殿ダム貯水池における選択取水設備の最適運用の検討

OPTIMAL OPERATION OF THE SELECTIVE WITHDRAWAL SYSTEM IN TONO DAM RESERVOIR

矢島　啓[1,2]・Andrea Castelletti[2,3]・Rodolfo Soncini-Sessa[3]
Hiroshi YAJIMA, Andrea CASTELLETTI and Rodolfo SONCINI-SESSA

[1]正会員　博(工)　鳥取大学准教授　工学研究科社会基盤工学専攻（〒680-8552 鳥取市湖山町南4-101）
[2]西オーストラリア大学 Centre for Water Research 特任研究員（35, Stirling Hwy, Crawley WA 6009, Australia）
[3]非会員　ミラノ工科大学 電子・情報技術学科（Piazza Leonardo da Vinci, 32, I-20133, Milano, Italy）

A reinforcement learning approach was developed and applied to design efficient management policies for selective withdrawal system with the purpose of meeting established water quality/quantity targets both in-reservoir and downstream. Structured design of experiment simulations was performed by a 1D coupled hydrodynamic-ecological model (DYRESM-CAEDYM) to generate a learning dataset over which a daily management policy was trained using a fitted-Q algorithm based on extremely randomized trees. The approach was demonstrated on the management of Tono Dam reservoir, which is now under construction. Preliminary results indicated that a potential great control over reservoir limnology and release quality can be gained by effectively exploiting - through the management policy - the operational flexibility provided by the selective withdrawal structures.

Key Words : *Tono dam reservoir, selective withdrawal system, optimal operation, fitted-Q learning, DYRESM-CAEDYM*

1．はじめに

日本では，1970年代後半以降に建設された多くのダムにおいて，選択取水設備（以下，SWS）が設置されている[1)]．また，その運用は，常時表層取水としているものが多い[2)]．これは冷水放流を防ぐためであるが，水量だけでなく水質においても高度な管理が求められている現在においては，水質管理も含んだ最適運用を行うことは容易ではない．また，矢島らが検討したように，SWS運用の違いは，規模の小さな貯水池の水質に大きな影響を与える[3)4)]．そのため，貯水池管理において，水量・水質の両面から最適化した運用を行うには，高度な最適運用システムが必要になると考えられる．

ダム貯水池を対象とした最適操作システムは，現在もなお活発に研究が行われている分野[5)]である．その一つとして，Stochastic Dynamic Programming (SDP) は貯水池操作において最も適したシステムの一つである[6)]．SDPは，連続した意志決定過程として運用ポリシーを定式化することに基づいており，システムの基本は，最適操作を探索する関数となる利得関数を定式化することにある．水資源システムの最適化問題においては，1960年代から(deterministic) Dynamic Programming (DP)が適用されている[7)]．それ以来，貯水池システム，特に発電の最適化問題にシステムは適用されてきた[8)9)]．また，1980年代の初めには，複数のダムを対象にした問題に対して，SDPに拡張されたシステムが検討されてきた[10)11)]．

SDPは広く水資源の最適システムに適用されているものの，実際の複雑なシステムに適用するのに２つの大きな問題がある．一つは，状態変数の数や運用ポリシーの決定空間が大きくなるに従い，計算時間が指数関数的に増大することである[12)]．もう一つは，システムの状態変化を表現できる陽的なモデルが必要であることである[13)]．これら2つは，水質まで考慮した貯水池の最適運用システムにとって，容易に解決できない問題となる．計算時間問題の克服のためには，Incremental DP[14)] などが

考えられたが，これらは決定論的問題に適用されてきた．また，モデル化については，通常，SDPは状態変数にもとづいた操作決定の効果を評価するための物理モデルが必要であるが，経験から学んだことをもとに評価する手法についての研究が始まった．それらは強化学習と呼ばれ，SDPの概念とシミュレーションを通した確率的近似と価値関数の近似に基づいた手法である[15)16)]．

経験による学習は，実際にダム貯水池操作を変更することにより行う方法とモデルを用いて行う方法が考えられる．ただし，貯水池操作の場合は，何らかの水理・水質を予測できるモデルを用いたオフライン方式の学習が現実的である．Castellettiらは，部分的にモデルを使用しない強化学習の一つであるQ学習法を，多目的に利用されている湖の運用問題に適用した[17)]．近年では，新たな手法としてfitted-Q学習と呼ばれ，Q学習と価値関数の関数近似を組み合わせた手法が提案されている[18)]．このfitted-Q学習によれば，関数に含まれるパラメータの同定に多くの時間を必要とするこれまでの手法に対して，決定木による近似を用いたパラメータに依存しない手法は，複雑な価値関数の形状を同定するのに有利となる．

本論文では，千代川水系袋川の鳥取県鳥取市国府町殿地内に建設中である殿ダム貯水池を対象に，水量と水質に関する複数の貯水池運用目標を同時に最適化するという複雑な問題解決のため，SDPでは計算時間の観点から現実的でなく，他の手法では，連続する状態変数を取り扱うのが難しいため，現状での唯一の適用手法と考えられるfitted-Q学習を用いた選択取水設備の運用の最適化に関する基礎的な検討を行う．

２．強化学習とfitted-Q学習

(1) 強化学習[19)]

強化学習は，数値化された報酬信号を最大にするために，何をすべきか（どのようにして状況に基づく動作選択を行うか）を学習する．すなわち，この学習法は，機械学習や人工ニューラルネットワークで学習されるような教師あり学習とは異なり，どの行動をとればいっそうの報酬に結びつくかを自ら学習する．

学習と意志決定を行う「エージェント」とこのエージェントが相互作用を行う対象である「環境」は，離散的な時間ステップ $t = 0, 1, 2, 3$ …の各々において相互作用を行う．各時間ステップt において，エージェントは何らかの環境の状態（state）の表現 $x_t \in S$ (Sは可能な状態の集合)を受け取り，これに基づいて行動 $u_t \in A(x_t)$を選択する ($A(x_t)$は状態において選択することが出来る行動の集合である)．1ステップ時間後に，エージェントはその行動の結果として数値化された報酬 $g_{t+1} \in R$ (Rは可能な報酬の集合)を受け取り，新しい状態 x_{t+1}にいることを知る（**図-1**参照）．エージェントは，自分の経験に基づき，どのように方策を変更することで，最終的に受け取る報酬の総量を最大化することができるかを学習する．

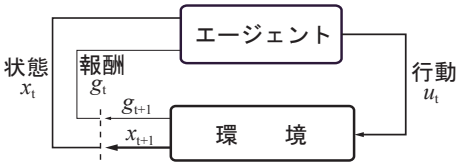


図-1 強化学習におけるエージェントと環境間の相互作用

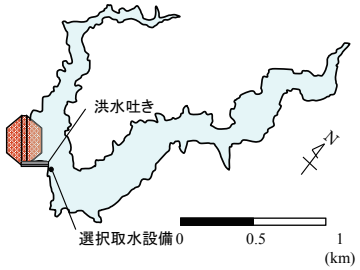


図-2　殿ダム貯水池平面図

(2) fitted-Q学習

fitted-Q学習は，他の強化学習と異なり，システムの陽的なモデルを必要としない学習法である．学習に必要な経験は，一連の次式で表されるデータセットである．

$$F = \left\{ \left\langle x_t^l, u_t^l \; x_{t+1}^l \; g_{t+1}^l \right\rangle, \quad l = 1, \cdots, \#F \right\} \tag{1}$$

ここで，$\#F$は4種類のデータセットの個数である．fitted-Q学習は，最適政策（コントロール）を求めるための近似関数を求めるが，以下のようなアルゴリズムによる．状態x_tで，行動u_tをとり，それ以降の最適政策をとったときの評価時間hに関する積算報酬の期待値もしくは「最適行動価値関数」を$Q_h^*(x_t, u_t)$とする．これは，再帰的に，

$$Q_h^*(x_t, u_t) = g(x_t, u_t) + \gamma \max_{u_{t+1}} Q_{h-1}^*(x_{t+1}, u_{t+1}) \tag{2}$$

と定義される．ここで，γは減衰係数を表し，将来の報酬がどの程度行動価値に影響を与えるかを制御し，通常，1より小さい値を持つ．関数$Q_{h-1}^*(x_t, u_t)$が既知の場合は，$Q_h^*(x_t, u_t)$はすべての状態変数の組み合わせ$<x_t^l, u_t^l>$, $l = 1$, …, $\#F$に対して学習を行う．学習が終了すると，状態変数とコントロール変数の総組み合わせである $s_x \times s_u$ の空間に対して，Q_h^*が連続となるように回帰分析を行う．このように，fitted-Q学習は，バッチ型の強化学習と考えることができ，すべてのデータセットFに対してバッチモードで計算を行い，Q_h^* とQ_{h-1}^*の差分がある基準以下になった時に計算を終了させる．

３．殿ダム貯水池への適用

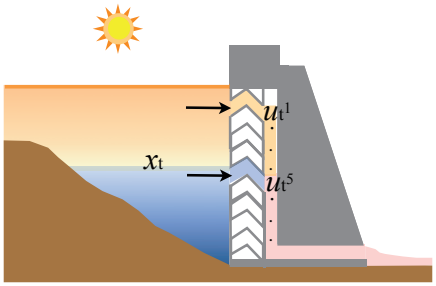


図-3 貯水池の状態変数とSWSのコントロールベクトル

(1) 殿ダムの概要

殿ダム貯水池は，平成23年完成予定の堤高75m，堤頂長294m，総貯水量1240万m^3の規模を有する国土交通省直轄のロックフィルダムである．**図-2**に貯水池の平面図を示す．流域面積は38.1km^2，湛水面積は0.64km^2の比較的小規模な貯水池である．取水放流設備としては，取水レベルが常時満水位と同高なゲートレス洪水吐き，発電及び利水を目的とした選択取水設備，緊急時対応の予備設備として低水放流設備が計画されている．選択取水設備は，常時満水位と最低水位間の19.8mに連続サイフォン形式のものが計画されており，表層と深水層の取水を組み合わせた異高同時取水を含んだ複雑な取水を行うことができると考えられている．

これまでに行われた検討[4]では，平水時は，異高同時選択取水を行うことにより，下層の貧酸素化や表層での植物プランクトンの増殖を抑え，冷水放流の可能性も低くすることができるとともに，出水時には，常時表層取水を行うことが望ましいことが明らかにされている．

(2) 選択取水設備の最適制御における変数

選択取水設備の最適制御を行うに当たって，コントロールベクトルu_tは，**図-3**に示すように各選択取水設備からの取水量となる．すなわち，

$$u_t = \left| u_t^1, \cdots, u_t^n \right| \tag{3}$$

である．ここで，u_t^nはn番目のSWSからの時間tとt+1の間の放流量，nはSWSの取水口の数である．

また，最適制御における状態変数x_tの代表的なものとして，貯水池内の水位や水温などが考えられる．ただし，本研究では鉛直1次元の水質予測モデルDYRESM-CAEDYM[3]を用いるため，貯水池内の水質項目を状態変数とするときには，その参照水深（あるいは水位）のみを規定することになる．

(3) 殿ダム貯水池の最適制御における変数の設定

本研究においては，まず，手法のシステムへの適用を試みることが一番の目的であることから，計算に必要な変数をできるだけ少なくすることが望ましい．これまでの研究[3) 4)]では，選択取水設備の運用法として，水深3mと13mの2箇所から同時に混合取水する異高同時取水が長期の水質保全からは望ましいという検討結果が得られている．そこで，この異高同時取水の運用をベンチマークとして比較できるよう，取水のコントロールベクトルとして水深3m，13mの2箇所のみを考慮し，どちらか一方からまたは両方から取水する制御を試みた．ただし，貯水位の変動に応じて取水レベルも変化するため，実際には複数の取水口をモデル化していることになる．

さらに，DYRESM-CAEDYMを用いた検討では，洪水時に洪水吐敷高より水位が高くなった場合は，放流量式に応じた洪水吐からの放流量を計算した．また，渇水時のように水位が大きく低下した場合は，物理的に取水が不可能な取水口が存在する．そのため，水深3mと13mの取水口の他に，通常取水範囲の最下層の取水口および堆砂域に存在する低水放流設備を加え，合計4つの変数を取水に関するコントロールベクトルu_tとした．

また，状態変数については，取水レベルの水深3mと13mは，表水層と深水層に位置する．そのため，それぞれの水質が大きく異なり，状態変数として設定することが適当と考えられた．そこで，この2水深における水温，全浮遊物質量Total Suspended Solid (TSS)および貯水位の合計5変数を状態変数x_tとして用いた．

(4) 最適化問題の目標設定

最適化のためには，まず，政策ポリシーの価値関数（最適化指標）を設定する必要がある．本研究では，殿ダム貯水池にコンフリクト関係が生じる可能性のある多目的間の最適化を目指すため，次の3つをステップ最適化指標とした．まず，1番目の指標J^{sed}は，貯水池内の堆砂防止と下流河川における土砂環境の配慮から，できるだけ多くのTSSを放流することを目標としている．また，2番目の指標J^{env}は，下流河川における水温環境をダム建設前の自然な状態にできるだけ保つため，ダム貯水池への流入水温と放流水温を等しくすることを目標としている．さらに，最後の指標J^{irr}は，灌漑用水などの下流河川の水需要を満たすことを目標としている．

$$J^{sed} = \underset{\varepsilon_1 \cdots \varepsilon_{h-1}}{E} \sum_{t=0}^{h-1} TSS_{t+1}^{out} \tag{4}$$

$$J^{env} = \underset{\varepsilon_1 \cdots \varepsilon_{h-1}}{E} \sum_{t=0}^{h-1} \left(T_{t+1}^{out} - T_{t+1}^{in} \right)^2 \tag{5}$$

$$J^{irr} = \underset{\varepsilon_1 \cdots \varepsilon_{h-1}}{E} \sum_{t=0}^{h-1} \left(w_t - r_{t+1} \right)^{1.3} + \tag{6}$$

ここで，tは時間ステップを表し，hは最適化指標の期待値を評価する時間ステップ（365日）である．TSS^{out}は，全設備からの放流水の平均TSS濃度（各放流量を加重平均），T^{out}は，全設備からの放流水の平均水温（各放流量を加重平均）である．また，T^{in}は，流入する2河川の各流量による加重平均を行った平均水温である．さらに，wは全放流量，rは下流における正常流量である．ただし，正常流は，5月3日～5月4日は0.694 m^3/s，5月5日～6月1日は0.795 m^3/s，6月2日～9月2日は0.698 m^3/s，9月3日～9月4日は0.450 m^3/s，9月5日～5月2日0.349 m^3/sと下流の水

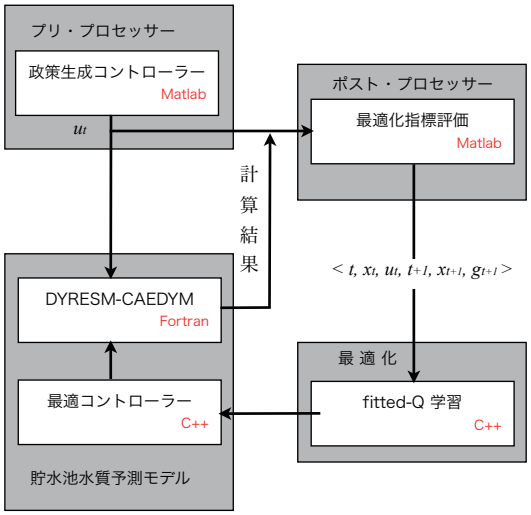


図-4 殿ダムにおける最適制御システムの構築ツール

需要に応じて年間を通じて変動する．また，(6)式における+は，右辺の括弧の中が正の場合のみ考慮することを表している．

最終的な最適制御の目標においては，3指標 J^{sed}，J^{env}，J^{irr}の重み付けが必要である．ここでは，試行的にそれぞれの重みを0.1，0.1，0.8とした．また，これら重み付けの係数と3指標の右辺のべき乗項の係数は，最終的に得られた結果に影響を及ぼす．これについては，ダム貯水池管理者あるいは関係者との合意形成による係数の同定が必要となる．

(5) 殿ダム貯水池における選択取水設備の最適制御システムの構築

最適制御システムの全体構成は，**図-4**に示すようである．ここで，まず，学習データセット$\left\langle x_t^l, u_t^l\ x_{t+1}^l\ g_{t+1}^l \right\rangle$が必要となる．そのため，これまでの研究[3)4)]で用いてきた1990年から1994年までの5年間の殿ダム貯水池への流入量・水質データおよびダム地点における気象データを使用した．また，SWSの制御ポリシーを作成するため，外部プリプロセッサーとしてMatlab（MathWorks社製ソフトウェア）を用いた$<u_t^l>$の作成を行った．その後，それらを入力条件として，DYRESM-CAEDYM（Fortran言語で記述）の計算を実行し，得られた計算結果から，外部ポストプロセッサーとしてMatlabを用いて前節に示した3つの最適化指標の評価を行い，fitted-Q学習に必要な学習データを作成した．fitted-Q学習終了後は，全空間 $s_x \times s_u$ に対する補間を行った最適コントローラー（C++言語で記述）を作成し，DYRESM-CAEYDMとのカップリングを行い，一連の計算として自動で最適制御を行いながらのDYRESM-CAEYDMの計算を可能とした． ただし，最適制御における選択取水設備の運用は，1日1回午前8時の状態変数から得られたコントロール政策に従い変化させ，翌日の同時刻までその政策を継続させた．

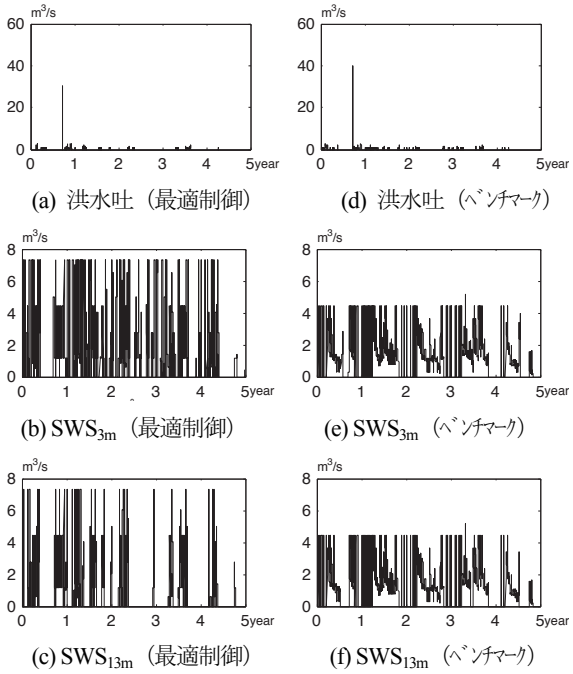


図-5 最適制御およびベンチマーク計算における取水量
（ただし，SWSの添え字が取水水深を表す）

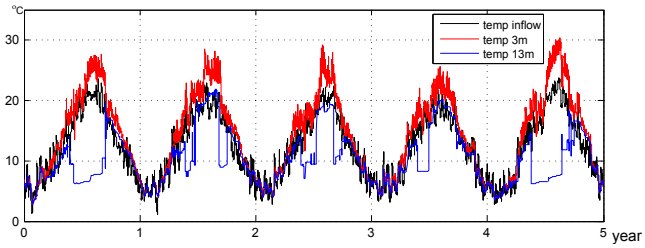
4．Fitted-Q学習による最適操作の評価

(1) 最適操作の概要

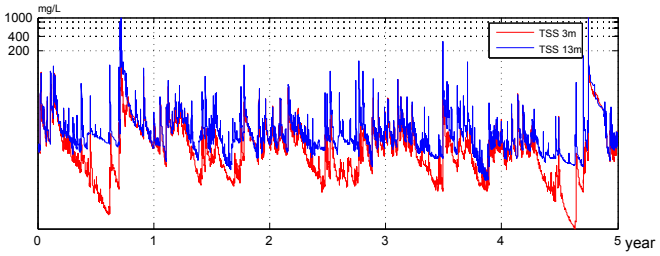
3章で述べたように，最適操作と比較するため，原則，水深3mと13mから同量を混合取水する方式をベンチマーク計算とした．**図-5**に計算を行った1990年から1994年までの5年間の，洪水吐，水深3mおよび13mの取水口からの取水(放流)量を示している．最適制御では，ベンチマーク計算より多くの放流量を選択取水設備から放流していることが分かる．特に，水深3mの取水口から取水している時間が長い．また，ベンチマーク計算は，取水量が最適制御より少ないため，貯水池内の水位が高くなり，出水時に洪水吐からの放流量が大きくなっている．

選択取水設備の操作状況を確認したところ，最適制御において水深3mと13mを同時に使用した異高同時取水は5年間の計算時間中の20.0%，水深3mだけを使用した取水は29.9%，13mだけを使用した取水が6.7%，それ以外が43.4%となっていた．

図-6に貯水池への流入河川水の水温および貯水池内の状態変数の変化を示す．同図(a)より，水深13mより3mの水温の方が流入河川水の水温と近似していることが明らかである．そのため，水温に関する最適化指標J^{env}に関しては，水深3mで取水するのが望ましい．また，同様に同図(b)から，土砂の最適化指標J^{sed}の観点からは，TSSの高い水深13mから取水するのが望ましい．従って，最終の最適目標において，2つの最適化指標はコンフリクト関係にあり，それぞれの最適化指標の重み付けが最適運用に影響を与えることが明らかになった．

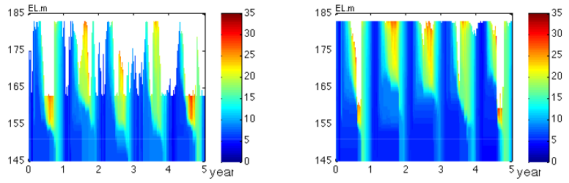


(a) 流入水温と貯水池内水深3mと13mの水温



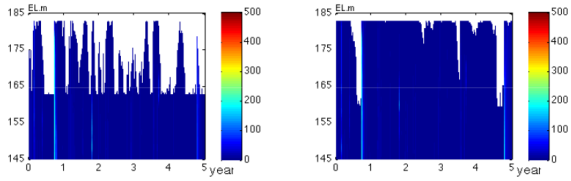
(b) 貯水池内水深3mと13mにおけるTSS

図-6　流入水温と最適運用における状態変数の推移



(a) 最適制御　　(b) ベンチマーク計算

図-7　貯水池内の水温比較（℃）



(a) 最適制御　　(b) ベンチマーク計算

図-8　貯水池内のTSS比較（mg/L）

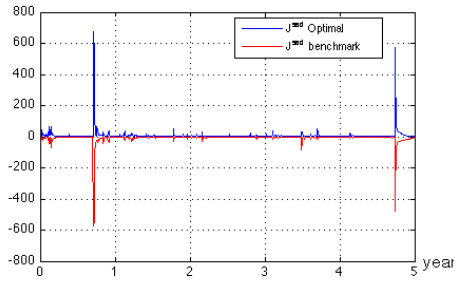
(2) 最適操作による貯水池内の水質に与える影響

図-7および**図-8**に，計算期間5年間の貯水池内の水温およびTSSの分布を示す．これらから，最適操作では，水量に関する最適化指標を改善するため，放流量を高める運用を行うとともに，J^{irr}に差がない複数の政策に対しては，放流量がランダムに決定される．その結果，常に貯水池水位が低い．また，水温分布については，最適操作を行っている方が放流量が大きいため，表層における温水層の容量が少ないことが予想されたが，**図-7**において明瞭な差はみられなかった．また，**図-8**から，貯水池内の土砂環境は，最適制御とベンチマーク計算で大きな違いはみられなかった．

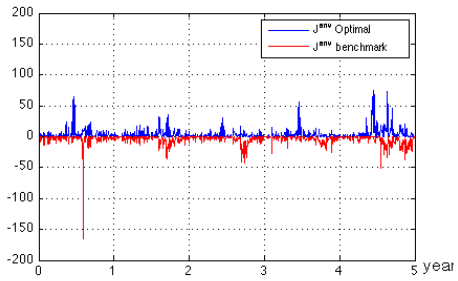
(3) 最適化指標の評価

3つの最適化指標に関する最終価値を**表-1**に示す．この表から，土砂に関する指標J^{sed}は最適運用に関するものが少し良いが，全体的に大きな差がみられない．そこで，それぞれの最適化指標の時間ステップごとの変化を考察する（**図-9**参照）．同図(a)に示すように，土砂に関する最適化指標J^{sed}は，最適運用とベンチマーク計算で差がみられない．これは，貯水池内のTSSが上昇するのは洪水直後であり，1日1回のSWSの運用変更では，貯水池内における濁水の貫入層から効率よく濁水を取水することが難しいためであると考えられる．

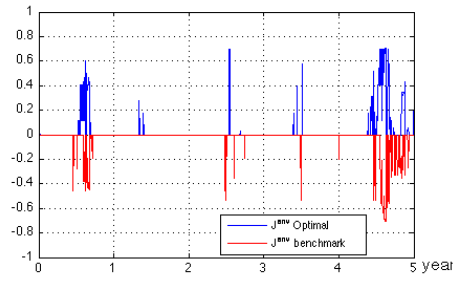
また，水温に関する最適化指標J^{env}は，最終価値では最適運用が2.1℃で，ベンチマーク計算の2.0℃よりも悪い．しかし，同図(b)に示すように，計算初年度の1990



(a) 土砂に関する時間ステップの最適化指標J^{sed}の変化



(b) 水温に関する時間ステップの最適化指標J^{env}の変化



(c) 水量に関する時間ステップの最適化指標J^{irr}の変化

図-9　最適化指標 J^{sed}，J^{env}，J^{irr}の時間変化
（ベンチマーク計算に対する値はマイナス表示）

表-1　3つの最適化指標の最終価値

最適化指標	最適運用	ベンチマーク計算
J^{sed} ↑	6.86 g/m^3	6.38 g/m^3
J^{env} ↓	2.1℃	2.0℃
J^{irr} ↓	0.03 m^3/s	0.03 m^3/s

注）最適化指標右に示す矢印は，最適化にともない指標が向かう方向を表している．

年の夏季の渇水時に，ベンチマーク計算では流入水温から大きく離れた放流を行っているが，最適運用ではそれを防ぐことができている．ただし，平常時は最適運用の方が時間ステップごとの最適化指標が大きくなっているため，最終価値での評価が悪かったものと考えられる．

さらに，同図(c)に示すように，水量に関する最適化指標J^{qr}は，最終価値とともに，時間ステップ毎の差もあまりみられない．これは，ダム計画自体が10ヶ年第1位相当の渇水時に対応した利水計画であるため，実際は，この指標があまり有効に働いていないと考えられた．

５．おわりに

ダム貯水池における選択取水設備の最適運用を目指し，殿ダム貯水池を対象に，強化学習法を発展させたfitted-Qアルゴリズムの適用に関する基本的な検討を行った．ここで，本研究で得られた主要な結論をまとめる．

1) Fortran言語で記述された1次元貯水池水理水質予測モデルDYRESM-CAEDYMとfitted-Q学習から得られC++言語で記述された最適化ポリシーをダイレクトにカップリング計算させながら，選択取水設備の最適運用を行うシステムを構築することができた．
2) 土砂と水温と水量の3つの観点からの最適化指標を作成した検討を行ったが，指標に含まれる係数や個々の指標の重み付けは，最終の最適運用に大きな影響を与えることが明らかとなった．
3) 作成した最適運用で得られた最終価値は，ベンチマーク計算で得られたものと大きな差がないものであった．これは，運用ポリシーの変更が1日1回ということに原因があると考えられた．

以上の成果を踏まえ，今後は，選択取水設備の運用ゲートの数を増やしたより現実に近い検討を行うと共に，運用ポリシー変更の時間間隔の短縮や，最適化ポリシー関数における状態変数，さらに，最適化指標に含まれる係数や重み付けについて検討を深めていく予定である．

謝辞：本研究を行うにあたり，国土交通省中国地方整備局殿ダム工事事務所からは必要なデータな提供を頂いた．また，システム構築にあたり，ミラノ工科大学助手Enrico Weber氏および修士学生Giovanni Garbarini氏の協力を得た．ここに記して謝意を表す．本論文のCentre for Water Research参照番号は，2326-HYである．

参考文献

1) 川崎秀明:ダム技術の動向と課題, ダム日本, No.700, pp.21-33, 2003.
2) 吉田延雄・中村徹:選択取水設備の運用効果について，平成11年度ダム水源地環境技術研究所所報, pp.30-37, 2000.
3) 矢島ら：異高同時選択取水によるダム貯水池の水質保全効果に関する研究，水工学論文集，第49巻，pp.1135-1140, 2005.
4) 矢島ら：選択取水方式がダム貯水池の長期･短期の水質保全に与える影響に関する研究，水工学論文集，第50巻，2006.
5) Labadie, J.: Optimal operation of multireservoir systems: State-of-the-art review, Journal of Water Research Planning and Management - ASCE, vol.130 (2), pp.93–111, 2004.
6) Soncini-Sessa, R., A. Castelletti, and E. Weber: Integrated and participatory water resources management. Theory, Elsevier, Amsterdam, 2007.
7) Hall, W., and N. Buras: The dynamic programming approach to water resources development, Journal of Geophysical Research, vol.66 (2), pp.510–520, 1961.
8) Hall, W., W. Butcher, and A. Esogbue: Optimization of the operation of a multi-purpose reservoir by dynamic programming, Water Resources Research, vol.4 (3), pp.471–477, 1968.
9) Trott, W., and W. Yeh: Optimization of multiple reservoir systems, Journal of the Hydraulic Division ASCE, vol.99, pp.1865– 1884, 1973.
10) Yakowitz, S.: Dynamic programming applications in water resources, Water Resources Research, vol.8 (4), pp.673–696, 1982.
11) Yeh, W.: Reservoir management and operations models: a state of the art review, Water Resources Research, vol.21 (12), pp.1797–1818, 1985.
12) Bellman, R.: Dynamic Programming, Princeton University Press, Princeton, 1957.
13) Bertsekas, D., and J. Tsitsiklis: Neuro-Dynamic Programming, Athena Scientific, Boston, 1996.
14) Larson, R.: State Incremental Dynamic Programming, American Elsevier, New York, 1968.
15) Kaelbling, L., M. Littman, and A. Moore: Reinforcement Learning: a survey, Journal of Artificial Intelligence Research, vol.4, pp.237–285, 1996.
16) Gosavi, A.: Simulation-based optimization: parametric optimization techniques and reinforcement learning, Kluwer Academic Publishers, Boston, 2003.
17) Castelletti, A., G. Corani, A. Rizzoli, R. Soncini-Sessa, and E. Weber: A reinforcement learning approach for the operational management of a water system, in Proceedings of IFAC Workshop Modelling and Control in Environmental Is-sues, August 22-23, Elsevier, Yokohama, 2001.
18) Ernst, D., P. Geurts, and L. Wehenkel: Tree-based batch mode reinforcement learning, Journal of Machine Learning Research, vol.6, pp.503–556, 2005.
19) Sutton, R. S. and A.G.Barto (三上貞芳・皆川雅章訳): 強化学習，森北出版株式会社，2000.

（2009. 9. 30受付）