Intermittent reservoir daily-inflow prediction using lumped and distributed data multi-linear regression models

R B Magar and V Jothiprakash*

Department of Civil Engineering, Indian Institute of Technology, Bombay 400 076, India.
** Corresponding author. e-mail: vprakash@iitb.ac.in; r_magar@iitb.ac.in*

In this study, multi-linear regression (MLR) approach is used to construct intermittent reservoir daily inflow forecasting system. To illustrate the applicability and effect of using lumped and distributed input data in MLR approach, Koyna river watershed in Maharashtra, India is chosen as a case study. The results are also compared with autoregressive integrated moving average (ARIMA) models. MLR attempts to model the relationship between two or more independent variables over a dependent variable by fitting a linear regression equation. The main aim of the present study is to see the consequences of development and applicability of simple models, when sufficient data length is available. Out of 47 years of daily historical rainfall and reservoir inflow data, 33 years of data is used for building the model and 14 years of data is used for validating the model. Based on the observed daily rainfall and reservoir inflow, various types of time-series, cause-effect and combined models are developed using lumped and distributed input data. Model performance was evaluated using various performance criteria and it was found that as in the present case, of well correlated input data, both lumped and distributed MLR models perform equally well. For the present case study considered, both MLR and ARIMA models performed equally sound due to availability of large dataset.

1. Introduction

Inflow forecast is a key component in planning, development, design, operation and maintenance of the available water resources. Inflow forecast models are useful in many water resources applications such as flood control, drought management, optimal reservoir operation and hydropower generation. There are many studies pertaining to the reservoir inflow prediction by considering the observed inflow as a time-series (Mays and Tung 1992). However, in many reservoirs in India, especially in intermittent rivers where the only source is monsoon rainfall, the inflow depends heavily on the rainfall and catchment characteristics. It is to be remembered that the transformation of rainfall into runoff involves much highly complex process, such as interception, depression storage, infiltration, overland flow, percolation, evaporation and transpiration (Singh 1988). Over a period of time, the relationship between rainfall and runoff becomes equilibrium when there is not much change in the catchment characteristics, i.e., the pattern between input and output remain similar. Fairly a large number of models have been developed and applied to simulate these processes. According to the use of observational data and description of the physical processes all

Keywords. Multi-linear regression; lumped and distributed data; time-series models; cause-effect models; combined models; ARIMA models; Koyna reservoir inflow; India.

watershed models can be categorized into four broad types (Jothiprakash and Magar 2009):

- Empirical models,
- Conceptual models,
- Physically based models, and
- Data driven models

Each of these type of models has its own advantages and disadvantages (Sorooshian *et al* 1993).

Large numbers of time-series models (Yevjevich 1963; Box and Jenkins 1976; Salas *et al* 1980) are available in the literature. All the time-series models generate the synthetic sequence based on the statistical parameters of the historical data. The results reported in many of the earlier studies are frequently benchmarked against those produced by Box–Jenkins autoregressive integrated moving average (ARIMA) modelling approach (Ahmed and Sarma 2007; Momani and Naill 2009). If fairly longer length of observed data is available, then it is assumed that the catchment characteristics are inherently captured in the observed input series and thus simple regression models could also result in better scenario. Regression is a basic statistical technique for the extrapolation of a dataset to other situations either in time or space. Regression relationship will always be associated with an estimate of the uncertainty associated with the prediction of the dependent variables (Beven 2000). Even though many types of models are available for representing rainfall-runoff process, the problem still remains unresolved and it is perhaps for this reason that the alternative modelling approaches are still being sought along with the empirical models. Black box or the empirical models which attempt to develop relationship among input and output variables provide first-hand information on the rainfall-runoff estimation. There are several empirical and semi-empirical rainfall-runoff models currently in use. Such models are developed by first assuming some mathematical *a priori*, and then the parameters of the model are estimated by minimizing a suitable objective function.

Diskin (1970) viewed a linear regression model as a simple conceptual model and explained the physical meaning of the regression coefficients. Loague and Freeze (1985) used regression, unit hydrograph and quasi-physically based models for upland catchments and concluded that regression models perform marginally better. Driver and Troutman (1989) used linear regression models for estimating urban storm-runoff quantity and quality. The use of various regression models in the data has been dealt in detail by Hirsch (1979) and Hirsch and Gilroy (1984). Chiew *et al* (1993) suggested minimization of the sum of the two objective functions; these objective functions were minimized using the Levenberg–Marquardt (LM) non-linear least-squares algorithm. Raman *et al* (1995) studied five regression models namely, runoff coefficient model, single linear regression, monthly linear regression model, monthly linear regression with stochastic description and double regression models. All these models were used to extend the monthly stream flow data at a site where the available historic rainfall and stream flow data are short for adequate system study. Jagdeesh *et al* (2000) used sum of squares of errors (SSE) between predicted and measured values particularly useful to take low monthly flows into account.

Jain and Prasad (2003) investigated two types of regression models namely, a linear multi-regression and nonlinear multi-regression models to model an event based rainfall-runoff process. Jothiprakash *et al* (2007) developed MLR models based on different input structure combinations namely, cause-effect, time-series and combined for modelling monthly rainfall–runoff relationship for Kanand watershed in Maharashtra, India. Sveinsson *et al* (2008) compared the forecasting performances of autoregressive (AR) model and linear regression model by means of absolute-force error, root-mean-square-force error, Akaike information criteria (AIC), bias, and coefficient of determination (R^2). Only seasonal inflow (May–June–July) was considered disregarding the rest of the time in a year. It was reported that multiple model combination approach was more effective than a single forecast model. The above studies prove that still multi-linear regression (MLR) models are very much sought because of having the advantage of a relationship between input and output. Present study is intended to study the MLR model development and its applicability for reservoir inflow forecasting using fairly longer length of daily lumped and distributed input data (for the same catchment). The main objective of the present study is to develop various types of MLR models using time-series data derived from the Koyna watershed in Maharashtra, India which is an intermittent river. Further ARIMA models also have been developed using time-series lumped data and compared with time-series MLR lumped data models to see the effect of longer length of data.

2. Multi-linear regression models

Since the basic characteristics of the watershed remain unaltered in years, there exists certain correlation between the input and output variables. MLR is the simplest and well developed representation of a casual, time invariant relationship

between an input function of time and corresponding output function. MLR models are considered as benchmark for comparison with other techniques in reservoir inflow forecasting (Chau *et al* 2005). MLR attempts to model the relationship between two or more independent variables and dependent variables by fitting a linear regression equation to observed data. Every value of the independent variable 'x' is associated with a value of the dependent variable 'y'. If y is a dependent variable (expected value) and $x_1, x_2, \ldots, x_n$ are independent variables, then the basic MLR model is given by

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n \qquad (1)$$

where a, b_i = regression constant determined using a least square method.

3. Study area

The area selected for the present study model application is the Koyna watershed, situated on the west coast of Maharashtra, India, lies between the latitude of 17°00′–17°59′N and longitude of 73°02′–73°35′E. The location of the study area along with nine rain-gauge stations in the Koyna watershed is shown in figure 1. The Koyna Dam is one among the 23,000 large dams in the world. The height of the Koyna Dam above foundation level is 103 m and the length of the dam at the crest is about 800 m. The Koyna project is a multipurpose project, but primarily designed as a hydroelectric project that supplies hydro-electric power to Maharashtra, India with an installed capacity of 1920 MW. The Koyna watershed has an elongated leaf shape, about 64 km in length and about 13 km width with an area of 891.78 km^2. The watershed is bounded by hills and broadly consists of 41% forest, 49% cultivated area, 6% waste land and 4% of others (CDO 1992). The water spread area at full reservoir level is 115.36 km^2 which is about 13% of the total catchment area. Nearly 99% of the annual rainfall in this basin occurs during south-west monsoon (June to October) and varies from 2972 to 6694 mm annually over the valley.

Daily rainfall data (January 1961–December 2007) available from nine rain-gauge stations and daily inflow data into the reservoir has been collected from the Koyna Irrigation Division Office, Government of Maharashtra, India and is used in this study. Table 1 shows important statistical properties of the daily rainfall and daily inflow

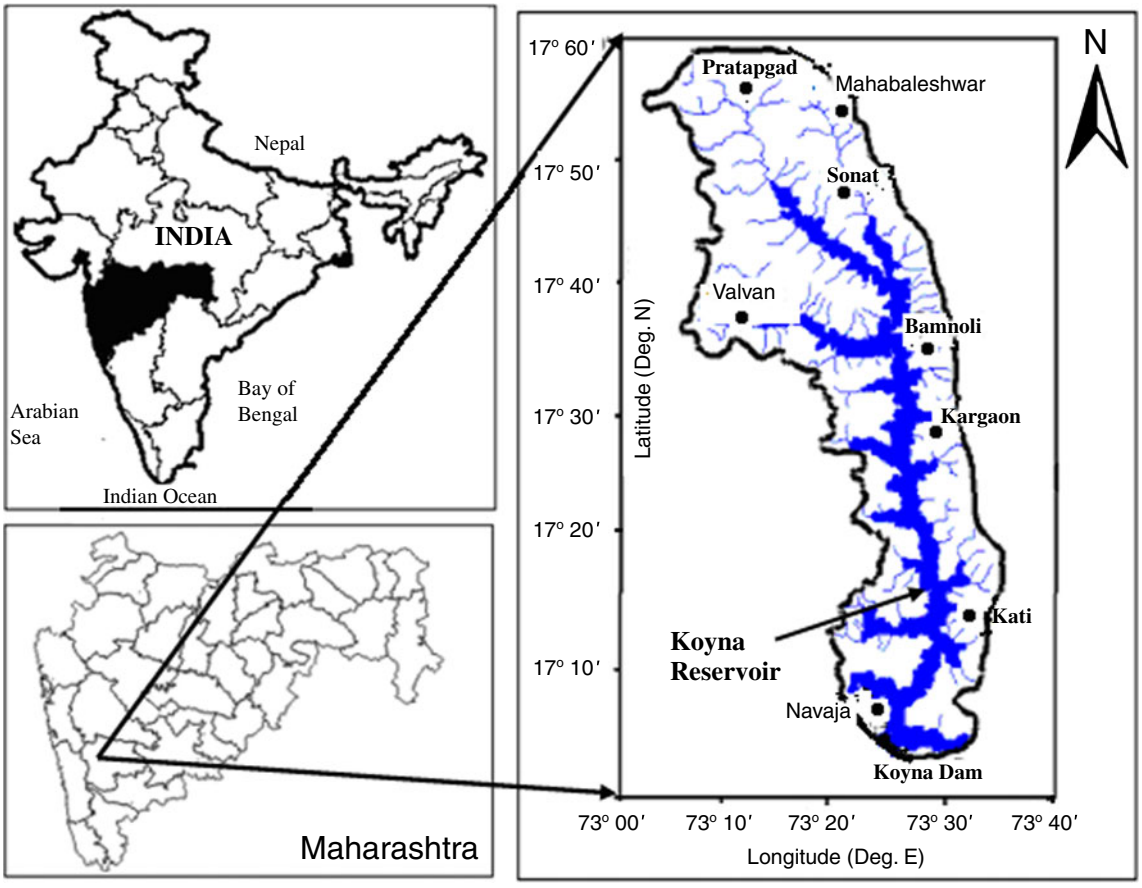

Figure 1. Location of the study area, the Koyna watershed.

Table 1. *Details of rain-gauge stations and statistical properties.*

Sl. no.	Rain-gauge station	Latitude (N)	Longitude (E)	Length of the data available	% Area contribution	Area (km^2)	Average annual rainfall (mm)	Std. dev. (mm)	Skewness	Kurtosis	Coeff. of var.
1	Mahabaleshwar	73° 40′ 21″	17° 55′ 23″	47 years	5.72	50.99	5407.20	1211.99	0.69	1.79	0.22
2	Valvan	73° 35′ 43″	17° 44′ 17″	37 years	13.34	119	3997.35	2236.14	1.38	1.28	0.56
3	Pratapgad	73° 34′ 43″	17° 56′ 02″	41 years	6.20	55.31	4743.68	1026.12	0.20	1.63	0.22
4	Navaja	73° 43′ 24″	17° 25′ 37″	36 years	7.73	68.92	3759.72	1835.49	0.42	1.32	0.49
5	Sonat	73° 42′ 30″	17° 50′ 14″	41 years	16.24	144.83	5669.49	2655.17	0.88	1.82	0.47
6	Kati	73° 49′ 36″	17° 29′ 18″	41 years	14.61	130.29	5323.20	2676.52	0.73	1.73	0.50
7	Kargaon	73° 76′ 47″	17° 39′ 17″	15 years	17.49	155.96	2855.97	1230.34	0.78	1.01	0.43
8	Bamnoli	73° 45′ 43″	17° 43′ 46″	41 years	14.84	132.32	3822.66	1604.43	0.81	1.62	0.42
9	Koyna	73° 44′ 28″	17° 23′ 33″	47 years	3.83	34.16	6315.30	2934.15	0.83	1.89	0.46
Average annual rainfall statistics						891.78	4654.95	1934.48	0.74	1.56	0.42
10	Inflow × 10^6 m^3	73° 45′ 08″	17° 25′ 24″	–	–	–	3808.17	1024.93	1.87	2.01	0.27

Table 2. *Correlation among rainfall-inflow stations (daily data).*

Station	Mahabaleshwar	Valvan	Pratapgad	Navaja	Sonat	Kati	Kargaon	Bamnoli	Koyna	Koyna Inflow
Mahabaleshwar	1.00	–	–	–	–	–	–	–	–	–
Valvan	0.96	1.00	–	–	–	–	–	–	–	–
Pratapgad	0.97	0.96	1.00	–	–	–	–	–	–	
Navaja	0.97	0.96	0.97	1.00	–	–	–	–	–	–
Sonat	0.97	0.97	0.97	0.96	1.00		–	–	–	–
Kati	0.97	0.97	0.97	0.96	0.96	1.00	–	–	–	–
Kargaon	0.89	0.89	0.89	0.90	0.90	0.88	1.00	–	–	–
Bamnoli	0.97	0.96	0.95	0.96	0.98	0.95	0.90	1.00	–	–
Koyna	0.98	0.97	0.95	0.97	0.97	0.96	0.91	0.97	1.00	–
Koyna inflow	0.95	0.90	0.94	0.93	0.91	0.93	0.80	0.90	0.91	1.00

series. The cross correlation among each rain-gauge station as well as inflow station is presented in table 2. From table 2, it can be seen that the correlation among each station is very good revealing that the daily rainfall is uniformly distributed over the catchment.

4. Model development

Historically, observed inflow values represent the hydrological state of the catchment which greatly determines a catchment's response to a rainfall event. Hence both rainfall and inflow are considered as critical input to the model development. In the Koyna catchment, there are nine rain-gauge stations measuring the rainfall data. Hence only rainfall (P) and runoff (Q) data are used for model development. Even though each station has a time-series of data, all the rainfall data are lumped using Thiessen polygon method with respect to time and a single time-series rainfall data has been used to predict the inflow and the model is considered as lumped data models. In fact averaging by Thiessen polygon method produced a smoothing of non-stationarities by averaging the fluctuations recorded at each rain-gauge station (Burlando *et al* 1993; Toth *et al* 2000). Spatially or distributed data models are developed by using the rainfall at each rain-gauge station 'as it is' as the input data. According to different input combinations to the models, various types of models developed in the present study are time-series models (forecasted values are based on observed current and past values), cause-effect models (output, the reservoir inflow is affected by precipitation alone over the entire catchment area) and combined models (output is affected by current and delayed rainfall as well as inflows).

4.1 *Daily lumped input data models*

Initially daily lumped input data is used for model development. Training dataset is a major part of the data that is used for training the network and for finding the governing pattern of the data and should be preferably high for better generalization ability compared to testing the data. The testing patterns are used for evaluating accuracy of the trained model. Training dataset should contain low, medium as well as peak values so that it can capture the fluctuations within the data. Hence, in the present study, the data was divided into two sets: a training set consisting of first 33 years (70%) and a testing set of the remaining 14 years (30%). In the present study, an attempt is made to develop a relationship between the inflow at the catchment outlet (at reservoir), using rainfall and inflow data available up to the current time 't'. Therefore, all developed models are basically approximators of the general function.

$$Q_{(t+n)} = f\left\{P_{\mathrm{obs}(t)}, P_{\mathrm{obs}(t-1)}, ..., P_{\mathrm{obs}(t-m)}, Q_{\mathrm{obs}(t)}, Q_{\mathrm{obs}(t-1)}, ..., Q_{\mathrm{obs}(t-n)}\right\} \quad (2)$$

$P_{\mathrm{obs}(t)}$ and $Q_{\mathrm{obs}(t)}$ represent observed rainfall and inflow during time period 't'. $Q_{(t+n)}$ is inflow to be predicted for the next time step. The prediction is done up to 3 days ahead. Initially ARIMA model has been developed for the observed time-series inflow.

Luk *et al* (2000) and Aqil *et al* (2007) reported that networks trained on transformed data achieve better performance. A logarithmic transformation has been used to bring the observed data to near normal distribution. Transformation is performed on each input output variable independently using the following equations.

$$Z_{p,t} = a \log_{10}\left(P_{\mathrm{obs}(t)} + b\right), \quad (3)$$

$$Z_{q,t} = a \log_{10}\left(Q_{\mathrm{obs}(t)} + b\right). \quad (4)$$

The predicted results were then back-transformed using the following equation

$$Q_{\mathrm{pred}(t)} = 10^{Z_{qt}/a} - b \quad (5)$$

where $Z_{p,t}$, $Z_{q,t}$ are the transformed values of the rainfall and inflow during time period 't', a and b are arbitrary constants assumed as 0.5 and 1, respectively.

Since these coefficients are arrived on trial-and-error basis, until the data followed normal distribution, it is assumed that there is no need of sensitivity analysis using these coefficients. The descriptive statistics of total observed, transformed dataset as well as training and testing dataset are shown in table 3. From table 3, it can be observed that the standard deviation, skewness, kurtosis show very high values in observed data and found to be greatly reduced after logarithmic transformation. From table 3 it can also be observed that the training, testing, and entire dataset (both rainfall and inflow) are statistically similar revealing that the data are from same population and are stationary. This conclusion is supported by many authors (Kisi 2007; Aytek and Alp 2008). The same dataset has been used for both ARIMA as well as MLR model development.

Determination of significant input variables that determines the model *a priori* is one of the most important steps in the linear model development process. Choosing a proper and relevant input variable that have an influence on the output is one of the most essential tasks in developing a

Table 3. *Statistical properties of raw and logarithmic transformed daily dataset.*

Statistical properties	Average daily rainfall (mm)			Inflow ($10^6 m^3$)			
	Entire dataset (observed) (1/1/1961–31/12/2007)	Entire dataset (transformed) (1/1/1961–31/12/2007)	Training dataset (transformed) (1/1/1961–24/11/1993)	Testing dataset (transformed) 25/11/1993–31/12/2007)	Entire dataset (observed) (1/1/1961–31/12/2007)	Entire dataset (transformed) (1/1/1961–31/12/2007)	Training dataset (transformed) (1/1/1961–24/11/1993)	Testing dataset (transformed) 25/11/1993–31/12/2007)
$\overline{X}$	13.223	0.226	0.228	0.224	10.491	0.222	0.223	0.220
S_x	32.424	0.348	0.350	0.344	25.981	0.324	0.314	0.310
C_{sx}	3.761	1.178	1.166	1.206	4.042	1.137	1.119	0.324
K_x	17.552	−0.140	−0.189	−0.020	20.902	−0.109	−0.175	0.048
$X_{\min}$	0	0	0	0	0	0	0	0
$X_{\max}$	343.171	1.276	1.262	1.273	328.581	1.268	1.211	1.262
C_v	2.45	1.536	1.536	1.537	2.472	1.459	1.472	1.454
No. of data	17166	17166	12016	5150	17166	17166	12016	5150
No. of zeros	12587	12587	8875	3712	14472	14472	10227	4245
% of the zeros to the total	73.33	73.33	73.86	72.08	84.31	84.31	85.11	82.43

$\overline{X}$: Mean, S_x: Standard deviation, C_{sx}: Skewness, K_x: Kurtosis, $X_{\min}$: Minimum observed value, $X_{\max}$: Maximum observed value, C_v: Coefficient of variation.

successful forecast of a rainfall-inflow dynamics (Maier and Dandy 1997). Generally, some degree of a *priori* knowledge is used to specify the initial set of inputs (Campolo *et al* 1999; Thirumalaiah and Deo 2000). When the relationship to be modelled is not well understood; then various techniques, such as cross correlation, autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are often employed (Sudheer *et al* 2002; Sivakumar *et al* 2002). Based on cross correlation, ACF and PACF, 20 different lumped data models with various input combinations have been formulated to develop the rainfall inflow relationship. As an example, sample lumped data input models (two in each type) are shown in Appendix I.

4.2 *Daily distributed input data models*

Daily distributed data models have been developed for establishing relationship between rainfall and inflow for the Koyna watershed. Unlike lumped data model, in distributed data model, the input variables from different stations are considered as individual inputs (used 'as it is') for the model development. Hence in this case only cause-effect and combined models have been developed for daily distributed data. The rainfall from nine rain-gauge stations are valued as $P_{1(t)}, P_{2(t)}, P_{3(t)}, P_{4(t)}, P_{5(t)}, P_{6(t)}, P_{7(t)}, P_{8(t)}, P_{9(t)}$ and inflow as $Q_{(t)}$. In this case also logarithmic transformed inputs are used. Twenty different models (same lags as that of lumped data models) with various combinations of input have been formulated to develop the rainfall inflow relationship. As an example, sample distributed data input models (two in cause-effect and one in combined) are shown in Appendix II. The equations given in Appendices I and II are same, the difference is: if the rainfall is lumped then it is one input and if rainfall is distributed then it will be nine inputs for a single lag. The general form of distributed data MLR model is

$$\begin{aligned} Q_{t+1} = f\{&(P_{1(t)}, P_{2(t)}, ..., P_{9(t)}), \\ &(P_{1(t-1)}, P_{2(t-1)}, \ldots, P_{9(t-1)}), \ldots, \\ &(P_{1(t-m)}, P_{2(t-m)}, \ldots, P_{9(t-m)}) \\ &\times Q_{\mathrm{obs}(t)}, Q_{\mathrm{obs}(t-1)}, \ldots, Q_{\mathrm{obs}(t-n)}\}. \end{aligned} \quad (6)$$

5. Model performance criteria

There is no single performance criterion available to select the best model. Many performance criteria are used to select a best model. All the performance criteria are estimated based on

the observed and predicted values. Each performance criteria indicates a particular capability of the model hence various indicators are used. In the present study, correlation coefficient (R), root mean square error (RMSE), Nash–Sutcliffe efficiency (E), Akaike information criteria (AIC) and Bayesian information criteria (BIC) are used and are explained in Appendix III. 'R' is commonly used statistical parameter and provides information on the strength of linear relationship between the observed and the computed values. 'R' can be used to determine whether two ranges of data move together; that is, where large values of one set are associated with large values of other (positive correlation) whether small values of one set are associated with large values of other (negative correlation), or whether values in both sets are unrelated (correlation near zero) (Srinivasulu and Jain 2006).

The Nash–Sutcliffe model efficiency (E) is used to assess the predictive power of hydrological models (Nash and Sutcliffe 1970). It is a normalized statistic that determines the relative magnitude of the residual variance (noise) compared to the observed data variance and indicates how well the plot of observed *versus* predicted data fits the 1:1 line. RMSE is probably the most easily interpreted statistic, since it has the same units as the variable. The RMSE is thus the difference, on average, of an observed data and the estimated variable. The RMSE is specially suited to iterative algorithms and is a better measure for high values. It offers a general picture of the errors involved in prediction. It needs to be noted that the measures involving the error-square terms are also sensitive to extreme values.

In order to make a better balance among generalization ability, parsimony and training speed, two additional indicators, the AIC and BIC, are used (Akaike 1974; Rissanen 1978). The goal is to minimize AIC to obtain a network with best generalization. Although the RMSE statistics are expected to progressively improve as more parameters are added to the model, the AIC and BIC statistics penalize the model for having more parameters and therefore tend to result in more parsimonious models (Hsu *et al* 1995). Model selection is performed by looking for the minimum BIC value. It turns out the final form of this criterion is rather similar to that of AIC but one can see that the penalty due to the number of model parameters is multiplied by 'ln'. As a consequence, BIC learns more than AIC towards lower-dimensional models.

6. Results and discussion

More than 20 developed models have been applied to the Koyna watershed data to select the best model and type of input data to forecast inflow into reservoir. The ARIMA time-series analysis used lags and shifts in the historical data (e.g., moving averages, seasonality) to predict the future values. It is to be noted that separate models are to be developed for different multi-time-step ahead daily inflow prediction. Trial-and-error procedure is adopted to select the best parameters (p, d and q). Various combinations of p, d and q are tried and the models that have resulted in better combination are only presented in table 4 with a lead period of 1 day, 2 days and 3 days. The values of the parameters are chosen such that the sum of squared residuals (SSR) between the observed data and the estimated values and AIC and BIC are as small as possible. The ARIMA models are developed using 70% length of the data and remaining 30% length of data is used for testing. This percentage was arrived after a number of trial-and-error run of the model by using various percentages of data for training and testing. The commercially available software SPSS 16.0 was used for ARIMA model development. The analysis was initialized with one parameter at a time, then their combination and so on.

From table 4, it is apparent that performances of the models are slightly deteriorating with increase in lead time. This may be due to poor correlation of current inflow with 2-day and 3-day lagged inflows. The prediction of 1 day ahead inflow is quite satisfactory because the input space contains the most recent information. It can also be observed that ARIMA(2, 1, 2) model performed better than any other combination for all lead periods and obtained the best statistics, i.e., maximum of R(0.66), E(0.56) and minimum of RMSE(14.99), AIC(13945.02) and BIC(13951.57). It is also found that the model performance is not increasing with the increase in p, q and d. The time-series and scatter plot of observed and predicted inflow (1 day lead period) during testing period resulted from ARIMA(2, 1, 2) model is presented in figure 2(a) and (b), respectively. From the time-series and scatter plot, it can be seen that only low flows are predicted reasonably accurate, medium inflows are overpredicted and high inflows are underpredicted. The reason may be due to non-linear behaviour of medium and high inflows. Nevertheless ARIMA model can provide first-hand information about the process of inflow prediction. Further models are developed using MLR techniques.

6.1 *Daily lumped data MLR models*

The commercially available software SPSS 16.0 has been used for MLR model development also. The performance of the lumped MLR models during

Table 4. *Performance measures of daily time-step ARIMA models.*

Models	Performance criteria	Development			Testing		
		Lead period			Lead period		
		1 day	2 days	3 days	1 day	2 days	3 days
ARIMA 1–1–1	R	0.62	0.56	0.52	0.60	0.54	0.50
	E	0.58	0.55	0.47	0.57	0.51	0.45
	RMSE	17.89	17.95	18.02	18.23	18.32	18.37
	AIC	34656.17	34696.40	34743.16	14952.8	14978.17	14992.21
	BIC	34663.56	34703.79	34753.44	14959.35	14984.71	14998.75
ARIMA 2–2–2	R	0.65	0.59	0.56	0.62	0.57	0.54
	E	0.57	0.56	0.52	0.55	0.52	0.48
	RMSE	14.97	15.13	15.05	15.32	15.43	15.49
	AIC	32515.17	32642.90	32579.21	14057.17	14094.02	14114.00
	BIC	32522.56	32650.30	32589.31	14063.72	14100.56	14120.55
ARIMA 1–2–1	R	0.63	0.58	0.56	0.64	0.57	0.54
	E	0.54	0.56	0.55	0.51	0.54	0.50
	RMSE	15.12	15.33	15.59	15.46	15.49	15.49
	AIC	32634.96	32800.69	33002.76	14104.02	14114	14114.00
	BIC	32642.35	32808.08	33012.90	14110.57	14120.55	14120.55
ARIMA 1–2–2	R	0.61	0.58	0.50	0.59	0.52	0.50
	E	0.52	0.49	0.45	0.51	0.50	0.43
	RMSE	15.26	15.78	15.88	15.31	15.39	15.46
	AIC	32745.70	33148.3	33224.20	14053.81	14080.65	14104.02
	BIC	32753.09	33155.69	33234.36	14060.35	14087.19	14110.57
ARIMA 2–1–1	R	0.64	0.55	0.51	0.63	0.52	0.50
	E	0.59	0.52	0.45	0.58	0.50	0.43
	RMSE	15.01	15.03	15.10	15.14	15.20	15.30
	AIC	32547.23	32563.23	32619.06	13996.3	14016.67	14050.44
	BIC	32554.62	32570.62	32629.17	14002.85	14023.22	14056.99
ARIMA 2–1–2	**R**	**0.64**	**0.60**	**0.62**	**0.66**	**0.62**	**0.60**
	E	**0.58**	**0.50**	**0.51**	**0.56**	**0.53**	**0.55**
	RMSE	**14.01**	**14.12**	**14.19**	**14.99**	**15.12**	**15.19**
	AIC	**31718.85**	**31812.82**	**31872.24**	**13945.02**	**13989.49**	**14013.28**
	BIC	**31726.25**	**31820.21**	**31882.28**	**13951.57**	**13996.04**	**14019.83**
ARIMA 3–2–2	R	0.45	0.47	0.44	0.43	0.45	0.43
	E	0.47	0.42	0.41	0.45	0.41	0.39
	RMSE	15.18	15.32	15.72	16.10	16.12	16.22
	AIC	32682.54	32792.85	33102.53	14312.92	14319.31	14351.16
	BIC	32689.94	32800.24	33112.68	14319.47	14325.86	14357.71
ARIMA 4–1–4	R	0.45	0.43	0.41	0.43	0.41	0.39
	E	0.42	0.40	0.40	0.40	0.40	0.45
	RMSE	18.38	18.54	18.61	18.08	18.10	18.21
	AIC	34980.83	35084.97	35130.24	14910.25	14915.95	14947.15
	BIC	34988.22	35092.36	35140.56	14916.8	14922.49	14953.70

training and testing period is presented in table 5. From table 5, it can be observed that the performance of each model during training and testing is similar indicating that the models are not overfitted and also the results are consistent and encouraging. The reason may be due to the statistical properties of training dataset and testing dataset are similar and length of input data used for model development is sufficiently longer. From table 5 it is apparent that the performances of all models are slightly deteriorating with increase in lead time. As the forecast lead period increases the correlation between desirable output and given input decreases leading to poor prediction.

For comparison, the performance of time-series models (DL-MLR model 1 to DL-MLR model 7) listed in table 5 are considered. From this it is observed that model performance is increasing up to a lagged input of six and then slightly decreases. Correlation coefficient (R) and

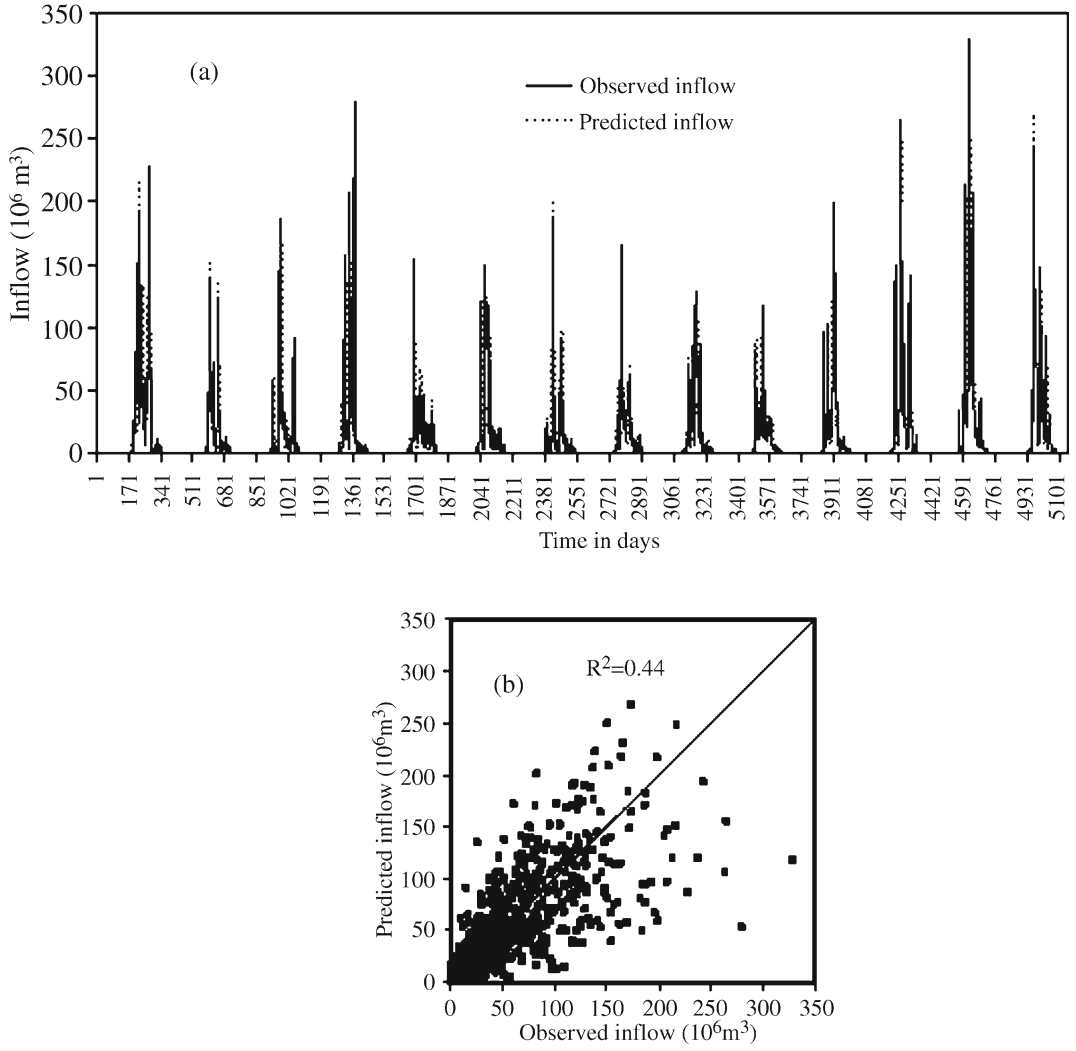

Figure 2. (**a**) Time-series and (**b**) scatter plot of ARIMA(2, 1, 2) model during testing period.

Nash–Sutcliffe efficiency (E) is gradually increasing and the RMSE, AIC, BIC values are decreasing with increase in number of inputs. Among seven models, DL-MLR model 6 with 1 day lead period which used input structure of $Q_{(t-5)}$, $Q_{(t-4)}, Q_{(t-3)}, Q_{(t-2)}, Q_{(t-1)}, Q_{(t)}$ has yielded a maximum R (0.67) and E (0.59) values and minimum RMSE (16.21), AIC (14349.81) and BIC (14356.35) values. Since AIC and BIC values are minimum than any other model, DL-MLR model 6 may be considered as a parsimonious model. From this result it is found that this lumped time-series MLR model behave same as that of ARIMA(2–1–2) model. This may be due to large dataset, which might have completely captured the stochasticity. However, overall the performance is not convincing, hence to improve the performance further, the causing parameter, viz., rainfall is introduced as the input in the model development and is named as cause-effect MLR models.

From table 5, it can be seen that the DL-MLR model 8 to DL-MLR model 16 are cause-effect models. In this model type, it is assumed that the output (inflow in this case) is caused by the lumped rainfall (exogenous input parameter) over the entire catchment area. In this type, the models are redeveloped and from the performances during training and testing it is found that there is gradual improvement with increase in numbers of input up to 7-day lags (i.e., from DL-model 8 to DL-model 15) and thereafter the performance is decreasing. Among the cause-effect models DL-MLR model 15 which used 8 inputs has obtained best statistics than any other model. In this type also the model performance is deteriorating as lead time increases from 1 day to 3 days. This could be attributed to the low dependency between the values separated by higher lags. It is also found that the time-series models are performing better than cause-effect models. The reason may be due

Table 5. *Performance measures of daily lumped data MLR models.*

Models	No. of input variables	Performance criteria	Training			Testing		
			Lead period			Lead period		
			1 day	2 days	3 days	1 day	2 days	3 days
Time-series models								
DL-MLR	1	R	0.56	0.53	0.50	0.51	0.48	0.50
Model 1		E	0.54	0.53	0.48	0.50	0.47	0.52
		RMSE	17.98	18.21	18.47	18.16	18.31	18.52
		AIC	34724.36	34878.17	35042.59	14935.44	14977.69	15036.78
		BIC	34731.75	34885.56	35049.99	14941.99	14984.23	15043.33
DL-MLR	2	R	0.62	0.58	0.55	0.58	0.48	0.50
Model 2		E	0.56	0.55	0.50	0.47	0.47	0.52
		RMSE	17.64	17.37	18.19	17.89	18.20	18.25
		AIC	34496.71	34307.06	34863.49	14857.04	14946.96	14960.83
		BIC	34504.11	36473.33	34870.88	14863.58	14953.51	14967.37
DL-MLR	3	R	0.64	0.62	0.59	0.61	0.59	0.55
Model 3		E	0.61	0.53	0.51	0.58	0.52	0.51
		RMSE	17.59	17.66	18.01	17.67	17.98	18.23
		AIC	34456.66	34505.58	34743.27	14793.94	14883.04	14952.86
		BIC	34464.05	34512.97	34750.66	14800.49	14889.59	14959.40
DL-MLR	4	R	0.66	0.63	0.60	0.62	0.60	0.57
Model 4		E	0.63	0.60	0.58	0.59	0.55	0.54
		RMSE	16.70	16.99	17.00	17.06	17.35	17.92
		AIC	33833.89	34045.40	34050.67	14613.81	14699.78	14865.31
		BIC	33841.28	34052.87	34058.07	14620.36	14706.32	14871.85
DL-MLR	5	R	0.67	0.65	0.61	0.65	0.64	0.60
Model 5		E	0.63	0.62	0.59	0.63	0.62	0.62
		RMSE	16.37	16.64	16.96	16.90	16.96	17.34
		AIC	33592.20	33790.66	34022.35	14565.01	14581.17	14697.64
		BIC	33599.60	33798.06	34029.75	14571.55	14587.72	14704.18
DL-MLR	6	R	0.70	0.68	0.64	*0.67*	0.62	0.59
Model 6		E	0.66	0.60	0.53	*0.59*	0.57	0.56
		RMSE	15.99	17.06	18.56	*16.21*	17.23	18.83
		AIC	33316.25	34090.42	35106.71	*14349.81*	14664.83	15121.76
		BIC	33323.65	34097.82	35114.10	*14356.35*	14671.38	15128.30
DL-MLR	7	R	0.65	0.62	0.58	0.61	0.57	0.50
Model 7		E	0.61	0.58	0.55	0.58	0.56	0.45
		RMSE	15.25	16.12	17.38	15.31	15.68	15.99
		AIC	32745.65	33410.11	34312.63	14055.61	14176.87	14280.63
		BIC	32753.04	33417.51	34320.03	14062.15	14183.41	14287.18
Cause-effect models								
DL-MLR	1	R	0.55	0.52	0.50	0.52	0.50	0.45
Model 8		E	0.45	0.43	0.44	0.40	0.37	0.35
		RMSE	19.17	19.23	19.25	19.46	20.14	20.80
		AIC	35495.64	35530	35544.27	15291.02	15467.53	15634.03
		BIC	35503.03	35537.4	35551.67	15297.57	15474.08	15640.58
DL-MLR	2	R	0.58	0.55	0.52	0.55	0.52	0.48
Model 9		E	0.55	0.54	0.42	0.46	0.50	0.47
		RMSE	19.17	19.23	19.25	19.46	20.14	20.80
		AIC	35495.64	35530	35544.27	15291.02	15467.53	15634.03
		BIC	35503.03	35537.4	35551.67	15297.57	15474.08	15640.58
DL-MLR	3	R	0.60	0.58	0.55	0.57	0.54	0.50
Model 10		E	0.58	0.56	0.45	0.48	0.45	0.48
		RMSE	19.12	19.26	19.18	19.23	19.36	19.69
		AIC	35462.89	35546.22	35497.44	15229.13	15262.94	15351.46
		BIC	35470.28	35553.61	35504.83	15235.68	15269.49	15358

Table 5. *(Continued.)*

Models	No. of input variables	Performance criteria	Training Lead period 1 day	Training Lead period 2 days	Training Lead period 3 days	Testing Lead period 1 day	Testing Lead period 2 days	Testing Lead period 3 days
DL-MLR	4	R	0.62	0.58	0.56	0.58	0.55	0.53
Model 11		E	0.67	0.55	0.46	0.49	0.47	0.49
		RMSE	18.86	19.20	19.00	19.12	19.41	19.72
		AIC	35298.25	35513.74	35386.87	15200.37	15277.39	15358.08
		BIC	35305.64	35521.13	35394.26	15206.91	15283.94	15364.63
DL-MLR	5	R	0.62	0.58	0.56	0.59	0.55	0.54
Model 12		E	0.66	0.55	0.47	0.50	0.48	0.50
		RMSE	19.12	19.33	19.03	19.17	19.15	19.94
		AIC	35462.89	35591.07	35406.46	15213.91	15208.17	15416.98
		BIC	35470.28	35598.47	35413.86	15220.46	15214.71	15423.53
DL-MLR	6	R	0.63	0.59	0.58	0.58	0.56	0.56
Model 13		E	0.62	0.56	0.49	0.52	0.49	0.52
		RMSE	18.92	19.41	19.15	19.18	19.41	19.46
		AIC	35331.74	35641.18	35477.81	15215.17	15277.39	15291.02
		BIC	35339.13	35648.57	35485.21	15221.72	15283.94	15297.57
DL-MLR	7	R	0.63	0.60	0.55	0.62	0.57	0.55
Model 14		E	0.61	0.56	0.50	0.54	0.50	0.52
		RMSE	18.89	19.46	19.20	19.30	18.98	19.04
		AIC	35313.25	35672.66	35513.74	15249.17	15162.72	15179.16
		BIC	35320.64	35680.05	35521.13	15255.71	15169.27	15185.71
DL-MLR	8	R	0.66	0.64	0.60	*0.63*	0.58	0.56
Model 15		E	0.61	0.60	0.58	*0.57*	0.50	0.53
		RMSE	14.95	15.87	17.63	*17.37*	17.77	17.98
		AIC	32503.09	33220.9	34488.8	*14706.18*	14822.06	14882.32
		BIC	32510.48	33228.3	34496.19	*14712.73*	14828.60	14888.87
DL-MLR	9	R	0.54	0.53	0.52	0.50	0.50	0.45
Model 16		E	0.45	0.44	0.45	0.45	0.38	0.35
		RMSE	19.71	20.07	18.021	18.53	18.60	18.89
		AIC	35827.94	36042.87	34746.78	15038.36	15056.43	15137.03
		BIC	35835.33	32791.42	34754.18	15044.90	15062.98	15143.57
Combined models								
DL-MLR	2	R	0.70	0.65	0.63	0.65	0.62	0.60
Model 17		E	0.63	0.60	0.62	0.57	0.61	0.59
		RMSE	14.10	15.31	18.02	18.53	18.60	18.89
		AIC	31802.67	32791.42	34746.78	15038.36	15056.43	15137.03
		BIC	31810.06	32798.81	34754.18	15044.90	15062.98	15143.57
DL-MLR	**3**	**R**	**0.84**	**0.82**	**0.77**	**0.80**	**0.79**	**0.74**
Model 18		**E**	**0.69**	**0.68**	**0.57**	**0.66**	**0.62**	**0.54**
		RMSE	**14.57**	**16.06**	**16.70**	**14.83**	**16.64**	**17.29**
		AIC	**32199.94**	**33363.02**	**33833.89**	**13891.64**	**14483.66**	**14680.47**
		BIC	**32207.33**	**33370.41**	**33841.28**	**13898.19**	**14490.21**	**14687.02**
DL-MLR	5	R	0.67	0.65	0.62	0.71	0.68	0.65
Model 19		E	0.64	0.61	0.56	0.65	0.62	0.59
		RMSE	14.20	15.24	18.02	16.93	15.32	16.05
		AIC	31890.20	32734.80	34746.78	14574	14057.80	14299.57
		BIC	31897.60	32742.19	34754.18	14580.55	14064.35	14306.12
DL-MLR	7	R	0.65	0.62	0.60	0.69	0.66	0.63
Model 20		E	0.62	0.59	0.50	0.66	0.65	0.61
		RMSE	14.55	14.91	18.29	16.33	15.64	15.74
		AIC	32180.12	32470.75	34928.99	14388.10	14165.53	14197.94
		BIC	32187.51	32478.14	34936.38	14394.65	14172.08	14204.48

to the better autocorrelation of inflow data than the serial correlation with rainfall.

Since it is found that either rainfall data (causing variable) alone or inflow data (effective variable) alone is insufficient to reproduce the inflows in an effective way, both rainfall and inflow are given as input and named as combined models. The performances of the developed combined DL-MLR models, i.e., DL-MLR model 17 to DL-MLR model 20 are shown in table 5. However, among the combined models, DL-MLR model 18 with 1 day ahead which used input structure as $P_{(t-1)}$, $P_{(t)}$ and $Q_{(t)}$ is showing better performance R (0.80), E (0.66). Hence, DL-MLR model 18 is selected as best MLR model among lumped input data MLR models (including time-series and cause-effect models). It is also to be noted that the number of data points during training and testing are different leading to skewed AIC and BIC values during training and testing. The combined input is responsible for the reduction in RMSE, AIC, BIC to greater extent. Thus it may be concluded that while developing a lumped reservoir inflow prediction model, having impulse response to rainfall, combined input models may result in better scenario. The two dimensions, i.e., rainfall and inflow as input has captured the non-linearity nature of inflow.

Figure 3 shows the scatter plots between the actual observed inflow and corresponding predictions by combined DL-MLR model 18 for different lead times of 1 day, 2 days and 3 days during testing period. Visual inspection of these figures reveals that the performances of the models are deteriorating with increase in lead period, especially peak values. In all the MLR models, 1 day ahead prediction was found to produce more acceptable results, may be because of higher correlation with 1 day ahead input and output. The poor performance of higher lead period may be due to non-linear relationship between current inflow and higher order input variables.

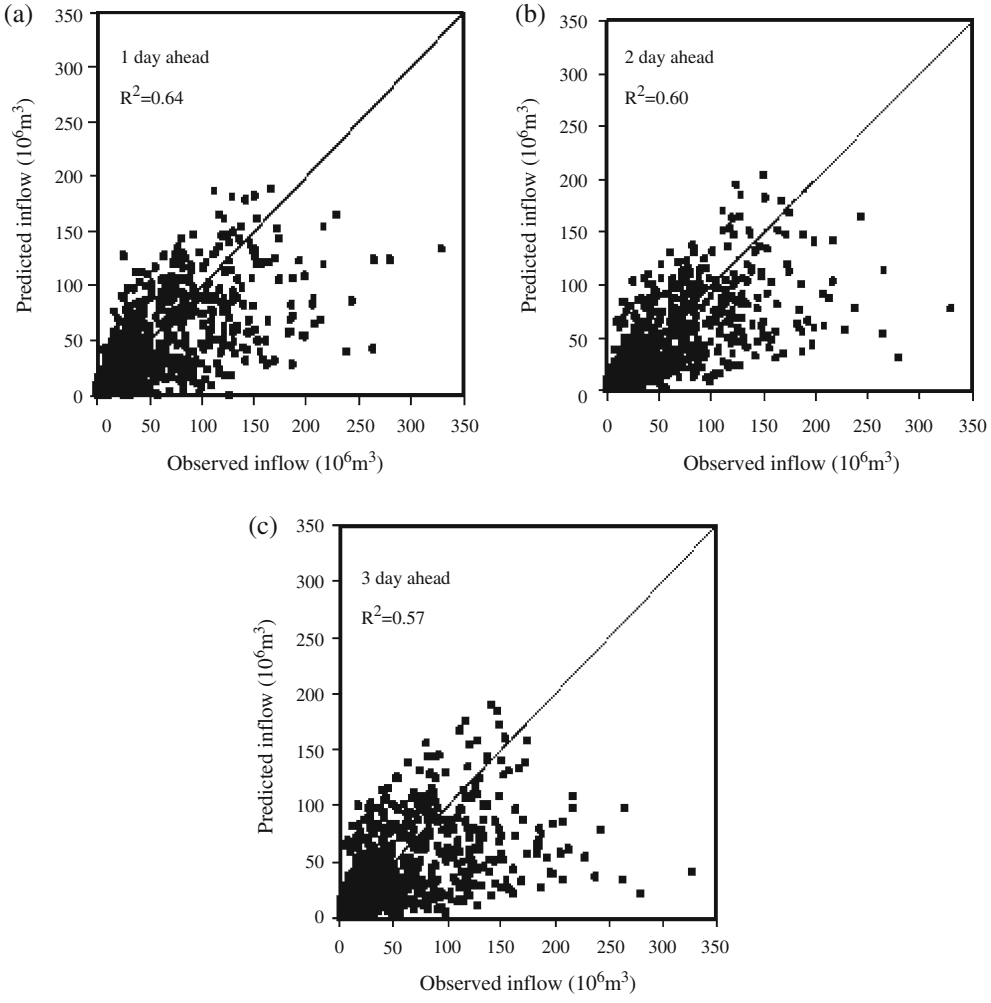

Figure 3. Scatter plot of observed and multi-time-step ahead predicted inflow by DL-MLR model 18 during testing period (combined input).

Table 6. *Performance measures of daily distributed data MLR models.*

Models	No. of input variables	Performance criteria	Training			Testing		
			Lead period			Lead period		
			1 day	2 days	3 days	1 day	2 days	3 days
Cause-effect models								
DD-MLR	9	R	0.45	0.42	0.40	0.42	0.41	0.45
Model 1		E	0.42	0.40	0.42	0.35	0.38	0.35
		RMSE	21.15	22.27	23.14	22.98	23.87	25.76
		AIC	11705.04	11902.93	12049.89	5152.19	5214.62	5339.82
		BIC	11711.29	11909.18	12056.14	5157.59	5220.02	5345.22
DD-MLR	18	R	0.51	0.48	0.42	0.48	0.43	0.46
Model 2		E	0.45	0.46	0.38	0.42	0.40	0.42
		RMSE	20.16	21.98	22.76	21.87	22.16	23.80
		AIC	11521.19	11852.66	11986.39	5070.85	5092.49	5339.82
		BIC	11527.44	11858.91	11992.64	5076.25	5220.02	5345.22
DD-MLR	27	R	0.55	0.52	0.50	0.50	0.48	0.43
Model 3		E	0.51	0.46	0.48	0.45	0.42	0.40
		RMSE	19.12	19.30	20.21	20.54	21.32	22.87
		AIC	11318.07	11354.00	11530.69	4967.76	5029.00	5144.30
		BIC	11324.32	11360.25	11536.94	4973.17	5034.40	5149.71
DD-MLR	36	R	0.60	0.58	0.53	0.58	0.55	0.53
Model 4		E	0.57	0.55	0.44	0.52	0.51	0.49
		RMSE	19.86	21.87	22.54	19.87	19.98	20.32
		AIC	11463.69	11833.42	11949.14	4913.27	4922.34	4950.07
		BIC	11469.95	11839.67	11955.39	4918.68	4927.75	4955.47
DD-MLR	45	R	0.64	0.62	0.59	*0.62*	0.56	0.55
Model 5		E	0.59	0.56	0.55	*0.55*	0.50	0.52
		RMSE	15.92	16.23	18.61	*18.33*	18.75	18.91
		AIC	10615.65	10689.61	11214.39	*4780.73*	4817.95	4831.91
		BIC	10621.91	10695.87	11220.64	*4786.13*	4823.36	4837.32
DD-MLR	54	R	0.58	0.55	0.50	0.50	0.48	0.43
Model 6		E	0.55	0.52	0.48	0.45	0.42	0.40
		RMSE	18.12	19.30	20.21	20.54	21.32	22.87
		AIC	11112.06	11354.00	11530.69	4967.76	5029.00	5144.30
		BIC	11118.31	11360.25	11536.94	4973.17	5034.40	5149.71
DD-MLR	63	R	0.54	0.50	0.50	0.52	0.51	0.42
Model 7		E	0.48	0.46	0.48	0.43	0.40	0.39
		RMSE	18.98	19.01	20.21	20.14	21.12	22.67
		AIC	11289.88	11295.94	11530.69	4935.45	5013.51	5129.87
		BIC	11296.14	11302.19	11536.94	4940.85	5018.92	5135.28
Combined models								
DD-MLR	10	R	0.65	0.62	0.60	0.61	0.56	0.54
Model 8		E	0.60	0.52	0.50	0.58	0.45	0.46
		RMSE	19.82	19.61	19.11	18.89	19.27	19.76
		AIC	11455.96	11415.11	11316.06	4830.17	4862.90	4904.15
		BIC	11462.21	11421.36	11322.31	4835.58	4868.30	4909.56
DD-MLR	19	R	0.67	0.65	0.62	0.64	0.59	0.55
Model 9		E	0.63	0.55	0.55	0.61	0.58	0.48
		RMSE	18.15	18.69	18.91	18.89	18.98	17.31
		AIC	11118.40	11230.84	11275.71	4830.17	4837.98	4686.66
		BIC	11124.65	11237.09	11281.97	4835.58	4843.39	4692.06
DD-MLR	20	R	0.69	0.66	0.61	0.63	0.61	0.59
Model 10		E	0.65	0.57	0.54	0.60	0.58	0.52
		RMSE	16.08	16.34	16.57	18.29	18.68	18.35
		AIC	10654.00	10715.52	10769.12	4777.14	4811.81	4782.52
		BIC	10660.26	10721.77	10775.37	4782.54	4817.21	4787.93

Table 6. *(Continued.)*

Models	No. of input variables	Performance criteria	Training			Testing		
			Lead period			Lead period		
			1 day	2 days	3 days	1 day	2 days	3 days
DD-MLR	21	R	0.72	0.69	0.63	0.69	0.65	0.61
Model 11		E	0.69	0.59	0.56	0.63	0.58	0.56
		RMSE	15.98	15.14	14.23	15.66	16.54	17.54
		AIC	10630.08	10423.00	10185.28	4522.07	4611.90	4708.35
		BIC	10636.33	10429.25	10191.53	4527.48	4617.30	4713.75
DD-MLR	22	R	0.73	0.68	0.59	0.69	0.66	0.63
Model 12		E	0.61	0.57	0.52	0.56	0.64	0.56
		RMSE	16.29	17.38	17.75	16.29	16.76	17.89
		AIC	10703.76	10952.15	11032.94	4586.88	4633.61	4740.81
		BIC	10710.02	10958.40	11039.19	4592.28	4639.01	4746.21
DD-MLR	23	R	0.70	0.65	0.57	0.62	0.60	0.58
Model 13		E	0.60	0.55	0.50	0.55	0.54	0.54
		RMSE	16.60	17.67	17.94	16.41	16.68	18.03
		AIC	10776.06	11015.61	11073.77	4598.93	4625.75	4753.62
		BIC	10782.31	11021.87	11080.02	4604.34	4631.15	4759.02
DD-MLR	24	R	0.72	0.70	0.68	0.68	0.65	0.63
Model 14		E	0.58	0.65	0.62	0.61	0.63	0.67
		RMSE	15.89	16.04	16.76	15.89	16.04	16.76
		AIC	10608.42	10644.45	10812.85	4546.03	4561.47	4633.61
		BIC	10614.67	10650.71	10819.10	4551.43	4566.87	4639.01
DD-MLR	25	R	0.76	0.72	0.66	0.75	0.71	0.70
Model 15		E	0.70	0.65	0.68	0.71	0.69	0.66
		RMSE	16.87	16.14	16.23	16.66	16.38	17.32
		AIC	10837.93	10668.29	10689.61	4623.78	4595.93	4687.61
		BIC	10844.19	10674.54	10695.87	4629.18	4601.33	4693.01
DD-MLR	28	R	0.75	0.72	0.66	0.75	0.71	0.70
Model 16		E	0.72	0.65	0.68	0.71	0.69	0.66
		RMSE	16.51	16.64	16.73	18.07	16.93	16.87
		AIC	10755.21	10785.29	10805.98	4757.26	4650.19	4644.36
		BIC	10761.46	10791.54	10812.23	4762.66	4655.59	4649.76
DD-MLR	**29**	**R**	**0.80**	**0.82**	**0.75**	**0.76**	**0.72**	**0.70**
Model 17		**E**	**0.67**	**0.68**	**0.65**	**0.62**	**0.61**	**0.52**
		RMSE	**16.03**	**16.62**	**18.56**	**16.87**	**19.25**	**21.76**
		AIC	**10642.06**	**10780.68**	**11204.07**	**4644.36**	**4861.19**	**5062.56**
		BIC	**10648.31**	**10786.93**	**11210.32**	**4649.76**	**4866.59**	**5067.96**
DD-MLR	30	R	0.76	0.73	0.67	0.72	0.65	0.62
Model 18		E	0.62	0.60	0.59	0.58	0.57	0.51
		RMSE	18.03	18.62	19.03	18.06	18.65	18.98
		AIC	11092.96	11216.45	11299.97	4756.35	4809.17	4837.98
		BIC	11099.21	11222.70	11306.23	4761.75	4814.57	4843.39
DD-MLR	39	R	0.70	0.66	0.63	0.65	0.60	0.57
Model 19		E	0.60	0.57	0.54	0.58	0.58	0.50
		RMSE	18.87	19.65	19.76	19.05	19.62	19.87
		AIC	11267.59	11422.93	11444.33	4844.03	4892.47	4913.27
		BIC	11273.85	11429.18	11450.59	4849.44	4897.88	4918.68
DD-MLR	48	R	0.68	0.61	0.57	0.61	0.57	0.55
Model 20		E	0.59	0.55	0.54	0.55	0.53	0.49
		RMSE	15.76	17.21	18.98	19.52	19.66	19.89
		AIC	10576.92	10914.46	11289.88	4884.08	4895.82	4914.93
		BIC	10583.17	10920.71	11296.14	4889.48	4901.22	4920.33

6.2 *Daily distributed data MLR models*

The daily distributed input MLR models are developed as discussed above. The distributed data model has increased number of parameters to be estimated for same number of lagged input as that of lumped data. This leads to increased complexity in model parameter estimation. Also the number of parameters became laborious and time consuming with lengthy equations while using it for testing. The other limitations of DD-MLR model is that the length of data used is 15 years (common for all stations). The performances of distributed data MLR models during training and testing period are analyzed and are depicted in table 6. From table 6, it is also noticed that the performances of all the models are slightly deteriorating when lead time is increased from 1 day to 3 days. On studying the distributed cause-effect models, namely, DD-MLR model 1 to DD-MLR model 7 (table 6), it can be observed that the performance of the models during training and testing are comparable and there is gradual improvement in performances of input up to 4-day lags (i.e., from DD-MLR model 1 to DD-MLR model 5) and thereafter the performance has slightly deteriorated. However, the performance of DD-MLR cause-effect models is inferior to DL-MLR as well as ARIMA models. This indicates that the accuracy of the parameter governs the prediction results in linear models.

Analyzing the results of combined models in table 6 (DD-MLR model 8 to DD-MLR model 20) it is apparent that all the combined models show satisfactory results during training and testing. However, combined DD-MLR model 17 outperformed all the models. Combined DD-MLR model 17 with 1 day lead period during testing which used 29 input variables showed best performances (testing) as evident from highest R (0.76) and E (0.62). Hence, DD-MLR model 17 is selected as the best model among combined models as well as among cause-effect models. In comparison with best cause-effect DD-MLR model 5 to best combined model DD-MLR model 17 'R' value is increased from 0.62 to 0.76 and 'E' value is increased from 0.55 to 0.62.

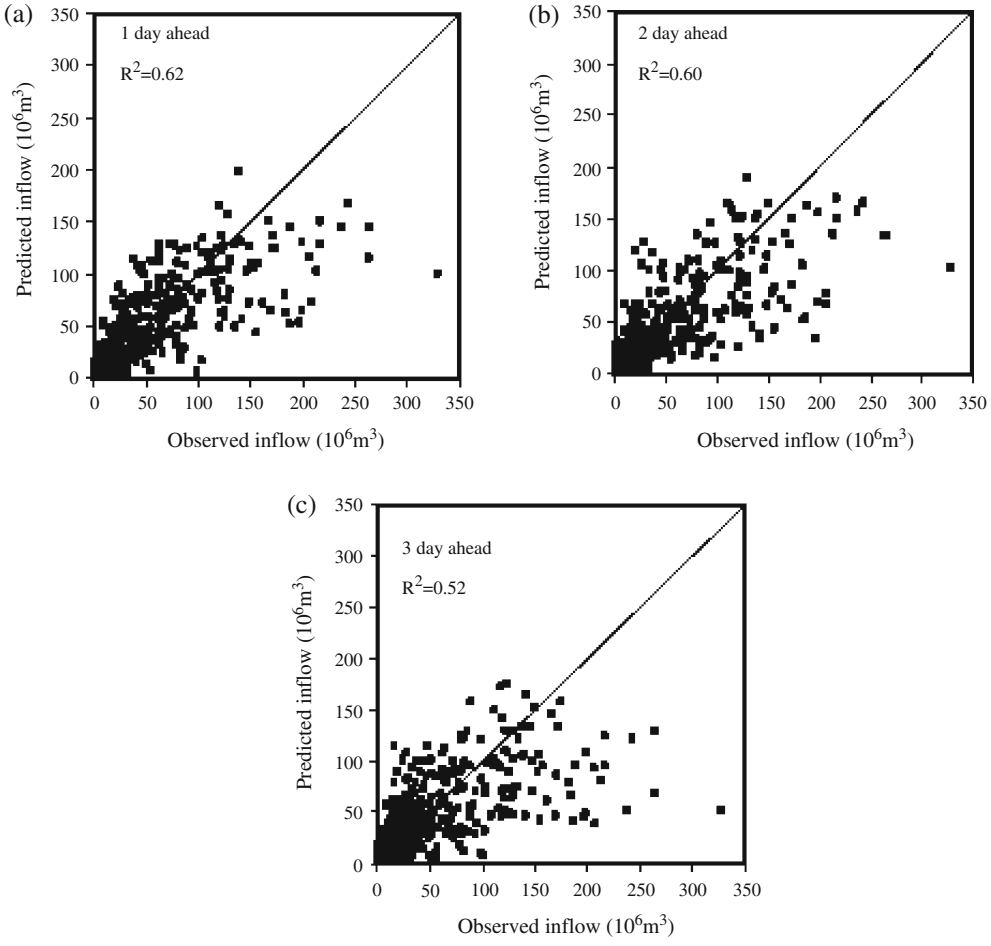

Figure 4. Scatter plot of observed and multi-time-step ahead predicted inflow by best DD-MLR model 17 during testing (combined input).

The number of input variables are reduced from 45 to 29 which show that distributed combined models performed better than distributed cause-effect models even with lesser number of inputs. The scatter plot of observed inflow and the predicted inflow from combined DD-MLR model 17 during testing period with lead period of 1 day, 2 days and 3 days is shown in figure 4. Visual inspection of figures reveals that as the lead period increases the performance deteriorated. The identified MLR model performed fairly better in the prediction of low and medium inflow but failed in prediction of non-linear peak inflows.

7. Conclusions

This study investigated the applicability and capability of multi-linear regression models in inflow forecasting for the Koyna watershed in Maharashtra, India into the Koyna reservoir. The dataset includes daily rainfall and inflow data for a period of 47 years. Seventy percent of dataset are used for model building and remaining is used for testing the models. Twenty models are developed based on different input structure combination as time-series, cause-effect and combined models and also their performances are evaluated and tested. The results of time-series input models are compared with ARIMA models. Based on the results it may be concluded that deterministic and stochastic models perform equally good, if data length is sufficiently longer. For lumped data, DL-MLR model 18 with input combination of *P(t−1), P(t), Q(t)* showed better performance with a highest R of 0.80 during testing. However, for distributed data models, DD-MLR model 17 having 29 inputs showed better performance of R (0.76) and E (0.62) during testing. It is also observed that even though performances of both lumped and distributed data are encouraging, lumped daily data models slightly performed better. The reason may be length of the data available for distributed model is less than lumped data models as well large number of input variables increased the complexity of the model and reduced the performance.

8. Practical significance

Even though large numbers of sophisticated rainfall runoff models are available in the literature, many suffer from various drawbacks such as data intensive; require large computational time and high skill for successful application. However, field engineers are familiar with the application of empirical models developed for particular watershed. Hence the approach of this study will help field engineers to develop such types of models. The best MLR equation is currently in use by the dam authorities at dam site.

Acknowledgements

The authors express their sincere thanks to the Executive Engineer, Irrigation Department and office personnel of the Koyna Dam Division, Government of Maharashtra, India for providing necessary data to carry out this work. The authors also acknowledge the Ministry of Water Resources, Government of India, New Delhi for sponsoring the research project through Indian National Committee on Hydrology.

Appendix I

Table A1. *Model types and input combinations (daily lumped data).*

Model type	Input variables	No. of input variables	Output variables		
Time-series models					
DL Model 3	$Q_{(t-2)}, Q_{(t-1)}, Q_{(t)}$	3	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$
DL Model 6	$Q_{(t-5)}, Q_{(t-4)}, Q_{(t-3)}, Q_{(t-2)}, Q_{(t-1)}, Q_{(t)}$	6	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$
Cause-effect models					
DL Model 10	$P_{(t-2)}, P_{(t-1)}, P_{(t)}$	3	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$
DL Model 14	$P_{(t-6)}, P_{(t-5)}, P_{(t-4)}, P_{(t-3)}, P_{(t-2)}, P_{(t-1)}, P_{(t)}$	7	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$
Combined models					
DL Model 19	$P_{(t-2)}, P_{(t-1)}, P_{(t)}, Q_{(t-1)}, Q_{(t)}$	5	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$
DL Model 20	$P_{(t-3)}, P_{(t-2)}, P_{(t-1)}, P_{(t)}, Q_{(t-2)}, Q_{(t-1)}, Q_{(t)}$	7	$Q_{(t+1)}$	$Q_{(t+2)}$	$Q_{(t+3)}$

*DL: Daily-lumped input data, time 't' is in days.

Appendix II

Table A2. *Model types and input combinations (daily distributed data).*

Model type	Model inputs	No. of input variables	Output variable
Cause-effect models			
DD Model 3	$P_{1(t-2)}, P_{1(t-1)}, P_{1(t)}, P_{2(t-2)}, P_{2(t-1)}, P_{2(t)}, P_{3(t-2)}, P_{3(t-1)}, P_{3(t)}$, $P_{4(t-2)}, P_{4(t-1)}, P_{4(t)}, P_{5(t-2)}, P_{5(t-1)}, P_{5(t)}, P_{6(t-2)}, P_{6(t-1)}, P_{6(t)}$, $P_{7(t-2)}, P_{7(t-1)}, P_{7(t)}, P_{8(t-2)}, P_{8(t-1)}, P_{8(t)}, P_{9(t-2)}, P_{9(t-1)}, P_{9(t)}$	27	$Q_{(t+1)}$ $Q_{(t+2)}$ $Q_{(t+3)}$
DD Model 7	$P_{1(t-6)}, P_{1(t-5)}, P_{1(t-4)}, P_{1(t-3)}, P_{1(t-2)}, P_{1(t-1)}, P_{1(t)}$, $P_{2(t-6)}, P_{2(t-5)}, P_{2(t-4)}, P_{2(t-3)}, P_{2(t-2)}, P_{2(t-1)}, P_{2(t)}$, $P_{3(t-6)}, P_{3(t-5)}, P_{3(t-4)}, P_{3(t-3)}, P_{3(t-2)}, P_{3(t-1)}, P_{3(t)}$, $P_{4(t-6)}, P_{4(t-5)}, P_{4(t-4)}, P_{4(t-3)}, P_{4(t-2)}, P_{4(t-1)}, P_{4(t)}$, $P_{5(t-6)}, P_{5(t-5)}, P_{5(t-4)}, P_{5(t-3)}, P_{5(t-2)}, P_{5(t-1)}, P_{5(t)}$, $P_{6(t-6)}, P_{6(t-5)}, P_{6(t-4)}, P_{6(t-3)}, P_{6(t-2)}, P_{6(t-1)}, P_{6(t)}$, $P_{7(t-6)}, P_{7(t-5)}, P_{7(t-4)}, P_{7(t-3)}, P_{7(t-2)}, P_{7(t-1)}, P_{7(t)}$, $P_{8(t-6)}, P_{8(t-5)}, P_{8(t-4)}, P_{8(t-3)}, P_{8(t-2)}, P_{8(t-1)}, P_{8(t)}$, $P_{9(t-6)}, P_{9(t-5)}, P_{9(t-4)}, P_{9(t-3)}, P_{9(t-2)}, P_{9(t-1)}, P_{9(t)}$	63	$Q_{(t+1)}$ $Q_{(t+2)}$ $Q_{(t+3)}$
Combined models			
DD Model 17	$P_{1(t-2)}, P_{1(t-1)}, P_{1(t)}, P_{2(t-2)}, P_{2(t-1)}, P_{2(t)}, P_{3(t-2)}, P_{3(t-1)}, P_{3(t)}$, $P_{4(t-2)}, P_{4(t-1)}, P_{4(t)}, P_{5(t-2)}, P_{5(t-1)}, P_{5(t)}, P_{6(t-2)}, P_{6(t-1)}, P_{6(t)}$, $P_{7(t-2)}, P_{7(t-1)}, P_{7(t)}, P_{8(t-2)}, P_{8(t-1)}, P_{8(t)}, P_{9(t-2)}, P_{9(t-1)}$, $P_{9(t)}, Q_{(t-1)}, Q_{(t)}$	29	$Q_{(t+1)}$ $Q_{(t+2)}$ $Q_{(t+3)}$

*DD: Daily-distributed data input, time 't' is in days.

Appendix III

Table A3. *Model performance criteria.*

Pearson's correlation coefficient (R)	$\dfrac{\sum_{t=1}^{N}\left[Q_{\text{obs}}(t)-\overline{Q}_{\text{obs}}\right]\left[Q_{\text{est}}(t)-\overline{Q}_{\text{est}}\right]}{\sqrt{\sum_{t=1}^{N}\left[Q_{\text{obs}}(t)-\overline{Q}_{\text{obs}}\right]^2\left[Q_{\text{est}}(t)-\overline{Q}_{\text{est}}\right]^2}}$
Nash–Sutcliffe efficiency (E)	$E=\dfrac{E_1-E_2}{E_1}$ $E_1=\sum_{t=1}^{N}\left[Q_{\text{obs}}(t)-\overline{Q}_{\text{obs}}\right]^2,\ E_2=\sum_{t=1}^{N}\left[Q_{\text{est}}(t)-Q_{\text{obs}}(t)\right]^2$
Root mean square error (RMSE)	$\text{RMSE}=\sqrt{\dfrac{\sum_{i=1}^{n}\left(Q_{\text{obs}}(t)-Q_{\text{est}}(t)\right)^2}{n}}$
Akaike information criterion (AIC)	$\text{AIC}=m\ln(\text{RMSE})+2n$
Bayesian information criterion (BIC)	$\text{BIC}=m\ln(\text{RMSE})+n\ln(m)$

where $Q_{\text{obs}}(t)$ is the observed runoff at time t and $Q_{\text{est}}(t)$ is the estimated runoff at time t. N is the total number of runoff data points estimated from the model. $\overline{Q}_{\text{obs}}$ the mean observed runoff and $\overline{Q}_{\text{est}}$ the mean estimated runoff. m is the number of input-output patterns, and n is the number of parameters to be estimated.

References

Ahmed J and Sarma A K 2007 Artificial neural network model for synthetic streamflow generation; *Water Resour. Mgmt.* **21(6)** 1015–1029.

Akaike H 1974 A new look at the statistical model identification; *IEEE transaction on Automatic Control* **AC-19(6)** 716–723.

Aqil M, Kita I, Yano A and Nishiyama S 2007 A comparative study of artificial neural networks and neuro-fuzzy in continuous modeling of the daily and hourly behaviour of runoff; *J. Hydrol.* **337(1–2)** 22–34.

Aytek A and Alp M 2008 An application of artificial intelligence for rainfall-runoff modelling; *J. Earth Syst. Sci.* **117(2)** 145–155.

Beven K J 2000 *Rainfall Runoff Modeling: The Primer*; John Wiley & Sons Ltd., Chichester, U.K.

Box G E P and Jenkins G M 1976 *Time series analysis forecasting and control*; 2nd edn, Holden-Day, San Francisco.

Burlando P, Rosso R, Cadavid L G and Salas J D 1993 Forecasting of short-term rainfall using ARMA models; *J. Hydrol.* **144(1–4)** 193–211.

Campolo M, Andreussi P and Soldati A 1999 River flood forecasting with neural network model; *Water Resour. Res.* **35(4)** 1191–1197.

CDO 1992 *Final Report on Revised Flood Study for Koyna Dam*, Government of Maharashtra, Irrigation Department.

Chau K W, Wu C L and Li Y S 2005 Comparison of several flood forecasting models in Yangtze River; *J. Hydrol. Eng. ASCE* **10(6)** 485–491.

Chiew F H S, Stewardson M J and McMahon T A 1993 Comparison of six rainfall-runoff modeling approaches; *J. Hydrol.* **147(1–4)** 1–36.

Diskin M H 1970 Definition and uses of the linear regression model; *Water Resour. Res.* **6(6)** 1668–1673.

Driver N E and Troutman B M 1989 Regression models for estimating urban storm-runoff quantity and quality in the US; *J. Hydrol.* **109(3–4)** 221–236.

Hirsch R M 1979 An evaluation on record reconstruction techniques; *Water Resour. Res.* **15(6)** 1781–1790.

Hirsch R M and Gilroy E J 1984 Methods of fitting a straight line data: Examples in water resources; *Water Resour. Bullet.* **20(5)** 705–711.

Hsu K L, Gupta H V and Sorooshian S 1995 Artificial neural network modeling of the rainfall-runoff process; *Water Resour. Res.* **31(10)** 2517–2530.

Jagdeesh A, Zhang B and Govindraju R S 2000 Comparison of ANNs and empirical approaches for predicting watershed runoff; *J. Water Resour. Plann. Mgmt. ASCE* **126(3)** 156–166.

Jain A and Prasad S K V 2003 Comparative analysis of event based rainfall modeling techniques – Deterministic, statistical and Artificial Neural Network; *J. Hydrol. Eng. ASCE* **8(2)** 93–98.

Jothiprakash V and Magar R 2009 Soft computing tools in rainfall-runoff modelling; *ISH J. Hydraul. Eng.* **15(SP-1)** 84–96.

Jothiprakash V, Magar R and Sunil K 2007 Rainfall-runoff modeling using Linear Regression – A case study of an intermittent river; *J. Indian Assoc. Env. Mgmt.* **34(3)** 125–131.

Kisi O 2007 Stream flow forecasting using different artificial neural network algorithms; *J. Hydrol. Eng. ASCE* **12(5)** 532–539.

Loague K M and Freeze R A 1985 A comparison of rainfall-runoff modelling techniques of small upland catchments; *Water Resour. Res.* **21(2)** 229–248.

Luk K C, Ball J E and Sharma A 2000 A study of optimal model lag and spatial inputs to artificial neural network for rainfall forecasting; *J. Hydrol.* **227(1–4)** 56–65.

Maier H R and Dandy G C 1997 Determining inputs for neural network models of multivariate time series; *Microcomput. Civil Eng.* **12(5)** 353–368.

Mays L W and Tung Y 1992 *Hydro-systems Engineering and Management*, McGraw-Hill Inc, USA.

Momani M and Naill P E 2009 Time series model for rainfall data in Jordan: Case study for using time series analysis; *American J. Env. Sci.* **5(5)** 599–604.

Nash J E and Sutcliffe J V 1970 River flow forecasting through conceptual models, part I – a discussion of principle; *J. Hydrol.* **10(3)** 282–290.

Raman H, Mohan S and Padalianathan 1995 Models for extending stream flow: A case study; *Hydrol. Sci. J.* **40** 381–393.

Rissanen J 1978 Modeling of short data description; *Automation* **14** 465–471.

Salas J D, Delleur V, Yevjevich V and Lane W L 1980 Applied Modeling of Hydrologic Time Series; Water Resour. Publication, CA.

Singh V P 1988 *Hydrologic Systems, Vol 1: Rainfall-runoff modeling*; Prentice Hall, NJ.

Sivakumar B, Jayawardena A W and Fernando T M G H 2002 River flow forecasting: Use of phase-space reconstruction and artificial neural networks approaches; *J. Hydrol.* **265(1–4)** 225–245.

Sorooshian S, Duan Q and Gupta V K 1993 Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting model; *Water Resour. Res.* **29(4)** 1185–1194.

Srinivasulu S and Jain A 2006 A comparative analysis of training methods for artificial neural network rainfall runoff models; *Applied Soft Compt.* **6(3)** 295–306.

Sudheer K P, Gosain A K and Ramasastri K S 2002 A data driven algorithm for constructing artificial neural network rainfall-runoff models; *Hydrol. Process.* **16(6)** 1325–1330.

Sveinsson O G B, Lall U, Fortin V, Perrault L, Gaudet J, Zebiak S and Kushnir Y 2008 Forecasting spring reservoir inflows in Churchill falls basin in Quebec, Canada; *J. Hydrol. Eng. ASCE* **13(6)** 426–437.

Thirumalaiah K and Deo M C 2000 Hydrological forecasting using neural networks; *J. Hydrol Eng. ASCE* **5(2)** 180–189.

Toth E, Brath and Montanari A 2000 Comparison of short-term rainfall prediction models for real-time flood forecasting; *J. Hydrol.* **239(1–4)** 132–147.

Yevjevich V M 1963 Fluctuations of wet and dry years: Part 1 – research data assembly and mathematical models; Coarado State University, Hydrology Paper No. 1.

MS received 25 March 2011; revised 29 July 2011; accepted 9 August 2011