Al-Khwarizmi Engineering Journal, Vol. 4, No. 2, PP 27- 36 (2008)

Al-khwarizmi Engineering Journal

Time Series Analysis of Total Suspended Solids Concentrations in Euphrates River in Al-Nasria Province

Prof. Dr. Rafa Al-Suhaili **Ass.Lecturer. Tariq J.Al-Mosewi**

Environment Engineering Deptartement/ *College of Engineering/ University of Baghdad*

(Received 6 May 2007; accepted 3 April 2008)

Abstract

The monthly time series of the Total Suspended Solids (TSS) concentrations in Euphrates River at Nasria was analyzed as a time series. The data used for the analysis was the monthly series during (1977-2000).

The series was tested for nonhomogenity and found to be nonhomogeneous. A significant positive jump was observed after 1988. This nonhomogenity was removed using a method suggested by Yevichevich (7). The homogeneous series was then normalized using Box and Cox (2) transformation. The periodic component of the series was fitted using harmonic analyses, and removed from the series to obtain the dependent stochastic component. This component was then modeled using first order autoregressive model (Markovian chain). The above analysis was conducted using the data for the period (1977-1997), the remaining 3-years (1998-2000) of the observed data was left for the verification of the model. The observed model was used to generate future series. Those series were compared with the observed series using t-test. The comparison indicates the capability of the model to produce acceptable future data.

Key words: Total suspended solids, nonhomogeneous, periodic component, and dependent stochastic.

Introduction

Water quality express the suitability of water to sustain various uses or processes, such as drinking water, irrigation water and nature conservation.

Solids can be dispersed in water in both suspended and dissolved forms. Solids suspended in surface water may consist of inorganic particles or of immiscible liquids. Domestic wastewater usually contains large quantities of suspended solids. Suspended solids material may be objectionable in water for several reasons. It is aesthetically displeasing and provides adsorption sites for chemical and biological agents. Suspended organic solids may be degrading biologically resulting in objectionable by products. Biologically active suspended solids may include disease causing organisms as well as organisms such as producing strain of algae.

During the last 25 years, Time Series Analysis had become one of the most important and widely used branches of Mathematical Statistics. The technique of time-series analysis uses estimated statistical parameters to build a mathematical model. This model is capable of describing the evolution of possible sequences of events in time, at the site of observations, which have the same statistical properties as the historical sample.

In this research, the data of TSS in Euphrates River will be utilized to build the mathematical model to predict the future concentration. Numbers of projects were by now under construction in the southern of Iraq after March 2003, such as water treatment plants and projects related to restoration of marshes and irrigation projects. The modeling of TSS could be used to predict future values, that are useful in the operation of such project. Moreover these data are useful also in planning, design of new projects.

Garry L.Grabow et,al. (4) used weekly total suspended solids data to build a mathematical model. The data were collected from a measurement station located downstream of a dairy farm. Two types of statistical methods were used to detect the changes in the TSS concentrations, ANCOVA and ANOVA.

Zhiyong Huang and Hiroshi Morimoto(8) designed a structure of input/mixing/output model

(called it the three-step model) and represented each process of material input, mixing and output by a stochastic numerical methods. Data of BOD and DO concentrations of river Cam used to build the model. The stochastic numerical methods applied in this paper are discrete time approximation methods. Additional equations representing mixing processes and biochemical reactions also used in model. The researcher concluded that the theory of stochastic differential equations is a beneficial tool for studying water pollution.

Kadri Yurkel and Ahmet Kurunc (3) had analyzed the daily discharge data of each month from three gauge stations on Cekerek Stream for forecasting using stochastic approaches. Initially non-parametric test (Mann-Kendall) was used to identify the trend during the study period. The two approaches of stochastic modeling, ARIMA and Thomas-Fiering models were used to simulate the monthly-minimum daily discharge data of each month. The error estimates (RMSE and MAE) forecasts from both approaches were compared to identify the most suitable approach for reliable forecast. The two error estimates calculated for two approaches indicate that ARIMA model appear slightly better than Thomas-Fiering model. However, both approaches were identified as an appropriate method for simulating the monthly-minimum daily discharge data of each month from three gauge stations on Cekerek Stream.

The use of time series analysis (stochastic analysis) in generating possible future suspended solids concentrations assumes that those concentrations are extracted from a common statistical population, and that the recorded historical data forms a sample from this population.

Generally, a hydrologic time series may consist of four components depending on the type of variable and the averaging time interval. In seasonal TSS concentration series four components may exist and the TSS is considered to arise from a combination of those components, which are termed the jump component (J_t), trend component (T_t), periodic or cyclic component (P_t) and stochastic or random component (ε_t).These components may be formulated by:

$$TSS_t = J_t + T_t + P_t + \varepsilon_t \quad \ldots(1)$$

The first three components represents the deterministic part of the process while the fourth component represents the non-deterministic part, therefore those three components should be detected and identified by suitable formulations and decomposed from the stochastic component.

Methodology

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

1- Filling-in of Missing Data

Data of monthly TSS concentration in Euphrates river were mentioned in reports of the Ministry of Water Resources. These reports indicated some months with missing data in irregular years. Table (1) shows the missing readings in TSS in a period from 1977 to 1983.

Filling-in the missing data was performed by using the regression relations method.

2- Test and Removal of Non-homogeneity

The modeling process required a set data to be homogenous. Hence, the first step before starting the analysis is to test the homogeneity of the data series. If the test indicates non-homogeneity, then this non-homogeneity should be removed. This was achieved by plotting the average monthly data and computing the annual mean and standard deviation for each year then using the spilt-sample approach which divides the entire sample into two sub-samples. Then testing the differences between the means and standard deviations of these two sub-samples at the 95 percent probability level of significance using the t-test method.

The data were tested for non-homogeneity and found to be non-homogeneous at year 1988, see figure (1). The calculated t-value is greater than the tabulated t-value.

For non-homogeneity removal, Yevjevich (7) suggest fitting linear regression equations for both annual means and annual standard deviations, then applying the following equation:-

$$Yj,t = \frac{Xj,t - \overline{Xj}}{Sj} Sd2 + Av2 \quad \ldots(2)$$

where:

j, t= the annual and seasonal positions of the observations, respectively.

Y=transformed series(homogeneous)

X = historical non-homogeneous series.

Av2,Sd2= the average and standard deviation of the second sub –sample respectively, and

Xj,Sj= linear regression equations for annual means and standard deviations against years.

Thus the data are divided into two sub-samples, the first (12) years long (1977-1988) and the second (9) years long (1989-1997). The first sub-sample will be transformed according to the above equation. The two sub-samples were then tested again using t-test to check the homogeneity.

The result of applying the above equation are shown in figure (2) and table (2) below, hence the

data are homogeneous. Since the calculated t-values is less than tabulated t-value.

3- Transformation to Normally Distributed data

It is of common practice in time- series analysis to transform the data to the normal distribution. This means, to remove the skewness in the data and try to make it nearly zero. For the normalization process several transformations may be used to normalize the data, but the most common one is the power transformation. The power transformation used in this research is the one suggested by Box and Cox (2) (see equation (3) below). The application of this transformation begins with the estimation of the transformation coefficient value (λ). This coefficient has a value between (-1) and (1) and is strongly related to the skewness coefficient (C_s).Table (3) shows values of C_s computed for each transformed series, using different λ-values.

The values above were found to best fitted by a second polynomial equation.

$$\lambda=0.0249C_s^2+0.0132C_s+0.01 \quad \ldots(3)$$

In order to find the λ-value that will normalize the data, the skewness coefficient Cs in the above equation is substituted by zero, which gives a λ value as 0.01. This value will be used to get the transformed series according to the Box and Cox transformation as follows:

$$y=\frac{(x-1)^{\lambda}}{\lambda} \quad \ldots(4)$$

where:

y: The transformed Series, x : The original Series Data.

Table (4) shows the monthly means and standard deviations for the original and transformed series.

4- Determination of the Independent Stochastic Component

The series obtained after the removal of non-homogeneity and non-stationary (periodic component of mean and standard deviation) is termed as the dependent stochastic component of the process and denoted as ($\varepsilon_{p,t}$). The values of monthly mean and standard deviations were used to find the value of the independent stochastic component by the following equation:-

$$\varepsilon_{p,t}=\frac{y_{p,t}-\mu_t}{\sigma_t} \quad \ldots(5)$$

where:

$\varepsilon_{p,t}$ = is the dependent stochastic component.

μ_t =is the mean value of $y_{p,t}$ data at position p(month).

σ_t =is the standard deviation value of $y_{p,t}$ data at position p(month).

The values of $\varepsilon_{p,t}$ may be fitted by a suitable model whose parameters will depend directly or indirectly on the amount of existing correlation represented by the lag r_k serial correlation coefficient model (r_k) ,figure (3 a). One of the most familiar models, are the autoregressive model. It is preferable to try the first degree model, and then check its adequacy to remove the dependency of the $\varepsilon_{p,t}$ series in the first degree model fails to remove the dependency, the second degree model will be used, and so on. The first degree autoregressive model required the calculation of lag-one (r1) serial correlation coefficient, which was found to be r1=0.664.

The model is represented as the relation between the dependent stochastic component ($\varepsilon_{p,t}$) and the independent stochastic component ($\xi_{p,t}$). The independent stochastic series ($\xi_{p,t}$) is a series of random numbers usually with zero mean and unit variance.

As mentioned above one of the most used models is the first order autoregressive model (Markov model). This model express the relationship between the $\varepsilon_{p,t}$ and $\xi_{p,t}$ as follows:

$$\varepsilon_{p,t}=a \times \varepsilon_{p,t-1}+\sqrt{1-a^2}\ \xi_{p,t} \quad \ldots(6)$$

where: $a= r_1$

Substituting the value of a =(0.664) in the above equation the independent stochastic component $\xi_{p,t}$, could be found using:

$$\xi_{p,t}=(\varepsilon_{p,t}-0.664 \times \varepsilon_{p,t-1})/0.747 \quad \ldots(7)$$

In order to test the validity of the proposed first order autoregressive model, the correlogram of the $\xi_{p,t}$ component should be found and tested. This correlogram is shown below which show that the first order autoregressive model, is suitable since the values of the serial correlation coefficient are fluctuated around the zero-values. Hence the proposed model was capable of removing the dependency between the values of $\varepsilon_{p,t,}$ figure (3 b).

5- Model Verification:

Upon the completion of the first four steps above, the model parameters were found. As mentioned before the observed data series of TSS was divided into two parts (1977-1997), was used for the analysis (i.e., models parameters

estimation), the other part (1998-2000), will be used now for model verification.

Usually in practice, the model is used to generate future values (series). The model validity will be decided upon the comparison between the statistical properties of the generated series with those of the observed one that was not used in the estimation of the parameters of the model.

The Microsoft Excel program was used for generating future series for the TSS-values. Three series were generated as shown in table (5). The generation process begins by generating a standardized normally distributed random series (i.e., with zero mean and unit variance) then, using those as $\xi_{p,t}$ values to generate the $\varepsilon_{p,t}$ values using the first autoregressive model. The TSS values were found using a reverse process of the analysis conducted in steps (2-4).

Table (6) shows the generated monthly TSS values using the three generated randomized series, proposed to be for years 1998, 1999 and 2000. The observed TSS concentrations for these 3-years are shown in table (7).

Figure (4) shows the values of mean monthly TSS concentrations calculated from the generated series rand 1, rand 2, and rand 3 numbers as well as those calculated using the observed TSS record for the period 1998 to 2000. Table 8 shows the values of the mean, standard deviation and skewness coefficient of observed data and those of the observed one.

Table (9) shows the results of the t-test for monthly means of observed and generated TSS Series.

Conclusions

1) The series of monthly TSS concentrations in Euphrates River at Al-Nasria Province is non-homogeneous. The non-homogeneity can be attributed to the disposal of effluent wastewater from the constructed treatment plants.
2) The suitable value of the power transformation parameter λ that can be used to transform data to the normal distribution was found to be 0.01.
3) The correlogram of the observed independent stochastic component indicate the capability of the first order autoregressive model to model to time-dependency of the dependant stochastic component.
4) The T-test result shows that the obtained model can presence future forecasted values for the monthly TSS values.

Table 1
The Missing Data of the Recorded Monthly TSS (mg/l) in a period from 1977 to 1983.

Year / Month	1977	1978	1979	1980	1981	1982	1983
Jan.	1056	1417	972	1264	1738	1168	4187.7
Feb.	×	1518	1204	1301	1728	1325	2011
March.	1145	1421	1222	×	1557	1650	1912
April.	×	1322	1209	1290	1578	1578	×
May.	1160	×	1248	×	1730	1730	×
June.	1416	1667	1584	1491	1545	1545	1511
July.	×	1273	1528	×	1194	×	1601
August.	1366	1412	1293	1658	1145	1600	1780
Sep.	1429	1316	1834	1560	1375	×	1620
Oct.	1447	1250	1301	1620	1228	1311	1555
Nov.	1457	×	1633	1720	1253	×	1858
Dec.	1573	1024	1169	1805	1233	1954	1340

× missing data

Table 2
Mean and Standard Deviation of each Sub-Samples before and after Applying the Procedure of Removal of Non-Homogeneity.

	Before Removal		*After Removal*	
	Mean	Standard deviation	Mean	Standard deviation
Set 1 data	1699.60	427.47	2821.42	386.52
Set 2 data	2821.41	386.50	2821.41	386.50

Table 3
Variation of Skewness Coefficient with Box and Cox Transformations Coefficient.

λ	-0.8	-0.6	-0.4	-0.2	0.2	0.4	0.6	0.8
C_s	0.005	0.016	0.005	0.039	-0.007	0.010	0.027	0.045

Table 4
Monthly Means and Standard Deviations for the Original Homogeneous series and the Normalized Series in a period from 1977 to 1997.

	Original Homogeneous Series (x)		*Normalized Series* (y)	
	Mean	**Standard Deviation**	**Mean**	**Standard Deviation**
January	2817.434	716.122	108.240	0.243
February	2896.002	726.797	108.263	0.284
March	2980.307	633.320	108.306	0.223
April	2913.758	778.573	108.268	0.286
May	3030.474	669.865	108.322	0.236
June	2722.360	449.878	108.216	0.176
July	2734.419	305.069	108.229	0.117
August	2929.858	426.885	108.300	0.150
September	2824.034	306.742	108.264	0.118
October	2678.520	374.626	108.203	0.140
November	2718.747	343.527	108.220	0.139
December	2611.145	562.362	108.163	0.222

Table 5
Values of Generated Randomized Numbers ($\xi_{p,t}$) for 1998,1999 and 2000 years.

	Rand 1			Rand 2			Rand 3		
	1998	1999	2000	1998	1999	2000	1998	1999	2000
Jan.	0.271	0.359	0.465	0.850	0.150	0.815	0.210	0.041	0.363
Feb.	0.313	0.476	0.559	0.075	0.987	0.184	0.132	0.724	0.944
Mar.	0.188	0.490	0.866	0.844	0.801	0.140	0.218	0.955	0.855
Apr.	0.980	0.807	0.849	0.105	0.848	0.974	0.670	0.173	0.363
May.	0.348	0.223	0.436	0.859	0.619	0.073	0.427	0.267	0.961
June	0.721	0.719	0.290	0.365	0.038	0.599	0.283	0.171	0.053
July.	0.594	0.060	0.727	0.426	0.997	0.727	0.383	0.262	0.953
Aug.	0.388	0.274	0.301	0.475	0.497	0.537	0.824	0.101	0.176
Sep.	0.273	0.277	0.640	0.799	0.893	0.667	0.453	0.278	0.047
Oct.	0.073	0.130	0.390	0.943	0.929	0.094	0.382	0.731	0.266
Nov.	0.295	0.520	0.634	0.295	0.932	0.087	0.568	0.208	0.770
Dec.	0.964	0.650	0.629	0.654	0.783	0.698	0.126	0.471	0.981

Table 6
Generated Monthly TSS Concentrations (mg/l) for years (1998, 1999, and 2000).

	TSS Conc. From *Rand 1*			Mean	TSS Conc. From *Rand 2*			Mean	TSS Conc. From *Rand 3*			Mean
	1998	1999	2000		1998	1999	2000		1998	1999	2000	
Jan.	3318	3159	3419	3298	3184	3439	3664	3429	3524	3600	3822	3649
Feb.	3915	3241	3630	3595	3599	3396	3716	3571	3751	3856	4254	3954
Mar.	3563	3206	3407	3392	3355	3332	3818	3502	3713	3670	3670	3685
Apr.	3939	3624	3810	3791	3717	3158	3735	3536	3644	4005	4029	3892
May.	3902	3565	3625	3697	4047	3625	4031	3901	3849	4132	4150	4044
June	3171	3304	3209	3228	3485	3040	3248	3257	3104	3503	3435	3347
July.	3025	3062	3016	3034	3117	3085	3098	3100	2936	3111	3261	3103
Aug.	3392	3477	3361	3410	3547	3322	3308	3392	3121	3484	3727	3444
Sep.	3185	3135	3113	3144	3220	3162	3080	3154	3034	3098	3459	3197
Oct.	3116	2938	3058	3037	3018	3057	2911	2995	2960	3055	3441	3152
Nov.	3096	2931	3192	3073	3187	2997	3031	3071	3064	3038	3340	3147
Dec.	2977	3123	3062	3054	3272	3165	2992	3143	3168	3220	3290	3226

Table 7
Observed TSS Concentrations (mg/l) from Period 1998 to 2000.

	1998	1999	2000
January	4870	3535	4470
February	5326	3570	4530
March	5975	3535	3997
April	6094	3680	3608
May	5879	3710	3519
June	5530	3704	4800
July	4320	3620	3355
August	6530	3635	5370
September	5092	3540	3073
October	5928	4416	3376
November	6975	4514	2895
December	7148	4612	3690

Table 8
Overall Properties of Observed Data and Generated Series.

	Observed data	Rand 1	Rand 2	Rand 3
Mean	4512	4425	4456	4654
Standard deviation	422.69	356.93	353.59	460.03
Skewness coefficient	0.6338	0.6915	0.4685	-0.1061

Tables 9
Results of the t-test for Monthly TSS Means, t tabulated= 2.776.

	T rand 1	T rand 2	T rand 3
January	*0.226*	*0.655*	*1.402*
February	*0.535*	*0.557*	*1.467*
March	*0.033*	*0.221*	*0.556*
April	*0.726*	*0.304*	*0.882*
May	*0.735*	*1.075*	*1.338*
June	*0.690*	*0.591*	*0.376*
July	*0.991*	*1.297*	*1.200*
August	*0.741*	*0.766*	*0.664*
September	*0.484*	*0.503*	*0.595*
October	*0.697*	*0.773*	*0.477*
November	*0.581*	*0.584*	*0.497*
December	*1.036*	*0.917*	*0.814*

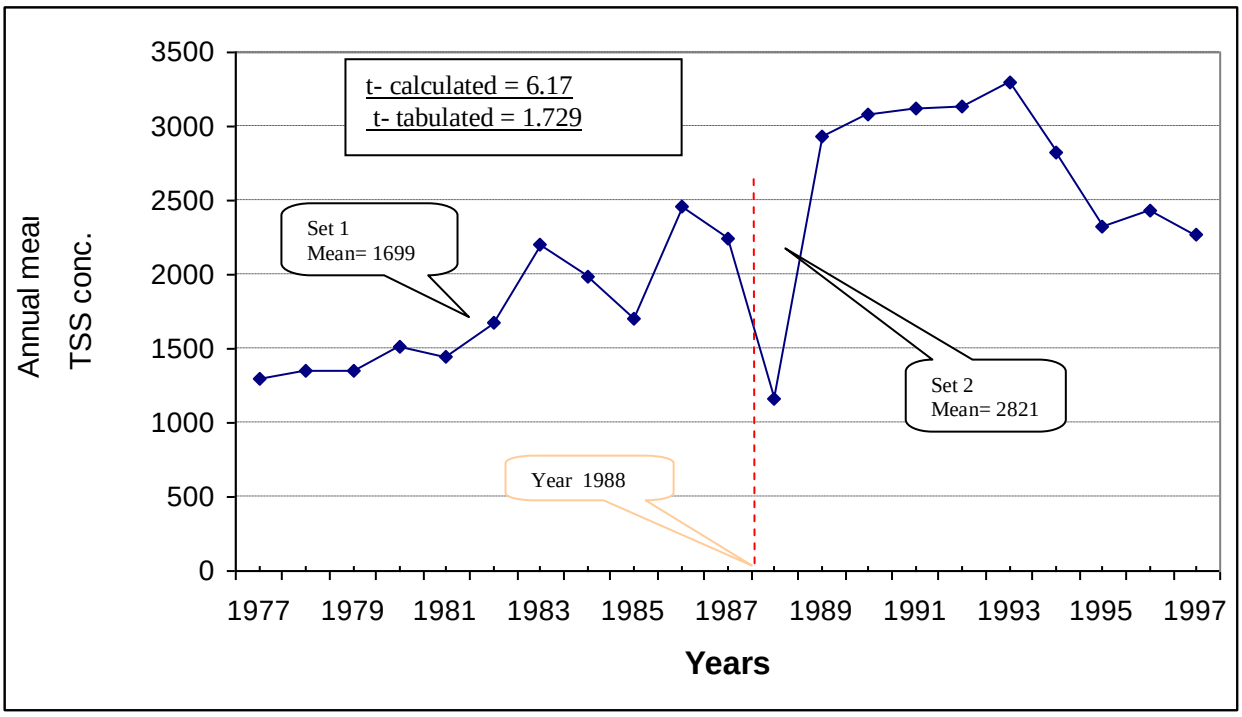

Fig. 1. Spilt Sample Test of the Original Historical Data.

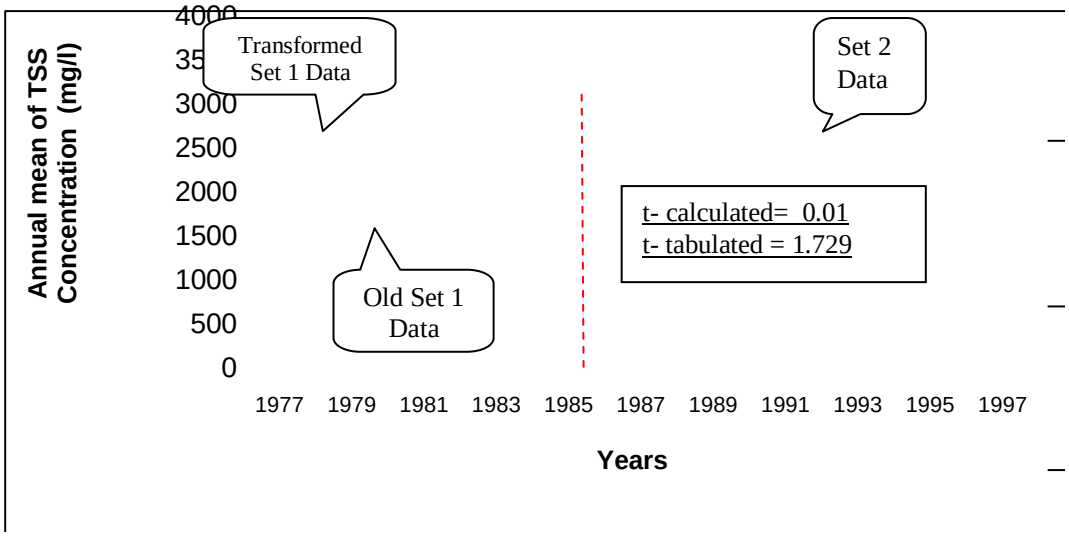

Fig. 2. Split-Sample Test for the Historical Data After Non-Homogeneity Removal.

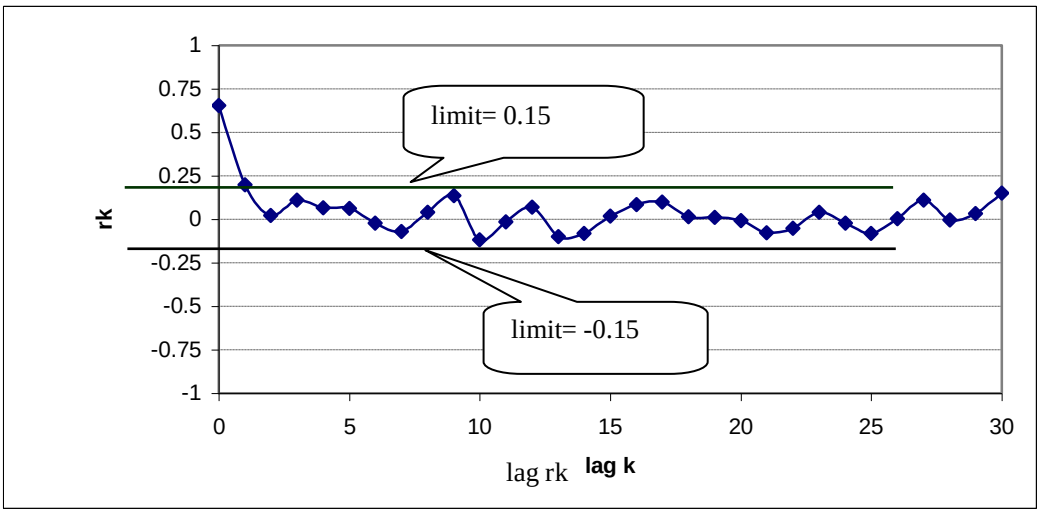

Fig. 3. a: **Corrologram of the Dependent Stochastic Component ($\varepsilon_{p,t}$), r_k is Correlation Coefficient, lag r_k is the Time of Creeping of Correlation Coefficient Values .**

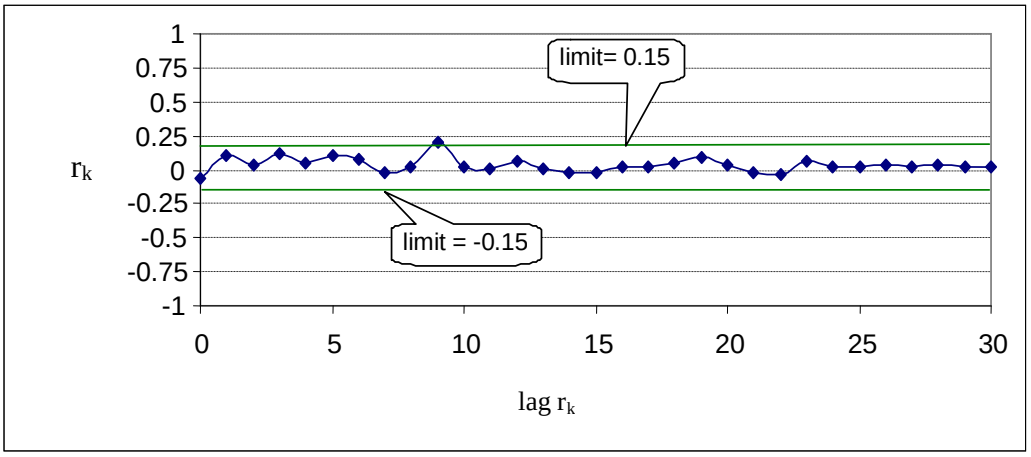

Fig. 3. b: Corrologram of the Independent Stochastic Component ($\xi_{p,t}$) Obtained Using First Order Autoregressive Model, r_k is Correlation Coefficient, lag r_k is the Time of Creeping of Correlation Coefficient Values.

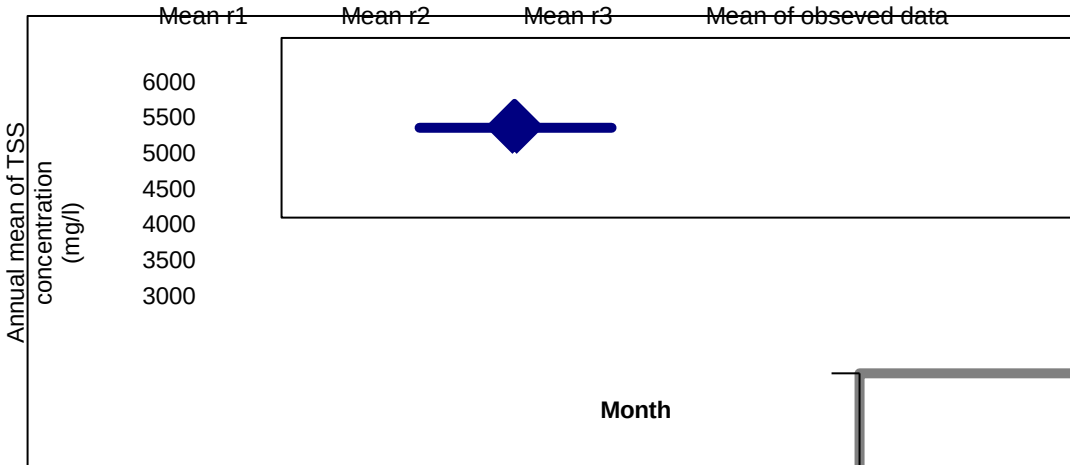

Fig. 4. Comparison of Observed Mean Monthly TSS Values, with Those of the Three Generated Series.

References

[1] Al-Suhaili R.H., (1985): "Stochastic Analysis of Daily Streamflow of Tigris River", M.Sc. Thesis, College of Engineering, University of Baghdad.

[2] Box, G.E.P., and, Jenkins, G.M., "Time Series Analysis,Forcasting and Control". Holdened Day, San Fransisco, 1976. P.575.

[3] Kadri Yurkel and Ahmet Kurunc.,(2006)" Performances of Stochastic Approaches in Generating Low Streamflow Data for Drought Analysis". Journal of Spatial Hydrology Vol. 5 no. 1. Internet.

[4] Garry L. Grabow, Dan Line, Jean Spooner and Laura Lombardo.," How to Detect a Water Quality Change using SAS an Example"PDF, Internet.

[5] Howard S.Peavy, Donald R.Rowe and Georage Techobanoglous., "Environmental Engineering", McGrew-Hill Book Co.,1986.

[6] Nemerow,N.L.,"Scientific Stream Pollution Analysis",McGrew-Hill Book Co.,1974.

[7] Yevjevich,W.,M.,"Structural Analysis of Hydrology time Series", Hydrology Paper No.56,Fort Collins, Colorado, Nov.,1972.

[8] Zhiyong Hung and Hiroshi Morimoto, (2006)" Water Pollution Models based on Stochastic Differential Equations". Department of Earth and Environmental Sciences Graduate School of Environmental Studies, Nagoya University, PDF, Internet.

[9] وزارة الموارد المائية ، قسم المدلولات المائية.، " التحاليل الكيمياوية لنهر الفرات في الناصرية"، تقارير دورية.

تحليل البيانات الشهرية لتراكيز المواد العالقة الكلية لنهر الفرات في مدينة الناصرية

أ.د. رافع هاشم السهيلي م.م. طارق جواد كاظم
قسم الهندسة البيئية/ *كلية الهندسة/ جامعة بغداد*

الخلاصة

تم في هذا البحث تحليل البيانات الشهرية لتراكيز المواد العالقة الكلية لنهر الفرات في مدينة الناصرية، حيث تم اخذ المعدل للشهري للبيانات المتوفرة من سنة (1977-2000) .

في البداية تم اختبار البيانات لمعرفة فيما اذا كانت متجانسة او غيرمتجاسة ووجد انها غير متجانسة عند سنة 1988. باستخدام طريقة Yevichevich تم ازالة عدم التجانس في البيانات. ثم تم توزيع البيانات طبيعياًباستخدام طريقة العالمين Box و Cox .

بعد ذلك تم اخذ القيم المعدلة لازالة المركبة الدورية عنها وذلك للحصول على الدالة المستقلة التي تم نمذجتها بموديل من نوع سلسلة Markovian.

ان التحليل اعلاه تم تطبيقه على البيانات من سنة 1977-1997 حيث تركت قيم التراكيز للثلاث سنين المتبقية (1998-2000) لاستخدامها في التحقق من نتائج الموديل.

ومن خلال النموذج الذي تم التوصل اليه تم توليد قيم لثلاث سنين مستقبلية لمقارنتها مع القيم المقاسة وباستخدام اختبار t-test وقد بينت النتائج امكانية الاعتماد على النموذج لاعطاء نتائج مستقبلية مقبولة.