J. Agr. Sci. Tech. (2010) Vol. 12: 47-60

A Comparison of the Efficiency of Parameter Estimation Methods in the Context of Streamflow Forecasting

L. Parviz[1], M. Kholghi[1]* and A. Hoorfar[1]

ABSTRACT

The forecasting of hydrological variables, such as streamflow, plays an important role in water resource planning and management. Recently, the development of stochastic models is regarded as a major step for this purpose. Streamflow forecasting using the ARIMA model can be conducted when unknown parameters are estimated correctly because parameter estimation is one of the crucial steps in modeling process. The main objective of this research is to explore the performance of parameter estimation methods in the ARIMA model. In this study, four parameter estimation methods have been used: (i) autocorrelation function based on model parameters; (ii) conditional likelihood; (iii) unconditional likelihood; and (iv) genetic algorithm. Streamflow data of Ouromieh River basin situated in Northwest Iran has been selected as a case study for this research. The results of these four parameter estimation methods have been compared using RMSE, RME, SE, MAE and minimizing the sum squares of error. This research indicates that the genetic algorithm and unconditional likelihood methods are, respectively, more appropriate in comparison with other methods but, due to the complexity of the model, genetic algorithm has high convergence to a global optimum.

Keywords*:* ARIMA model, Conditional likelihood, Forecasting, Genetic algorithm, Parameter estimation.

INTRODUCTION

Effective planning, management, and control of water resources systems require considerable data on numerous hydrological variables such as streamflow, rainfall and temperature. Invariably, the data sets are recorded in time and are referred to as time series. These series are analyzed using statistical methods to evaluate the parameter of interest so as to arrive at a suitable decision support system for management and control purposes [8]. Among several time series models, the ARIMA (Autoregressive Integrated Moving Average) model has been attractive to researchers for its power in streamflow forecasting.

Generally, in the stochastic modeling process the objective is to develop a simple model with the parsimony rule of the stochastic model. In order to achieve this objective, the model parameter estimation that is one of the modeling processes plays the main role for best fitting with observed data. This is because incorrect parameter estimation methods lead to bias and unacceptable forecasting.

Carlson *et al.* (1970) were the first researchers to analyses the time series of annual streamflow using ARIMA [1]. The basis and modeling procedure of classic ARIMA models are described by Box and Jenkins (1976), Davis and Brockwell (1978). Delleur *et al.* (1976) and Mcleod *et al.* (1977) have used PARIMA (Periodic Autoregressive Integrated Moving Average)

[1] Department of Irrigation and Reclamation Engineering, College of Soil and Water Engineering, Campus of Agriculture and Natural Resources, University of Tehran, Karaj, Islamic Republic of Iran.
* Corresponding author, e-mail: kholghi@ut.ac.ir

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

in the modeling of a streamflow management basin. An important extension of ARIMA models was introduced by Granger and Joyeux (1980), and by Hosking (1981) who proposed the FARIMA (Fractional Autoregressive Integrated Moving Average) model [3, 10]. Sharman and Breakenbridge (1994) reviewed the form of the likelihood function for ARMA signal models and then they described how a genetic algorithm may be employed to search the likelihood space with the aim of finding the maximum point. The use of parallel processing techniques to speed up the search procedures has been examined in this research [2]. Anderson *et al.* (1999) provided a parameter estimation technique that considers two types of periodic time series model, those with a finite fourth moment and models with finite variance but an infinite fourth moment. The results regarding the infinite fourth moment case are of particular interest [5]. Shin and Lee (1999) have established the consistency of the maximum likelihood estimators for the ARIMA model with time trends. General uniform approximations are established for the quadratic forms which appear in the Gaussian likelihood [7]. Lu and Chon (2000) introduced a new method for ARIMA parameter estimation. Theer algorithm was based on the GMDH (Group Method of Data Handling), first introduced by Ivakhnenko (1966 and 1971). Computer simulations show that in cases with noise contamination and incorrect model order assumptions, the GMDH usually performs better than either the FOS or the least-squares methods in providing only the parameters that are associated with the true model terms [6]. Valenzuela *et al.* (2003) have obtained an expert system based on paradigms of artificial intelligence, such as genetic algorithm, so that model can be identified automatically [12]. Wurtz *et al*. (2003) have used GARCH/APARCH errors for parameter estimation of ARIMA models and optimization (maximization) of the constrained log-likelihood function with the help of a SQP solver [13]. Chorng Shyong *et al.* (2004) have provided a genetic algorithms based model identification to overcome the problem of local optima which was suitable for any ARIMA model. The results show that the GA-based model identification method can present better solutions, and is suitable for any ARIMA models [14]. Jonsttir *et al.* (2006) used a parameter estimation method for stochastic rainfall-runoff model. The parameter estimation method was a maximum likelihood method where the maximum likelihood function is evaluated using the Kalman filter technique. The maximum likelihood method estimated the parameters in a prediction error setting; they also estimated the parameters by an output error method. The model performs well and parameter estimation methods are promising for future model development [16]. The point of maximum likelihood method in a failure domain yields the highest value of the probability density function in the failure domain. Obadage and Harnpornchai (2006) have proposed a genetic algorithm with an adoptive penalty scheme as a tool for the determination of the maximum likelihood point. The genetic algorithm can be used as a tool for increasing the computational efficiency in the element and system reliability analysis [17].

The main objective of this research is the comparison of four parameter estimation methods of the ARIMA model. Basically, the maximum likelihood function methods are used in this regard. In this study, this classic parameter estimation method is compared with a new optimization method like the genetic algorithm (*GA*). The streamflow time series have been used because streamflow forecasting has always been a challenging task for water resource engineers and managers and a major component of water resource planning and management.

MATERIALS AND METHODS

ARIMA is the method first introduced by Box–Jenkins to analyze stationary time

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

series data, and has since been used in various fields [14]. The generalized form of ARIMA can be described as:

$$\Phi(B^s)\phi(B)(1-B^s)^D(1-B)^d Z_t = \Theta(B^s)\theta(B)\varepsilon_t \quad (1$$

$$B^n Z_t = Z_{t-n} \quad (2$$

$$\Phi(B) = 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps} \quad (3$$

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \quad (4$$

$$\Theta(B) = 1 - \Theta_1 B^s - \cdots - \Theta_Q B^{Qs} \quad (5$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \quad (6$$

where Z_t is a discrete time observation process, ε_t is random series with mean zero and variance σ_ε^2, B denotes the backward shift operator, d and D denotes the non-seasonal and seasonal order of differences taken respectively (ARIMA models can be fitted to stationary hydrological series. For the transformation of non stationary into stationary series, the nonstationarity was removed by alternative methods such as differencing of theoriginal series). Φ (B), θ (B), Φ (B) and Θ (B) are polynomials in B and B^s of finite order p and q, P and Q, respectively, and usually abbreviated as SARIMA $(p, d, q)(P, D, Q)_s$.

When there is no seasonal effect, a SARIMA (Seasonal Autoregressive Integrated Moving Average) model reduces to pure ARIMA (p, d, q), and when the time series data set is stationary a pure ARIMA reduces to ARMA (p, q) [14].

The popularity of the ARIMA model is due to its flexibility, and the inclusion of both autoregressive and moving average terms. The ARIMA approach has several advantages over others such as a moving average, exponential smoothing and, in particular, its forecasting capability and its richer information on time related changes. It can also handle serial correlation among observations, which is found in most time series. It also provides systematic searching in each stage (of 3 stages) for an appropriate model. Other aspects of this model are complexity and requiring a great deal of experience [15].

The different methods are for parameter estimation of the ARIMA model some of which are described briefly below, namely the autocorrelation function formula based on model parameters, conditional likelihood, unconditional likelihood and genetic algorithm (GA).

Autocorrelation Function Formula Based on Model Parameters

In this method using autocovariance and autocorrelation function (ACF) after the identification of the model, some equations can be achieved. Using these equations, the parameters of model can be found. For example, using Equation (7-9) shows the relationship between ACF (known) and the parameters (unknown) of the ARIMA (1,1) model and solving them, the parameters of

$$\rho_0 = 1 \quad (7$$

$$\rho_1 = \frac{(\phi_1 - \theta_1)(1 - \phi_1\theta_1)}{1 + \theta_1^{\,2} - 2\phi_1\theta_1} \quad (8$$

$$\rho_k = \phi_1 \rho_{k-1} \qquad k \geq 2 \quad (9$$

the model are estimated [4].

Maximum Likelihood Estimation Method (*MLE*)

Maximum likelihood estimation begins with writing a mathematical expression known as a likelihood function of the sample data. Loosely speaking, the likelihood of a set of data is the probability of obtaining the particular set of data the chosen probability distribution model (the likelihood principle expresses the notion that the whole information that data have about parameters has been hidden in the likelihood function). The idea behind maximum parameter estimation is to determine the parameters (unknown) that maximize the probability (likelihood) of sample data. From a stated

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

point of view, the method of maximum parameter estimation is considered to be more robust and to yield estimates with good strand properties.

If the number of random samples from a society is taken into account, the value of multiplying the probability density function for random quantities $x_1, x_2 ..., x_N$ is introduced as the likelihood function for these quantities.

$$L = \prod_{i=1}^{N} f(x_i, \alpha) \quad (10$$

Where α is unknown parameters, L is the likelihood function and $f(x,\alpha)$ the distribution density function. The normal distribution is the standard distribution in this regard. This method is based on maximizing the likelihood function based on the selective parameters. For ease of calculation, maximization is conducted on a function logarithm, because the maximizing function logarithm is equal to the maximizing likelihood function. This method can be used from the error frequency function which follows the normal distribution with mean zero and variance σ_ε^2. The errors probability density function is obtained from Equation (11) and, after applying the likelihood function [Equation (12)] and then logarithm operation, it will result in relationship 13.

$$f(\varepsilon_t) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{\frac{-(\varepsilon_t - \bar{\varepsilon})^2}{2\sigma_\varepsilon^2}} = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{\frac{-\varepsilon_t^2}{2\sigma_\varepsilon^2}} \quad (11$$

$$L = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{\frac{-\varepsilon_1^2}{2\sigma_\varepsilon^2}} \times ... \times \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} e^{\frac{-\varepsilon_N^2}{2\sigma_\varepsilon^2}}$$

$$= \frac{1}{(\sqrt{2\pi\sigma_\varepsilon^2})^N} e^{\frac{-\sum_{t=1}^{N} \varepsilon_t^2}{2\sigma_\varepsilon^2}} \quad (12$$

$$LL = -NLn(\sqrt{2\pi}\sigma_\varepsilon) - \frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^{N} \varepsilon_t^2 \quad (13$$

The first term of Equation (13) is constant and so maximization of log-likelihood led to the reduction of the sum squares of error. For minimizing the sum squares of error, the three methods conditional likelihood, unconditional likelihood and genetic algorithm were introduced in the next section.

Conditional Likelihood

In this method for calculating the sum squares of error, ε can be derived from the ARIMA model like Equation (14) (after transformation of nonstationary series into stationary ones).

$$\varepsilon_t = \tilde{W}_t - \phi_1 \tilde{W}_{t-1} - ... - \phi_p \tilde{W}_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

and

$$E[W_t] = \mu, \tilde{W}_t = W_t - \mu, W_t = \nabla^d Z_t \quad (14$$

Using equation 14 in the explicit form it is hard to calculate ε. One of the solutions is the determination p number of W and q number of ε. In this regard, the calculation of $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ conditional on initial values subsequently is necessary. For each parameter group that minimized the sum squares of errors, those parameter groups are chosen as selective parameters. In computing initial values, unconditional exception of W and ε can be used. Unconditional exception of errors is zero; where the unconditional exception of W is zero, then the initial values of W are equal to zero, otherwise the mean of series are used instead of each of the components W.

Unconditional Likelihood or Calculating of Non Conditional Sum Squares

This method used precise initial values of W and error. In this regard, two backward [Equation (15)] and forward [Equation (16)] equations have been used, respectively.

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

$$\phi(B)\tilde{W}_t = \theta(B)\varepsilon_{B_t} \quad (15$$

$$\phi(F)\tilde{W}_t = \theta(F)\varepsilon_{F_t} \quad (16$$

For calculation of the previous time series, the backward equation has been used and for the stationary characteristics of the AR model, W_t estimations for limit values of W in $t= -Q$ (Q is the time which W is about zero) is zero. Then, the forward equation is used for ε estimations on the basis of precise time series. For each parameter group that has lower sum squares, that parameters group is selected. It should be noted that using backward and forward equations is possible in respect to their expectations. For the reversion calculation, Equations (17 and 18) have been used [1]:

$$\left[\varepsilon_{-j}\middle|\phi,\theta,W\right]=0 \quad j = 0, 1, 2,\ldots \quad (17$$

$$\left[e_{-j}\middle|\theta,\phi,W\right]=0 \quad j>q-1,\ldots \quad (18$$

Genetic Algorithm (GA)

The most popular technique in evolutionary computation research has been the genetic algorithm that was introduced by Hollend (1975). Evolutionary computation techniques abstract these evolutionary principles into algorithms that may be used to search for the optimal solution to a problem.

In a typical evolutionary algorithm, a genetic representation scheme is chosen by the researcher to define the set of solution that forms the search space for the algorithm. Any individual solution in the space has a specific representation. A number of individual solutions are created to form an initial population. The following steps are then repeated iteratively until a solution has been found which satisfies a pre-determined termination criterion (achieving a stopping criterion such as the time limit, the number of generation). Each individual is evaluated using a fitness function that is specified to the problem being solved. Based upon their fitness values, a number of individuals are chosen to be parents (selection). New individuals or offspring are produced from those parents using crossover operators. The crossover operators act upon the information available in the representations of the parents to product new individual consistent with the representation scheme. These new individuals may be radically different from, slightly different from, or even the same as their parents. A crossover operation is conducted probably and with regard to crossover probability. The fitness values of the offspring are determined. To prevent optimized results converging with local optimums, a mutation operator has been used. Some of the chromosome (individual) genes after the crossover process have been randomly altered; mutation applies to genes, which form the chromosomes. In binary genetic algorithms, the gene which is selected for mutation is changed from 1 to 0 and vice versa. The function of the two last operators which imitate biological processes start producing the second generation. Finally, survivors are selected from the old population and the new offspring to form the new population of the next generation. The mechanisms determining which and how many parents to select, how many offspring to create, and which individuals will move into the next generation together represent a selection method. The key aspect distinguishing an evolutionary from a traditional search algorithm is that it is population-based. Rather than moving from one point in the search space to another during each phase of the search, as is done in iterative improvement algorithm, a population-based search moves from one set of points to another set of points. At any given time, the points in the set may be sampled from different areas of the search space. The operation process of GA has been represented in Figure 1 [12,14].

Genetic algorithms are particularly efficient in optimization problems, especially when the respective objective functions exhibit many local optima or discontinuous derivatives. In this research

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

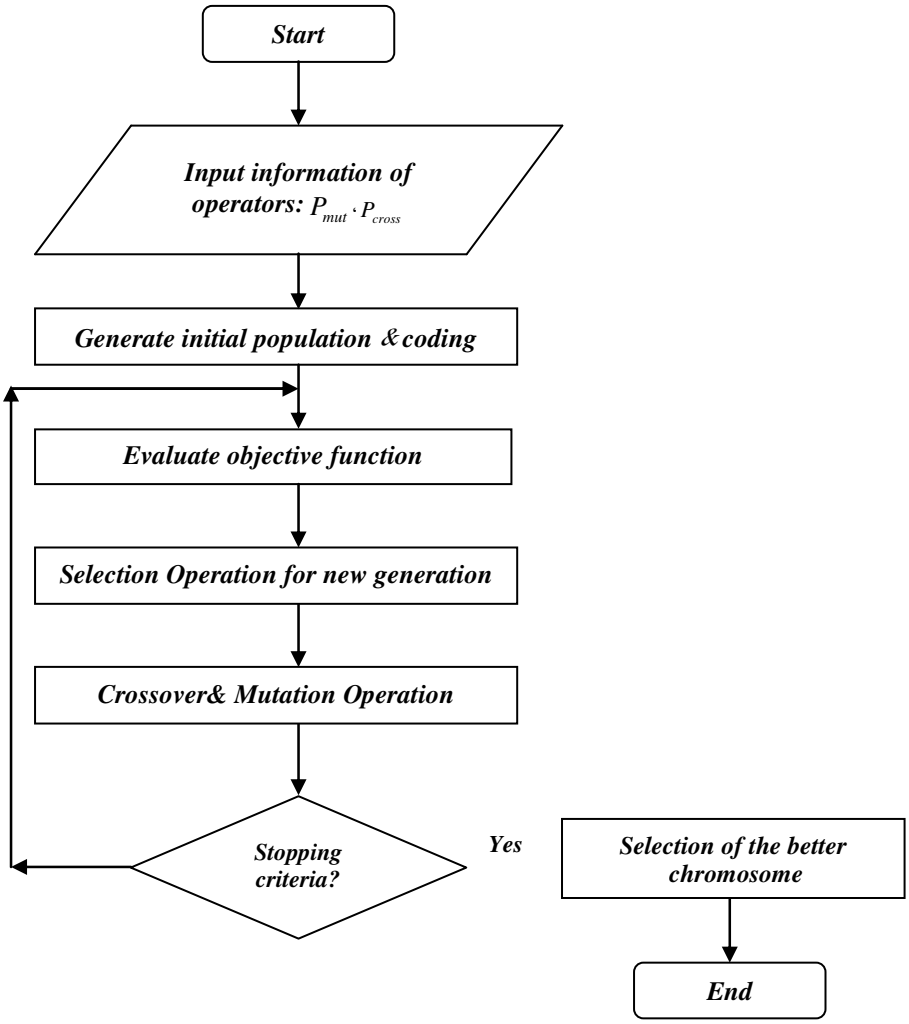

Figure 1. Representation of genetic algorithm (GA)
A process [11]

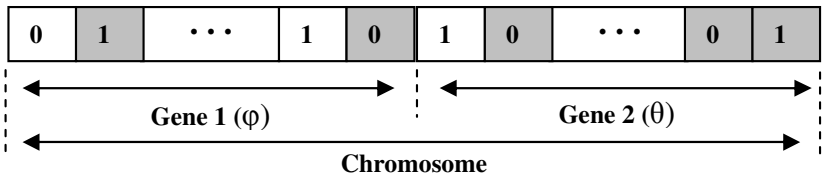

Figure 2. Element of each chromosome.

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

decision variables are the parameter of ARIMA model (φ, θ) and objective function is the basis of minimizing the sum squares of error. The criteria for getting optimum results are the number of generations.

Study Watershed

Ouromieh Lake watershed is one of the sixth major basins in Iran. It is located in the North West of Iran and covers an area of 51,866 Km^2. The coordinates of basin are between 35, 39 and 38, 30 N 44, 33 and 47, 53 E. The annual mean precipitation is variable from 203 to 688 mm and annual evaporation is 1,499 mm. The annual_mean temperature of Ouromieh station (with 1,313 m height) is 10.83.

In recent years according to a decrease in precipitation, drought threatens the Ouromieh Lake watershed. The most important problem of this watershed is related to the lack of observed streamflow data. River water is mostly used for irrigation in addition to drinking and fishing. Existence of important dams such as Shahid Madani on Ajichai River with 7,700 km^2 and Nahand on Nahandchai River, and environmental conditions like increasing the level of salinity and pollution by decreasing the flow of the rivers are the reasons for selection of this basin for this study. Stations chosen to be used in this research are Vanyar station on Ajichai River with 21 years' streamflow data from 1980 till 2000 and Nahand dam entrance station on Nahandchai River with 16 years' data from 1981 till 1996. In Figure 3, the hydrometrical station and Ouromieh Lake basin have been illustrated on a scale of 1:100,000.

RESULTS AND DISCUSION

When any type of stochastic model is being developed to model a given time series it is recommended to follow the identification, estimation and diagnostic check stages of model construction. Figure 4 shows this iterative process to create the ARIMA model, and this algorithm will be continued until validation of model is determined.

Using this modeling process, the first step is to check the normality of the time series. For checking the normality of the Vanyar and Nahant time series, a probability plot of streamflow was used. A probability plot

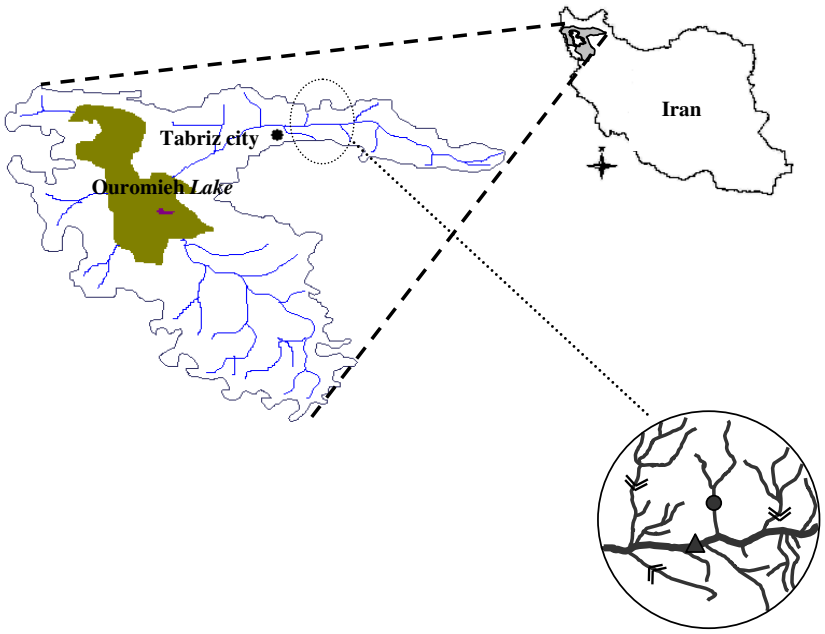

Figure 3. Situation of watershed and stations.

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

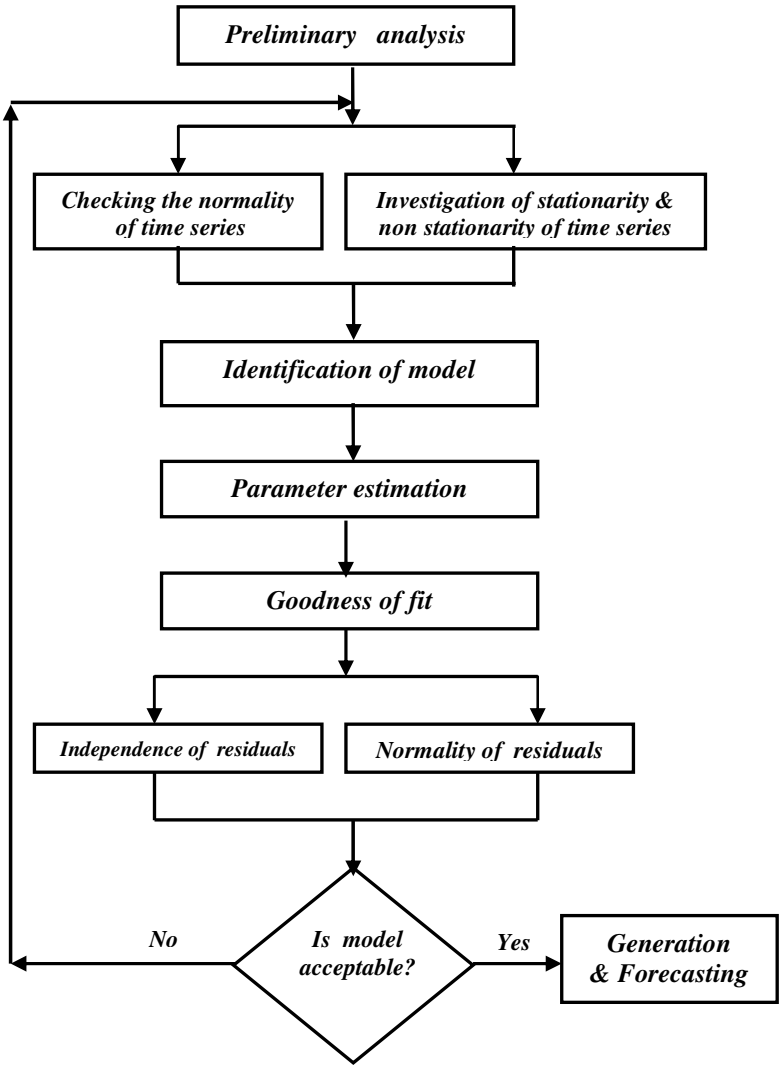

Figure 4. Iterative process of auto-regressive integrated moving average (ARIMA) modelling.

displays the percentiles (95% confidence). Using the probability plot it is possible to assess whether a particular distribution fits the time series. In general, the closer the points fall to the fitted line, the better the fit. Figures 5-7 indicated that the Vanyar and Nahand time series follow normal and log-normal distribution.

For an ARIMA (p, d, q) model, it is necessary to obtain the order of the model. The next step is to identify the order of differencing (d) needed to make the series stationary. In this regard two methods can be used. (1) From a plot of the normalized series we can observe whether there is any non stationary element in the level or both in the level and slope. The first case may indicate the need for first differencing, the second for differencing twice.

(2) Based upon the information given by the coefficients of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), the following steps can be defined to determine the stationarity of the time

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

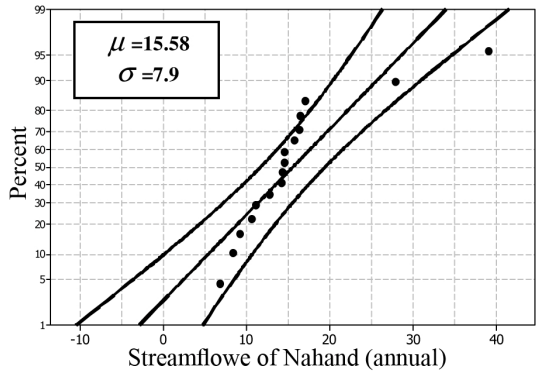

Figure 5. Probability plot of Nahand streamflow.

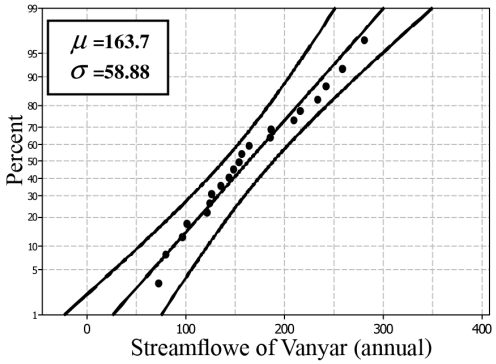

Figure 6. Probability plot of Vanyar streamflow.

series. (2-1) If the time series has many high positive autocorrelation coefficients, then it probably needs a higher order of differentiation. (2-2) Starting from an original time series with positive autocorrelation, if after differentiating the first autocorrelation coefficient is close to zero or negative, then the series does not need a higher order of differentiation. (2-3) The optimum order of differentiation is frequently the one in which the standard deviation of the series is smaller [18].

From the plotting of two normalized time series it is obvious that first differencing is adequate for a stationary series. The results

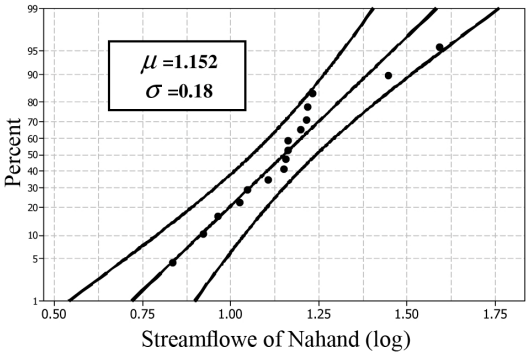

Figure 7. Probability plot of Nahand streamflow (log-normal)**.**

of the second method verified this matter and Tables 1 and 2 show the results of method 2.

After this step the Correlogram method, the sample PACF and the sample ACF are used in appropriate differenced series for identifying the orders *p* and *q* of the ARIMA (p, q) model. However this is complicated

$$AIC(p,q) = NLn(\sigma_\varepsilon^{\,2}) + 2(p+q) \qquad (19$$

$$SBC(p,q) = NLn(\sigma_\varepsilon^{\,2}) + (p+q)Ln(N) \quad (20$$

and not easily conducted (when the time series data sets have a mixed ARIMA effect, the plot cannot provide clear lags to identify. In addition, the lags of a mixed ARIMA model usually involve subjective judgment which makes the results unstable). Various minimizing AIC, SBC. Since SBC had been proved to be strongly consistent, it determines the true model asymptotically, and is preferred to AIC for comparing different models [14].

In the case of Vaniar station data, this test for the ARIMA (1, 1, 1) model has had the

Table 1. Variation of auto-correlation function (ACF) with difference (d) of 1, 2 and 3.

Station	ρ1(d= 1)	P1(d= 2)	ρ1(d= 3)
Nahand (Differenced series)	-0.431	-0.751	-0.77
Vanyar (Differenced series)	-0.182	-0.675	-0.761

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

Table 2. Variation of standard deviation with difference (d) of 1, 2 and 3.

Station	d= 1	d= 2	d= 3
Nahand	0.3	0.56	1.08
Vanyar	58.87	172.2	392.14

lowest quantity among models with various orders (AIC= 214.46, SBC= 242.7) and in the case of the Nahand dam entrance station series, this test has also had the lowest quantity for the ARIMA(1, 1, 1) model (AIC= 111.4, SBC= 121.48).

After primary analysis of time series and identification, the next step is parameter estimation. For the model parameter estimation a computer program in Mathematical Laboratory (MATLAB 7) was written with the objective of minimizing the sum squares of error using the maximization likelihood and genetic algorithm methods. Genetic algorithm has tournament for selection function, uniform for mutation function and single point for crossover function. The optimized paramters of the genetic algorithm with regard to mimimization of the objective function has been gotten after several runs. These results are given in Table 3. The results of parameter estimation for the model using four methods are given in Table 4.

For the comparison among the mentioned methods of parameter estimation, two approaches have been used.

(1) Forecasting time series using estimated parameters and comparison of these series with some criteria.

(2) Determination of the parameters set which has the lowest the sum squares of error ($\sum_{t=1}^{n} \varepsilon^2{}_t$)

For evaluating the performance of forecasted values, it is common to use 25% of data for this purpose. In this research forecasting of time series using etimated parameters with several parameter estimation methods was conducted by ITSM software. Forecasted values in the Vanyar station series are related to the years from 2001 up to 2004 and, in the case of the Nahand dam entrance station series, are related to the years from 1997 up to 1999. Results have been shown graphically in Figures 8 and 9.

By observing the plots it can be understood that with a noticeable decreasing or increasing (trend) in the forecasted years, differences between observed and simulated flows will be increased. Because these models use pervious data for parameter estimation and such trends have not been noted in observed data like the flow of 1997 and 1998 at the Nahand station that has a noticeable increase in observed flows in forecasted years than in the pervious time series (Mean= 15.58).

Table 3. Optimal values of genetic algorithm (GA) parameters.

Parameter	value
Probability of mutation	.001
Probability of crossover	.8
Initial population	20
Number of generations	500

Model Performance Indicators

Table 4. The results of parameter estimation using different methods.

The methods of parameter estimation	Vanyar station	Nahand station
	Φ, θ	Φ, θ
Autocorrelation function formula based on model parameters	-0.33, 0.58	-0.199, 0.962
Conditional likelihood	-0.4, 0.6	- 0.3, 0.6
Unconditional likelihood	-0.3, 0.8	-0.23, 0.8
Genetic algorithm	- 0.216, 0.8178	-0.166, 0.907

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

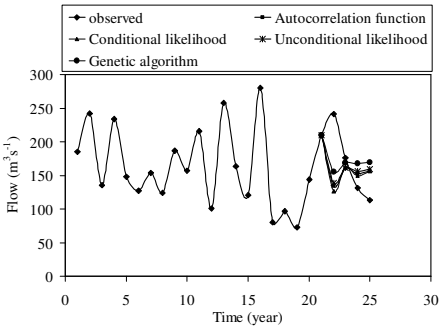

Figure 8. Forecasted and observed values of discharge (Ajichai River, Vaniar station)

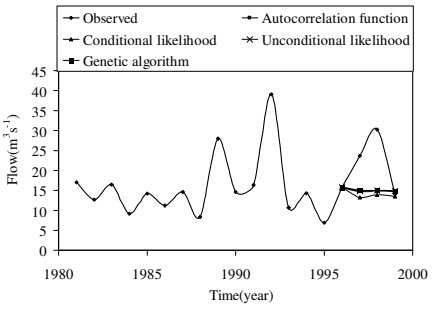

Figure 9. Forecasted and observed values of discharge (Ajichai River, Nahand station)

To evaluate the adequacy of the model with the proposed parameter estimation methods, the performance of the models should be analytically measured. Such criteria for the goodness of fit are obtained as:

Where $\hat{Q}_i$ is the forecasted streamflow (in this case forecasting using the mentioned

$$CE = 1 - \frac{\sum_{i=1}^{N}(\hat{Q}_i - Q_i)^2}{\sum_{i=1}^{N}(Q_i - \overline{Q})^2} \quad (21$$

$$SE = \frac{\left[\frac{\sum_{i=1}^{N}(\hat{Q}_i - Q_i)^2}{N}\right]^{\frac{1}{2}}}{\overline{Q}} \quad (22$$

$$MAE = \frac{1}{n}\sum_{i=1}^{N}\left[\left|Q_i - \hat{Q}_i\right|\right] \quad (23$$

$$RME = \frac{\sum_{i=1}^{N}\frac{\left|Q_i - \hat{Q}_i\right|}{Q_i}}{N} \quad (24$$

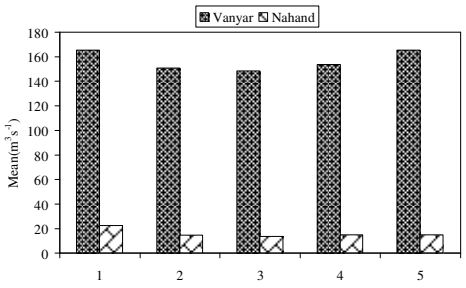

Figure 10. Comparison between statistics descriptions, mean, ($m^3 s^{-1}$).

methods of parameter estimation in modeling process), Q_i the observed streamflow, $\overline{Q}$ the mean observed streamflow and *N* the number of observed data items. The smallest RMSE (Root Mean Square Error) determines the method having the most accurate local or small-scale estimates. The smallest MAE (Mean Absolute Error) is indicative of the most accurate global estimates [9]. An RME (Relative Mean Error) value near zero implies that the model is providing a good estimate of observed values.

The minimum and maximum of these criteria are related to the genetic algorithm and conditional likelihood methods. The results in Table 5 illustrated that these criteria decreased from method one to four except in the second method, and it is an indication of the accuracy of parameter estimation methods from the first method to the fourth.

In the second approach, the sum squares of error were calculated for each group of parameters using the above mentioned methods. The minimum of the sum squares of error for example in Vanyar station according to Table 6 was related to the genetic algorithm method.

Another comparison is between statistical properties (mean and standard deviation) between observed and simulated series. Figure 10 shows that statistical properties like the mean with an indication that genetic algorithm has a lower difference between observed and forecasted streamflow using

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

Table 5. Performance of parameter estimation methods.

Method	Vanyar station				Nahand station			
	RMSE	RME	SE	MAE	RMSE	RME	SE	MAE
Autocorrelation function(1)	59.73	0.27	0.36	36.93	10.32	0.31	0.45	8.35
Conditional likelihood(2)	62.25	0.271	0.37	36.84	11.25	0.34	0.49	9.14
Unconditional Likelihood (3)	58.17	0.27	0.35	35.62	10.19	0.3	0.44	8.26
Genetic algorithm(4)	54.45	0.26	0.32	32.33	10.18	0.3	0.44	8.26

Table 6. Sum squares of error (Vaniar station).

Method	Minimum sum squares of error
Conditional likelihood	79938.75
Unconditional likelihood	76931.64
Genetic algorithm	76875.27

the GA parameter estimation method because of the precisely estimated parameter.

Maximum likelihood estimation is a reasonably well-principled way to work out what computation is needed for learning some kinds of model from the data. Some advantages of the maximum likelihood method over other methods are: it has a lower variance than other methods; the method is statistically well founded; and it uses all the sequences information. In this research using the maximum likelihood estimation method for parameter estimation of the ARIMA model led to the least squares in which, for minimizing the sum squares of error, the three methods of condition likelihood, unconditional likelihood and genetic algorithm have been used.

The main advantages of back box models in hydrology are that they are not as data demanding as physical models. In principles, the parameters in a physically based model can be estimated by field measurements, but such an ideal situation requires comprehensive field data which cover all the parameters. Because of the large number of parameters in a physically based model, parameter estimation cannot be done by free optimization for all parameters.

This research indicates that genetic algorithm and unconditional likelihood methods are respectively more appropriate in comparison with other methods. These results are given using some criteria such as RMSE, SE and minimizing the sum squares of error. Also the comparison between statistical properties indicated the increasing accuracy of the first parameter estimation method until the fourth except the second method.

The autocorrelation function formula method based on model parameters gave acceptable results but, by increasing the orders of model, deriving equations of parameters and solving of them is complicated. In the condition likelihood method, calculation and writing the program is simple but for accurate estimation of the sum squares of error, this method dose not use precise previous time series. The unconditional likelihood method using two forward and backward equations can overcome this difficulty. In addition, maximum likelihood is very CPU intensive and thus extremely slow.

Genetic algorithm does not work with parameters but it works with the coding of parameters set. In this regard, search space exponentially increased and has high confidence for achieving the global optimum. Genetic algorithm by speeding up the search producer and getting the global optimum can conquer the difficulties of other parameter estimation methods especially when the orders and the number parameters of the model are increased.

It must be mentioned that these methods for parameter estimation have used observed data and the persistence of trend (noticeable decrease or increase) in the observed data of forecasted years can be effected the forecasted data.

One of the modeling steps is checking the normality of time series. Then, using appropriate transformation for this regard is necessary. One of the reasons of high difference between observed and forecasted

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

flows in Nahand may be the need for other normal transformation.

CONCLUSIONS

Parameter estimation is one of the main steps of time series modeling. This research investigates using a stochastic model by different parameter estimation methods for forecasting annual streamflow. Four parameter estimation methods - autocorrelation function based on model parameters, conditional likelihood, unconditional likelihood and a genetic algorithm- have been used. Some criteria such as RMSE, MRE, SE, MAE and the minimum sum squares of error have been used for comparison of the performance of parameter estimation methods. This research indicates that genetic algorithm and unconditional likelihood methods are respectively more appropriate in comparison with other methods but, due to the complexity of the model, genetic algorithm (GA) has a high convergence speed to global optimum.

It must be mentioned that, for getting precise forecasted values, appropriate normal transformation using parameters estimation method has an important affect. Therefore, we can propose this method for ARIMA streamflow modeling.

REFERENCES

1. Box, E. P. and Jenkins, G. M. 1976. *Time Series Analysis: Forecasting and Control*. Prentice-Hall, Englewood Cliffs, NJ.
2. Sharman, K. C. and Breakenbridge, E. 1994. Estimation of Signal Parameters Using the Maximum Likelihood Method. Mathematical Aspects of Digital Signal Processing, *IEE Colloquium,* **8**: 1-4.
3. Taqqu, S. 1997. Fractionally Differenced ARIMA Models Applied to Hydrologic Time Series: Identification, Estimation and Simulation. *Water Resour. Res.,* **33(5):** 1035-1044.
4. Salas, J. D., Delleur, J. W., Yevjevich, V. and Lane, W. L. 1988. *Applied Modeling of Hydrologyic Time Series*. Water Recourse Publication, PP.192-194.
5. Anderson, P. L., Meerschaert, M. M. and Vecchia, A. V. 1999. Innovations Algorithm for Periodically Stationary Time Series. Stochastic Processes and Their Applications, **83**:149-169.
6. Lu, S. and Chon, K. H. 2000. A New Algorithm for ARIMA Model Parameter Estimation Using Groupmethod of Data Handling. *Bioengin. Con., Proc. of the IEEE 26th Annual Northeast*. PP.127-130.
7. Shine, D. W. and Lee, J. H. 2000. Consistency of the Maximum Likelihood Estimators for Nonstationary ARIMA Regressions with Time Trends. *Journal of Statistical Planning and Inference*, **87**: 55-68.
8. Khalil, M., Panu, U. S. and Lennox, W. C. 2001. Groups and Neural Networks Based Streamflow Data Infilling Procedures. *J. Hydrol.*, **241**:153-176.
9. Schloeder, C. A., Zimmerman, N. E. and Jacobs, M. J. 2001. Comparison of Methods for Interpolating Soil Properties Using Limited Data. *Soil Sci. Soc. Amer. J.,* **65**: 470-479.
10. Chen, B. S., Lee, B. K. and Peng, S. P. 2002. Maximum Likelihood Parameter Estimation of *F*-ARIMA Process Using the Genetic Algorithm, In: "The Frequency Domain". *IEEE Trans. on processing*, **50(9):**2208-2220.
11. Karamouz, M. and Kerachian, R. 2003. *Water Quality Planning and Management*. Amirkabir University of Technology Puplication. PP.130-134.
12. Valenzuela, O., Marquez, L., Pasadas, M. and Rojas, I. 2003. Automatic Identification of ARIMA Time Series by Expert Systems Using Paradigms of Artificial Intelligence. *Mongrafias del Seminaro Garcia de Galdeeano*, **31**:425-435.
13. Wurtz, D., Chalabi, Y. and Luksan, L.2003. Parameter Estimation of ARMA Models with GARCH/APARCH Errors an R and S Plus Software Implementation. *J. Stat. Soft.,* **5(2):** 1-41.
14. Shyonong, C., Huang, J. J. and Tzeng, G. H. 2005. Model Identification of ARIMA Family Using Genetic Algorithms. *Applied Mathematics and Computation* **164(3):** 885-912.
15. Kurunce, A., Yurekli, K. and Cevil, O. 2005. Performance of Two Stochastic Approaches

[DOR: 20.1001.1.16807073.2010.12.1.8.4]

for Forecasting Water Quality and Streamflow Data from Yeslirmak River, Turkey. *Environ. Model. Soft., 20*: 1195-1200.
16. Jonsdottir, H., Madsen, H. and Palsson, O. P. 2006. Parameter Estimation in Stochastic Rainfall-runoff Models. J. Hydrol., **326**: 379- 393.
17. Obadage, A.S. and Harnpornchai, N. 2006. Determination of Point of Maximum Likelihood in Failure Domain Using Genetic Algorithms. *Int. J. of Pressure Vessels and Piping,* **83**: 276-282.
18. Valenzuela, O., Rojas, I., Rojas, F., Pomares, H., Herrera. L. J., Guuillen, A., Marquez, L. and Pasadas, M. 2008. Hybridization of Intelligent Techniques and ARIMA Models for Time Series Prediction. *Fuzzy Sets and Systems*, **159**: 821-845.

مقایسه بین کارایی روش های تخمین پارامترها در پیش بینی جریان رودخانه

ل. پرویز، م. خلقی و ا. هورفر

چکیده

پیش بینی متغیرهای هیدرولوژیکی مانند جریان رودخانه نقش مهمی در برنامه ریزی و مدیریت منابع آب دارد. در سال های اخیر استفاده از مدل های استوکستیکی جهت این امر رواج بسیاری یافته اند. پیش بینی توسط مدل ARIMA از طریق تخمین پارامترهای مجهول امکان پذیر می باشد زیرا تخمین پارامتر به عنوان یکی از اساسی ترین مراحل مدلسازی می باشد. هدف اصلی این تحقیق بررسی عملکرد روش های تخمین پارامتر در مدل ARIMA می باشد. در این مطاله چهار روش تخمین پارامتر شامل: ۱. استفاده از فرمول های ضرایب خود همبستگی بر حسب پارامترهای مدل ۲. درستنمایی شرطی ۳. درستنمایی غیر شرطی ۴. الگوریتم ژنتیک مورد استفاده قرار گرفتند. آمار آبدهی حوضه آبریز دریاچه ارومیه در شمال غرب ایران به عنوان منطقه مورد مطالعاتی در نظر گرفته شد. نتایج این چهار روش توسط معیارهای RMSE, RME, SE, MAE و مینیمم مجموع مربعات خطا مقایسه شدند. نتایج مبین کارایی روش های الگوریتم ژنتیک و درستنمایی غیر شرطی نسبت به سایر روش های تخمین پارامترها می باشند ولی با افزایش پیچیدگی، مدل الگوریتم ژنتیک سرعت همگرایی بالایی برای رسیدن به بهینه مطلق را دارد.

[DOR: 20.1001.1.16807073.2010.12.1.8.4]