Performance Evaluation of Al-Karkh Water Treatment Plant Using Model-driven and Data-Driven Models

Basim Hussein Khudhair¹, Awatif Soaded Alsaqqar² and Rehab Karim Jbbar³

¹University of Baghdad/ Engineering College
²Uruk University/Engineering College
³Ministry of water resources/groundwater authority

E-mail: obaidy@coeng.uobaghdad.edu.iq

Abstract. There is a great operational risk to control the day-to-day management in water treatment plants, so water companies are looking for solutions to predict how the treatment processes may be improved due to the increased pressure to remain competitive. This study focused on the mathematical modeling of water treatment processes with the primary motivation to provide tools that can be used to predict the performance of the treatment to enable better control of uncertainty and risk. This research included choosing the most important variables affecting quality standards using the correlation test. According to this test, it was found that the important parameters of raw water: Total Hardness, Calcium, Magnesium, Total Solids, Nitrite, Nitrates, Ammonia, and Silica are to be used to construct the specific model, while pH, Fluoride, Aluminium, Nitrite, Ammonia, Silica, and Orthophosphate of the treated water were eliminated from the analysis. For modeling the coagulation and flocculation process temperature, Alkalinity and pH of raw water were the depended variables of the model. As for the modeling process turbidity of the treated water was used as the output variable. In general, the linear models including model-driven type, (Multivariate multiple regression, MMR and Multiple linear regression, MLR) have slightly higher prediction efficiencies than the, data-driven type (artificial neural network, ANNM). The coefficients of determination (R²) reached 66 to 85% for the MMR and MLR models and 65 to 81% for the ANN models.

Keywords: Operational performance evaluation, water treatment plant, multiple linear regression, multivariate multiple regression, artificial neural network model, model-driven and data-driven modeling types.

1. Introduction
In general, the quality of drinking water causes many important problems if it contains a wide range of chemical and microbial components that affect human health. Such components detected in treated water are often expensive, complex, and time consuming in water treatment processes [1]. Moreover, a resource should focus on improving the water supply system if a problem arises, such as poor water quality that can be economically harmful, and therefore, water quality management and treatment need to be improved to provide safe drinking water economically. In order to meet these challenges, regular raw water assessments, operational monitoring issues, and treatment processes are required, and therefore, water processes modeling can help to improve water treatment, Juntunen et al. in 2012, studied the application of a water treatment process model using linear and nonlinear modeling techniques. They concluded that residual aluminum and turbidity of the treated water were the output variables of the tested model and the procedure of data analysis provided an efficient mean of modeling the water treatment process [2]. Solaimany et al. applied a reliable forecast model (artificial neural network, ANN) to predict the quality of influent water to Sanandaj water treatment plant (WTP) to form a base for controlling the operational processes in this plant. The chosen model provided an effective diagnostic and analysis tool to simulate and understand the non-linear behavior of raw water properties affecting the correlation coefficient, (r) that reached 0.93 between the observed and predicted output variables [3].
Mariana et al. studied various optimization models to assess the minimum efficiency of sewage treatment plants located in the same watershed. The results indicated that the models showed satisfactory performance where equity measures were in the objective function and resulted in very close values in the efficiencies of the plants [4]. Tang et al. studied the different types of microorganisms in activated sludge (AS) which provided the key role for rapid assessment on the performance of sewage treatment plants performance [5].

Kandris et al. employed data-driven algorithms as inverse process models for the coagulation treatment process in order to specify the optimum coagulant dose in Simbirizzi WTP in Italy and Aposelemi WTP in Crete. The inverse process model considered all available process inputs, the desired output process (treated water turbidity) and searched for the optimal process control parameter, i.e. the coagulant dose, operational constraints and standards of treated water quality. Two candidate data-driven algorithms were trained to capture the complex and nonlinear relationships between physical, chemical and operational parameters of coagulation: random forests (RFs) and Gaussian Process regression (GPR) models. Results from the application of these models indicated that a reduction of chemical costs equal to 6% and 8% would be probable for the two case studies respectively. He underscored that the large quantity of passive data that amassed daily during the operation of WTPs can be turned into actionable intelligence that supports decision-making and enhances adaptive planning for water utility operators [6]. The main aim of this research is to study the feasibility of using; Model-driven and Data-Driven models to simulate the performance of water treatment plants to increase process stability and optimize the usage of resources, so, Al-Karkh Water Treatment Plant was chosen to be the case study in this research.

2. Materials and Methods
The following sections give the general case study description, preliminary data analysis, and research methodology to simulate the performance of Al-Karkh Water Treatment Plant (WTP) using Model-driven and Data-Driven models.

2.1. Case study description
Al-Karkh water treatment plant (WTP) in Baghdad City is designed as a conventional WTP, which consists of sedimentation, coagulation and flocculation, filtration and disinfection units designed for impurities removal of different sizes and types. The data of water quality parameters available in this research included twenty-one physical and chemical parameters for raw and treated water: temperature, turbidity, alkalinity, total hardness, calcium, chloride, magnesium, pH, color, conductivity, sulfate, total solids, suspended solids, iron, fluoride, aluminum, nitrite, nitrate, ammonia, silica and orthophosphate. The water quality dataset in the research was provided from Baghdad Mayorality over the period January 2016 to September 2020.

2.2. Preliminary data analysis
The formulation of the engineering model requires input data to make predictions. In this research, available water quality parameters were used. However, not all of these data provide useful information therefore; it could be useless to include all of them in a model thus saving money and time spent on data collection. Several statistical techniques are available to study the relationships between these parameters; as these parameters are to be classified into categories and scale variables. When the variables are scaled, then the correlation test could be performed. This test can measure the linear relation between two scale variables and to estimate the correlation coefficient, where Pearson’s r is used. The values of -1, 0 and 1 indicate strong negative correlation, no correlation and strong positive correlation, respectively. Generally, weak, moderate and strong linear association would take values between (0-0.5), (0.5-0.85) and (0.85-1) respectively [7]. To specify whether the linear association is statistically significant, the t-test can be used by considering the p-value. The linear relationships are
considered statistically significant, if a critical value of 0.05 (for confidence level of 95%) is greater than the p-value of the t-statistic [8].

2.3. Research methodology

In this research, two types of model-based modeling techniques and data type models were used to evaluate the performance of Al-Karkh WTP. The first step was to model raw water quality parameters to ensure the performance evaluation of the coagulation and flocculation processes. Since water quality parameters such as temperature, alkalinity, and pH may dictate the effectiveness of the coagulation process, they have been used as the output variables in the models used. pH influences the destabilization process as it controls both the solubility of the coagulant. Temperature affects the viscosity of water, and therefore also affects the coagulation process, where low-temperature water can reduce hydrolysis [9]. The second step of the evaluation was the modeling of treated water parameters to ensure the efficiency of the filtration process. Turbidity was chosen be the output variable as it decreases the clarity of the treated water (in the physical sense caused by colloidal or suspended particles). It is used as an efficiency parameter of drinking water from the coagulation and filtration processes [2]. Fig. 1 represents the stages that were performed to establish the selected models.

![Diagram of the stages for the development of the selected models.](image)

**Figure 1.** Schematic diagram of the stages for the development of the selected models.

3. Concept of Modeling Techniques

As mentioned above, two types of modeling techniques were used in this study, which are model-driven and data-driven. Dasu and Johnson considered the statistical models and the artificial intelligence-based models as a model-driven type and data-driven type respectively [10]. In the first, the model structure is
often decided by the experts, while in the second, the model structure is decided by the available data. Modeling techniques which were applied in water treatment plants are multivariate multiple regression and multiple linear regression as model-driven type and artificial neural network as data-driven type. The calibrating process of these models is done by using the statistical software SPSS 23.

### 3.1. Multivariate multiple regression (MMR)

To model the linear relationship between two or more independent variables with two or more dependent variables, the MMR is used. This statistical test is multivariate as there is more than one dependent variable and more than one independent variable [11]. A linear model with a single dependent variable $Y$ takes the form (Eq. 1):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon$$

Where: $\beta_0, \beta_1, \ldots, \beta_n$ are to be estimated, $X_1, \ldots X_n$ are independent variables (water quality parameters) and $\epsilon$ is a random error term. For the MMR with $m$ dependent variables and $n$ independent variables the complete model becomes:

\[
\begin{align*}
Y_1 &= \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon_2 \\
\vdots  \\
Y_m &= \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \epsilon_m
\end{align*}
\]

The least square technique is commonly used to calibrate the linear model since water quality parameters such as temperature, alkalinity, and pH may dictate the effectiveness of the coagulation process, they have been used as output variables in this model [12].

### 3.2. Multiple linear regression (MLR)

Multiple linear regression models represent the relationship between two or more independent variables with a dependent variable by fitting a linear equation to the available data. In this model the level of variation between the variables could be determined. Equation (1) defines the trend of MLR as $Y$ the dependent variable and $X$ the independent variable [13, 14]. Multiple linear regression is used to predict one scale dependent variable from two or more independent variables by fitting a linear equation to the available data, this is considered a simple regression process. The level of variation between the variables is determined by this test [15].

### 3.3. Artificial neural network model (ANNM)

To predict the output variables from the input variables, neural networks that use a method similar to the human nervous system process can be used. The human brain is similar in two ways; through a learning process, the network acquires knowledge and the gained knowledge is stored in interneuron connection strengths known as connection weights, and the structure of the ANNM is defined by the available data and does not require any assumptions [16]. Generally, the models simulate non-linear relationship within water treatment processes and are simulated by artificial neurons which are located in different layers and connected to each other by connection weights [7]. The model is basically composed of three layers: input, hidden and output layers respectively as shown in Fig. 2.

a) The input layer contains the independent variables ($X_i$).

b) The hidden layer contains unobservable nodes which receive signals from the input layer. The signals represent the summation of the inputs ($X_i$) multiplied by their connection weights. Then, these nodes use pre-defined mathematical functions (activation functions) to produce output signals.

c) The output layer contains neurons that define the predicted output variables after receiving signals from the hidden layer.

Building the model is done by dividing the data into: (training), (testing) and (holdout), which means (set of data used to train the model), (another part of dataset used to prevent overtraining by tracking errors during training) and (data not used in the model building process so its error gives an honest estimate of the model predictive ability) respectively. The calibration of this model means finding the
connection weights and the number of hidden neurons by minimizing the differences between the predicted and actual output values [17].

Figure 2. Structure of the neural networks.

4. Results and Discussions

4.1. Results of the preliminary analysis
The correlation test was applied to the raw and treated water parameters respectively to study the relationships between input and output variables.

4.1.1. Raw water parameters. The results of the correlation test are given in Table 1, which shows the Pearson’s correlation coefficient between raw water depended variables: temperature, alkalinity and pH and the remaining water quality parameters. It appears from this table that the parameters T. Hardness, calcium, magnesium, total solids, nitrite, nitrate, ammonia and silica have statistically significant correlation (i.e. p-values associated with the correlation smaller than 0.05 or 0.01), therefore they were used to build the models, while the remaining parameters were discarded from the analysis [18].

| Parameter          | Temperature | Alkalinity | pH    |
|--------------------|-------------|------------|-------|
| Turbidity          | -0.148      | 0.23       | 0.026 |
| T. Hardness        | -0.606 b    | 0.538 b    | 0.344 |
| Calcium            | -0.681 b    | 0.697 b    | 0.384 |
| Chloride           | -0.347      | 0.258      | 0.103 |
| Magnesium          | -0.488 a    | 0.319      | 0.293 |
| Conductivity       | -0.342      | 0.269      | 0.09  |
| Sulfate            | -0.396      | 0.201      | 0.138 |
| Total Solids       | -0.499 a    | 0.448 a    | 0.274 |
| Suspended Solids   | 0.059       | 0.187      | -0.022|
| Iron               | 0.042       | 0.199      | -0.062|
| Fluoride           | -0.331 c    | -0.005 c   | 0.242 |
| Aluminum           | c           | c          | c     |
| Nitrite            | -0.592 b    | 0.626 b    | 0.349 |
| Nitrate            | -0.460 a    | 0.763 b    | 0.558 b|
| Ammonia            | 0.26        | -0.559 b   | -0.229|
| Silica             | 0.442 a     | 0.006      | -0.531 b|
| Orthophosphate     | 0.00        | -0.369     | -0.1  |

a Correlation is statistically significant at 0.05 level from t-test.
b Correlation is statistically significant at 0.01 level from t-test.
c Cannot be computed because at least one of the variables is constant.
4.1.2 Treated water parameters. Table 2 shows the correlation coefficients between treated water turbidity as the output variable from the filtration process and the remaining parameters as inputs. It was found that most of the parameters were statistically significant except pH, fluoride, aluminum, nitrite, nitrate, ammonia, silica and orthophosphate that were neglected.

| Parameter   | Turbidity | Parameter   | Turbidity |
|-------------|-----------|-------------|-----------|
| Temperature | -0.649 b  | Total Solids | 0.58 b    |
| Alkalinity  | 0.636 b   | Iron        | 0.676 b   |
| T. Hardness | 0.652 b   | Fluoride    | 0.283     |
| Calcium     | 0.734 b   | Aluminum    | -0.393    |
| Chloride    | 0.418 a   | Nitrite     | c         |
| Magnesium   | 0.497 a   | Nitrate     | 0.223     |
| pH          | 0.212     | Ammonia     | c         |
| Conductivity| 0.429 a   | Silica      | -0.304    |
| Sulfate     | 0.471 a   | Orthophosphate | 0.064   |

a Correlation is statistically significant at 0.05 level from t-test.
b Correlation is statistically significant at 0.01 level from t-test.
c Cannot be computed because at least one of the variables is constant.

4.2 The calibration of the models

4.2.1 Multivariate multiple regression (MMR). The input variables that were selected for raw water are shown in Table 3 with their coefficients are introduced in the model equations for predicting the dependent parameters. Three equations were formed using eq. 1 to predict the outputs temperature, alkalinity and pH as the significance of the inputs in each equation is assessed by p-value (i.e. sig.) that appears in Table 3. As an example, silica contributes in the prediction of temperature since it has a p-value less than 0.05 [18].

4.2.2 Multiple linear regression (MLR)
The calibrating process of MLR is to find the model coefficients (i.e. β), which are shown in Table 4 as unstandardized coefficients that can be easily used to predict output variables. In this model, treated water turbidity was the output variable and ten water quality parameters were the input variables (which passed the correlation test). The equation for this model is then written using the coefficients in Table 4 to develop the prediction model. In addition, Table 4 displays the standardized coefficients (B), which is used to compare the relative strength of the various inputs within the MLR model. Instead of water parameters units, these coefficients are measured in standard deviations, so they can be compared to one another. Temperature has the smallest B coefficient of -0.099 and T. Hardness has the largest 2.733. Thus, with the other inputs held constant, an increase in the standard deviation by one in temperature leads to a 0.099 standard deviation decrease in the predicted turbidity. While, an increase in the standard deviation by one in T. Hardness leads to a 2.733 standard deviation increase in turbidity [16].

To study the significance of the overall model, the p-value of the F-test was used as shown in Table 5. The model is statistically significant as it has a p-value at the significant level of 0.01 greater than the significant level 0.05 [8].
Table 3: Parameter estimates of the MMR model for raw water.

| Dependent Variable | Parameter     | $\beta$ | Std. Error | t     | Sig. |
|--------------------|---------------|---------|------------|-------|------|
| Temperature        | Intercept     | 47.400  | 13.429     | 3.530 | .003 |
|                    | T. Hardness   | .645    | .499       | 1.294 | .215 |
|                    | Calcium       | -2.793  | 1.407      | -1.985| .066 |
|                    | Magnesium     | -2.710  | 1.960      | -1.383| .187 |
|                    | Total Solids  | .085    | .048       | 1.769 | .097 |
|                    | Nitrite       | -80.215 | 643.569    | -0.125| .902 |
|                    | Nitrate       | -.138   | 4.332      | -0.032| .975 |
|                    | Ammonia       | -54.331 | 80.286     | -0.677| .509 |
|                    | Silica        | 3.317   | 1.334      | 2.486 | .025 |
| Alkalinity         | Intercept     | 63.745  | 18.252     | 3.492 | .003 |
|                    | T. Hardness   | .001    | .678       | .001  | .999 |
|                    | Calcium       | .341    | 1.912      | .178  | .861 |
|                    | Magnesium     | -.049   | 2.664      | -.018 | .986 |
|                    | Total Solids  | .022    | .065       | .341  | .738 |
|                    | Nitrite       | 1047.369| 874.685    | 1.197 | .250 |
|                    | Nitrate       | 20.449  | 5.888      | 3.473 | .003 |
|                    | Ammonia       | -152.190| 109.118    | -1.395| .183 |
|                    | Silica        | 2.220   | 1.813      | 1.224 | .240 |
| pH                 | Intercept     | 8.021   | .241       | 33.286| .000 |
|                    | T. Hardness   | -.009   | .009       | -1.060| .306 |
|                    | Calcium       | .020    | .025       | .800  | .436 |
|                    | Magnesium     | .054    | .035       | 1.544 | .144 |
|                    | Total Solids  | .000    | .001       | -0.473| .643 |
|                    | Nitrite       | 2.273   | 11.548     | .197  | .847 |
|                    | Nitrate       | .199    | .078       | 2.554 | .022 |
|                    | Ammonia       | -1.170  | 1.441      | -0.812| .429 |
|                    | Silica        | -.062   | .024       | -2.591| .020 |

Table 4: Coefficients of the unstandardized and standardized regression function for treated water.

| Model          | Unstandardized Coefficients | Standardized Coefficients | t     | Sig. |
|----------------|-----------------------------|---------------------------|-------|------|
| (Constant)     | -7.338                      | 4.395                     | -1.670| .119 |
| Temperature    | -.012                       | .042                      | -.099 | -.301| .769 |
| Alkalinity     | .011                        | .022                      | .163  | .503 | .623 |
| T. Hardness    | .057                        | .087                      | 2.733 | .659 | .522 |
| Calcium        | -.124                       | .219                      | -1.308| -.569| .579 |
| Chloride       | -.117                       | .091                      | -1.723| -.128| .223 |
| Magnesium      | -.181                       | .371                      | -.987 | -.489| .633 |
| Conductivity   | .009                        | .013                      | 1.153 | .653 | .525 |
| Sulfate        | -.009                       | .015                      | -.524 | -.586| .568 |
| Total Solids   | .007                        | .009                      | .652  | .750 | .467 |
| Iron           | 24.910                      | 7.218                     | 3.451 | .004 |
Table 5: Results of the F-test.

| Model  | Sum of Squares | df  | Mean Square | F      | Sig. |
|--------|----------------|-----|-------------|--------|------|
| 1      | Regression     | 8.025 | 10          | .802   | 6.386 | .001b |
|        | Residual       | 1.634 | 13          | .126   |       |      |
|        | Total          | 9.658 | 23          |        |      |      |

a. Dependent Variable: Turbidity  
b. Predictors: (Constant), Iron, Alkalinity, Conductivity, Temperature, Magnesium, Total Solids, Calcium, Sulfate, Chloride, T. Hardness

4.2.3 Artificial neural network model (ANNM). In this study, the developed procedure for training ANNM is as follows:

a. **Partition dataset**: To find the best partition of the dataset for (training, testing and holdout) samples, a number of trials for partitioning dataset were performed depending on the sum-of-squares error of the testing and training samples. For raw water parameters, eight input variables were used: T. Hardness, Calcium, Magnesium, Total Solids, Nitrite, Nitrate, Ammonia, and Silica to predict three output raw water variables: Temperature, Alkalinity, and pH. Fig. 3 shows that the best partition was (37.5, 30, 32.5) % as the sum-of-squares error was the lowest. As for the treated water parameters, ten input variables were used: Temperature, Alkalinity, T. Hardness, Calcium, Chloride, Magnesium, Conductivity, Sulfate, Total Solids, Iron and Silica where the output treated water variable was the turbidity. Fig. 4 shows that two partitions have the lowest sum-of-squares error (50, 30, 20) % and (50, 35, 15) % and was equal for training and testing samples, but (50, 30, 20) % was considered in model development as it has the greater coefficient of determination ($R^2$) 0.814.

b. **Activation function**: In this step, the effect of changing the activation functions of the output neurons on the error function was examined after obtaining the best partition dataset. The activation functions for scale output variables: identify hyperbolic tangent and sigmoid functions that are used in the software package SPSS 23. Fig. 5 and 6 show that the sigmoid function has the lowest error for both raw and treated water parameters and was used to develop the AAN models.

![Figure 3. Sum-of-squares error with different partitions for raw water parameters.](image-url)
Figure 4. Sum-of-squares error with different partitions for treated water parameters.

Figure 5. Sum-of-squares error with different activation functions for raw water parameters.

Figure 6. Sum-of-squares error with different activation functions for treated water parameters.
c. **Neurons number in the hidden layer:** This step involves selecting the number of neurons in the hidden layer that give the lowest error for the training and testing samples. High numbers in the ANNM lead to lose the generalization ability, while few neurons may lead to incorrect predictions [7]. It can be observed that eight and six neurons are the best numbers for raw and treated water parameters respectively as shown in Figs. 7 and 8.

![Figure 7. Sum-of-squares error with different hidden neurons number for raw water parameters.](image)

The last step, in the development of the ANNM is to find the connection weights of the network. The optimization algorithm scaled conjugate gradient was used and the standardized method for rescaling input data was selected. The following equations are used to model the non-linear relationship between the output and input data.

\[ C = \sum_{i=1}^{n} X_i W_{i(H-H)} + (W_o(H-H)) \]  
(2)

\[ H_j = \tanh(C) \]  
(3)

\[ H_k = \sum_j H_j W_{j(H-o)} + (W_o(H-o)) \]  
(4)

\[ Y = \frac{1}{1+e^{-H_k}} \]  
(5)
Where: \(n\) number of parameters, \(W_{i(I-H)}\) and \(W_{o(I-H)}\) the connection weights and bias weight between input and hidden layers respectively, \(j\) the number of hidden neurons, \(H_j\) hidden neurons output value, \(H_k\) the input to the output neuron, \(W_{j(H-O)}\) and \(W_{o(H-O)}\) the connection weights and bias weight between hidden and output layers respectively and \(Y\) is the output neuron value [8]. The water quality parameters used in this model were temperature, alkalinity, and pH for the raw water as the input data and the output variable was the turbidity for the treated water.

4.3 Performance Evaluation of the Models
The performance process involves the evaluation of the agreement degree between the available measurements and model predictions. The model is said to have high predictive power, if the agreement is significant, while the model fails with lack of agreement [19]. The commonly used statistical technique for scale output variables is the coefficient of determination (R\(^2\)). In addition, the prediction quality of regression based models is often evaluated using the root mean square error or R\(^2\) [17]. For raw water treatment processes, the results showed that the MMR model, gave R\(^2\) for temperature, alkalinity and pH 0.728, 0.854 and 0.662 respectively. While from the ANN model R\(^2\) was 0.652, 0.811 and 0.690 for temperature, alkalinity and pH respectively. As for the treated water from the treatment processes, the MLR model had R\(^2\) of 0.831 for turbidity as the output variable (as shown in Table 6). Finally, the R\(^2\) value of the ANNM model to predicted turbidity for the treated water was 0.772.

### Table 6: Coefficient of determination for the MLR model for treated water.

| Model Summary | R | R\(^2\) | Adjusted R\(^2\) | Std. Error of the Estimate | R\(^2\) Change | F Change | df1 | df2 | Sig. F Change |
|---------------|---|--------|----------------|--------------------------|---------------|---------|-----|-----|--------------|
| 1             | .912\(^a\) | .831 | .701 | .3545 | .831 | 6.386 | 10 | 13 | .001 |

\(^a\) Predictors: (Constant), Iron, Alkalinity, Conductivity, Temperature, Magnesium, Total Solids, Calcium, Sulfate, Chloride, T. Hardness

5. Conclusion
In this research, both linear (MMR and MLR) and nonlinear (ANN) modeling techniques were used to simulate the performance of Al-Karkh water treatment plant, for both raw and treated water parameters affecting the coagulation & flocculation and the filtration processes.

1) For modeling the coagulation and flocculation process temperature, alkalinity and pH of raw water were the depended variables of the MMR and ANN models.
2) As for modeling, the filtration process turbidity of the treated water was used in the MMR and ANN models as an output variable.
3) The linear models, MMR and MLR have slightly higher prediction efficiencies than the ANN model. The coefficients of determination (R\(^2\)) for raw water reached 66 to 85% for the MMR model and 65 to 81% for the ANN model. As for the models used to simulate the treatment processes R\(^2\) was 83% in the MLR model and 77% for the ANN model.

References
[1] World Health Organization, WHO 2006 Guidelines for drinking-water quality, 1, Recommendations, 3rd edition.
[2] Juntunen P, Liukkonen M, Pelo M., Lehtola JM and Hiltunen Y 2012 Modelling of water quality: an application to a water treatment process. applied computational intelligence and soft computing. Article ID 846321, 1-9 p.
[3] Solaimany-Aminabad M, Maleki A and Hadi M 2014 Application of artificial neural network (ANN) for the prediction of water treatment plant influent characteristics. Journal of Advances
in Environmental Health Research, 1(2): 1-12.

[4] Santoro M, Reis MC, J AT D and Mendonça A S F, 2016 Performance evaluation of optimization models in the determination of wastewater treatment efficiencies inside watersheds. RBRH, 21(4):694-706.

[5] Tang Liu S L Maosheng Zheng Q Ch and Jinren Ni 2016 Performance Assessment of Full-Scale Wastewater Treatment Plants Based on Seasonal Variability of Microbial Communities via High-Throughput Sequencing. PLOS ONE, 6: 1-15.

[6] Kandris K., Romas E., Tzimas A and Gavalakis E 2019 Employing data-driven models in the optimization of chemical usage in water treatment plants. 16th International Conference on Environmental Science and Technology Rhodes, Greece, 4-7.

[7] Tran D H 2007 Investigation of deterioration models for stormwater pipe systems. Doctoral dissertation, School of Architectural, Civil and Mechanical Engineering, Faculty of Health, Engineering and Science, Victoria University, Australia.

[8] Montgomery D C, Runger G C and Hubele N F 2004 Engineering Statistics. (Florida:John Wiley & Sons).

[9] Environmental protection agency (EPA) United States. https://iaspub.epa.gov/tdb/pages/treatment/treatmentOverview.do#content 2018.

[10] Dasu T and Johnson T 2003 Exploratory Data Mining and Data Cleaning. (New York:John Wiley and Sons Inc).

[11] Dattalo P 2013 Analysis of Multiple Dependent Variables. Published to Oxford Scholarship Online.

[12] Johnson A R and Wichern W D 2007 Applied Multivariate Statistical Analysis, Sixth Edition. Prentice-Hall.

[13] Parmar K S and Bhardwaj R 2015 River water prediction modeling using neural networks, fuzzy and wavelet coupled model. Water Resour Management, 29(1):17–33.

[14] Izquierdo J, Montalvo I, Tavera M 2008 Optimization in water systems: a PSO approach. Centro Multidisciplinar de Modelación de Fluidos, Universidad Politécnica de Valencia, SpringSim, pp. 239–246.

[15] Abba I S, Hadi J S and Abdullahi J 2017 River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW. 24-25.

[16] IBM_SPSS_Statistics_Core_System_Users_Guide, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf

[17] Kley G and Caradot N 2013 D 1.2 Review of sewer deterioration models. KWB report, project SEMA, Berlin, Germany.

[18] Vuckovic A, Radiivojevic V, Chen A, Popovic D 2002 Automatic recognition of alertness and drowsiness from EEG by an artificial neural network. Medical Engineering & Physics 24, 349–60

[19] Gardner RH, and Urban DL. 2003 Model Validation and Testing: Past Lessons, Present Concerns, Future Prospects. In Canham, C.D., Cole, J.J. & Lauenroth. W.K. (eds.), Models in Ecosystem Science. New Jersey: Princeton University Press.

Acknowledgment
The authors wish to thank the Mayorality of Baghdad/Baghdad Water Department, for the opportunity to obtain data, and to be given the facilities to be in touch with the station manager. The author would like to thank the staff of the sanitary engineering laboratory and the Civil Engineering Department/Engineering College-University of Baghdad for their valuable support to complete this work.