

Software Application for Data Driven Prediction Models for Intermittent Streamflow for Narmada River Basin

Ila Dashora
Research Scholar
AHEC, IIT Roorkee
Uttarakhand-247667, India

S. K. Singal
Principle Scientific Officer;
AHEC, IIT Roorkee
Uttarakhand-247667, India

D. K. Srivastav
Retd. Professor, Dept of
Hydrology, IIT Roorkee
Uttarakhand-247667, India

ABSTRACT

Synthetic generation of streamflow data facilitates the planning and operation of water resource projects. Significance of streamflow forecasting for intermittent river increases many fold in order to use available water yearlong for multipurpose water resources project. In the present study, monthly streamflow data has been used for intermittent river Gai in Narmada river basin. The performance of stochastic stream flow generation models— seasonal autoregressive integrated moving average (SARIMA) and Thomas-Fiering model are compared with Artificial Neural Network (ANN) approach. The performance of these models is evaluated on the basis of root mean square error (RMSE) and coefficient of determination (R^2). The study reveals that SARIMA performs better than Thomas-Fiering and ANN models. Thomas Fiering model is least reliable model among other two models. However Thomas-Fiering model performed well in case of high flow prediction whereas SARIMA and ANN performed well for lower and moderate flow. The predicted data can be used for the small hydropower projects development.

Keywords

Seasonal Autoregressive Integrated Moving Average, Neural Network, Stochastic Models, Prediction, Small Hydropower, Power Potential

1. INTRODUCTION

Streamflow prediction plays vital role for water resource project planning, operation and management. Short term forecasting for hourly or daily period forecasting is important in flood mitigation whereas long term forecasting like monthly, seasonal or annual forecasting helps in operation of multipurpose water resource projects like irrigation water management, municipal water supply, small hydropower and drought mitigation. Since the streamflow data has all the basic information related to flow regime, it helps for designing of water resource structures. These water resource structures are designed considering severity and undulation of streamflow and profound lack of long and continuous data availability lead the apprehension for prediction model development. As the streamflow is purely random process and has significant variability in time and space, the prediction model has to be accurate and less uncertain.

The prediction models can be categorized two class (a) knowledge driven models and (b) data-driven models. Knowledge driven models are successfully implemented for known physical catchment characteristics like area, shape slope, stream-length, altitude etc. Rainfall-runoff modeling and empirical relations are paradigm of such models. On contrary data driven or black box model, do not consider the internal mechanism of system between input and output of data. These models execute well even with the limited availability of physiographic catchment information. Artificial

intelligence techniques, regression models and stochastic models are the type of black-box or data-driven models.

Stochastic processes can be either linear or nonlinear processes. For linear stochastic processes like ARIMA (auto regressive integrated moving average), correlation coefficient is considered as reliability criteria (Haan, 1977). Researchers preferred ARIMA model due to its systematic procedure of estimation (i) identification, (ii) estimation and (iii) diagnostic checking which was described by Box and Jenkins (1976). Exclusively ARIMA and its seasonal variation were applied for streamflow prediction of univariate time series recently by Abrahart and See (2000) and Yürekli et al (2005). Velicer and Harrop (1983) worked on the adequate number of observations required to build an accurate model, further Wei (1990) found that ARIMA model efficiently predicted for more than 50 observations. Efficient execution and prediction by ARIMA model requires plenty of research experience since it is a complex modeling technique. This is the other drawback associated with this technique.

Regression based technique i.e. Thomas-Fiering model is monthly prediction model, developed by Thomas and Fiering (1962). In recent time, Kurunç, et al. (2005) found the Thomas Fiering models is successful tool for water quality prediction for Yeşilırmak River. Martins et al. (2011) worked on streamflow prediction using Thomas-Fiering models for long term prediction. Arselan and Cheleng (2012) used Thomas Fiering model for streamflow prediction after removal of persistence and found that this model works effectively specially for drought time.

Artificial intelligent technique, ANN has the ability to estimate the desired accuracy that makes this technique extensively useful for streamflow prediction. ANN has been successfully employed for many hydrological applications like rainfall-runoff modeling, water surface level prediction, ground water level prediction and streamflow prediction. Past research validates the transcendence of ANN over conventional stochastic and regression models for univariate streamflow time series. Huang et al. (2004) concluded that ANN provides better forecasting performance than ARIMA model and ANN is a conclusive approach and it does not explain randomness of streamflow processes. Ahmed and Sarma (2007) worked on the performance of ANN over traditional stochastic models- ARMA and Thomas Fiering models and found that ANN is superior to stochastic models.

The objective of this study is to predict long term (monthly) streamflow for intermittent river using SARIMA, Thomas-Fiering model and ANN models. Evaluation of suitable and accurate prediction model is the preliminary step towards the development of small hydropower and multipurpose water resource projects. Prediction is effectual tool for the development of water resource projects as these projects requires minimum 30 years of the discharge data.

2. STUDY AREA

The Narmada River originates in the Amarkantak plateau of Maikala range in the Madhya Pradesh (MP) state of India at an elevation of 1051 m above MSL and falls into the Gulf of Cambay in the Arabian Sea, after travelling 1312 km. The river basin is located between 72° 32'E to 81° 45'E longitudes and between 21° 20'N to 23° 45'N latitudes (NIH 2014). This river covers drainage area of 95726 sq. km. During monsoon season, peak discharges varies from 10,000 m³/s to 60,000 m³/s (Kale et al., 1994). It is the largest west flowing river of the Indian peninsula. Goi River is an intermittent river of lower Narmada River basin, covering 787 km² catchment area. The gauging site Dhulsar on Goi River is located at the latitude of 22°12'00" and longitude of 74°52'00". Discharge data of 8 years from June 2000 to May 2008 are available for this site considered in the present study. In this study, 8 years long monthly data has been used for synthetic data generation. The Narmada river basin and the study area are shown in the Figure 1.

3. STREAMFLOW PREDICTION

This study is analytical outcome of monthly streamflow prediction for intermittent river in Goi River in Narmada River basin using ARIMA, Thomas Fiering and ANN models. The river flow is recorded twice or thrice on daily basis, but mean monthly discharge is taken into account in the study.

3.1. Data Preparation

The requisite criterion for synthetic generation model is normality condition. Standardization and Box-Cox

transformations are two methods for converting any time series to normal time series. In this study, the Box-Cox transformation method has been preferred for normalization. The Box-Cox transformation given in Eq. (1) is applied for monthly flow period to make the data applicable for ANN model, as this method fails for zero discharge. After ANN model fitting and prediction the data is transformed to original form.

$$W = \begin{cases} (q^{\lambda} - 1) / \lambda & \lambda \neq 0 \\ \ln(q) & \lambda = 0 \end{cases} \quad \text{Eq. (1)}$$

Where, W is power transformed series and λ is transformation parameter, calculated by optimization for satisfying the normality condition.

3.2 Flow Duration Curve for Hydropower Potential Assessment

Figure 2 shows the cumulative probability distribution functions or flow duration curve (FDC) for Dhulsar gauging site. According to Vogel & Fennessey (1994) FDC is a cumulative frequency plot that shows the percentage of time that discharge in the stream is equaled or exceeded during a specific time. In FDC, the discharge corresponding to the particular percentage exceedence of time is known as dependable discharge. The power generation by hydropower plant can be estimated using following Eq. (2)

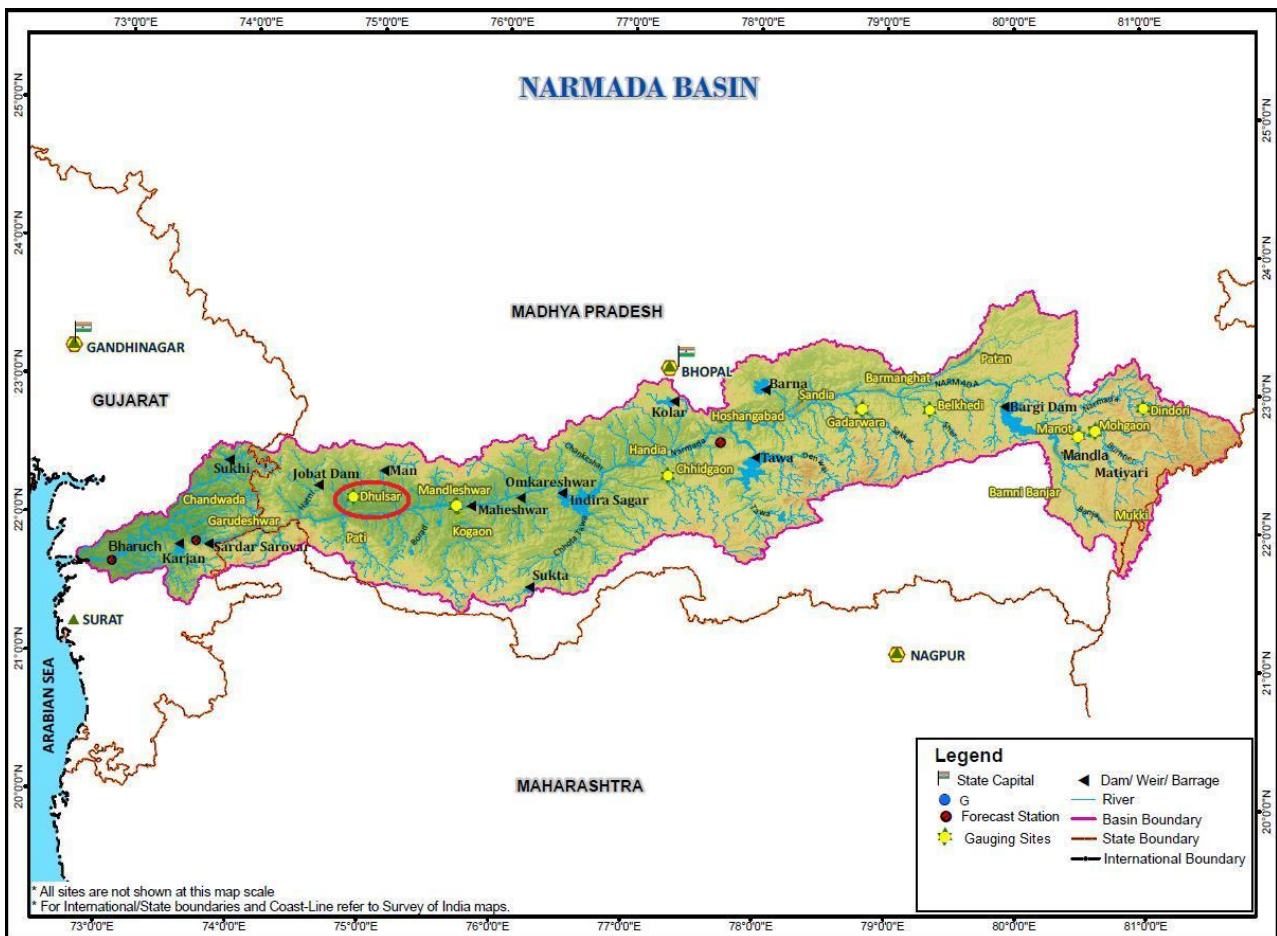


Figure 1: Narmada River basin showing major gauging sites, dams and barrage, India WIRS, (2014).

$$P = \frac{9.81 \times \eta \times Q \times H}{1000} \quad \text{Eq. (2)}$$

Where P = power (KW); Q = discharge (cumecs); H = net available head; and η = overall turbine and generator efficiency. For estimation of annual energy generation, the flow duration is divided into firm energy, secondary energy and dump energy as shown in Figure 2. Firm energy (E^F) generation takes place on the basis of firm discharge which is 75% dependable flow, whereas secondary energy (E^S) is generated at 50% dependable flow. Peak discharge (0-25% dependable flow) is not available throughout the year, so it comes under dump energy.

$$\text{Total energy, } E^T = E^F + E^S \quad \text{Eq. (3)}$$

Firm energy,

$$E^F = 9.81 \times H \times Q_{100} \times \eta \times (1 \times 365 \times 24) \quad \text{Eq. (4)}$$

As the study area is for intermittent river, the availability of the discharge for harnessing hydropower will be available in the monsoon months only. Because of this reason overall performance of the plant can reduce up to 50%. The other factors that may reduce the generation capacity are time required for operation and maintenance, silt problem at site etc.

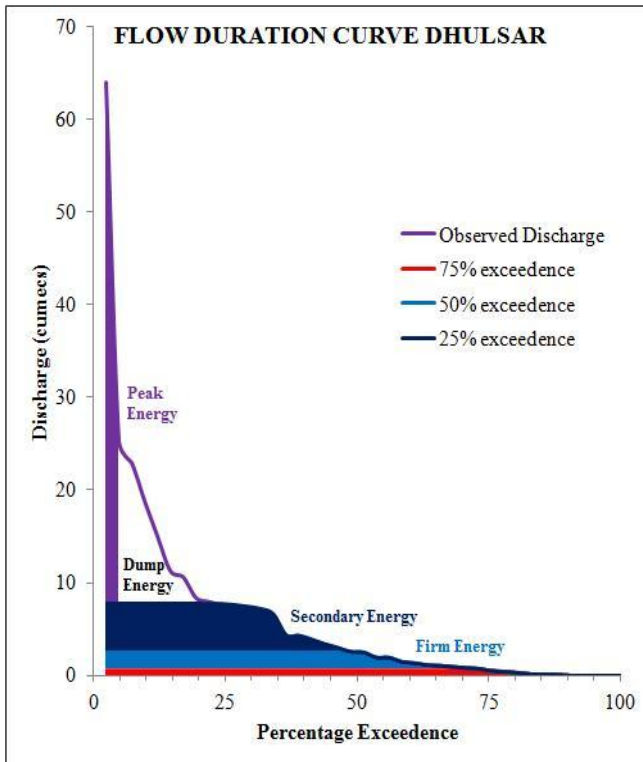


Figure 2: Cumulative Probability Distribution Function for Dhulsar Gauging Site (Flow Duration Curve)

4. MONTHLY TIME SERIES ANALYSIS

The present study is carried out to estimate the performance of ANN over the conventional forecasting models like SARIMA and Thomas-Fiering models. The efficiency of these models are compared using root mean square error (RMSE) and regression coefficient (R^2), as explained in later sections.

4.1 Thomas-Fiering Model

Thomas Fiering model (Thomas and Fiering, 1962) is monthly synthetic data generation model and useful in operation related studies of multipurpose water resource project. The transformed historical streamflow in month, m and year, y is given by $q_{m,y}$. The repetitive relation for synthetic monthly data $q_{m,y}$ generation is as follows;

$$q_{m+1,y} = \bar{q}_{m+1} + r_m \left(\frac{s_{m+1}}{s_m} \right) (q_{m,y} - \bar{q}_m) + s_{m+1} (1 - r_m^2)^{1/2} \zeta_{m,y} \quad \text{Eq. (5)}$$

where \bar{q}_m is average of observed historical monthly streamflow series for month m ; s_m^2 is variance of observed monthly streamflow for month m ; $\zeta_{m,y}$ are independent standard normal random variables and r_m is correlations between month m and $m+1$ of the observed streamflow. In the above equation, $\bar{q}_{1,y+1}$ is understood to be $q_{1,y+1}$ when $m = 12$. The synthetic generated monthly flow is transformed to original form using the Eq. (6).

$$Q_{m,y} = \exp(q_{m,y}) \quad \text{Eq. (6)}$$

Thomas Fiering model is based on 12 regression equations having statistical properties of the discharge time series, like standard deviation, average and correlation coefficient.

4.2 Seasonal-Autoregressive Integrated Moving Average (SARIMA) Model

Autoregressive Integrated Moving Average (ARIMA) model is a stochastic model that has random elements of past observations and random errors. The streamflow time series has four components namely trends (T), periodicity (P), dependency (D) and independency (V_t). Trends in the streamflow time series can be determined using Mann-Kendall test. Fourier transformation is applied for removal of periodic components. The remaining two components, dependent and independent components are collectively known as stochastic components. The stochastic model, ARIMA, is applicable for streamflow prediction. The original process that generates the streamflow using ARIMA is given in Eq. (7),

$$S_{t+1} = \sum_{i=1}^p \phi_i S_{t+1-i} + V_t - \sum_{j=1}^q \theta_j V_{t-j} \quad \text{Eq. (7)}$$

ARIMA (p, d, q) is a non seasonal synthetic data generation model. Here p is the order of auto-regression, d is the amount of differencing, and q is the order of the moving average. The ARIMA process turns simply into autoregressive (AR) model if the moving average (MA) component is unavailable and vice-versa. The time series has to be stationary for ARIMA modeling. Differencing or standardization turns non-stationarity into stationarity. For stationary time series autocorrelation function (ACF) and partial autocorrelation function (PACF) should lie in between upper and lower limit of confidence. ARMA model transforms into ARIMA models when the time series is non-stationary and differencing is non zero ($d \neq 0$). Linear differencing operator (Δ) is illustrated in Eq. (8)

$$\Delta S_t = S_t - S_{t-1} = S_t - B S_t = (1 - B) S_t \quad \text{Eq. (8)}$$

Stationary time series (W_t) obtained after d^{th} difference (Δ^d) for S_t , is mentioned in Eq. (9)

$$W_t = \Delta^d S_t = (1 - B)^d S_t \quad \text{Eq. (9)}$$

Eq. (10) is the general form of the general form of ARIMA (p, d, q)

$$\phi_p(B)(1 - B)^d S_t = \mu + \theta_q(B)\varepsilon_t \quad \text{Eq. (10)}$$

If the ARIMA (p,d,q) does not fits well then for monthly time series fitting and forecasting leads toward the SARIMA (p,d,q)(P,D,Q)_m process. Here m=12 for monthly streamflow time series. SARIMA solution for S_t is shown in the following Eq. (11),

$$\phi_p(B)\Phi_m(B^m)\nabla_m^D \nabla^d S_t = \theta_q(B)\Theta_m(B^m)\varepsilon_t \quad \text{Eq. (11)}$$

The Box–Jenkins (1976) presented a methodology that includes three iterative steps (i) Model identification, (ii) Parameter estimation and (iii) Diagnostic checking of fitted parameters as shown in Figure. 3.

Pankratz (1983) prepared a guideline for model identification on the basis of shape and characteristics of autocorrelation function (ACF) and partial autocorrelation function (PACF) as shown in Table 1.

Table 1: Guidelines to identify model parameters on the basis of ACF and PACF (Machiwel and Jha, 2012)

S. No	Model parameter	Characteristics of ACF	Characteristics of PACF
1	One autoregressive (p)	Exponential decay	Spike at lag 1, no correlations for other lags
2	Two autoregressive (p)	A sine-wave shape pattern or a set of exponential decays	Spikes at lags 1 and 2, no correlation for other lags
3	One moving average (q)	Spike at lag 1, no correlation for other lags	Damps out exponentially
4	Two moving average (q)	Spikes at lags 1 and 2 no correlation for other	A sine-wave shape pattern or a set of exponential decays lags
5	One autoregressive (p) and one moving average (q)	Exponential decay starting at lag 1	Exponential decay starting at lag 1

Das (2000) mentioned that autocorrelation function r_k and partial autocorrelation function $\rho_{k,k}$ are the keys to identify p and q value for a reasonable ARIMA model. The ACF and PACF equations are mentioned in Eq. (12) and Eq. (13)

$$r_k = \frac{\sum_{t=1}^{N-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^N (Y_t - \bar{Y})^2} \quad \forall_t \quad \text{Eq. (12)}$$

Where, r_k is an autocorrelation function at k lag; \bar{Y} is the mean value of time series Y_t and N is data length.

$$\rho_{k,k} = \frac{r_k - \sum_{j=1}^{k-1} \rho_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} \rho_{k-1,j} r_j} \quad \forall_t \quad \text{Eq. (13)}$$

Where, $\rho_{k,k}$ is partial autocorrelation function at lag k; r_k is autocorrelation function at lag k; and $\rho_{k,j} = \rho_{k-1,j} - \rho_{k,k} \rho_{k-1,k-j}$; where j = 1, 2, ..., k-1.

For tentative model selection, ACF and PACF are to be plotted against lag k. ACF helps to identify the value of q whereas PACF is useful for identification of order of p. Among various combinations of p and q, the best ARIMA model can be selected on the basis of minimal Akaike's information criteria (AIC) (Akaike, 1978) as given in Eq. (14)

$$AIC = 2 \ln(ML^*) + 2 \frac{(p + q + 1)}{n + 1} \quad \text{Eq. (14)}$$

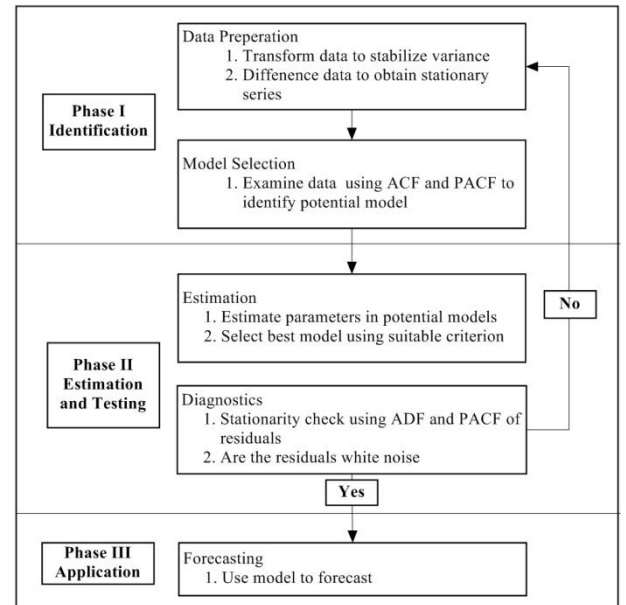


Figure 3: Schematic representation of the ARIMA methodology for time series modelling

Schlittgen and Streitberg (2001) mentioned that AIC always estimates right order (p and q) of models whereas BIC is suitable for larger sample size. The other criterion for model selection is residual analysis using ACF and PACF. The residual for the best fitted model should lie in between the upper and lower limit confidence bound. The residuals should be independent, normally distributed and random. The other minimization function is Bayesian Information Criterion developed by Brockwell and Davis (1991) and can be expressed as given in Eq. (15)

$$BIC = 2 \ln(ML^*) + \frac{(p+q) \ln(n+1)}{2(n+1)} \quad \text{Eq. (15)}$$

4.3 ANN Model

ANN works as biological nervous system to pass the information. Neurons are the units of neural network system that are properly arranged in layers. The common ANN structure has three layers (i) input layer, (ii) hidden layer and (iii) output layer. Each layer has different number of nodes. Input layer has single node to receive input data and it does not modifies or processes data.

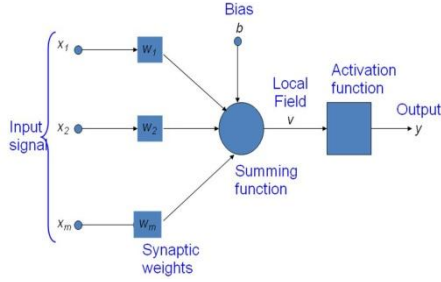


Figure 4: Neural Network architecture and working

It duplicates the signal according to multiple outputs to feed hidden layers. The second step in neural network computing is processing in hidden layer. This layer is the workstation of neural network structure. Here input signals are multiplied by weights that are set of encoded numbers stored in hidden layer as shown in Figure 4. Hidden layer has ability to manipulate data for proper selection of weights. The values of weights are added and then this single value passes through transfer function. The output layer has several numbers of nodes and that also uses transfer function. In this study, tan-sigmoid transfer function is used. This transfer function significantly transforms all the input values that lie between $-\infty$ to $+\infty$ into -1 to 1 . The mathematical and graphical representation of tan-sigmoid function is shown in Eq. (16)

$$f(s_i) = \tan \text{sig}(s_i) = \frac{2}{1 + \exp(-2s_i)} - 1 \quad \text{Eq. (16)}$$

where $S_i = \sum_{i=1}^n w_i x_i$ is the input signal referred as the weighted sum of incoming information. Burian et al. (2001) found that as the number of nodes in hidden and output layers decreases, prediction accuracy increases. The input data set is divided into three portions for training, validation and testing. The objective to train input data is to reduce global error as mentioned in Eq. (17)

$$\varepsilon = \frac{1}{n} \sum_{n=1}^n \varepsilon_n \quad \text{and} \quad \varepsilon_n = \frac{1}{2} \sum_{i=1}^i (O_i - T_i)^2 \quad \text{Eq. (17)}$$

Where, n is number of training iteration, ε_n is error at the end of training n , i is number of output nodes, O_i is neural network output and T_i is training output. Adjustment of weight and biases reduces the error while training. Karul et al. (2000) found that Levenberg-Marquardt algorithm is suitable for second order training speed problem as it requires a greater amount of memory than other algorithms. This algorithm solves the following Eq. (18)

$$(x_k - x_{k+1}) (J^T J + \mu I) = J^T e \quad \text{Eq. (18)}$$

Here $(x_k - x_{k+1})$ are the weight updating vectors to be computed, J is Jacobian matrix, μ is damping factor, e is error vector that contains output error for each input vector. The damping factor μ is adjusted to zero after each iteration. When $\mu=0$, the algorithm turns to Gauss-Newton algorithm where iteration starts giving insufficient reduction in residuals.

All the mention steps for SARIMA application are performed using SPSS and *gretl* software packages. The ANN application for prediction is done using MATLAB 2013(a) software package. Matlab 2013(a) is used for the model fitting and prediction for Thomas-Fiering model.

4.4 Performance Criteria

Performance of model can be evaluated by performance criteria because prediction accuracy of any model is highly dependent on model structure, iterations and computational techniques. Abrahart and See (2000) applied five different global evaluation measures for comparison of performance of streamflow prediction model. These criteria are mean absolute error (MAE), root mean square error (RMSE), mean higher order error function for peak flow prediction, model efficiency, and percentage of predictions grouped according to degree of error. Karunanithi et al. (1994) suggested that two measures should be used in order to get different information about predictive ability of models. Keeping this in view, root mean square error and coefficient of determination has been used as performance measures in this study.

1. Coefficient of determination

$$R^2 = \left[\frac{\sum_{i=1}^n (y_i - \bar{y})(F_i - \bar{F})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (F_i - \bar{F})^2}} \right]^2 \quad \text{Eq. (19)}$$

2. Root mean square error

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - F_i)^2} \quad \text{Eq. (20)}$$

Where, Y_i is the observed flow, F_i is the output response from the models, \bar{Y} is the mean of the observed flow and \bar{F} is mean of forecasted flow. Karunanithi et al., (1994) stated that RMSE is a good criterion for indicating the goodness of fit at high and moderate streamflow time series.

5. RESULTS OF STREAMFLOW PREDICTION MODELS

Firstly talking about the results of Thomas-Fiering model fitting and prediction. As reported earlier, Thomas-Fiering model is a regression based prediction model that includes correlation coefficient, standard deviation and average discharge of consecutive months. **Error! Reference source not found.** Figure 5 shows that this model not only works well for low and moderate flow, but performs pretty well for high flow also.

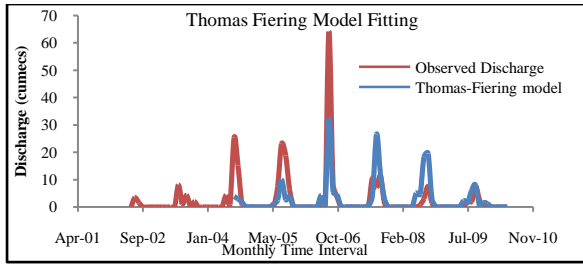


Figure 5: Thomas-Fiering model fitting

The statistics of the observed and estimated time series given in Table 2, shows that mean discharges of both time series are nearly similar but the standard deviation has notable difference. This result indicates fewer fluctuations in observed and estimated time series due to more similarity in moderate flow pattern than in high flow pattern.

Table 2: Statistics of observed and calculated flow using Thomas-Fiering model

Parameters	Observed	Modeled
Mean	2.688318	2.638065
Standard Deviation	7.887753	5.581939
R2	0.4055	
RMSE	6.5363	

The $R^2 = 0.4438$ is less for both modeled and observed time series indicating the moderate slope as shown in Figure 6. These statistical parameters prove that modeled time series has also done well.

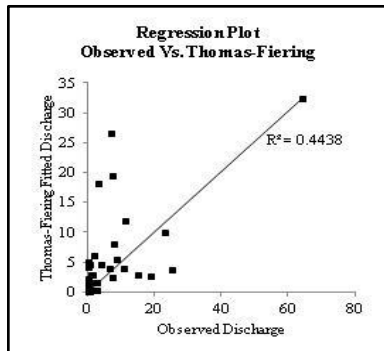


Figure 6: Regression plot between observed and fitted flow for Thomas Fiering model

Now moving the discussion towards second selected stochastic model i.e. SARIMA. The model fitting and prediction are divided into three sequential steps; identification, estimation and prediction. The results of mentioned three steps are discussed in subsequent parts respectively.

5.1 Identification (Model Identification ACF and PACF)

The correlogram test of observed time series is appropriate approach to verify stationarity. The correlogram is the plot of autocorrelation function (ACF) and partial autocorrelation function (PACF) against lags. If time series is non-stationary

then standardization or differencing are two approaches for stationary time series. For optimal order of differencing the standard deviation has to be minimal for the differenced time series. The order of differencing should be minimal to avoid over differencing. ACF and PACF plots. Figure 7 shows non-stationarity in time series as ACF is not bounded in between the confidence interval. The notable feature of this ACF plot of Figure 7 is the presence of periodicity, that advice to add seasonality in ARIMA model, thus the ARIMA(p,d,q) process turns to seasonal ARIMA(p,d,q)(P,D,Q).

5.2 Estimation and Testing (AIC and Residual analysis)

As mentioned in previous section, there are three criteria for selecting best fit SARIMA. The first criteria is minimal Akaike's Information Criteria (AIC) and Bayesian information criterion (BIC), second criteria is stationary residual of selected model and third one is minimum RMSE (performance criteria). The best fit model should meet all the three criteria simultaneously. Different combinations of p, d, q, P, D and Q are tested for minimum AIC, RMSE and stationary residual correlogram.

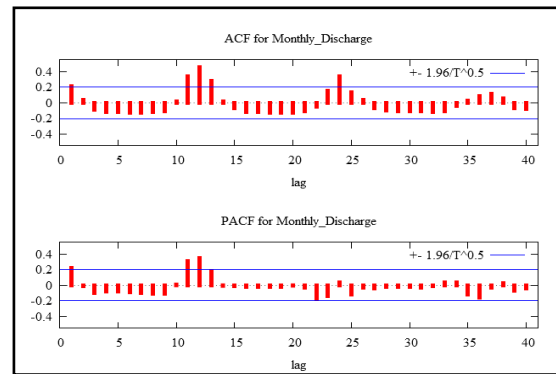


Figure. 7 ACF and PACF analysis of Dhulsar streamflow time series

Figure 8 shows the stationarity for the residuals of ARIMA (0,0,0)(1,0,0), that proves that selected model is best prediction model among other SARIMA models. The seasonal ARIMA model equation can be written in three different forms as: differenced equation or as infinite sum of current and weighted previous values of error or as an infinite sum of weighted previous observations plus the current value of error (Martins et al., 2011).

Suitable requirement for prediction helps to form the seasonal ARIMA equation. The differenced generalized equation for ARIMA (0,0,0)(1,0,0) is shown in Eq. (21)

$$(1 - \Phi_{1,12} B^{12}) s_t = \varepsilon_t \quad \text{Eq. (21)}$$

Where s_{t-j} = time series component $j=0,1,2,\dots$

ε_{t-j} = white noise in time series; $j=0,1,2,\dots$;

Φ = Seasonal AR component of time series

The Eq. 21 shows that only seasonal autoregressive component is present and other components are not present in the equation mentioned. The final values of estimated parameters of ARIMA (0,0,0)(1,0,0) are shown in Table 3. The Log-likelihood and AIC are minimum for this selected ARIMA (0,0,0)(1,0,0) model.

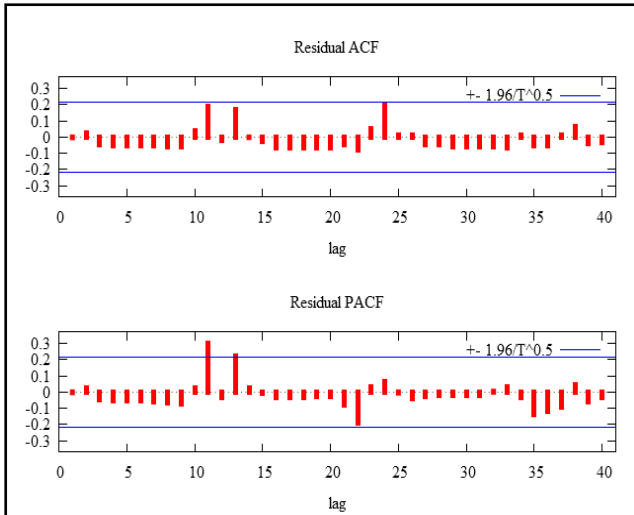


Figure 8: Residual ACF and PACF for model ARIMA (0,0,0)(1,0,0)

Table 3: Final model parameters estimated and model fitting statistics

Model Fitting Parameters				
Types	Coefficient	SE Coefficient	z	p-value
SAR1	0.4074	0.0906	4.492	0.00000705
Constant	2.5376	1.2547	2.022	0.0431
Model Fitting Statistics				
Log-likelihood		-288.674	AIC	583.3473
Hannan-Quinn		586.2788	BIC	590.6397

Further assurance is required for independent, homoscedastic and normally distributed residuals that were tested using Ljung-Box test (1978). The results of Ljung-Box test indicate that all correlations are not different from zero at 95% confidence interval as shown in Table 4. This test also shows that the selected ARIMA (0,0,0)(1,0,0) models is most suitable model for monthly stream flow prediction.

Table 4: Ljung-Box test statistic for residuals of selected model residuals

p value	DOF	χ^2 value	Critical value
0.3934	11	4.575	11.6131
0.2853	23	13.091	26.3362
0.7032	35	22.465	30.1081
0.9392	47	32.2676	33.0017

5.3 Prediction

Figure 9 is the time series plot, shows that ARIMA (0,0,0)(1,0,0) model follows the moderate flow properly, although there is weak prediction at high flow discharge. Even though the selected model performed well still there is lack of accuracy due to less availability of discharge data. The model prediction can only be said accurate when long

period of data is available for validation. Figure 10 shows the regression plot between observed and ARIMA fitted discharge. The $R^2 = 0.6231$ is satisfactory regression coefficient while RMSE = 6.1206 calculated lesser than Thomas-Fiering model.

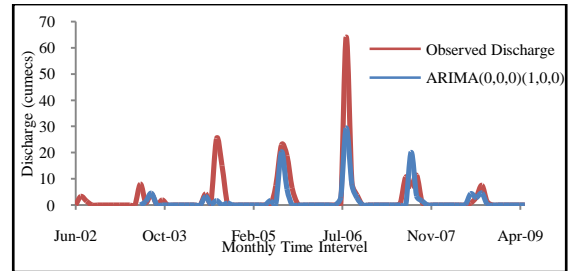


Figure 9: Streamflow fitting by ARIMA (0,0,0)(1,0,0)

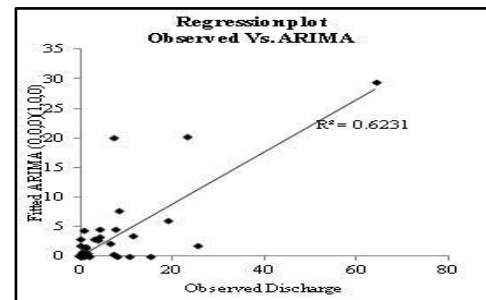


Figure 10: Regression plot between fitted and observed discharge for ARIMA model

Now approaching towards the third selected data driven model i.e. ANN approach. In case of ANN model fitting and prediction, data is divided into various ratios for testing, validation and training purpose, with the delayed and hidden neurons. Initially the discharge is transformed to normally distributed data using Box-Cox transformation and after the ANN model fitting the dataset is transformed to original time series. Among other combinations of training, validation and testing ratios, the overall performance of model 75:15:15 with 4 delays and 10 hidden neurons is found most suitable. The performance criteria RMSE is found as 8.084749 that is higher than the other two selected models but on the other hand $R^2 = 0.80993$ is also highest. Figure 11 shows the time series plot of observed and ANN fitted data, indicates that fitted flow is following the observed flow for low and moderate flow, while in case of high flows, the performance is not satisfactory.

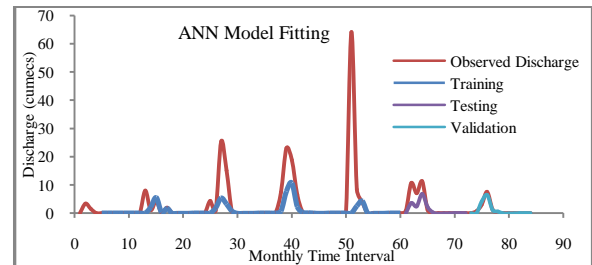


Figure 11: Time series plot for ANN (75:15:15) model with 4 delay and 10 hidden neurons

Figure 12 is regression plot between observed and fitted flow while data is divided for training, validation and testing. The regression coefficient for all the divided parts and overall data set is found very well. This proves that this model is best for prediction purpose

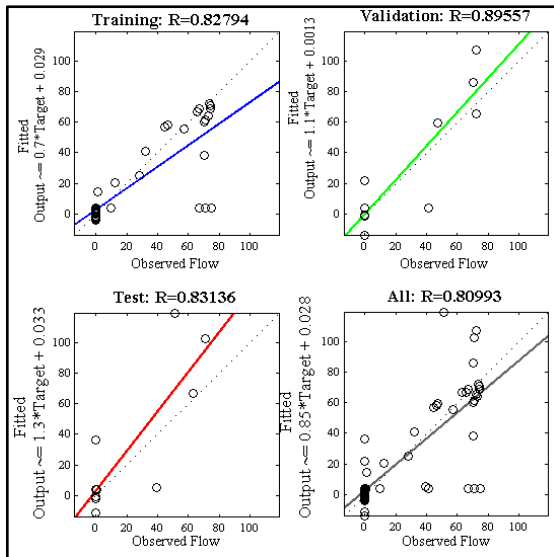


Figure 12: Regression plot of Train, validate and test streamflow discharge for ANN

Figure 13 shows the prediction for next year using all selected models. The performances of all these selected models are listed in Table 5. It can conclude from Table 5 that for model fitting and prediction ARIMA (0,0,0)(1,0,0) performed well. Figure 13 for one year prediction shows that Thomas Fiering model is performing fine even there is more fluctuations but ARIMA and ANN models are good for low and moderate flow.

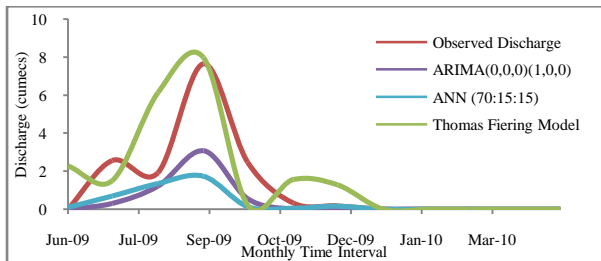


Figure 13: Time series plot of one year prediction using all selected forecasting models

Table 5: Performance statistics for selected models

	Criteria	Thomas Fiering Model	ARIMA (0,0,0) (1,0,0)	ANN (70:15:15)
Model Fitting	RMSE	6.5363	6.1206	8.084749
	R ²	0.4055	0.6231	0.80993
Prediction	RMSE	1.6322	1.5909	1.9236
	R ²	0.5816	0.9252	0.7317

6. CONCLUSIONS AND RECOMMENDATIONS

In the present study, comparative analysis of SARIMA, Thomas-Fiering and ANN model has been carried out. These models are found as most sophisticated extrapolation method for prediction. These models can predict any time series, with any pattern of change and do not require the forecaster to

choose value of any parameter. The limitation with these models is the requirement of a long time series. These models are also known as Black Box model. Even the prediction is too accurate but these models do not guarantee the prediction accuracy. In this study, Thomas Fiering, seasonal ARIMA and ANN models were tested for monthly stream flow prediction. The performances of these models are tested using different performance criteria. The following conclusions are drawn in this study:

1. Streamflow time series achieved stationarity after first order differencing that was confirmed by ACF and PACF analysis of difference time series.
2. Even though many ARIMA models fit to the time series but the appropriate model can only be selected by residual analysis and Ljung-Box test, which declares the residual as stationary and independent.
3. ARIMA(0,0,0)(1,0,0) is found to be the most suitable model among the other combinations of seasonal ARIMA.
4. Thomas-Fiering model, regression based model, cannot work for lean period, so only 7 regression equations were used for successful prediction of streamflow.
5. ANN is a nonlinear autoregression based technique, so Box-Cox Transformation is must for accurate prediction.
6. ANN model with the 70:15:15, training, testing and validation ration and 4 delays and 10 hidden neurons is found the best among other ratios of testing, validation and training.
7. ARIMA (0,0,0)(1,0,0) performed well among other models for the intermittent river prediction, as this model followed peak or high discharge more as compared with other models.
8. As the development of water resources project requires 30 years of discharge data, prediction is effectual method for data extension. The SARIMA is found appropriate prediction tool for monthly discharge data.

7. ACKNOWLEDGEMENTS

The authors greatly acknowledge the financial support from the Ministry of Human Resource Development, Govt. of India in the form of research scholarship to carry out this work. The authors also acknowledge the use of facilities as available in AHEC, IIT Roorkee.

8. REFERENCES

- [1] Abrahart, R.J., & See, L. 2000. Comparing neural network (NN) and auto regressive moving average (ARMA) techniques for the provision of continuous river flow forecasts in two contrasting catchments. Hydrol. Process. 14, 2157–2172.
- [2] Ahmed, J.A. & Sarma, A.K., 2007. Artificial neural network model for synthetic stream-flow generation. Water Resour Manage., 21,1015–1029.
- [3] Akaike, H., (1978. A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Stat. Math, 30A, 9–14.
- [4] Arselan, & Cheleng A., 2012. Stream flow Simulation and Synthetic Flow Calculation by Modified Thomas Fiering Model. Al-Rafadain Engineering Journal, 20(4), 118.

- [5] Box, G.E.P. & Jenkins, G.M., 1976. Time Series Analysis Forecasting and Control. Holden-Day, San Francisco, pp. 32-100.
- [6] Box, G.E.P. & Ljung, G.M. 1978. On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303.
- [7] Brockwell, P.J. & Davis, R.A., 1991. Time Series: Theory and Methods. 2nd Edition, Springer Series in Statistics, Springer, New York.
- [8] Burian, J.S., Durrans, S.R., Nix, S.J., & Pitt, R.E., 2001. Training artificial neural networks to perform rainfall disaggregation. *J. Hydrol. Engg.* 6(1), 43–51
- [9] Das, G., 2000. Hydrology and soil conservation engineering. Prentice Hall of India, New Delhi
- [10] Haan, C.T., 1977. Statistics Methods in Hydrology. Iowa State Press, Iowa, p. 378.
- [11] Huang, W.R., Xu, B., & Hilton, A. 2004. Forecasting Flows in Apalachicola River Using Neural Networks. *Hydrological Processes*. 18, 2545-2564.
- [12] India WIRS, 2014. Water Resources Information System of India website. <http://india-wris.nrsc.gov.in/wrpinfo/index.php?title=Narmada>
- [13] Kale, V.S., Ely, L.L., Enzel, Y. & Baker, V.R., 1994. Geomorphic and hydrologic aspects of monsoon floods on the Narmada and Tapi rivers in central India. *Geomorphology*, 10,157-168.
- [14] Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., & Germen, E., 2000. Case studies on the use of neural networks in eutrophication modeling. *Ecological Modelling*, 134, 145–152.
- [15] Karunanithi, N., Grenney, W.J., Whitley, D., & Bovee, K., 1994. Neural networks for river flow prediction. *J. Comput. Civ. Engg.*, 8(2), 201–220.
- [16] Machiwele, D., & Jha, M.K., 2012. Hydrologic Time Series Analysis, Capital Publishing Company. pp-91. DOI 10.1007/978-94-007-1861-6_5.
- [17] Martins, O.Y., Ahaneku, I.E., & Mohammed S.A., 2011. Parametric Linear Stochastic Modelling of Benue River Flow Process. *Open Journal of Marine Science*, 3, 73-81
- [18] NIH website 2014. http://www.nih.ernet.in/rbis/basin%20maps/narmada_about.htm.
- [19] Pankratz, A. 1983. Forecasting with Univariate Box-Jenkins Models: Concepts and Cases. Wiley, New York.
- [20] Schlittgen, J. and Streitberg, B.H.J. 2001. *Zeitreihenanalyse*. Oldenbourg, Munich, Germany.
- [21] Thomas, H.A. & Fiering, M.B., 1962. Mathematical synthesis of stream-flow sequences for the analysis of river basin by simulation. In A. Maas et al. Design of Water Resource Systems, chapter 12: Harvard University Press.
- [22] Velicer, W.F., & Harrop, J., 1983. The reliability and accuracy of time series model identification. *Eval. Rev.*, 7, 551–560.
- [23] Vogel, R. M., and N. M. Fennessey 1994, Flow-Duration Curves .2. New Interpretation and Confidence-Intervals, *J. Water Resour. Plan. Manage.-ASCE*, 120, 485-504.
- [24] Wei, W.W.S., 1990. Time Series Analysis. Addison-Wesley Publishing Company, Inc, New York, p. 478.
- [25] Yurekli, K., Kurunc, A., & Ozturk, F., 2005. Application of Linear Stochastic Models to Monthly Flow Data of Kelkit Stream, *Ecological Modelling*, 183, 67-75.