International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 12, Issue 1 [January. 2023] PP: 355-361

Application of Autoregressive Integrated Moving AverageModel for Stream Flows

Dr. NaciBüyükkaracığan
SBMYO,Selcuk University, Konya, TURKEY

ABSTRACT:*The effective use of water resources is very important for the continuity of human life. Rapid population growth, unplanned urbanization and technological developments negatively affect the availability of water resources. In this context, it is necessary to model water resources under global climate change conditions and to produce forecasts for the future. Estimating and modeling river flows is of vital importance, especially in Turkey, where flood disasters are frequently encountered, as well as scarce water resources. Autoregressive Integrated Moving Average (ARIMA) model is an accepted method for its accuracy and efficiency in hydrological time series analysis. In this study, modeling and estimation of the annual flows used in the design and operation of hydraulic structures were carried out by using the stochastic structure. For this purpose, data from a flow observation station on Kızılırmak, the longest river in Turkey, were used. Before defining the model, preliminary studies such as data preparation, normality and stationarity tests were carried out, and parameter estimation was carried out according to model suitability criteria. It was decided that the ARIMA(1,1,1) model is the optimum model using the Akaike Information Criteria (AIC). At the end of the study, it was seen that the ARIMA(1,1,1) model, which is not used in stable synthetic hydrological series production, can be used in annual flow estimates.*

Keywords: *Akaike information criterion, autoregressive model, ARIMA model, stochastic model, synthetic hydrological series*

Date of Submission: 18-01-2023 Date of acceptance: 03-02-2023

I. INTRODUCTION

Time series analysis is a numerical method used in modeling and forecasting monthly or annual precipitation. A time series is a dataset. This series consists of sequential data points measured at successive times. In cases where its elements occur as intrinsically dependent, this series is called stochastic process. Time series analysis is used to fit the series to an appropriate model to reveal meaningful statistics of the series and other characteristics of the data. For this purpose, it is necessary to estimate the model parameters. Thus, it is possible to predict the future of the series and to determine how the observation series may continue in the future [1-2].

Flow information of the studied stream is needed in the design of water structures. Researchers produce synthetic stream series using simulation methods when available records are incomplete or insufficient. Simulation is the mathematical expression of the behavior of a water supply system over a certain period of time and can be used to calculate daily, monthly or seasonal flows, to determine the flow rate of a hydrograph, or to complete missing values in flow records [3].

Stochastic models, which are widely used in hydrology for the modeling of time series, were first created by Box and Jenkins. For this reason, the models are also included in the literature as Box-Jenkins Models [4]. It is one of the methods that can be used in modelling. The most important parameter of these models is the autocorrelation coefficient, which shows the dependence between observations [5]. ARIMA has become very popular for modeling flow and precipitation data due to its ease of development and implementation [6]. ARIMA can be a powerful model for forecasting evapotranspiration in hydrometeorology, irrigation water requirement and Rainfall Forecasting [7-10].

In the literature, there are many models of stochastic processes. Kahya et al. (1998) applied ARIMA models for annual average flows measured at 4 flow observation stations in the Yeşilırmak basin [11]. Huang et al. (2004) compared artificial neural networks (ARIMA) models to estimate the flow values of the Apalachicola river in Florida, USA. They found that ARIMA models were more successful than others [12]. Al-Aboodi et al. (2017) established estimation models with ARIMA, ANN and ANFIS, the monthly average flow rates of the Euphrates river passing through the city of Thi-Qar in southeastern Iraq [13]. Altunkaynak and Başakın (2018) tried to estimate the daily flows of Colombia river in America with ANN, ARIMA and ANFIS [14].Kurak(2013) established ARIMA models of monthly groundwater levels recorded in 2 different wells in Izmir. As a result of the research, the random walk model of first order differentiated and fully standardized

water levels in the well found ARIMA(0,1,0) fit [15].Fashae et al. (2019) worked to compare artificial neural network (ANN) and ARIMA to model the Opeki River discharge [16]. According to the results, ARIMA outperformed ANN. Kır (2020) made the estimation of monthly and annual precipitation in Antalya with seasonal ARIMA models [17].

II. DATA and METHODOLOGY

Within the scope of this study, monthly average flow data of the Yamula Flow Observation Station (AGI) numbered E15A001 in the Yamula Sub-basin located within the boundaries of the Kızılırmak basin shown in Fig 1. In this study, the annual average flow data of the station between 1980 and 2015 were used for modelling.

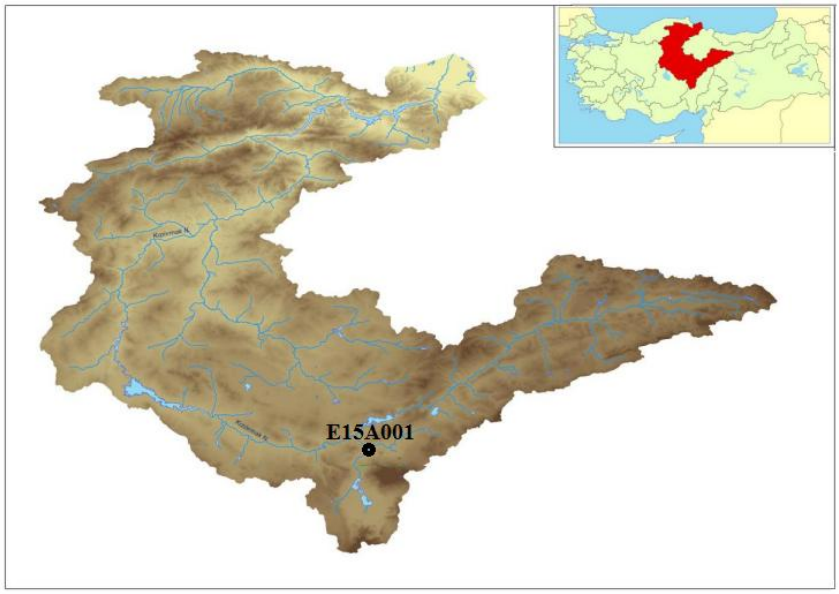

Fig.1 Kızılırmak Basin and Yamula Flow Observation Station[20].

ARIMA Model

In a stable ARIMA(p,d,q) model, p; autoregressive component, q; moving average component and d; Represents the number of differencing operations. The difference (d) of the annual hydrological series x_t(t=1, ..., n) is calculated by the following equation:

$$u_t = \sum_{j=1}^{p} \varphi_j . u_{t\ j} + \varepsilon_t \quad \sum_{j=1}^{q} \theta_j . \varepsilon_{t\ j} \quad (1)$$

The purpose of differentiation is to eliminate instabilities in the series. The ARMA(1,1) model applied to ut is stable with the condition $|\phi_1|<1$. In a normalized series, if there is an instability in the level and the slope of the trend of the series elements, $w_t=u_t-u_{t-1}=(x_t-x_{t-1})-(x_{t-1}-x_{t-2})=x_t-2.x_{t-1}+x_{t-}2$. The difference is taken as 2 consecutive times.

The undifferentiated xt series can be written as follows, considering xt=ut+xt-1 and ut-1=xt-1-xt-2 for d=1.

$$x_t=x_{t-1}+\phi_1.(x_{t-1}-x_{t-2})+.....+\phi_p.(x_{t-p}-x_{t-p-1})+\varepsilon_t-\theta_1.\varepsilon_{t-1}-.....-\theta_q.\varepsilon_{t-q} \quad (2)$$

It consists of the integration of the x_t series and the ut series. This stochastic process, consisting of an infinite sum of u_t's, is therefore referred to as the ARIMA model.

Annual ARIMA Modeling Process Steps

Preliminary Analysis

Step (1). The conformity of the original (historical) x_t time series to the normal distribution should be checked with the Skewness Test.

The skewness coefficient (γ) is calculated with the following equation.

$$\gamma = \frac{N\sum_{t=1}^{N}(x_t - \bar{x})^3}{(N-1)(N-2)\left[\frac{1}{(N-1)}\sum_{t=1}^{N}(x_t - \bar{x})^2\right]^{3/2}} \quad (3)$$

wherex_t is the serial elements, the sample mean and N is the total number of elements. The fact that the skewness coefficient remains within the following limits indicates that the series is normally distributed.

$$\left\{-u_{1-\alpha/2}\sqrt{\frac{6}{N}} \;\; ; \;\; u_{1-\alpha/2}\sqrt{\frac{6}{N}}\right\} \quad (4)$$

$\alpha$is the chosen significance level, and an evaluation is made for γ at confidence limits (1-α). "$u_{1-\alpha/2}$" is the standard normal variable with a probability value of "1-α/2". Formula 5 is generally accurate enough for samples with N>150. For smaller samples, the "γ"is compared with the N>150 values in the Table given by Snedecor and Cohran [18]. If $\gamma<\gamma_\alpha(N)$, the series is considered normally distributed.

If it is concluded in the tests that the historical series is normally distributed, the steps are continued with Step (1c), if it does not comply with the normal distribution, the steps are continued with Step (1b). The time series of the two-parameter lognormal probability distribution function represents the frequency distribution very well. Accordingly, if it is assumed that the time series conforms to the lognormal-2 distribution, the probability density function of the series can be written as follows.

$$f(x) = \frac{1}{x.\sigma_y\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\frac{\log(x)-\mu_y}{\sigma_y}\right]^2\right\} \quad (6)$$

The mean of the values obtained with the y=log(x) function is μ_y and the standard deviation is σ_y. Thus, by applying the y=log(x) transformation to the non-normally distributed series, the standard normal series with the mean "0" and the standard deviation "1" is obtained with the (7).

$$y = \frac{\log(x)-\mu_y}{\sigma_y} \quad (7)$$

The conformity of the values found with the y=log(x) transformation to the normal distribution is checked. If the distribution is normal, the next step is taken, if not, other transformations are used.

Step (1c). For the elimination of low-frequency components, the series "d." difference is taken. In case of indecision in level and trend, the difference is taken twice in succession.

Step (1d). In this step, the series is plotted and only the low order models get an idea.

Step (1e). The $r_k$autocorrelation coefficients of the series are calculated with the following formula.

$$r_k = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x}).(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \quad (8)$$

The "r_k" values found are checked with Anderson limit values. The fact that "r_k" is outside the 95% confidence limits indicates that the consecutive values in the series are interdependent. After the "r_k" values of the series, the partial autocorrelation coefficients (ϕ_1, ϕ_2, ...,ϕ_L) for the L≅0.3N element can be found with the help of the Durbin Formulas given below.

$$\phi_1(1) = r_1, \quad \phi_2(2) = \frac{r_2 - r_1^2}{1 - r_1^2} \quad , \phi_2(1) = \frac{r_1.(1-r_2)}{1-r_1^2} \quad (9a)$$

$$\phi_{k+1}(k+1) = \frac{r_{k+1} - \sum_{j=1}^{k}\phi_k(j).r_{k+1-j}}{1 - \sum_{j=1}^{k}\phi_k(j).r_j} \quad (9b)$$

$$\phi_{k+1}(j) = \phi_k(j) - \phi_{k+1}(k+1).\phi_k(k-j+1) \quad (9c)$$

Here, partial autocorrelation coefficients are values with $\phi_{k+1}(k+1)$ notation. When the subscript and the number in parentheses are the same, values to be used as partial autocorrelation coefficients are obtained.
Step (1f). A preliminary evaluation is made for the model to be selected as a result of the calculations. If all autocorrelation values after any "k" delay degree in the correlogram remain within the confidence limits, this indicates a moving average process from the "q" degree and usually q=1 is taken [19]. Thus, a preselection is made for the "p" and "q" degrees of the ARMA(p,q) model.

Parameter estimation
Step (2a). The mean " $\bar{y}$ " and variance "σ^2" of the series are calculated with the following equations.

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \quad , \quad \sigma^2 = \frac{1}{N-1}\sum_{t=1}^{N}(y_t - \bar{y})^2 \quad (10)$$

Step (2b). The z_t series is obtained with the following equation.
$z_t = y_t - \bar{y}$, , t = 1, 2, ..., N(11)

Step (2c). For the calculation of the "ϕ" and "θ" parameters of the selected model, both values that make the sum of the squares of the series minimum must be found.

$$S = \sum_{t=1}^{N} \varepsilon_t^{\,2} \quad (12)$$

The [ϕ,θ] pair, which gives the minimum value from the calculated "S" values of the [ϕ,θ] parameters in various combinations, is considered as the exact parameters.
Here, ε_t residual series is calculated as follows for ARMA(p,q) models and the first ε_t value is taken as zero until the largest "p" or "q" value.

$$\varepsilon_t = z_t - \sum_{i=1}^{p} \phi_i . z_{t-i} + \sum_{i=1}^{q} \theta_i . \varepsilon_{t-i} \quad (13)$$

The stability conditions of the first and second order model parameters can be checked practically by the following expressions.

Goodness of Fit test of Model
Step (3a). The internal dependence of the residual series ε_t is checked with the Port Monteau Test using the "Q" statistic calculated by the equation below.

$$Q = N\sum_{k=1}^{L} r_k^{\,2}(\varepsilon_t) \quad (14)$$

Step (3b). The suitability of the grade of the selected model is decided by comparing the model with models with an upper and a lower order using the Akaike Information Criteria (AIC). The AIC value is calculated as follows.

$$AIC(p,q) = N.\ln(\sigma_\varepsilon^2) + 2.(p+q) \quad (15)$$

Here $\sigma_\varepsilon^2 = S/N$ is found by the equation. The model that gives the minimum AIC value is selected as the best model. If the AIC value of the predicted model is significantly greater than the AIC value of the compared models, the modeling process is done from the beginning by changing the model degree.

III. RESULTS AND DISCUSSION

Step (1a). Since γ=-0.217 found by equation (4) was|-0.128|<0.755 at α=0.02 significance level, the series was considered normally distributed and proceeds to Step (1c).
Step (1c). Once the difference was sufficient to eliminate the low-frequency components in the series.
Step (1d). Annual flow rates and graphs of differential values are given in Fig. 2 and Fig. 3. Since a value above (below) the mean was followed by a value above (below) the mean, there wasa positive time dependence in the series (Fig. 2).

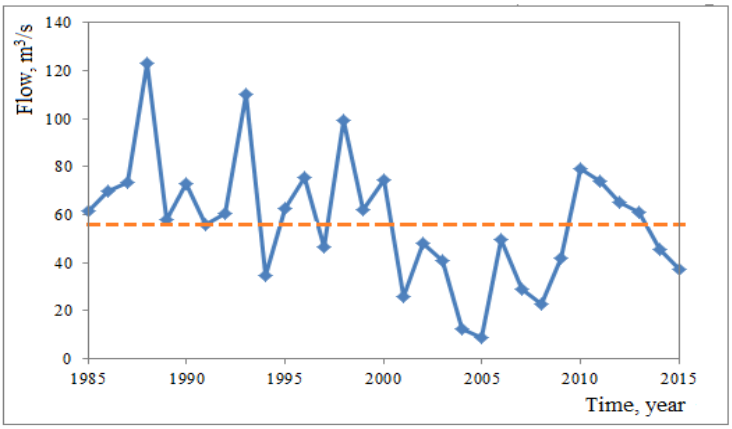

Fig.2 Time series of annual average flow rates

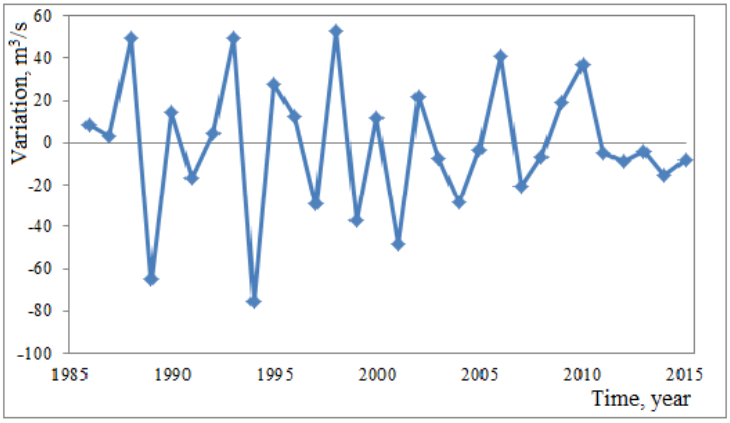

Fig.3 Time series of values taken with one-time subtraction

Step (1e). The rk (k=1,2,...,18) values and 95% confidence limits were determined by the equations (8) and (9), and the correlogram was drawn (Fig. 4).

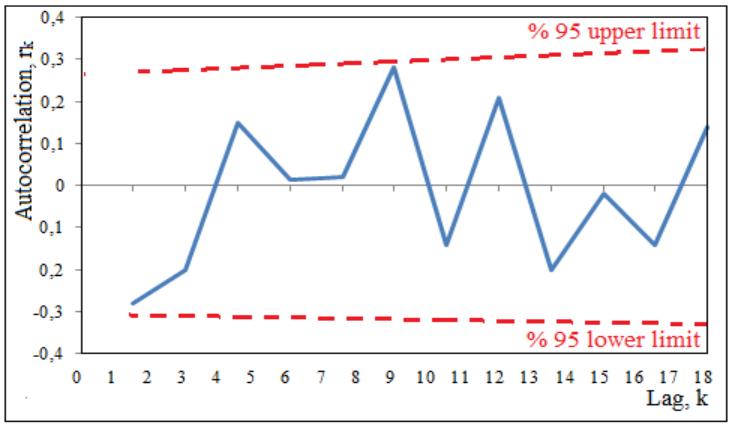

Fig. 4 Correlogram of subtracted values and 95% confidence limits

the$\phi_k(k)$ coefficients and 95% confidence limits of the series were calculated with (10) and (11). The partial correlogram exceeds the lower confidence limit at k=9 lag (Fig. 5). At the chosen significance level α=0.05, this wasacceptable (0.05x18≅1).

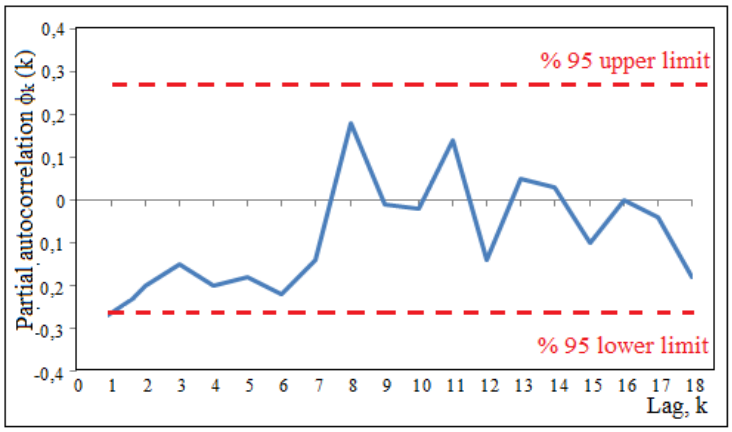

Fig.5 Partial correlogram of subtracted values and 95% confidence limits

Step (1f). Since the correlograms were significant at the k=1 lag value with 95% confidence, the ARMA(1,1) model was applied as a preselection to the series whose difference was taken.

Parameter estimation

Step (2a). The mean and variance of the series whose difference was taken were calculated $\bar{y}$ =-0.273 and σ^2=79.481 with (12).
Step (2b). The zt series was found by equation (13) as follows.
z_1 = (1.089) - (-0.273) = 1.363, z_2 = (-2.049) - (-0.273) = -1.776, …, z_{54} = (6.037) - (-0.273) = 6.310
Step (2c). The parameters of the ARMA(1,1) model were ϕ_1=0.3228 and θ_1=0.9028, which make the equation (12) minimum, and the conditions in (13) were met (−1<0.3228<1 ve −1<0.9028<1). Accordingly, the sum of the squares of the residual series was S=2924.14 and its variance was σ_ε^2=57.88.

Goodness of Fit test

Step (3a). In order to control the internal dependence of the ε_t residual series, the autocorrelation coefficients of the ε_t values (with the largest lag value L=0.15N) were found by the equation (8) and the "Q" statistic in (16) was calculated with the Port Monteau test.

$$Q = N\sum_{k=1}^{L} r_k^{\,2}(\varepsilon_t) = 54.\sum_{k=1}^{8} r_k^{\,2}(\varepsilon_t)$$

Q = 1.7895 < 12.34 = $x^2_{0.95}$so, the series are now independent.
Since γ=0.399 found by equation (4) was 0.482<0.761 at α=0.02 significance level, so the series were normally distributed.

Step (3b). Since the partial correlogram was significant at k=1 (Fig. 5), it was not considered appropriate to establish and compare the ARMA(0,1) model without considering the autoregressive component. The parameters of the ARMA(2,1) model compared with ARMA(1,1) were found to be ϕ_1=0.3228, ϕ_2=0.2228, θ_1=0.9028, S=2924.14 andσ_ε^2=58.86.
ARMA(1,1), AIC(1,1) = 54.ln(57.88) + 2.(1+1) = 218.3257
ARMA(2,1), AIC(2,1) = 54.ln(58.86) + 2.(1+1) = 220.0283

Accordingly, the ARMA(1,1) model (with the minimum AIC value) selected by preliminary evaluation from the correlograms was the most appropriate model. Thus, the ARMA(1,1) model of the ut values obtained by taking the difference of the annual average flows once was given below.

u_t= (0.3228).u_{t-1} + ε_t – (0.9028).
The ARIMA(1,1,1) model of the station could be written as follows with the necessary transformations.
x_t= (1.3228).x_{t-1} - (0.3228).x_{t-2} + ε_t – (0.9028).ε_{t-1}

IV. CONCLUSION

The ARIMA Model is widely used in time series analysis due to its simple applicability and considering the internal dependence of the data series.

In this study, the ARIMA(p,d,q) model of annual average flows measured at Yamula station on the Kızılırmak River was established. Since the skewness coefficients of the data were found and normally distributed, no transformation was performed. In the calculations, no significant level of instability was observed in the level and slope. Therefore, only one difference was taken to dispose of the low-frequency components in the series. The ARMA(1,1) model was established by making a preliminary evaluation from the correlogram of the u_t values.

Akaike Information Criteria (AIC) values were used between the selected model and models with an upper and lower autoregressive degree. The optimum model with the minimum AIC value was determined as the ARMA(1,1) model selected in the preliminary evaluation.

The ARIMA(1,1,1) model has not used in the production of stable synthetic hydrological series due to the xt series being unstable. However, it has shown that it was sufficient and reliable for the annual flow estimates of the station used in the study.

REFRENCES

[1]. Salas, J. D. [1980] "Applied modeling of hydrologic time series" Water Resources Publication.

[2]. Bayazit, M. [1980] "Statistical methods in hydrology" Istanbul Technical University, Turkey (in Turkish).

[3]. Viessman, W.,Lewis, G.L., Knapp, J.W. [1989] "Introductiontohydrology" Harper&Row, Singapur.

[4]. Hipel, K. W. andMcLeod, A.I. [1981] "Time Series modellingforwaterresourcesandenvironmentalengineers" TheNetherlands:Elsevier.

[5]. Bartlett, M. S. [1946] "On thetheoreticalspecification of sampling properties of autocorrelated time series", J. Roy, Statistical Soc., vol. 8:pp. 27-41.

[6]. Somvanshi, V. K., Pandey, O. P., Agrawal, P. K., Kalanker, N. V., Prakash, M. R., & Chand, R. [1981] "Modeling and prediction of rainfall using artificial neural network and ARIMA techniques" J. Ind. Geophys. Union, 10(2): pp.141-151.

[7]. MahdizadehKhasraghi M, GholamiSefidkouhi MA, Valipour M. [2014] "Simulation of open- and closed-end border irrigation systems using SIRMOD" Arch. Agron. Soil Sci. DOI: 10.1080/03650340.2014.981163

[8]. Valipour M. [2014] "Application of new mass transfer formulae for computation of evapotranspiration" J. Appl. Water Eng. Res. 2(1): pp.33– 46.

[9]. Valipour, M. [2015] "A comprehensive study on irrigation management in Asia and Oceania" Archives of Agronomy and Soil Science, *61*(9): pp. 1247-1271

[10]. Dayal, D., Swain, S., Gautam, A. K., Palmate, S. S., Pandey, A., & Mishra, S. K. [2019] "Development of ARIMA model for monthly rainfall forecasting over an Indian River Basin" In World environmental and water resources congress 2019: Watershed management, irrigation and drainage, and water resources planning and management, pp. 264-271.

[11]. Kâhya, E., Karabörk, Ç.,Kalaycı, S., [1998]"Yeşilırmakhavzasindaarimaveçokdeğişkenlistokastikmodellemeuygulamaları" II. National HyrometeolojiSysmposium: pp.195-203.

[12]. Huang W., Xu B. and Chan-Hilton A. [2004] "Forecasting flows in Apalachicola River using neural networks" Hydrological Processes, 18: pp. 2545–2564.

[13]. Al-Aboodi A. H., Dakheel A. A. and Ibrahim H. T. [2017] "Comparison of data-driven modelling techniques for predicting river flow in an arid region" International Journal of Applied Engineering Research, Vol. 12, No.11: pp. 2647-2655.

[14]. Altunkaynak, A., &Başakin, E. [2018] "Zaman serilerikullanılaraknehirakımtahminivefarklıyöntemlerlekarşılaştırılması" Erzincan University Journal of Science and Technology, 11(1): pp.92-101.

[15]. Kurak, M., [2013] "İzmir'e Su sağlayankuyulardayeraltisuyuseviyesalinimlarininstokastikanalizi" DokuzEylülÜniversitesi, Fen BilimleriEnstitüsü, İnşaatMühendisliğiAnabilim Dalı, MS Thesis, pp. 5-11.

[16]. Fashae, O. A., Olusola, A. O., Ndubuisi, I., &Udomboso, C. G. [2019] "Comparing ANN and ARIMA model in predicting the discharge of River Opeki from 2010 to 2020" River research and applications, Vol35(2): pp.169-177.

[17]. Kır, E. G., &Güldal, V. [2020] "Antalya iliaylikveyillikyağişlarinin zaman serisimodellemesi" Su Kaynakları, Vol.5(2): 9-15.

[18]. Karabörk, M. Ç., &Kahya, E. [1999] "Multivariate stochastic modeling of monthly streamflow of rivers in the Sakarya basin. Turkish Journal of Engineering and Environmental Sciences, Vol.23(2): pp. 133-148.

[19]. Salas, J. D. [1980] "Applied modeling of hydrologic time series" Water Resources Publication.

[20]. SYGM, (2022). Date of Access:14.11.2022 https://www.tarimorman.gov.tr/SYGM/Belgeler/Ta%C5%9Fk%C4%B1n%20Y%C3%B6netim%20Planlar%C4%B1/KIZILIRMAK%20HAVZASI%20TA%C5%9EKIN%20YONETIM%20PLANI%20Y%C3%96NET%C4%B0C%C4%B0%20%C3%96ZET%C4%B0.pdf