Stochastic Model for Monthly Flow to Haditha Reservoir

Riyadh Hamad Mohammed
University of Babylon

Abstract

In this study, time series analysis is applied to records of mean inflow to Haditha reservoir in the west of Iraq for the period from water year 1988/1989 to the water year 2007/2008. This method involves decomposition of the historical series into trend, seasonality and residual components. Split-Sample method is used to test the homogeneity of the series, and it is found homogenous .Box-Cox transformation method is used to transform the series to normal distribution. Removal of periodicity from series is done by the nonparametric method.

Two stochastic known as ARMA models are fitted to this series. These are AR (1) and AR (2) models. The unconditional sum of squares method is used to estimate the parameters of the models and to compute the sum of squared errors for each of the fitted model. It is found that the model which corresponds to the minimum sum of squared errors is the AR (1) model with autoregressive average parameters Φ =0.817.The adequacy of this model is checked by the diagnostic checking. Forecasts of monthly inflow for the period from October, 2008, to September, 2012 are found.

الخـلاصـة

في هذه الدراسة تم تطبيق تحليل السلاسل الزمنية على التصاريف الشهرية الداخلة إلى خزان حديثة في غرب العراق للفترة من تقالسلسلة الزمنية متجانسة. طبقت طريقة(Box-Cox transformation) لتحويل السلسلة الزمنية إلى التوزيع الطبيعي. تم إزالة مركبة الدورية بتطبيق طريقة (nonparametric).

تمت مطابقة نوعين من النماذج ألتصادفية المعروفة بنموذج(ARMA) ، وهذين الموذجين هما النموذج AR(1)والنموذج AR(2) .إن طريقة مجموع المربعات غير المشروطة قد استخدمت لتقدير معالم النماذج وحساب مجموع مربعات الأخطاء لكل نموذج. وقد وجد إن النموذج المقابل لأقل مجموع مربعات أخطاء هو AR(1) بمعلم θ=0.817 . إن كفاية هذا النموذج قد تم تدقيقها بفحص التشخيص (diagnostic).كما إن الفحص المعروف بفحص مخطط الذبذبة لم يبين وجود عدم عشوائية في المتبقيات للنموذج المذكور. وقد استخدمت المركبة المتبقية لغرض بناء النموذج التنبئي. ومن ثم تم إيجاد القيم المستقبلية للتصاريف للفترة من تشرين الأول /2008 ولغاية أيلول / 2012.

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)**.**Al-Suhaili (1986)** used single site 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 for 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 single site autoregressive models with first order AR(1), and multisite model with first order AR(1)to model of the monthly water quality data of eight hydro chemical 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 water river in Babylon, Najaf, and Diwaniya governorates. **Ali (2009)** fitted Box-Jenkins models to the seasonal time series of monthly inflow to Bekhme reservoir in the north of Iraq for the period from water year 1933/1934 to water year 2001/2002.

Application of ARMA Model

The present study is applying the well-known mixed autoregressive moving average ARMA model to flow rate for Haditha dam, to forecast 5 years of future time series and compared with flow rate standards. 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 removed by non-parametric method and the resulting series is called stochastic series which is used for building 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 a 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 is done by applying the non-parametric method for the normalized series.
5. Estimation of model parameter is done by plotting the corrlograms of the data and by the least squares algorithm for conditional model.
6. The diagnostic check of a model is done by one of some tests and the independency of the residuals is tested.
7. Forecasting and verifying the forecasted models.

Test and Removal of Non-Homogeneity

Hydraulic and water quality time series are commonly non-stationary with trends and seasonality (**Zou and Yu, 1996**). This section includes non-homogeneity test of the mean and standard deviation, and its removal when the series is found to be non-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 10^{th} years long (1988-1997) and the second is ten years long also (1997-2008). Figures (1.a) and (1.b) 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 samples is given as follows:

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

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

where:

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

n_1,n_2: number of years in the first and second samples ,respectively.

X_i,X_j: annual value of the first and second samples, respectively.

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

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

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

where:

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

V_1, V_2: degrees of freedom of the first and second sub-samples, respectively.

If the calculated value of (t and f) is more than the critical values, then the trend is found in the mean and variance, respectively. The series is homogenous because the (t-calculated =0.704 is less than the t-critical=1.960 and the f-calculated =0.88 is less than the f-critical=3.179).

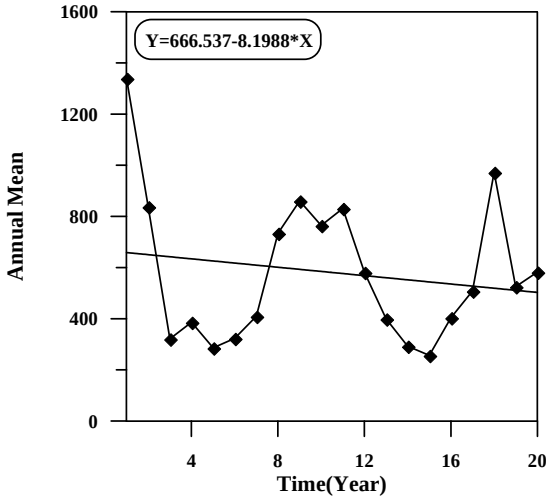

Fig. (1.a)

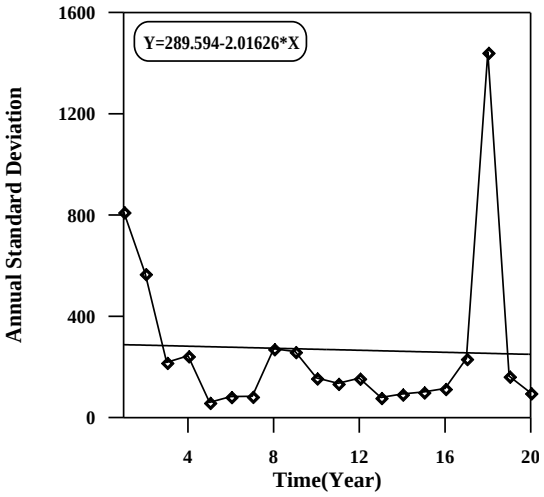

Fig. (1.b)

Fig. (1): Annual Mean and Standard Deviations of Observed Data before Removing Variation in Mean from 1988-2008.

Transforming the Data to Normal Distribution

Time series observation of a given phenomenon required a certain type of transformation (**Hipel et.al., 1977**). Before transformation the data to normal distribution, the coefficient of skewness (C_s) and the coefficient of kurtosis (C_k) must be determined. The values of the four parameters of the homogenous data are (Mean=580.45, S_d=489.50, C_s=5.329,C_k=45.607).

It is often better to transform the 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 and is given as follows:

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

$$F_{i,j} = LogY_{i,j} \qquad \lambda = 0$$

……..……...……...**(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:

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

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 using equation (4), (**Abed, 2007**). Then by fitting equation (6) to these eight points, the value of λ is found as equal to (B_0) corresponding to C_s=0. Table (1) show the effect of (λ) values on the first four parameters mean, S_d, C_s, C_k of data.

Table (1): Effect of (λ) Value on the First Four Parameters of Data.

(λ) values	**Mean**	**S_d**	**C_s**	**C_k**
λ=0.8	195.107	117.87	3.856	26.444
λ=0.6	70.528	29.5846	2.667	13.940
λ=0.4	27.873	7.7156	1.763	6.690
λ=0.2	12.304	2.0814	1.092	2.845
λ=-0.2	3.536	0.1646	0.178	0.132
λ=-0.4	2.283	0.0479	-0.163	-0.083
λ=-0.6	1.623	0.0142	-0.47	0.094
λ=-0.8	1.240	0.0043	-0.761	0.572

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) value for (C_s=0) is chosen as the required normalizing coefficient.

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

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

Where:

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

m: is the rank of (x) in ascending order;

N: is the number of tested data (N=240)

The resulting plot is shown in Figure (2). The Figure show good agreement.

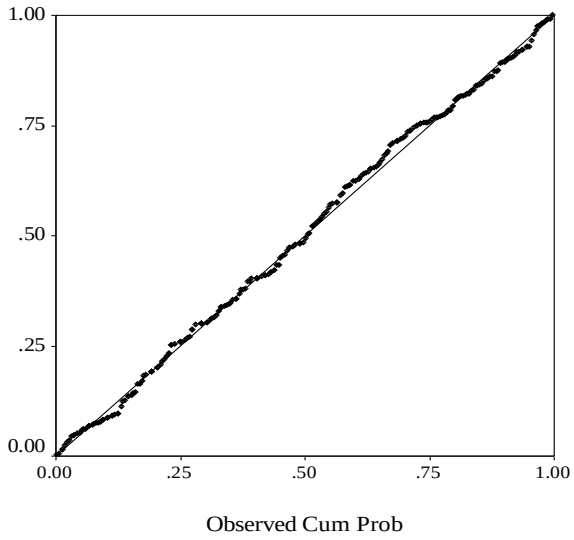

Fig. (2): Normal Distributed Test of Data.

Detection and Removal of Periodic Components

If a time series contains 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 (3).

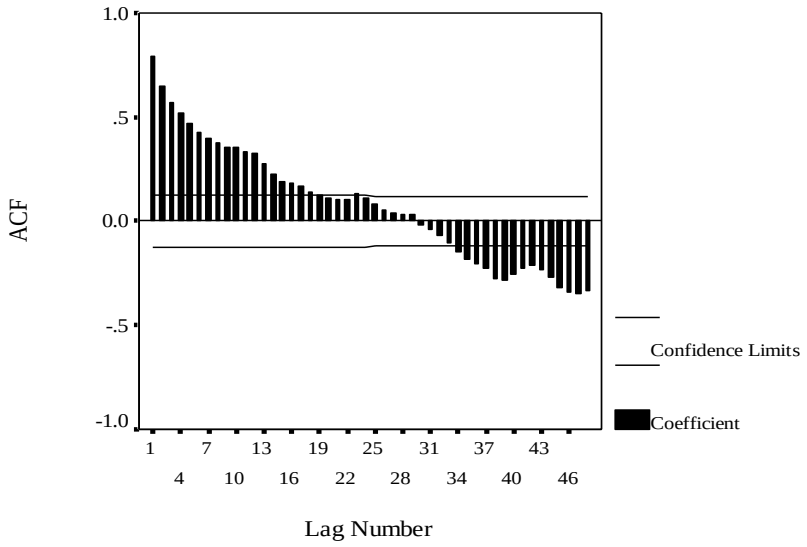

Fig.(3): Correlogram of Normalized Series.

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:

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

μ_j, σ_j: 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) which 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) are important guides to the properties of time series because they provide insight into the probability model which generated the data. Figure(4) shows the behavior of the (ACF)and (PACF) of dependent stochastic series ($W_{i,j}$) with 95% confidence limits .The suggested model was AR (1), and AR (2) because the (ACF) tail off while its (PACF) has a cut off after lag(2).

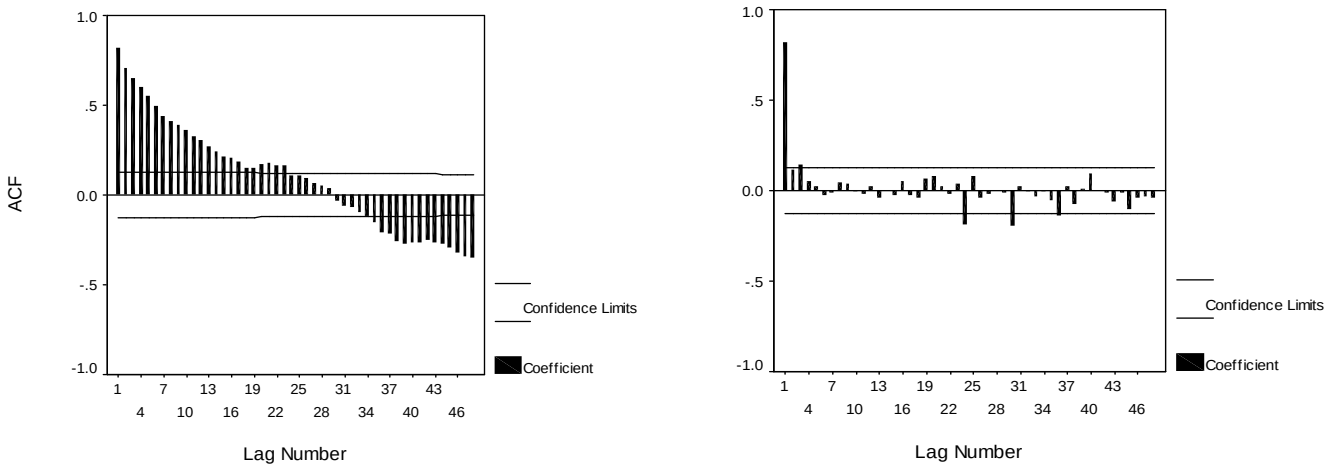

Fig.(4): ACF &PACF for Data.

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 is based on the first estimated autocorrelations as follows (**Box and Jenkins, 1976**):

AR (1) process:

$$\Phi_1=r_1 \qquad (-1<\Phi_1>1) \quad \text{…..……..…..……..(9)}$$

Where:

r_1: lag one estimated autocorrelation coefficient.

In this model the value of the process is expressed 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:

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

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 (0.817), then the model equation is:

$W_2=0.817*W_{2-1}+a_2$ ……(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 independent, 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 =45.25 with, M=N/5) is less than (χ^2-table=63.98 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 is computed for lag (M=N/5) and shown in figure (5). The figure shows that the most of the computed lags lie inside the tolerance interval ($\pm 2/\sqrt{N}$, at 95% confidence limits). Hence, the suggested model can be considered as appropriate model because of its capability of removing the dependency from data.

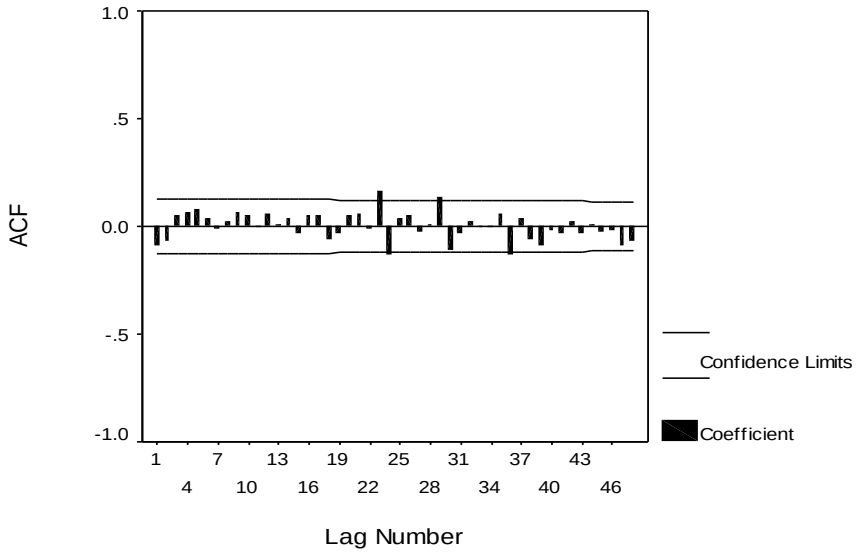

Fig.(5): Autocorrelogram of Residual Series Parameter.

Forecasting

Forecasting monthly data are computed for the period from 2008 to 2012 by applying forecasting equation: 0.817

$\hat{Z}_t(\ell) = 0.817 Z_t$ for ℓ =1 ……..(13)

$\hat{Z}_t(\ell) = 0.817 \hat{Z}_t(\ell - 1)$ for ℓ =2,3,…,36 ……..(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 (2). In this Table, Z_t (col.2) is calculated from model forecasting equation when the origin time is (t=240) 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=-0.3007.The results of Table(2) are plotted as shown in Fig.(7).

Table (2): Calculation of Forecasting Value of Data.

Time (month)	Z_t	μ_j	σ_j	F_t	Forecasted data(m^3/s.)
240	0.8968				
241	0.7327	2.7840	0.0608	2.8384	595
242	0.5986	2.7933	0.0645	2.8405	604
243	0.4890	2.7906	0.0898	2.8443	620
244	0.3995	2.7968	0.1186	2.8547	666
245	0.3264	2.8168	0.1045	2.8585	684
246	0.2667	2.8066	0.1011	2.8395	599
247	0.2179	2.7884	0.1261	2.8220	533
248	0.1780	2.7666	0.0984	2.7880	429
249	0.1454	2.8041	0.0785	2.8181	519
250	0.1188	2.8073	0.0682	2.8173	517
251	0.0971	2.8203	0.0651	2.8280	555
252	0.0793	2.8054	0.0584	2.8110	496
253	0.0648	2.7840	0.0608	2.7888	431
254	0.0529	2.7933	0.0645	2.7974	455
255	0.0433	2.7906	0.0898	2.7954	449
256	0.0353	2.7968	0.1186	2.8020	468
257	0.0289	2.8168	0.1045	2.8205	528
258	0.0236	2.8066	0.1011	2.8095	491
259	0.0193	2.7884	0.1261	2.7914	438
260	0.0157	2.7666	0.0984	2.76845	381
261	0.0129	2.8041	0.0785	2.8054	478
262	0.0105	2.8073	0.0682	2.8082	487
263	0.00856	2.8203	0.0651	2.8210	529
264	0.0070	2.8054	0.0584	2.8059	480
265	0.0057	2.7840	0.0608	2.7844	420
266	0.0047	2.7933	0.0645	2.7936	444

Time (month)	Z_t	μ_j	σ_j	F_t	Forecasted data ($m^3/s.$)
267	0.0038	2.7906	0.0898	2.7911	437
268	0.0031	2.7968	0.1186	2.7973	455
269	0.0026	2.8168	0.1045	2.8172	516
270	0.0021	2.8066	0.1011	2.8068	483
271	0.0017	2.7884	0.1261	2.7887	431
272	0.0014	2.7666	0.0984	2.7667	377
273	0.0011	2.8041	0.0785	2.8042	475
274	0.0009	2.8073	0.0682	2.8074	485
275	0.0008	2.8203	0.0651	2.8204	527
276	0.0006	2.8054	0.0584	2.8054	479
277	0.0005	2.7840	0.0608	2.7840	419
278	0.00041	2.7933	0.0645	2.7933	443
279	0.00034	2.7906	0.0898	2.7910	436
280	0.00027	2.7968	0.1186	2.7969	453
281	0.00022	2.8168	0.1045	2.8169	515
282	0.00018	2.8066	0.1011	2.8066	482
283	0.00015	2.7884	0.1261	2.7884	430
284	0.00012	2.7666	0.0984	2.7666	377
285	0.00010	2.8041	0.0785	2.8041	475
286	8.22E-05	2.8073	0.0682	2.8074	485
287	6.72E-05	2.8203	0.0651	2.8203	527
288	5.49E-05	2.8054	0.0584	2.8054	478
289	4.48E-05	2.7840	0.0608	2.7840	418
290	3.66E-05	2.7933	0.0645	2.7933	443
291	2.99E-05	2.7906	0.0898	2.7906	436
292	2.44E-05	2.7968	0.1186	2.7968	453
293	2E-05	2.8168	0.1045	2.8168	515
294	1.63E-05	2.8066	0.1011	2.8066	482
295	1.33E-05	2.7884	0.1261	2.7884	430
296	1.09E-05	2.7666	0.0984	2.7666	377
297	8.9E-06	2.8041	0.0785	2.8041	475
298	7.27E-06	2.8073	0.0682	2.8073	485
299	5.94E-06	2.8203	0.0651	2.8203	527
300	4.85E-06	2.8054	0.0584	2.8054	478

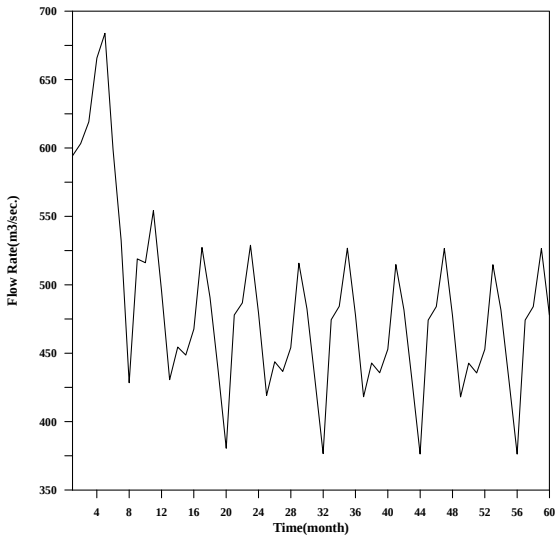

Fig.(7): Forecasted Monthly Inflow to Haditha Reservoir.

Conclusion

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

1. The series was found to be homogenous in mean and standard deviation.
2. The series show seasonal pattern, the pattern may be due to the influence of the annual cyclic pattern of the hydrological inputs to the reservoir.
3. Two of ARMA models are fitted to the series and it is found that sum of square errors of the AR(1) model with autoregressive average parameters of θ=0.817 is less than the other model AR(2). Which state that the data are dependent on the data of the last (1) month.

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.

Ali,S.,T.(2009)**:"Fitting Seasonal Stochastic Model to Inflow of Bekhme Reservoir**.", The Iraqi Journal for Mechanical and Material Engineering , Special Issue(A)

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):"**Advance in Box-Jenkins Modeling; 1-Model Construction**.",Journal of water Resource Research,Vol.13,No.3.

Al-Husseini,A.H.,(2000):"**Hydrochemical Forecasting Model of Shutt Al-Hilla.**",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.

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 Stream flow of Tigris River.**",M.Sc.Thesis , College of Engineering, University of Baghdad.

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