



Implicit stochastic optimization for deriving operating rules for a multi-purpose multi-reservoir system

Safayat Ali Shaikh¹ · Tapas Pattanayek¹

Received: 21 October 2021 / Accepted: 2 August 2022 / Published online: 20 August 2022
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

Abstract

In this study, optimal operating policies for a multi-purpose multi-reservoir system have been derived using Implicit Stochastic Optimization (ISO) method. In order to do so, four models: (1) all possible regression, (2) stepwise regression, (3) decomposition and (4) simulation, have been proposed that deals with the statistical significance as well as the physical significance of the presence of independent variables in a particular regression model. The first two models deal with the variable selection based on statistical significance. In the decomposition models, operation problem is divided into sub-problems to generate smaller subsets for regression retaining variables of physical significance. The simulation model examines the performance and the predictive efficiency of different regression models. Fifty six years historical inflow data (1962–2017) have been used for this analyses. These four models along with 1000 years monthly synthetic inflow sequences, generated by the multivariate gamma AR(1) model, are applied to the Damodar Valley (DV), a multipurpose multi-reservoir system in India. The effect of variability in those inflow values on the release decisions are taken into account implicitly and the results are presented. Analysis of result indicates that decomposition model with less number of predictor variables is the most preferred model for DV system.

Keywords Regression analysis · Implicit stochastic optimization · Simulation · Reservoir operation · Optimal operating policies

Introduction

In the deterministic analysis, the reservoir operation problem is solved for a particular (known) inflow sequence, among numerous other possible sequences. Hence, the effect of variation in the inflow pattern is not reflected in the solution obtained from the deterministic analysis and it represents only a small part of a more wide range of possible reservoir behaviors. In the implicit analysis, the randomness of the inflows to the reservoir, which makes the reservoir operation a stochastic process, is considered indirectly by solving the operation problem for a number of possible sequences of these random inflows, in a deterministic frame work. This results in a number of release sequences and a regression

model is used with these sequences to obtain an average operating policy.

The major issue in the implicit analysis is the selection of a proper regression model, particularly regarding the selection of specific independent variables to be used in the regression equation, from a number of candidate variables. This selection poses a formidable task for a multireservoir system where a number of combinations of the independent variables are possible. In general, there is no specific criteria for selecting a particular regression model and usually the selection is based on the significance of the regression coefficients in terms of certain statistical parameters.

In this study, a generalized approach is proposed here for the selection of appropriate variables in the regression in the implicit stochastic Dynamic Programming (DP) framework. The approach selects the appropriate model from the judgements based on statistical and physical significance of the presence of a variable in the regression model and performance based on objective function values through simulation. In the regression analysis fundamental contribution have been made by introducing new concepts: (1)

✉ Safayat Ali Shaikh
safayat@aliah.ac.in

Tapas Pattanayek
pattanayek@aliah.ac.in

¹ Department of Civil Engineering, Aliah University, Kolkata, India

all possible combination of independent variables and (2) stepwise inclusion and elimination of independent variables in regression equations, in order to determining strength and character of relationship between optimal storage state (dependent variable) with initial storage states and inflow (independent variables) in reservoir management.

Literature survey

The implicit approach was first suggested by Hall and Howell (1963). As stated by McKerchar (1975) this approach makes one important assumption that the average operating policy obtained is equivalent to the policy obtained from an explicit stochastic dynamic programming model, which by definition optimizes the expected value of the objective function. Although no mathematical proof is available, this assumption appears reasonable for the case where the loss function is smooth and convex. Hall and Howell (1963) suggested that as the historic inflow records are of limited length, synthetic inflow sequence which resemble the historic data in terms of certain statistical parameters, could be used in deriving the average operating policy.

Hall and Howell (1963) used a deterministic dynamic programming to find the optimal pattern of release decisions for known sequences of inflow data and combines these release data into a mean operating policy using a regression model. The variability of inflow has been considered implicitly. Young (1967) used the implicit approach to determine the operating rules for the nonseasonal operation of a single reservoir. Minimization of the losses associated with the failure to supply target outputs was used as the objective function and the optimal releases were determined for a simulated sequence of inflow data. A multiple regression technique was used to relate the optimal releases as a function of the storage and inflows of the current period. McKerchar (1975) extended Young (1967) model to determine the monthly release policies of a two reservoir system. A multivariate streamflow model was used to generate synthetic sequences of inflows. In the multiple regression analysis, he considered optimal monthly releases as a linear function of the two storage only (inflow variables were not considered) and obtained twelve sets of equations, one for each month. Trott (1979) used the implicit approach to determine the real time operating rules for a multireservoir system. A Linear Programming–Dynamic Programming model was used to determine the optimal operating policies which maximize the on-peak energy generation. Historical inflows were used in place of generated data. He considered release as a linear function of storage, previous periods inflow, estimated current periods inflow and releases from other reservoirs in the system. A two-stage least square technique was used for the multiple regression analysis.

Bhaskar and Whitlatch (1980) used the implicit approach for the optimal operation of a single reservoir and emphasized that the policies derived from the regression analysis should be verified for their performance through simulation. They have used both the linear and non-linear regression model where the storage and inflows of the current period and previous periods were used as the independent variables. A stepwise regression procedure was adopted to determine the regression coefficients. Karamouz and Houck (1982) examined the applicability of the operating rules obtained through regression. They proposed an iterative technique to refine the regression coefficients where the DP-regression–simulation cycle was executed repeatedly. They have used the technique for the annual and monthly operation of a single reservoir. Karamouz et al. (1992) applied ISO schemes through three step cycle to improve the initial operating rules for a multireservoir system. The optimal solutions are then analyzed in a regression procedure to obtain a set of operating rules.

Celeste et al. (2009) used ISO to determine monthly operating rules for a reservoir system located in the semiarid Northeast of Brazil. ISO employs a deterministic optimization model to find optimal reservoir allocations under several possible inflow scenarios. Liu et al. (2014) derived operating rules using a simulation-based optimization method in the context of implicit stochastic optimization of China's Three Gorges Reservoir system. Parameter uncertainty for reservoir operating rules is analyzed using Linear Regression (LR) and Bayesian simulation (BS) and established that LR performed less than the Celeste et al. (2016) assesses the suitability of ISO against Stochastic Dynamic Programming (SDP) to set up reservoir release policies and established the performance of both SDP and ISO is superior to that of Standard Operating Procedure followed and close to that of perfect-forecast deterministic optimization (PFDO). Furthermore, the simple ISO shows to perform similarly to the more complex SDP. Moreira and Celeste (2017) applies ISO to develop monthly operating rules using the forecast of the mean inflow for a future horizon instead of the current-month inflow. With a hundred different 100-year monthly synthetically generated inflow scenario release policies derived using SDP and PFDO. The comparison between ISO and SDP shows small differences between both, justifying the adoption of ISO for its simplified mathematics as opposed to SDP. Celeste and Ahmed (2018) focused on assessing the applicability of an ISO procedure to derive rule curves for two different dams of contrasting reservoir scales in terms of physical and operational characteristics. They established that the ISO provided operating rules similar to, and even less vulnerable than, those derived by stochastic dynamic programming. Avila et al. (2020) applied ISO with copula functions to simulate long-term operating policies for a hydropower reservoir located in the Northeastern region of

Brazil. Overall, ISO is considered as one of the most reliable techniques to derive long-term reservoir operating rules for reservoirs. In this review it is found that in ISO for reservoir management some researcher used regression analysis and some other used synthetically generated inflow sequence to deal stochasticity of inflow.

Study area

In this study, DV system (as shown in Fig. 1), a four reservoir multi-purpose system in Jharkhand, India has been considered. There are four reservoirs: Konar, Tilaiya, Panchet and Maithon in DV system as shown in Fig. 1. Reservoir Konar and Panchet are located over river Damodar and reservoir Tilaiya and Maithon are over river Barakar, a tributary of Damodar. Durgapur Barrage is located 17 km downstream from the point of confluence of river Damodar and Barakar near Dishergarh in Bardhaman District in West Bengal. Tilaiya, Panchet and Mithon are associated with hydroelectric power plants of capacities 4 MW, 40 MW and 60 MW respectively. Total catchment area of the river Damodar and its tributaries is about 6159 km². The catchment is irregular in shape and somewhat elongated in lower reach.

The mean annual rainfall are 125.9 cm (at Barakar basin), 127.1 cm (at Damodar basin), lower valley (down stream of point of confluence of Damodar and Barakar) 132.89 cm. About 82% of the total rainfall occurs during south west monsoon period (July to September). During the post-monsoon (October–November) and pre-monsoon (April–May)

months the amount of rainfall is about 8 percent and 7 percent respectively. During the summer season (May) mean daily maximum temperature exceeds 40 °C and in winter mean daily maximum temperature is around below 24 °C. During the summer season (May) mean daily maximum temperature exceeds 40 °C and in winter mean daily maximum temperature is around below 24 °C. Due to high temperature, humidity and wind, almost 50% of the total annual evaporation takes place during March to June. The area faces almost 80% mean relative humidity during south west monsoon period.

In DV system there is Municipal and Industrial (M &I) demand at: (1) downstream of Konar up to Panchet, (2) downstream of Tilaiya up to Maithon and (3) downstream of Panchet and Mithon up to Durgapur Barrage. As M &I demand vary from reservoir to reservoir, it is constant for a particular reservoir throughout the year. Irrigation demand is placed downstream of Durgapur barrage and Irrigation demand is variable throughout the year. For sustenance of aquatic life and ecology in the streams, quantity of water for municipal and industrial (M &I) use is considered to be adequate to meet this mandatory requirement and a minimum quantity (2.1 m³/s) is to be released from the barrage to maintain the river ecology.

Statistical parameters like: maximum, minimum, mean, standard deviation, skewness and correlation coefficient for all monthly inflows have been computed from historical inflow of 56 years (1962–2017) and presented in Table 1. With the help of multivariate gamma AR(1) model, 1000 years synthetic sequence resemble to historic sequence been generated. Matalas (1967) proposed the multivariate AR(1) model preserves the lag-zero and lag-one autocorrelations and cross correlations. In this study, the randomness of inflows (hydrologic timeseries) have been tested by computing lag-zero autocorrelation which is almost near to zero. The monthly mean values show that maximum inflow in the system took place from June to October and in maximum and minimum values of these months also indicate about very high variation. In these 5 months appreciable amount of skewness are also observed which indicate a non-normal distribution. Some months (July, August, October, November) show strong co-relation. Variation of monthly inflows into the four reservoirs during June to October have been presented in Fig. 2.

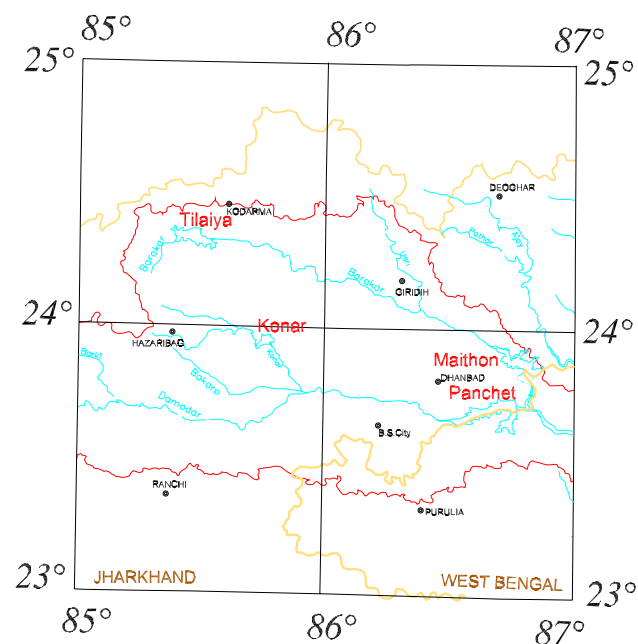
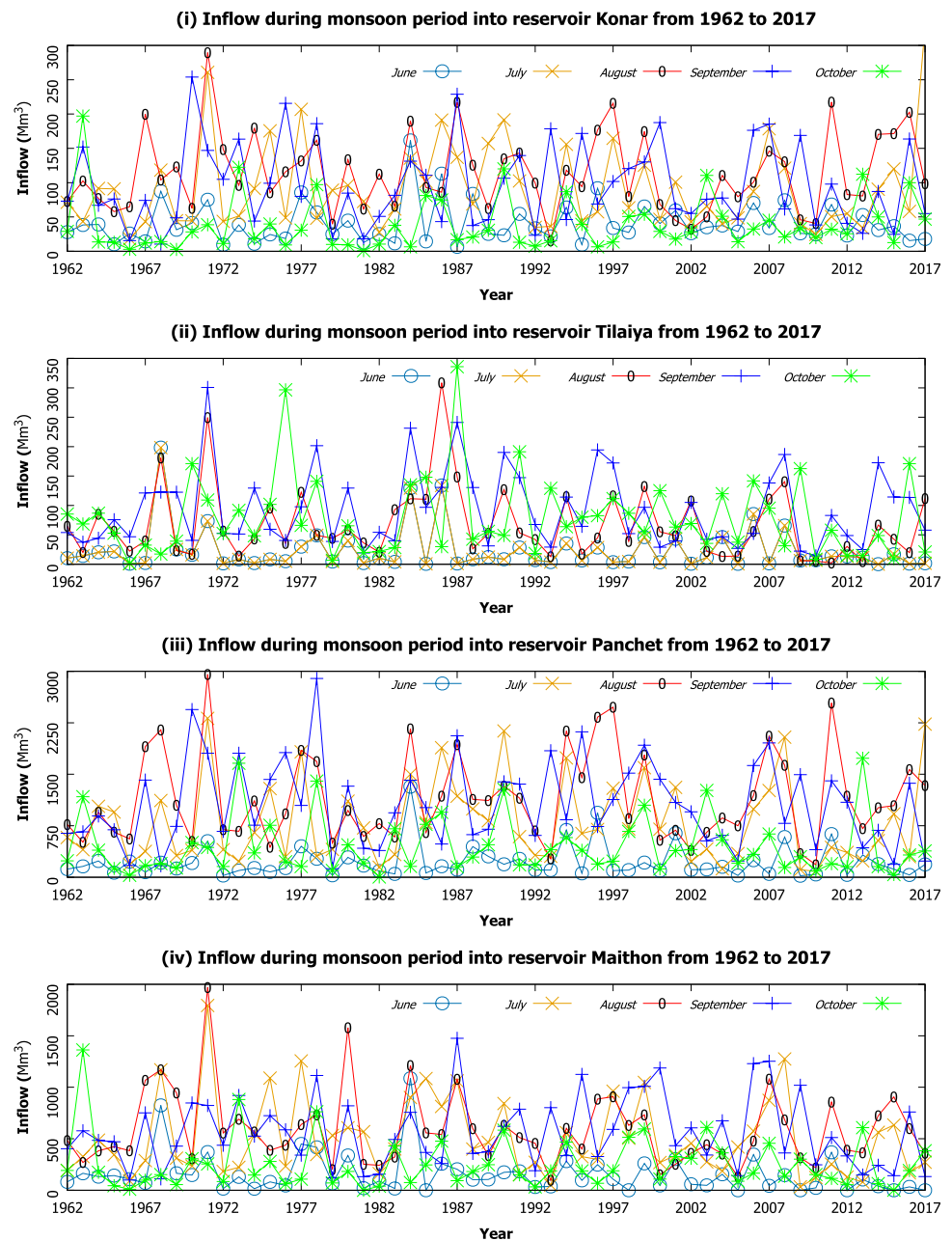


Fig. 1 Damodar valley system

Methodology

In this research paper, optimal operating policies of a multi-purpose multi-reservoir system have been developed through implicit optimization method. At the beginning variable selection techniques, in terms of (1) All Possible Regression (APR) and (2) Stepwise Regression (SR) are discussed.

Fig. 2 Monthly inflow distribution during monsoon period



It is assumed that the regression equations are linear and a constant term (intercept) is present in the equation. The decomposition models and the simulation technique are then described in relation to the DV system. Finally, this approach is applied to the DV system with a number of synthetic inflow sequences generated by the multivariate gamma AR(1) model as input and the results are presented.

Here a subset of independent (predictor) variables is to be selected for use in a regression equation among many potential variables present in the system. Although different methods for subset selection are available which use different statistics for test of significance and in practice all of them may not lead to the same solution when applied to

the same problem (Draper and Smith 1968; Kennedy and Gentle 1980). Hence, performances of different model building methods are to be examined before selecting the final equation. In the present study, two widely used methods for regression analysis in statistics: APR and SR, have been applied for reservoir management and tested.

All possible regression (APR) technique

One of the most comprehensive but cumbersome ways of selecting subset regression is to compute all possible subset regressions. In all possible regression, the pool of q variables is divided into subsets which involve n variables,

Table 1 Statistical parameters of inflow of DV system

Period	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<i>Reservoir Konar</i>												
Ymax	31.51	29.34	32.24	25.35	22.64	161.5	328.2	289.5	253.8	197.1	44.39	35.05
Ymin	0.009	0.000	0.000	0.000	0.430	6.519	18.09	15.16	10.42	0.888	0.510	0.000
Mean	6.305	5.834	5.978	6.211	6.931	40.81	91.71	114.0	96.45	38.71	8.103	6.314
σ	6.530	5.990	6.160	5.700	7.660	29.51	62.82	57.04	60.91	38.77	8.033	6.601
γ	1.760	1.803	1.838	1.593	1.401	1.739	1.576	0.740	0.689	1.935	2.116	1.999
ρ	0.290	0.602	0.590	0.732	0.473	0.321	0.802	0.734	0.678	0.811	0.853	0.465
<i>Reservoir Tilaiya</i>												
Ymax	24.40	14.71	12.33	14.62	11.38	198.4	308.5	300.6	335.9	175.5	27.58	36.09
Ymin	0.000	0.000	0.000	0.000	0.000	0.260	2.220	14.22	1.300	0.160	0.000	0.000
Mean	4.563	2.853	1.750	1.136	2.232	23.88	66.88	92.67	79.40	31.08	5.221	4.921
σ	5.625	3.618	2.700	2.600	3.143	37.41	60.14	64.88	66.69	33.71	5.457	6.503
γ	2.069	1.503	2.380	3.557	1.520	2.900	1.847	1.113	1.773	2.171	2.012	2.629
ρ	0.102	0.694	0.538	0.344	0.751	0.678	0.703	0.745	0.632	0.652	0.753	0.714
<i>Reservoir Panchet</i>												
Ymax	134.5	150.8	89.70	195.8	146.9	1320	2315	2956	2896	1734	486.6	216.8
Ymin	0.000	0.000	0.000	0.000	0.470	22.54	85.40	182.5	124.7	4.215	0.000	0.000
Mean	18.36	19.36	17.07	23.99	29.49	224.0	807.6	1146	1063	445.9	76.92	25.36
σ	21.67	23.83	17.75	30.19	26.18	244.1	602.9	656.1	630.0	412.9	80.36	32.71
γ	3.137	3.521	1.918	3.639	2.037	2.483	0.975	0.860	0.637	1.609	2.827	3.921
ρ	0.215	0.207	0.324	0.284	0.010	0.436	0.672	0.652	0.796	0.891	0.763	0.241
<i>Reservoir Maithon</i>												
Ymax	112.9	101.07	66.98	60.17	53.04	1087	1794	1970	1475	1363	215.8	65.17
Ymin	0.000	0.000	0.000	0.000	0.000	0.000	33.08	96.90	98.89	0.000	0.000	0.000
Mean	11.55	13.84	9.925	5.967	8.307	156.9	501.0	586.7	562.6	257.6	30.09	13.28
σ	21.25	20.95	15.60	12.59	12.96	196.8	377.3	359.3	346.24	251.77	39.03	18.76
γ	2.768	2.016	1.726	2.674	1.694	2.787	1.249	1.568	0.648	2.084	2.328	1.217
ρ	0.133	0.141	0.671	0.238	0.768	0.552	0.732	0.798	0.621	0.765	0.831	0.453

$n = 1, 2, \dots, q$. In each subset, regressions are done for all possible combinations of the predictors, and each regression equation is ordered according to some criterion; usually the criterion is the value of R^2 (coefficient of determination) achieved by the least square fit (Draper and Smith 1968, Chapter-8). The leaders in this ordering within each subset are then selected for further examination and a decision is made as to which equation is best to be used, after examining the R^2 [correlation coefficients computed for optimal storage states (dependent variables), initial storage states and inflows into reservoir] values.

The major drawback of this approach is the number of regressions that must be calculated: $2^q - 1$ for a pool of q predictor variables. Contrasting with the computational effort involved is the valuable information provided by examining all possible subsets. Most other variable selection procedures yield only one or perhaps a few candidates for the final prediction equation. By examining summary information one can identify the better subsets and then choose one or more according to the nature and purpose of the investigation. Considerable attention has been directed

at identifying the best (largest R^2 or, equivalently, smallest residual sum of squares) subsets of a given size; i.e., of all subsets having a fixed number of predictor variables, identifying the one that has the largest R^2 (Gunst and Mason 1980, Chapter-8).

The main advantages in computing all possible regression or the selection criteria for all the regression, have led to research for more efficient means of determining best subsets. The disadvantage of this approach is that only the single best subset of a given size is identified where as other good subsets that could be trivially inferior to the best ones might not be identified, yet these other subsets may be preferable due to problem related considerations.

Stepwise regression (SR) technique

Stepwise regression procedure is a selection technique that sequentially add or delete single predictor variables to the regression equation, based on certain criterion. Since a series of steps are involved and since each step leads directly

to the next, these methods involve the calculation of a much smaller number of equations than $2^q - 1$ required for the APR approach.

At each step of the SR a predictor variable is added to the regression if its t -statistic is the largest one calculated and exceeds critical value, t_c . Next each predictor variable already chosen is reconsidered and eliminated from the selected subset if its t -statistic is the smallest one and does not exceed t_c value. This process continues until all q predictor variables are in the equation or until the selection criteria are no longer met. At any step, the SR method allows one to judge the contribution of each predictor in the subset as though it were the most recent variable to enter the regression equation.

The advantages of SR is that (1) the particular variable which does not have any significant contribution is eliminated at the entry level for given step and (2) it puts more power and information readily available than ordinary multiple regression method. The major disadvantages of SR is that the end result is a single subset which may not be necessarily the best for a given size (Gunst and Mason 1980, Chapter-8).

Application to DV reservoir system

In the implicit stochastic DP framework, the operation problem of the DV system of reservoirs is first solved for 100 sets of synthetic inflow sequences (generated by the multivariate gamma AR(1) model, using (1) Discrete Dynamic Programming (DDP) algorithm as proposed by Bellman and Dreyfus (1962) considering one reservoir at a time approach and (2) Discrete Differential Dynamic Programming(DDDP) algorithm as proposed by Heidari et al. (1971) with the help of initial trajectories as obtained from DDP based on the following system dynamics (Eqs. 1–4), system constraints (Eqs. 5–6), performance function (Eq. 7), objective function (Eq. 8) and recursive equation (Eq. 9):

$$x_{t+1}(1) = x_t(1) + y_t(1) - u_t(1) - ev_t(1) \tag{1}$$

$$x_{t+1}(2) = x_t(2) + y_t(2) - u_t(2) - ev_t(2) \tag{2}$$

$$x_{t+1}(3) = x_t(3) + y_t(3) + u_t(1) - en_t(3) - u_t(3) - ev_t(3) \tag{3}$$

$$x_{t+1}(4) = x_t(4) + y_t(4) + u_t(2) - en_t(4) - u_t(4) - ev_t(4) \tag{4}$$

$$C_{\min} \leq \mathbf{x}_t \leq C_{\max} \text{ (on storage)} \tag{5}$$

$$\mathbf{u}_{t,\min} \leq \mathbf{u}_t \leq \mathbf{u}_{t,\max} \text{ (on release)} \tag{6}$$

$$L_t = [(D_t(1) - u_t(1))/D_t(1)]^2 + [D_t(2) - u_t(2)]/D_t(2)]^2 + [D_t^c - u_t(3) - u_t(4)]/D_t^c]^2 \tag{7}$$

$$\min_{u_t} \sum_{t=1}^N L_t(u_t) \tag{8}$$

$$V_t(\mathbf{x}_t^i) = \min_{\mathbf{x}_{t+1}^k} \{L_t + V_{t+1}(\mathbf{x}_{t+1}^k)\} \tag{9}$$

$$i = 1, 2, \dots, I, t = 1, 2, \dots, T$$

$x_t(i)$, $u_t(i)$, $y_t(i)$, $ev_t(i)$ and $en_t(i)$ represent the volume of storage, the volume of release, the volume of inflow, the amount of evaporation loss to the i th reservoir during time period t and the en-route loss in the reach between $(i - 2)$ th reservoir and i th reservoir. C_{\min} and C_{\max} represent vectors of minimum and maximum storage capacities of the four reservoirs, $\mathbf{u}_{t,\min}$ is the vector of minimum mandatory releases, and $\mathbf{u}_{t,\max}$ is the vector of maximum permissible releases, L_t is the single stage loss function, $D_t(1)$ and $D_t(2)$ represent the water supply target levels for reservoir Konar and reservoir Tilaiya respectively and D_t^c represents the combined water supply target level for reservoir Panchet and Maithon where $V_{N+1}(\mathbf{x}_{N+1}^k) = 0$ for all k (number of iterations) and $\mathbf{x}_{t+1}^i, \mathbf{x}_{t+1}^k, \mathbf{u}_t = \mathbf{f}(\mathbf{x}_t^i, \mathbf{x}_{t+1}^k)$ feasible. (Bracketted number 1, 2, 3 and 4 attached with all notations represent reservoir Konar, Tilaiya, Panchet and Maithon respectively). In system dynamics (Eqs. 1–4), seepage loss has been neglected as well as reduction of live storage space due to sedimentation is assumed as zero.

The 100 sets of optimal trajectories as obtained from DDDP, the average operating policies are determined for each reservoir for each month. As the final storage state (\mathbf{x}_{t+1}) during stage t as used as the decision variable in the deterministic model, the same is used as the dependent variable (response) in the regression analysis. The independent (predictor) variables are $\mathbf{x}_t(i)$, $\mathbf{y}_t(i)$, $i = 1, \dots, M$ and $y_{t-1}(j)$ for the particular reservoir j . For convenience, these independent variables are assigned name as follows:

Independent variable	$x_t(1)$	$x_t(2)$	$x_t(3)$	$x_t(4)$	$y_t(1)$	$y_t(2)$	$y_t(3)$	$y_t(4)$	$y_{t-1}(j)$
Name assigned	1	2	3	4	5	6	7	8	9

thus making a total of 9 candidate independent variables. A linear equation (with a constant term) is assumed for regression. In the present case, dealing with non-linear equations is extremely difficult as the number of predictor variables are many, and so it is not attempted here.

For the selection of the appropriate predictor variables from a total of 9 candidate variables, the proposed approach starts with the APR as described in “All possible regression”. In APR, out of several subsets four subsets: SS1, SS2, SS3 and SS4 with independent variable 2, 4, 6 and 8 respectively, are used for comparison. Then the SR (as described in “Stepwise regression (SR)”) is applied and continued till all 9 independent variables are in the equation based on the selection criteria (critical value, t_c) are met. As these two techniques deals with the statistical significance of the regressions, another group of subsets is formulated based on the decomposition of the original four reservoir problem into smaller sub-problems, which is described next.

Decomposition model

Decomposition models are based on the physical significance of the system variables in terms of continuity of flow and configuration of the system. In this study for DV system, the continuity equations (Eqs. 1–4) indicate that for any reservoir, the final storage is primarily a function of the initial storage state and the inflow state of that reservoir, during stage t . As reservoir is of finite size, any amount above the maximum storage capacity will be spilled. Similarly, as downstream channel is of finite capacity, outflow will not exceed optimal channel capacity. In this problem down stream demand is predefined and sum of the square of deviation of outflow from target is minimized. Considering the integrated operation of the reservoirs, the other influencing variables are the state of the other reservoirs upstream or downstream, connected in series or parallel, although, their order of influence can not be judged precisely. Based on these considerations, four different configurations: DM-A, DM-B, DM-C and DM-D, of decomposition models are constructed as shown in Fig. 3.

DM-A represents the simplest (elementary) model which uses a single reservoir unit and the subset contains only two predictor variables: storage and inflow. Four models (DM-A21, DM-A22, DM-A23 and DM-M24) are constructed in this configuration. The DM-B configuration is constructed considering the unit of two reservoirs and the subsets consist of 4 independent variables. Two models are constructed (DM-B41 and DM-B42) in this configuration. The DM-C configuration consists of three reservoirs and accordingly, the subset contains six predictor variables. In this configuration also, two models are constructed (DM-C61 and DM-C62). Finally, DM-D81 considers the original configuration of the four reservoir system, and accordingly, the subset contains eight independent variables (Table 2).

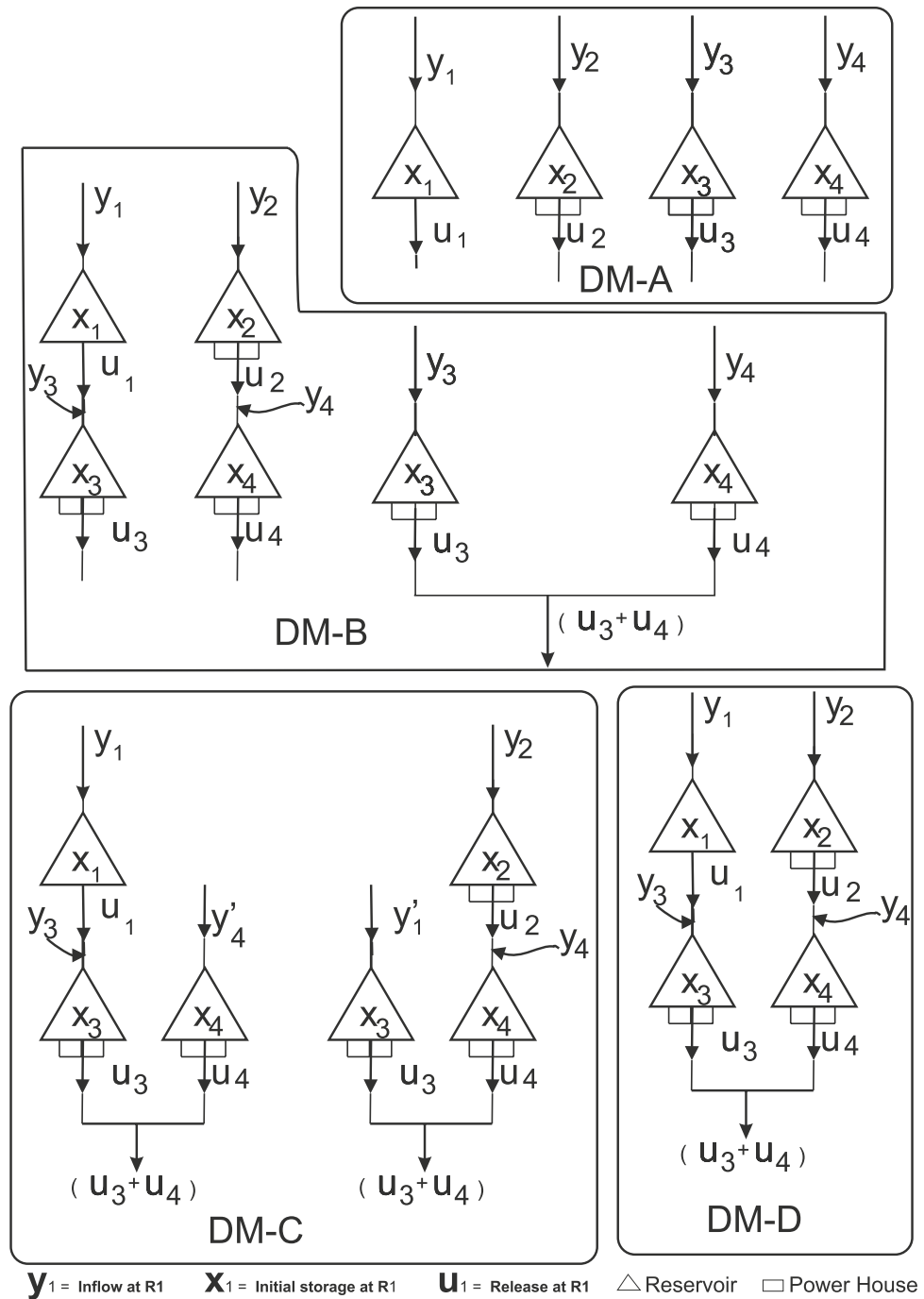
For the downstream reservoirs, y'_t represents the net inflow whereas y_t represents the local (intermediate) inflow. Each of them (y_t and y'_t) is used in the equation separately to observe the corresponding effects. Actually, in the present case, the contribution of the y_t to the net inflow y'_t is much more than the upstream release. Also, presence of y_t in the regression equation avoids the forecast errors associated with the inflow to the upstream reservoir. Previous period's inflow instead of current period's inflow is also tested. However, this is not tested for the other three configurations (DM-B, DM-C and DM-D). In the above table number 1, 2, 3 and 4 represent reservoir Konar, Tilaiya, Panchet and Maithon respectively and $b_{j,t}(i)$ is j -regression coefficient for the i th reservoir for stage t . The decomposition models (DM-A, DM-B, DM-C and DM-D) are actually the stepwise inclusion of different reservoir units maintaining the spatial continuity of the system.

In decomposition model one after another reservoir unit is included in the system maintaining the spatial continuity. The main advantages of the decomposition model are the effect of one reservoir on the other reservoir can be studied clearly. The main disadvantage is with the increase of number of reservoir in the system, number of regression equation would be more, therefore number of predictor variables would increase with increasing number of reservoir in different configuration. The APR and the SR serve as guidelines, in terms of statistical significance, in order to select the appropriate decomposition model.

Simulation

From the regression analysis, a model can be selected by considering the physical and statistical significance, however, the real test of how good the resulting regression model will depend on the ability of the model to predict the dependent variable for observations on the independent variables that were not used in estimating the regression coefficients (Haan, 1979, Chapter-10). To make a comparison of this nature, the historical inflow sequence, which is not used in forming the data set for regression, is used with the regression coefficients obtained from different decomposition models (DM-A21, DM-A22, DM-A23, DM-A24, DM-B41, DM-B42, DM-C61, DM-C62, DM-C81). The resulting release sequences are then used to compute the average values of the objective function (as mentioned in Eq. 8). With the help of simulation model, performance and predictive efficiency of different regression models are computed and a model with less number of independent variables is selected among different models having relatively similar or closer objective function value.

Fig. 3 Decomposition models with different configurations



Results and discussion

APR technique

The APR technique yields a number of solutions based on different subsets instead of a particular solution. As mentioned in “All possible regression”, the best of the each subset is selected based on the largest value of *R*² for each reservoir for each month in percent. For subsets SS1, SS2, SS3 and SS4, the corresponding *R*² values as obtained from

the decomposition models DM-A, DM-B, DM-C and DM-D having same subsets have also been computed. Variable names corresponding the largest *R*² values as obtained from those subsets and decomposition models are presented in Table 3. In all the following tables, variables number are written without separating by comma.

From the Table 3, 1 it is observed that most of the variance in the response can be explained by two or at the most four variables and the regression model should contain the

Table 2 Regrssion equations of different decomposition models

	DM-A21	DM-A22
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1)$	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1)$
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2)$	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2)$
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_t(3)$	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_t'(3)$
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_t(4)$	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_t'(4)$
	DM-A23	DM-A24
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_{t-1}(1)$	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_{t-1}(1)$
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_{t-1}(2)$	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_{t-1}(2)$
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_{t-1}(3)$	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_{t-1}'(3)$
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_{t-1}(4)$	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_{t-1}'(4)$
	DM-B41	
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1) + b_{3,t}(1)x_t(3) + b_{4,t}(1)y_t(3)$	
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2) + b_{3,t}(2)x_t(4) + b_{4,t}(2)y_t(4)$	
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_t(3) + b_{3,t}(3)x_t(3) + b_{4,t}(3)y_t(3)$	
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_t(4) + b_{3,t}(4)x_t(4) + b_{4,t}(4)y_t(4)$	
	DM-B42	
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1) + b_{3,t}(1)x_t(3) + b_{4,t}(1)y_t(3)$	
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2) + b_{3,t}(2)x_t(4) + b_{4,t}(2)y_t(4)$	
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(3) + b_{2,t}(3)y_t(3) + b_{3,t}(3)x_t(3) + b_{4,t}(3)y_t'(3)$	
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(4) + b_{2,t}(4)y_t(4) + b_{3,t}(4)x_t(4) + b_{4,t}(4)y_t'(4)$	
	DM-C61	
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1) + b_{3,t}(1)x_t(3) + b_{4,t}(1)y_t(3) + b_{5,t}(1)x_t(4) + b_{6,t}(1)y_t(4)$	
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2) + b_{3,t}(2)x_t(4) + b_{4,t}(2)y_t(4) + b_{5,t}(2)x_t(3) + b_{6,t}(2)y_t(3)$	
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(1) + b_{2,t}(3)y_t(1) + b_{3,t}(3)x_t(3) + b_{4,t}(3)y_t(3) + b_{5,t}(3)x_t(4) + b_{6,t}(3)y_t(4)$	
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(2) + b_{2,t}(4)y_t(2) + b_{3,t}(4)x_t(4) + b_{4,t}(4)y_t(4) + b_{5,t}(4)x_t(3) + b_{6,t}(4)y_t(3)$	
	DM-C62	
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1) + b_{3,t}(1)x_t(3) + b_{4,t}(1)y_t(3) + b_{5,t}(1)x_t(4) + b_{6,t}(1)y_t(4)$	
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2) + b_{3,t}(2)x_t(4) + b_{4,t}(2)y_t(4) + b_{5,t}(2)x_t(3) + b_{6,t}(2)y_t(3)$	
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(1) + b_{2,t}(3)y_t(1) + b_{3,t}(3)x_t(3) + b_{4,t}(3)y_t(3) + b_{5,t}(3)x_t(4) + b_{6,t}(3)y_t'(4)$	
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(2) + b_{2,t}(4)y_t(2) + b_{3,t}(4)x_t(4) + b_{4,t}(4)y_t(4) + b_{5,t}(4)x_t(3) + b_{6,t}(4)y_t'(3)$	
	DM-D81	
1	$x_{t+1}(1) = b_{0,t}(1) + b_{1,t}(1)x_t(1) + b_{2,t}(1)y_t(1) + b_{3,t}(1)x_t(3) + b_{4,t}(1)y_t(3) + b_{5,t}(1)x_t(4) + b_{6,t}(1)y_t(4) + b_{7,t}(1)x_t(4) + b_{8,t}(1)y_t(4)$	
2	$x_{t+1}(2) = b_{0,t}(2) + b_{1,t}(2)x_t(2) + b_{2,t}(2)y_t(2) + b_{3,t}(2)x_t(4) + b_{4,t}(2)y_t(4) + b_{5,t}(2)x_t(3) + b_{6,t}(2)y_t(3) + b_{7,t}(2)x_t(4) + b_{8,t}(2)y_t(4)$	
3	$x_{t+1}(3) = b_{0,t}(3) + b_{1,t}(3)x_t(1) + b_{2,t}(3)y_t(1) + b_{3,t}(3)x_t(3) + b_{4,t}(3)y_t(3) + b_{5,t}(3)x_t(4) + b_{6,t}(3)y_t(4) + b_{7,t}(3)x_t(4) + b_{8,t}(3)y_t(4)$	
4	$x_{t+1}(4) = b_{0,t}(4) + b_{1,t}(4)x_t(2) + b_{2,t}(4)y_t(2) + b_{3,t}(4)x_t(4) + b_{4,t}(4)y_t(4) + b_{5,t}(4)x_t(3) + b_{6,t}(4)y_t(3) + b_{7,t}(4)x_t(4) + b_{8,t}(4)y_t(4)$	

two primary variables, storage and inflow for the particular reservoir. It also demonstrates the limitation of selection process based on the highest R^2 value e.g, for the month of February and for subset 2, the highest R^2 criterion selects variables 1 [$x_t(1)$] and 2 [$x_t(2)$], whereas the corresponding decomposition model with variables 1 [$x_t(1)$], and 5 [$y_t(1)$] yielded almost the same R^2 value, but it contains variables physically more significant than subset of variables 1 and 2.

It is also observed that the 9th variable, i.e, previous period's inflow for the corresponding reservoirs appears in fourth subset and onward. It indicates that predictor variable y_{t-1} is having very little influence on the response and so it's presence in the regression models can be avoided. This is partly due to the presence of both y_t and y_{t-1} in the data matrix which are correlated. Although in the subset of two predictor variables, y_t is always present, but not y_{t-1} .

Table 3 Selected variables from different subset and decomposition model based on highest R^2 value

Month	Reservoir Konar				Reservoir Tilaiya			
	APR subset corresponding variables				APR subset corresponding variable			
	SS1	SS2	SS3	SS4	SS1	SS2	SS3	SS4
Oct	15	1245	124568	12456789	26	1269	126789	12346789
Nov	15	1345	134589	12345789	26	2469	234569	12345689
Dec	14	1459	145679	13456789	26	2469	124569	12456789
Jan	13	1346	123579	12345679	26	2369	234569	12345679
Feb	13	1357	123569	12345789	26	2369	235679	12345679
Mar	13	1359	123569	12345689	23	2369	123679	12356789
Apr	15	1359	123459	12345789	26	2369	236789	12346789
May	15	1579	145789	12345789	26	2469	234679	23456789
Jun	15	1568	145678	13456789	26	1268	124568	12456789
Jul	15	1459	134569	12345689	26	1269	124679	12345689
Aug	15	1569	124569	12345679	26	1269	124679	12345679
Sep	15	1459	124569	12456789	26	1269	124679	12346789
	Decomposition model				Decomposition model			
	15	1357	134578	12345678	26	2468	234678	12345678
Month	Reservoir Panchet				Reservoir Maithon			
	SS1	SS2	SS3	SS4	SS1	SS2	SS3	SS4
	SS1	SS2	SS3	SS4	SS1	SS2	SS3	SS4
Oct	56	3578	134578	12345678	47	4578	124578	12456789
Nov	37	3567	135679	12356789	48	1468	346789	13456789
Dec	37	1378	135678	12345678	47	3478	346789	13456789
Jan	37	2347	234679	23456789	48	3478	123478	12346789
Feb	37	1347	134567	13456789	48	3478	345789	12345789
Mar	37	2347	234789	23456789	48	1248	123478	12346789
Apr	37	3467	234678	12345678	47	3478	123478	12345789
May	37	1378	134578	12345789	48	2458	245678	12456789
Jun	37	3479	345679	23456789	48	4678	234678	12345678
Jul	67	1267	125678	12356789	27	1567	125679	12456789
Aug	17	1479	134579	12345679	47	1457	134579	13456789
Sep	17	1579	156789	12456789	17	1479	145789	12456789
	Decomposition model				Decomposition model			
	37	1357	134578	12345678	48	2468	234678	12345678

However, at this point, y_{t-1} is not discarded as a predictor variable and it will be used for further examination.

SR technique

In the SR technique a particular model has been selected based on t -statistic. In Table 4 the particular variables selected are shown in terms of their numbers for each reservoir for each month. It is observed from Table 4 that the x_t and y_t values for the corresponding reservoirs (1,5 for reservoir 1 and 2, 4 for reservoir 2) are always present in the selected models, for almost every month. Regarding the downstream reservoirs Table 4 showed that x_t and y_t values for reservoir 3 (numbers 3, 7) and reservoir 4

(numbers 4, 8) are present in the models almost always. The 9th variable (y_{t-1}) occurred in some months for different reservoirs. In a few places y_{t-1} is present in absence of y_t for a particular reservoir. In general, although the effect of y_{t-1} is not very prominent, it is retained for further examination. Regarding the number of variables in a subset, in most of the cases stepwise regression suggested a subset of minimum four variables and higher (some times it includes all the variables).

In general the results from APR and SR always indicate the use of elementary variables (storage and inflow). Regarding the inclusion of other system variables many times model suggested by these two techniques are similar to the decomposition models, although in some cases they suggested inclusion of variables of lesser physical

Table 4 Selected variables from stepwise regression

Month	Variable number								Variable number									
	Reservoir Konar								Reservoir Tilaiya									
Oct	1	2		4	5	6		8		1	2		5	6		9		
Nov	1		3	4	5			8	9		2	3	4	5	6	9		
Dec	1			4	5	6	7		9	1	2		4	5	6	9		
Jan	1		3	4	5	6	7		9	1	2	3	4	5	6	7	9	
Feb	1	2	3		5		7		9		2	3		5	6	7	9	
Mar	1	2	3		5	6			9	1	2	3			6		9	
Apr	1	2	3	4	5		7	8	9		2	3			6		9	
May	1	2	3	4	5		7	8	9		2	3	4	5	6		9	
Jun	1			4	5	6	7	8		1	2			5	6		8	
Jul	1		3	4	5	6			9	1	2		4	5	6		8	9
Aug	1		3		5	6			9	1	2		4		6			9
Sep	1	2		4	5	6			9	1	2		4		6	7		9

Month	Variable number								Variable number									
	Reservoir Panchet								Reservoir Maithon									
Oct	1		3		5		7	8		1	2		4	5		7	8	9
Nov	1		3		5	6	7		9	1		3	4		6	7	8	9
Dec	1		3		5	6	7	8		1		3	4	5	6	7	8	9
Jan		2	3	4			7		9	1	2	3	4			7	8	
Feb	1		3	4	5	6	7					3	4	5		7	8	
Mar		2	3	4			7	8	9	1	2	3	4			7	8	
Apr			3	4		6	7			1	2	3	4	5		7	8	
May	1		3		5		7	8			2		4	5	6	7	8	
Jun		2	3	4	5	6	7		9		2	3	4		6	7	8	
Jul	1	2		4		6	7			1	2		4	5	6	7	8	9
Aug	1		3	4	5		7		9	1		3	4	5		7	8	9
Sep	1				5		7			1			4	5			8	9

importance. Hence, the selection of final regression model will be based on the set of decomposition models starting from simplest model with two variables to the full model with all reservoirs.

Decomposition models and simulation

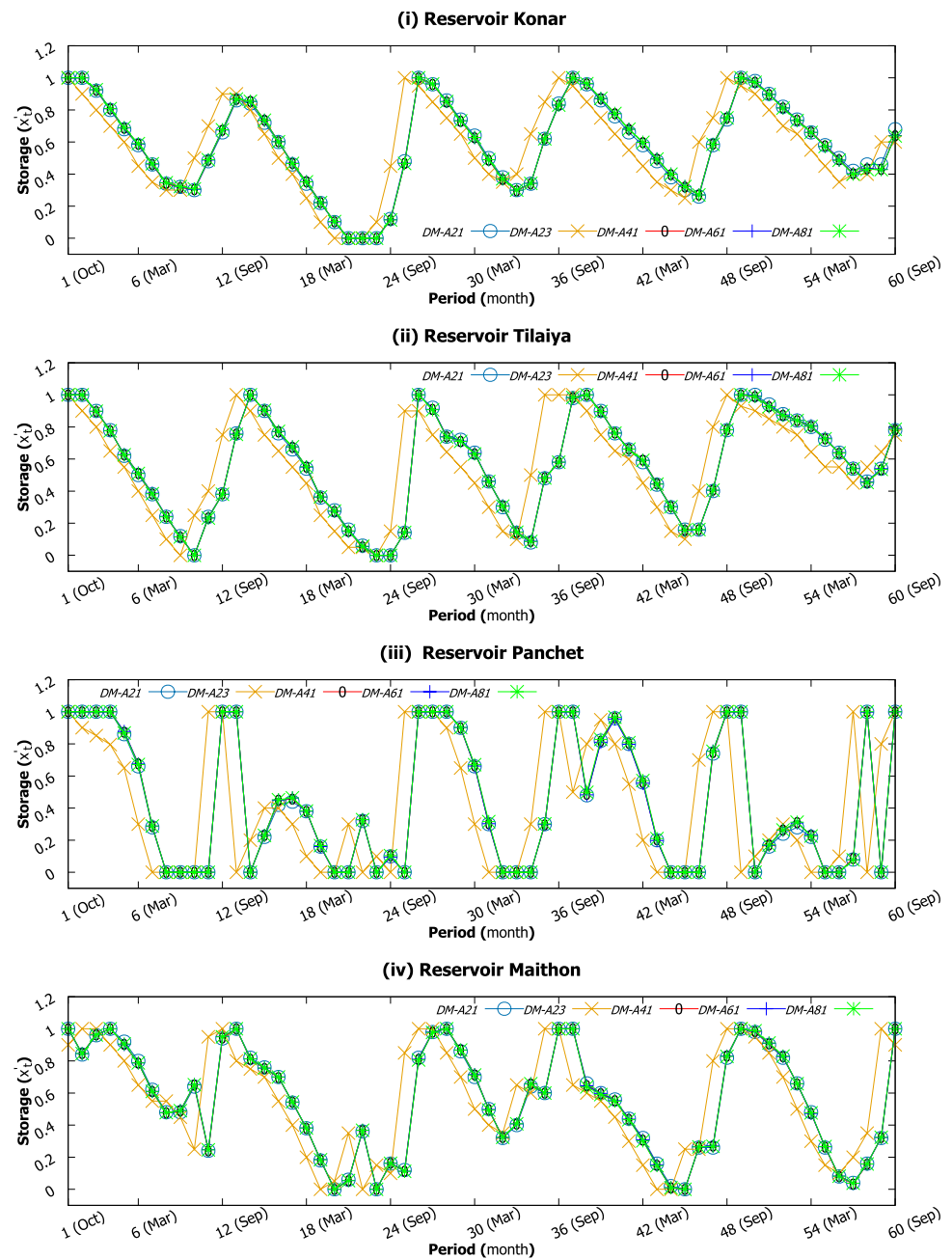
The decomposition models are based on the physical significance of the system variables and from the APR and SR analysis, it is found that these models are also having acceptable statistical significance. Hence, final selection

of a particular regression equation can be made from these decomposition models. The APR analysis suggested for models with at the most four variables, whereas the SR method sometimes suggested for higher models. Before final selection, another comparison is made with these decomposition models through simulation. All the nine regression models are used with the historical inflow sequences (1961–2018) to determine the corresponding releases and the value of objective function as shown in Table 5 have been computed and the lowest value is obtained from DM-B41.

Table 5 Decomposition models in terms objective function value

Model	Objective function value	Model	Objective function value	Model	Objective function value
DM-A21	1.9447173×10^{12}	DM-A24	1.953566×10^{12}	DM-C61	1.959352×10^{12}
DM-A22	1.937229×10^{12}	DM-B41	1.803144×10^{12}	DM-C62	1.951731×10^{12}
DM-A23	1.972152×10^{12}	DM-B42	1.822271×10^{12}	DM-D81	1.966327×10^{12}

Fig. 4 Trajectories of different decomposition models



Storage behaviors of the four reservoirs corresponding to the different decomposition models are shown in Fig. 4. The trajectories are shown for the first five years only. Although no specific inferences can be drawn, these figures show that trajectories obtained from DM-A23 (based on y_{t-1}) is somewhat different than the others which are almost identical. Due to fixed M & I demand, nature of trajectories for two upstream reservoirs are different than that of two downstream reservoirs because of variable irrigation demand.

Selection of the final model from competing models of different size is actually a compromise between two opposite criteria (Draper and Smith 1968, Chapter-6): (1) to select

a model with less of number of variables because of the difficulties involved in obtaining information on a large number of predictors (specifically for inflows) and (2) to select a model with as many variables as possible for better representation of the system and for better prediction. The simpler models (DM-A and DM-B) are having the advantage of using less number of variables: for any reservoir, decisions are based on the state of that particular reservoir. Forecast errors or the numerical errors associated with the state variables of the other reservoirs do not affect the decision much. However, these models cannot incorporate the other significant configurational aspects, like the joint operation

of downstream reservoirs, which sometime may have greater consequence.

On the other hand, higher models (DM-C and DM-D) show better approximation of the system regarding the system configuration and the influence of other reservoirs on a particular reservoir. Apart from this, the numerical errors associated with the use of regression is also more in case of higher models. For example, for the complete model DM-D81, during a particular stage t , for any reservoir, the final storage state x_{t+1} is based on the storage state x_t of all reservoirs which were again computed from some regression equations during stage $t - 1$. Thus as stage t increases, numerical errors also increase. Also, as the equations for DM-D81 involve more number of random variables y_t , forecast errors associated with these variables will also affect the decision and the error will be propagated.

Based on statistical significance, physical significance and objective function values obtained through simulation, and other justifications discussed above, model DM-B41 is selected.

Conclusion

In the implicit stochastic DP framework, a generalized approach is proposed here for the selection of appropriate variables in the regression. The variable selection techniques based on statistical significance, APR and SR, are powerful tools but sometimes important variables from physical justification of the problem may be discarded. So, care should be taken regarding the final selection. The approach selects the appropriate model from the judgments based on statistical and physical significance of the presence of a variable in the regression model, performance of the model through simulation and from the view of utility. The statistical performances of the different decomposition models which are based on the spatial decomposition of the original problem and maintain the essence of physical significance of the operation problem, are also quite justified and acceptable. Therefore, final selections can be made from this configuration of models.

To obtain precise information about some variable (e.g, inflow) is very difficult and there is possibility of propagation of error, therefore models with less number of predictor variables are selected from a group of models with similar performance. Here simulation serves as an important basis for selecting models from almost similar models (in terms of statistical or physical significance), DM-41 is selected as most preferred model. All these conclusions are system dependent in general and true for the DV system in particular.

References

- Avila L, Mine MR, Kaviski E (2020) Probabilistic long-term reservoir operation employing copulas and implicit stochastic optimization. *Stochastic Env Res Risk Assess* 34(7):931–947
- Bellman R, Dreyfus S (1962) *Applied dynamic programming*. Princeton, NJ
- Bhaskar NR, Whitlatch EE Jr (1980) Derivation of monthly reservoir release policies. *Water Resour Res* 16(6):987–993
- Celeste AB, Ahmed ES (2018) Assessment of stochastic operation optimization for reservoirs of contrasting scales. *Water Resour Manage* 32(11):3751–3763
- Celeste AB, Santana RF, Santos WM (2016) Dynamic and implicit stochastic optimization for reservoir operation: Case study in Sergipe. *Ciencia Natura* 38(1):273–280
- Celeste AB, Wilson FC, Rosires CC (2009) Implicit stochastic optimization for deriving reservoir operating rules in semiarid Brazil. *Pesquisa Operacional* 29(1):223–234. <https://doi.org/10.1590/SS0101-74382009000100011>
- Draper NR, Smith H (1968) *Applied regression analysis*. John Wiley & Sons Inc., New York
- Gunst RF, Mason RL (1980) *Regression analysis and its application: a data-oriented approach*. Marcel Dekker Inc., New York
- Hall WA, Howell D (1963) The optimization of single-purpose reservoir design with application of dynamic programming to synthetic hydrology samples. *J Hydrol* 1:355–363
- Heidari M, Chow VT, Kokotovic PV, Meredith DD (1971) Discrete differential dynamic programming approach to water resources systems optimization. *Water Resour Res* 7(2):273–282
- Karamouz M, Houck MH (1982) Annual and monthly reservoir operating rules generated by deterministic optimization. *Water Resour Res* 18(5):1337–1344
- Karamouz M, Mark HH, Delleur JW (1992) Optimization and simulation of multiple reservoir systems. *J Water Resour Plann Manage Div, Am Soc Civ Eng*. 118(1):71–81
- Kennedy WJ Jr, Gentle JE (1980) *Statistical Computing*. Marcel Dekker Inc., New York
- Liu P, Liping L, Gaojie C, David ER (2014) Parameter uncertainty analysis of reservoir operating rules based on implicit stochastic optimization. *J Hydrol* 514:102–113
- Matalas NC (1967) Mathematical assessment of synthetic hydrology. *Water Resour Res* 3(4):937–945
- McKerchar AI (1975) Optimal monthly operation of interconnected hydroelectric power storages. *J Hydrol* 25:137–158
- Moreira RM, Celeste AB (2017) Performance evaluation of implicit stochastic reservoir operation optimization supported by long-term mean inflow forecast. *Stochastic Environ Res Risk Assess* 31(9):2357–2364
- Trott WJ (1979) Determination of real time operating rules for a multiple reservoir system. University of California, Dissertation
- Young GK (1967) Finding reservoir operating rules. *J Hydraul Div Am Soc Civ Eng*. 93(HY6):297–321

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.