Prediction of Monthly Fluoride Content in Tigris River using SARIMA Model in R Software

Asst. Prof. Dr. Awatif Soaded Alsaqqar
Department of Civil Engineering
College of Engineering
Baghdad University
Email: d.alsaqqar@yahoo.com

Asst. Prof. Dr. Basim Hussein Khudair
Department of Civil Engineering
College of Engineering
Baghdad University
Email: basim22003@yahoo.com

Asst. Lect. Rasha Attwan Faraj
Department of Civil Engineering
College of Engineering
Baghdad University
Email: rashaatwan@yahoo.com

ABSTRACT

The need to create the optimal water quality management process has motivated researchers to pursue prediction modeling development. One of the widely important forecasting models is the sessional autoregressive integrated moving average (SARIMA) model. In the present study, a SARIMA model was developed in R software to fit a time series data of monthly fluoride content collected from six stations on Tigris River for the period from 2004 to 2014. The adequate SARIMA model that has the least Akaike's information criterion (AIC) and mean squared error (MSE) was found to be SARIMA (2, 0, 0) (0,1,1). The model parameters were identified and diagnosed to derive the forecasting equations at each selected location. The correlation coefficient between the actual and predicted values for fluoride concentration at the six locations, Al-Karakh, East Tigris, Al-Wathbah, AL-Karamah, Al-Rashid and Al-Wahda WTP intakes, was 0.93, 0.82, 0.86, 0.90, 0.83 and 0.89, respectively. Model verification results indicated that the model forecasting outputs rationally estimated the actual monthly fluoride content in the selected locations.

Keywords: water quality management, time series analysis and prediction, SARIMA model, R software.

التنبؤ بمحتوى الفلورايد الشهري في نهر دجلة باستخدام (SARIMA) موديل في برنامج (R)

أ.م. د. عواطف سؤدد عبدالحميد
قسم الهندسة المدنية
كلية الهندسة/جامعة بغداد

أ.م.د. باسم حسين خضير
قسم الهندسة المدنية
كلية الهندسة/جامعة بغداد

م.م. رشا عطوان فرج
قسم الهندسة المدنية
كلية الهندسة/جامعة بغداد

الخلاصة

الحاجة لابتكار نظام افضل لادارة نوعية المياه قد حفزت الباحثين لمواصلة تطوير نماذج التنبؤ. واحد من نماذج التنبؤ المهمة والواسعة الانتشار(SARIMA). في هذه الدراسة, تم تطوير نموذج (SARIMA) بأستخدام برنامج (R) لموافقة السلاسل الزمنية لمحتوى الفلورايد الشهري لستة مواقع في نهر دجلة خلال الفترة الزمنية من (2004-2014). وقد وجد النموذج الأنسب لاحتوائه على أقل قيمة (AIC) و قيمة معدل مربع الخطأ (MSE) وقيمة (SARIMA) لتكون (0،1،1) (2،0،0). معالم النموذج قد عرفت وشخصت لاشتقاق معادلات التنبؤ في كل المواقع المحددة. معامل الارتباط بين القيم الفعلية والمتوقعة لتركيز الفلوريد في المواقع الستة، مآخذ الكرخ، شرق دجلة، الوثبة، الكرامة، الرشيد و الوحدة كان 0.93، 0.82، 0.86، 0.90 ، 0.83 و 0.89 على التوالي. قد بينت نتائج أختبار النموذج بأن نتائج التنبؤ للنموذج قد قدرت محتوى الفلورايد الشهري الفعلي في المواقع المختارة جيدا.

الكلمات الرئيسية: ادارة نوعية المياه، تحليل والتنبؤ للسلاسل الزمنية ، نموذج SARIMA، برنامج R .

1. INTRODUCTION

Water quality is a critical subject of the ongoing environmental concerns. Impairment of water quality has motivated researchers to develop methods to monitor water characteristics. Usually monitoring water quality in lakes, streams and rivers is performed through the conventional methods which involve field measurements and laboratory analysis. The conventional methods are useful for small water surfaces. However, for large areas, the conventional methods are expensive and incapable for monitoring and evaluating of regional water quality when numerous sampling locations need to be evaluated periodically , **Hirsch et al., 1982**. Therefore, thoughtful water quality management efforts have been taken into consideration in many countries to be utilized in conjunction with conventional methods.
Water quality management includes modeling, forecasting and analysis of water bodies' quality. Developing precise forecasting model of future water parameter is the essence of optimal water quality management. Modeling methods have been improved with the continued development of computer science and statistics particularly for discovering time series data patterns. Time series prediction is ongoing area of forecasting approaches. Past collected observations of the same parameter are investigated to create a model that describes the underlying relationship. Afterwards, the developed model is used to estimate the parameter in the future.

One of the most useful and important time series prediction models is the autoregressive integrated moving average (ARIMA) model , **Sowell, 1992**. The advantages of the ARIMA model can be attributed to its statistical characteristics, represented by the famous Box–Jenkins methodology ,**Harvey, 1990** in the model building process. Moreover, ARIMA models can be performed to represent several exponential smoothing applications including environmental applications ,**Williams** and **Hoel, 1999**. The development of ARIMA models have continued over decades. The basic ARIMA model has been implemented to include pure autoregressive (AR) or pure moving average (MA). Then, it was developed to combine both autoregressive and moving average (ARMA) compounds. To implant this model for non-stationary time series data, the integrated compound was included in the model to be the ARIMA model. Afterwards, studies have investigated the effect of seasonality on fitting an accurate ARIMA model and found that seasonality causes weakly stationary condition. Therefore, seasonal compound was added to the original ARIMA model to decompose a time series data uniquely into equally independent additive seasonal, trend, and irregular noise components **Hillmer** and **Tiao, 1982, Williams** and **Hoel, 2003**. Recently, the seasonal ARIMA (SARIMA) model has been used for water quality management applications**, Lehmann** and **Rode, 2001, Kurunç et al., 2005**.

Various programing languages and statistical software were utilized to build SARIMA models such as Matlab language, SAS and Stata software. To date, R software is rarely used in modeling processes of water management applications. R is a language and environment similar to the S language which is used for statistical applications. R offers a variation of statistical techniques that can be utilized for linear and nonlinear modelling, statistical tests and time series analysis. Unlike the other statistical software and environments, R software is a free software which gives the advantage of being an open source in the modeling methodology.

In the present study, Time series analysis of monthly fluoride content was carried on using R software. Also, R software was used to code and construct a SARIMA model to forecast monthly fluoride concentration in Tigris River.

2. MATERIALS AND METHOD

2.1 Study Area and Data Collection

The study area in this work is Tigris River in Baghdad City. This river is considered the main source for Baghdad City water supply systems. Sample collection locations were chosen to represent water quality in the main flow stream, including upstream, downstream and between at different water treatment plant intakes. Al-Karakh and East Tigris WTP intakes represent the upstream flow locations, Al-Wathbah and AL-Karamah WTP intakes represent in between upstream and downstream location, and Al-Rashid and Al-Wahda WTP intake represent downstream locations. The data used in this study were provided from Baghdad Mayoralty (Amanat Baghdad)–Water office which represents the fluoride content at the selected locations which were measured monthly according to the standard methods for water and wastewater examination , **APHA et al., 2005**. Monthly time series data of fluoride concentration from 2004 to 2014 was used in this study.

3. MATHEMATICAL STRUCTURE OF SARIMA MODEL

This model assumes that the predicted value of a variable is a linear function of multi previous observations with random errors, **Box et al., 2011**. The model structure is identified as the below expression:

ARIMA $(p, d, q)(P, D, Q)_s$

This expression can be broken down into two terms, non-seasonal and seasonal terms and the mathematical formula of a SARIMA is presented in Eq. (1):

$$\varphi_p(B)\emptyset_P P(B^s)\nabla^d\nabla_s^D y_t = \vartheta_q(B)\theta_Q(B^s)e_t \quad (1)$$

Where:

p and φ_p: the order of auto-regressive (AR) and the AR operator of order p, respectively.

P and $\emptyset_P$: the order of seasonal auto-regressive and the seasonal AR parameter of order P, respectively.

B: the backshift operator of y_t.

d and ∇^d: the order of integration and the differencing operator, respectively.

D and ∇_s^D: the order of seasonal integration and the seasonal differencing operator, respectively.

q and ϑ_q: order of moving average and the MA operator of order q, respectively.

Q and θ_Q: order of seasonal moving average (MA) and the seasonal MA parameter of order Q, respectively.

s : The seasonal period.

y_t : The value at time point t.

e_t : The white noise (random walk) of the stochastic model.

To fit predictive equations for fluoride content in the selected locations, a SARIMA model was constructed by proceeding multi steps, model identification, parameter estimation, diagnostic checking and model validation.

3.1 SARIMA Model Development

As mentioned above a SARIMA model was constructed for fluoride concentration at each location. An R software code was developed to construct the SARIMA model. Time series inputs for fluoride concentration from 2004 to 2012 were used to develop the SARIMA model. Data from 2013 to 2014 was used to validate the constructed model. To fit the optimal SARIMA, model parameters, *p, P, d, D, q, Q, s,* were identified. Parameter identification was performed based on examining the autocorrelation (ACF) and partial autocorrelation (PACF) functions of the transformed data. Additionally, it was examined to give the minimum Akaike's information criterion (AIC) and mean squared error (MSE). After identifying the model, the model was diagnosed and verified for being appropriately fitting the series. The diagnostic process was conducted by examining the model residuals based on its ACF, normal quantile-quantile plot (Q-Q plot) and Ljung–Box statistics results (**Box et al., 2011**). Validation of the model was performed by running correlation analysis to determine whether predicted and actual data are significantly different or not.

4. RESULTS AND DISCUSSION:

Fig. 1 shows monthly data variation of fluoride concentration at each location for the study period. The monthly data was examined for its stationary condition. This was achieved by performing Augmented Dickey-Fuller test to decide if transformation of the data is neede , **Harvey, 1990**. The time series input was found to be not stationary (p>0.05). Seasonality could result in nonstationary condition. Mainly, this is possible due to the difference between the average values at some particular times within the seasonal period and the average values at other times. Therefore, the monthly data was differentiated to yield stationary input series with respect to yearly periodicity by seasonally transformation.

4.1 SARIMA model construction

To fit the optimal SARIMA model parameters for fluoride data series, several trails were performed. The model parameters were chosen based on the statistical residual diagnostic test. Time series inputs from 2004 to 2012 were used for model calibration and to find the adequate model that fits fluoride content in each location. Data from 2013 to 2014 were used for model verification. The data was differentiated to overcome seasonal effects.

To identify and find the persistence model structure, the ACF, PACF, MSE and AIC were tested. The best fitted model was chosen to give the least ACF, PACF, MSE and AIC values. The most appropriate SARIMA model included non-sessional and sessional compounds. The model parameters, *p, d, q, P, D, Q, s,* were 2, 0, 0, 0, 1, 1, 12, respectively, for all locations. The

SARIMA (2, 0, 0) $(0, 1, 1)_{12}$ equation was derived from equation 1 based on the selected model compounds. Equation 2 is the SARIMA (2, 0, 0) $(0, 1, 1)_{12}$ equation

$$y_t=\varphi_1 y_{t-1}+\varphi_2 y_{t-2}+y_{t-12}-\varphi_1 y_{t-13}- \quad \varphi_2 y_{t-14}+e_t+\theta_1 e_{t-12} \qquad (2)$$

To compute the model parameters, the computational steps proposed by **, Box** and **Jenkins, 1976**. were followed. The residuals of the computed parameters were diagnostic to make sure that the adequate model was selected for their independence, constant variance and normality. **Fig. 2** shows the results of the statistical residual diagnostic test for the selected model of the monthly fluoride content collected at the East Tigris WTP intake. This test showed similar results for the other locations. According to, **El-Din** and **Smith, 2002**, to estimate a decent prediction model, the residuals of the fitted model should satisfy the white noise method requirements which should be uncorrelated and normally distributed around a zero mean. The statistic test showed that the ACF values of the residuals and *p* value were adequately distributed within confidence limits (98%) and no significant "spikes" among 35 lags which confirms the normality of the residuals. Also, the results exhibited no significant correlation between the residuals of the fluoride content at each location. Moreover, the normal quantile-quantile plot (Q-Q plot) exhibits that the residuals were laying closely on the theoretical line, which obviously supports the normality of the residuals. The diagnostic test showed that the selected SARIMA model is appropriately fitting the time series data.

Based on the above analyses, the model parameters were selected. **Table 1** demonstrates the sessional compounds as well as the MSE and AIC for the selected model at each location. The results exhibit that the fitted model for fluoride content at Al-Wathbah WTP intake has the highest MSE among the other location. The forecasting equation for each location can be determined by substituting the corresponding computed parameters to each location from **Table 1** in Eq. (2).

4.2 Monthly Fluoride Content Prediction and Verification

Using the fitted SARIMA model, monthly fluoride concentration at each location was predicted for the period from 2013 to 2016 as shown in **Fig. 3**. Predicted data from 2013 to 2014 was compared to the actual data to verify the fitted model. **Fig. 4** shows the correlation between the actual data and the predicted data at each location. The correlation coefficient between the actual and predicted values for fluoride concentration at Al-Karakh, East Tigris, Al-Wathbah, AL-Karamah, Al-Rashid and Al-Wahda WTP intakes was 0.93, 0.82, 0.86, 0.90, 0.83 and 0.89, respectively. These values are acceptable in common model applications, **Faruk, 2010**, **Zhang, 2003**, **Nourani et al., 2011**. Model verification results state that the model forecasting outputs reasonably estimated the actual monthly fluoride content in the selected locations.

5. CONCLUSIONS

In the last few decades, time series analysis and prediction have been considered a dynamic study area. Researchers have never stopped to improve the accuracy efficiency of prediction models of water quality. Using R software, an effective SARIMA model was

developed for monthly fluoride concentration in Tigris River. The model validation demonstrated that the fitted model is adequately forecasting monthly fluoride content. The outcomes of this study are necessary in the environmental applications especially with the raised concerns regarding water bodies impairment phenomenon.

REFERENCES

- APHA, AWWA and WEF, 2005, *Standard methods for the examination of water and wastewater*, American Public Heath Association, Washington, DC.

- Box, G.E. and Jenkins, G.M., 1976, *Time series analysis: forecasting and control*, revised ed, Holden-Day.

- Box, G.E., Jenkins, G.M. and Reinsel, G.C., 2011, *Time series analysis: forecasting and control*, John Wiley & Sons.

- El-Din, A.G. and Smith, D.W., 2002, *A combined transfer-function noise model to predict the dynamic behavior of a full-scale primary sedimentation tank*, Water Research 36(15), 3747-3764.

- Faruk, D.Ö., 2010, *A hybrid neural network and ARIMA model for water quality time series prediction*, Engineering Applications of Artificial Intelligence 23(4), 586-594.

- Harvey, A.C., 1990, *Forecasting, structural time series models and the Kalman filter*, Cambridge university press.

- Hillmer, S.C. and Tiao, G.C.,1982, *An ARIMA-model-based approach to seasonal adjustment*, Journal of the American Statistical Association 77(377), 63-70.

- Hirsch, R.M., Slack, J.R. and Smith, R.A., 1982, *Techniques of trend analysis for monthly water quality data*, Water resources research 18(1), 107-121.

- Kurunç, A., Yürekli, K. and Çevik, O., 2005, *Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilırmak River, Turkey*, Environmental Modelling & Software 20(9), 1195-1200.

- Lehmann, A. and Rode, M., 2001, *Long-term behaviour and cross-correlation water quality analysis of the river Elbe, Germany*, Water Research 35(9), 2153-2160.

- Nourani, V., Kisi, Ö. and Komasi, M., 2011, *Two hybrid Artificial Intelligence approaches for modeling rainfall–runoff process*, Journal of Hydrology 402(1–2), 41-59.

- Sowell, F., 1992, *Modeling long-run behavior with the fractional ARIMA model*,

Journal of Monetary Economics 29(2), 277-302.

- Williams, B.M. and Hoel, L.A., 1999, *Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process*.

- Williams, B.M. and Hoel, L.A., 2003, *Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results*. Journal of transportation engineering 129(6), 664-672.

- Zhang, G.P., 2003, *Time series forecasting using a hybrid ARIMA and neural network model*, Neurocomputing 50, 159-175.

Table 1. Summary of the statistical parameters of the selected SARIMA model fitted to fluoride concentration at all locations.

Plant	φ_1	φ_2	θ_1	AIC	MSE
AL_KARAKH	0.4974	0.3788	-1	-7.0644	0.000261
EAST TIGRIS	0.4413	0.327	-0.9999	-6.5953	0.00041
AL-WATHBAH	0.4458	0.2069	-1	-4.8524	0.00235
AL-KARAMAH	0.461	-0.0879	-0.7834	-6.0299	0.00074
AL_RASHID	0.2707	0.0539	-0.7947	-6.0299	0.0013
AL_WAHDA	0.3943	-0.0758	-0.775	-6.1358	0.00082

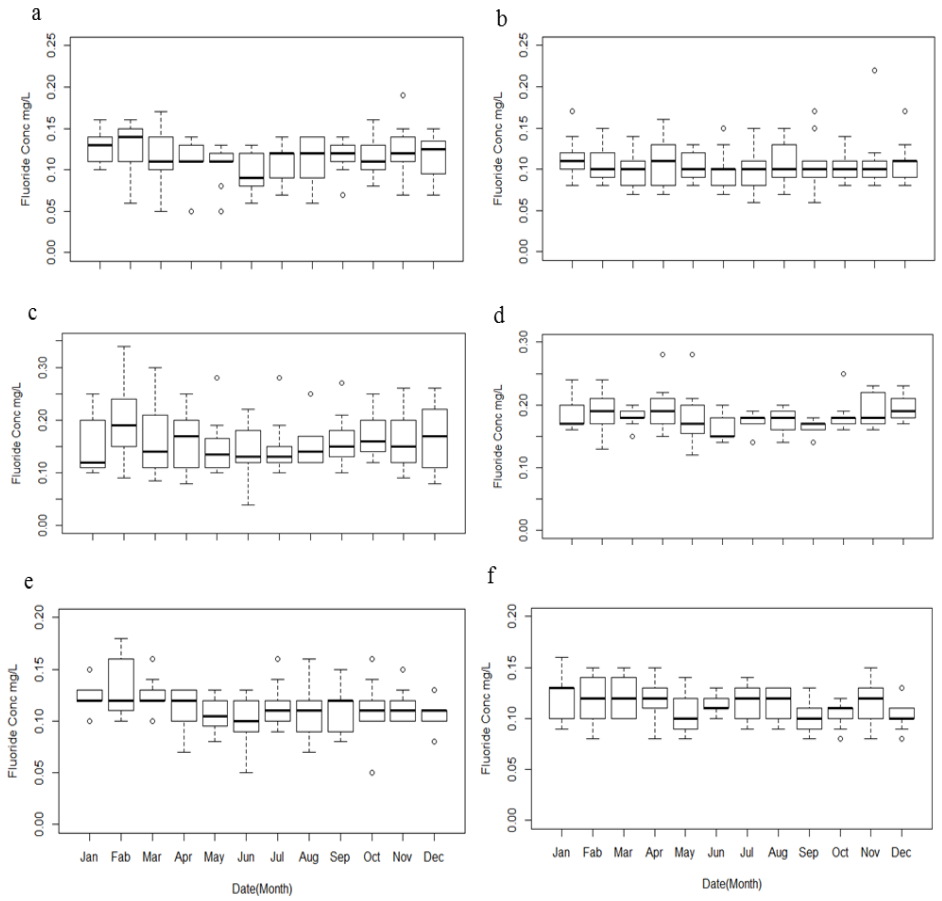

Figure 1. Box plots of monthly fluoride concentration data from 2004 to 2014 at different WTP intakes. a) Al-Karakh WTP intake. b) East Tigris WTP intake. c) Al-Wathbah WTP intake. d) Al-Karamah WTP intake. e) AL-Rashid WTP intake. f) Al-Wahda WTP intake.

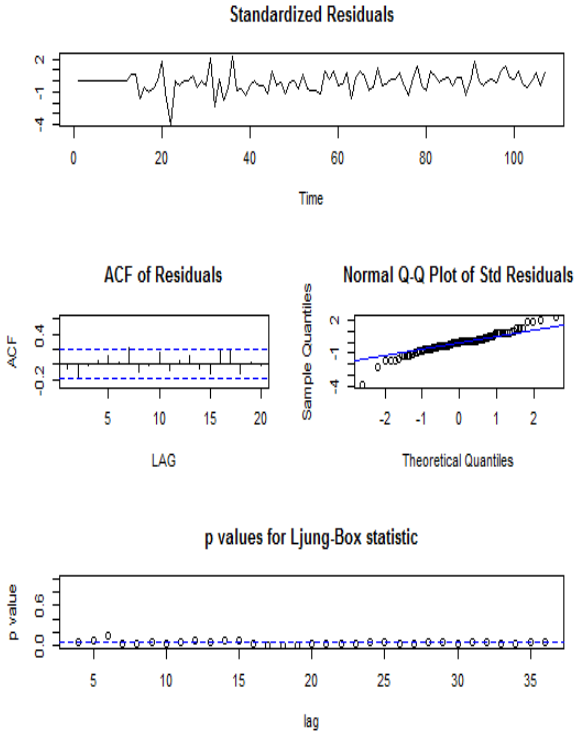

Figure 2. The residual statistical results of the fitted model for monthly fluoride content at the East Tigris WTP intake.

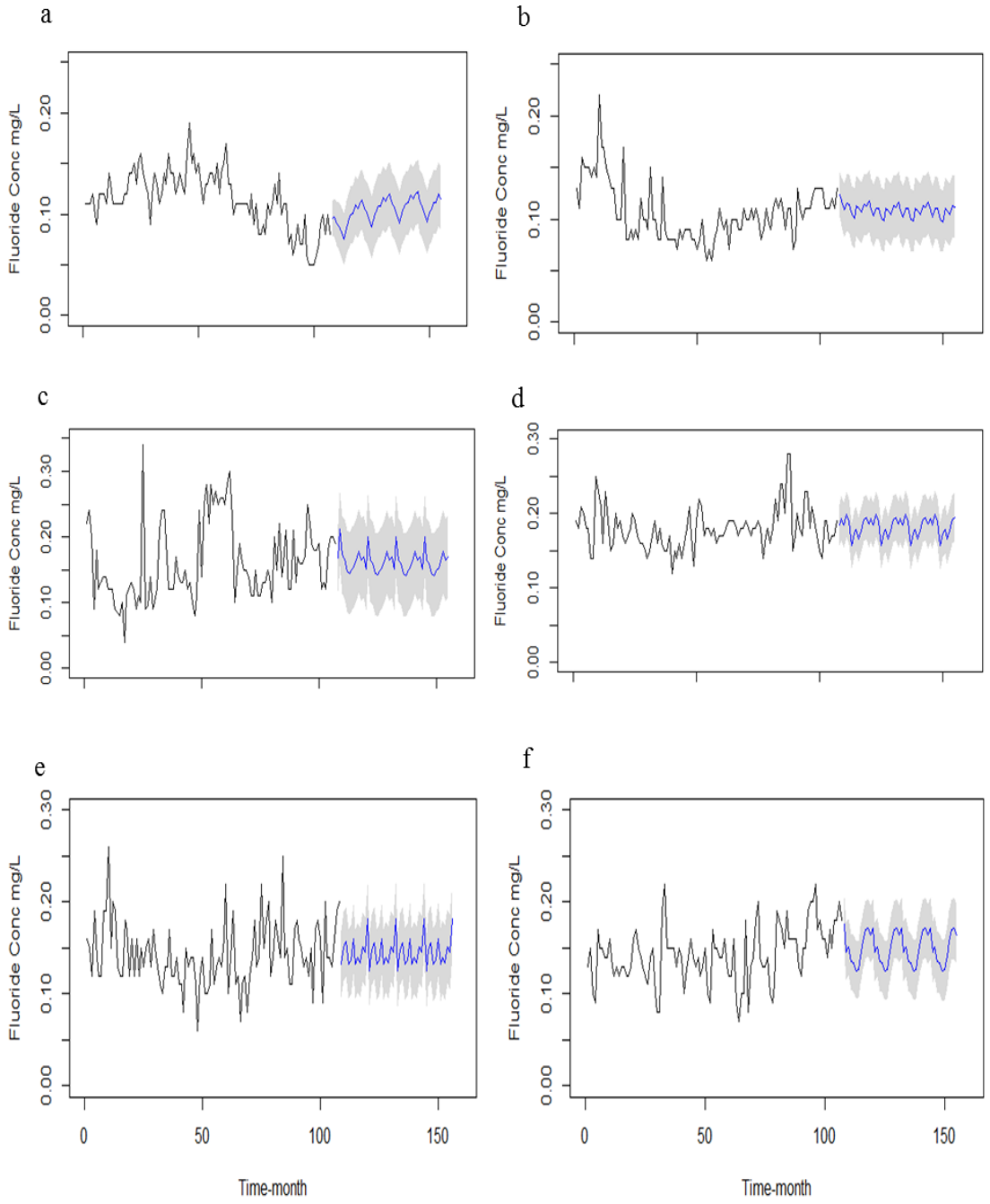

Figure 3. Prediction of monthly fluoride content at various locations in Tigris River. Black solid line is the actual data from 2004 to 2012 used to calibrate the model. Blue solid line is the predicted data for the period from 2013 to 2016. Gray areas shows the upper and lower limits of the predicted values based on 95% confidence intervals. a) Al-Karakh WTP intake. b) East Tigris WTP intake. c) Al-Wathbah WTP intake. d) Al-Karamah WTP intake. e) AL-Rashid WTP intake. f) Al-Wahda WTP intake.

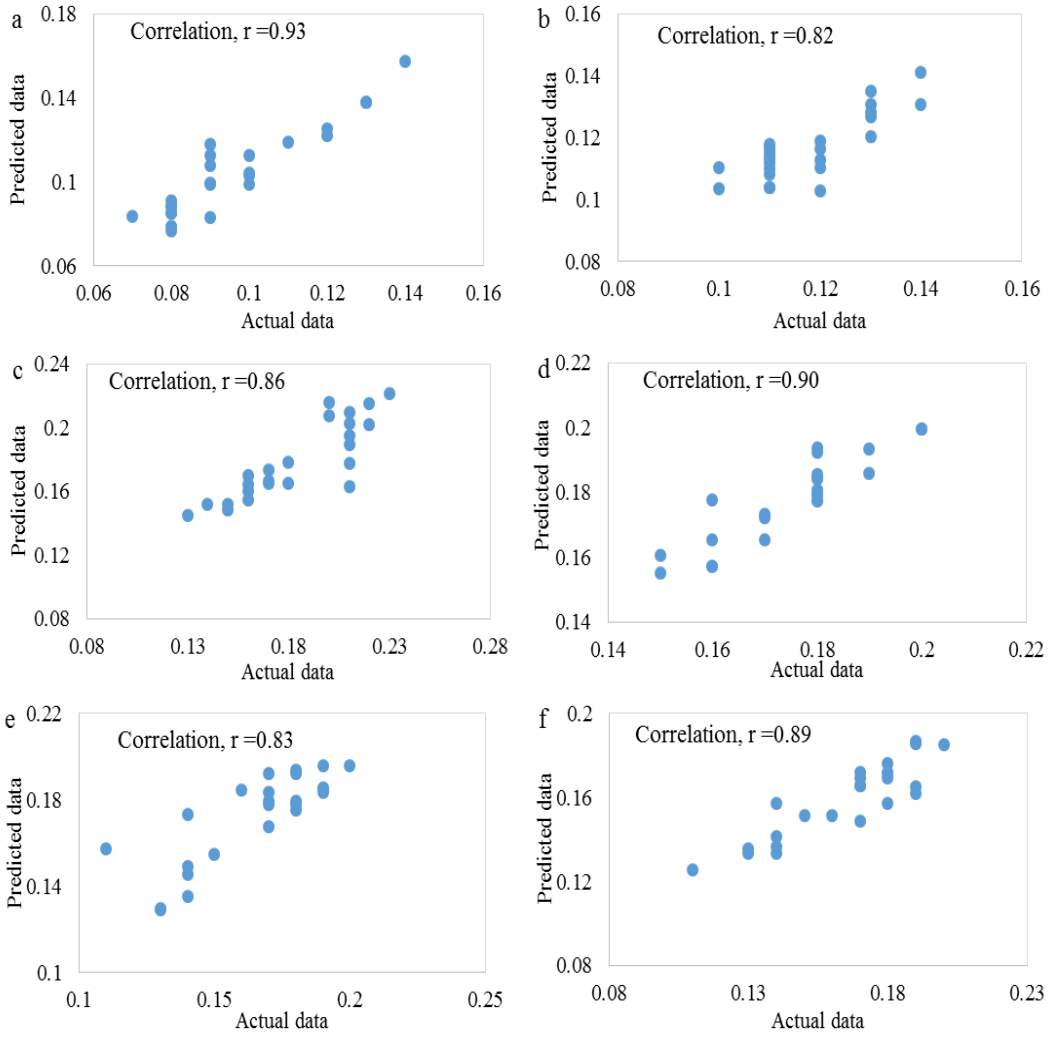

Figure 4. Actual versus predicted fluoride concentration data for correlation analysis at the selected locations. a) Al-Karakh WTP intake. b) East Tigris WTP intake. c) Al-Wathbah WTP intake. d) Al-Karamah WTP intake. e) AL-Rashid WTP intake. f) Al-Wahda WTP intake.