Statistical modeling of monthly streamflow using time series and artificial neural network models: Hindiya Barrage as a case study


a Al-Mussaib Technical Institute, Al-Furat Al-Awsat Technical University, Babylon 51009, Iraq
b Department of Environment Engineering, Babylon University, Babylon 51001, Iraq
c School of Civil Engineering and Built Environment, Liverpool John Moores University, Liverpool, UK
*Corresponding author. E-mail: khalid_alhilli@yahoo.com

Abstract

Autoregressive Integrated Moving Average (ARIMA) Box-Jenkins models combine the autoregressive and moving average models to a stationary time series after the appropriate transformation, while the nonlinear autoregressive (N.A.R.) or the autoregressive neural network (ARNN) models are of the kind of multi-layer perceptron (M.L.P.), which compose an input layer, hidden layer and an output layer. Monthly streamflow at the downstream of the Euphrates River (Hindiya Barrage) /Iraq for the period January 2000 to December 2019 was modeled utilizing ARIMA and N.A.R. time series models. The predicted Box-Jenkins model was ARIMA (1,1,0) (0,1,1), while the predicted artificial neural network (N.A.R.) model was (M.L.P. 1-3-1). The results of the study indicate that the traditional Box-Jenkins model was more accurate than the N.A.R. model in modeling the monthly streamflow of the studied case. Performing a one-step-ahead forecast during the year 2019, the forecast accuracy between the forecasted and recorded monthly streamflow for both models was as follows: the Box-Jenkins model gave root mean squared error (RMSE = 48.7) and the coefficient of determination (R² = 0.801), while the (NAR) model gave (RMSE = 93.4) and (R² = 0.269). Future projection of the monthly stream flow through the year 2025, utilizing the Box-Jenkins model, indicated the existence of long-term periodicity.

Key words: ANN, Box-Jenkins, monthly stream flow, RMSE, time series

Highlights

• Box-Jenkins models and artificial neural network (ANN) were used to model the flow of Hindiya Barrage.
• The collected data covered the period of January 2000 to December 2019.
• Box-Jenkins model was more accurate than ANN in modeling the monthly flow of the studied river.
• The outcomes indicated long-term periodicity (until 2025).

INTRODUCTION

Streamflow is an important issue in the design, operation, and control of many vital projects in the water resources and sanitary/environmental engineering specialities; examples are reservoir storage capacity, water treatment plants (WTPs), and wastewater treatment plants (WWTPs). For this reason, recorded data of streamflow (hourly, daily, weekly, monthly, and yearly) are required...
for the optimum design, operation, and control of these vital projects. Kawamura (2000) showed that during the preliminary studies to design a (WTP), the project engineer must evaluate the potential sources of water; one of the elements of this evaluation is the quantity of water required, which is directly related to the streamflow of the nearby river (delivering raw water). The importance of streamflow in the design, operation, and control of a WWTP is best illustrated by the direct relation between the self-purification of rivers to the river streamflow (as shown in the Streeter-Phelps equation) and the direct effect of the dilution factor (Metcalf et al. 1979). Monthly streamflow of the Sefidrood river, Iran, and Sangeen river, Canada, were studied and modeled by applying autoregressive (AR), moving average (MA) models and compared with two artificial intelligence (AI) approaches, namely, multivariate adaptive regression splines (MARS) and gene expression programming (GEP). Results indicate that AI models outperformed the conventional AR and MA models (Mehdizadeh et al. 2019). Frausto-Solis et al. (2008) compared ARIMA versus ANN in forecasting streamflow. Their results indicate that ARIMA has a higher forecast accuracy than ANN methodology. Moeni et al. (2017) compared the Seasonal Autoregressive Integrated Moving Average (SARIMA) with the Artificial Neural Network-Genetic Algorithm (ANN-GA) method in forecasting the monthly streamflow. Their results confirm that the SARIMA model have much more accuracy than the ANN-GA model in short-term and long-term forecasting. Joodavi et al. (2020) presented a methodology based on the combination of numerical groundwater flow simulations and reservoir operation optimization models to develop an optimization model for the management of the off-stream Bar-Reservoir operation (Iran), taking into account the lake bed seepage and the random inflow to the reservoir, the planned objective was to satisfy water demand for a total amount of 12 million m$^3$/year for drinking and industrial purposes. A two-stage time series model for the monthly flows of the Lim River basin in South-Eastern Europe for the period 1950–2012 was carried out by Stojkovic et al. (2020). The model took into consideration climate change and consisted of several components (trend, long-term periodicity, seasonality, and the stochastic component). It was designed to estimate future water availability. Water demand of any city varies according to the variation of climatic variables (Zubaidi et al. 2020a, 2020b, 2020c), such as rainfall, temperature, humidity, and evaporation (Zubaidi et al. 2019a, 2019b, 2020d, 2020e). Feng & Niu (2021) proposed a hybrid artificial neural network model enhanced by the addition of a cooperation search algorithm for nonlinear river stream flow time series forecasting. Mohammadi et al. (2020) developed novel robust models to improve the accuracy of daily streamflow time series modeling. Niu & Feng (2021) evaluated the performance of five artificial intelligence models in forecasting the daily streamflow time series. Adnan et al. (2020) evaluated the abilities of three models to predict monthly streamflow; they are Group Method of Data Handling-Neural Networks (GMDH-NN), Dynamic Evolving Neural-Fuzzy Inference System (DENFIS), and Multivariate Adaptive Regression Spline (MARS) methods. Peng et al. (2020) developed a monthly streamflow prediction model based on the Random Forest Algorithm and Phase Space Reconstruction Theory. Khazaei et al. (2020) simulated daily runoff correlated to weather generators utilizing the LARS-WG model. Avand & Moradi (2020) used machine learning models (including LARS-WG), remote sensing, and GIS to study the effects of changing climatic variables and land uses on flood probability.

To the best of the authors’ knowledge, this is the first time that statistical modeling models were built for monthly streamflow at the Euphrates River (Hindiya Barrage) in Iraq. Generally, this paper aims to firstly investigate the monthly streamflow at downstream of the Euphrates River (Hindiya Barrage) /Iraq for the period of January 2000 to December 2018. Secondly, building a Box-Jenkins ARIMA forecasting model for the data recorded during the above period, checking its adequacy and proving its goodness of fit. Then, performing a one-step-ahead forecast during the year 2019 (12 months). Thirdly, train, test and validate the above-recorded data using the multi-layer perceptron (MLP) and feed-forward neural network (FNN). Then performing a one-step-ahead forecast during the year 2019 (12 months). Finally, compare the forecast accuracy for the two methods relying on the root mean squared error (RMSE)
and the coefficient of determination ($R^2$). Average monthly streamflow data at the downstream of the Euphrates River (Hindiya Barrage), Iraq for the period January 2000 to December 2019 were selected as a case study (these data were obtained from the Ministry of Water Resources/Al-Mussaib Water Resources Directorate, Iraq). The data for the period 2000–2018 was adopted in model building and that during 2019 were adopted for calculating model forecast accuracy.

Figure 1 is an aerial view (Google Earth) of the study area with Latitude 32°43′01″N, Longitude 44°16′01″E.

**Figure 1 |** Aerial view of Hindiya Barrage.

**MATERIALS AND METHODS**

Univariate Box-Jenkins ARIMA family of time series models

A time series of a finite number of successive observations consisting of the data $Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots, Y_n$ is called a univariate time series. Autoregressive Integrated Moving Average (ARIMA) models describe a collection of time series models that can be very simple or complicated (Brown & Mac Berthouex 2002). A seasonal ARIMA model for $Y_t$ is written as (Geurts 1977):

$$\theta(B)\phi(B^S)(\nabla^d \nabla^D Y^A_t - \mu) = \theta(B)\phi(B^S)\alpha_t$$

where $Y^A_t$ represent some appropriate transformation of $Y_t$, $t$ is the discrete-time, $S$ is the seasonal length (equals12 for monthly data), $B$ is the backshift operator defined by $BY_t = Y_{t-1}$ and $B^SY_t = Y_{t-S}$, $\mu$ is the mean level of the series, usually taken as the average of the $W_t$ series (if $D + d > 0$, often $\mu \approx 0$), $\alpha_t$ is normally independently distributed white noise series (no autocorrelation) (with mean = 0 and variance = $\sigma^2_{\alpha}$, it is written as NID (0, $\sigma^2_{\alpha}$).

$$\Theta(B) = 1 - \Theta_1B - \Theta_2B^2 \ldots \ldots \ldots - \Theta_pB^p$$

where $\Theta$ (B) is a non-seasonal autoregressive (AR) operator of order (P).

$$\Phi(B^S) = 1 - \Phi_1B^S - \Phi_2B^{2S} \ldots \ldots \ldots - \Phi_pB^{pS}$$

where $\Phi$ ($B^S$) is the seasonal AR operator of order (P), $\nabla^d$ = non-seasonal differencing operator of order (d) to produce non-seasonal stationarity (trend removal), usually $d = 0, 1$ or 2, $\nabla^D_S$ = seasonal
differencing operator of order \( (D) \) to produce seasonal stationarity (seasonality removal), usually \( D = 0, 1 \) or 2.

\[
\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q
\]

(4)

where \( \theta(B) \) = non-seasonal moving average (MA) operator of order\( (q) \).

\[
\psi(B) = 1 - \psi_1 B^S - \psi_2 B^{2S} - \ldots - \psi_Q B^{QS}
\]

(5)

where \( \psi(B) \) = seasonal (MA) operator of order\( (Q) \), and \( W_t \) = stationary time series formed by non-seasonal and/or seasonal differencing of \( Y_t^{A} \) time-series. ARIMA models (Box et al. 2015) can handle the problem of statistical modeling of any time-dependent phenomenon, including model building, forecasting and diagnostic checking, taking into account stationarity, missing data, outlier observations, intervention, normality, and independence of residuals. A stationary time series is defined as a time series without trend and seasonality; that is, with zero mean and small variance (Box et al. 2015). For the precise explanation of the three steps of model building (identification, estimation, and diagnostic check), the interested reader may refer to (Ljung & Box 1978; Ang & Tang 2007). The above three steps of Box-Jenkins ARIMA modeling (model building, forecasting and diagnostic checking) can be easily executed using IBM SPSS version 20 software (Yaffee & McGee 2000).

**Artificial neural network (ANN) models**

The recommended modeling procedure here, according to the published literature (Faraway & Chatfield 1998; Tealab et al. 2017), is the nonlinear autoregressive (NAR) or the autoregressive neural network (ARNN). It is of the kind of multi-layer perceptron (MLP), which comprises an input layer, hidden layer and an output layer. The input layer holds the target vector (time series), while the output layer computes the estimator vector (time series). The hidden and output layers are governed by an activation function, which may be defined by a (logistic, tanh, softmax…), and a weight function (uniform and Gaussian). Figure 2 is a schematic representation of a perceptron learning process.

**Figure 2 | Perceptron learning process.**

The nonlinear autoregressive model of order \( p \), NAR (p), is defined as:

\[
Y_t = h(Y_{t-1}, \ldots, Y_{t-p}) + \varepsilon_t
\]

(6)

where \( h(\cdot) \) is a nonlinear function; it is assumed that \( \{\varepsilon_t\} \) is a sequence of random independent variables and identically distributed with zero mean and a finite variance \( \sigma^2 \). The ARNN is a feed-forward
network that constitutes a nonlinear approximation \( h(\cdot) \), which is defined as:

\[
\hat{Y}_t = \hat{h}(Y_{t-1}, \ldots, Y_{t-p})
\]

\[
\hat{Y}_t = \beta_0 + \sum_{i=1}^{l} \beta_i f\left( \alpha_i + \sum_{j=1}^{p} \omega_{ij} Y_{t-j} \right)
\]

where \( f(\cdot) \) function is the activation function and \( \theta = (\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q, \omega_{11}, \ldots, \omega_{pq}) \) is the parameter vector of the neural network, which is calculated by minimizing the sum of squared errors:

\[
SSE = \sum_{t=1}^{N} (Y_t - \hat{Y}_t)^2
\]

where \( \hat{Y}_t \) is the estimator of the target variable \( Y_t \). The recorded data (for the purpose of analysis) is divided into three portions, \( N_1, N_2 \) and \( N_3 \). \( N_1 \) of the data are used for training, \( N_2 \) are used for data validation to check the prediction accuracy for the model selection, and \( N_3 \) of the data are employed for the out-of-sample predictions (forecasts) by the calibrated model. The smallest number of RMSEs for the validation data becomes a desirable model (Kajitani et al. 2005). During the training stage, several algorithms were adopted for optimization and truncating the iterations (cycles) after reaching the specified error criteria, strictly speaking, Broyden-Fletcher-Goldfarb-Shanno (BFGS), Scaled Conjugate and Gradient Descent algorithms (Bishop 1995; Becerikli et al. 2003). The model that gives the maximum correlation coefficient (R) and the minimum sum of squared errors (SSE) will be relied on in the analysis (KILIŞ & Sciences 2005). The above modeling procedure was performed and analyzed in this research adopting STATISTICA version 12 software.

Forecasting

After model building, it is necessary to make a one-step-ahead forecast. To check the forecasting accuracy, several formulas are calculated and checked according to the forecast accuracy, these are coefficient of determination \( (R^2) \), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute error (MaxAE), and maximum absolute percentage error (MaxAPE). The above formulas can be found in any statistical textbook. For example, to define (RMSE) and \( R^2 \), then:

\[
RMSE = \sqrt{\frac{\sum_{t=1}^{N} (Y_t - F_t)^2}{N}}
\]

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}}
\]

\[
SS_{res} = \sum_{t=1}^{N} (Y_t - F_t)^2 = \sum_{t=1}^{N} (E_t)^2
\]

\[
SS_{tot} = \sum_{t=1}^{N} (Y_t - \bar{Y})^2
\]

where \( N \) = the number of months to be forecasted in the future (normally 12), \( Y_t \) = observed (recorded) stream flow at month \( t \) (\( m^3/s \)), \( F_t \) = forecasted streamflow at month \( t \) (\( m^3/s \)), \( \bar{Y} \) = average
of observed values \((m^3/s)\), \(\bar{F}\) = average of forecasted values \((m^3/s)\), \(E_t\) = error (residual) at time \(t\) \((m^3/s)\). These accuracy formulas are calculated and updated from the same recommended software, IBM SPSS version 20 and STATISTICA version 12.

**RESULTS AND DISCUSSION**

Applying the Box-Jenkins modeling procedure to the recorded data, using IBM SPSS version 20 software referring to the route (Analyze/Forecasting/Create Models/Expert Modeler), the best-fit model relying on the RMSE criteria was: ARIMA \((1,1,0) (0,1,1)\).

The appropriate equation was:

\[
(1 + 0.174B)\nabla^1 \nabla^1_{12} \ln \hat{Y}_t = (1 - 0.951B^{12})a_t
\]

(14)

The original data suggested a natural logarithmic transformation to enhance the normality of residuals. The above model gave \(R^2 = 0.844\), RMSE = 42.049 and Ljung-Box \(Q(18) = 21.906\) with degrees of freedom \((DF) = 16\), also the significance of \(Q\) was 14.6% > 5%, indicating that residuals from the model were uncorrelated (random). Figure 3 depicts the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the residuals. Figure 4 depicts the normal probability paper of the residuals.

![Figure 3 | ACF and PACF of the residuals.](image-url)

From the above figures, it is evident that the residuals are normally independently distributed. Figure 5 describes the predicted values of the streamflow time series together with the recorded values. It states that there was a clear, strong correlation between predicted and recorded values.

Figure 6 depicts the forecasted monthly streamflow for the 12 months during 2019, together with the recorded values. The calculated RMSE = 48.7 and the calculated \(R^2 = 0.801\).

Applying the (ANN) modeling procedure to the above data using the STATISTICA version 12 software, referring to the route (Data Mining/Neural Networks/Time Series (Regression)), taking into consideration that the recorded data was considered as the Target, the input layer was taken as the
Figure 4 | Normal probability paper of the residuals.

Figure 5 | Box-Jenkins ARIMA model predicted time series vs. recorded time series for the period 2000–2018.

Figure 6 | Recorded vs. Forecasted (by Box-Jenkins) monthly stream flow during 2019.
Target with a lag of specified length, the input layer data were classified as (70% Training + 15% Test + 15% Validation) randomly. The results in the output layer will be compared with the target. Table 1 lists the statistics of the predicted ANN models for different lags between (1–12), noting that one cycle of seasonality = 12 months.

Table 1 | Showing statistics of the predicted ANN models

<table>
<thead>
<tr>
<th>Lag</th>
<th>Predicted model</th>
<th>Training (R)</th>
<th>Test (R)</th>
<th>Validation (R)</th>
<th>Training (S.S.E.)</th>
<th>Test (S.S.E.)</th>
<th>Validation (S.S.E.)</th>
<th>Training algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MLP 1-3-1</td>
<td>0.560163</td>
<td>0.651468</td>
<td>0.898259</td>
<td>4,257.256</td>
<td>3,018.890</td>
<td>821.1875</td>
<td>BFGS 4</td>
</tr>
<tr>
<td>2</td>
<td>MLP 2-6-1</td>
<td>0.544030</td>
<td>0.435790</td>
<td>0.868627</td>
<td>5,330.841</td>
<td>4,504.276</td>
<td>2,921.298</td>
<td>BFGS 2</td>
</tr>
<tr>
<td>3</td>
<td>MLP 3-2-1</td>
<td>0.601277</td>
<td>0.735790</td>
<td>0.860533</td>
<td>3,917.588</td>
<td>2,356.584</td>
<td>919.7809</td>
<td>BFGS 33</td>
</tr>
<tr>
<td>4</td>
<td>MLP 4-8-1</td>
<td>0.616656</td>
<td>0.756074</td>
<td>0.867327</td>
<td>3,765.428</td>
<td>2,182.703</td>
<td>864.7852</td>
<td>BFGS 34</td>
</tr>
<tr>
<td>5</td>
<td>MLP 5-5-1</td>
<td>0.518009</td>
<td>0.519080</td>
<td>0.816368</td>
<td>4,448.370</td>
<td>4,219.433</td>
<td>1,246.969</td>
<td>BFGS 3</td>
</tr>
<tr>
<td>6</td>
<td>MLP 6-2-1</td>
<td>0.614488</td>
<td>0.729865</td>
<td>0.800189</td>
<td>3,755.190</td>
<td>2,356.035</td>
<td>1,026.629</td>
<td>BFGS 41</td>
</tr>
<tr>
<td>7</td>
<td>MLP 7-7-1</td>
<td>0.662119</td>
<td>0.706750</td>
<td>0.796674</td>
<td>3,394.344</td>
<td>2,612.846</td>
<td>1,170.198</td>
<td>BFGS 35</td>
</tr>
<tr>
<td>8</td>
<td>MLP 8-3-1</td>
<td>0.657709</td>
<td>0.637734</td>
<td>0.784143</td>
<td>3,426.545</td>
<td>3,208.390</td>
<td>1,128.715</td>
<td>BFGS 38</td>
</tr>
<tr>
<td>9</td>
<td>MLP 9-6-1</td>
<td>0.548164</td>
<td>0.650932</td>
<td>0.846927</td>
<td>4,201.155</td>
<td>2,708.997</td>
<td>909.2035</td>
<td>BFGS 7</td>
</tr>
<tr>
<td>10</td>
<td>MLP 10-4-1</td>
<td>0.498419</td>
<td>0.499938</td>
<td>0.743900</td>
<td>4,498.307</td>
<td>3,815.725</td>
<td>1,191.457</td>
<td>BFGS 5</td>
</tr>
<tr>
<td>11</td>
<td>MLP 11-6-1</td>
<td>0.533244</td>
<td>0.475782</td>
<td>0.759208</td>
<td>4,305.496</td>
<td>3,775.697</td>
<td>1,342.761</td>
<td>BFGS 7</td>
</tr>
<tr>
<td>12</td>
<td>MLP 12-8-1</td>
<td>0.494050</td>
<td>0.656991</td>
<td>0.723700</td>
<td>4,364.795</td>
<td>2,680.175</td>
<td>1,302.532</td>
<td>BFGS 3</td>
</tr>
</tbody>
</table>

Table 1 reflects the fact that all predicted models are not strongly correlated with the Target, indicating a small coefficient of determination ($R^2$). It is evident from Table 1 that the maximum (R) and the minimum SSE (validation stage) occurred at lag 1 (MLP 1-3-1); for this, it is selected as the best model. The notation (MLP 1-3-1), for example, refers to a multi-layer perceptron with 1 (number of inputs), 3 (number of hidden units), and 1 (number of outputs); also, the notation BFGS 4 refers to the Broyden-Fletcher-Goldfarb-Shanno training algorithm with 4 cycles of iteration. Figure 7 is a time-series graph that compares the prediction of the model at lag 1 (MLP 1-3-1) with the recorded data (Target). Figure 8 compares the (Target) streamflow $Q$ (x-axis) with the output streamflow.
from the selected ANN model (y-axis). It is evident from Figures 7 and 8 that the predictions are not strongly correlated to the original (recorded) data, as prescribed in Table 1. The ANN time series model gave RMSE = 84.63 and $R^2 = 0.365$, indicating that it is less efficient than the Box-Jenkins time series model predicted above (which gave RMSE = 42.049 and $R^2 = 0.844$). Figure 9 depicts the forecasted monthly streamflow for the 12 months during 2019 (by the ANN time series model), together with the recorded values. The calculated (RMSE = 93.4) and the calculated ($R^2 = 0.269$). From this figure, it is evident that the ANN model does not simulate the seasonality of the recorded data. Figures 6 and 9 give an indication that Box-Jenkins models are competent in simulating seasonality, which is not the case for ANN models. Comparing the forecast accuracy (RMSE and $R^2$) for the ANN model during 2019 with that resulted from the Box-Jenkins model, it is evident that the latter is more accurate. This result is in accordance with the results documented by other authors (Moeeni et al. 2017; Mehdizadeh et al. 2019). Figure 10 illustrates the future forecasts resulting from the Box-Jenkins model mentioned above by Equations (14) and (15) together with the recorded and fit monthly streamflow for the period 2000–2018. The long-term periodicity is clearly observed; this is in accordance with Stojković et al. (2020).
CONCLUSION

From the results of the study, it can be concluded that the Box-Jenkins model was more accurate than the ANN model in forecasting future monthly streamflow downstream of the Euphrates River (Hindiya Barrage)/Iraq for the period 2000–2019, that the ANN model was not able to simulate seasonality, which was not the case for the Box-Jenkins model. Future forecast of monthly streamflow by Box-Jenkins models indicates the existence of some long-period trend in the form of long-term periodicity, which already exists in the recorded data. It is advisable to study the monthly streamflow in terms of the climatic variables through multiple regression and apply the same modeling procedure; that is, the Box-Jenkins (transfer function models) and the ANN (regression models), also the Support Vector Regression and the Random Forest analysis may be adopted in the future studies.

ACKNOWLEDGEMENTS

The authors would like to express their appreciation and gratitude to the Ministry of Higher Education and Scientific Research, Al-Furat Al-Awsat Technical University, and the Ministry of Water Resources/Al-Mussaib Water Resources Directorate/Iraq for the facilities provided to complete this research. The authors are also grateful to the colleagues and the technician team who provided insight and expertise that greatly assisted the completion of the research.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES


First received 29 December 2020; accepted in revised form 2 February 2021. Available online 15 February 2021.