Forecasting Monthly Water Quality Index Using a Seasonal ARIMA Model for Tigris River at Al-Rashediya Water Station in Baghdad City

Muna Yousif Abdul-Ahad[1], Shaymaa Nashat Subhee[1]
[1] Petroleum and Gas Refining Engineering, Al-Farabi University College, Baghdad , Iraq.
E-mail: Shaymaa.en98@gmail.com

Abstract
In this study, the quality of TGRIS River is studied at the intake of Al-Rashediya Water Station using time series analysis. 14 measured parameters of water quality, daily periods for 9 years (2013-2021), monthly mean averaged were studied which are: K^+, Na^+, T.S.S, T.D.S, $SO4^{2-}$, Cl^-, Mg^{2+}, Ca2+, T.H, Alk., E.C, pH, Turb, and Temp., from which WQI was calculated. Investigation of observed WQI time series shows that there is a simple seasonal behavior. The order of model for WQI time series was determined using auto correlation function (ACF) and partial auto correlation function (PACF). ARIMA (0, 1, 1) (autoregressive, integrated, moving average) model was found suitable to generate and forecast the quality of the river water. The fit statistic for, Stationary R-squared, R-squared, RMSE, MAPE, MaxAPE, MAE, MaxAE, and Normalized BIC criteria were used for evaluating the generation and forecasting results. Their MEAN generated for the model fit were 0.250, 0.338, 106.248, 43.119, 217.295, 73.758, 355.509, 9.419, respectively. The model statistics result for Ljung-Box Q (18) (statistics, DF, and Sig.) were 17.156,17, and 0.444 respectively.
The above results show that time series modeling is quite capable of water quality forecasting.
In this study of the Forecasted WQI model of the becoming 24 months for the years (2022 and 2023) were predicted, shows an increasing trend, which must be considered and managed.

Key words: Water quality index , Time series , ARIMA , forecasting.

1. Introduction

Water quality affects life for its direct impact on human health. salinity, urban and domestic wastewater entrance into surface streams, agricultural drainage, geological structures, ground water usage, and a wide range of chemical compounds throughputs [1],[2]. Different methods and approaches to investigate and forecast the quality of water are used. Also, the majority of water software such as SWAT, QUAL2K MIKE-11, QGIS, SAGA GIS, HEC-RAS, iRIC, PRMS, SPSS, and Python, are used as tools to assess the quality of streams.

Applying time series in water quality modeling and forecasting are very useful methods for understanding and analyzing the process of different phenomena. It is also helpful in generating past observations for forecasting the future values based on the past memory. It is a string of data over time with an equal interval between all data which can be daily, weekly, monthly as well as yearly time steps.

Studies on water quality parameters reviews are available in literature can be useful as references.

Applied Time Series Analysis on surface water quality Hirsch used new methods to analyze monthly water quality data for monotonic trends**[3].** Also, temporal changes in water quality parameters such as pH, Alkalinity, total Phosphorous and Nitrate concentrations have been studied using data series of Niagara[4]. Time-series analysis using ARIMA approaches have been used to examine water quality [5]-[8]. Yu examined surface water quality data of the Arkansas, Verdigris, and Neosho as well as Walnut River basin to study trends in 17 major

constituents using 4 different nonparametric methods[9]. Robson and Neal studied the trend of ten-year upland stream and bulk deposition water quality data from Plynlimon, mid Wales through the seasonal Kendall test [10]. Time series analysis is used to understand and model the stochastic mechanism of hydrologic phenomena and to forecast the future values of the phenomena [11].

Time series forecasting is done with the help of ARIMA. In ARIMA model, AR stands for auto-regression and MA stands for moving average. In ARIMA the non-seasonal part is represented using (p.d, q) where p is the number of autoregressive values, d is the order of differencing and q is the number of moving average values [12].

Primarily Graphical and statistical time series techniques have been used to analyze the trends and specified time changes, in river water quality data. The information obtained may be associated with some socio-economic variables, such as industrial or agricultural development, urban increase and wastewater discharge around or upstream of the measure station. Such a study may now be applied to more rural stations in order to compare the evolution of water quality and perhaps, historical monthly average values to evaluate the seasonality effect on annual trends [13].

Many studies focused on water quality parameters by time series, are: Applied Time Series Analysis on Surface Water Quality Hirsch used new methods to analyze monthly water quality data for monotonic trends[3]. Also, temporal changes in water quality parameters such as K^+, Na^+, T.S.S, T.D.S, $SO4^{2-}$, Cl^-, Mg^{2+}, Ca^{2+}, T.H, Alk., E.C, PH, Turb, and Temp. concentrations have been studied using data series of Niagara [4] . Yu analyzed surface water quality data of the Arkansas, Verdigris and Neosho as well as Walnut River basin to study trends in 17 major constituents using 4 different nonparametric methods**[9].** The trend of upland stream and water quality data from Plynlimon, mid Wales were examined. Robson and Neal applying the seasonal Kendall test[10]. studied the time series of water quality parameters and the discharge of Strymon River in Greece from 1980 to 1997. Gangyan investigated the temporal sediment load characteristics of the Yangtze River using the turning point test, Kendall's rank correlation test[14]. Jassby developed a time series model for Secchi depth in Lake Tahoe, USA[15]. Panda studied the trends in sediment load of a tropical river basin in India[16].

The objective of this study is forecasting using time series analysis (ARIMA modelling)for the becoming years (2022-2023) to provide information on the physicochemical characteristics of Tigris River water quality within Al-Rashediya station in Baghdad city, and the impacts of unregulated waste discharge on the quality of the river as well as to discuss its suitability for human consumption based on predicted and forecasted water quality index values (WQI),where a large data matrix, obtained during 9-years (2013-2021) monitoring program, is subjected to time series analysis (ARIMA modeling) technique.

2. Water quality Index Application and Formulation

In the formulation of WQI, the importance of various parameters depends on the intended use of water, and its suitability for human consumption. The standard permissible values of various parameters for the drinking water used in this study are those recommended by the Iraqi drinking water standards (Drinking-Water Standard IQS: 417,2001), and by the (World Health Organization WHO, 2004).

For calculating the Water Quality Index, a set of fourteen water quality parameters have been collected from Al-Rashediya station. The overall Water Quality Index (WQI) was calculated using the Weighted Arithmetic

Index method from (Flaieh et al. (2014), and Kizar (2018)). The quality rating scale for each parameter (qi) was calculated by using Eq. (1):

$$q_i = \left(\frac{c_i}{s_i}\right) \times 100 \quad (1)$$

A quality rating scale (qi) for each parameter is assigned by dividing its observed concentration (Ci) in each water sample by its respective standard value (Si) and the result is multiplied by 100. Relative

$$w_i = \frac{1}{s_i} \quad (2)$$

The overall Water Quality Index (WQI) was calculated by aggregating the quality rating scale (qi) with the unit weight (w_i) linearly in Eq. (3) as follows:

$$WQI = \left(\sum_{i=1}^{n} w_i \times q_i\right) \quad (3)$$

Generally, WQI is to be discussed for a specific and intended use of water. In this study the WQI for drinking purposes is considered a permissible WQI if its value is 100 using Eq. (4):

$$\text{Over all WQI} = \frac{\left(\sum_{i=1}^{n} w_i \times q_i\right)}{\sum w_i} \quad (4)$$

The data used to calibrate and validate the time series analysis were collected from Baghdad Water Governorate. In Al-Rashediya station, data were collected from raw water (river water near intakes of the water treatment plant). SPSS© v.26, and Microsoft Office Excel© 2021, software packages were used to implement all the mathematical and time series analyses.

3. Applied time-series analysis

Time-series analysis using ARIMA approaches have been used to examine water quality **[5]-[8]** ARIMA models are capable of reproducing the main statistical characteristics of a hydrologic or environmental time series.

Time series models used to generate synthetic time series can be classified into autoregressive models (AR (p)), moving average models (MA (q)), and their combination, autoregressive moving average (ARMA (p, q)) with variations, such as autoregressive, integrated moving average (ARIMA) models (p, d, q) and others, where p and q are the orders of autoregressive and moving average terms, respectively, and "d " is an order of differencing.

Many studies have been written about the theory of ARIMA modeling as well as its applications (Pankratz, 1983; Vandaele, 1983; Nelson, 1973; Box and Jenkins, 1976).

The basic stages in ARIMA modeling are composed of:

(1) identifying the autocorrelation and partial autocorrelation of time series,

(2) estimating the orders of the identified model.

(3) verify the model through standard tests.

(4) At the next stage forecasting the water quality parameters can be done.

4. Methodology

The data obtained from the Al-Rashediya station observed daily for years (2013-2021), which were transferred to an average monthly data, and their descriptive statistics will be presented.

The Expert Modeler in SPSS will be used to estimate an appropriate ARIMA model based on the data for the monthly water quality index. The procedure for the model is based on a series of steps, including proper transformation and differencing, detection of ARIMA pattern, estimation of the parameters, and diagnostic checking of the residuals through the Ljung-Box statistic. Based on the procedure of the Expert Modeler, the fitted model for the data will be validated. The Box-Jenkins

methodology will be followed to validate the suggested model. For pattern detection, which is the first step of the Box-Jenkins methodology, the correlograms of the ACF and PACF with the 95% confidence interval will be used. To test if the model fulfils the requirements of the third step of the Box-Jenkins methodology, the Ljung-Box statistic will be used. five percent significance level will be used and the critical value is extracted from the chi-squared distribution table. To illustrate goodness of fit, Akaike's information criterion will be used. The estimated ARIMA model will be used to forecast 108 months of the occurred water quality indices, in a span of years (2013-2021). the out-of-sample forecasts, the mean percentage error (MPE) and the mean absolute percentage error (MAPE) will be computed the for-forecasting accuracy of the fitted model. the line charts of the forecasts will be displayed in the analysis.

5. Analysis

Descriptive Statistics & Line Charts

Table 1 presents the descriptive statistics for WQI. the data for each index consist of 108 months over a span of years(2013-2021). Figure 1 illustrates the line chart where WQI is plotted against time. The line chart shows that the mean and variance are changing over time, i.e. they are non-stationary. The decaying autocorrelation as lags increase, illustrated in the ACF and PACF correlograms in Figure2 and figure3 for WQI, further supports that the indices are non-stationary.

Table 1. The Descriptive Statistics for WQI.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
wqi.r	108	810.76	59.08	869.84	183.8009	12.49844	129.88761	16870.791	2.701	.233	9.409	.461
Valid N (listwise)	108											

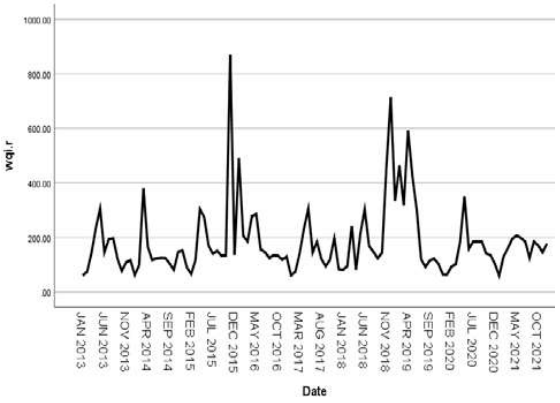


Fig 1. The line chart of WQI against time.

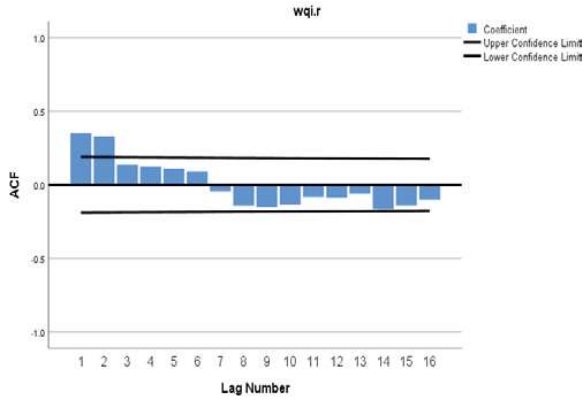


Fig 2. ACF correlograms for the line chart of WQI against time.

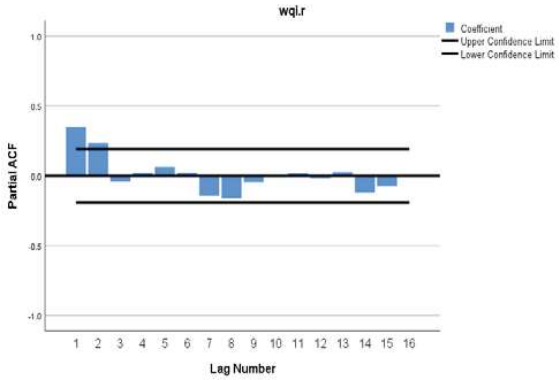


Fig 3. PACF correlogram for the line chart of WQI against time .

Modeling the WQI

For WQI, the Expert Modeler proposes an ARIMA (0,1,1) model (Table 2) , and shows the model statistics(Table3) with a natural

Table 3 The model statistics

Model Statistics

Model	Number of Predictors	Model Fit statistics: Stationary R-squared	R-squared	RMSE	MAPE	MAE	MaxAPE	MaxAE	Normalized BIC	Ljung-Box Q(18): Statistics	DF	Sig.	Number of Outliers
wqi.r-Model_1	0	.250	.338	106.248	43.119	73.758	217.295	355.509	9.419	17.156	17	.444	1

logarithmic transformation, to reduce the influence of outliers, as the best fit (Table 4). In the model, there is no autoregressive component, one moving average component, and the model is integrated of the first order. the ARIMA (0,1,1) for WQI is validated by Box-Jenkins's methodology.

Table 2. Model description

Model Description			Model Type
Model ID	wqi.r	Model_1	ARIMA(0,1,1)(0,0,0)

Table 4. The Best Fit, with a natural logarithmic transformation, to reduce the influence of outliers.

Outliers

			Estimate	SE	t	Sig.
wqi.r-Model_1	Nov 2015	Additive	1.705	.383	4.446	.000

The Expert Modeler suggests that the natural logarithmic transformation of WQI is stationary after the first difference, which is shown in Figure 4.

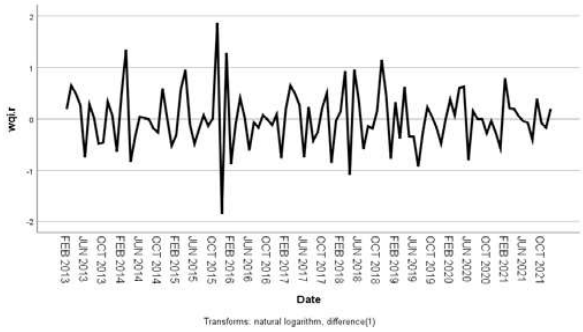


Fig 4. the natural logarithmic transformation of WQI ($lnWQI_t$—$ln\ WQI_{t-1}$ against time)

According to the line chart of figure 4, the mean and variance are constant over time and the covariance is time invariant, i.e. the time series is stationary.

Figure 5 presents the correlogram for the ACF of $WQI_t - WQI_{t-1}$, where the resemblance of a white noise process further confirms that the time series is stationary.

The correlograms of the ACF and the PACF of the stationary process for WQI illustrated in Figure 5 and Figure 6, respectively, shows a significant spike at the first lag in the ACF and an exponential decline as lags increase in the PACF, confirming that the time series is a MA (1) process. The patterns in the correlograms suggests that there should not be an autoregressive process included in the ARIMA model for WQI.

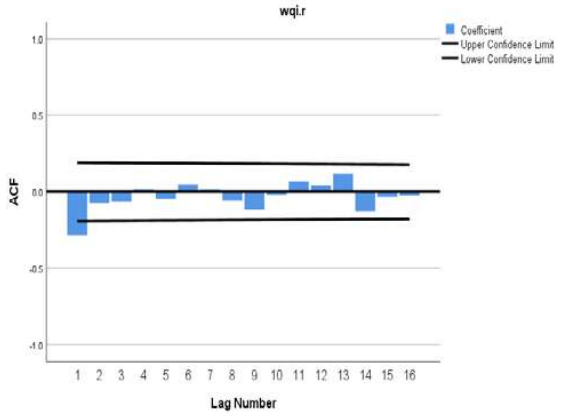


Fig 5. The ACF correlogram of lnWQI$_t$—ln WQI$_{t-1}$ with the 95% confidence limit.

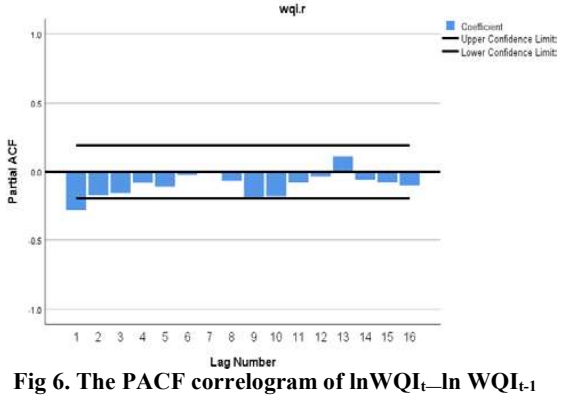


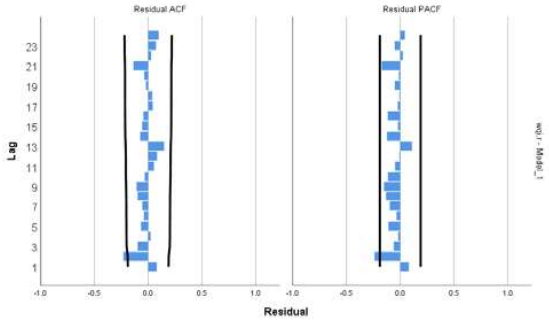
Fig 6. The PACF correlogram of lnWQI$_t$—ln WQI$_{t-1}$ with the 95% confidence limit.

in the Box-Jenkins methodology after the identification of p, d and q, is to estimate the model's parameters. The parameter for the ARIMA (0, 1, 1) of WQI is statistically significantly different from zero, and the estimated parameter is presented in Table 5.

Table 5. ARIMA model parameters

ARIMA Model Parameters								
					Estimate	SE	T	Sig.
wqi.r-Model_1	wqi.r	Natural Logarithm	Difference		1			
			MA	Lag 1	.300	.093	3.219	.002

The ACF and PACF correlograms of the residuals for the ARIMA model, presented in Figure 7 (a) and (b), further confirms that the residuals follow a white noise process since the lags are around zero, the variance is constant, no serial correlation, and most lags are statistically insignificant. The white noise residuals in the Box-Jenkins methodology, and the ARIMA (0, 1, 1) of WQI may be used to forecast the time series.



(a) (b)

Fig 7. (a) The ACF correlogram of the residuals for the ARIMA (0, 1, 1) of WQI with the 95% confidence interval, (b) The ACF correlogram of the residuals for the ARIMA (0, 1, 1) of WQI with the 95% confidence interval.

The ARIMA (0, 1, 1) is used to forecast the 24 months of the years (2022-2023) for the WQI as in figure8.

the out-of-sample forecast accuracy of the ARIMA(0, 1, 1) In Table 6, the ARIMA model of WQI with its corresponding AIC, MPE and MAPE are presented.

In Table 6, the forecasted values of the 24 months of years(2022-2023) estimated with the ARIMA(0, 1, 1) and the actual outcomes are presented for each month. In Figure 8, a visual demonstration of the fitted values in relation to the observed values for WQI is shown. The figure also displays the forecasted values of the 24 months of years (2022-2023) For WQI.

Table 6. Actual outcomes and forecasts for the WQI ARIMA(0, 1, 1).

FORECAST													
MODEL		Jan-22	Feb-22	Mar-22	Apr-22	May-22	Jun-22	Jul-22	Aug-22	Sep-22	Oct-22	Nov-22	Dec-22
WQI.R-MODEL_1	Forecast	189.54	200.3	211.69	223.71	236.43	249.86	264.06	279.06	294.92	311.68	329.39	348.1
	UCL	434.38	534.65	637.25	744	855.92	973.72	1097.95	1229.08	1367.52	1513.67	1667.91	1830.63
	LCL	65.98	53.61	44.98	38.52	33.49	29.44	26.11	23.32	20.96	18.94	17.18	15.66
MODEL		Jan-23	Feb-23	Mar-23	Apr-23	May-23	Jun-23	Jul-23	Aug-23	Sep-23	Oct-23	Nov-23	Dec-23
WQI.R-MODEL_1	Forecast	367.88	388.78	410.88	434.22	458.9	484.97	512.53	541.65	572.43	604.95	639.33	675.65
	UCL	2002.18	2182.96	2373.33	2573.69	2784.43	3005.95	3238.66	3482.97	3739.32	4008.14	4289.9	4585.03
	LCL	14.32	13.13	12.08	11.14	10.29	9.54	8.85	8.23	7.67	7.15	6.68	6.25

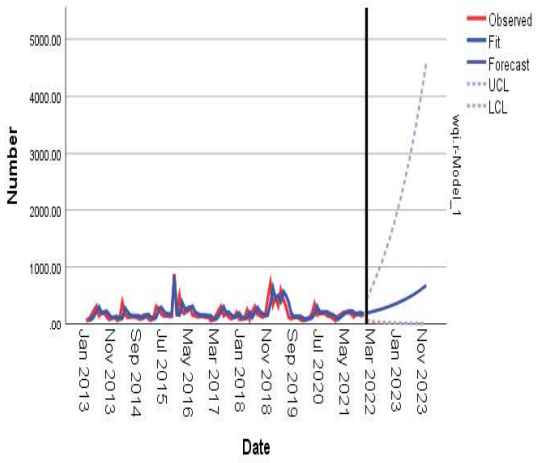


Fig 8. Line chart of WQI, the fitted values of the ARIMA (0, 1, 1), and the forecasts.

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

6. Conclusion

In this study, the best-fitted ARIMA model for WQI was estimated using the Expert Modeler in SPSS(ARIMA modeling). The model was used to forecast the 108 months of years(2013-2021) for the WQI. To validate the model, the Box-Jenkins methodology was followed. When the Expert Modeler estimates an appropriate ARIMA model for a time series, one of the steps in the procedure is to determine the stationary process of the time series, followed by pattern detection in the ACF and PACF. Furthermore, diagnostic checking of the models' residuals was conducted through the Ljung-Box statistic and graphical analysis of the correlograms of the residuals. The residuals of the model resembled white noise process and the model could therefore be used to forecast the TWENTY-FOUR months of YEARS (2022- 2023) for WQI. The validation of the suggested model by following the Box-Jenkins methodology, advocates that the Expert Modeler was appropriate in estimating ARIMA model for the used time series. the best-fitted model in this study assumed to be sufficiently accurate in forecasting WQI for the years (2022-2023). it is possible to conclude that a well-informed guess could be better than a random guess. Further research should focus on comparison of forecasting performances of models such as exponential smoothing to the forecasting performance of an ARIMA model. When forecasting WQI, different time periods or different observational frequencies could be of interest and could give better forecasting results.

References

[1] G. a. D.Alexakis, "Water quality models: An overview," *European Water ,* vol. 37, pp. 33-46, 2012.

[2] M. ,. P. ,. D. A. G. Tsakiris, "Assessing the water potential of karstic saline springs by applying a fuzzy approach: The case of Almyros (Heraklion, Crete).," *Desalination ,* vol. 237, pp. 54-64, 2009.

[3] D. A. V. a. G. Papamichail, "s,P.E. Stochastic models for Strymon river flow and water quality parameter," in *Proc. of International Conference*, 2000.

[4] S. Z. ,. D. W. Yun-Sheng Yu, "Non-parametric trend analysis of water quality data of rivers in Kansas," *Journal of Hydrology ,* vol. 150 , no. 1, pp. 61-80, September 1993.

[5] C. N. A. J. Robson, "Water quality trends at an upland site in wales, UK, 1983–1993," *Hydrological processes,* vol. 10, no. 2, pp. 183-203, February 1996.

[6] M. G. P. G. K. V. Abdollah Taheri Tizro, "Time series analysis of water quality parametersournal of Applied Research in Water and Wastewater," *Journal of Applied Research in Water and Wastewater,* vol. 1, pp. 43-52, 2014.

[7] D. F. B. Naresh Patnaik, "Weather Forecasting in Coastal Districts of Odisha and Andhra Pradesh by Using Time Series Analysis," *International Journal of Emerging Research in Management &Technology,* vol. 6, no. 1, 2017.

[8] G. a. R. V. C, "Time series analysis on chlorides, nitrates, ammonium and dissolved oxygen concentrations in the Seine," *The Science of the Total Environment ,* vol. 208, no. 1-2, pp. 59-69, 1997.

[9] N. K. J. a. V. B. Zhou Gangyan, " Hydrological Sciences Journal," *Stochastic modelling of the sediment load of the upper Yangtze River (China),* Vols. 93-105, p. 47, 2002.

[10] J. E. R. a. C. R. G. Alan D Jassby, "Determining long-term water quality change in the presence of climatic variability: Lake Tahoe (USA)," *Canadian Journal of Fisheries and Aquatic Sciences,* vol. 60, pp. 1452-1461, December 2003.

[11] A. a. S. Dileep K.Panda, "Recent trends in sediment load of the tropical (Peninsular) river basins of India," *Global and Planetary Change,* vol. 75, no. 3–4, pp. 108-118, February 2011.

[12] M. S. ,. S. P. a. D. A. George Tsakiris, "Assessing the water potential of karstic saline springs by applying a fuzzy approach: The case of Almyros (Heraklion, Crete)," *Desalination,* vol. 237, pp. 54-64, 2009.

[13] G. a. D.Alexakis, "Water quality models: An overview," *European Water,* vol. 37, pp. 33-46, 2012.

[14] Hirsch, R. M., Slack, J. R., Smith, R. A.,, "Techniques of trend analysis for monthly water quality data," *Water Resources Research,* vol. 18, no. 1, pp. 107-121, 1982.

[15] S. ,. K. A.H.El-Shaarawi, "A Statistical Evaluation of Trends in the Water Quality of the Niagara River," *Journal of Great Lakes Research,* vol. 9, no. 2, pp. 234-240, 1983.

[16] A. a. M. R. Lehmann, "Long-term behaviour and cross-correlation water quality analysis of the River Elbe Germany," *Water Research,* vol. 35, p. 2153–2160, 2001.

[17] D.O.Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," *Engineering Applications of Artificial Intelligence,* vol. 23, p. 586–594, 2010.

[18] P. Hanh, "Analysis of variation and relation of climate, hydrology and water quality in the lower Mekong River," *Water Science Technology,* vol. 62, no. 7, p. 1587–1594, 2010.