مجلة جامعة بابل / العلوم الصرفة والتطبيقية / العدد (1) / المجلد (17) : 2009

Stochastic Model for Monthly Flow to Bekhem Reservoir at the North of Iraq

Ruqaya K.M. Al-Masudi
College of Energy, Babylon Univ., Babil, Iraq

Abstract

In this study (ARMA) method of time series analysis is applied to monthly flow for Bekhem reservoir. This method involves decomposition of the historical series into trend, seasonality and residual components. Stochastic ARMA models are then fitted to the residuals. Forecasts of the future value of the flow rate for 5th years starting from the last observation in Sep. 1982 to the observation in Oct. 1988 are also made by using ARMA.

الخلاصة

لقد تم في هذه الدراسة استخدام واحدة من طرق تحليل السلاسل الزمنية لغرض دراسة المعدلات الشهرية للجريان إلى خزان بخمة. تتضمن هذه الطريقة تحليل السلسلة الزمنية إلى مركبة الاتجاه ومركبة الدورية والمركبة المتبقية.وقد استخدمت المركبة المتبقية لغرض بناء النموذج التنبئي.ثم تم إيجاد القيم المستقبلية للتصاريف لمدة خمس سنوات تبدأ من سنة 1983 ولغاية 1987 .

Introduction

Hydrologic time series is defined as continuous sequential observations which are usually expressed as an average value over equal intervals of time such as mean: daily, monthly, or annual flows. The process of averaging is called discretization, and the resulting series called discrete time series (**Chatfield, 1982**).The main objective of the present study is applying the well-known mixed autoregressive moving average ARMA model to flow rate for Bekhem dam in the north of Iraq. Then using this model to forecast 5th years of future time series compared with flow rate standards .**Al-Suhaili (1986)** used singlesite AR(1), autoregressive integrated moving average(ARIMA (1,0,1)) and (matalas model) for four Tigris river flow stations. These models were used for daily stream flow of Tigris River at four stations from the period (1936-1982).**Al-Husseini (2000)** used AR (1), MA (1) and ARMA (1, 1) as univariate models and first order multivariate model to fit stochastic component of eight years (1992-1998) of mean monthly water quality parameter at Al-Hilla station .**Al-Mousawi (2003)** applied two stochastic singlesite autoregressive models with first order AR(1), and multisite model with first order AR(1)to model of monthly water quality data of eight hydrochemical parameter with discharges of four stations on Hilla river for the period (1987-2001).**Abed (2007)** applied ARMA and ARIMA models to monthly records of some physical and chemical properties of river water in Babylon, Najaf, and Diwaniya governorates.

The Study Area

The Bekhme dam is located at about 7km upstream of Bekhme village and approximately 2km downstream from the confluence with the tributary Ruwanduz River which flows in from the left-bank side as shown in figure(1),(**Planning Report on Bekhme Dam Project,1986,quoted in Hussien,2006**).

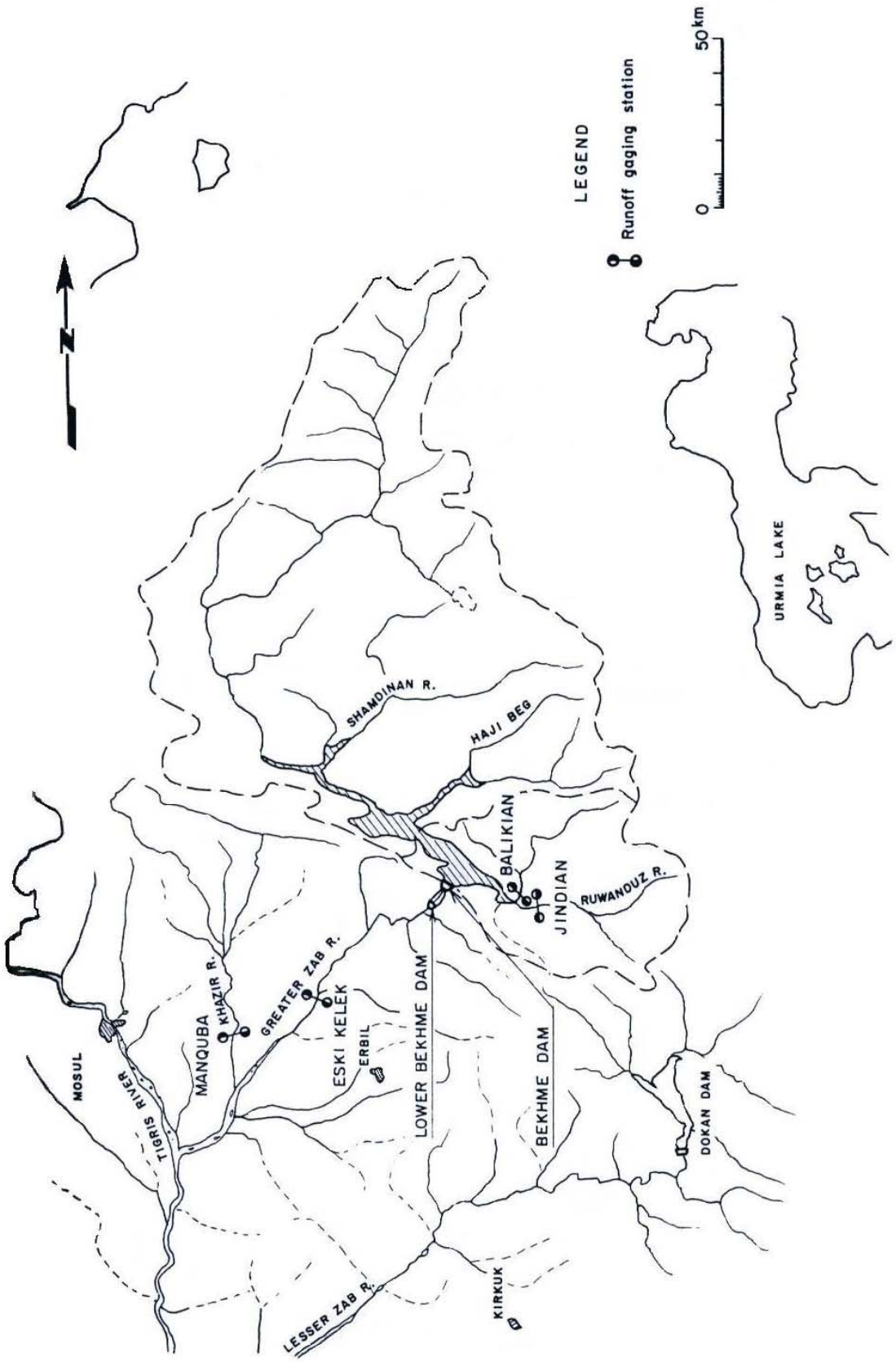

Fig.(1):The Study Area

Application of ARMA Model

Before building of ARMA model historical annual series must be tested for jump and trend components by statistical tests and these components must be removed by using an appropriate method. Then a suitable transformation is selected to convert the data to the normal distribution. The periodic component can be detected and removed by non-parametric method and the resulting series is called stochastic series which is used for building of stochastic models (**Chatfield, 1982**).

The procedure used for data analysis can be summarized by the following steps:

1. Test the homogeneity by some statistical testes, if the non-homogeneity exist, it will be removed by suitable method.
2. Use a suitable transformation like (Box-Cox, Square root, or Log transformation) to normalize the homogenous data.
3. To detect the periodicity plot the correlogram of monthly means and standard deviation of normalized data.
4. Removal of periodicity by applying non-parametric method for the normalized series.
5. Estimation of model parameter by plotting the corrlograms of the data and by the least squares algorithm for conditional model.
6. The diagnostic check of model by one of some tests and testing the independency of the residuals.
7. Forecasting and verifying the forecasted models.

Test and Removal of No-homogeneity

Hydraulic and water quality time series are commonly non-stationary with trends and seasonality (**Zou and Yu, 1996**). This section includes no-homogeneity test of the mean and standard deviation, and its removal when the series is found to be no-homogeneous. Split-Sample method is used to ascertain whether or not the differences between the means and standard deviations of two sub-samples are significantly different from zero at (97.5%) confidence limit, therefore, the data is divided into two sub-samples of years, the first is 25^{th} years long (1933-1957) and the second is five years long also (1958-1982). Figures (2&3) show the annual mean and standard deviation, respectively.

The test is applied as follows:

1. **t-test:** the test for homogeneity in mean of two sub-samples is given as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S \times \sqrt{\frac{n_1 + n_2}{n_1 \times n_2}}} \quad \text{.....(1)}$$

where:
$$S = \sqrt{\frac{\sum_{i=1}^{n_1}(X_i - \overline{X}_1)^2 + \sum_{j=1}^{n_2}(X_j - \overline{X}_2)^2}{n_1 + n_2 - 2}} \quad \text{.....(2)}$$

where: $\overline{X}_1, \overline{X}_2 =$ means of first and second sub-samples, respectively.

n_1, n_2= number of years in the first and second sub-samples ,respectively.

X_i, X_j=annual value of the first and second sub-samples, respectively.

2. **f-test:** the test for homogeneity in variance of two sub-samples is given as follows:

$$f = \frac{S_1^2}{S_2^2} \quad \text{.....(3)}$$

with: $V_1=n_1-1$, $V_2=n_2-1$

where:

S_1, S_2= standard deviation of the first and second sub-samples, respectively.

V_1, V_2= degree of freedom of the first and second sub-samples, respectively.

If the calculated values of (t and f) is more than the critical values, then the trend is found in mean and variance, respectively. The results are summarized in table (1).

The regression coefficients of the second sub-samples are calculated and the results are summarized in table (2).

Table (1): Result of non-homogeneity test where critical (t) and (f) value (2.064) and (2.27).

Parameter	Data
t-calculated	0.425
Change in mean	None
f-calculated	0.666
Change in sd	none

Table (2): Regression coefficient.

Parameter	Reg. coeff. of mean		Reg. coeff. of sd	
Data	0.474	367.248	1.153	325.93

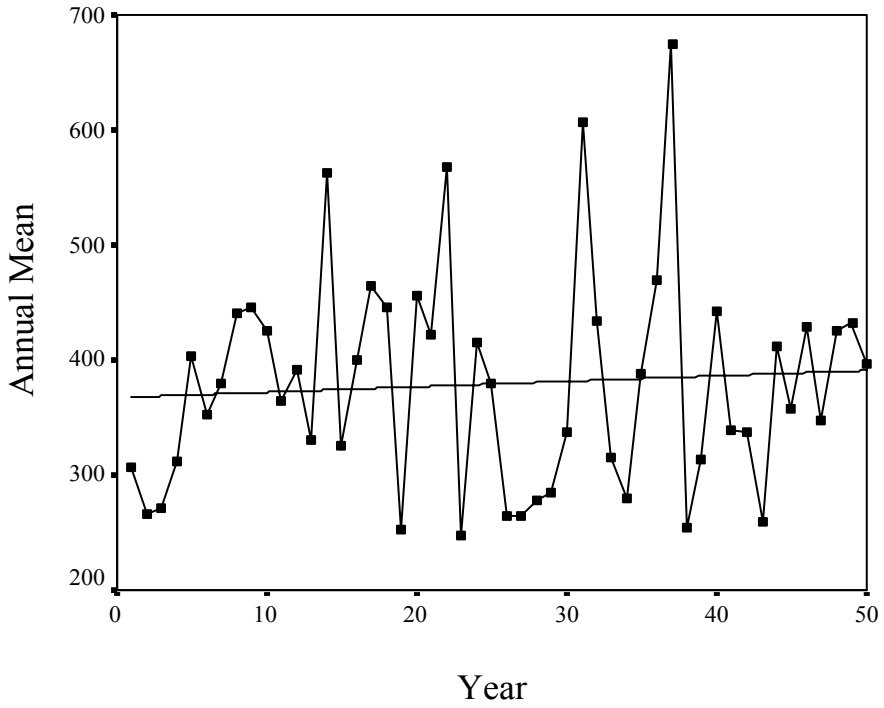

Fig. (2): Annual means of observed data before removing variation in mean from 1933-1982.

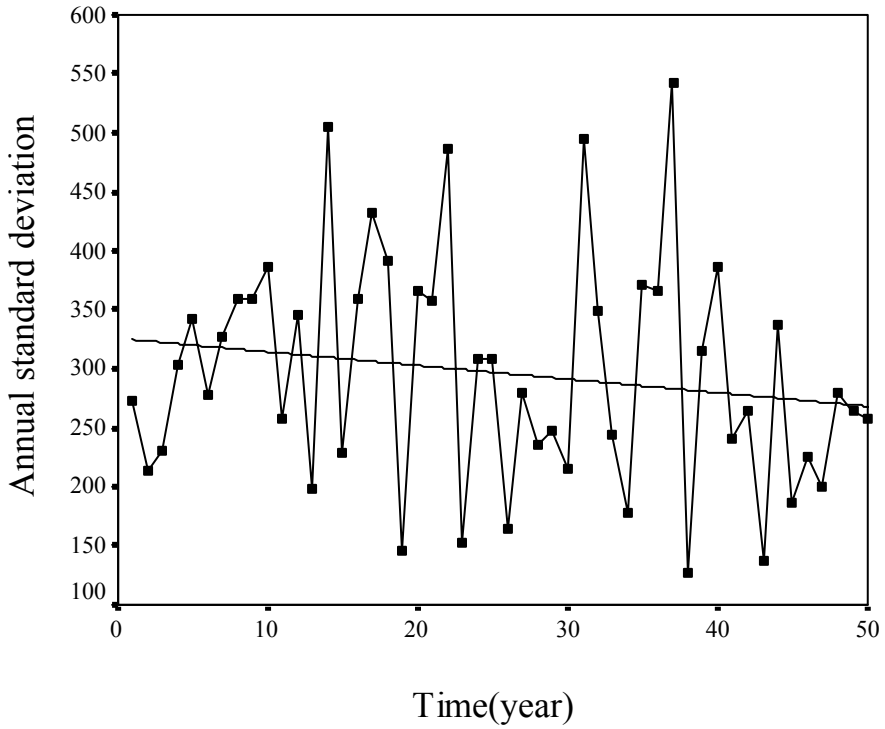

Fig. (3): Annual standard deviations of observed data before removing variation in mean from 1933-1982.

Transforming data to normal distribution

Time series observation of a given phenomenon required a certain type transformation (**Hipel et.al., 1977**). Before transformation data to normal distribution , the coefficients of skewness (C_s) and the coefficients of kurtosis (C_k) must be determined.The values of the four moments of the homogenous data are (Mean=379.34, Sd=93.69, C_s=.086,C_k=3.87).These values indicate that the data are closer to normal distribution because of the small (C_s) and (C_k) near 3.

It is often better to transform data to normal distribution to utilize its simple properties, its familiarity to most engineers, and to obtain satisfactory fit to data. The other reason for transforming the data includes stabilizing the variance and improving the normality assumption of the noise series (**Box and Jenkins, 1976**). Several transformations may be used to normalize the data but the most common and useful class of transformations for stabilizing the variance is the Box-Cox transformation as follows:

$$F_{i,j} = \frac{(Y_{i,j}^{\lambda} - 1)}{\lambda} \qquad \lambda \neq 0 \qquad \text{.....(4)}$$

$$F_{i,j} = LogY_{i,j} \qquad \lambda = 0 \qquad \text{.....(5)}$$

Where:

$F_{i,j}$ is the transformed series;

$Y_{i,j}$ is the varieties of given series;

λ is the constant of transformation.

The relationship between λ and C_s is being some form of second degree polynomial as:

$$\lambda = B_0 + B_1 C_s + B_2 C_s^2 \quad \text{.....(6)}$$

The value of (λ) is found by choosing random eight values of (λ) between (-1) and (1) and computing the corresponding (C_s) values for the series after transforming it by equation (4), (**Abed, 2007**). Then by fitting equation (5) to these eight points, the value of λ is found as equal to (B_0) corresponding to C_s=0. Table (3) show the effect of (λ) values on the first four moments mean, Sd, C_s, C_k of data.

Table(3):Effect of (λ) value on the first four moments of data.

(λ) values	**Mean**	**Sd**	**C_s**	**C_k**
λ=0.2	10.621	2.463	0.498	-0.759
λ=0.4	22.508	8.018	0.727	-0.392
λ=0.6	53.043	26.662	0.964	0.147
λ=0.8	136.670	90.508	1.212	0.876
λ=-0.2	3.361	0.248	0.053	-1.028
λ=-0.4	2.225	0.081	-0.169	-0.939
λ=-0.6	1.604	0.027	-0.397	-0.698
λ=-0.8	1.233	0.009	-0.631	-0.281

For normally distributed data (C_s=0) and (C_k≈3),the value of (λ=-0.242) with mean=3.06,Sd=0.196 ,C_s=0.007 ,and C_k=0.199.Normally distributed data has skewness equal to (0) and kurtosis equal to (3) , however, was found that it is not possible to (λ) values which simultaneous satisfy the two conditions (C_s=0 , C_k=3) of normality ,(**Chow et.al., 1988**) does not recommend moments of order higher than (3) for statistical analysis of hydrologic data , therefore, the (λ) values for (C_s=0) are chosen as the required normalizing coefficients.

The normality was tested by plotting observed cumulative probanility against expected cumulative probability using the following formula which is a **Weibull** formula:

$$P(x) = \frac{m}{N+1} \quad \text{.....(7)}$$

Where:

P(x) is the observed cumulative probability of the value (x) of transformed data;

m is the rank (x) in ascending order;

N is the number of tested data (N=600)

The resulting plot is shown in figure (4). The figure show good agreement.

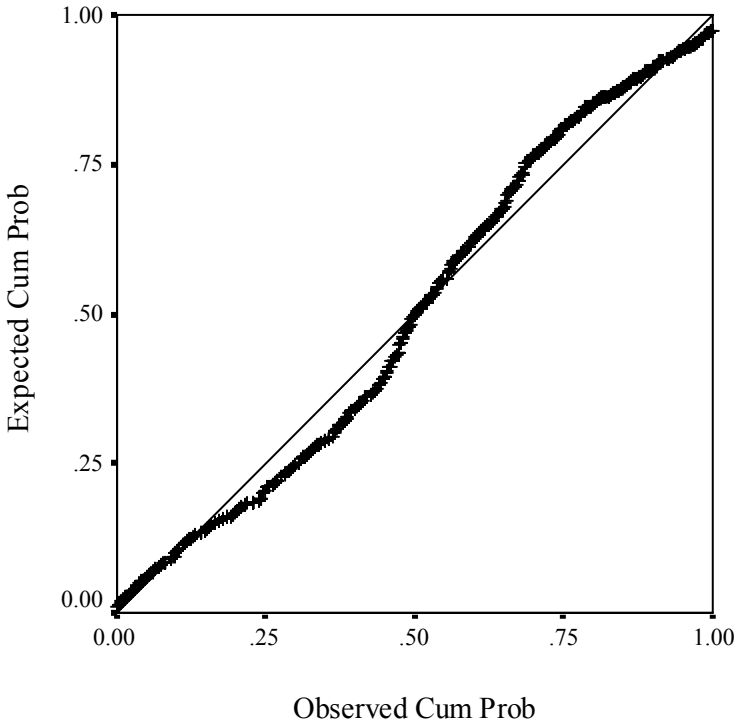

Fig. (4): Normal distributed test of Data.

Detection and Removal of Periodic Components

If a time series contain a seasonal fluctuation then the correlogram which is a plot of autocorrelation coefficient against the lag will also exhibit an oscillation at the same frequency (**Chtafield, 1982**). Then the periodicity can be detected by the correlogram.The correlogram of normalized data is shown in figure (5).

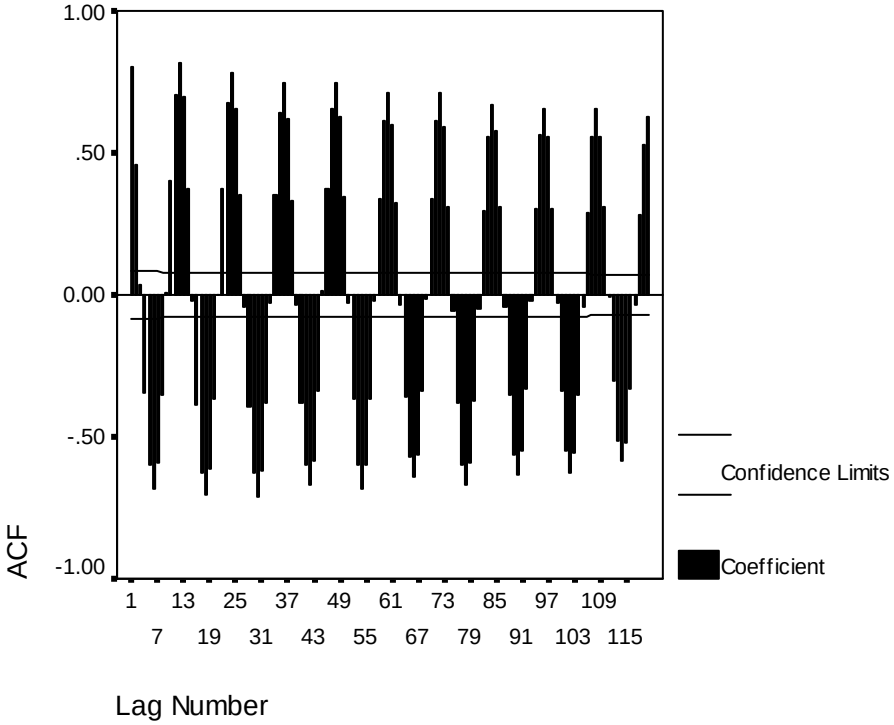

Fig.(5): Correlogram of normalized series.

Removal of periodicity from data is done by the nonparametric method by using the following equation:

$$W_{i,j} = \frac{F_{i,j} - \mu_j}{\sigma_j} \quad \text{.....(8)}$$

Where: $W_{i,j}$ is the dependent stochastic components of year i, and month j.

μ_j, σ_j are the mean and standard deviation for month (j) respectively.

The result series is called stochastic series, which contains a dependent part on time which may be represented by AR(p), MA(q) or ARMA(p,q) models, and an independent part (a_t) can be described by some probability distribution function.

Stochastic Model

This section contains the following steps:

Identification of the Model

Autocorrelation function (ACF) and partial autocorrelation function (PACF) is an important guide to the properties of time series because they provide insight into the probability model which generated the data. Figure(6) show the behavior of the (ACF)and (PACF) of dependent stochastic series ($W_{i,j}$) with 95% confidence limits .The suggested model was AR(1) because the (ACF) tail off while its (PACF) has a cut off after lag(1) .

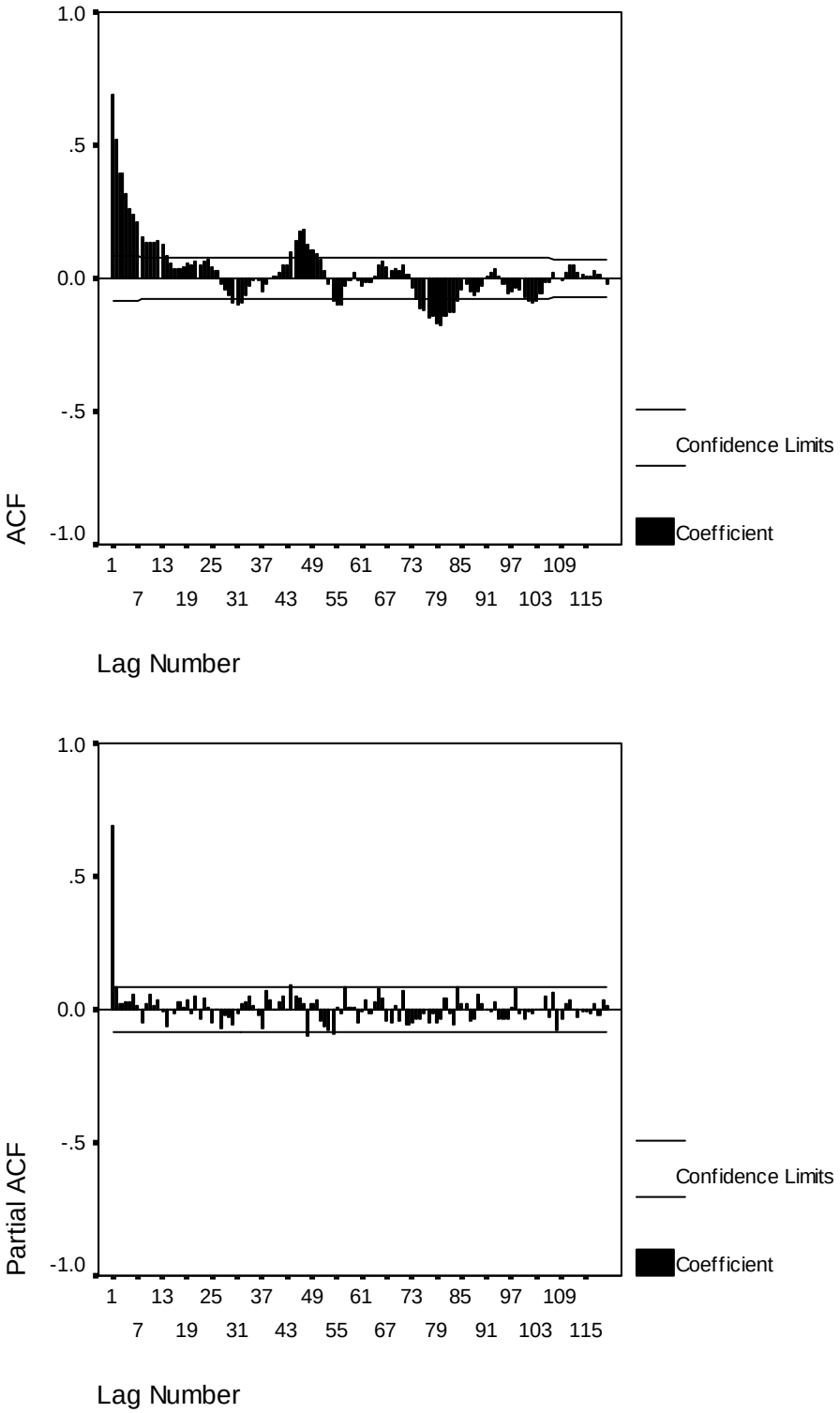

Fig.(6): ACF &PACF for Data.

Estimation of Model Parameters

The identification process having led to a tentative formulation for the model, it is needed to obtain efficient estimation of the model parameters. The method of estimation based on the first estimated autocorrelations as follows (**Box and Jenkins, 1976**):

AR (1) process:

$\Phi_1=r_1 \quad (-1<\Phi_1>1)$ …..……..…..……..(9)

Where: r_1=lag one estimated autocorrelation coefficient.

In this model the value of the process is express as a finite linear aggregate of previous value of the process and a shock (a_t).By denoting the value of a process at equally spaced time t,t-1,t-2,…. by W_t, W_{t-1}, W_{t-2},…. Then:

$W_t= \Phi_1 W_{t-1}+ \Phi_2 W_{t-2}+…..+ \Phi_p W_{t-p}+ a_t$ ……..……..……(10)

Where: Φ_1 , Φ_2 ,…….., Φ_p are the autoregressive model parameter.

p is the model order.

W_t is the value of stochastic series at time (t).

a_t is the shock (independent part of stochastic process).

Therefore the AR (1) model parameter (ϕ_1) equal to (0.698), then the model equation is:

$$W_t = 0.698W_{t-1} + a_t \quad \text{…….……(11)}$$

Model Diagnostic Checking

The diagnostic checking are then applied to test the adequacy of the fitted model. Therefore the following statistical tests are used:

1. Port Manteau Lak of fit Test

It is a test of the residual independency and uses the Q-statistic defined as follows:

$$Q = N\sum_{k=1}^{M} r_k^2(a_t) \quad \text{…….……..……..(12)}$$

Where:

$r_k(a_t)$ is the autocorrelation coefficient of the residual (a_t)at lag k.

M is the maximum lag considered (N/5).

If the a_t is independency, then the calculated Q, which is approximately chi-squared distributed with (M-p) degree of freedom, should be less than $\chi^2_{(M-p-q)}$ degree of freedom(**Jayawardena and Lai, 1989**) .Therefore the model is succeeded because the (Q-calculated =115.19 with, M=N/5) is less than (χ^2-table=146.57 with, degree of freedom =M-p-q ,and confidence limit=95%).

2. Residual autocorrelation Function (RACF) Test

The second test is the independency of the resulting (a_t) series, the correlogram of this series are computed for lag (M=N/5) are shown in figure (7) . The figure shown that the most of computed lags lie inside the tolerance interval (±2/√N, at 95% confidence limits). Hence, the suggested model can be considered as appropriate model because of its capability of removing the dependency from data.

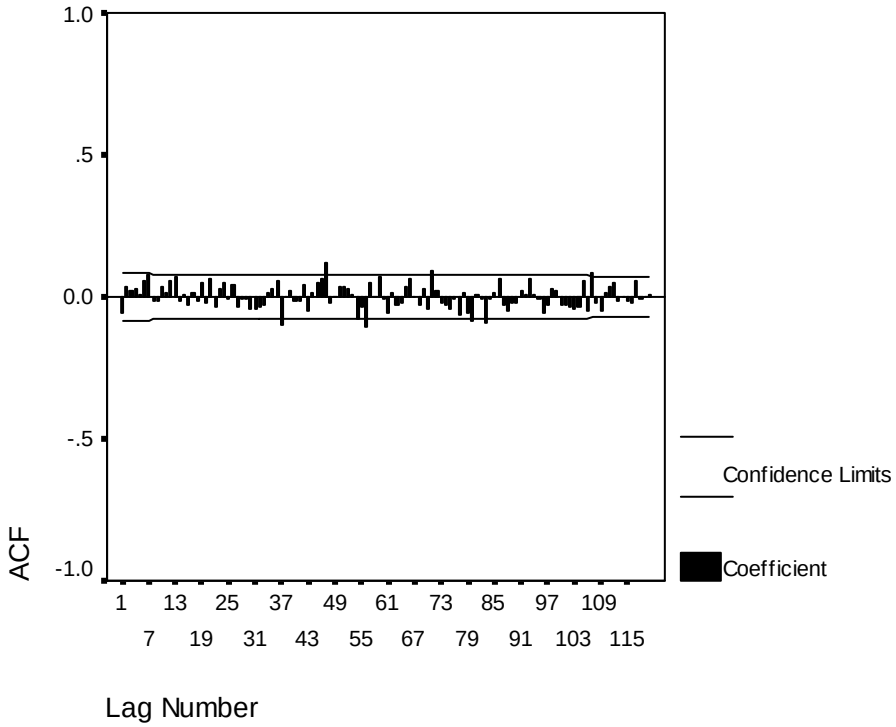

Fig.(7): Autocorrelogram of residual series parameter.

Forecasting

Forecasting monthly data are computed for the period from 1983 to 1988 by applying forecasting equation:

$$\hat{Z}_t(1) = 0.698 Z_t \quad \text{.....(13)}$$

$$\hat{Z}_t(\ell) = 0.698 \hat{Z}_t(\ell - 1) \quad \text{for } \ell = 2,3,\ldots,36 \quad \text{.....(14)}$$

Where: $\hat{Z}_t(\ell)$ = Forecasted series at origin time t and lead time ℓ.

To illustrate the forecasting procedure, the calculation of parameter is given in table (4). In this table Z_t (col.2) is calculated from model forecasting equation when the origin time is (t=600) and lead time (l=1, 2,…, 60). F_t (col.5)is calculated by reversing the standardization presses, then $F_t=Z_t*\sigma_j+\mu_j$ .Forecasted series(col.6) is calculated by reversing Box-Cox transformation with λ=-0.242.Figure(8) show the forecasted values obtained by the model and the observed values.

Table(4):Calculation of forecasting value of data.

Time (month)	Z_t	μ_j	σ_j	F_t	Forecasted data(m^3/s.)	Observed data(m^3/s.)
600	1.2984					
601	0.8920	2.824	0.097	2.9095	154	155
602	0.6128	2.894	0.109	2.9600	183	208
603	0.4210	2.95	0.104	2.9932	205	177
604	0.2892	3.011	0.113	3.0433	247	205
605	0.1987	3.119	0.103	3.1392	362	289
606	0.1365	3.219	0.084	3.2303	539	552
607	0.0938	3.325	0.057	3.3302	875	773
608	0.0644	3.322	0.063	3.3259	856	839
609	0.0443	3.222	0.065	3.2248	526	551
610	0.0304	3.062	0.084	3.0645	268	212
611	0.0209	2.911	0.084	2.9127	155	156
612	0.0144	2.831	0.078	2.8320	119	145
613	0.0099	2.824	0.097	2.8249	116	117
614	0.0068	2.894	0.109	2.8946	146	157
615	0.0047	2.95	0.104	2.9504	176	206
616	0.0032	3.011	0.113	3.0113	219	201
617	0.0022	3.119	0.103	3.1191	333	260
618	0.0015	3.219	0.084	3.2190	512	491
619	0.0010	3.325	0.057	3.3250	852	704
620	0.0007	3.322	0.063	3.3220	839	588
621	0.0005	3.222	0.065	3.2220	518	513
622	0.0003	3.062	0.084	3.0620	265	239
623	0.0002	2.911	0.084	2.9109	154	118
624	0.0002	2.831	0.078	2.8309	118	95
625	0.0001	2.824	0.097	2.8239	115	95
626	7.49E-05	2.894	0.109	2.8939	145	302
627	5.14E-05	2.95	0.104	2.9499	176	206
628	3.53E-05	3.011	0.113	3.0109	220	317
629	2.43E-05	3.119	0.103	3.1189	333	702
630	1.67E-05	3.219	0.084	3.2189	512	767
631	1.15E-05	3.325	0.057	3.3249	853	1127
632	7.87E-06	3.322	0.063	3.3219	839	950
633	5.41E-06	3.222	0.065	3.2219	518	491
634	3.71E-06	3.062	0.084	3.0619	265	204
635	2.55E-06	2.911	0.084	2.9109	154	134
636	1.75E-06	2.831	0.078	2.8309	118	117
637	1.2E-06	2.824	0.097	2.8239	116	103
638	8.28E-07	2.894	0.109	2.8939	145	120
639	5.68E-07	2.95	0.104	2.9499	176	184
640	3.91E-07	3.011	0.113	3.0109	219	277
641	2.68E-07	3.119	0.103	3.1189	333	394
642	1.84E-07	3.219	0.084	3.2189	511	410
643	1.27E-07	3.325	0.057	3.3249	852	731
644	8.7E-08	3.322	0.063	3.3219	839	728
645	5.98E-08	3.222	0.065	3.2219	519	383

646	4.11E-08	3.062	0.084	3.0619	265	207
647	2.82E-08	2.911	0.084	2.9109	154	92
648	1.94E-08	2.831	0.078	2.8309	118	102
649	1.33E-08	2.824	0.097	2.8239	116	117
650	9.15E-09	2.894	0.109	2.8939	145	260
651	6.28E-09	2.95	0.104	2.9499	176	229
652	4.32E-09	3.011	0.113	3.0109	219	329
653	2.97E-09	3.119	0.103	3.1189	333	422
654	2.04E-09	3.219	0.084	3.2189	512	681
655	1.4E-09	3.325	0.057	3.3249	852	805
656	9.62E-10	3.322	0.063	3.3219	839	980
657	6.61E-10	3.222	0.065	3.2219	519	603
658	4.54E-10	3.062	0.084	3.0619	266	321
659	3.12E-10	2.911	0.084	2.9109	154	179
660	2.14E-10	2.831	0.078	2.8309	118	130

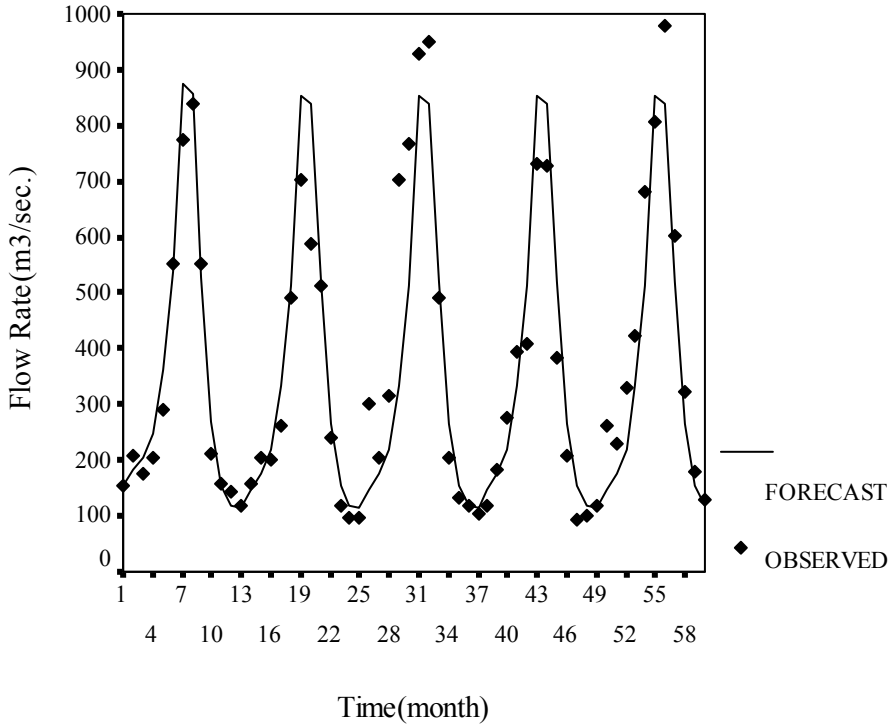

Fig.(8): Comparison between forecasted and observed series.

Conclusion

Based on this study results, the following conclusions are drawn:

1. The data was found to be homogenous in mean and standard deviation.
2. The data show seasonal pattern, the pattern may be due to the influence of the annual cyclic pattern of the hydrological inputs to the reservoir.
3. According to ARMA method, the monthly flow rate give the autoregressive AR (1) model, which state that the data are dependent on the data of the last (1) month.

Acknowledgement

Thanks are due to Dr. Salah Tawfeek, Assis. Professor of Civil Engineering, Babylon University, for their encouragement, reading and comments concerned with this study.

References

Abed,Z. A.Al-Ridah ,(2007):"Stochastic Models of Some Properties of water in the middle of Euphrates Region in Iraq.",M.Sc.Thesis ,College of Engineering, University of Babylon.

Al-Husseini,A.H.,(2000):"Hydrochemical Forecasting Model of Shutt Al-Hilla.", M.Sc. Thesis , College of Engineering, University of Babylon.

Al-Mousawi,E.M.,(2003):"Multisite Stochastic Model of water Quality Properties at Selected Regions.",M.Sc.Thesis , College of Engineering, University of Babylon. Al-Suhaili,R.H.,(1986):" Stochastic Analysis of Daily Streamflow of Tigris River.",M.Sc.Thesis , College of Engineering, University of Baghdad.

Box,G.E.P.and Jenkins,G.M.,(1976):"Time Series Analysis; Forecasting and Control.", Holden Day , San Francisco, California.

Chatfield, C.,(1982):"The Analysis of Time Series; An Introduction.", Chapman and Hill, 2nd Edition.

Chow,V.T.,Maidment,D.R.,and Mays,L.W.,(1988):"Applied Hydrology.",McGraw-Hill.

Hipel, K.W., McLeeod, A.I. and Lennox, W.C.,(1977):"A dvance in Box-Jenkins Modeling;1-Model Construction.", Journal of water Resource Research, Vol.13, No.3.

Hussien, W.A.B.,(2006):"Reliability of Iraqi Reservoirs.", M.Sc. Thesis, College of Engineering, University of Babylon.

Jayawardena, A.W., and Lai,F.,(1989):" Time Series of Water Quality Data in Pearl River-China."Journal of Environmental Engineering, ASCE,Vol.115,No3.

Zou, S., and Yu, Y.S.,(1996):"A dynamic Factor Model for Multivariate Water Quality Time Series with Trend.",Journal of Hydrology,178.