Environ Sci Pollut Res (2015) 22:397–414
DOI 10.1007/s11356-014-3346-1

RESEARCH ARTICLE

Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management

Kulwinder Singh Parmar · Rashmi Bhardwaj

Received: 5 April 2014 / Accepted: 14 July 2014 / Published online: 1 August 2014

Abstract River water is a major resource of drinking water on earth. Management of river water is highly needed for surviving. Yamuna is the main river of India, and monthly variation of water quality of river Yamuna, using statistical methods have been compared at different sites for each water parameters. Regression, correlation coefficient, autoregressive integrated moving average (ARIMA), box-Jenkins, residual autocorrelation function (ACF), residual partial autocorrelation function (PACF), lag, fractal, Hurst exponent, and predictability index have been estimated to analyze trend and prediction of water quality. Predictive model is useful at 95 % confidence limits and all water parameters reveal platykurtic curve. Brownian motion (true random walk) behavior exists at different sites for BOD, AMM, and total Kjeldahl nitrogen (TKN). Quality of Yamuna River water at Hathnikund is good, declines at Nizamuddin, Mazawali, Agra D/S, and regains good quality again at Juhikha. For all sites, almost all parameters except potential of hydrogen (pH), water temperature (WT) crosses the prescribed limits of World Health Organization (WHO)/United States Environmental Protection Agency (EPA).

Keywords Statistical analysis · Fractal dimension · Time series analysis · Water resource management

Responsible editor: Michael Matthies

K. S. Parmar · R. Bhardwaj (✉)
Non-Linear Dynamics Research Laboratory, Department of Mathematics, University School of Basic and Applied Sciences, Guru Gobind Singh Indraprastha University, Dwarka, Delhi 110078, India
e-mail: rashmib22@gmail.com

K. S. Parmar
e-mail: kulmaths@gmail.com

Introduction

River water plays an important role in the supply of drinking water, and the quality of drinking water depends upon the quality of nearer river water. Yamuna is the largest tributary river of the Ganga in India. It originates from Yamunotri glacier at a height of 6,387 m on the southwestern slopes of Banderpooch peaks (38° 59′ N, 78° 27′ E) in the lower Himalayas in Uttarakhand. It travels a total length of 1,376 km by crossing several states, Uttarakhand, Haryana, Himachal Pradesh, Delhi, and Uttar Pradesh and has a mixing of drainage system of 366,233 km^2 before merging with the Ganga at Allahabad, i.e., a total of 40.2 % of the entire Ganga basin. The river accounts for more than 70 % of Delhi's water supplies and about 57 million people depend on river water for their daily usage (CPCB 2006).

Central Pollution Control Board (CPCB) monitors the water quality parameters at different sites of Yamuna River. Five sample sites are chosen according to utilization of river water, namely Hathnikund, Nizamuddin, Mazawali, Agra, and Juhikha. Hathnikund is approximately 157 km downstream from Yamunotri and 2 km upstream from Tajewala barrage. Nizamuddin is approximately 14 km downstream from Wazirabad barrage at Delhi. The distance from Hathnikund to Wazirabad is 224 km. The water quality at Hathnikund has the impact of industrial and sewerage discharge from Haryana and Delhi. Mazawali is about 84 km downstream from Wazirabad barrage and have the impact of wastewater discharge from Shahdara drain and Hindon river. Agra D/S at west Burzi of Taj Mahal monument is about 310 km downstream from Wazirabad barrage and depicts the impact of sewerage water discharge from Agra city and industrial waste from Mathura refinery. Juhikha is about 613 km downstream from Wazirabad barrage and assesses the impact of river Chambal confluence on river Yamuna (CPCB 2006).

Pollution in river water is continuously increasing due to urbanization, industrialization, etc., and most of the rivers are at dying position, which is an alarming signal (Parmar et al. 2009). Industrial wastes, municipal sewage, and agricultural runoff effect physicochemical parameters of river water (Akoto and Adiyiah 2007; Alam et al. 2007; Hermans et al. 2007; Shukla et al. 2008). Trihalomethane compounds were determined in the drinking water samples at consumption sites and treatment plants of Okinawa and Samoa Islands and observed that the chloroform, bromodichloromethane compound exceed the level of Japan water quality and WHO (World Health Organization) standards (APHA 1995; WHO 1971). Water quality modeling, using multiple linear regression, structural equation, trend and time series analysis are major tools for application in water quality management (Chenini and Khemiri 2009; Fang et al. 2010; Singh et al. 2004; Su et al. 2011; Vassilis et al. 2001; Bhardwaj and Parmar 2014; Panepinto and Genon 2010; Amiri and Nakane 2009; Boskidis et al. 2012).

Climatic dynamic plays an important role in determining the water quality. Using fractal dimensional analysis, trend and time series data of three major dynamic components temperature, pressure and precipitation of the climate analyzed (Dutta et al. 2013; Bhardwaj and Parmar (2013a); Rangarajan 1997). Regional climatic models would not be able to predict local climate as it deals, with averaged quantities and that precipitation during the southwest monsoon is affected by temperature and pressure variability during the preceding winter (Kahya and Kalayci 2004; McCleary and Hay 1980; Mousavi et al. 2008; Movahed and Hermanisc 2008; Park and Park 2009; Rangarajan and Ding 2000; Rangarajan and Sant 2004; Toprak 2009; Toprak et al. 2009; Calvo et al. 2012; Yarar 2014).

The quality of Yamuna River water depends upon quality of water parameters, potential of hydrogen (pH), chemical oxygen demand (COD), biochemical oxygen demand (BOD), dissolved oxygen (DO), water temperature (WT), free ammonia (AMM), and total Kjeldahl nitrogen (TKN). In this paper, statistical analysis, regression analysis, trend, time-series, autoregressive integrated moving average (ARIMA), residual autocorrelation function (ACF), residual partial autocorrelation function (PACF), lag, Hurst exponent, fractal dimension, and predictability index of these water parameters have been estimated at the five sample sites, Hathnikund (S_1), Nizamuddin (S_2), Mazawali (S_3), Agra D/S (S_4), and Juhikha (S_5) of Yamuna River which crosses different states of India as shown in Fig. 1. Monthly average values of last 10 years of water quality parameters at these sites have been considered for study.

Mathematical modeling

Statistical analysis

Measure of central tendency and dispersion are used to calculate mean, median, mode, standard deviation, kurtosis, skewness, and coefficient of variation. Mean explains average value. Median gives the middle values of an ordered sequence or positional average. Mode is defined as the value which occurs the maximum number of time that is having the maximum frequency. Standard deviation gives measure of spread of the sample. Kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. Skewness describes the symmetry of data. Coefficient of variation gives the relative measure of sample (Bhardwaj and Parmar (2013b) Box et al. 2008; Rangarajan and Ding 2000; Diodato et al. 2014).

R squared is an estimate of the proportion of the total variation in the series which is explained by the model and measure is useful when the series is stationary. Stationary *R* squared is a measure that compares the stationary part of the model to a simple mean model and is preferable to ordinary *R* squared when there is a trend or seasonal pattern. Stationary *R* squared can be negative with a range of negative infinity to 1. Negative values mean that the model under consideration is worse than the baseline model. Positive values mean that the model under consideration is better than the baseline model (Box et al. 2008; DeLurgio 1998; McCleary and Hay 1980).

In each of the forthcoming definitions, y_t is the actual value, f_t is the forecasted value, $e_t = y_t - f_t$ is the forecast error, and n is the size of the test set. Also, $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ is the test mean and $\sigma^2 = \frac{1}{n-1}\sum_{t=1}^{n} (y_t - \bar{y})^2$ is the test variance.

The mean absolute error is defined as

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{1}$$

Mean absolute percentage error, measure is given by

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \times 100 \tag{2}$$

Mean percentage error, is defined as

$$\mathrm{MPE} = \frac{1}{n}\sum_{t=1}^{n} \left(\frac{e_t}{y_t}\right) \times 100 \tag{3}$$

Mathematically, root mean square error is

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{4}$$

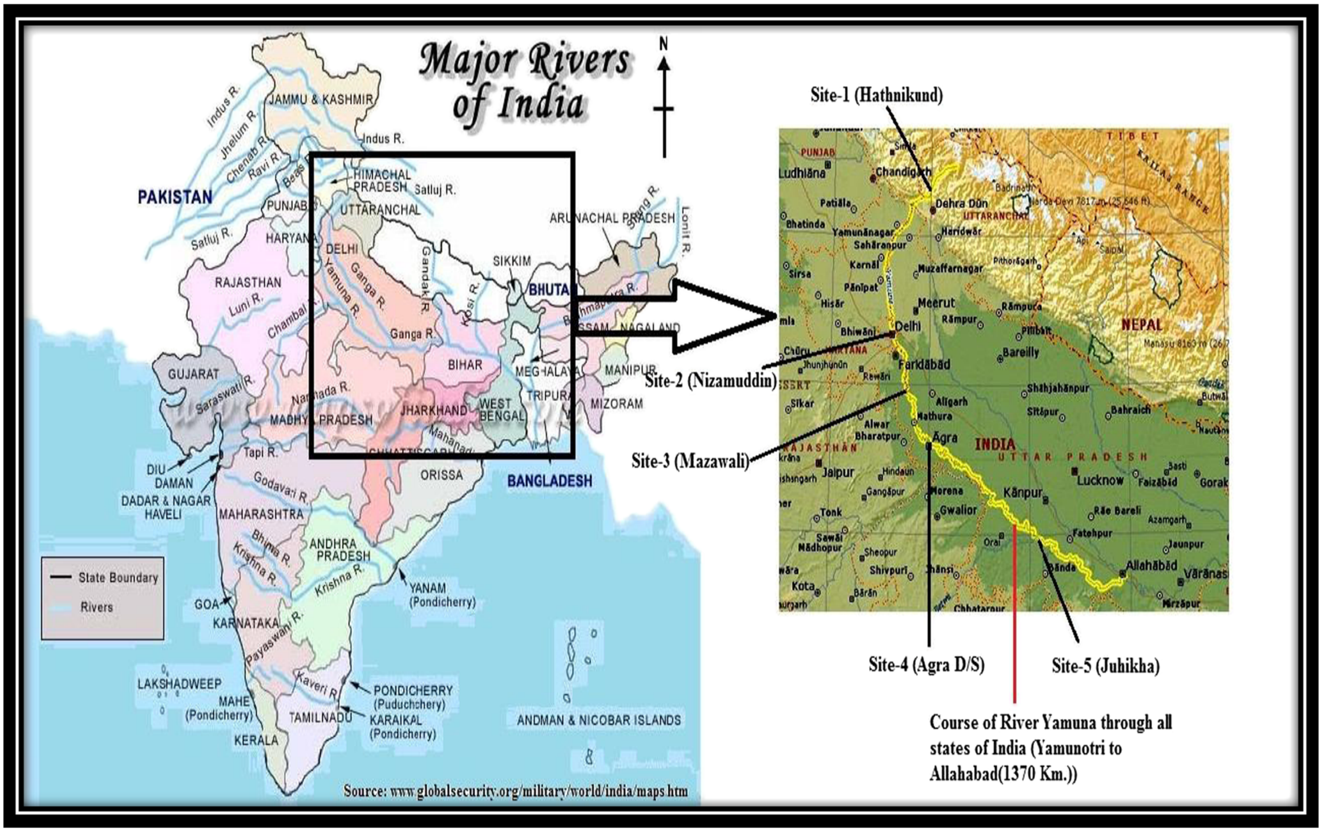

Fig. 1 Flow of Yamuna River in India with five main stations, site1—Hathnikund (S_1), site2—Nizamuddin (S_2), site3—Mazawali (S_3), site4—Agra D/S (S_4), and site5—Juhikha (S_5)

Regression analysis

It is a technique used for modeling and analyzing the variables present in a sample. Regression analysis helps in understanding the variation in value of the dependent variable as independent variables is varied, while the other independent variables are held fixed. Regression line of Y (dependent variable) on X (independent variable) defined as (Chenini and Khemiri 2009)

$$Y = b_{yx}X + C \tag{5}$$

where C is the intercept,

$$\begin{aligned} b_{yx} &= \text{regression coefficient} = r \times \frac{\sigma_y}{\sigma_x} \\ r &= \text{Correlation coefficient} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\left(E(X^2) - E(X)^2\right)\left(E(Y^2) - E(Y)^2\right)}} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \end{aligned} \tag{6}$$

σ_Y, σ_X are standard deviation of variables Y and X, respectively, and $E(X)$, $E(Y)$, $E(XY)$ are expected value of variables X,Y and XY, respectively.

Time series analysis

Time series is a sequence of data points, measured at successive times spaced at uniform time intervals. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data and to forecast future events based on known past events to predict data points before these are measured. Time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart (Weng, et al. 2008).

Autoregressive integrated moving average

Auto regressive integrated moving average (ARIMA) model of a time series is defined by three terms p, d, q. Identification of a time series is the process of finding integer values of p, d, and q. When the value is 0, the element is not needed in the model. The middle element, d, is investigated before p and q. The goal is to determine if the process is stationary and, if not, to make it stationary before determining the values of p and q. A stationary process has a constant mean and variance over the time period of the study. The representation of an autoregressive model in time series (Box et al. 2008; DeLurgio 1998; McCleary and Hay 1980), well-known as AR(p), is

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_1 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t \tag{7}$$

where the term ε_t is the source of randomness and is called white noise α_i are constants.

In ARIMA models, a non-stationary time series is made stationary by applying finite differencing of the data points. The mathematical formulation of the ARIMA (p,d,q) model using lag polynomials is given below

$$\phi(L)(1-L)^d y_t = \theta(L)\varepsilon_t, \text{i.e.}$$
$$\left(1-\sum_{i=1}^{p} \phi_i L^i\right)(1-L)^d y_t = \left(1+\sum_{j=1}^{q} \theta_j L^j\right)\varepsilon_t \tag{8}$$

Here, p, d, and q are integers greater than or equal to zero and refer to the order of the autoregressive, integrated, and moving average parts of the model, respectively. The integer d controls the level of differencing. Generally, d=1 is enough in most cases. When d=0, then it reduces to an ARMA(p,q) model. An ARIMA(p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model. ARIMA(0,1,0), i.e., y_t=y_{t-1}+ε is a special one and called as the random walk model.

Autocorrelation functions and partial autocorrelation functions

To determine a proper model for a given time series data, it is necessary to carry out the autocorrelation functions (ACF) and partial autocorrelation functions (PACF) analysis. These statistical measures reflect how the observations in a time series are related to each other. For modeling and forecasting purpose, it is often useful to plot the ACF and PACF against consecutive time lags. These plots help in determining the order of AR and MA terms. Below, we give their mathematical definitions:

For a time series{$x(t)$, t=0,1, 2,…} the autocovariance at lag k is defined as

$$\gamma_k = \mathrm{Cov}(x_t, x_{t+k}) = E[(x_t-\mu)(x_{t+k}-\mu)] \tag{9}$$

The *autocorrelation coefficient* at lag k is defined as

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{10}$$

Here, μ is the mean of the time series, i.e., μ=$E[x_t]$. The autocovariance at lag zero, i.e., γ_0 is the variance of the time series. Autocorrelation coefficient ρ_k is dimensionless and so is independent of the scale of measurement also, $-1\leq\rho_k\leq 1$. Statisticians Box and Jenkins termed γ_k as the theoretical autocovariance function (ACVF) and ρ_k as the theoretical autocorrelation function (ACF).

Partial autocorrelation function (PACF) is used to measure the correlation between an observation k period ago and the current observation, after controlling for observations at intermediate lags (i.e., at lags <k). At lag 1, PACF(1) is same as ACF(1).

Stochastic process governing a time series is unknown, and so, it is not possible to determine the actual or theoretical ACF and PACF values. Rather, these values are to be estimated from the training data, i.e., the known time series at hand. The estimated ACF and PACF values from the training data are respectively termed as sample ACF and PACF. The most appropriate sample estimate for the ACVF at lag k is

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t-\mu)(x_{t+k}-\mu) \tag{11}$$

Then the estimate for the sample ACF at lag k is given by

$$r_k = \frac{c_k}{c_0} \tag{12}$$

Here, { $x(t)$, t=0,1,2,.......} is the training series of size n with mean μ.

Figure 2 explains Box and Jenkins methodology procedure; the sample ACF plot is useful in determining the type of model to fit to a time series of length N. Since ACF is symmetrical about lag zero, it is only required to plot the sample ACF for positive lags, from lag one onwards to a maximum lag of about $N/4$. The sample PACF plot helps in identifying the maximum order of an AR process.

Hurst exponent (H)

It refers the index of dependence. It quantifies the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction. The value of the Hurst exponent ranges between 0 and 1. A value of 0.5 indicates a true random walk (a Brownian time series). In a random walk, there is no correlation between any element and a future element. A Hurst exponent value H, 0.5<H<1 indicates "persistent behavior" (a positive autocorrelation). If there is an increase

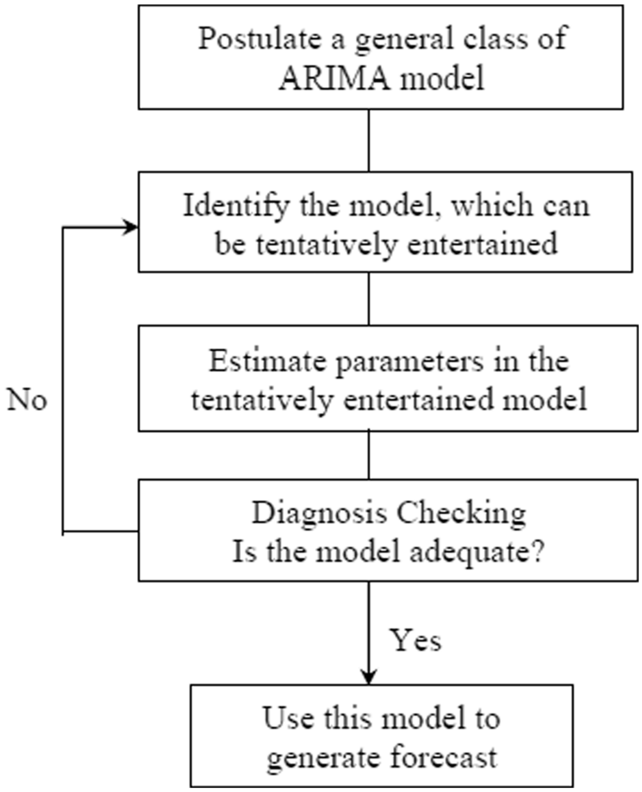

Fig. 2 Box-Jenkins methodology for optimal model selection

from time step t_{i-1} to t_i, there will probably be an increase from t_i to t_{i+1}. The same is true of decreases, where a decrease will tend to follow a decrease. A Hurst exponent value H, $0<H<0.5$ will exist for a time series with “anti-persistent behavior” (or negative autocorrelation). Here, an increase will tend to be followed by a decrease or decrease will be followed by an increase. This behavior is sometimes called “mean reversion” (Rangarajan and Sant 2004).

$$H = \left|\frac{b_{yx}-1}{2}\right| \tag{13}$$

Also, Hurst exponent can be calculated using power law decay (Rangarajan and Ding 2000)

$$p(k) = Ck^{-\alpha} \tag{14}$$

where C is a constant and $p\ (k)$ is the autocorrelation function with lag k. The Hurst exponent is related to the exponent alpha in the equation by the relation

$$H = 1-\frac{\alpha}{2} \tag{15}$$

Fractal dimension (D)

It is a statistical quantity, which gives an indication of how completely a fractal appears to fill space, as one zooms down to finer and finer scales.

$$D = 2-H \tag{16}$$

Also fractal dimension is calculated from the Hausdorff dimension. The Hausdorff dimension D_H, in a metric space, is defined as (Rangarajan and Ding 2000; Rangarajan and Sant 2004)

$$D_H = -\lim_{\varepsilon\to 0}\frac{\ln[N(\varepsilon)]}{\ln\varepsilon} \tag{17}$$

where $N(\varepsilon)$ is the number of open balls of a radius ε needed to cover the entire set. An open ball with center P and radius ε in a metric space with metric d is defined as set of all points x such that $d(P,x)<\varepsilon$.

Predictability index

It describes the behavior of time series (Rangarajan 1997; Rangarajan and Ding 2000; Rangarajan and Sant 2004).

$$PI = 2|D-1.5| \tag{18}$$

Predictability index (PI) value increases when D value becomes less than or greater than 1.5. In the former case, persistence behavior is observed while in the later, an anti-persistence. If one of these indices comes close to 0, then the corresponding process approximates the Brownian motion and is therefore unpredictable.

Results and discussion

Using statistical, time series, and fractal analysis, the quality of water at different sites S_1, S_2, S_3, S_4, and S_5 of full stretch river Yamuna has been discussed. Figure 3 depicts the average value, positional average, mode, standard deviation, skewness, kurtosis, and coefficient of variation for all parameters pH, COD, BOD, AMM, TKN, DO, and WT at sample sites S_1, S_2, S_3, S_4, and S_5, respectively. Table 1 explains trend and time series analysis of ARIMA model, stationary R squared, R squared, RMSE, MAPE, MaxAPE, MAE, Ljung-Box, residual ACF, and residual PACF for all water quality parameters at all sample sites. Figure 4 shows the plot of ACF, PACF, time series, observed data, best fit, lower confidence limit (LCL), and upper confidence limit (UCL). Table 2 gives regression equation, coefficient of determination, Hurst exponent, fractal dimension, and predictability index, and Table 3 depicts fractal and predictability analysis behavior for S_1–S_2, S_1–S_3, S_1–S_4, S_1–S_5, S_2–S_3, S_2–S_4, S_2–S_5, S_3–S_4, S_3–S_5, and S_4–S_5. By using equations (1)–(18), the following are observed:

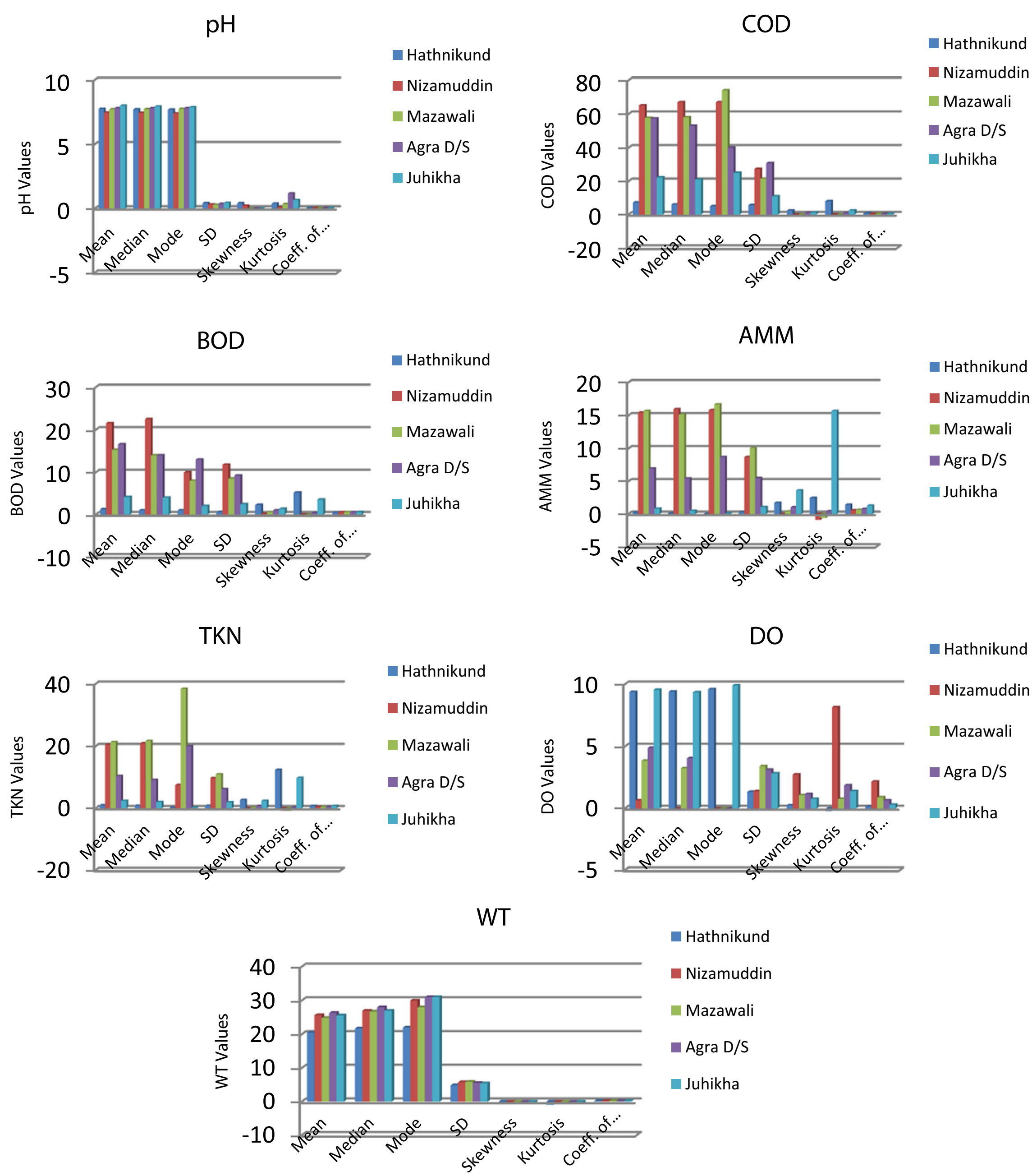

Fig. 3 Graphical representation of statistical analysis of water quality parameters at sample sites of Yamuna River

pH: For all sites, the mean, median, and mode remain within prescribed limits of WHO/EPA and exhibit normal behavior, standard deviation, and skewness values that are close to zero, which show that curve is symmetrical and platykurtic. Prediction model is better than the baseline model as stationary *R* squared and *R* squared values exhibit the similar behavior. RMSE values are low, so dependent series is closed with its model-predicted level. Using Ljung-Box model, for all sites, value of statistics lies between 18 to 29, significance level varies from 0.03 to 0.39, and simple ARIMA model was used for prediction. It is observed that value of pH lies between 7 to 9, and quality of water remains same at all sites, which is calculated at 95 % confidence limits. Anti-persistence behavior exists at all sites except for S_2–S_5 which shows persistence behavior.

COD: For all sites, behavior is not normal, spread of data points is high, and curve is symmetrical and platykurtic, but for S_1, it is nonsymmetrical and leptokurtic. COD crosses the prescribed limits of WHO/EPA at all sites

Table 1 Trend and time series analysis of water quality parameters

Fit statistic		Hathnikund (S_1)	Nizamuddin (S_2)	Mazawali (S_3)	Agra D/S (S_4)	Juhikha (S_5)
Stationary *R* squared	Mean for pH	0.225	0.31	2.22E-13	0.045	0.271
R squared		0.225	0.089	2.22E-13	0.045	0.136
RMSE		0.373	0.309	0.295	0.363	0.413
MAPE		3.811	3.347	2.93	3.407	3.899
MaxAPE		12.886	11.305	12.178	17.046	26.158
MAE		0.297	0.251	0.225	0.265	0.311
Statistics	Ljung-Box Q(18) for pH	19.847	18.061	20.071	29.212	23.9
DF		17	17	18	17	17
Sig.		0.282	0.385	0.329	0.033	0.122
Stationary *R* squared	Mean for COD	0.823	0.229	0.35	0.189	0.069
R squared		0.022	0.229	0.35	0.172	0.059
RMSE		5.63	24.771	17.472	28.175	10.602
MAPE		83.196	60.779	30.648	49.141	50.135
MaxAPE		670.741	1.63E+03	169.07	262.859	485.991
MAE		3.812	19.088	13.961	21.828	7.752
Statistics	Ljung-Box Q(18) for COD	22.358	11.975	24.688	11.133	17.256
DF		17	15	16	17	17
Sig.		0.171	0.681	0.76	0.85	0.437
Stationary *R* squared	Mean for BOD	0.422	0.434	0.089	0.322	0.39
R squared		0.016	0.434	0.086	0.06	0.045
RMSE		0.575	8.962	8.177	8.976	2.381
MAPE		26.594	68.162	67.176	56.364	60.258
MaxAPE		64.781	706.023	516.894	553.774	341.494
MAE		0.388	7.347	6.639	6.928	1.852
Statistics	Ljung-Box Q(18) for BOD	24.471	18.227	33.665	15.96	29.257
DF		17	16	17	17	17
Sig.		0.107	0.311	0.009	0.527	0.032
Stationary *R* squared	Mean for AMM	0.423	0.302	0.394	0.174	−1.01E-15
R squared		0.008	0.302	0.394	0.171	−6.47E-06
RMSE		0.287	7.354	7.964	5.013	1.058
MAPE		101.265	145.958	131.082	157.18	191.539
MaxAPE		1.48E+03	4.43E+03	3.80E+03	3.74E+03	1.95E+03
MAE		0.216	5.933	6.078	4.091	0.632
Statistics	Ljung-Box Q(18) for AMM	19.521	34.414	13.041	36.819	18.763
DF		17	15	14	17	18
Sig.		0.299	0.003	0.523	0.004	0.407
Stationary *R* squared	Mean for TKN	0.441	0.255	0.416	0.172	−1.22E-15
R squared		0.047	0.255	0.416	0.169	−0.002
RMSE		0.811	8.513	8.502	5.68	1.98
MAPE		74.887	62.387	51.205	79.105	121.042
MaxAPE		416.091	548.081	563.396	1.62E+03	873.403
MAE		0.522	6.744	6.739	4.475	1.384
Statistics	Ljung-Box Q(18) for TKN	22.828	22.456	17.593	29.374	16.836
DF		17	17	15	17	18
Sig.		0.155	0.168	0.285	0.031	0.534
Stationary *R* squared	Mean for DO	0.298	0.152	0.189	0.22	0.175
R squared		0.298	0.152	0.189	0.22	0.162
RMSE		1.148	1.295	3.086	2.816	2.612
MAPE		9.563	93.576	94.939	60.331	23.028

Table 1 (continued)

Fit statistic		Hathnikund (S_1)	Nizamuddin (S_2)	Mazawali (S_3)	Agra D/S (S_4)	Juhikha (S_5)
MaxAPE		49.075	988.732	2.06E+03	1.30E+03	84.5
MAE		0.877	0.827	2.343	2.007	2.013
Statistics	Ljung-Box Q(18) for DO	36.042	16.859	32.406	12.421	39.549
DF		15	17	17	14	16
Sig.		0.002	0.464	0.013	0.573	0.001
Stationary *R* squared	Mean for WT	0.561	0.588	0.569	0.706	0.546
R squared		0.561	0.588	0.569	0.706	0.546
RMSE		3.277	3.754	3.915	3.084	3.705
MAPE		13.44	11.844	13.305	9.937	12.5
MaxAPE		70.796	90.229	76.704	69.855	78.318
MAE		2.523	2.678	2.96	2.371	2.931
Statistics	Ljung-Box Q(18) for WT	24.724	23.76	32.84	72.304	76.098
DF		15	15	15	13	15
Sig.		0.054	0.069	0.005	0	0

with maximum at S_2 and minimum at S_1. Time series model is better than the baseline model as stationary *R* squared and *R* squared values exhibit the similar behavior. RMSE value is low so dependent series is closed with its model-predicted level. Using Ljung-Box model, for all sites, value of statistics lies from 11 to 25, significance level ranging from 0.17 to 0.85, and simple ARIMA model was used for prediction. Using plots of ACF, PACF, lag, and time series, it is observed that value of COD lies between 0 to 18 for S_1, 0 to 120 for S_2, 0 to 100 for S_3, 0 to 150 for S_4, and 0 to 60 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence interval. It is observed that persistence behavior exist for S_1–S_2, S_1–S_3, and S_1–S_4; anti-persistence for S_2–S_3, S_2–S_4, S_2–S_5, S_3–S_4, S_3–S_5, and S_4–S_5; and Brownian time series (true random walk) for S_1–S_5.

BOD: Behavior is not normal, spread of data points is high except for S_1, and spread of data is low. BOD exhibits symmetrical behavior. Curve is platykurtic for all sites except for S_1 and S_5, which shows leptokurtic behavior. BOD remains within prescribed limits of WHO/EPA at S_1 and S_5, but for S_2, S_3, S_4 crosses prescribed limits with maximum at S_2. Model is better than the baseline model as stationary *R* squared and *R* squared value exhibit the similar behavior. RMSE value is low, so dependent series is closed with its model-predicted level for all sites. From Ljung-Box model, for all sites, statistics lies between 15 and 34, significance varies from 0.01 to 0.53, and simple ARIMA model was used for prediction. ACF, PACF, lag, and time series explains that value of BOD lies between 0 to 2 for S_1, 0 to 50 for S_2, 0 to 35 for S_3, 0 to 40 for S_4, and 0 to 6 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence limits. It is observed that anti-persistence behavior exist for S_2–S_3, S_2–S_4, S_2–S_5, S_3–S_4, S_3–S_5, and S_4–S_5 and Brownian time series (true random walk) for S_1–S_2, S_1–S_3, S_1–S_4, and S_1–S_5.

AMM: For all sites, behavior exhibits normal. Spread of data is high for all sites except for S_1 and S_5. Curve is symmetrical and platykurtic expect for S_5; it is nonsymmetrical and leptokurtic. AMM crosses the prescribed limits of WHO/EPA at all sites except S_1 and S_5 with maximum at S_2, S_3. For all sites, stationary *R* squared and *R* squared value reveal the similar behavior so prediction model is better than the baseline model. RMSE value is low; thus, dependent series is closed with its model-predicted level except at S_2, S_3, and S_4. Using Ljung-Box model, statistics lies between 13 and 37, significance level varies from 0.00 to 0.52, and simple ARIMA model was used for prediction. ACF, PACF, lag, and time series shows that value of AMM lies between 0 to 1 for S_1, 0 to 30 for S_2, 0 to 35 for S_3,0 to 20 for S_4, and 0 to 4 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence interval. It is observed that persistence behavior exist for S_1–S_5, S_3–S_5, and S_4–S_5; anti-persistence for S_2–S_3, S_2–S_4, S_2–S_5, and S_3–S_4; and Brownian time series (true random walk) for S_1–S_2, S_1–S_3, and S_1–S_4.

TKN: For all sites, curve is not normal. Data spread is high, symmetric, platykurtic for all sites except S_1 and S_5, which has spread low, nonsymmetrical, and leptokurtic. TKN crosses the prescribed limits of WHO/EPA at all sites except for S_1 and S_5 with maximum at S_3. Prediction model is better than the baseline model as stationary *R* squared and *R* squared value reveal the similar behavior. RMSE values are low, so dependent series is closed with its model-predicted level except for S_2, S_3, and S_4. Using Ljung-Box model, value of statistics ranges from 16 to

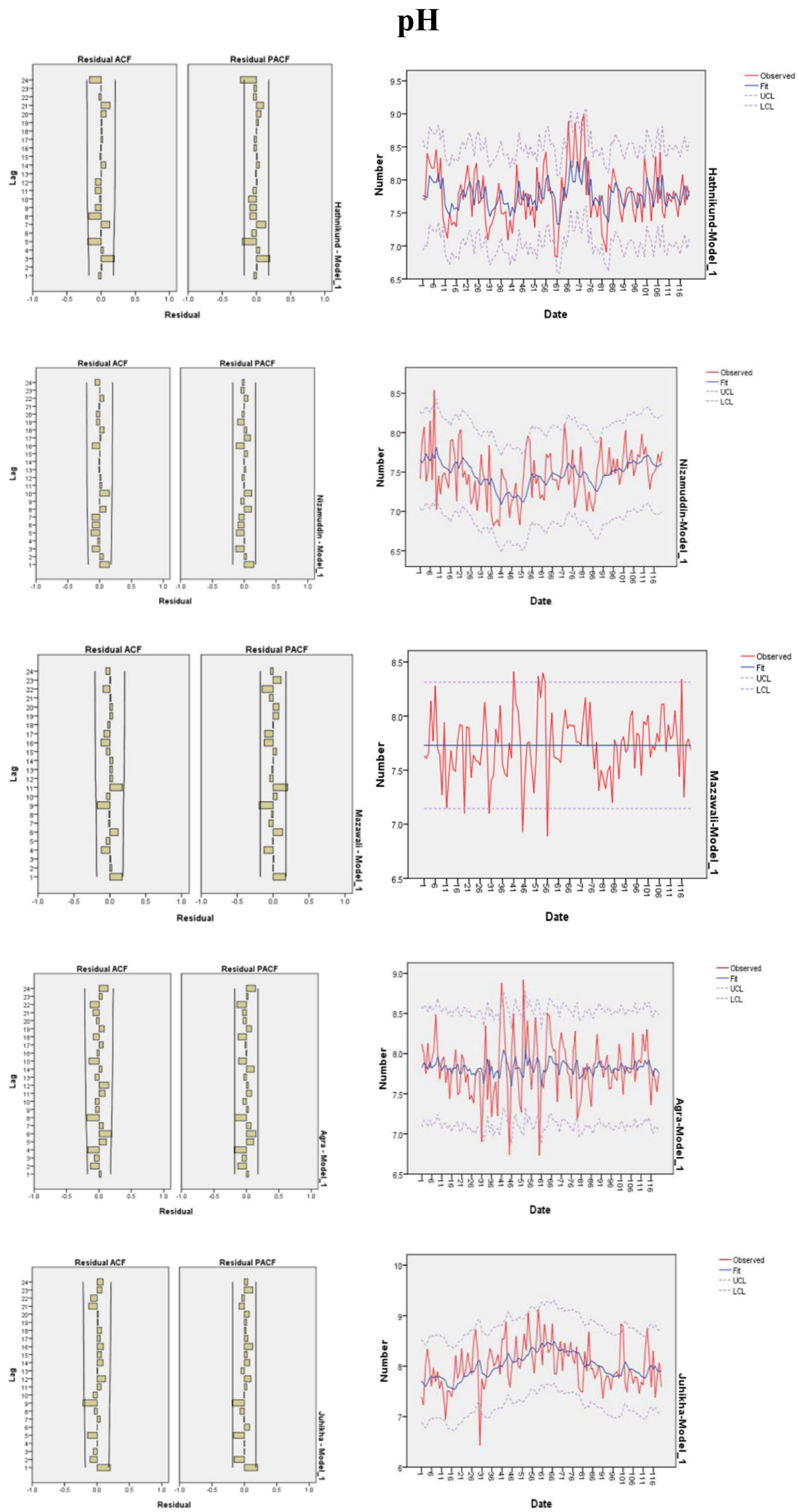

Fig. 4 Graphical representation of trend, time series analysis (ACF, PACF, observed, best fit, LCL, UCL) of water quality parameters

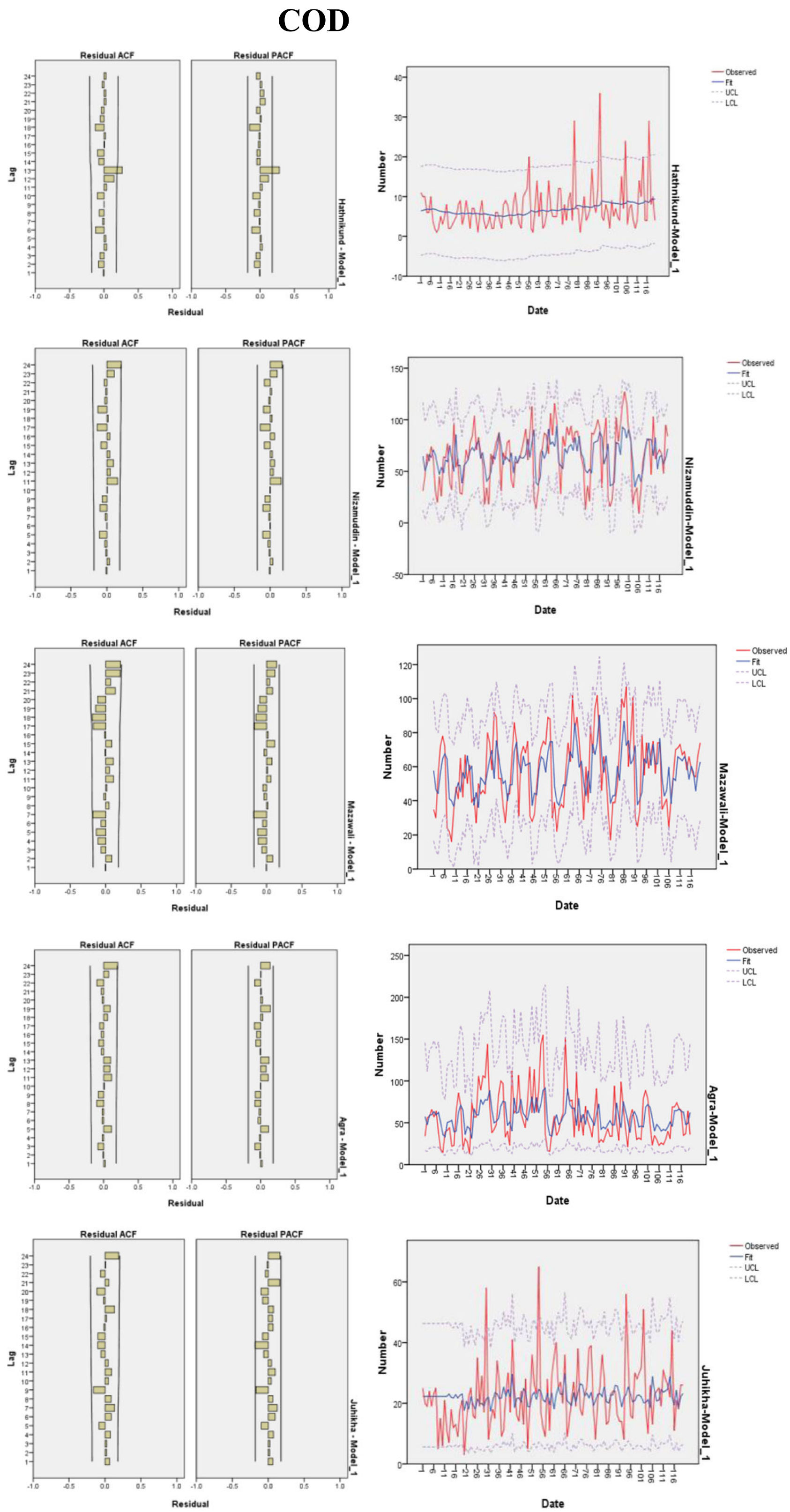

Fig. 4 (continued)

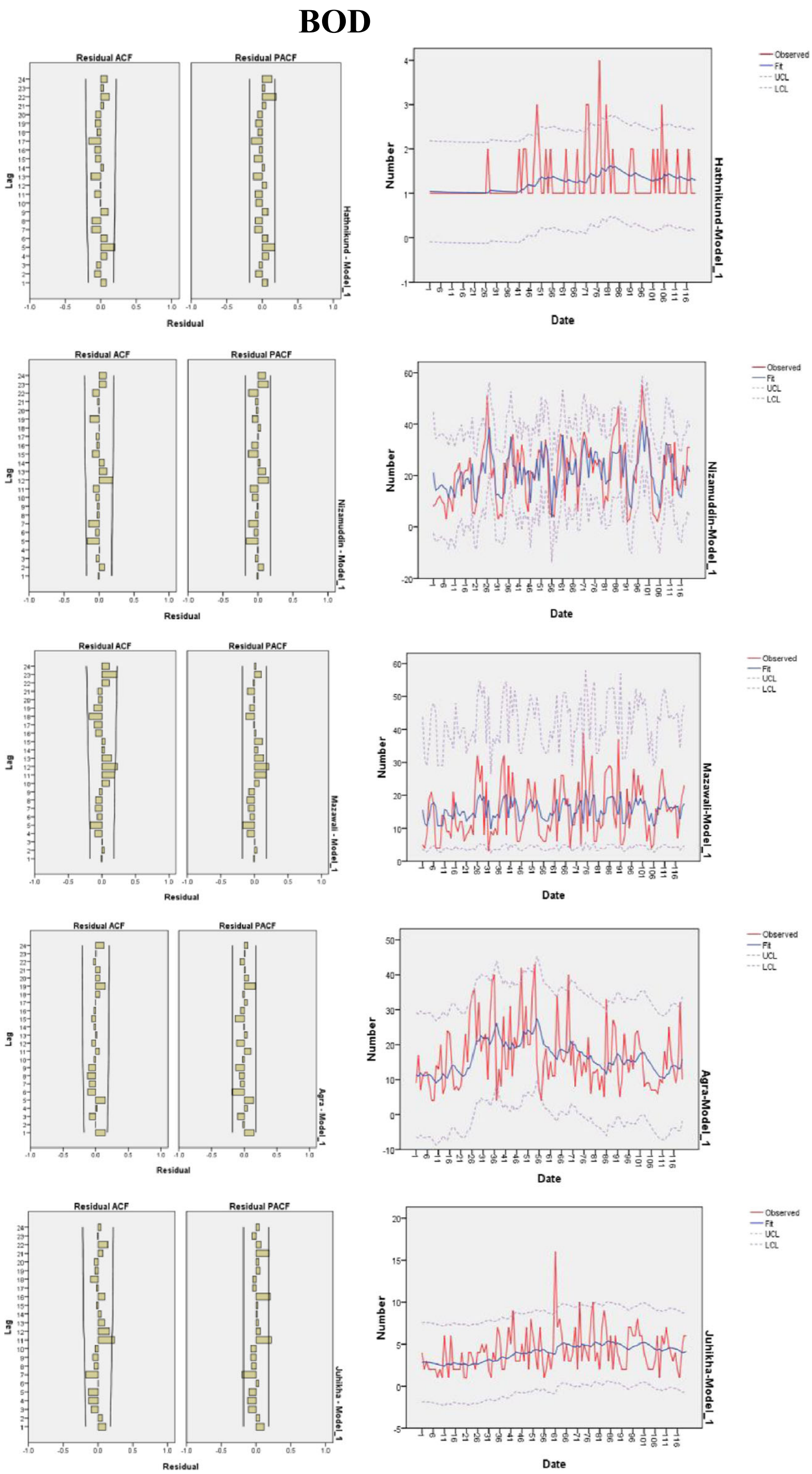

Fig. 4 (continued)

29, significance level lies from 0.03 to 0.52, and simple ARIMA model was used for prediction. ACF, PACF, lag, and time series shows that value of TKN lies between 0 to 3 for S_1, 0 to 40 for S_2, 0 to 40 for S_3, 0 to 25 for S_4, and 0 to 75 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence limits. It is observed that persistence behavior exist for S_1–S_5; anti-persistence for S_2–S_3, S_2–S_4, S_2–S_5, S_3–S_4, S_3–S_5, and S_4–S_5; and Brownian time series (true random walk) for S_1–S_2, S_1–S_3, and S_1–S_4.

AMM

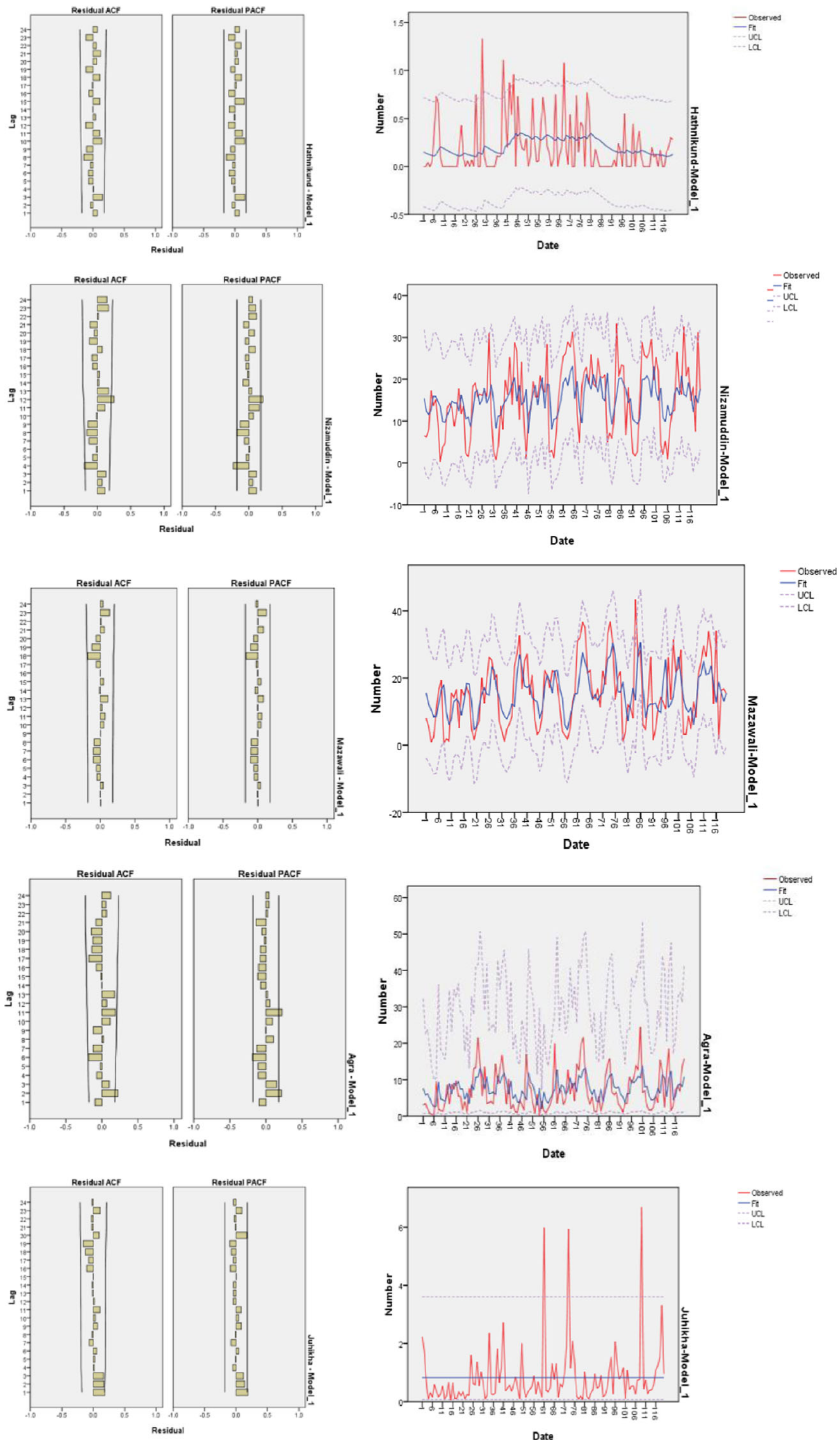

Fig. 4 (continued)

DO: Mean, median, and mode are same and thus behaves normally for S_1, S_2 and S_5. At all sites, data spread is low and symmetrical. Curve is platykurtic at all sites except for S_2, which has leptokurtic. WHO/EPA standards are not satisfied by DO at S_2, S_3, and S_4 except for S_1 and S_5.

For all sites, time series model is better than the baseline model as stationary *R* squared and *R* squared value exhibit the similar behavior. RMSE values are low, so dependent series is closed with its model-predicted level except for S_3, S_4, and S_5. From Ljung-Box model, for all

TKN

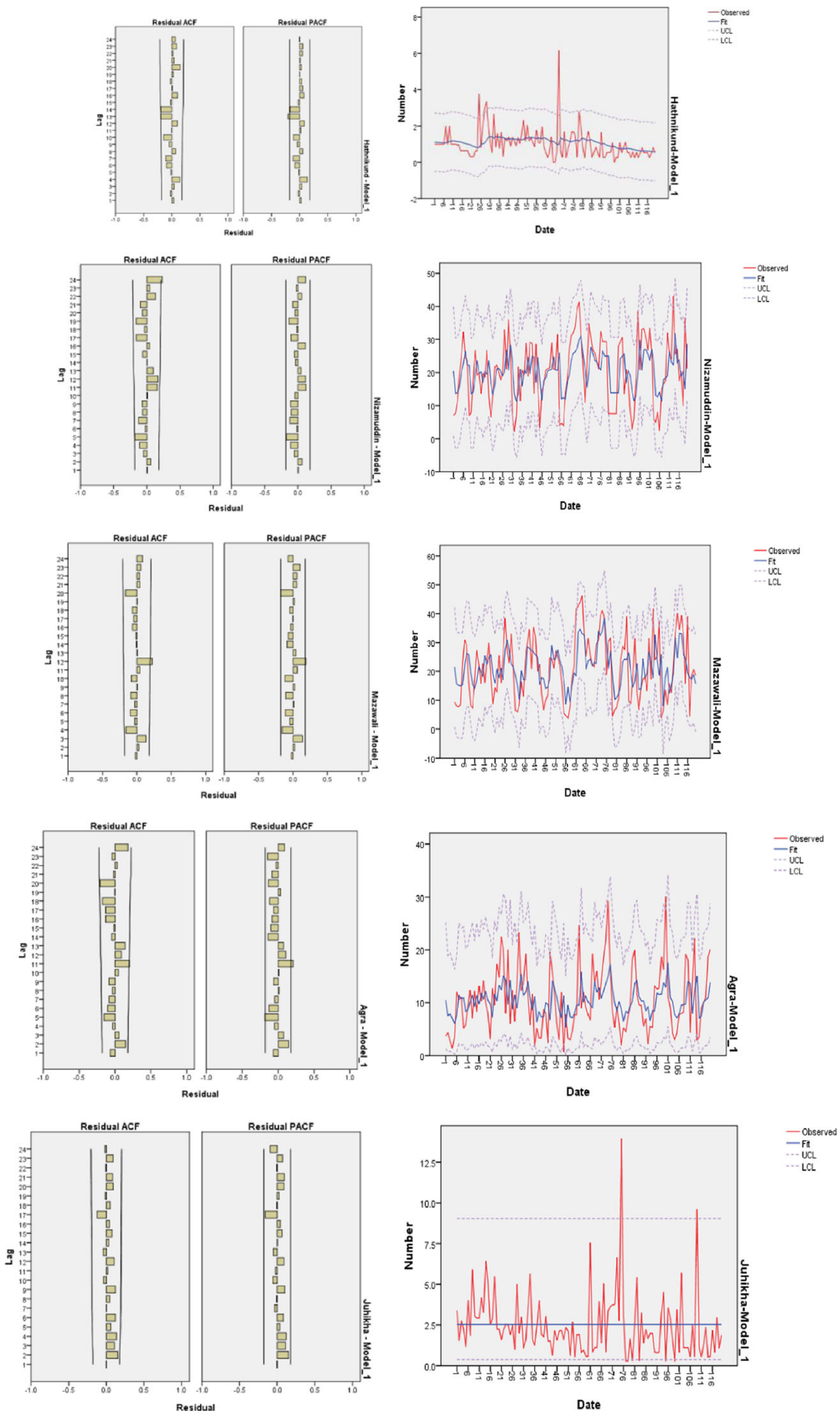

Fig. 4 (continued)

sites, value of statistics lies between 12 and 40, significance level between 0.00 and 0.57, and simple ARIMA model was used for prediction. Using plots of ACF, PACF, lag, and time series, it is observed that value of DO lies between 7 to 13 for S_1, 0 to 5 for S_2, 0 to 14 for S_3, 0 to 13 for S_4, and 5 to 15 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence scale. It is observed that persistence behavior exist for S_1–S_2, S_1–S_3, S_1–S_4, S_2–S_5, and S_3–S_5 and anti-persistence for S_1–S_5, S_2–S_3, S_2–S_4, S_3–S_4, and S_4–S_5.

DO

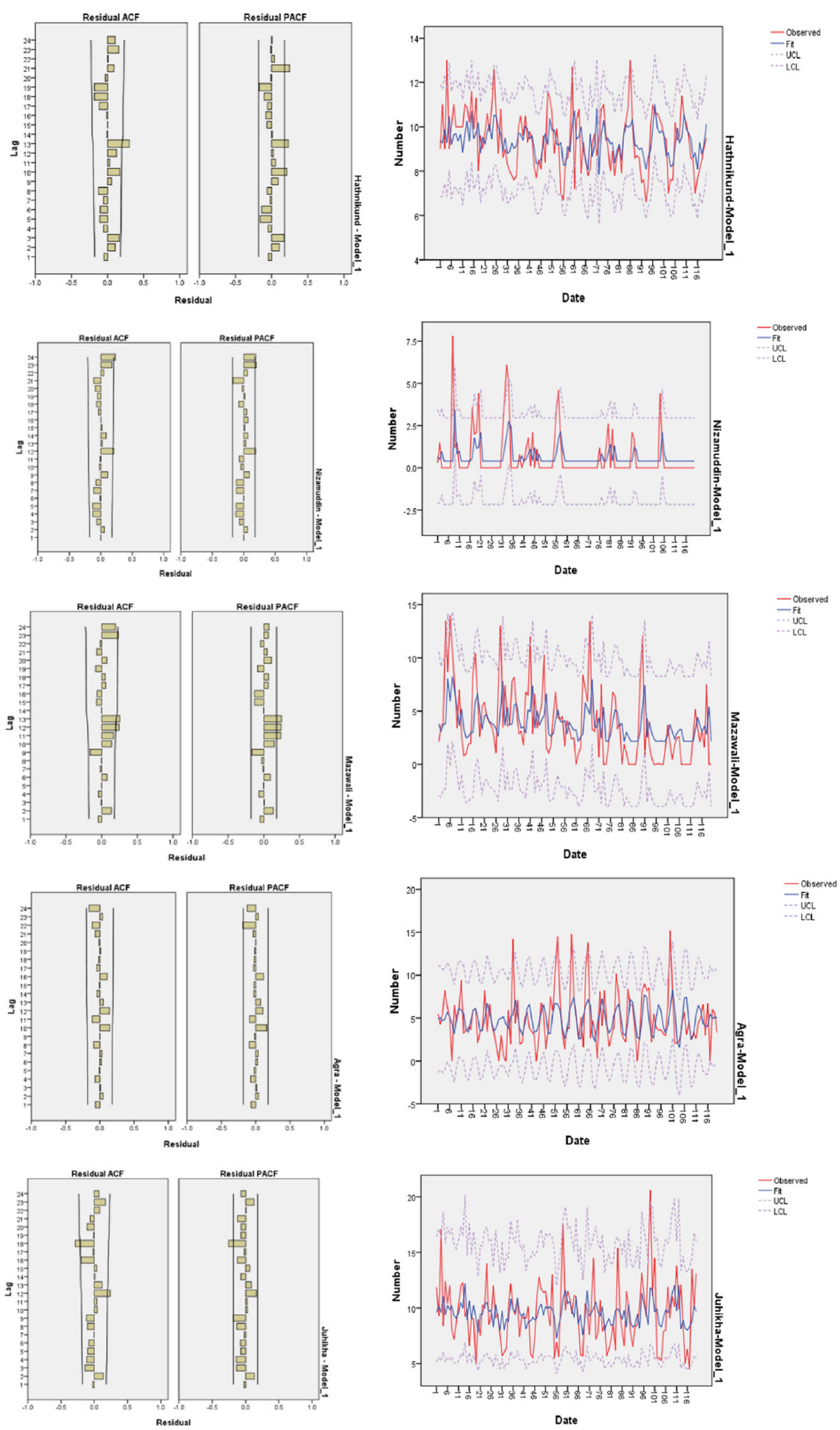

Fig. 4 (continued)

WT: For all sites, mean, median, and mode remains within the prescribed limits of WHO/EPA, exhibits normal behavior, standard deviation is high, and spread is same and symmetrical. Curve is platykurtic except for S_2, which has leptokurtic. Model is better than the baseline model as stationary *R* squared and *R* squared value behave alike. RMSE values are low, so dependent series is closed with its model-predicted level. Using Ljung-Box model, value of statistics lies between 23 and 76, significance level between 0 and 0.07, and simple ARIMA

WT

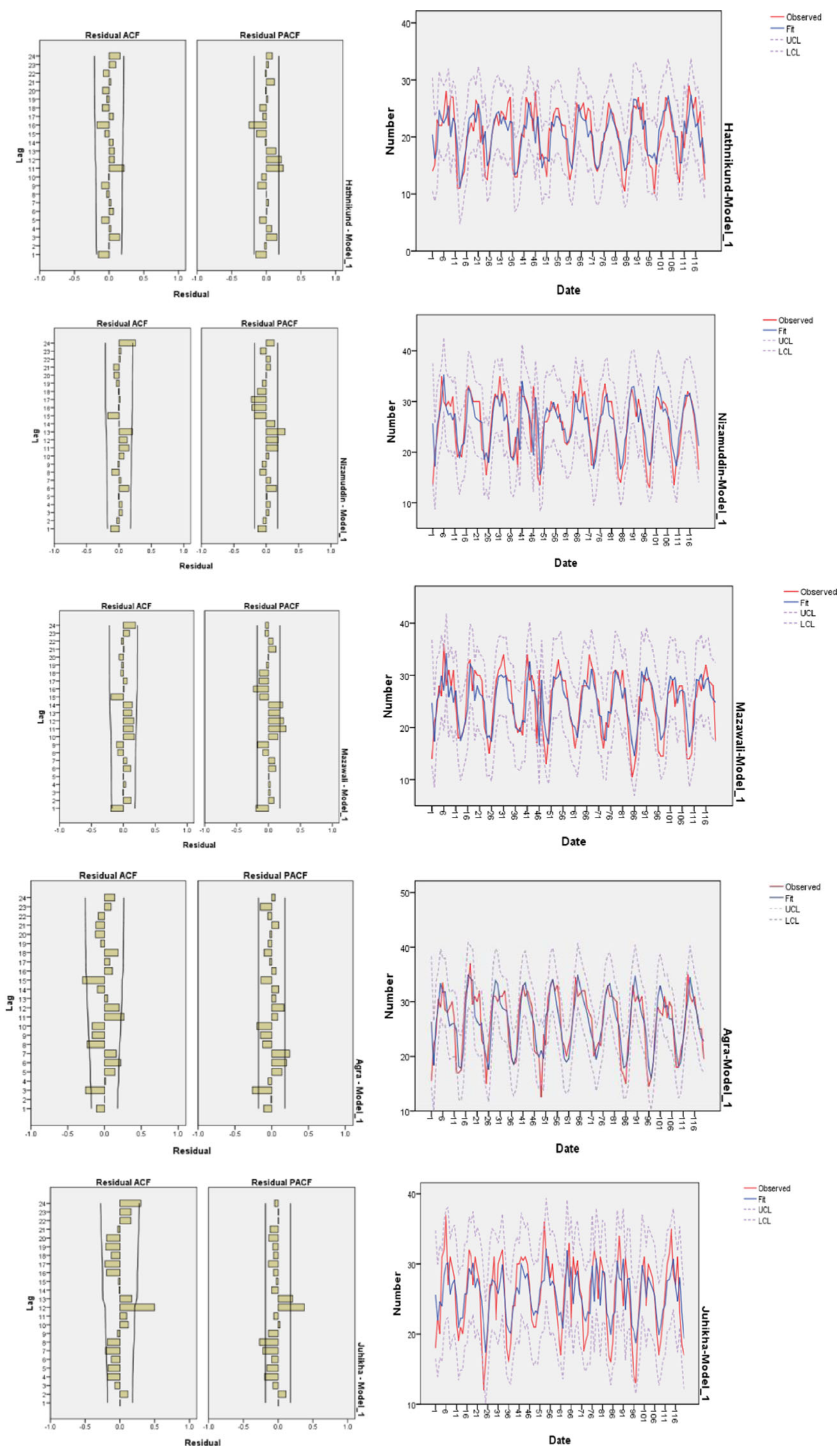

Fig. 4 (continued)

model was used for prediction. Using plots of ACF, PACF, lag, and time series, it is observed that value of WT lies between 10 to 28 for S_1, 15 to 35 for S_2, 15 to 35 for S_3, 13 to 35 for S_4, and 18 to 30 for S_5 and the quality of water gets effected at all sites, which is calculated at 95 % confidence limits. It is observed that anti-persistence behavior exist for S_1–S_2, S_1–S_3, S_1–S_4, S_1–S_5, S_2–S_3, S_2–S_4, S_2–S_5, S_3–S_4, S_3–S_5, and S_4–S_5.

Table 2 Fractal analysis of water quality parameters

pH	Hathnikund (S_1)	Sample Stations	Regression equation	r^2	b_{yx}	*H*(abs)	*D*(fractal)	PI
		Nizamuddin (S_2)	y=0.3819x+4.9138	0.086	0.3819	0.3091	1.69095	0.3819
		Mazawali (S_3)	y=0.4431x+4.343	0.0962	0.4431	0.2785	1.72155	0.4431
		Agra D/S (S_4)	y=0.2913x+5.4889	0.0652	0.2913	0.3544	1.64565	0.2913
		Juhikha (S_5)	y=0.2973x+5.3873	0.0981	0.2973	0.3514	1.64865	0.2973
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.4903x+3.6829	0.1998	0.4903	0.2549	1.74515	0.4903
		Agra D/S (S_4)	y=0.2516x+5.5042	0.0825	0.2516	0.3742	1.6258	0.2516
		Juhikha (S_5)	y=−0.0438x+7.823	0.0036	−0.0438	0.5219	1.4781	0.0438
	Mazawali (S_3)	Agra D/S (S_4)	y=0.3834x+4.7303	0.2304	0.3834	0.3083	1.6917	0.3834
		Juhikha (S_5)	y=0.2973x+5.3873	0.0981	0.2973	0.3514	1.64865	0.2973
	Agra D/S (S_4)	Juhikha (S_5)	y=0.2015x+6.2096	0.0586	0.2015	0.3993	1.60075	0.2015
COD	Hathnikund (S_1)	Nizamuddin (S_2)	*y*=−0.0354x+9.5039	0.029	0.0354	0.5177	1.4823	0.0354
		Mazawali (S_3)	y=−0.0475x+9.9418	0.0322	0.0475	0.5238	1.47625	0.0475
		Agra D/S (S_4)	y=−0.0295x+8.8894	0.0255	0.0295	0.5148	1.48525	0.0295
		Juhikha (S_5)	y=−0.0042x+7.2931	6.00E-05	0.0042	0.5021	1.4979	0.0042
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.8917x+13.614	0.4903	0.8917	0.0542	1.94585	0.8917
		Agra D/S (S_4)	y=0.4034x+41.955	0.2068	0.4034	0.2983	1.7017	0.4034
		Juhikha (S_5)	y=0.9577x+43.749	0.1452	0.9577	0.0212	1.97885	0.9577
	Mazawali (S_3)	Agra D/S (S_4)	y=0.2972x+40.67	0.182	0.2972	0.3514	1.6486	0.2972
		Juhikha (S_5)	y=0.4171x+48.412	0.0447	0.4171	0.2915	1.70855	0.4171
	Agra D/S (S_4)	Juhikha (S_5)	y=0.8701x+37.915	0.0943	0.8701	0.065	1.93505	0.8701
BOD	Hathnikund (S_1)	Nizamuddin (S_2)	y=−0.0035x+1.3509	0.0051	−0.0035	0.5018	1.49825	0.0035
		Mazawali (S_3)	y=−0.0024x+1.3113	0.0012	−0.0024	0.5012	1.4988	0.0024
		Agra D/S (S_4)	y=−0.002x+1.3089	0.0011	−0.002	0.501	1.499	0.002
		Juhikha (S_5)	y=0.0124x+1.2233	0.0027	0.0124	0.4938	1.5062	0.0124
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.9087x+7.6396	0.4296	0.9087	0.0457	1.95435	0.9087
		Agra D/S (S_4)	y=0.4025x+14.877	0.0995	0.4025	0.2988	1.70125	0.4025
		Juhikha (S_5)	y=1.7739x+14.182	0.1338	1.7739	0.387	1.61305	0.2261
	Mazawali (S_3)	Agra D/S (S_4)	y=0.2257x+11.57	0.0602	0.2257	0.3872	1.61285	0.2257
		Juhikha (S_5)	y=1.2634x+10.063	0.1305	1.2634	0.1317	1.8683	0.7366
	Agra D/S (S_4)	Juhikha (S_5)	y=0.0497x+3.3338	0.0356	0.0497	0.4752	1.52485	0.0497
AMM	Hathnikund (S_1)	Nizamuddin (S_2)	y=0.0019x+0.1742	0.0033	0.0019	0.4991	1.50095	0.0019
		Mazawali (S_3)	y=0.0007x+0.1927	0.0006	0.0007	0.4997	1.50035	0.0007
		Agra D/S (S_4)	y=−0.0013x+0.2128	0.0006	−0.0013	0.5007	1.49935	0.0013
		Juhikha (S_5)	y=−0.0295x+0.2281	0.0118	−0.0295	0.5148	1.48525	0.0295
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.6122x+5.8183	0.5015	0.6122	0.1939	1.8061	0.6122
		Agra D/S (S_4)	y=0.7183x+10.444	0.2053	0.7183	0.1409	1.85915	0.7183
		Juhikha (S_5)	y=1.7587x+13.977	0.0458	1.7587	0.3794	1.62065	0.2413
	Mazawali (S_3)	Agra D/S (S_4)	y=0.9935x+8.8038	0.2935	0.9935	0.0032	1.99675	0.9935
		Juhikha (S_5)	y=2.3513x+13.756	0.0612	2.3513	0.6757	1.32435	0.3513
	Agra D/S (S_4)	Juhikha (S_5)	y=2.1327x+5.1774	0.1692	2.1327	0.5664	1.43365	0.1327
TKN	Hathnikund (S_1)	Nizamuddin (S_2)	y=−0.0066x+1.1562	0.006	−0.0066	0.5033	1.4967	0.0066
		Mazawali (S_3)	y=−0.0017x+1.0572	0.0005	−0.0017	0.5009	1.49915	0.0017
		Agra D/S (S_4)	y=0.002x+1.0005	0.0002	0.002	0.499	1.501	0.002
		Juhikha (S_5)	y=−0.0225x+1.0766	0.0029	−0.0225	0.5113	1.48875	0.0225
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.6009x+7.6638	0.4517	0.6009	0.1996	1.80045	0.6009
		Agra D/S (S_4)	y=0.6299x+13.9	0.1583	0.6299	0.1851	1.81495	0.6299
		Juhikha (S_5)	y=0.2399x+19.912	0.0023	0.2399	0.3801	1.61995	0.2399
	Mazawali (S_3)	Agra D/S (S_4)	y=0.8403x+12.558	0.2251	0.8403	0.0799	1.92015	0.8403
		Juhikha (S_5)	y=1.0935x+18.686	0.0388	1.0935	0.0468	1.95325	0.9065

Table 2 (continued)

	Agra D/S (S_4)	Juhikha (S_5)	y=1.0383x+7.9361	0.1096	1.0383	0.0192	1.98085	0.9617
DO	Hathnikund (S_1)	Nizamuddin (S_2)	y=−0.2495x+9.5318	0.0675	−0.2495	0.6248	1.37525	0.2495
		Mazawali (S_3)	y=−0.0339x+9.5014	0.0074	−0.0339	0.517	1.48305	0.0339
		Agra D/S (S_4)	y=−0.0293x+9.5138	0.0047	−0.0293	0.5147	1.48535	0.0293
		Juhikha (S_5)	y=0.1416x+8.0185	0.0888	0.1416	0.4292	1.5708	0.1416
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.0737x+0.3616	0.0322	0.0737	0.4632	1.53685	0.0737
		Agra D/S (S_4)	y=0.0285x+0.5058	0.0041	0.0285	0.4858	1.51425	0.0285
		Juhikha (S_5)	y=−0.1584x+2.158	0.1025	−0.1584	0.5792	1.4208	0.1584
	Mazawali (S_3)	Agra D/S (S_4)	y=0.2259x+2.7451	0.043	0.2259	0.3871	1.61295	0.2259
		Juhikha (S_5)	y=−0.1472x+5.2531	0.0149	−0.1472	0.5736	1.4264	0.1472
	Agra D/S (S_4)	Juhikha (S_5)	y=0.1429x+3.5158	0.0166	0.1429	0.4286	1.57145	0.1429
WT	Hathnikund (S_1)	Nizamuddin (S_2)	y=0.6731x+3.2178	0.6325	0.6731	0.1635	1.83655	0.6731
		Mazawali (S_3)	y=0.6436x+4.5158	0.6012	0.6436	0.1782	1.8218	0.6436
		Agra D/S (S_4)	y=0.6696x+2.8635	0.5825	0.6696	0.1652	1.8348	0.6696
		Juhikha (S_5)	y=0.6322x+4.3292	0.4926	0.6322	0.1839	1.8161	0.6322
	Nizamuddin (S_2)	Mazawali (S_3)	y=0.8978x+3.3784	0.838	0.8978	0.0511	1.9489	0.8978
		Agra D/S (S_4)	y=0.8755x+2.6159	0.7134	0.8755	0.0623	1.93775	0.8755
		Juhikha (S_5)	y=0.8236x+4.6097	0.5989	0.8236	0.0882	1.9118	0.8236
	Mazawali (S_3)	Agra D/S (S_4)	y=0.9063x+0.9649	0.7353	0.9063	0.0469	1.95315	0.9063
		Juhikha (S_5)	y=0.8269x+3.685	0.5807	0.8269	0.0866	1.91345	0.8269
	Agra D/S (S_4)	Juhikha (S_5)	y=0.8569x+4.4215	0.6966	0.8569	0.0716	1.92845	0.8569

Conclusion

River water quality management, using statistical, trend, and time series analysis has been studied for full stretch Yamuna River. It is observed that for most of the sites, RMSE value are comparatively very low which shows that dependent series is closed with the model predicted level; thus, predictive model is useful at 95 % confidence limits, and all water parameters exhibits platykurtic curve. For COD, BOD, AMM, and TKN parameters, the observed values are increasing from Hathnikund to Nizamuddin and almost remains constant between Nizamuddin to Mazawali and Agra D/S, but then it decreases again at Juhikha but not maintains the same water quality standard as at Hathnikund. ACF and PACF plots of original data indicates that the data is stationary and therefore does not require differencing (d=0); thus, series is serially independent. Water quality does not remain same at all sites for all parameters except for pH. Brownian motion (true random walk) behavior exists at different sites for BOD, AMM, and TKN; therefore, water quality trend is unpredictable.

In comparison to all sites, quality of Yamuna River water at Hathnikund is good, declines at Nizamuddin, Mazawali, Agra D/S, and gain good quality again at Juhikha. The quality of water declines at Nizamuddin, Mazawali, Agra D/S because of the mixing of municipal, agricultural, drains and industrial waste in large scale at these sites, and as Yamuna River reaches at Juhikha after traveling a long distance, then most of the polluted river water parameters settled down or wash out; thus, water again gain good quality at Juhikha. For all sites, almost all parameters except pH and WT crosses the prescribed limits of WHO/EPA; thus, water is not fit for drinking, agriculture, and industrial use.

Table 3 Fractal Analysis of Water Quality Parameters for Different Sites of Yamuna River (AP- Anti persistence, P- Persistence, B- Brownian time series motion)

Sites/parameters	pH	COD	BOD	AMM	TKN	DO	WT
S_1 to S_2	AP	P	B	B	B	P	AP
S_1 to S_3	AP	P	B	B	B	P	AP
S_1 to S_4	AP	P	B	B	B	P	AP
S_1 to S_5	AP	B	B	P	P	AP	AP
S_2 to S_3	AP	AP	AP	AP	AP	AP	AP
S_2 to S_4	AP	AP	AP	AP	AP	AP	AP
S_2 to S_5	P	AP	AP	AP	AP	P	AP
S_3 to S_4	AP	AP	AP	AP	AP	AP	AP
S_3 to S_5	AP	AP	AP	P	AP	P	AP
S_4 to S_5	AP	AP	AP	P	AP	AP	AP

Acknowledgments Authors are thankful to University Grant Commission (UGC) (F. 41-803/2012 (SR)), Government of India for financial support; Central Pollution Control Board (CPCB), Government of India for providing the research data; and Guru Gobind Singh Indraprastha University, New Delhi (India) for providing research facilities. First author is also thankful to Sant Baba Bhag Singh Institute of Engineering and Technology for providing study leave to pursue research degree.

References

Akoto O, Adiyiah J (2007) Chemical analysis of drinking water from some communities in the Brong Ahafo region. Int J Environ Sci Technol 4(2):211–214

Alam MJB, Islam MR, Muyen Z, Mamun M, Islam S (2007) Water quality parameters along rivers. Int J Environ Sci Technol 4(1):159–167

Amiri BJ, Nakane K (2009) Modeling the linkage between river water quality and landscape metrics in the Chugoku district of Japan. Water Resour Manag 23:931–956

APHA (1995) Standard methods for examination of water and waste water American Public Health Association. Washington D.C.19th Edn

Bhardwaj R, Parmar KS (2013a) Water quality index and fractal dimension analysis of water parameters. Int J Environ Sci Technol 10(1): 151–164

Bhardwaj R, Parmar KS (2013b) Wavelet and statistical analysis of river water quality parameters. App Math Comput 219(20):10172–10182

Bhardwaj R, Parmar KS (2014) Water quality management using statistical analysis and time-series prediction model. Appl Water Sci. doi: 10.1007/s13201-014-0159-9

Boskidis I, Gikas GD, Sylaios GK, Tsihrintzis VA (2012) Hydrologic and water quality modeling of lower Nestos river basin. Water Resour Manag 26:3023–3051

Box GEP, Jenkins GM, Reinsel GC (2008) Time series analysis: forecasting and control, 4th edn. John Wiely & Sons; Inc, U.K

Calvo IP, Estrada JCG, Savic D (2012) Heuristic modelling of the water resources management in the Guadalquivir river basin, southern Spain. Water Resour Manag 26:185–209

Chenini I, Khemiri S (2009) Evaluation of ground water quality using multiple linear regression and structural equation modeling. Int J Environ Sci Technol 6(3):509–519

CPCB (2006) Water Quality Status of Yamuna River (1999–2005): Central Pollution Control Board, Ministry of Environment & Forests, Assessment and Development of River Basin Series: ADSORBS/41/2006-07

DeLurgio SA (1998) Forecasting Principles and Applications; 1st Edition. Irwin McGraw-Hill Publishers

Diodato N, Guerriero L, Fiorillo F, Esposito L, Revellino P, Grelle G, Guadagno FM (2014) Predicting monthly spring discharges using a simple statistical model. Water Resour Manage 28:969–978

Dutta D, Wilson K, Welsh YD, Nicholls D, Kim S, Cetin L (2013) A new river system modelling tool for sustainable operational management of water resources. J Environ Manag 121:13–28

Fang H, Wang X, Lou L, Zhou Z, Wu J (2010) Spatial variation and source apportionment of water pollution in Qiantang River (China) using statistical techniques. Water Res 44(5):1562–1572

Hermans C, Erickson J, Noordewier T, Sheldon A, Kline M (2007) Collaborative environmental planning in river management: an application of multicriteria decision analysis in the White River Watershed in Vermont. J Environ Manag 84:534–546

Kahya E, Kalayci S (2004) Trend analysis of streamflow in Turkey. J Hydrol 289:128–144

McCleary R and Hay RA (1980) Applied Time Series Analysis for the Social Sciences,Beverly Hills, Sage

Mousavi M, Kiani S, Lotfi S, Naeemi N, Honarmand M (2008) Transient and spatial modeling and simulation of polybrominated diphenyl ethers reaction and transport in air, water and soil. Int J Environ Sci Technol 5(3):323–330

Movahed M, Hermanisc E (2008) Fractal analysis of river flow fluctuations. Physica A 387(4):915–932

Panepinto D, Genon G (2010) Modeling of Po River water quality in Torino (Italy). Water Resour Manag 24:2937–2958

Park J, Park C (2009) Robust estimation of the Hurst parameter and selection of an onset scaling. Stat Sinica 19(4):1531–1555

Parmar KS, Chugh P, Minhas P, Sahota HS (2009) Alarming pollution levels in rivers of Punjab. Indian J Env Prot 29(11):953–959

Rangarajan G (1997) A climate predictability index and its applications. Geophys Res Lett 24(10):1239–1242

Rangarajan G, Ding M (2000) Integrated approach to the assessment of long range correlation in time series data. Phys Rev E 61(5):4991–5001

Rangarajan G, Sant DA (2004) Fractal dimensional analysis of Indian climatic dynamics. Chaos Solitons Fractals 19(2):285–291

Shukla JB, Misra AK, Chandra P (2008) Mathematical modeling and analysis of the depletion of dissolved oxygen in eutrophied water bodies affected by organic pollutant. Non-linear Anal: Real World Appl 9(5):1851–1865

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

Su S, Li D, Zhang Q, Xiao R, Huang F, Wu J (2011) Temporal trend and source apportionment of water pollution in different functional zones of Qiantang River. China Water Res 45(4):1781–1795

Toprak ZF (2009) Flow discharge modeling in open canals using a new fuzzy modeling technique (SMRGT). CLEAN-Soil Air Water 37(9):742–752

Toprak ZF, Eris E, Agiralioglu N, Cigizoglu HK, Yilmaz L, Aksoy H, Coskun G, Andic G, Alganci U (2009) Modeling monthly mean flow in a poorly gauged basin by fuzzy logic, *CLEAN-Soil*, *Air*. Water 37(7):555–564

Vassilis Z, Antonopoulos M, Mitsiou AK (2001) Statistical and trend analysis of water quality and quantity data for the Strymon River in Greece. Hydrology Earth Syst Sci 5(4):679–691

Weng YC, Chang NB, Lee TY (2008) Nonlinear time series analysis of ground-level ozone dynamics in Southern Taiwan. J Environ Manag 87:405–414

WHO (1971) International standards for drinking water. World Health Organization, Geneva

Yarar A (2014) A hybrid wavelet and neuro-fuzzy model for forecasting the monthly streamflow data. Water Resour Manag 28:553–565