who ensures that learning curves of the agents do not show instabilities and unreasonable behavior. Tuning involves setting $\alpha = \frac{A_0}{NOV(s,a)}$, determining a value for A_0 (where $A_0 < 1$), and adjusting γ to reduce the effects of initial observations of the agent on the final solution. Heuristic action

taking methods such as ε-greedy (Gosavi, 2003) (illustrated later) can be employed to help the agent make better decisions and sufficiently visit the near-optimal solutions.

The water resources literature includes application of RL to different problems. Bergez et al. (2001) compared the performance of RL with DP in identifying optimal starting irrigation strategies when a limited amount of water is available for irrigation. They concluded that RL outperforms DP when the budget of simulations is limited due to computational constraints. Lee and Labadie (2007) applied Q-Learning to a two-reservoir system in South Korea to find the optimal operational policy. Performance of Q-learning was compared with some SDP-based optimization methods including Sampling Stochastic Dynamic Programming (SSDP) and implicit stochastic optimization using Dynamic Programming (DP). Their results showed that Q-learning outperformed the other two methods. Mahootchi et al. (2007) used an opposition-based technique to speed up the convergence of Q-learning in reservoir operations problems. Castelletti et al. (2010) developed a new method based on RL and tree-based regression to improve the computational efficiency of the learning process. In comparison with SDP, the tree-based RL was found to have a better performance and to be computationally more tractable.

The reviewed water resources studies have used RL as a heuristic optimization method to solve non-linear optimization problems in a social planner mode, where a single central decision maker (learning agent) optimizes its decisions to maximize the compound benefits of the system. Due to its nature, RL also has a strong potential to solve multi-agent problems in which each agent or group of agents simultaneously and independently maximize their own objective functions, creating externalities for each other. This study shows how to take advantage of this capability of RL in an example multi-agent multi-period problem. In this hypothetical numerical problem, the operators of three interconnected hydropower reservoirs try to coordinate their strategies in order to maximize their benefits while challenged with finding a fair and efficient allocation scheme to share the incremental benefits of cooperation.

4. Numerical example

The proposed GT–RL method (Fig. 1) is applied to a hypothetical multi-reservoir system. The example system (Fig. 2) includes three reservoirs, connected through channels. These reservoirs are used for hydroelectricity production. For simplicity, the benefit of power generation is assumed to be a linear function of release. It is also assumed that these reservoirs are high-head single-purpose units with negligible head-storage effects (similar to the single-purpose high-head hydropower production reservoirs in California (Madani et al., 2014a)).

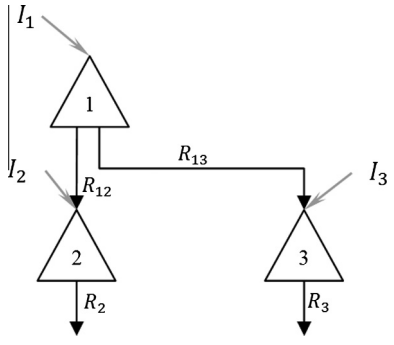

Fig. 2. Schematic representation of the example problem.

The following two sections present the objective function of the agents and their operations constraints in the hypothetical example. Sections 4.3–4.5 describe the procedure of finding the best response strategies of the agents and calculation of the obtainable benefits under all possible coalitions (grand coalition, partial coalitions and singleton coalitions). Sections 4.6 and 4.7 present the resulting game theoretic allocation solutions and the procedure for finding the most stable allocation solution.

4.1. Objective function

The objective of each agent i is to maximize the annual revenue of hydroelectricity generation from the reservoir it is operating by finding the best monthly release policy (rule curve):

$$\text{Max} \sum_{j=1}^{N} \sum_{t=1}^{T} b_i^t \times R_{ij}^t \tag{9}$$

where R_{ij}^t is the amount of release from reservoir i to reservoir j through the ij channel in month t (j is emitted from the equation when there is no reservoir at the end of the release channel), b_i^t is the benefit per unit of water release in month t (Table 1), and N is the number of channels below reservoir i.

It is assumed that the average cost per unit energy produced is constant in the whole system (irrespective of the choice of reservoir for energy generation). In this case, revenue maximization is a surrogate for profit maximization.

4.2. Major optimization constraints

4.2.1. Continuity equation

This equality constraint is imposed for conservation of mass:

$$S_i^{t+1} = S_i^t - \sum_{j=1}^{N} R_{ij}^t + I_i^t - v_i^t + \sum_{j=1}^{N} R_{ji}^t, \quad \forall i = 1 \ldots N \;\&\; t = 1 \ldots T \tag{10}$$

where I_i^t is the amount of inflow to reservoir i in month t (Table 2) and S_i^t is the volume of stored water in reservoir i, v_i^t is the water loss at reservoir i in month t, including evaporative and seepage losses. The value of release from one reservoir to another is always is zero when the two reservoirs are not connected directly.

4.2.2. Storage limits

The water storage in the reservoir should be kept within a certain range to maintain the required head for hydropower operations without creating an excessive risk of flooding:

$$S_{i,min} \leqslant S_i^t \leqslant S_{i,max} \tag{11}$$

where $S_{i,max}$ and $S_{i,min}$ are the maximum and minimum volume of stored water in reservoir i, respectively (Table 3).

Table 1
Hydroelectricity production profit per unit of water release ($/cubic meters).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
R_{12}	0.210	0.200	0.190	0.180	0.170	0.160	0.010	0.010	0.010	0.010	0.010	0.010
R_{13}	0.210	0.200	0.190	0.180	0.170	0.160	0.010	0.010	0.010	0.010	0.010	0.010
R_2	0.005	0.005	0.005	0.005	0.005	0.005	0.180	0.190	0.200	0.210	0.220	0.230
R_3	0.005	0.005	0.005	0.005	0.005	0.005	0.180	0.190	0.200	0.210	0.220	0.230

Table 2
Monthly reservoir inflows (MCM).

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Reservoir 1	117.3	143.0	62.3	31.2	20.8	11.7	4.5	4.8	5.2	4.8	7.5	61.6
Reservoir 2	85.5	99.8	56.0	44.8	34.2	17.5	10.3	7.6	6.3	5.5	6.8	38.6
Reservoir 3	42.1	47.4	22.1	16.2	12.7	17.4	14.8	5.3	5.3	4.3	5.6	19.7

Table 3
Upper and lower bounds of release and storage (MCM).

		R_{max}	R_{min}	S_{max}	S_{min}
Reservoir 1	To reservoir 2	100	0	300	0
	To reservoir 3	70	0	0	0
Reservoir 2		200	0	150	0
Reservoir 3		150	0	100	0

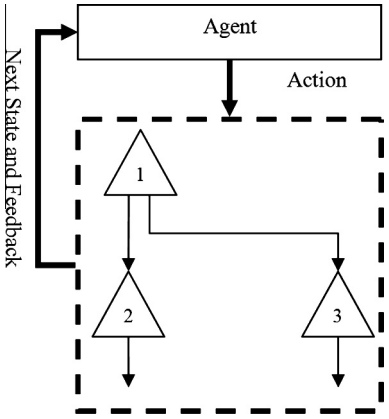

Fig. 3. Single-agent (grand coalition operations).

4.2.3. Release limits

The release constraints are imposed by the maximum capacity of outlets or downstream reaches and the minimum required environmental flows (if any).

$$R_{ij,Min} \leqslant \sum_{j=1}^{N} R_{ij}^{t} \leqslant R_{ij,Max} \tag{12}$$

where $R_{ij,Max}$ and $R_{ij,Min}$ are maximum and minimum releases from reservoir i, into channel ij, respectively (Table 3).

4.3. Optimization in the social planner mode (grand coalition)

Conventional systems engineering applications in the water resources management context normally optimize the system in the social planner mode. The inherent assumption of this type of planning is that the interest is in maximizing the social welfare of the system based on collective action. So, in this case, the social planner assumes that the parties are willing to fully cooperate based on group (collective) rationality in order to maximize the total benefits of the system (Madani and Dinar, 2012).

Past applications of RL in water resources planning also maximize the benefits in the social planner mode, putting one agent in charge of the whole system. Similarly, to calculate the total obtainable benefits by the grand coalition, we can assume that the 3-reservoir system is run by a single agent (Fig. 3) who tries to maximize the total benefits of the system. In the grand coalition, the state of the social planner agent is a vector of storages in reservoirs 1–3 and its action is a vector of releases from these reservoirs.

Here, the number of learning episodes was set to 100, each containing 100 years of simulation. This means that the agent starts by learning through 100 years of simulations (episode 1). The system is reset then to a new (random) initial state and the learning process continues for another 100 years (episode 2). This process continues for 100 episodes. Selection of the number of required episodes and simulation steps (years in this case) is one of the model tuning steps. Longer episodes and simulation periods are more desirable in general when there is no computational budget limitation. The number and length of learning episodes are selected based on the preliminary (exploratory) model runs and must to be sufficiently large to ensure convergence to the optimal solution. Exploratory runs in this study suggested that 100 episodes would be sufficient for RL's convergence in all three modeling modes. As another tuning element, A_0 was also set to 0.9 for all model runs in this study.

The steps of learning process of a single (social planner) agent in charge of the whole system are as follows:

1. For each episode $n = 1,\ldots,100$
   1.1. For each year $y = 1,\ldots,100$
      1.1.1. For each month $t = 1,\ldots,12$
         1.1.1.1. *Take action:* The social planner agent takes action a from a set of admissible (feasible) actions defined by the modeler. This action is a vector of releases from all three reservoirs.

         In this study, we use ε-greedy as a heuristic action taking method to improve the learning process. Based on the ε-greedy method (Gosavi, 2003), the agent might take a random action with a probability of ε (exploration) or takes the best (current) action with a probability of $1-\varepsilon$ (exploitation). The current best action is the best action based on agent's current knowledge and might change as the learning proceeds. Exploration ensures that the agent is not stuck in a local optimum by giving the agent a chance of taking random actions. The probability of exploration diminishes in the later stages of the learning process. Here, we assumed that ε decreases linearly from 1.0 in episode 1 to 0.5 in episode 100 as the agent gains more experience.

         It is noteworthy that a sufficiently large number of episodes helps ensuring that the agent is provided with sufficient exploration opportunities to avoid local optima. Also, given that each episode is independent from other episodes and has random initial conditions (e.g., reservoir storage level), having more episodes minimizes the effect of the initial conditions on the optimum policy developed by the agent and lets the agent explore different initial conditions.

         1.1.1.2. *Calculate the immediate reward:* Assuming that the agent is in state s and will move to state s' by taking action a (a vector of R_{ij}^{t}), the feedback of action a in state s ($r_{ss'}^{t}$) is calculated using Eq. (13):

$$r_{ss'}^{t} = \sum_{i=1}^{N}\sum_{j=1}^{N} b_i^t \times R_{ij}^{t} \tag{13}$$

If the decision taken in step 1.1.1.1 is not feasible, the agent does not receive a full reward. In other words, the actual reward will be less than the perceived reward. For example, if the agent decides to release 20 units of water while the available water is 10 units only, the agent gets rewarded for releasing 10 units (all the available water) only which will be less than the perceived reward for 20 units. Based on this feedback the agent learns to avoid infeasible release strategies in the future years/episodes whenever making decisions in the exploitation mode (based on past decisions) as opposed to the exploration (random) mode that ignores previous feedback.

1.1.1.3. *Update the Q-factor:* Based on the feedback received in step 1.1.1.2, the agent updates its *Q*-factor using the following equation:

$$Q^t_{new}(s,a) = Q^t_{old}(s,a) + \alpha \times \left[r^t_{ss'}(a) + \gamma \max_{b \in A(s')} Q^{t+1}_{old}(s',b) - Q^t_{old}(s,a) \right] \quad (14)$$

where superscripts t and $t+1$ refer to the current month and next month, respectively; and subscripts *new* and *old* refer to the current year and previous year, respectively. For example, combination $t+1$, *old* corresponds to the $t+1$ (next) month in the *old* (previous) year.

Based on preliminary model runs, the discount factor (γ) was set to 0.7 in this case to ensure appropriate convergence.

1.1.1.4. *Update the policy:* Using the updated *Q*-factor in time-step t, the agent updates its policy (release decision) for state s using the following equation:

$$\pi^t(s) = \arg\max_a Q^t(s,a) \quad (15)$$

It most be noted that the agent's policy in a given time step is essentially a rule curve that specifics the best vector of releases (from the reservoirs controlled by that agent) for different vectors of storages in these reservoirs in that time step. In this case, the social planner agent's policy specifies the best release levels from the three reservoirs for each unique combination of storage levels in these reservoirs.

1.1.2. End loop

1.2. End loop

1.3. *Evaluate policy performance:* At the end of each episode, performance of the developed policy is tested by applying this policy in a simulation model which is run for a sufficiently long period. To ensure that initial conditions have no effect on the policy performance, simulations continue until the annual revenue becomes constant. The average annual revenue of the system over the simulation period is then considered as a measure to evaluate the performance of the developed policy.

2. End loop

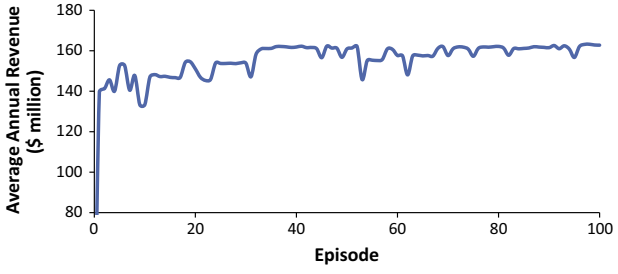

Fig. 4. Learning curve of the social planner agent.

Fig. 4 shows the learning curve of the social planner agent. This curve shows how the performance of the agent's selected policy changes in each episode. In this numerical example, the social planner agent could develop a policy resulting in an average annual revenue of $163 million.

4.4. Optimal operations of partial coalitions

In absence of the grand coalition, parties might form partial coalitions or act individually (in singleton coalitions). So, there is a need to calculate the obtainable benefits of all possible coalitions in the game. To find the obtainable benefits under partial coalitions, the best response strategies of the cooperating and non-cooperating parties need to be identified. Given that a coalition needs at least two members, in the example case, only one partial coalition (with two members) is possible at a time. In this case, it is assumed that one agent is running the reservoirs belonging to the coalition members and another agent operates the non-cooperative reservoir. So, the problem is solved in a multi-agent mode with two agents. For example, for coalition {1,3}, one agent is assigned to reservoirs 1 and 3 and another agent is assigned to reservoir 2 (Fig. 5). The first agent maximizes the overall revenue of the coalition, while the other agent is maximizes its benefit from reservoir 2. In this case, the state of agent 1 is the vector of storages of reservoirs 1 and 3 and its action is the vector of releases of reservoirs 1 and 3. Storage of reservoir 2 and its release are the state and action of agent 2. The maximum obtainable benefit of the two sides occurs at the non-cooperative Nash equilibrium in which both parties are using their best response strategies.

The learning process is the same as in the social planner condition. The only difference is that here there are multiple (two in this case) agents. Same as in the social planner mode, in this numerical example, the learning procedure included 100 episodes, each one containing 35 years of simulations. Here, the learning process for the partial coalition mode is explained for an example case where coalition {1,3} operates against singleton coalition {2}.

1. For each episode ($n = 1,\ldots,100$)
   1.1. For each year ($y = 1,\ldots,35$)
      1.1.1. For each month ($t = 1,\ldots,12$)
         1.1.1.1. *Take action:* Agent 1 (representing operators 1 and 3) takes action a_1. The action is a vector of releases from reservoirs 1 and 3. At the same time, agent 2 (operator 2) takes action a_2, which is the release from reservoir 2. Actions are made based on the ε-greedy method explained earlier.

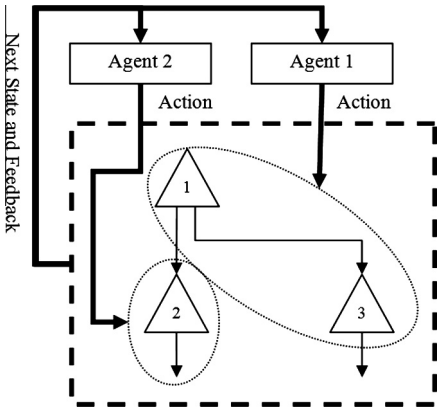

Fig. 5. Multi-agent operations for coalition of operators 1 and 3 against operator 2.

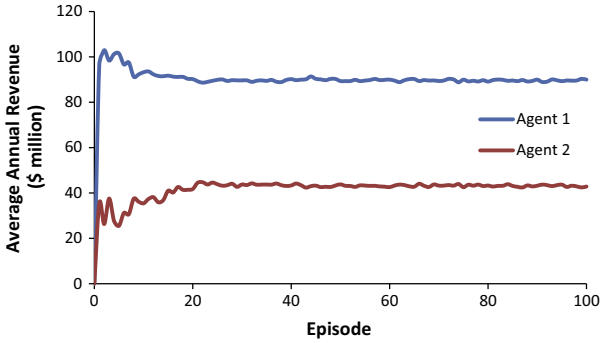

Fig. 6. Learning curves for agent 1, running coalition {1,3}, and agent 2, in charge of coalition {2}.

1.1.1.2. *Calculate the immediate reward:* Assuming that agent i is in state s_i and will go to state s_i' by taking action a_i, the feedback of action a_i $\left(r^t_{i,s_i s_i'}\right)$ is calculated using Eq. (16):

$$r^t_{i,s_i s_i'} = \begin{cases} \sum_{j=1,3} b^t_i \times R^t_{ij}, & i = 1 \\ \sum_{j=2} b^t_i \times R^t_{ij}, & i = 2 \end{cases} \quad (16)$$

1.1.1.3. *Update the Q-factor:* Having the feedback calculated in step 1.1.1.2, each agent i updates its Q-factor using the following equation:

$$Q^t_{i_{new}}(s_i, a_i) = Q^t_{i_{old}}(s_i, a_i) + \alpha \times \left[r^t_{i,s_i s_i'}(a_i) + \gamma \max_{b \in A(s_i')} Q^{t+1}_{i_{old}}(s_i', b) - Q^t_{i_{old}}(s_i, a_i)\right], \quad \forall i \quad (17)$$

Based on preliminary model runs, the discount factor (γ) was set to 0.9 in this case to ensure appropriate convergence.

1.1.1.4. *Update policy:* Using the updated Q-factor, each agent's policy is updated using the following equation:

$$\pi^t_i(s_i) = \arg\max_a Q^t(s_i, a), \quad \forall i \quad (18)$$

1.1.2. End loop

1.2. End loop

1.3. *Evaluate policy performance:* Similar to the previous case, at the end of each episode, performance of the developed policies are tested by applying these policies in a simulation model which is run for a sufficiently long period. To ensure that initial conditions have no effect on policy performance, simulations continue until the annual revenues become constant. The total average annual revenues of reservoirs 1 and 3 over the simulation period is considered as a performance measure for agent 1 to evaluate its policy performance, and the average annual revenue of reservoir 2 over the simulation period is used by agent 2 to evaluate the performance of its policy.

2. End loop

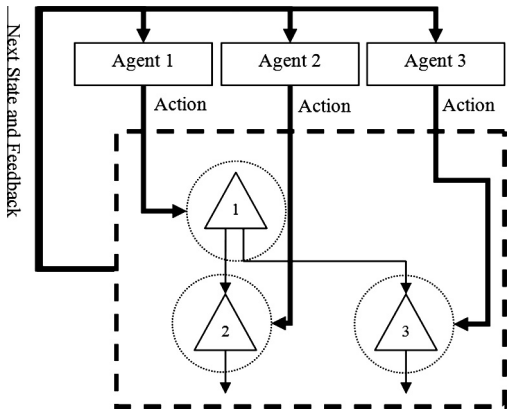

Fig. 7. Multi-agent operations in the fully non-cooperative mode.

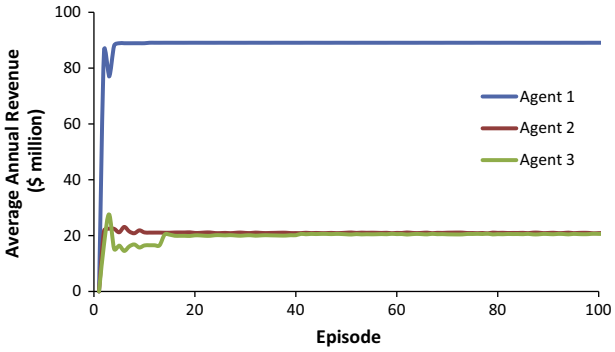

Fig. 8. Learning curves of the three agents in the fully non-cooperative mode.

Fig. 6 shows the learning curves of agent 1, in charge of the {1,3} coalition and agent 2, operating reservoir 2. This figure shows how the two agents develop their best response strategies over time in the competitive environment. Agent 1, has found a policy with a total average annual revenue of $90.18 million for reservoirs 1 and 3, while agent 2 has selected a policy leading to average annual revenue of $43.11 million for reservoir 2. The total obtained benefits in this case is $133.29 million per year which is 82% of the total obtainable benefits in the social planner (full cooperation) mode.

The learning outcomes of other partial coalitions can be evaluated similarly. The obtainable revenue of each reservoir through partial coalitions is report in Table 4.

4.5. Optimization in the non-cooperative mode (singleton coalitions)

In the non-cooperative mode, three independent agents operate the system by making uncoordinated decisions in a competitive environment. Each agent is responsible for only one reservoir (Fig. 7). The learning process is exactly the same as the process explained for partial coalitions in Section 4.4. The only difference is that in this case there are three agents (instead of two) that make

Table 4
Obtainable benefits of each reservoir for different levels of cooperation ($ million).

Optimization model	Coalitions	Optimization model	Reservoir 1	Reservoir 2	Reservoir 3	Total	% Of the total achievable benefits
Three-agent RL	{1}, {2}, {3}	Three-agent RL	89.09	20.89	20.61	130.6	80
Two-agent RL	{1,2} and {3}	Two-agent RL	76.49	50.4	30.47	157.36	97
	{1,3} and {2}	Two-agent RL	73.27	43.11	16.91	133.29	82
	{2,3} and {1}	Two-agent RL	89.91	41.35	15.23	146.49	90
Single-agent RL	{1,2,3}	Single agent RL	77.34	57.01	28.66	163.01	100

decisions simultaneously. The action of agent i is the release from reservoir i and its state is storage in that reservoir. The learning process in this condition also involved 100 episodes, each containing 50 years. The discount factor (γ) was set to 0.7 to ensure convergence.

Fig. 8 indicates the learning curves of the three agents in the non-cooperative mode. Agents 1, 2 and 3 managed to develop best response policies that lead to \$80.09 million, \$20.89 million, and \$20.61 million in average revenue per year at the non-cooperative Nash equilibrium. The overall revenue of the system in this case is \$130.6 million per year which is about 80% of the total obtainable revenue in social planner condition.

Comparison of Figs. 4, 6 and 8 suggests that convergence time is correlated with the size of the state–action space as the fastest convergence occurs in the non-cooperative mode with the smallest state–action space and the slowest convergence occurs in the social planner (fully cooperative) mode with the biggest state–action space.

4.6. Game theoretic allocations

Table 4 shows the average annual obtainable benefit from each reservoir under different levels of cooperation. The maximum total gain is possible under the grand coalition when operation strategies are fully coordinated. In this case, the total obtainable benefit of the system is 20% more than the non-cooperative case in which operation strategies are uncoordinated. Partial coalitions can also result in some incremental benefits, depending on which operators form the coalition. The maximum obtainable benefit through partial coalition occurs under coalition {1,2}, reflecting the lower level of contribution by operator 3 to the grand coalition in comparison to the other two operators.

Looking at the obtainable benefits of each reservoir under different levels of cooperation, one can realize why the social planner's (system's optimal) solution is not stable in practice (Madani and Dinar, 2012; Read et al., 2014). When maximizing social welfare, the central planner is not concerned about the gains of the individuals as the main objective is to maximize the system's benefits regardless of how welfare is distributed among the individuals. In this case, some members of the system might need to sacrifice individual benefits to help with materializing the social planner's (system's) objective. For example, in order to achieve the maximum system benefits, operator 1 should reduce its benefits by \$11.75 million. Therefore, this player has no incentive to participate in the grand coalition. For the same reason, this operator has no incentive to join any partial coalition, making coalitions {1,2,3}, {1,2} and {1,3} infeasible. The only remaining non-singleton coalition is {2,3} which cannot be formed because for this coalition to form, operator 3 has to be willing to accept lower benefits than the status quo. This means that in this example rational operators will not accept the system's optimal solution and will prefer to operate in the fully non-cooperative mode, resulting in the non-cooperative Nash equilibrium.

However, if utility is transferrable, side payments might create cooperation incentives to stabilize the grand coalition. In this case, the main challenge is to find the fair and efficient levels of side-payments. CGT solution methods can address this challenge and help determining the appropriate shares of the agents from the incremental benefits of cooperation. In order to calculate these shares, CGT solutions must not only consider the parties' gains under the status quo, but also their gains under partial coalitions and their levels of contribution to the grand as well as partial coalitions. While collecting this information has been challenging in multi-agent multi-period problems, application of the suggested RL method helped obtaining this information (Table 4), as explained in Sections 4.3–4.5. Thus the CGT method explained in Section 2 can be applied to calculate the fair and efficient allocations of the obtainable benefits under cooperation.

Table 5 shows the suggested shares of the cooperating operators from the grand coalition based on the Nash–Harsanyi, Shapley and Nucleolus methods. The amount of side-payments can be calculated based on these allocation solutions. For example, based on the Nash–Harsanyi solution, operator 2 has to pay \$22.56 (99.9–77.34) million per year to operator 1 and \$2.76 (31.42–28.66) million per year to reservoir 3 to convince them to stay in the grand coalition, resulting in win–win solutions for all parties.

Table 5
Calculated shares of the cooperating parties from the grand coalition based on different cooperative game theory methods (million \$).

Method	Operator 1	Operator 2	Operator 3	Total
Nash–Harsanyi	99.90	31.70	31.42	163.01
Shapley	94.44	43.54	25.04	163.01
Nucleolus	98.67	35.97	28.37	163.01

4.7. Stability assessment

All the allocation solutions presented in Table 5 belong to the core of the cooperative game. So, theoretically they leave no incentive for the operators to depart from the grand coalition to act independently or in partial coalitions. Nevertheless, parties have different preferences over the proposed allocation solutions which have been developed based on different notions of fairness. Therefore, there is a need to evaluate the acceptability (stability) of the allocation solutions to find the best solution in this case. Here, the stability of the allocation solutions is evaluated using three commonly used stability analysis methods.

4.7.1. Plurality

The plurality method (Sheikhmohammady and Madani, 2008) is the easiest method for evaluation of the acceptability of allocation solutions (Dinar and Howitt, 1997; Madani and Dinar, 2012; Madani et al., 2014c). As a social choice rule (Madani et al., 2014b), this qualitative voting method selects the favorite option of the majority as the socially optimal solution. In this example, each player prefers the allocation solution which results in a higher gain. Table 6 shows how the three operators rank the allocation solutions. According to this table, the Nash–Harsanyi solution method is the most popular (potentially acceptable) method with two supporters (operators 1 and 3).

4.7.2. Propensity of disrupt

Socially optimal solutions are not necessarily feasible in practice (Read et al., 2014; Asgari et al., 2014), especially when the stakeholders' powers are heterogonous (Madani et al., 2014c). To address this fact, the stability analysis can be complemented by application of quantitative methods that carefully consider the

Table 6
Application of the plurality rule to the cooperative allocation solutions (lower ranks are more desirable).

	Allocation method	Gain (\$ million)	Rank
Operator 1	Nash–Harsanyi	99.90	1
	Shapley	98.67	2
	Nucleolus	94.44	3
Operator 2	Nash–Harsanyi	31.70	3
	Shapley	35.97	2
	Nucleolus	43.54	1
Operator 3	Nash–Harsanyi	31.42	1
	Shapley	28.37	2
	Nucleolus	25.04	3
	Selected method		Nash–Harsanyi

power distribution of players. Propensity to disrupt (PTD) (Gately, 1974) is a commonly used quantitative method (Straffin and Heaney, 1981; Dinar and Howitt, 1997; Teasley and McKinney, 2011; Madani and Dinar, 2012; Asgari et al., 2014) to evaluate the stability of game theoretic allocations with respect to the cooperating agents' powers in the grand coalition. Based on the PTD concept, an agent has a stronger negotiation power when its contribution to the grand coalition is relatively higher than the contributions of other agents. The PTD of operator i is defined as the ratio of what the other operators would lose if this operator leaves the grand coalition to what this agent will lose if it leaves the grand coalition:

$$PTD_i = \frac{\sum_{j\neq i} u_j^* - v(N-\{i\})}{u_i^* - v(i)} \tag{19}$$

The higher the PTD of an operator, the higher its power and the lower its willingness to cooperate. Table 7 shows the PTD of the operators under different allocation schemes. Based on this table, the Nucleolus method is not stable as it results in relatively high PTDs for all agents. Selection of the more stable solution out of the Shapley and Nash–Haranyi solutions would be challenging based on the PTD method. While Nash–Harsanyi results in lower PTDs for operators 1 and 3, it results in a higher PTD for operator 2. The opposite is true for the Shapley solution. Therefore, based on the PTD concept, both Nash–Harsanyi and Shapley solutions can be selected as potentially stable solutions out of the suggested allocation solutions to the study problem.

4.7.3. Stability index

Stability index, or coefficient of variation of agents' power indices, is another commonly used quantitative indicator of the stability of allocation solutions (Dinar and Howitt, 1997; Teasley and McKinney, 2011; Madani and Dinar, 2012; Madani et al., 2014c; Read et al., 2014). The power index of operator i (α_i) is defined as the ratio of the operator i's loss if it leaves the grand coalition to the summation of other operators' losses when they leave the grand coalition (Shapley and Shubik, 1954; Loehman et al., 1979):

$$\alpha_i = \frac{u_i^* - v(i)}{\sum_{j=1}^{n}(u_j^* - v(j))} \tag{20}$$

The lower the power index for an operator, the lower its tendency for cooperation. The stability index is reflective of the quality of power distribution (Read et al., 2014). A lower stability index shows better (equal) distribution of powers among the cooperating agents. Thus, the lower the stability index under an allocation scheme, the more stable the allocation solution under that scheme. Table 8 shows the power indices of the operators for each allocation solution. Based on the table values, Nash–Harsanyi solution is the most stable solution because it distributes the power equally among the operators, resulting in the stability index of zero.

Based on the stability analysis results discussed in Sections 4.7.1–4.7.3 it can be concluded that the Nash–Harsanyi solution is the most stable among the proposed allocation solutions.

Table 7
Propensity to disrupt (PTD) of the operators under different allocation schemes (lower PTD shows higher willingness to cooperate).

	Allocation method	Gain ($ million)	PTD
Operator 1	Nash–Harsanyi	99.90	0.6
	Shapley	98.67	0.8
	Nucleolus	94.44	2.2
Operator 2	Nash–Harsanyi	31.70	3.8
	Shapley	35.97	2.4
	Nucleolus	43.54	1.3
Operator 3	Nash–Harsanyi	31.42	0.4
	Shapley	28.37	1.0
	Nucleolus	25.04	2.5

Table 8
The power index of each player using different institution (higher power index shows a higher willingness to cooperate).

	Allocation method	Gain ($ million)	Power index
Operator 1	Nash–Harsanyi	99.90	0.3
	Shapley	98.67	0.3
	Nucleolus	94.44	0.2
Operator 2	Nash–Harsanyi	31.70	0.3
	Shapley	35.97	0.5
	Nucleolus	43.54	0.7
Operator 3	Nash–Harsanyi	31.42	0.3
	Shapley	28.37	0.2
	Nucleolus	25.04	0.1

5. Conclusions

This paper proposed a new framework for developing cooperative institutions to increase the efficiency of multi-agent multi-period water management problems. The proposed framework combines reinforcement learning (RL) and cooperative game theory (CGT) to address two of the main weaknesses of the dominant approaches in the literature for maximization and distribution of the benefits of multi-beneficiary water resource systems.

The first weakness is related to common application of social planner (systems) approaches with the objective of maximizing the total benefits of the system regardless of how the benefits are shared among the beneficiaries. While the resulting solutions from these approaches are attractive in theory, they are not easy to implement. This is because these methods assume a perfect cooperation among the stakeholders of the system and ignore the dynamics of decision making at the individuals' level. Therefore, they do not provide the required incentives (for changing behavior and implementing the system's optimal solution) to the individual agents who base their decisions on individual rationality rather than group rationality.

Cooperative game theoretic solutions can complement the conventional social planner solution by providing benefit sharing mechanisms that provide strong incentives to individual decision makers to facilitation cooperation in order to achieve the system's optimal solution. However, obtaining the required information for application of game theoretic solutions is very challenging and can be computationally intensive. This has resulted in the second weakness of the previous game theoretic applications in the water resources literature that had to make simplifying assumptions about the achievable benefits of the parties under different levels of cooperation. In order to fill this information gap, this study proposed application of RL, which helps understanding the behavior dynamics of the operating agents in interactive systems.

Combination of RL and CGT, as suggested by this paper, provides an opportunity to explore feasible coordination policies that maximize the total benefits of the system while taking into account the fair allocation of incremental benefits. To show the applicability of the proposed framework to interactive multi-agent multi-period water resource management problems, the suggested game theory–reinforcement learning (GT–RL) method was applied to a hypothetical multi-operator multi-reservoir system. The presented numerical example was relatively small in this proof-of-concept study, but the developed framework is general and can be applied to larger systems, involving additional decision-making agents, objectives, decision variables, and constraints. With the

existing computational capacity and possibility of running the Q-learning algorithm on parallel processors, GT–RL can solve much more complex problems in a reasonable time.

In this study, reservoir operators were assumed to have identical objectives. However, the proposed framework is applicable to problems with heterogeneous beneficiaries who purse different objectives and might have different power levels. In the numerical example, reservoir inflows were assumed to be deterministic and fixed in different time-steps. Future studies can apply the GT–RL method to solve problems with stochastic or changing (but deterministic) inflows. For such problems, the inflows must change in each iteration with a given probability, so convergence requires more and longer episodes. Here, utility was assumed to be transferrable, making side-payments possible to achieve the coordinated operations which maximize the social welfare. Future studies might consider solving non-transferrable utility problems in which side-payments are not feasible and increasing benefits is possible only through coordination of actions.

Acknowledgements

The authors would like to thank the two anonymous reviewers and the Editor for their constructive comments.

References

Arnold, E., Tatjewsli, P., Wolochowisc, P., 1994. Two methods for large-scale nonlinear optimization and their comparison on a case study of hydropower optimization. J. Optim. Theory Appl. 81 (2), 221–224.

Asgari, S., Afshar, A., Madani, K., 2014. A cooperative game theoretic framework for joint resource management in construction. J. Constr. Eng. Manage. 140 (3), 04013066. http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000818.

Bergez, J., Eigenraam, M., Garica, F., 2001. Comparison between dynamic programming and reinforcement learning: a case study on maize irrigation management. In: Proceedings of the 3rd European Conference on Information Technology in Agriculture, Montpellier, Fr, June 18–20, pp. 343–348.

Castelletti, A., Galelli, S., Restelli, M., 2010. Tree-based reinforcement learning for optimal water reservoir operation. Water Resour. Res. 46 (9), W09507.

Dinar, A., 2001. Scale and equity in water resource development: a Nash bargaining model. Nat. Resour. Model. 14 (4), 477–494.

Dinar, A., Howitt, R.E., 1997. Mechanisms for allocation of environmental control cost: empirical tests of acceptability and stability. J. Environ. Manage. 49, 183–203.

Gately, D., 1974. Sharing the gains from regional cooperation: a game theoretic application to planning investment in electric power. Int. Econ. Rev. 15 (1), 195–208.

Gosavi, A., 2003. Simulation-based Optimization: Parametric Optimization Techniques and Reinforcement Learning. Norwell, Massachusetts, USA.

Harsanyi, J., 1959. A bargaining model for the cooperative *n*-person game. In: Contributions to the Theory of Games 4. Princeton University Press, Princeton, New Jersey, pp. 324–356.

Hiew, K., 1987. Optimization Algorithms for Large Scale Multi-Reservoir Hydropower Systems. PhD Dissertation. Department of Civil Engineering, Colorado State University, USA.

Imen, S., Madani, K., Chang, N.-B., 2012. Bringing environmental benefits into Caspian Sea negotiations for resources allocation: cooperative game theory insights. In: Proceedings of the World Environmental and Water Resources Congress 2012. ASCE, Albuquerque, New Mexico, pp. 2264–2271. http://dx.doi.org/10.1061/9780784412312.228.

Kelman, J., Stedinger, J.R., Cooper, L.A., 1990. Sampling stochastic dynamic programming applied to reservoir operation. Water Resour. Res. 26 (3), 447–454.

Labadie, J.W., 2004. Optimal operation of multireservoir systems. J. Water Resour. Plann. Manage. 130 (2), 93–111.

Lee, J.H., Labadie, J.W., 2007. Stochastic optimization of multireservoir systems via reinforcement learning. Water Resour. Res. 43 (11), W11408.

Loehman, E., Orlando, J., Tschirhart, J., Winstion, A., 1979. Cost allocation for a regional wastewater treatment system. Water Resour. Res. 15, 193–202.

Loucks, D., Dorfman, P., 1975. An evaluation of some linear decision rules in chance-constrained models for reservoir planning and operation. Water Resour. Res. 11 (6), 777–782.

Lund, J., Ferreira, I., 1996. Operating rule optimization for Missouri River reservoir system. J. Water Resour. Plann. Manage. 122 (4), 287–295.

Madani, K., 2010. Game theory and water resources. J. Hydrol. 381 (3–4), 225–238.

Madani, K., 2011. Hydropower licensing and climate change: insights from cooperative game theory. Adv. Water Resour. 34 (2), 174–183.

Madani, K., Dinar, A., 2012. Cooperative institutions for sustainable management of common pool resources: Application to Groundwater. Water Resour. Res. 48 (9). http://dx.doi.org/10.1029/2011WR010849.

Madani, K., Guégan, M., Uvo, C., 2014a. Climate change impacts on high-elevation hydroelectricity in California. J. Hydrol. 510 (14), 153–163.

Madani, K., Read, L., Shalikarian, L., 2014b. Voting under uncertainty: a stochastic framework for analyzing group decision making problems. Water Resour. Manage 28 (7), 1839–1856.

Madani, K., Zarezadeh, M., Morid, S., 2014c. A new framework for resolving conflicts over transboundary rivers using bankruptcy methods. Hydrol. Earth Syst. Sci. 18, 3055–3068. http://dx.doi.org/10.5194/hess-18-1-2014.

Mahootchi, M., 2009. Storage System Management Using Reinforcement Learning Techniques and Nonlinear Models. PhD Dissertation. University of Waterloo, Canada.

Mahootchi, M., Tizhoosh, H.R., Ponnambalam, K., 2007. Opposition-based reinforcement learning in the management of water resources. Proc. IEEE Symp. Approx. Dyn. Program. Reinf. Learn., 217–224.

Mousavi, S.J., Karamouz, M., 2003. Computational improvement for dynamic programming models by diagnosing infeasible storage combinations. Adv. Water Resour. 26 (8), 851–859.

Nakayama, M., Suzuki, M., 1976. The cost assignment of cooperative water resource development: a game theoretic approach. Manage. Sci. 22 (10), 1081–1086.

Nash, J., 1953. Two-person cooperative games. Econometrica 21, 128–140.

Needham, J., Watkins Jr., D., Lund, J., Nanda, S., 2000. Linear programming for flood control in the Iowa and Des Moines Rivers. J. Water Resour. Plann. Manage. 126 (3), 118–127.

Parrachino, I., Dinar, A., Patrone, F., 2006. Cooperative game theory and its application to natural, environmental, and water resource issues: 3. Application to water resources. The World Bank.

Ratner, A., Yaron, D., 1990. Regional cooperation in the use of irrigation water, efficiency and game theory analysis of income distribution. Agric. Econ. 4 (1).

Read, L., Madani, K., Inanloo, B., 2014. Optimality versus stabiloty in water resource allocation negotiations. J. Environ. Manage. 133, 343–354. http://dx.doi.org/10.1016/j.jenvman.2013.11.045.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Annu. Math. Stat. 22 (3), 400.

Sadegh, M., Mahjouri, N., Kerachian, R., 2010. Optimal inter-basin water allocation using crisp and fuzzy Shapley games. Water Resour. Manage. 24 (10), 2291–2310.

Schmeidler, D., 1969. The nucleolus of a characteristic function game. SIAM J. Appl. Math. 17, 1163–1170.

Shapley, L.S., 1953. A value for *n*-person games. Annu. Math. Stud. 28, 307–318.

Shapley, L., Shubik, M., 1954. A method for evaluating the distribution of power in a committee system. Am. Polit. Sci. Rev. 48, 787–792.

Sheikhmohammady, M., Madani, K., 2008. Bargaining over the Caspian Sea – the largest lake on the earth. In: Proceeding of the 2008 World Environmental and Water Resources Congress.

Straffin, P.D., Heaney, J.P., 1981. Game theory and the Tennessee Valley Authority. Int. J. Game Theory 10, 35–43.

Sutten, R., Barto, A., 2000. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Szidarovszky, F., Duckstein, L., Bogardi, I., 1984. Multiobjective management of mining under water hazard by game theory. Eur. J. Oper. Res. 15 (2), 251–258.

Teasley, R., McKinney, D., 2011. Calculating the benefits of transboundary river basin cooperation: Syr Darya Basin. J. Water Resour. Plann. Manage. 137 (6), 481–490.

Tejada-Guibert, J.A., Stedinger, J.R., 1990. Optimization of value of CVP'S hydropower production. J. Water Resour. Plann. Manage. 116 (1), 52–70.

Thomas, H., Watermeyer, P., 1962. Mathematical Models: A Stochastic Sequential Approach. Harvard University Press, Cambridge, MA.

Wang, L., Fang, L., Hipel, K., 2008. Basin-wide cooperative water resources allocation. Eur. J. Oper. Res. 190 (3), 798–817.

Watkins, C., 1989. Learning from Delayed Rewards. PhD Thesis. University of Cambridge, England.

Watkins, C., Dayan, P., 1992. Q-learning. Mach. Learn. 8, 279–292.

Willis, R., Finneya, B.A., Chu, W., 1984. Monte Carlo optimization for reservoir operation. Water Resour. Res. 20 (9), 1177–1182.

Young, H.P., Okada, N., Hashimoto, N., 1982. Cost allocation in water resources development. Water Resour. Res. 18 (3), 463–475.