Modeling of Water Quality Parameters using Data-Driven Models, 
A Case Study Abo-Ziriq Marsh in South of Iraq

Mustafa Al-Mukhtar and Fuaad Al-Yaseen
1Department of Civil Engineering, University of Technology, Baghdad, Iraq
2Directorate of Thi Qar Municipality, Nasiriyah, Iraq
42148@student.utechnology.edu.iq, 9647830918992

Abstract: Total Dissolved Solids (TDS) and Electrical Conductivity (EC) are important parameters in determining water quality such as drinking water and agricultural water because it directly affects the concentration of salt in water. High values of these parameters cause low water quality. In addition to that, they play a significant role in hydrous life, effective water resources management and health studies. Thus, it is of critical importance to identify the optimum modeling method that would be capable to capture the behavior of these parameters. The aim of this study was to assess the ability of using three different models of artificial intelligence techniques: Adaptive Neural based Fuzzy Inference System (ANFIS), Artificial Neural Networks (ANNs) and Multiple Regression Model (MLR) to predict and estimate Total Dissolved Solids (TDS) and Electrical Conductivity (EC) in Abu-Ziriq Marsh South of Iraq. As so 84 monthly TDS and EC values from 2009-2018 were used in the evaluation. The collected data was randomly split into 75% for training and 25% for testing. The most effective input parameters to Model TDS and EC were determined based on cross-correlation test. The three performance criteria: Correlation Coefficient (CC), Root Mean Square Error (RMSE) and Nash-Sutcliffe Effectivity coefficient (NSE) were used to evaluate performance criteria of developed models. It was found that Nitrate (NO₃⁻), Calcium (Ca²⁺), Magnesium (Mg²⁺), Total Hardness (T.H), Sulfate (SO₄²⁻) and Chloride (Cl⁻) are the most influential inputs on TDS. While Calcium (Ca²⁺), Magnesium (Mg²⁺), Total Hardness (T.H), Sulfate (SO₄²⁻) and Chloride (Cl⁻) are the most effective on EC. Comparison of the results showed that the three models can satisfactory estimate the total dissolved solids and electrical conductivity but ANFIS model was superior in modelling the two water quality parameters. ANFIS is recommended to be used as a predictive model for TDS and EC in the Iraqi Marshes.

Key words: Total dissolved solids, electrical conductivity, data-driven models, Abu-Ziriq Marsh, water quality parameters, magnesium

INTRODUCTION

Preserving water quality has become an urgent issue, since, it affects human health and hydrous ecosystems. With the continuous increase in population, there is an increasing need for water resources. Contamination of water sources resulting from some natural processes including air inputs or climatic conditions and through human pollutants such as non-treatment of sewage discharge and industrial activities which might add further stress to water quality (Singh et al., 2004). The considered important indicators of water quality are the Electrical Conductivity (EC) and the Total Dissolved Solids (TDS). High values of these parameters cause low water quality because they directly affect the concentration of salt in water. In addition to that, assessment of water quality plays a significant role in hydrous life and water management (Dogan et al., 2009). Although, EC and TDS are measurable parameters, their direct estimations are costly and take a long time. Therefore, convenient, cost-effective, fast and reliable methods are needed for estimations and prediction of EC and (Basant et al., 2010). Recently, the use of data-driven models such as ANN, ANFIS, Gene Expression Programming (GEP) and Support Vector Machines (SVM) have become viable alternative (Firat and Gungor, 2010; Kashani and Dinapashch, 2012; Li et al., 2013; Verma et al., 2013; Wang et al., 2014; Yilmaz and Kaynar, 2011). Recently, artificial intelligence has been used in many private of literature water quality modeling and water management applications (Bowers and Shedrow, 2000; Chen and Liu, 2014; Kim, 2011; Kuo et al., 2007;
developed an adaptive C neuro-fuzzy inference system model to estimate electrical conductivity in groundwater. It was shown that the ANFIS Model outperforms the traditional methods of modelling EC based on TDS in the water. Singh et al. (2009) used two ANNs Models for computing the Dissolved Oxygen (DO) and Biochemical Oxygen Demand (BOD) levels of the Gomti River in India. In their study, 11 parameters were used as input variables and two variables as output at the Gomti River. The result showed that the ANN Model can be used successfully in estimating water quality parameters. Kisi and Ay (2012) used the ANFIS and Radial Basis Neural network (RBN) Models to predict DO values by using different input parameters including discharge, pH and temperature and EC at Fountain Creek Stream-Gauging Station which covers 9 years of daily data and EC at Fountain Creek Stream-Gauging Station which covers 9 years of daily data. The results showed the RBN Model was better than ANFIS Model in the prediction of DO values. Wen et al. (2013) developed ANN Model to estimate the DO values of Heihe River in Northwestern China. The input parameters of the neural network were EC pH, total hardness, Chloride (Cl\(^{-}\)), total hardness, Calcium (Ca\(^{2+}\)), total alkalinity, Nitrate Nitrogen (NO\(_3\)-N) and ammonical Nitrogen (NH\(_4\)-N) with one output DO. The result indicated that the ANN Model can be used successfully to estimate DO concentrations. Orouji et al. (2013) utilized the ANFIS and Genetic Programming (GP) as two data-driven models to predict and simulate water quality parameters (i.e., EC and TDS) of the Astane station in Sefidrood River, Iran. Both models of the data-driven succeeded in determining the water quality parameters. Ghavidel and Montasen (2014) used four artificial intelligence approaches, namely two ANFIS including ANFIS with Grid Partition (ANFIS-GP) and ANFIS with Subtractive Clustering (ANFIS-SC), ANN and Gene Expression Programming (GEP) for the estimation of TDS in the Zarinsroud basin in Northwest of Iran. The result indicated that the GEP can be used successfully over other data-driven models. Edwin et al. explored the ability of ANN to predict dissolved oxygen in Lake Victoria basin, Kenya. Four input variables of temperature, turbidity, pH and EC were used. The data consisted of 113 monthly values for the input variables and output variable from 2009-2013 which were split into training and testing datasets. The results obtained during training and testing revealed that the ANN can be used as a monitoring tool in the prediction of dissolved oxygen. Obviously, there is no specific method attained a universal acceptance in terms of its applicability, therefore, further evaluation is needed based on data specific area.

The main objective of this study is to identify the optimum model which could be of use to model the water quality parameters in Abu Zirig Marsh South of Iraq. Thence, three different algorithms (i.e., ANFIS, ANN and MLR) methods were investigated to model both water quality parameters, i.e., TDS and EC. The study area was selected based on its importance in terms of the amount of inflow water, representing a good example of the ecological system and its role in the Iraqi Marshes revive. The statistical metric, i.e., cross correlation was employed to select the best input parameters with a significant level of 0.01 and 0.05. The models were assessed based on three evaluation criteria which are correlation coefficient, root mean square error, Nash coefficient efficiency.

**MATERIALS AND METHODS**

**Adaptive Neuro-Fuzzy Inference System (ANFIS):**

ANFIS is an advanced feed network containing several layers and analyzes each incoming signal node and performs a specific function. Square node and circle node nodes are used to illustrate different qualities of adaptive learning. To obtain the required input and output attributes, adaptive learning parameters are developed on the basis of gradual learning rules. The ANFIS membership functions are based on the rules and membership functions of the data (Jang, 1993). Essentially, the fuzzy inference system explained here contains two inputs (x\(_1\) and x\(_2\)) and only one output (y\(_1\)). It is assumed that the rule base contains two fuzzy IF-THEN rules of a first-order Sugeno fuzzy (Takagi and Sugeno, 1985):

\[ \text{Rule 1: If} \quad x_1 \quad \text{is} \quad A_1 \quad \text{and} \quad x_2 \quad \text{is} \quad B_1, \quad \text{then} \quad y_1 = f_1 = p_1 x_1 + q_1 x_2 + r_1 \]  

\[ \text{Rule 2: If} \quad x_1 \quad \text{is} \quad A_2 \quad \text{and} \quad x_2 \quad \text{is} \quad B_2, \quad \text{then} \quad y_1 = f_2 = p_2 x_1 + q_2 x_2 + r_1, \]  

Where:

- \( A_1 \) and \( B_1 \) = The fuzzy sets
- \( p_1, q_1 \) and \( r_1 \) = The design parameters to be identified during calibrations and validation processes

The architecture of ANFIS is shown in Fig. 1 in which circles nodes and squares describe adaptive nodes. The following paragraph provides a brief introduction to the ANFIS Model.

**Input nodes (layer 1):** Each node \( i \) of this layer is a square node with a node function. In fuzzy system, for input values \( x_1 \) and \( x_2 \), the inferred output \( y \) is estimated by using Eq. 3 (Lin and Lee 1996):
Fig. 1: Architecture of the Adaptive Network-based Fuzzy Interface System (ANFIS) (Jang, 1993)

Fig. 2: A simple structure of the artificial neuron (Haykin, 1999)

\[ y = \frac{(\mu_1 x_1 + \mu_2 x_2)}{(\mu_1 + \mu_2)} \]  \hspace{1cm} (3)

where, \( \mu_j \) are firing strengths of \( R_j \), \( j = 1, 2 \) given by Eq. 11:

\[ \mu_j = \mu A_j^1(x_1) + \mu A_j^2(x_2), \quad j = 1, 2 \]  \hspace{1cm} (4)

In this study, ANFIS, ANN and MLR modeling were implemented using MATLAB v2017 Software.

**Artificial Neural Network (ANNs):** An artificial neuron is the primary building step for all ANN. It has the same design and characteristics in natural neurons in biological neural networks. Figure 2 shows the architecture of the artificial neuron with inputs variable, weights, transfer function, activation functions, threshold and output.

The artificial neuron is fed by numbers of inputs. Depending on the value of the weight, the effect of the transfer function and output, the effect of all inputs on the neuron will be differ furthermore the calculation of the transfer function and output. Generally, greater weight values result in higher power and affect the associated inputs. Since, all the inputs are multiplied by its corresponding weight, the weights will influence the neuron's output. The transfer function as a summation of the weighted inputs is used to produce the net input to the neuron (Haykin, 1999) as provided in Eq. 5:

\[ \text{net}_j = \sum \omega_{ij} x_i + b \]  \hspace{1cm} (5)

Where:
- \( j \) = The actual neuron number
- \( x_i \) = An input value, \( i \) from 1 to \( n \)
- \( \omega_{ij} \) = A weight value
- \( b \) = Equal to the negative threshold value of a neuron and called the bias of the neuron
And:

\[ x_j = \phi(u_j - \theta_j) \]  \hspace{1cm} (6)

Where:

\[ x_j = \text{Output signal} \]

* \( \theta_j \) = The bias term of the j neuron (Haykin 1999; Melesse and Hanley, 2005)

The logistic sigmoid function (Bilgili et al., 2007) is used for this purpose, expressed as given in Eq. 7:

\[ \phi_x = \frac{1}{1+e^x} \]  \hspace{1cm} (7)

The structure of ANN Model with a three-layer used in this study is shown in Fig. 3. The type of ANN is feed-forward back propagation algorithm and using the Levenberg-Marquardt Training algorithm (Train LM). The transfer function between layer one and layer two was LOGSIG. The optimal number of neurons in the hidden layer was selected using the trial and error method by experimenting with changing the number of neurons in the hidden layer from 1-20.

**Multiple Linear Regression (MLR):** In MLR is defined that the dependent variable as a linear function of one variable. In Eq. 8 and 9 describe the simple linear regression as well as the measured and calculated values of the dependent variable (Brown and Berthouex 2002):

\[ Y = a + bX \]  \hspace{1cm} (8)

\[ Y_i = a + bX + \varepsilon_i \]  \hspace{1cm} (9)

Where:

\( Y \) = The measured value

\( Y_i \) = The calculated value

\( a \) = The constant

\( b \) = The error associated with estimating of \( Y_i \) and the value of \( X = x_i \) is the given value of the independent variable

The constants \( a \) and \( b \) are estimated by ordinary least squares. If \( \varepsilon_i = 0 \), calculated value \( (Y) \) is equal to measured value \( (Y) \). MLR is very similar to simple linear regression but the difference in MLR is that the dependent variable is a function for more than one independent variable MLR model can be specified as given in Eq. 10:

\[ Y_i = a + b_1X_1 + b_2X_2 + ... + b_nX_n + \varepsilon_i \]  \hspace{1cm} (10)

where, \( Y, a \) and \( \varepsilon \) have described above, \( b_1, b_2, ..., b_n \) are the partial regression (slope) parameter for \( X_1, X_2, ..., X_n \). The main purpose of using MLR is to find the linear relationship between dependent and independent variables and to obtain a linear model using regression coefficients as well as to calculate the dependent variable. For the best-calculated value of the dependent variable, \( \varepsilon \), can be specified as given in Eq. 11:

\[ \sum_{i=1}^{n}(\varepsilon_i)^2 = \sum_{i=1}^{n}(Y_i - a + b_1X_1 + b_2X_2 + ... + b_nX_n)^2 \]  \hspace{1cm} (11)

**Study area and data**

**Abu-Ziriq Marsh description:** Abu Ziriq Marsh which covers 120 km², it is about 3% of all Marshes area, lies at the tail end of Al Gharraf River Southerly of Al Islah District at a location of latitude 31°09' 54.9"N, longitude 46°36'33"E. The main source of water supply to the Marsh is through Shatt Abo-Lihia and the channel of this river runs through the Marsh until it dissipates at the tail end into the central Marshes. The two main towns around the Marsh are Al-Islah in the North and Al-Fuhod in the...
Table 1: Monthly records of water quality parameters per year

<table>
<thead>
<tr>
<th>Sample of water</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2006</td>
</tr>
<tr>
<td>9</td>
<td>2010</td>
</tr>
<tr>
<td>12</td>
<td>2013</td>
</tr>
<tr>
<td>12</td>
<td>2014</td>
</tr>
<tr>
<td>12</td>
<td>2015</td>
</tr>
<tr>
<td>8</td>
<td>2016</td>
</tr>
<tr>
<td>12</td>
<td>2017</td>
</tr>
<tr>
<td>8</td>
<td>2018</td>
</tr>
</tbody>
</table>

South of Thi Qar governorate Fig. 4. Scattered villages of fishermen are located all along the embankments that surround the Marsh. Implying the vitality role of Abu Zirig Marsh in sustaining the daily life of the local residents.

**Water sampling procedure**: The dataset utilized in this study was collected and observed consistently, every month at Abu-Zirig Marsh by the Ministry of the Environment, Department of Protection and Improvement Environment in the South of Iraq. The final dataset of water quality consisted 84 monthly records of data between years, 2009 and 2018 as shown in Table 1. Each record consists of 8 parameters, namely: NO₃, Ca²⁺, Mg²⁺, T. H, SO₄²⁻, Cl⁻, EC and TDS. These variables are used to develop the ANFIS, ANN and MLR Models. Table 2 indicates the statistical parameters of water quality in the wetland. In this study, the total Abu-Zirig water quality dataset (84 samples) were randomly divided into two groups, calibration and validation. The calibration and validation datasets comprised of 63 (75%) and 21 (25%) samples, respectively.

**Performance measures**: Several criteria have been widely used in the literature for the assessment of model performance (Srinivasulu and Jain, 2006). In this study, the following three criteria were employed.

**Root Mean Squared Error (RMSE)**: The Root Mean Squared Error (RMSE) is an error index type parameter commonly used in hydrological modeling:

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (M_i - \hat{P}_i)^2}{N}}
\]  

Where:
- \( \mu \) = Average measured values
- \( N \) = Number of data set
- \( M_i \) = Some Measured value
- \( \hat{P}_i \) = The corresponding model Predicted or estimate value

For RMSE, a value of zero is the optimum.

**Correlation Coefficient (CC)**: Correlation Coefficient (CC) is a standard regression type parameter and defined as a measure of the strength of the linear relationship between the measured and predicted or estimate datasets:

\[
CC = \frac{\sum_{i=1}^{N} (M_i - \mu)(\hat{P}_i - \hat{P})}{\sqrt{\sum_{i=1}^{N} (M_i - \mu)^2 (\hat{P}_i - \hat{P})^2}}
\]
Table 2: Summary of statistical parameters of input and output variables (n = 84)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Range</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>SD</th>
<th>CV%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO\textsubscript{2}</td>
<td>ppm</td>
<td>3.50</td>
<td>0.30</td>
<td>3.80</td>
<td>1.4534</td>
<td>0.52253</td>
<td>0.273</td>
</tr>
<tr>
<td>Ca$^{2+}$</td>
<td>ppm</td>
<td>756.00</td>
<td>64.00</td>
<td>800.00</td>
<td>160.8452</td>
<td>94.25513</td>
<td>89.95700</td>
</tr>
<tr>
<td>Mg$^{2+}$</td>
<td>ppm</td>
<td>310.00</td>
<td>20.00</td>
<td>330.00</td>
<td>783.2024</td>
<td>394.79950</td>
<td>429.0847</td>
</tr>
<tr>
<td>T.H</td>
<td>ppm</td>
<td>1840.00</td>
<td>120.00</td>
<td>2160.00</td>
<td>1300.00</td>
<td>430.2500</td>
<td>77427.587</td>
</tr>
<tr>
<td>SO$^{4-}$</td>
<td>ppm</td>
<td>1201.00</td>
<td>59.00</td>
<td>1300.00</td>
<td>430.2500</td>
<td>77427.587</td>
<td>97626.158</td>
</tr>
<tr>
<td>Cl$^{-}$</td>
<td>ppm</td>
<td>1472.00</td>
<td>150.00</td>
<td>1622.00</td>
<td>481.2857</td>
<td>312.45185</td>
<td>97626.158</td>
</tr>
<tr>
<td>EC</td>
<td>µS/cm</td>
<td>6620.00</td>
<td>1200.00</td>
<td>7820.00</td>
<td>3072.1310</td>
<td>1676.7130</td>
<td>281.1366</td>
</tr>
<tr>
<td>TDS</td>
<td>ppm</td>
<td>4006.00</td>
<td>614.00</td>
<td>4620.00</td>
<td>1781.1905</td>
<td>960.22337</td>
<td>522.209313</td>
</tr>
</tbody>
</table>

SD: Standard Deviation; CV: Coefficient of Variation; Ca: Calcium; Cl: Chloride; EC: Electrical Conductivity; NO\textsubscript{2}: Nitrate; Mg: Magnesium; SO\textsubscript{4}: Sulfate; TDS: Total Dissolved Solids

Where:

\[ N = \text{The Number of input samples} \]
\[ M_i \text{ and } P_i = \text{The Measured and network output value from the elements, respectively} \]
\[ \mu \text{ and } \mu^\prime = \text{Average, respectively} \]

The Nash–Sutcliffe coefficient of Efficiency (NSE): The Nash–Sutcliffe Co-efficient of Efficiency (NSE) is a dimensionless type parameter widely used as a metric of model efficiency (Srinivasulu and Jain, 2006):

\[
\text{NSE} = \frac{\sum_{i=1}^{n}(P_i - M_i)^2 - \sum_{i=1}^{n}(P_i - \mu)^2}{\sum_{i=1}^{n}(P_i - \mu)^2}
\] (14)

NSE range from negative to 1 with better models giving NSE values as close to 1 as possible.

RESULTS AND DISCUSSION

Model structure: Given its importance in terms of the Abu-Ziriq Marsh water quality, Electrical Conductivity (EC) and Total Dissolved Solids (TDS) were chosen as the water quality parameters of interest. The chemical parameters, namely, NO\textsubscript{2}, Ca$^{2+}$, Mg$^{2+}$, T.H, SO$^{4-}$, Cl$^{-}$, EC, and TDS were assessed (Table 2) for Abu-Ziriq Marsh water samples collected on the monthly basis by the Ministry of the Environment, Department of Protection and Improvement Environment in the South of Iraq over the period of January, 2009, August, 2018 at the Abu-Ziriq station. An important thing to do in developing a prediction model is to choose the correct input parameters. The parameters for EC and TDS based on strong Pearson while the weak cross-correlation parameters are neglected (Table 3). Cross-correlation is used for measuring the similarity of two series as a function of the displacement of one relative to the other (Bracewell, 1965). Table 3 tabulates the correlation matrix between the water quality parameters. Based on Pearson correlation coefficient with p<0.01, the parameters used as inputs in modeling TDS were the concentrations of Ca$^{2+}$, Mg$^{2+}$, T.H, SO$^{4-}$ and Cl$^{-}$ (Table 3). Apparently, there was no remarkable difference between the model structure of EC and TDS. The only difference was the component of NO\textsubscript{2}. This might be attributed due to the equivalent characteristic of both parameters, i.e., they are both a measure of the amount of dissolved solids.

Models performance: In this study, Nitrate (NO\textsubscript{2}), Calcium (Ca$^{2+}$), Magnesium (Mg$^{2+}$), Total Hardness (T.H), Sulfate (SO$^{4-}$), Chloride (Cl$^{-}$), Electrical Conductivity (EC) and Total Dissolved Solids (TDS) in Abu-Ziriq, South of Iraq were used to develop artificial intelligence techniques. The TDS and EC Models were created by utilizing ANFIS, ANNs and MLR.

In ANFIS modeling, membership function types for input and output parameters were considered as Sugeno fuzzy Gaussian (gaussmf), backpropagation algorithm and linear MFs, respectively. The number of membership functions for each input of ANFIS for TDS and EC were set to (2, 2, 2, 1, 2, 3) and (2, 3, 1, 3, 2), respectively. The performance of the ANFIS Model for the calibration and validation of datasets are given in Table 4. Figure 5 shows the observed versus predicted TDS from ANFIS Model during the calibration and validation periods. As it can be raised from figure there was a satisfactory matching between both data sets. Moreover, values of RMSE, CC and NSE were 169.30, 0.98 and 0.98, respectively for the calibration and 193.59, 0.98 and 0.97, respectively for the validation of datasets (Table 4). While Fig. 6 shows the observed versus predicted EC from ANFIS Model during the calibration and validation periods. As it can be noticed from figure there was a satisfactory matching between both data sets. This was manifested through values of RMSE CC and NSE which were 273.45, 0.98 and 0.97, respectively for calibration data set and 246.42, 0.99 and 0.98, respectively for validation data set.

In ANN modeling feed forward-backpropagation algorithm, Levenberg-Marquardt Training algorithm (TrainLM) were constructed to estimate TDS and EC values. The transfer function between layer one and layer two was LOGSIG. The optimal number of neurons in the hidden layer was selected using the trial and error method.
Table 3: Correlation matrix among water quality parameters

<table>
<thead>
<tr>
<th>NO₃</th>
<th>Ca²⁺</th>
<th>Mg²⁺</th>
<th>T.H</th>
<th>SO₄⁻</th>
<th>Cl⁻</th>
<th>EC</th>
<th>TDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.225</td>
<td>0.139</td>
<td>0.149</td>
<td>0.126</td>
<td>0.209</td>
<td>0.225</td>
<td>1</td>
</tr>
<tr>
<td>0.225</td>
<td>1</td>
<td>0.582</td>
<td>0.943</td>
<td>1</td>
<td>0.789</td>
<td>1</td>
<td>0.103</td>
</tr>
<tr>
<td>0.139</td>
<td>0.582</td>
<td>1</td>
<td>0.878</td>
<td>0.894</td>
<td>1</td>
<td>0.293</td>
<td></td>
</tr>
<tr>
<td>0.149</td>
<td>0.943</td>
<td>0.878</td>
<td>1</td>
<td>0.884</td>
<td>1</td>
<td>0.590</td>
<td></td>
</tr>
<tr>
<td>0.126</td>
<td>1</td>
<td>0.894</td>
<td>0.884</td>
<td>1</td>
<td>0.255</td>
<td>1</td>
<td>0.193</td>
</tr>
<tr>
<td>0.209</td>
<td>0.789</td>
<td>1</td>
<td>0.884</td>
<td>0.855</td>
<td>1</td>
<td>0.220</td>
<td></td>
</tr>
<tr>
<td>0.225</td>
<td>0.103</td>
<td>0.293</td>
<td>0.590</td>
<td>0.566</td>
<td>0.988</td>
<td>1</td>
<td>0.193</td>
</tr>
</tbody>
</table>

Ca²⁺: Calcium, Cl⁻: Chlorine; EC: Electrical Conductivity; NO₃: Nitrate; Mg²⁺: Magnesium; SO₄⁻: Sulfate; TDS: Total Dissolved Solids. All the values were significant at α = 0.05.

Table 4: Comparison of ANFIS, ANN and MLR Models performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Estimated</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Calibration</td>
<td>NSE</td>
</tr>
<tr>
<td>TDS</td>
<td>MLR</td>
<td>184.58 ppm</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>185.89 ppm</td>
</tr>
<tr>
<td></td>
<td>ANFIS</td>
<td>146.30 ppm</td>
</tr>
<tr>
<td>EC</td>
<td>MLR</td>
<td>297.13 μS/cm</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>314.88 μS/cm</td>
</tr>
<tr>
<td></td>
<td>ANFIS</td>
<td>273.45 μS/cm</td>
</tr>
</tbody>
</table>

Fig. 5: Comparative plots of observed and predicted TDS values using ANFIS Model for: a) Calibration data set and b) Validation.

by experimenting with changing the number of neurons in the hidden layer from 1-20. The optimal number of neurons in the hidden layers providing the optimal structure was determined as 8 for TDS and 5 for EC. Therefore, ANN (6, 8, 1) was selected as the optimum ANN Model for TDS and ANN (5, 5, 1) for EC. The performance of the ANN Model for the calibration and validation of datasets are given in Table 4. Figure 7
shows the observed versus predicted TDS from ANN Model during the calibration and validation periods. As it can be shown from, figure there was an adequate consistency between both data sets. In addition, values of RMSE, CC and NSE were 185.89, 0.97 and 0.95, respectively for calibration data set and 202.39, 0.98 and 0.96, respectively for validation data set. On the other side, Fig. 8 shows the observed versus predicted EC from ANN Model during the calibration and validation periods. As it can be seen from figure, both data sets were in a good consistency.

Moreover, values of RMSE, CC and NSE were 314.88, 0.98 and 0.96, respectively for calibration data set and 568.3, 0.98 and 0.89, respectively for validation data set. The performance of the MLR Model for the calibration and validation of datasets are given in Table 4. Figure 9 shows the comparative plots of the results obtained from MLR Model for TDS during the calibration and validation periods. RMSE, CC and NSE values set were 184.58, 0.97 and 0.95 for the calibration data, respectively. While these values for the validation dataset were 196.89 ppm, 0.99 and 0.96, respectively.

![Fig. 6: Comparative plots of observed and predicted EC values using ANFIS model for a) calibration data set and b) validation](image)

![Fig. 7: Continue](image)
Fig. 7: Comparative plots of observed and predicted TDS values using ANN Model for: a) Calibration data set and b) Validation

Fig. 8: Comparative plots of observed and predicted EC values using ANN Model for: a) Calibration data set and b) Validation

In addition, MLR Model was used to estimate EC, the performance model for the calibration and validation of data sets were plotted as shown in Fig. 10. It can be noticed that the both data sets were consistent. In other words, MLR Model’s performance was satisfactory in modeling EC. However, RMSE, CC and NSE values set were 297.13 µsec/cm, 0.98 and 0.96, respectively for the calibration while these values for the validation of dataset were 537.53 µsec/cm, 0.98 and 0.90, respectively. Nemati et al. (2015) used ANFIS, ANN and MLR Models to estimate water quality parameter in the Tai Po River, Hong Kong. They found that MLR Model had not have the high accuracy to estimate DO.

From aforementioned, it can be concluded that the ANFIS Model outperformed the ANN and MLR Models. This might be attributed due to its sophisticated
Fig. 9: Comparative plots of observed and predicted TDS values using MLR Model for: a) Calibration data set and b) Validation

Fig. 10: Comparative plots of observed and predicted EC values using MLR Model for: a) Calibration data set and b) Validation
structure and the capability of eliminating the noisy data (Al-Mukhtar, 2018). For better visualization of the models performances, the time series of ANFIS, ANN and MLR for validation data set were depicted as shown in Fig. 11. It can be seen that, the ANFIS Model provided much closer values to the observed or actual TDS and EC values than ANN and MLR Models. The neuro-fuzzy systems have an advantage of both ANFIS and ANNs, i.e., benefiting from the training ability of the ANN and the fuzzy IF-THEN rule generation and parameter optimization (Kosko, 1997).

Recently, ANFIS Model has been successfully applied in water quality modeling with respect to other models. Hence, our findings are in parallel with previous studies (Ay and Kisi, 2017; Heddam, 2014; Najah et al., 2014; Yan et al., 2010) where they proved the superior performance of ANFIS in modeling hydrological and water quality parameters.

**CONCLUSION**

Total Dissolved Solids (TDS) and Electrical Conductivity (EC) are important parameters in determining water quality such as drinking water and agricultural water because it directly affects the concentration of salt in water. High values of these parameters cause low water quality. In addition to that, they play a significant role in hydrous life, effective water resources management. In recent years, many studies have successfully used different models to calculate and predict EC and TDS. In this study, MLR, ANFIS and ANN Models were developed to estimate EC and TDS values in Abu-Ziriq Marsh, South of Iraq. The performance of the developed models was evaluated through three performance criteria: Correlation Coefficient (CC), Root Mean Square Error (RMSE) and the Nash-Sutcliffe Efficiency coefficient (NSE). The comparison shows that the significantly correlated parameters of EC were the concentrations of Ca$^{2+}$, Mg$^{2+}$, T. H, SO$^{4}$, and Cl$^{-}$. While, those of TDS were the concentrations of NO$_{3}$, Ca$^{2+}$, Mg$^{2+}$, T. H, SO$^{4}$, and Cl$^{-}$. Results proved that ANFIS Model is close to TDS and EC actual or observed values. In other words, ANFIS Model led in the best fit with the observed data.

**RECOMMENDATION**

ANFIS is recommended to be used as a predictive model in the Iraqi marshes.

**ACKNOWLEDGEMENTS**

The researchers are grateful to the Ministry of the Environment, Department of Protection and Improvement Environment in the South of Iraq for their generous support in providing the data. The researchers are grateful to Mr. Arif Shamsi, manager of the Marshlands Department in Thi Qar Governorate and Miss Farah AL-Ghuraby for given the required permissions to use water quality data and encouragement to conduct such studies for the benefit of science and society.

**REFERENCES**


