Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model

Yafra Khan
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
yafra.khan@gmail.com

Chai Soo See
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
sschai@fit.unimas.my

***Abstract*— The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity. The effects of un-clean water are far-reaching, impacting every aspect of life. Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. This issue has been addressed in many previous researches, however, more work needs to be done in terms of effectiveness, reliability, accuracy as well as usability of the current water quality management methodologies. The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. This research uses the water quality historical data of the year of 2014, with 6-minutes time interval. Data is obtained from the United States Geological Survey (USGS) online resource called National Water Information System (NWIS). For this paper, the data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis. Previous works about Water Quality prediction have also been analyzed and future improvements have been proposed in this paper.**

Keywords- Artificial Neural Networks, Environmental Modeling, Machine Learning, Time-Series Analysis

I. INTRODUCTION

Natural water resources like groundwater and surface water have always been the cheapest and most widely available resources of fresh water. However, these resources are also most likely to become contaminated due to various factors including human, industrial and commercial activities as well as natural processes. In addition to that, poor sanitation infrastructure and lack of awareness also contributes immensely to drinking water contamination [1]. The effects of water quality deterioration are far-reaching, impacting health, environment and infrastructure in a very adverse manner. According to United Nations (UN), waterborne diseases cause death of more than 1.5 million people each year, much greater than deaths caused by accidents, crimes and terrorism combined[2]. Therefore, it is very crucial to devise novel approaches and methodologies for analyzing water quality and to forecast future water quality trends.

In order to carry out useful and efficient water quality analysis and predicting the water quality patterns, it is very significant to include a temporal dimension to the analysis, so that the seasonal variation of water quality is addressed [3]. Moreover, recent studies have shown that a suitable hybrid of multiple models for forecasting and prediction gives better results than using a single one[4][5]. Different methodologies have been proposed and applied for analysis and monitoring of water quality as well as time series analysis. The methodologies range from statistical techniques, visual modeling, analysis algorithms and prediction algorithms and decision making. Multivariate statistical techniques like Principal Component Analysis (PCA) has been used in order to determine relationship among different water quality parameters[3]. The geo-statistical techniques that have been used include kriging, transitional probability, multivariate interpolation, regression analysis etc.[4]. The algorithms for analysis and prediction might include Artificial Intelligence (AI) techniques like Bayesian Networks (BN), Artificial Neural Networks (ANN) [5] Neuro-Fuzzy Inference[3], Support Vector Regression (SVR)[6], Decision Support System (DSS) and Auto-Regressive Moving Average (ARMA)[7]. However, the non-linear nature of water quality data, as in this research, makes it very complex to map input-output data and predict future water quality [8].

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological or chemical factors which influence water quality. There are certain quality standards set up by international organizations like World Health Organization (WHO) and Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document "Parameters of Water Quality", EPA mentions a total of 101 parameters which have an effect upon water quality in one way or another [9]. However, some parameters have a greater and more visible effect on water quality than others.

This paper intends to address this issue by suggesting a model based upon Machine Learning techniques in order to

predict the future water quality trends of a particular area with the help of current water quality data. Artificial Neural Networks (ANN) with Nonlinear Autoregressive (NAR) time series model is used in order to develop a comprehensive methodology for efficient water quality prediction and analysis. There are four selected water quality parameters used in this study i.e. Chlorophyll, Specific Conductance, Dissolved Oxygen and Turbidity. The goal of this research is to develop efficient models to predict values of water quality parameters based upon their present values.

II. Data Acquisition and Study Area

Previous studies have shown that the richness and quality of data determines the accuracy and reliability of analysis[10]. Since most of the water monitoring organizations have lack of detail and insufficient observations [11], we have opted for the acquisition of data from one of the most reliable water resources in the world which is usually pre-processed and frequently updated. The sample data for this research has been acquired from U.S. Geological Survey's (USGS) National Water Information System (NWIS) which is an open data repository supporting acquisition, processing and long-term storage of water quality data across the U.S. The study area of this research lies in Island Park village, situated in the South-Western Nassau County with Latitude 40°36'31.8", Longitude 73°39'22.0" in the state of New York (Figure 1). The measurements of the data for the monitoring station of Hog Island Channel have been used in this study, where water samples are collected and monitored by USGS using different techniques. For measurement, Satellite telemeter is mainly used with readings collected from 1.6 Ft. above bottom. Data from 2014 with the time-interval of 6 minutes has been acquired in order to carry out an efficient prediction process using this time-series data that includes date/time, parameters and their measurements along with measurement units.

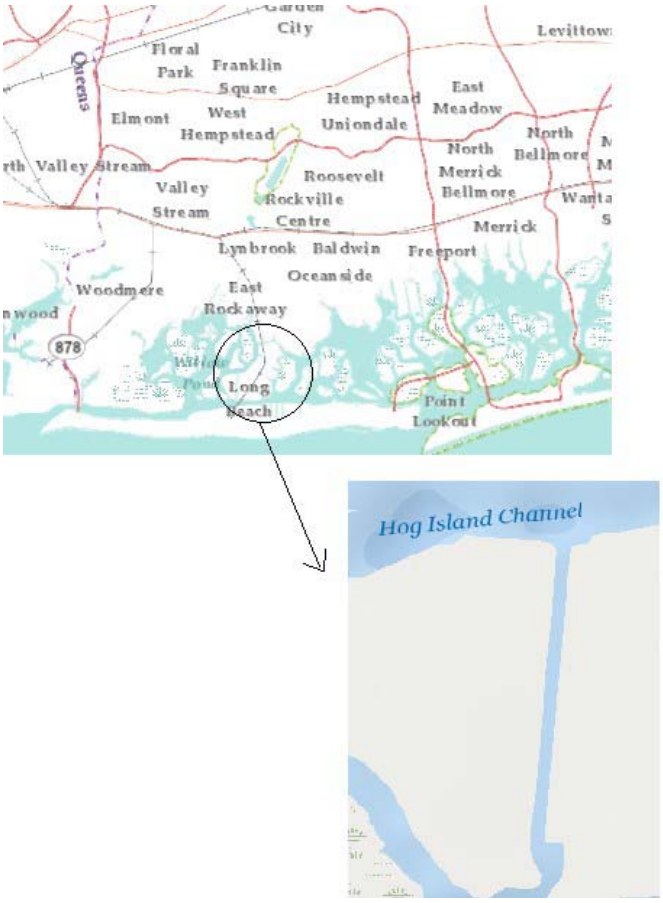

Figure 1 Area covered in the Island Park and Hog Island Channel Monitoring Station

III. Theoretical Background of Applied Methodology

The methodology used in this study comprises of Machine learning with training and testing data from USGS online data repository. The theoretical background of the methodology is as follows:

A. Artificial Neural Network

ANN has been widely acknowledged as a methodology for classification of complex datasets such as those of environmental processes. It has the ability to efficiently describe the non-linear relationship of the complex water quality datasets [12]. Moreover, it has strong adaptability to depict the changes that might occur in the water environment of a particular area. The algorithmic architecture of ANN attempts to simulate the structure and networks in a human brain, with an input layer, hidden layer and output layer each consisting of nodes. There might be one or more hidden layers, depending upon the problem at hand.

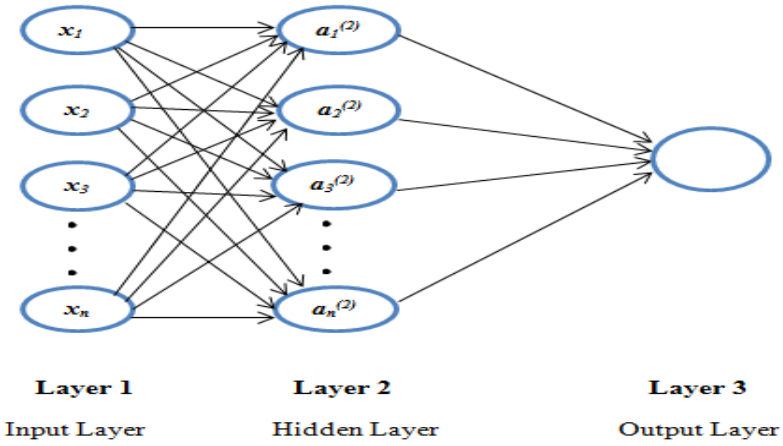

Figure 2 Structure of Artificial Neural Network

In addition to that, there are connections between the nodes with varying "weights"[13]. For this particular research, a feed-forward and back-propagation Neural Network with three layers has been used, i.e. one input, one hidden and one output layer. In the feed forward process, the weights are multiplied by the inputs and the resultant value is moved forward towards the next layer, until it reaches the output layer, as follows:

$$z_i = \sum_{j=1}^{m} w_{ij} x_{ij}$$

Where w_{ij} is the weight transferred from ***jth*** input to the ***ith*** node, is the input and ***zi*** is the summation of outputs of the ***ith*** node. In this study, for correlation between the parameters, the input layer initially consists of 10 units denoting the water quality parameters. The back-propagation process determines the error value by calculating the difference between estimated value and expected value, starting from output layer towards the input layer[5]. It is denoted by the symbol ***δ(l)j***, which is equal to error of node *j* in layer *l*. For a training set (*xj,yj*), the error term is:

$$\delta(l)j = zj - yj$$

It is an iterative process, so after adjustment of the weights, the process is run repeatedly until convergence.

B. Neural Network for Time Series

A time series is a sequence of values x_t, recorded at a specific time t. Time series which records the observations continuously in some specific intervals of time is known as *Continuous-time time series* [14]. Data used in this study also comes in the category of Continuous-time time series, as it consists of the values of water quality factors observed with the time-interval of 6 minutes. Since ANN is used to interpret non-linear relationship of the data, the time series model used in this study is Non-linear Autoregressive (NAR) model. This is a non-linear model is used to define the input and output in terms of time, which is easily estimated in terms of regression[15]. The Nonlinear Autoregressive Neural Network (NRA-NN) exploits the benefits of both NRA and ANN. In this scenario, the general ANN model slightly changes to take the mathematical form:

$$y_t(t) = \sum_{j=1}^{m} w_{ij} y_{ij}\ (t-1)$$

Where $y_t(t)$ is the output time series and y_{ij}***(t-1)*** is the input time series.

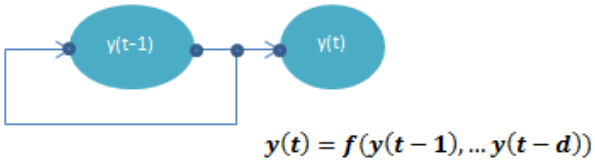

Figure 3 NAR Model

This model is described as a network of three layers of processing units connected by acyclic links [16], as indicated in figure 3.

For this study, data was divided into training data (60%), testing data (20%) and validation data (20%). As seasonal variation affects the time series forecasting, it was made sure that data for training, testing and validation is from the same or nearly similar seasons. For better analysis and results, data was scaled to fall between the ranges of [0,1].

IV. RESULTS AND DISCUSSION

A test was conducted in order to forecast the selected water quality factors based upon their past values. This has been done using Artificial Neural Network (ANN) with time series Nonlinear Autoregressive (NAR) model. Some statistics about the selected water quality parameters for the year 2014 were collected from USGS, including Minimum Value, Maximum Value and Mid-Range value, in order to depict the range of values (Table 1). This test consists of four models, each used for forecasting the four water quality factors that have been selected i.e. Turbidity, Dissolved Oxygen concentration, Chlorophyll and Specific Conductance.

In these tests, the input and output are represented by the values of same parameters at different times. The samples include the data ranging from January to March 2014, with 6-minute time interval. A feed-forward Neural Network with NAR time series model has been used with the training algorithm of Scaled Conjugate Gradient (SCG) and the activation function of Log Sigmoid. After running the test, the performance parameters of Regression(R), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) have been calculated. The performance is shown graphically with MSE and Regression analysis of four models (Figure 4, 5, 6, 7). The values of the performance measures for four ANN models for training and testing processes are shown in the table (Table 2).

The graphs for Regression Analysis show how well the data fits into the function, for training, testing and validation. The closer the value of Regression is to 1, the better the function fits and hence it indicates better prediction accuracy. The graphs for MSE show the amount of epochs (iterations) it takes for the function to converge and the related MSE for training, testing and validation. We can see in Figure 4(a) that it takes 57 epochs (iterations) to converge, with best performance on epoch 51. We can also see that MSE graph for training, testing and validation almost overlapping, indicating a balanced accuracy without being over fitted. In figure 4(b), we can analyze that most data for Chlorophyll prediction fits into the range of 0 and 0.5, though there are a few outliers.

Here, Regression for both training and testing is 0.8, hence indicating that data fits well. The graph of MSE for Specific Conductance (Figure 5(a)) shows a very drastic difference between training and validation, indicating a high variance (over fitting). Here, the number of epochs is 67, with the best value at epoch 61. On the other hand, if we see Regression Analysis for Conductance (Figure 5(b)), we can analyze that the data almost entirely fits the function with regression values of 0.99 for both training and testing, hence indicating high prediction accuracy for training. However, there needs to be a balance between bias and variance in order for the function to work well, where high bias indicates under-fitting and high variance indicates over-fitting. In case of Dissolved Oxygen, the graph of MSE (Figure 6(a)) shows the lines of training, validation and testing with almost similar values, indicating a balanced function fit. The Regression graph for Dissolved Oxygen (Figure 6(b)) shows that the data points fit almost entirely, with negligible quantity of outliers. The MSE for Turbidity (Figure 7(a)) shows a slightly less generalized model, with training error being less than the testing and validation error. We can see that it has only 22 epochs, hence it takes much less time for the function to converge than other models. In case of Regression of Turbidity (Figure 7(b)), the graph shows that's most of the points fit well, lying between -0.2 to 0.2, with few outliers. The performance measures and analysis can be verified by looking at Table 2.

Parameter	Minimum Value	Maximum Value	Mid-Range Value
Chlorophyll (µg/L)	0.7	140	70.35
Specific Conductance (µS/cm)	38900	49100	44000
Dissolved Oxygen (Mg/L)	3.6	18.0	10.8
Turbidity (FNU)	<0.1	120	--

Table 1: Characteristics of Water Quality Data for 2014

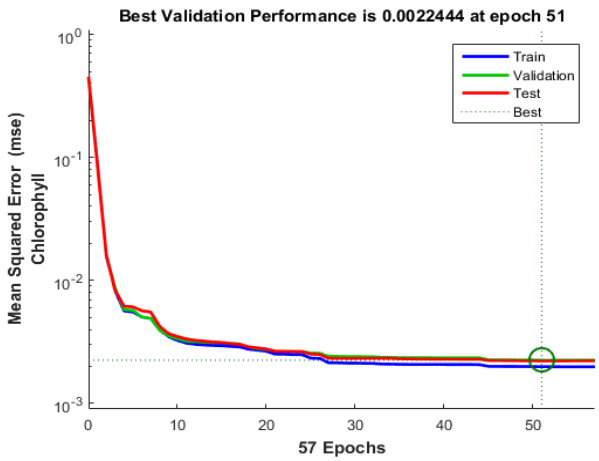

Figure 4(a) Mean Squared Error for Chlorophyll

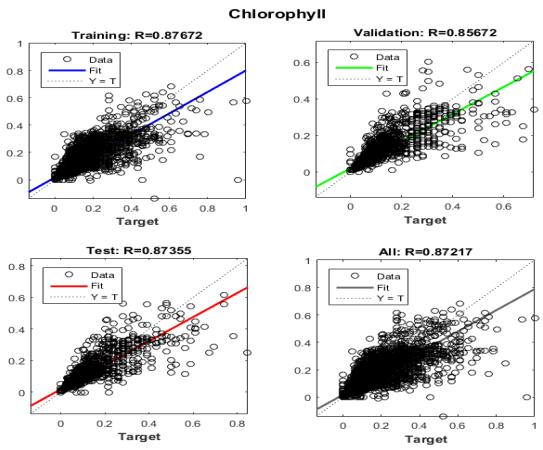

Figure 4(b) Regression for Chlorophyll

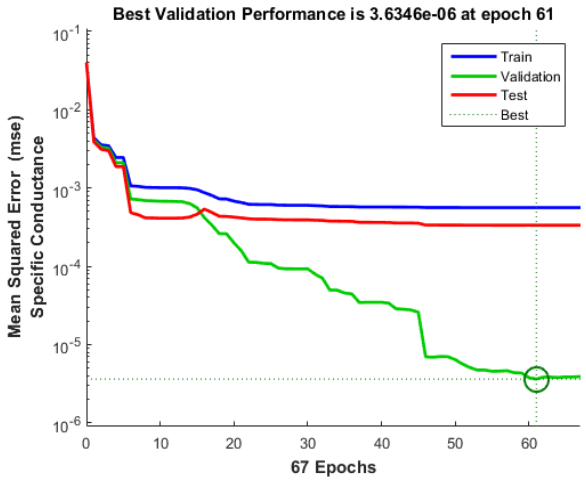

Figure 5(a) Mean Squared Error for Conductance

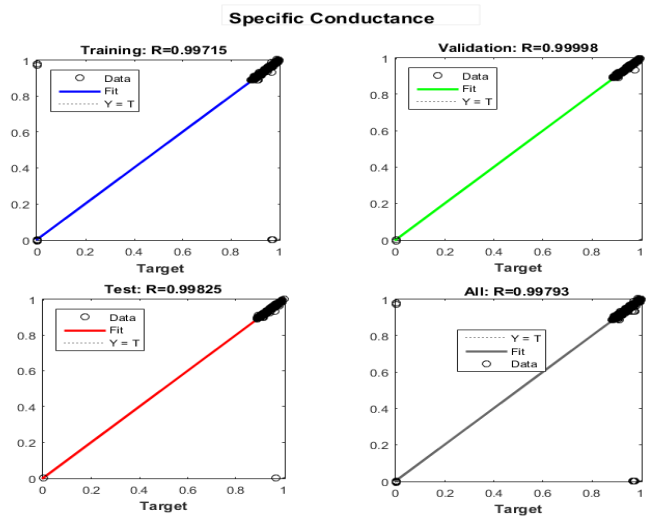

Figure 5(b) Mean Squared Error for Conductance

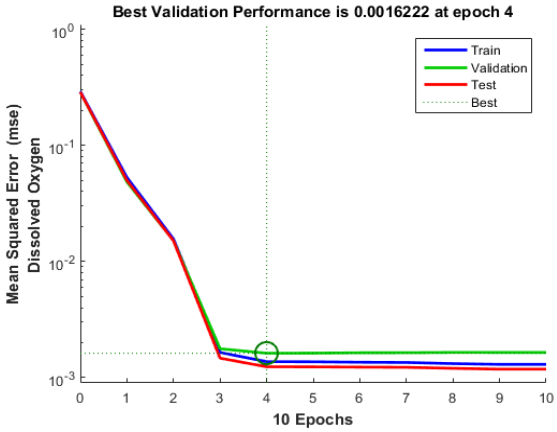

Figure 6(a) Mean Squared Error for Dissolved Oxygen

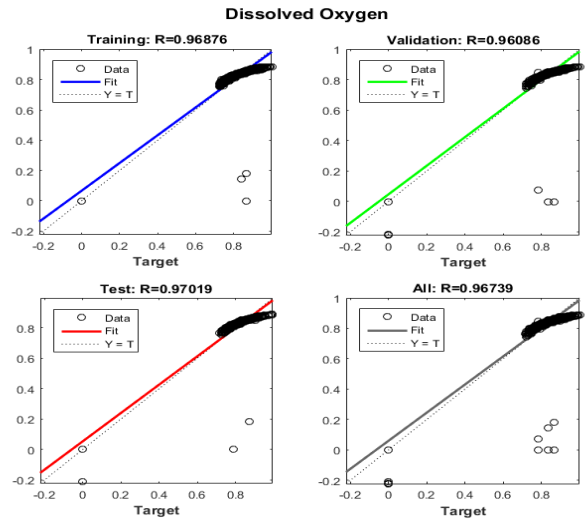

Figure 7(b) Regression for Dissolved Oxygen

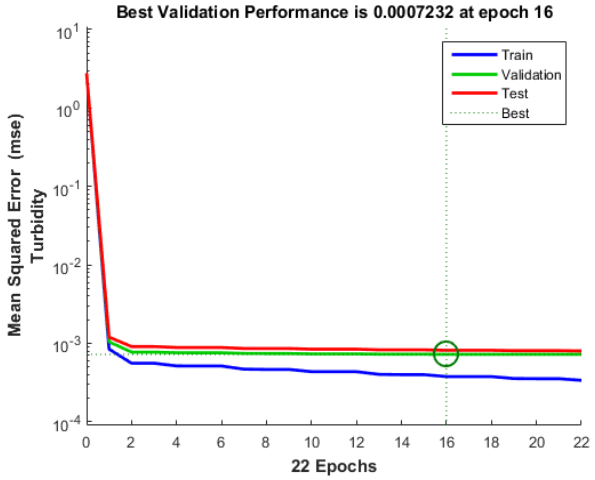

Figure 7(a) Mean Squared Error for Turbidity

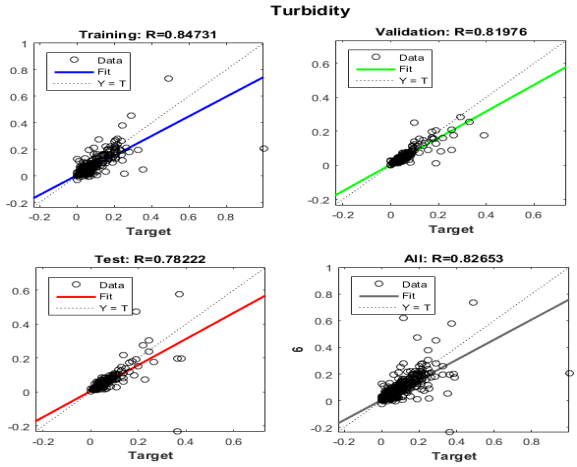

Figure 7(b) Regression for Turbidity

			Training Data			Testing Data		
Parameters	**Unit**	**Model**	**R**	**MSE**	**RMSE**	**R**	**MSE**	**RMSE**
Chlorophyll	µg/L	NAR	0.876	0.00198	0.0444	0.873	0.00221	0.0470
Specific Conductance	µS/cm	NAR	0.997	0.00056	0.0236	0.998	0.00033	0.0181
Dissolved Oxygen	mg/L	NAR	0.968	0.00137	0.0370	0.970	0.00123	0.0350
Turbidity	FNU	NAR	0.847	0.00037	0.0192	0.782	0.00081	0.0285

Table 2: Performance Measures for ANN Time-Series

V. CONCLUSION

This paper analyzes and forecasts the values of water quality parameters, in order to determine the concentration of Chlorophyll, Dissolved Oxygen, Turbidity and Specific Conductance and analyzes the results. The time series data used has been acquired from USGS National Water Information System (NWIS), with data from the year of 2014. The specified monitoring station is a channel situated in the State of New York. The measurements of water quality parameters were scaled between 0 and 1 for better data handling. Artificial Neural Network (ANN) with Nonlinear Autoregressive (NAR) time series has been used with Scaled Conjugate gradient (SCG) as training algorithm. Four ANN models depicting the four selected water quality parameters have been developed and analyzed. The performance measures that are used to depict the result are Regression, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

The results of the conducted tests provide an insight about the prediction efficiency and accuracy of the proposed model with the help of performance measures. The proposed model comprising of ANN-NAR proves to a reliable one with the prediction accuracy indicating much improved values, with the lowest MSE being $3.7x10^{-4}$ for turbidity and the best Regression value for Specific Conductance (0.99). The future of water quality modeling seems to be very bright and remarkable with the continuous improvement in technology day by day. Besides further improvements in prediction accuracy, there needs to be a more user-centric approach towards tackling the water quality issues, by involving all the relevant stakeholders, using user-friendly tools and an interactive environment so that the solution actually benefits the target users in tackling water quality issues.

REFERENCES

[1] P. Zeilhofer, "GIS applications for mapping and spatial modeling of urban-use water quality: a case study in District of Cuiabá, Mato Grosso, Brazil," Cad. Saúde …, vol. 23, no. 4, pp. 875–884, 2007.

[2] UN water, "Clean water for a healthy world," Development, pp. 1–16, 2010.

[3] K. Farrell-Poe, "Water Quality & Monitoring," pp. 1–18, 2000.

[4] T. Taskaya-Temizel and M. C. Casey, "A comparative study of autoregressive neural network hybrids," *Neural Networks*, vol. 18, no. 5–6, pp. 781–789, 2005.

[5] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data," *Appl. Soft Comput.*, vol. 23, no. January 2016, pp. 27–38, 2014

[6] Y. Wang, Y. Wang, M. Ran, Y. Liu, Z. Zhang, L. Guo, Y. Zhao, and P. Wang, "Identifying Potential Pollution Sources in River Basin via Water Quality Reasoning Based Expert System," 2013 Fourth Int. Conf. Digit. Manuf. Autom., pp. 671–674, 2013.

[7] S. Maiti and R. K. Tiwari, "A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction," Environ. Earth Sci., vol. 71, no. 7, pp. 3147–3160, 2013.

[8] C. Min, "An Improved Recurrent Support Vector Regression Algorithm for Water Quality Prediction," vol. 12, pp. 4455–4462, 2011.

[9] D. Hou, X. Song, G. Zhang, H. Zhang, and H. Loaiciga, "An early warning and control system for urban, drinking water quality protection: China's experience.," Environ. Sci. Pollut. Res. Int., vol. 20, no. 7, pp. 4496–508, 2013.

[10] A. J, "SAS Global Forum 2008 Data Mining and Predictive Modeling Data mining application of non-linear mixed modeling in water quality analysis SAS Global Forum 2008 Data Mining and Predictive Modeling," Forum Am. Bar Assoc., 2008.

[11] The Environmental and Protection Agency, "Parameters of water quality," Environ. Prot., p. 133, 2001.

[12] C. Leansing, T. Hartvigsen, and J. Reitan, "T He E Ffect of D Ata Q Uality on D Ata M Ining – I Mproving P Rediction a Ccuracy By G Eneric D Ata," Proc. 15th Int. Conf. Inf. Qual., 2010.

[13] Y. Park, K. H. Cho, J. Park, S. M. Cha, and J. H. Kim, "Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea.," Sci. Total Environ., vol. 502, pp. 31–41, Jan. 2015.

[14] S. Song, X. Zheng, and F. Li, "Surface water quality forecasting based on ANN and GIS for the Chanzhi Reservoir, China," 2nd Int. Conf. Inf. Sci. Eng. ICISE2010 - Proc., pp. 4094–4097, 2010.

[15] D. Graupe, "PRINCIPLES OF ARTIFICIAL NEURAL NETWORKS," Advanced Series on Circuits and Systems, vol. 6. World Scientific, University of lllinois, Chicago, USA, 2007.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[16] P. Brockwell and R. Davis, Introduction to Time Series and Forecasting. 2002.

[17] E. Zivot and J. Wang, "Nonlinear Time Series Models," *Model. Financ. Time Ser. with S-PLUS®*, pp. 651–709, 2006

[18] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 479–489, 2010