Sustainable Water Resources Management (2020) 6:74
https://doi.org/10.1007/s40899-020-00430-7

ORIGINAL ARTICLE

Simulation and prediction of surface water quality using stochastic models

Mostafa Dastorani[1] · Mohammad Mirzavand[2] · Mohammad T. Dastorani[3] · Hassan Khosravi[4]

Received: 1 July 2019 / Accepted: 8 July 2020

Abstract
In recent years, surface water quality has decreased due to the increasing demand for water and increasing the use of fertilizers, pesticides and the discharge of domestic and municipal wastewater to surface water. The purpose of this research is a comparison of the efficiency of different time-series models in modeling and prediction of monthly water quality performance in Harmaleh area of Khuzestan in the southwest of Iran. Water quality parameters including Ca, HCO_3, SO_4, Ec, pH, Mg, Cl, Na, and TDS for the period of 2001 to 2014 were evaluated. Five time-series models (AR, MA, ARMA, ARIMA, and SARIMA) with 12 different structures were assessed by R software. First, the data were normalized using Kolmogorov–Smirnov test. Also, the adequacy of data was tested by Hurst's coefficient. The Hurst coefficient was > 0.5 for all investigated parameters, which indicated suitable length of the time series for the modeling. As the components of trend, jump, and seasonality are usually specific, modeling of them is not required, but modeling of stochastic components is of importance in water resources simulation and management. Therefore, using the R software, deterministic parts of the time series (e.g., trend, jump, and seasonality) were eliminated and non-deterministic component (e.g., randomness) was simulated (from 2011 to 2014), and finally, the data were predicted (from 2015 to 2018) based on the optimized models. The optimized models were selected based on auto-correlation function (ACF) and partial auto-correlation function (PACF) as well as the use of Akaike information criteria (AIC) and coefficient of determination. Results showed that in 66% of data ARMA [with the same rate of ARMA (1, 2), ARMA (2, 1), and ARMA (2, 2)], in 22% of data AR (1), and in 11% of data ARIMA (1, 1, 2) models presented the highest efficiency in monthly water quality simulation. Finally, each quality parameter was also predicted for the next 4 years (2015–2018) based on the selected optimized models. Results indicated that the values of SO_4 and pH, respectively, showed the highest and lowest correlation with the related observations with a coefficient of determination of 0.54 and 0.19. Overall, modeling of water quality using stochastic models could save time and costs, especially when time series of parameters are long and adequate.

Keywords Water quality · Stochastic models · Simulation · Prediction

Introduction

There is water all around us. There are vast oceans, large lakes, big rivers, small ponds, and tiny streams. All of these matter to us and other creatures on this planet. Our river systems connect to make watersheds. Little streams feed large rivers, which can then feed lakes or oceans. The contamination of the smallest stream will affect everything downstream. We often get our drinking water from lakes and rivers. Although we treat our drinking water, we should protect it at its initial source: up to the tiniest stream. Lots of other animals and plants depend upon the watersheds which we inhabit. Cleaner water means a healthier food chain (from bugs to fish, to birds, to people), and poor water quality

✉ Mostafa Dastorani
m.dastorani@hsu.ac.ir

[1] Faculty of Geography and Environmental Sciences, Hakim Sabzevari University, Sabzevar, Iran

[2] Natural Resources and Earth Sciences, University of Kashan, Kashan, Iran

[3] Faculty of Natural Resources and Environment, Ferdowsi University of Mashhad, Mashhad, Iran

[4] Department of Arid and Mountainous Regions Reclamation, Faculty of Natural Resources, University of Tehran, Tehran, Iran

Published online: 03 August 2020

tends to affect disadvantaged communities disproportionately (Mirzavandet al. 2012).

Water is essential to human life and the health of the environment. As a valuable natural resource, it comprises marine, estuarine, freshwater (river and lakes), and groundwater environments that stretch across coastal and inland areas. Water has two dimensions that are closely linked: quantity and quality. Water quality is commonly defined by its physical, chemical, biological, and esthetic (appearance and smell) characteristics. A healthy environment is one in which the water quality supports a rich and varied community of organisms and protects public health. Water quality in a body of water influences the way in which communities use the water for activities such as drinking, swimming, or commercial purposes (Mirzavand et al. 2020a, b).

In the twenty-first century, a decrease in water quality is one of the major concerns in the world (Perrin et al. 2014). Surface water quality is a significant concern in the world at all levels of public and decision makers (He et al. 2015). Consequently, efficient management of water quality becomes very crucial (Singh Patel and Ramachandran. 2015). Variations in water quality and quantity within a regional or local basin can be attributed directly or indirectly to humans activities, e.g., climate change (Delpla et al. 2009; Piao et al. 2010; Solheim et al. 2010; Whitehead et al. 2009), land-use type (Tong and Chen. 2002; Wan et al. 2014), and industrial or agricultural activities (Carr and Neary 2006; Dabrowski et al. 2009). Temporal and spatial variations of surface water quality have been evaluated by many researchers (Antonopoulos et al. 2001; Luo et al. 2011; Pejman et al. 2009; Singh et al. 2004). Recognizing, modeling, and forecasting of surface water quality which enable long-term planning are essential for more efficient utilization of water (Mirzavand and Ghazavi 2015).

Many methods such as ANN, fuzzy, ANFIS, and other techniques have been used for water quality modeling (Nayak et al. 2004; Partal and Kisi 2007; Debele et al. 2009; Yeniguna and Ecer 2012; Boskidis et al. 2012; Dökmen and Aslan 2013). When we have just one series of data (dependent variable), stochastic time-series theory is usually applied to predict hydrological parameters (Bras and Rodriguez Iturbe 1985; Brockwell and Davis 2010; Lin and Lee 1992). Stochastic time-series models are accessible and useful tools for medium-range forecasting and for generating synthetic data (Adhikary et al. 2012). Time-series method can be used to evaluate the temporal variation of surface water quality. A time-series model is, in fact, an empirical model for stochastically simulating and forecasting the behavior of uncertain hydrological systems (Kim et al. 2005). A number of stochastic time-series models such as the Markov, Box–Jenkins (BJ) seasonal auto-regressive integrated moving average (ARIMA), depersonalized auto-regressive moving average (ARMA) (Mirzavand and Ghazavi 2015; Dasturani et al. 2016), periodic auto-regressive (PAR), transfer function noise (TFN), and periodic transfer function noise (PTFN), are usually used in this regard (Brockwell and Davis 2010; Box et al. 1994; Hipel and Mcleod 1994). The selection of an appropriate method for modeling a particular problem depends on factors such as the number of series to be modeled, required accuracy level, modeling costs, ease of use of the models, and ease of interpretation of the results (Mondal and Wasimi 2007).

The main aim of this study is to evaluate the performances of time-series models (AR, MA, ARMA, ARIMA, and SARIMA) for surface water quality forecasting in semiarid environment of Harmaleh area of Khuzestan in the southwest of Iran. The finding of this research can help to predict the quality of waters for different uses with acceptable accuracy using appropriate models.

Materials and methods

The study area

The study area in this research is the Harmaleh area located in Khuzestan province of Iran between longitude 47° 42′ 50° E and latitude 29° 56′ 33° N (Fig. 1). Khuzestan province with the area of 64,057 km^2 in the southwest of Iran has been surrounded by the Zagros Mountains from its north and east directions. Altitude varies from zero to more than 4000 m a.s.l, and it is one of the wealthiest provinces of Iran in water resources because five large rivers originated from Zagros Mountains irrigate Khuzestan plain and then are directly or indirectly drained to the Persian Gulf. The climate condition in this province is mostly semi-desert although due to variations of altitude, temperature, and precipitation in north, south, and west parts of it, warm as well as moderate climate conditions are seen different parts of this province. Annual average rainfall in southwest region is about 150 mm, in central parts is about 200 mm, and in north parts is about 300 mm, and toward the higher elevation of northern east parts, annual average precipitation reaches the amount of 1000 mm. The temperature is relatively high, and the monthly mean in July gets more than 30 °C. Due to the climatic condition and soil diversity, some parts of Khuzestanis are covered with relatively spars forestlands with shrubs, small trees, and trees as well as rangelands. The geology formation of Harmaleh is conglomerate of Bakhtiari Formation; Quaternary deposits and gray thick-bedded to massive Orbitolina limestone were observed in the river direction. The main land uses in Harmaleh are farmland (mainly wheat in winter and spring and corn and vegetables in summer and autumn).

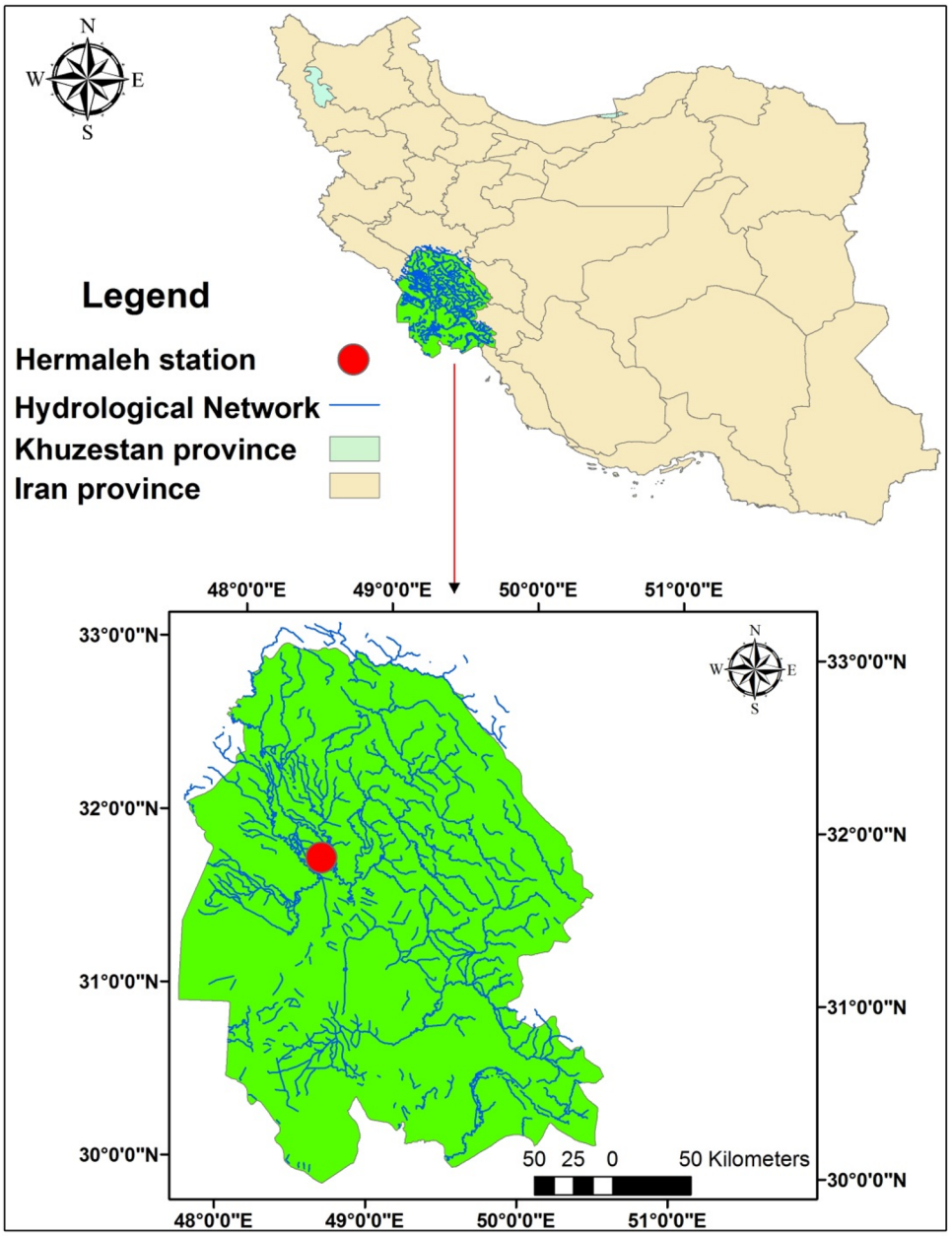

Fig. 1 The study area of Harmaleh in Khuzestan province located in southwestern Iran

Data collection and analysis

In order to assess the stochastic models ability in water quality prediction, data collected and analyzed by Ministry of Energy of Iran were used. The parameters of Ca, HCO_3, SO_4, Ec, pH, Mg, Cl, Na, and TDS from 2001 to 2014 were used. During 2001 to 2014, surface water samples from Harmaleh station were collected (Fig. 1 and Table 1). Before sampling, the bottles were rinsed with distilled water. Water samples were collected in 250-mL PVC bottles for anions and cations analysis. The electrical conductivity (EC) and pH values were measured in the field by portable HACH-150 and Metrohm instrument. After sampling, the cations samples were filtered with a Millipore filter of 0.45 µm and acidified to pH < 2 using pure HCl. Samples were stored at 4 °C and transferred to the Ministry of Energy of Iran Laboratory. The anions and cations were analyzed by ion chromatography (IC) and inductively coupled plasma mass spectrometry (ICP-MS). In addition, CO_3^{2-} and HCO_3^{-} were measured by digital titration using 0.16 N sulfuric acid and phenolphthalein and bromocresol green-methyl red powder indicators, respectively (Table 1). The computed charge balance error (Eq. 1) is within the acceptable range of ± 5% (Clark 2015).

Table 1 Statistical characteristics of physicochemical parameters and major ion concentrations of Harmaleh station (concentration in mg/L)

Date	Statistical characteristics	pH	Ec (μS/cm)	TDS	Na	Ca	Mg	Cl	ALK (HCO_3)	SO_4
2001	Max	8.3	1494	1174	5.6	15	4.44	4.32	4.45	7.71
	Min	7.6	550	375	1.5	6	1.7	1.4	2.26	2.22
	Mean	8	995	678	3.24	10	2.81	2.81	3	4.32
2002	Max	8.5	1236	988	3.84	5.9	4.37	3.31	3.63	5.79
	Min	7.5	536	351	1.3	3	1.17	1.24	2.1	1.17
	Mean	8	829	563	2.65	3.92	2.33	2.41	2.96	3.10
2003	Max	8.2	1176	820	3.84	5.5	3.4	2.93	4.33	5.2
	Min	7.3	515	338	1.3	2.91	0.86	1.1	3	0.84
	Mean	7.88	821	560	2.63	3.94	2.29	2	3.48	3.22
2004	Max	8.2	1110	767	4	5.45	3	3.75	3.61	5.4
	Min	7.7	610	390	1.45	3.5	0.99	1.51	2.33	1.2
	Mean	8	804	549	2.59	4	1.90	2.43	3	2.95
2005	Max	8.3	1390	970	5.96	6.15	4	5.98	4	6.5
	Min	7.6	569	386	1.21	2.15	0.56	1.06	1.54	1.2
	Mean	8	946	673	2.95	4.23	2.51	2.81	3.10	3.79
2006	Max	8.6	1277	894	3.3	8.61	6.11	3.43	3.91	5.33
	Min	7.4	341	243	1.2	1.24	0.24	1.3	1.14	1.4
	Mean	7.88	859	550	2.38	3.90	2.39	2.49	3	2.90
2007	Max	9	1342	939	5.53	5.31	4	4.95	3.63	7.33
	Min	7.4	440	242	0.94	3.14	1.32	1.45	2.56	0.82
	Mean	7.9	843	565	2.74	3.80	2.56	2.57	3.10	3.29
2008	Max	8.3	1348	990	5.29	6	5.12	5	5.7	6.88
	Min	6.8	898	494	2.15	2.72	0.90	2.94	1.99	0.99
	Mean	7.49	1151	756	4.23	4.89	3.22	3.62	3.81	4.78
2009	Max	8.2	1673	1255	6.11	7.9	5.83	6.12	4.44	8.74
	Min	7	883	600	2.94	2.94	2.05	2.74	2.78	2.69
	Mean	7.62	1342	924	4.95	5.78	3.47	4.55	3.55	6.12
2010	Max	8.1	1758	1145	8.3	6.8	5.15	6.9	4	7.43
	Min	7	657	420	2.12	2.9	1.25	2	1.89	1.47
	Mean	7.83	1109	727	3.90	4.77	2.97	3.85	3.14	4.31
2011	Max	8	1449	1096	5.21	9.23	4.1	4.5	4.63	7
	Min	7.5	869	560	2.93	3.36	1.36	2.94	3.11	2.2
	Mean	7.71	1204	834	4	5.81	3	3.70	3.95	5
2012	Max	8.1	1644	1150	6.75	7.44	9.69	6.22	4.95	10
	Min	7.4	520	338	1.18	2.54	1.55	1.13	2.41	2.6
	Mean	7.77	1330	894	4.85	5.26	4.63	4.38	3.51	6.62
2013	Max	8	1587	1100	5.47	8.11	6.8	4.85	4.32	8.48
	Min	7	1052	796	3.87	5	3	3.63	0.34	4.23
	Mean	764	1334	960	4.78	6.30	4.04	4.18	3.60	6.55
2014	Max	8	1515	1100	5.47	7.4	4.38	4.5	4.3	8.44
	Min	7.3	754	421	2	3.76	2	2.2	2.87	1.47
	Mean	7.67	1232	854	4.29	5.70	3.45	3.73	3.81	5.69

$$\mathrm{CBE} = (\%) = \frac{\sum \mathrm{cat} - \sum \mathrm{an}}{\sum \mathrm{cat} + \sum \mathrm{an}} \times 100. \quad (1)$$

Methodology

Data pre-processing

First, data adequacy test was performed using Horst's test. The Horst coefficient equal to 0.5 shows the independent and normal time series. Also, since the coefficient was more than 0.5, the data had no problem with data length. The Horst coefficient can be calculated using Eq. (2).

$$K = ((Log(R//(Log(N/2)) \quad (2)$$

$$R = S^{+} - S^{-}$$

$$S_n = \sum_{k=1}^{n} \left(X_k - \overline{X}\right),$$

where N is the number of data for each parameter and σ is the standard deviation. For calculation of S^{+} and S^{-}, the average of each parameter was minus from each data series and the cumulative series were plotted, and the maximum and minimum values were determined as S^{+} and S^{-}, respectively (Mirzavand and Ghazavi 2015).

After that, data were normalized using Kolmogorov–Smirnov test (P value < 0.5). Then, $N/4$ of the data were selected as a testing part and remaining ones used for simulation and optimized model selection. Hence, using the R software, time series were decomposed into deterministic components (jump, trend, and seasonality) and non-deterministic component (randomness). Then, modeling was done based on the randomness component. After that, monthly simulation was done (from 2011 to 2014) for determining the optimized models. Finally, monthly prediction was done (from 2015 to 2018) based on the optimized models for different hydrogeochemical parameters.

Modeling

Generally, the models for time-series data can have different forms and represent different non-deterministic processes (Sokolnikov 2013). Most modeling of time series takes place based on a linear technique. For example, AR, MA, and ARMA models have linear base (Klose et al. 2004). In this research, AR, MA, ARMA, ARIMA, and SARIMA models on 12 different structures based on trial and error were examined and used to assess the ability of these models in monthly water quality indicator prediction.

AR model In a series where persistence is present, the event outcome of $(t+1)$th period is dependent on the current tth period magnitude and those preceding values; then for such a series, the observed sequence $X_1, X_2, \ldots, X_t$ is used to fit an AR model.

The auto-regressive model can be expressed as Eq. (3):

$$z_t = \emptyset_1 z_{t-1} + \emptyset_2 z_{t-2} + \cdots + \emptyset_p z_{t-p} + a_t, \quad (3)$$

where $\emptyset_1, \emptyset_2, \ldots, \varphi_p$ are model parameters and coefficients and a_t is the random component of the data that follows a normal distribution with mean, p is the desired delay and z, is desired time series data (Mirzavand and Ghazavi 2015).

MA model Moving average models are simple covariance stationary and ergodic models that can be used for a wide variety of auto-correlation patterns (Mirzavand and Ghazavi 2015).

Moving average (MA) model can be expressed as Eq. (4):

$$Z_t = \delta + \sum_{i=1}^{p} \varphi_i z_{t-1} + \sum_{j=1}^{q} \phi_j e_{t-j} + e_t \quad (4)$$

where $\theta_1, \theta_2, \ldots, \theta_q$ are model parameters and Q coefficient model and a_t is the random component of the data that follows a normal distribution with mean (Hannan 1971).

ARMA model The ARMA model is a synthesis of AR and MA models. ARMA model forms a type of linear models which are widely applicable and parsimonious in parameterization. ARMA (p, q) model can be expressed as Eq. (5):

$$Z_t = \delta + \sum_{i=1}^{p} \varphi_i z_{t-1} + \sum_{j=1}^{q} \phi_j e_{t-j} + e_t, \quad (5)$$

where δ is the stationary part of the ARMA model, φ_i points out the ith autoregressive coefficient, ϕ_j is the jth moving average coefficient, it shows the error part at period t, and Z_t refers to the value of water quality indicator observed or predicted at time period t (Erdem and Shi 2011; Behnia and Rezaeian 2015).

ARIMA and SARIMA models Auto-regressive integrated moving average (ARIMA) models are one of the well-known linear models for time-series modeling and prediction. ARIMA models have been originated from the synthesis of AR and MA models. ARIMA is used to model time-series data behavior and to make predictions (Shirmohammadi et al. 2013). ARIMA modeling uses correlational methods and could be used to model arrays that may not be observable in plotted data (Box et al. 1994). In ARIMA, the future

amount of a parameter is assumed to be a linear function of past observations and random errors (Behnia and Rezaeian 2015). A SARIMA model can be explained as ARIMA (p, d, q) (P, D, Q)s, where (p, d, q) is the non-seasonal component of the model and (P, D, Q)s is the seasonal component of the model in which p is the order of non-seasonal auto-regression, d is the number of regular differencing, q is the order of non-seasonal moving average, P is the order of seasonal auto-regression, D is the number of seasonal differencing, Q is the order of sesonalal moving average, and s is the length of season (Faruk 2010).

Model selection

To determine the best model, partial auto-correlation function (PACF) and auto-correlation function (ACF) are usually used in most of the carried out researches. But, to improve the model selection accuracy, Akaike information criteria (AIC) and Coefficient of determination (R^2) have also been used in this research in addition to PACF and ACF.

AIC and R^2 can be expressed as Eqs. (6) and (7) (Hu 2007):

$$\mathrm{AIC}(k) = nln(\mathrm{MSE}) + 2k \tag{6}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(qi-\overline{q}\right)\left(\widehat{qi}-\overline{\hat{q}}\right)\right]^2}{\sum_{i=1}^{n}\left(qi-\overline{q}\right)^2\sum_{i=1}^{n}\left(\widehat{qi}-\overline{\hat{q}}\right)^2} \tag{7}$$

where n is the number of data points (which used for calibration), and k is the number of free parameters used in the modeling process. MSE stands for a mean square error. qi and $\widehat{qi}$, are observed value, and the estimated values and $\overline{\hat{q}}$ and $\overline{q}$ are the estimated mean values and computational model outputs respectively. Typically, the desired model gives higher R^2 or the lowest value of AIC. Also, the parameters of the selected model should not exceed from + 1 or − 1. The autocorrelation statistics and corresponding 95% confidence bands from lag-0 to lag-20 were estimated for the surface water quality time series (for example for TDS is shown in Fig. 2). For the surface water quality, the partial autocorrelation function (PACF) indicated significant correlation up to lag-2 for this time series within the confidence limits, and auto-correlation function (ACF) decreases exponentially.

Results and discussion

Table 2 shows the values of R^2 and AIC performance criteria for the results presented by different stochastic models for prediction of water quality indicators. Quality of the results given by different stochastic models against the observed values is shown in Figs. 3, 4, 5, 6, 7, 8, 9, 10 and 11 for the various hydrogeochemical parameters.

As given in Table 2 and also Fig. 3, models have shown different behaviors in predicting Ca. The values of R^2 are different from 0.611 for the outputs of ARMA (2, 1) and ARMA (2, 2) to 0.516 for the SARIMA models as the

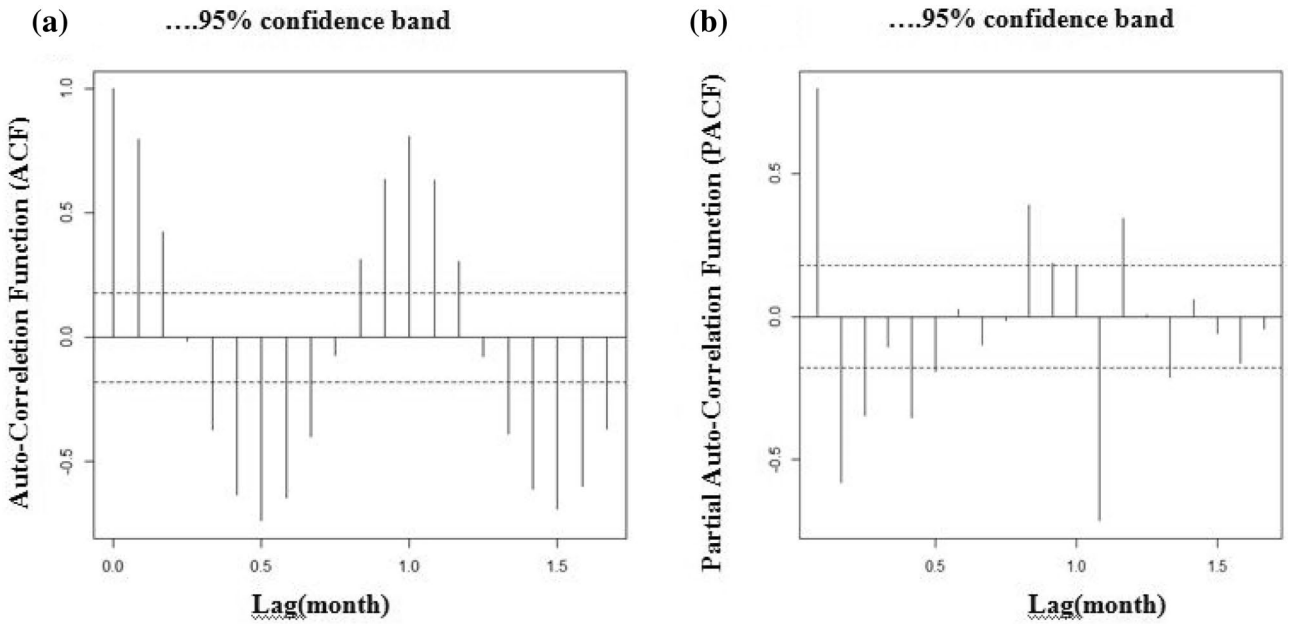

Fig.2 **a** Auto-correlation and **b** partial auto-correlation functions of the monthly data of TDS

Table 2 The values of performance criteria for the results of time-series models for water quality indicator prediction

Model	Ca			Cl			EC			HCO3			Mg		
	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2
AR (1)	ϕ_1 0.2982	273.59	0.606	0.095	221.95	0.498	0.1943	1296.6	0.577	-0.0217	165.65	0.695	0.0691	259.9	0.592
AR (2)	ϕ_1 0.3166	275.26	0.607	0.106	222.15	0.492	0.2104	1297.9	0.575	-0.022	161.63	0.688	0.069	261.9	0.592
	ϕ_2 − 0.0592			− 0.1369			− 0.0849			− 0.2504			0.0007		
MA (1)	θ_1 0.2793	274.08	0.607	0.1258	221.16	0.496	0.2118	1296.23	0.577	− 0.0431	165.61	0.694	0.0688	259.9	0.592
MA (2)	θ_1 0.3106	274.8	0.605	0.063	220.54	0.486	0.2039	1298.15	0.576	− 0.3933	150.44	0.684	0.0686	261.9	0.592
	θ_2 0.1343			− 0.3024			− 0.0425			− 0.6067			0.008		
ARMA (1, 1)	ϕ_1 0.2193	275.44	0.606	− 0.4778	222.41	0.496	− 0.3391	1298.12	0.577	− 0.8953	164.05	0.694	0.0802	261.9	0.592
	θ_1 0.0884			0.6174			0.537			1			− 0.0111		
ARMA (1, 2)	ϕ_1 − 0.019	276.8	0.605	0.5581	208.01	0.493	0.6228	1284.91	0.582	0.3874	146.31	0.689	0.6201	248.78	0.590
	θ_1 0.3285			− 0.6579			− 0.6461			− 0.6422			− 0.7680		
	θ_2 0.1383			− 0.3421			− 0.3536			− 0.3587			− 0.2320		
ARMA (2, 1)	ϕ_1 1.1885	261.56	0.611	0.9117	206.65	0.488	1.0219	1280.91	0.582	0.7339	146.53	0.69	− 0.0131	263.9	0.592
	ϕ_2 − 0.473			− 0.284			− 0.3618			− 0.2203			0.0055		
	θ_1 − 1			− 1			− 1			− 1			0.0822		
ARMA (2, 2)	ϕ_1 1.4546	260.16	0.611	1.6603	205.7	0.515	1.4218	1279.7	0.581	0.4546	148.27	0.689	− 0.0506	260.67	0.600
	ϕ_2 − 0.7079			− 0.7684			− 0.6669			− 0.0503			− 0.8404		
	θ_1 − 1.3444			− 1.9962			− 1.4829			− 0.7068			0.1138		
	θ_2 0.8827			0.9966			0.4829			− 0.2932			1		
ARIMA (1, 1, 2)	ϕ_1 0.3444	277.54	0.606	− 0.4639	225.55	0.495	− 0.11	1289.94	0.577	− 0.8941	167.9	0.694	− 0.8068	263.12	0.592
	d 0.2424			1			1			1			1		
	θ_1 1			− 0.3902			− 0.6730			0.0000			− 0.1045		
	θ_2 − 0.073			− 0.6098			− 0.327			− 1			− 0.8955		
ARIMA (1, 2, 1)	ϕ_1 − 0.3058	305.8	0.604	− 0.3704	265.73	0.49	− 0.3425	1312.38	0.577	− 0.3778	220.5	0.661	− 0.4563	297.49	0.596
	d 2			2			2			2			2		
	θ_1 − 1			− 1			− 1			− 1			− 1		
SARIMA (1, 1, 0) (1, 1, 1) (12)	ϕ_1 − 0.2515	293.53	0.532	− 0.2776	260.05	0.257	− 0.3332	1184.63	0.461	− 0.3307	223.22	0.557	− 0.4453	287.49	0.534
	d 1			1			1			1			1		
	Φ_1 0.1473			− 0.1556			− 0.1418			0.0793			− 0.2605		
	D 1			1			1			1			1		
	Θ_1 − 0.7018			− 0.9999			− 0.9993			− 1			− 1		

Table 2 (continued)

Model	Ca			Cl			EC			HCO3			Mg		
	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2
SARIMA (1, 1, 1) (1, 1, 1) (12)	φ_1 0.345	273.17	0.516	0.0986	231.77	0.288	0.215	1161.34	0.447	− 0.0134	186.53	0.597	0.1269	266.87	0.538
	d 1			1			1			1			1		
	θ_1 − 1			− 1			− 1			− 0.9998			− 1		
	Φ_2 0.1102			− 0.2212			− 0.2098			0.0694			− 0.2068		
	D 1			1			1			1			1		
	Θ_2 − 0.6349			− 0.7831			− 0.7445			− 0.9998			− 0.9998		

Model	Na				pH			SO4			TDS		
	Model coefficient		AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2
AR (1)	φ_1	0.1733	262.41	0.530	− 0.0758	42.84	0.442	0.18.62	345.82	0.737	0.1344	1255.29	0.643
AR (2)	φ_1	0.18.41	264.01	0.526	− 0.0897	42.09	0.446	0.1813	347.76	0.738	0.2116	1257.07	0.643
	φ_2	− 0.0644			− 0.1692			0.0265			− 0.0469		
MA (1)	θ_1	0.1791	262.26	0.528	0.1219	42.53	0.443	0.16.30	346.21	0.737	0.1990	1255.3	0.643
MA (2)	θ_1	0.1826	264.25	0.529	− 0.4015	30.95	0.450	0.1973	346.92	0.738	0.2141	1257.01	0.642
	θ_2	0.0123			− 0.8953			0.1346			0.0793		
ARMA (1, 1)	φ_1	0.0244	264.26	0.528	0.5960	27.51	0.468	0.2328	347.8	0.738	0.1143	1257.21	0.643
	θ_1	0.1561			− 1			− 0.0479			0.0928		
ARMA (1, 2)	φ_1	0.6286	251.11	0.544	0.4514	27.59	0.472	− 0.1829	348.77	0.737	− 0.0688	1258.99	0.642
	θ_1	− 0.6919			− 0.7750			0.3727			0.2798		
	θ_2	− 0.3080			− 0.2250			0.1671			0.0917		
ARMA (2, 1)	φ_1	1.0013	246.52	0.549	0.6832	27.27	0.474	− 0.3159	349.37	0.738	1.0272	1239.55	0.657
	φ_2	− 0.3536			− 0.1548			0.1512			− 0.3661		
	θ_1	− 1			− 1			0.4935			− 1		
ARMA (2, 2)	φ_1	1.4240	243.59	0.547	− 0.2317	22.62	0.469	− 0.3997	347.39	0.711	− 0.2229	1255.22	0.645
	φ_2	− 0.6846			0.4235			− 0.6381			− 0.7508		
	θ_1	− 1.5018			0.0000			0.5512			0.3704		
	θ_2	0.5018			− 1			0.8595			1		
ARIMA (1, 1, 2)	φ_1	0.0627	266.79	0.529	− 0.8279	43.75	0.441	0.2641	349.37	0.738	0.1472	1249.32	0.643
	d	1			1			1			1		
	θ_1	− 0.8715			0.0000			− 1.0670			− 0.9291		
	θ_2	− 0.1285			− 1			0.0670			− 0.0708		

Table 2 (continued)

Model	Na				pH			SO4			TDS		
	Model coefficient		AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2	Model coefficient	AIC	R^2
ARIMA (1, 2, 1)	φ_1	− 0.3678	299.56	0.511	− 0.4501	98.57	0.418	− 0.4161	376.18	0.736	− 0.3208	1270.14	0.644
	d	2			2			2			2		
	θ_1	− 1			− 1			− 1			− 1		
SARIMA (1, 1, 0) (1, 1, 1) (12)	φ_1	− 0.3700	290.33	0.313	− 0.4575	103.46	0.236	− 0.4871	360.06	0.603	− 0.3450	1149.63	0.568
	d	1			1			1			1		
	Φ_1	− 0.1448			0.1875			− 0.4539			− 0.0628		
	D	1			1			1			1		
	Θ_1	− 0.8147			− 0.6025			− 0.5305			− 0.9999		
SARIMA (1, 1, 1) (1, 1, 1) (12)	φ_1	0.1953	267.69	0.343	− 0.0094	71.09	0.161	0.1586	339.54	0.622	0.2215	1127.56	0.516
	d	1			1			1			1		
	θ_1	− 1			− 1			− 1			-1		
	Φ_2	− 0.1630			− 0.0087			− 0.1959			− 0.1462		
	D	1			1			1			1		
	Θ_2	− 0.7848			− 0.4711			− 0.5544			− 0.7257		

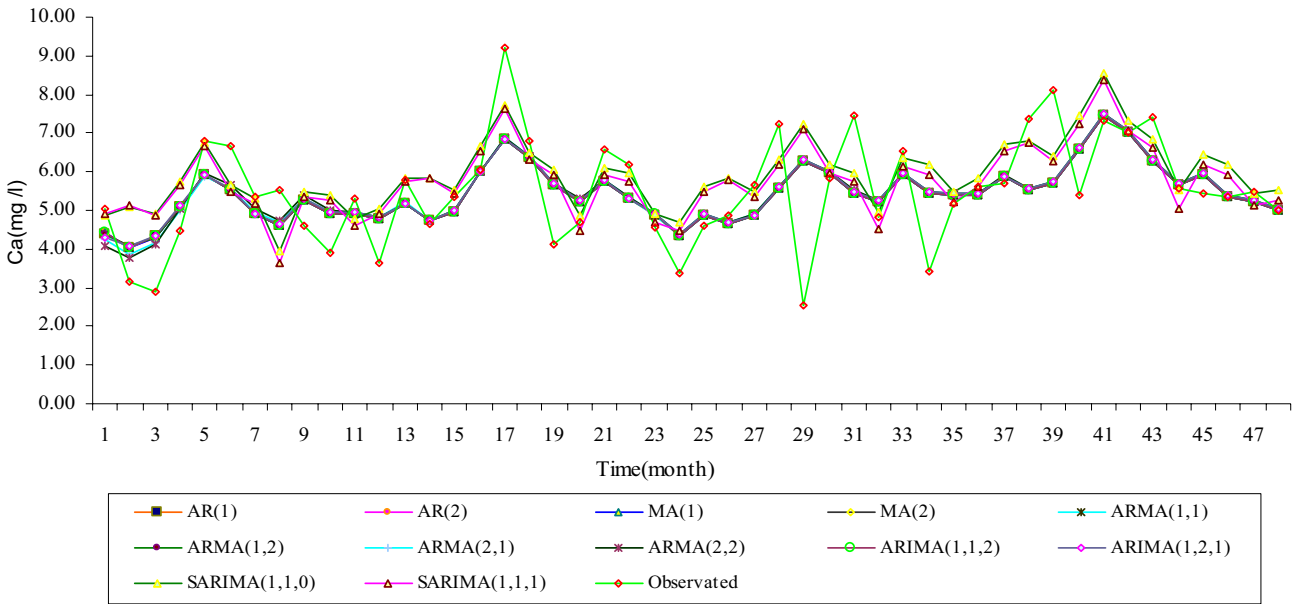

Fig. 3 Models prediction versus observed Ca data

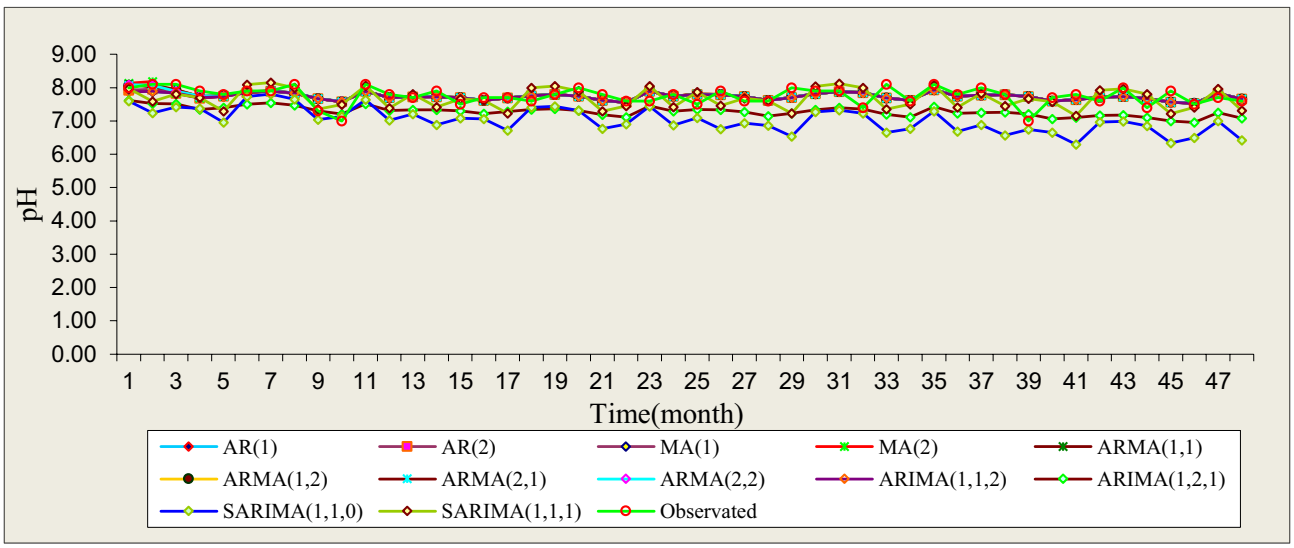

Fig. 4 Models prediction versus observed pH data

maximum and minimum values, respectively. The values of AIC also vary from 260.16 to 305.8 for the results produced by ARMA (2, 2) and ARIMA (1, 2, 1) models, respectively. Moreover, based on Fig. 3, in the prediction of Ca, results of different models are almost near to each other, but not close enough to the observed values, especially in the months when the maximum and minimum values (extreme values) occur. To sum up, the best model for simulation and prediction of Ca is AR (1) with R^2 equal to 0.60.

In the prediction of pH, the results presented by most of the used models are relatively good and almost close to the observed values and also close to each other. However, the results produced by SARIMA models are different from those presented by other models and even weaker in accuracy compared to the related observations. Figure 4 clearly shows this condition. Pursuant to Table 2, the optimized model for simulation and prediction of pH variations is ARMA (2, 2) with R^2 equal to 0.46. For example, the stochastic models could not do good simulation for pH variation.

For prediction of TDS, the outputs of all models are quite close to each other and following almost the average

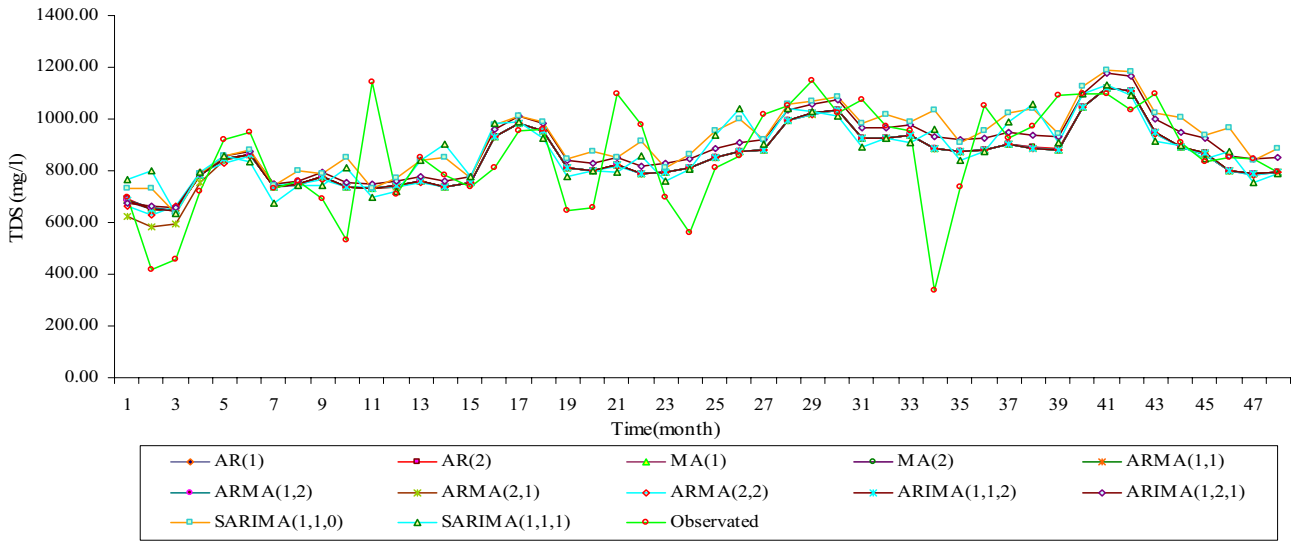

Fig. 5 Models prediction versus observed TDS data

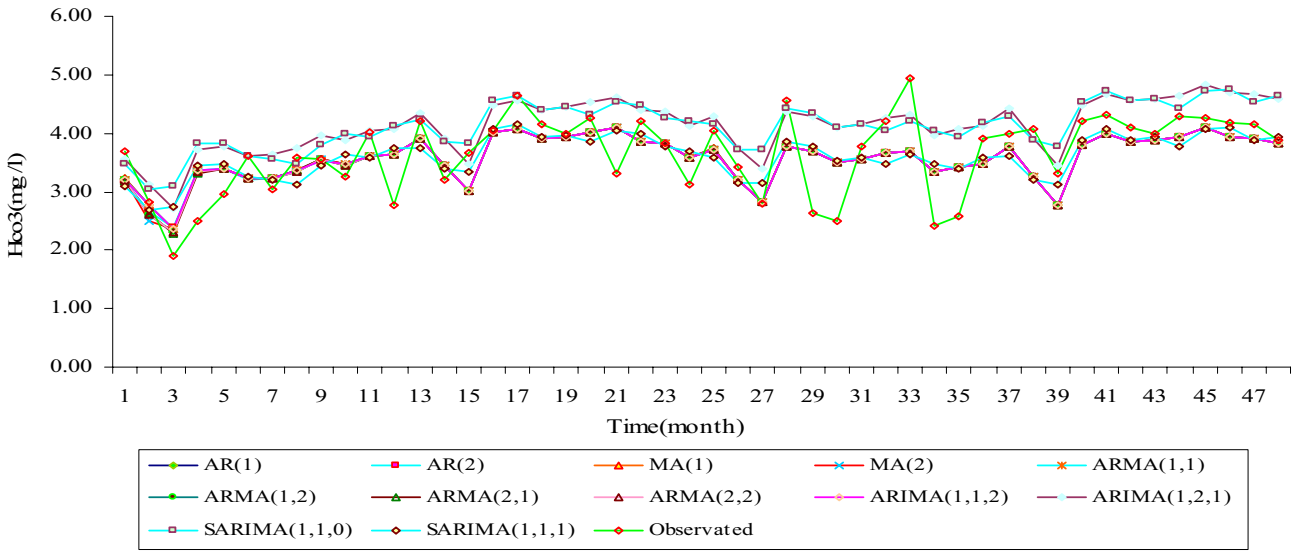

Fig. 6 Models prediction versus observed HCO_3 data

line of TDS values all over the modeled period, and not able to predict fluctuations with acceptable accuracy (Fig. 5). Based on the results (Table 2), the best model for simulation and prediction of TDS variations is ARIMA (1, 1, 2) with R^2 equal to 0.64. As shown in Fig. 5, the observed data showed sum jumps in which none of the models could not simulate this term.

As Fig. 6 indicates, in the prediction of HCO_3, the behaviors of all models are almost similar. Where the monthly data of HCO_3 are almost flat with small or no fluctuations, the predictions of all models are reliable. However, when it shows more fluctuations, the model's prediction accuracy decreases.

In the prediction of SO_4, as Fig. 7 shows the model's performance is almost similar to the HCO_3 prediction condition, which means that models can present good predictions for the data near the average values, but are not able to make acceptable predictions of the maximum and minimum points. However, the best model for simulation of SO_4 is AR (1) with R^2 equal to 0.73.

A similar statement can be made about the prediction of Mg, Na, Cl, and EC (Figs. 8, 9, 10, 11). Although the ability of different models in prediction of these parameters is more or less different in all of these, the performance of the models decreases in the prediction of maximum and minimum values.

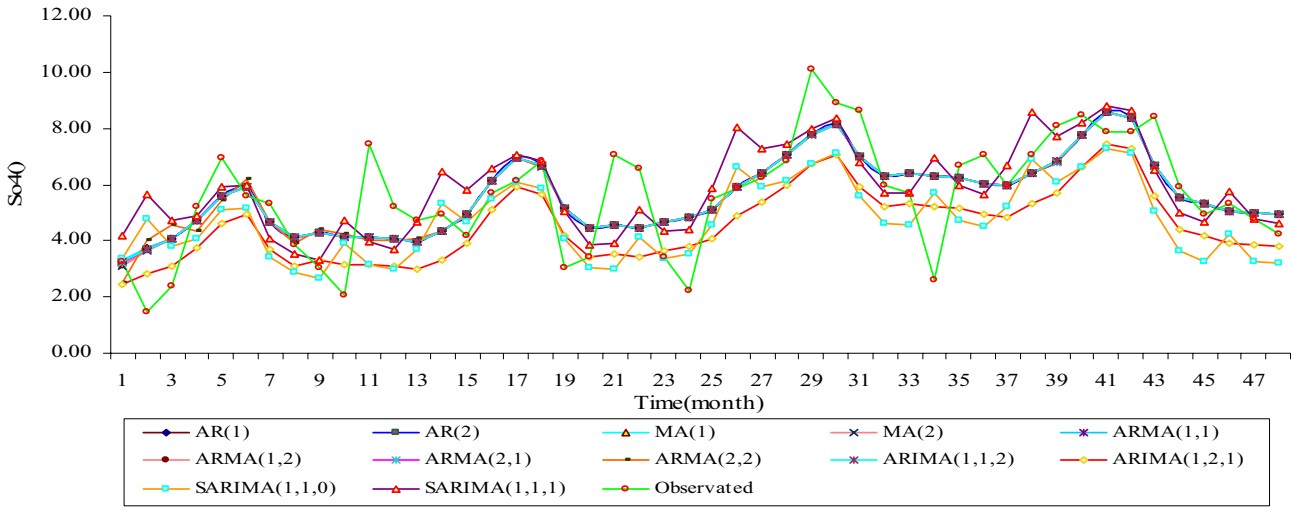

Fig. 7 Models prediction versus observed SO_4 data

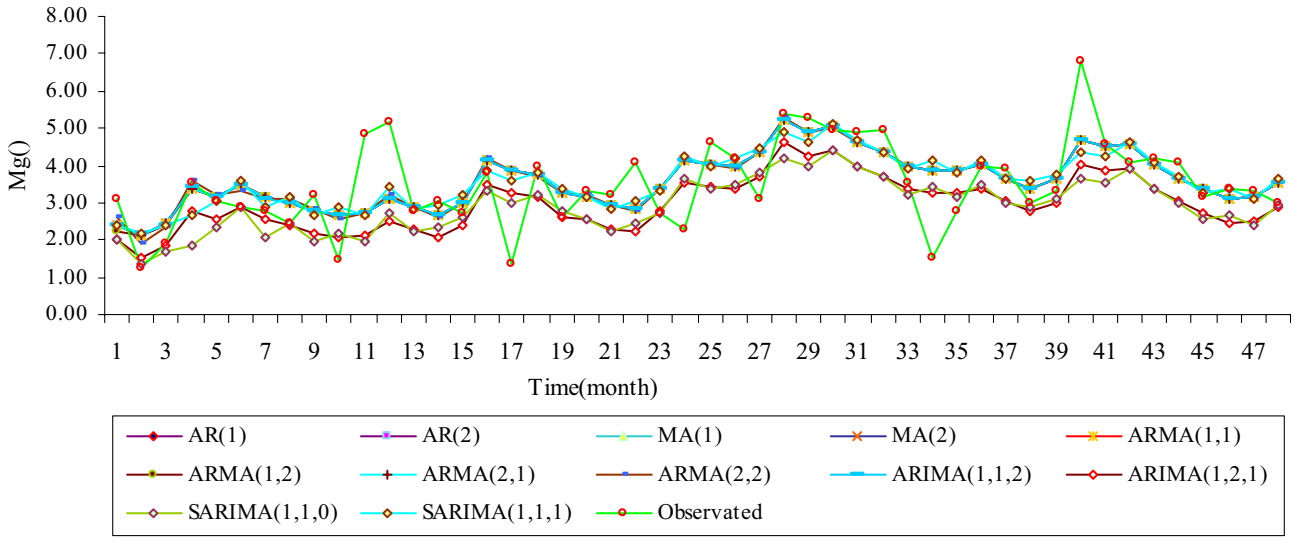

Fig. 8 Models prediction versus observed Mg data

According to the results, the water quality data had an ascending trend before removing deterministic components of the data (e.g., for TDS parameter shown in Fig. 7). But these parameters have seasonal and random ingredients. So, at first, the seasonal and trend (determine components) were omitted, and the random component (stochastic component) was simulated and predicted for each parameter. Overall, ARMA was the best model for six parameters out of nine based on ACF and PACF. The Akaike criterion and the correlation coefficient were used to select more accurate simulation models. The obtained results from this research about the applicability of time-series models in prediction of water quality parameters are in line with studies carried out by Singh Parmar and Bhardwaj (2014) and TaheriTizro et al (2014). But in the type of the best selected models, the findings of the present study differ from those of Singh Parmar and Bhardwaj (2014) and TaheriTizro et al. (2014). The main reason for this incompatibility can be limited types of models used in those studies in comparison with this one and the differences between geological, anthropogenic activities, and climate conditions of the mentioned two studies and the present one which affect the statistical nature of data.

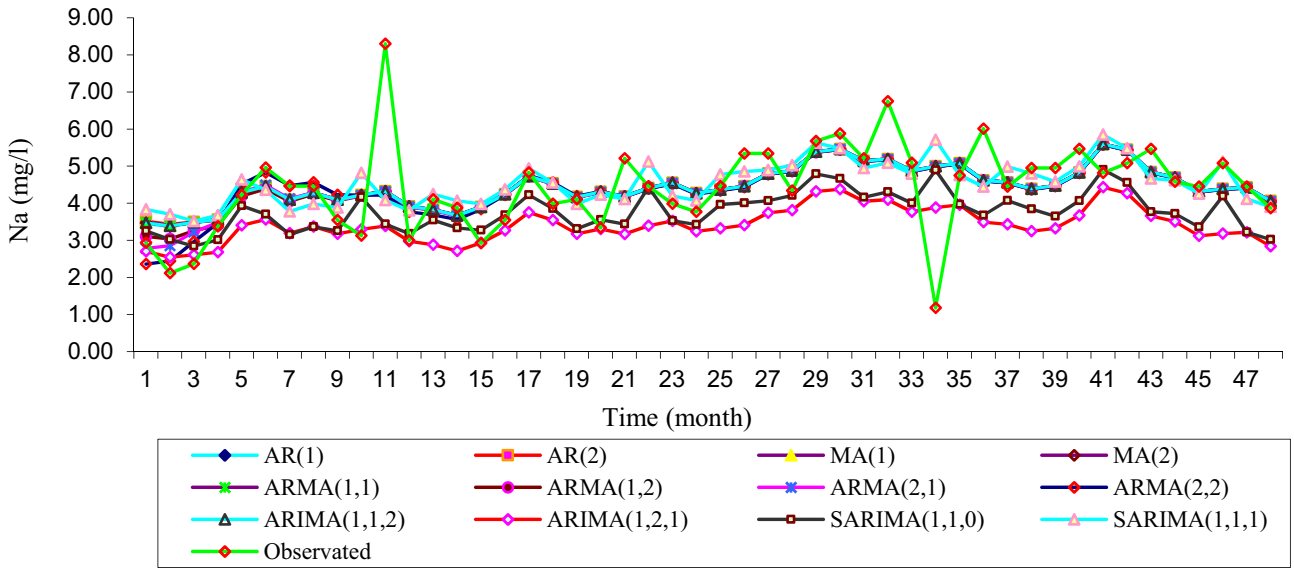

Fig. 9 Models prediction versus observed Na data

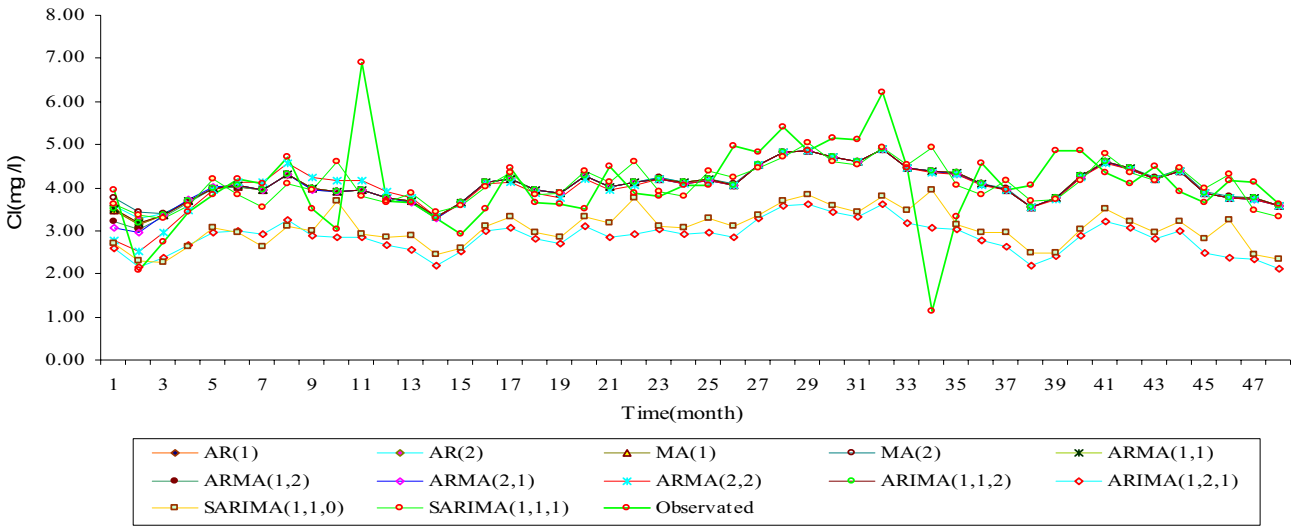

Fig. 10 Models prediction versus observed Cl data

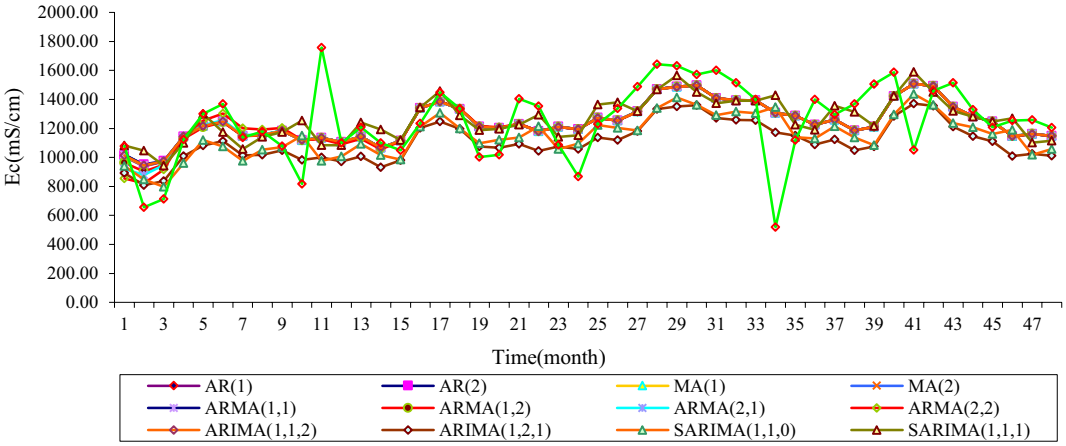

Fig. 11 Models prediction versus observed Ec data

Regarding the results produced by different models for different water quality parameters in this research, the best model for each parameter was selected based on Akaike information criteria (AIC) and coefficient of determination (R^2), and then, the water quality parameters were predicted for the next period of 2015–2018 based on the best model for each parameter. Table 3 shows these monthly prediction values.

Conclusions

In this research, it has been tried to simulate nine water quality parameters including Ca, HCO_3, SO_4, EC, pH, Mg, Cl, Na, and TDS using five time-series models (AR, MA, ARMA, ARIMA, and SARIMA) with 12 different structures for the time period of 2001–2014. The purpose is to evaluate and select the best model for prediction of the mentioned parameters in the future time periods. Time-series R software was used to simulate the water quality parameters using monthly data (2001–2014). Time-series data have four components (trend component, seasonal ingredient, jump, and random component). In this study, the time series was decomposed, and the random element for each parameter was modeled. Five models with 12 different structures were examined. Overall, ARMA was the best model for six parameters out of nine based on ACF and PACF. To sum up, it can be concluded that the simplicity of times-series models is an advantage to use these models for prediction purposes. However, this research showed that their results are generally satisfactory when the data range of variations is not too wide. In other word, the results of these models are acceptable when the data fluctuations are limited, as the predictions made by these models for maximum and minimum values in this research are relatively weak.

Table 3 Predicted monthly values of water quality indices using the optimized models for the period of 2015–2018

Time year	Ca	Cl	Ec	HCO_3	Mg	Na	pH	SO_4	TDS
	AR(1)	ARMA (2,1)	ARMA (1,2)	ARMA (2,1)	ARMA (2,2)	ARMA (1,2)	ARMA (2,2)	AR(1)	ARIMA (1,1,2)
2015									
1	4.36	3.15	936.21	2.92	2.25	3.15	7.82	3.3	677.65
2	4.06	2.71	840.44	2.53	2.04	3.05	8	3.73	652.56
3	4.34	3.07	915.61	2.27	2.32	3.29	7.79	4.08	647.23
4	5.12	3.54	1111.15	3.33	3.34	3.44	7.76	4.72	781.13
5	5.94	3.93	1198.8	3.4	3.12	4.22	7.7	5.59	847.97
6	5.56	4.03	1238.05	3.22	3.46	4.41	7.89	5.9	864.34
7	4.93	3.96	1143.67	3.24	3.11	4.07	7.87	4.68	736.14
8	4.63	4.33	1151.65	3.35	2.98	4.26	7.86	4.1	746.62
9	5.29	3.97	1183.04	3.55	2.78	4.08	7.65	4.3	776.59
10	4.94	3.91	1118.22	3.46	2.65	4.2	7.6	4.15	736.36
11	4.94	3.94	1135.52	3.61	2.73	4.31	7.88	4.13	731.87
12	4.81	3.76	1107.03	3.63	3.11	3.9	7.71	4.07	740.52
2016									
1	5.18	3.66	1143.36	3.9	2.87	3.81	7.72	3.99	758.59
2	4.74	3.3	1067.62	3.44	2.68	3.66	7.74	4.34	738.53
3	4.99	3.63	1116.34	3.01	3	3.88	7.7	4.93	752.86
4	6.01	4.11	1340.74	4.01	4.11	4.23	7.63	6.13	933.84
5	6.85	4.19	1385.04	4.08	3.87	4.72	7.69	6.93	985.56
6	6.38	3.93	1334.31	3.91	3.74	4.52	7.77	6.65	954.26
7	5.69	3.85	1212.11	3.94	3.25	4.16	7.79	5.18	811.72
8	5.25	4.27	1204.33	4.02	3.16	4.3	7.75	4.47	802.53
9	5.77	4.01	1229.98	4.1	2.92	4.17	7.62	4.57	821.04
10	5.31	4.1	1181.09	3.86	2.82	4.4	7.56	4.46	787.11
11	4.91	4.2	1211.82	3.82	3.38	4.54	7.89	4.66	795.2
12	4.36	4.1	1195.69	3.59	4.16	4.26	7.75	4.84	812.48
2017									
1	4.88	4.17	1273.65	3.73	4.02	4.35	7.8	5.12	853.92
2	4.67	4.05	1255.74	3.2	3.97	4.45	7.79	5.92	872.44
3	4.87	4.51	1317.9	2.81	4.34	4.79	7.74	6.42	881.11
4	5.59	4.8	1471.41	3.78	5.25	4.86	7.62	7.04	993.92
5	6.28	4.84	1487.97	3.69	4.88	5.38	7.71	7.79	1020.56
6	5.98	4.7	1498.28	3.5	5.03	5.45	7.82	8.15	1036.09
7	5.48	4.6	1409.58	3.55	4.62	5.13	7.88	6.99	926.6
8	5.23	4.89	1395	3.65	4.33	5.18	7.86	6.28	925.2
9	5.93	4.44	1394.52	3.68	3.99	4.86	7.7	6.39	936.13
10	5.44	4.36	1308.64	3.34	3.88	4.98	7.62	6.28	885.07
11	5.4	4.32	1288.12	3.42	3.87	5.07	7.94	6.25	874.66
12	5.41	4.07	1225.64	3.48	4.01	4.6	7.74	6.04	879.61
2018									
1	5.88	3.95	1260.61	3.77	3.67	4.55	7.78	5.95	902.46
2	5.55	3.53	1185.82	3.25	3.39	4.38	7.79	6.41	888.86
3	5.74	3.74	1211.78	2.78	3.62	4.46	7.74	6.81	883.11
4	6.61	4.24	1422.82	3.8	4.68	4.82	7.6	7.75	1045.13
5	7.48	4.58	1509.1	3.98	4.52	5.58	7.65	8.56	1122.72
6	7.05	4.43	1493.53	3.86	4.57	5.43	7.71	8.38	1110.09
7	6.3	4.2	1349.29	3.89	4	4.83	7.73	6.69	946.56
8	5.68	4.39	1283.46	3.93	3.67	4.68	7.67	5.54	894.41
9	5.93	3.87	1247.64	4.11	3.36	4.3	7.57	5.33	866.13

Table 3 (continued)

Time year	Ca	Cl	Ec	HCO_3	Mg	Na	pH	SO_4	TDS
	AR(1)	ARMA (2,1)	ARMA (1,2)	ARMA (2,1)	ARMA (2,2)	ARMA (1,2)	ARMA (2,2)	AR(1)	ARIMA (1,1,2)
10	5.37	3.77	1147.39	3.95	3.11	4.37	7.53	5.05	800.24
11	5.24	3.74	1163.03	3.92	3.14	4.41	7.83	5.01	786.37
12	5.01	3.56	1148.73	3.82	3.52	4.05	7.67	4.96	793.23

References

Adhikary SK, RahmanGupta MDA (2012) A stochastic modeling technique for predicting groundwater table fluctuations with time series analysis. Int J Appl Sci Eng Res 1(2):238–249. https://doi.org/10.6088/ijaser.0020101024

Antonopoulos VZ, Papamichail DM, Mitsiou KA (2001) Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrol Earth Syst Sci 5(4):679–691

Behnia N, Rezaeian F (2015) Coupling wavelet transform with time series models to estimate groundwater level. Arab J Geosci 8(10):8441–8447

Boskidis I, Gikas GD, Sylaios GK, Tsihrintzis VA (2012) Hydrologic and water quality modeling of Lower Nestos river basin. Water Resour Manag 26(10):3023–3051. https://doi.org/10.1007/s11269-012-0064-7

Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis: forecasting and control. Prentice Hall, Upper Saddle River

Bras RL, Rodriguez-Iturbe I (1985) Random functions and hydrology. Addison-Wesley, Reading

Brockwell PJ, Davis RA (2010) Introduction to time series and forecasting. Springer, New York

Carr GM, Neary PJ (2006) Water quality for ecosystem and human health. United Nations Environment Programme Global Environment Monitoring System (GEMS)/Water Programme, pp 8–32

Clark ID (2015) Groundwater geochemistry and isotopes. Taylor & Francis Group, Baco Raton

Dabrowski JM, Murray K, Ashton PJ, Leaner JJ (2009) Agricultural impacts on water quality and implications for virtual water trading decisions. Ecol Econ 68(4):1074–1082

Dasturani M, Mirzavand M, Dasturani MT (2016) Comparative study among different time series models applied to monthly rainfall forecasting. Nat Hazards. https://doi.org/10.1007/s11069-016-2163-x

Debele B, Srinivasan R, Parlange JY (2009) Hourly analyses of hydrological and water quality simulations using the ESWAT model. Water Resour Manag 23(2):303–324

Delpla I, Jung AV, Baures E, Clement M, Thomas O (2009) Impacts of climate change on surface water quality in relation to drinking water production. Environ Int 35(8):1225–1233

Dökmen F, Aslan Z (2013) Evaluation of the parameters of water quality with wavelet techniques. Water Resour Manag 27(14):4977–4988. https://doi.org/10.1007/s11269-013-0454-5

Erdem E, Shi J (2011) ARMA based approaches for forecasting the tuple of wind speed and direction. Appl Energy 84(2):1405–1414. https://doi.org/10.1016/j.apenergy.2010.10.031

Faruk D (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng Appl Artif Intell 23(4):586–594. https://doi.org/10.1016/j.engappai.2009.09.015

Hannan EJ (1971) Multiple time series. Wiley, New York

He T, Lu Y, Cui Y, Luo Y, Wang M, Meng W, Zhang K, Zhao F (2015) Detecting gradual and abrupt changes in water quality time series in response to regional payment programs for watershed services in an agricultural area. J Hydrol 525(2015):457–471

Hipel KW, Mcleod AI (1994) Time series modeling of water resources and environmental systems. Elsevier Science, Amsterdam

Hu Sh (2007) Akaike information criterion, enter for research in scientific computation. North Carolina State University, Raleigh, p 19

Kim SJ, Hyun Y, Lee KK (2005) Time series modeling for evaluation of groundwater discharge rates into an urban subway system. Geosci J 9(1):15–22. https://doi.org/10.1007/BF02910550

Klose C, Pircher M, Sharma S (2004) Univariate time-series forecasting. 406347/UK "ÖkonometrischePrognose" in SS04

Lin GF, Lee FC (1992) An aggregation–disaggregation approach for hydrologic time series modelling. J Hydrol 138(3–4):543–557. https://doi.org/10.1016/0022-1694(92)90136-J

Luo PP, He B, Takara K, Razafindrabe BHN, Nover D, Yamashiki Y (2011) Spatiotemporal trend analysis of recent river water quality conditions in Japan. J Environ Monit 13(10):2819–2829

Mirzavand M, Ghazavi R (2015) A stochastic modeling technique for groundwater level forecasting in an arid environment using time series methods. Water Resour Manag 29:1315–1328

Mirzavand M, Ghasemieh H, Vali AA (2012) Analysis of hydromorphometric features by use the geographic information system and statistical analysis (case study: Maragh basin). J Adv Appl Sci Res 4(4):22–29

Mirzavand M, Ghasemieh H, Sadatinejad SJ, Bagheri R (2020a) Delineating the source and mechanism of groundwater salinization in crucial declining aquifer using multi-chemo-isotopes approaches. J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.124877

Mirzavand M, HodaGhasemieh SeyyedJavadSadatinejad, Bagheri R (2020b) An overview on source, mechanism and investigation approaches in groundwater salinization studies. Int J Environ Sci Technol. https://doi.org/10.1007/s13762-020-02647-7

Mondal MS, Wasimi SA (2007) Choice of model type in stochastic river hydrology. In: Proceedings of the 1st international conference on water & flood management (ICWFM-2007), Dhaka, 12–14 March 2007, pp 633–640

Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS (2004) A neuro-fuzzy computing technique for modeling hydrological time series. J Hydrol 291(1–2):52–66

Partal T, Kisi O (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. J Hydrol 342(1–2):199–212

Pejman AH, Bidhendi GN, Karbassi AR, Mehrdadi N, Bidhendi ME (2009) Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques. Int J Environ Sci Technol 6(3):467–476

Perrin JL, Raïs N, Chahinian N, Moulin P, Ijjaali M (2014) Water quality assessment of highly polluted rivers in a semi-arid Mediterranean zoneOuedFez and Sebou River (Morocco). J Hydrol 510(2014):26–34

Piao SL, Ciais P, Huang Y, Shen ZB, Peng SS, Li JS, Zhou LP, Liu HY, Ma YC, Ding YH, Friedlingstein P, Liu CZ, Tan K, Yu YQ, Zhang TY, Fang JY (2010) The impacts of climate change on water resources and agriculture in China. Nature 467(7311):43–51

Shirmohammadi B, Vafakhah M, Moosavi V, Moghaddamnia A (2013) Application of several data-driven techniques for predicting groundwater level. Water Resour Manag 27:419–432. https://doi.org/10.1007/s11269-012-0194-y

Singh KP, Malik A, Mohan D, Sinha S (2004) Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India): a case study. Water Res 38(18):3980–3992

Singh Parmar K, Bhardwaj R (2014) Water quality management using statistical analysis and time-series prediction model. Appl Water Sci 4(4):425–434. https://doi.org/10.1007/s13201-014-0159-9

Singh Patel Sh, Ramachandran P (2015) A comparison of machine learning techniques for modeling river flow time series: the case of upper Cauvery river basin. Water Resour Manag 29(2):589–602. https://doi.org/10.1007/s11269-014-0705-0

Sokolnikov AN (2013) THZ identification for defense and security purposes identifying materials, substances and items. World Scientific Publishing CO.Pte. Ltd., Singapore

Solheim AL, Austnes K, Eriksen TE, Seifert I, Holen S (2010) Climate change impacts on water quality and biodiversity, Background Report for EEA European Environment State and Outlook Report. ETC Water Tech Rep 1/2010 1–68

TaheriTizro A, Ghashghaie M, Georgiou P, Voudouris K (2014) Time series analysis of water quality parameters. J Appl Res Water Wastewater 1(1):43–52

Tong STY, Chen WL (2002) Modeling the relationship between land use and surface water quality. J Environ Manag 66(4):377–393

Wan RR, Cai SS, Li HP, Yang GS, Li ZF (2014) Inferring land use and land cover impact on stream water quality using a Bayesian hierarchical modeling approach in the Xitiao River Watershed, China. J Environ Manag 133(2014):1–11

Whitehead PG, Wilby RL, Battarbee RW, Kernan M, Wade AJ (2009) A review of the potential impacts of climate change on surface water quality. Hydrol Sci J 54(1):101–123

Yeniguna K, Ecer R (2012) Overlay mapping trend analysis technique and its application in Euphrates Basin. Meteorol Appl 20(4):427–438. https://doi.org/10.1002/met.1304