Water Practice & Technology

IWA PUBLISHING

Water Practice & Technology Vol 18 No 1, 185 doi: 10.2166/wpt.2023.002

Time series trend modelling and forecasting of selected water quality parameters in the Mthatha River Catchment, South Africa

Oseni Taiwo Amoo [a,*], Abdultaofeek Abayomi [b], Akinola Ikudayisi [c] and Nombuyiselo Makupula[a]

[a] Risk and Vulnerable Science Centre, Walter Sisulu University, Mthatha, Eastern Cape, South Africa.
[b] Department of Information and Communication Technology, Mangosuthu University of Technology, Durban, South Africa
[c] Civil Engineering Department, Walter Sisulu University, Buffalo City Campus, East London, Eastern Cape, South Africa
*Corresponding author. E-mail: ejire36@gmail.com

OTA, 0000-0003-1713-3814; AA, 0000-0003-3129-5246; AI, 0000-0002-7992-8789

ABSTRACT

Over recent decades, water quality at the Mthatha River Catchment (MRC) within the Eastern Cape Province of South Africa has been threatened by various pollutants. The continuous effluent concentration discharges from the Mthatha Prison and the Efata School for the Blind and Deaf have caused ineffable damage to the Mthatha River's water quality. Thus, the time series-measured data between 2012 and 2020 were analysed to determine the trends and enable forecasting of selected water quality parameters using the Thomas–Fiering (T–F) stochastic model. The Kendall's τ test trends show an increase in the coefficient of variation of 0.498 and 0.394 at the Mthatha Prison and Efata School, respectively, for abrupt changes, whereas the mean monthly T–F forecasted model shows a good correlation value range from 0.79 to 0.87 for the various predicted variables. The simulated predicted models and trends could serve as a measure to forecast selected water quality parameters' occurrence and a likelihood period when the river pollutants could be controlled. Water managers and researchers would find usefulness in the employed tools for an effective control planning of the river pollutants.

Key words: Mann–Kendall, pollutants, river chemistry, stochastic model, water quality

HIGHLIGHTS

- Innovative ways for analysing and forecasting water quality (WQ) parameters.
- Prerequisite for prompt mitigation measures in maintaining a healthy river state.
- Contribution to the statistical approach, visual modelling, and predictive algorithms.

INTRODUCTION

The present study aims to assess the variability time series trend and forecasting of selected water quality (WQ) parameters in the Mthatha River Catchment (MRC) area, Eastern Cape of South Africa. The rampant densification of the Mthatha town going by its growing population, in Eastern Cape, South Africa, has been challenged by the Department of Water and Sanitation (DWS) to address the pollutant issues that are coming from effluent concentration discharges from the town and more particularly from the Mthatha Prison and the Efata School for the Blind and Deaf, which have caused ineffable impaired damage to the MRC. Thus, to minimise the hazardous health impacts and aesthetically offensive odour which could pose danger to the lives and health of the community if not properly managed and disposed of has necessitated the need to provide vital information for monitoring and regulating the Mthatha River water pollutants. Hence, this study tends to answer the following research questions: what factors are responsible for the vulnerability of the Mthatha Prison Waste Stabilisation Pond (WSP) in meeting the growing population treatment demand, what parameters are responsible for determining the temporal trend pattern for the selected WQ parameters, and how to explore the use of the stochastic T–F model for forecasting WQ parameters. The spatiotemporal dimension of the available data and the need for a precautionary high-end sensitivity for analytical data denote the novelty in exploring T–F Morkov-chain characteristics in projecting available data.

There are currently two types of models for modelling and predicting WQ, which include mechanism-oriented and non-mechanism-oriented models (Aldhyani *et al.* 2020). The mechanism model is more advanced; it uses advanced system structured data to simulate WQ and is thus a multifunctional model that can be used for any water body, whereas the non-mechanical-oriented model demonstrates the application of geostatistical methodologies in visual modelling and multivariate statistical interpolation approaches (Yang *et al.* 2018). However, most of the previous research studies for the prediction and modelling of WQ parameters had been based on statistical approaches, varying analysing algorithms, and predictive algorithms in large-scale-based computation needs (Ahmed *et al.* 2019; Manzione & Castrignanò 2019; Aldhyani *et al.* 2020). Most of these models' limitation lies in the amount of available data in implementing the desired techniques and approaches (Podvezko & Sivilevičius 2013). The study's contribution to knowledge lies in its applicability for short-term prediction and simplicity in monitoring and forecasting long-term WQ parameters useful for detecting likely future trend for various WQ variables.

Similarly, modern researchers' utilisation of internet of things (IoT) and big data have made predictions of WQ relying on new technology applications such as fuzzy logic, stochastic, artificial intelligence (AI)/machine learning (ML) systems, deep learning, or a combination of these methods depending on the problem statement had made it good to real-time situations analysis and prediction (Ahmed *et al.* 2019; Bui *et al.* 2020; Rao *et al.* 2022). These methods had made it possible to identify prospective factors that have an impact on water environment systems and are useful tools for water resource management but required special training, and expert knowledge for usage and in a complex situation to understand.

Frollini *et al.* (2021) show an excellent overview of the statistical methods for trend estimation and detection of monitored data in environmental time series for WQ and atmospheric deposition. Kamilaris & Ostermann (2018) evaluated six methods for trend detection in real-life data for hydrologic time series analysis while Chen *et al.* (2018) proposed multivariate statistical techniques such as cluster analysis (CA), principal component analysis (PCA), and factor analysis (FA) to identify and classify the main pollutants in a river system. In another study, Tripathi & Singal (2019) applied the PCA to aid the interpretation of complicated data matrices while Nath *et al.* (2018) used it to gain a better understanding of the WQ and ecological status in a river system. The partial least square (PLS) approach is recommended since it solves multi-variable issues such as the negative effects of multiplexing between variables in WQ prediction (Du Plessis *et al.* 2014). The PLS model is made up of a structural component that consists of a latent variable relationship and a measurement component that not only indicates how they are related but also reveals the influence of weighted weight that may be used to estimate uncertainty in latent variable case values (Du Plessis 2015). These are some of the previous related studies carried out globally, regionally, and specific to areas in and around the study area.

Because of intra-annual variation, outliers, and non-detected data, WQ data are frequently not normally distributed, and nonparametric statistical analyses were often used (Bracmort *et al.* 2006; Rao *et al.* 2022). The use of the nonparametric Mann–Kendall test for detecting monotonic trends in WQ modelling is reported by Rao *et al.* (2022), Kumar & Rathnam (2019), and Wang *et al.* (2019). The Mann–Kendall τ value has been used to measure the correlation between the observed parameters and the period. A trend exists in a dataset if there is a significant correlation (positive or negative) between the observations and time. Mann–Kendall compares the relative magnitudes of sample data rather than the data values (Pohlert 2016), whereas Sen's slope estimator is used to estimate the true slope of an existing trend such as change per year, where the trend can be assumed to be linear (Sen 1968). Sen's and Mann–Kendall sensor trend detection are used for trend analysis in forecasting catchment hydrologic flow regimes (Govindarajulu 1992). Hamed & Rao (1998) advocate that the series under testing should be normally distributed. Trend or non-stationary in the dataset is normally introduced through human activities such as personnel, technical, and/or nature-induced climate change.

The future behaviour of WQ variables depends on a lot of complex chemical interactions and the nature of the catchment response in space and time which induces complexity that is yet to be predicted in hydrology (Sathish & Babu 2017). Most of these complexities can be represented through a simple process of systematic model representation of the catchment indices (Kang & Lin 2007; Freni *et al.* 2011). Among the many stochastic model's representations of the catchment is the assumption that the process follows a Markov process $X(t)$ ordered in a discrete-time variable 't' ranging from 1, 2, 3,…, n. This implies that the probability of moving WQ parameters to the next state depends only on the current state and not on the previous states. Markov chains are a convenient model for describing many phenomena and are often used in synthetic flow and rainfall generation and optimisation models (Loucks & Van Beek 2005). Alfa *et al.* (2018) applied the modified Thomas–Fiering (T–F)'s

autoregressive Markov model to generate synthetic stream flows by extending the existing 19 years' stream flow data at the Oforachi Bridge hydrometric station in Ofu, Kogi State, Nigeria. The predicted flows obtained when subjected to statistical parameters showed that the historic stream flow data were preserved in the synthetic stream flow generated by the model. Similarly, Maroof *et al.* (2015) employed the T–F model to extend the 12 months discharge data at Ero-Omola falls as a basis to study its hydropower development potential. Celeste *et al.* (2004) also utilised T–F stochastic model for synthetic stream flow generation to determine monthly inflow scenarios for the watershed of the reservoir that supplies the city of Matsuyama, Ehime Prefecture.

Conversely, Brunner *et al.* (2019) postulate that a long-trend forecast with the stochastic model is safer as the prediction intervals allow for greater uncertainty growth in the future. The trend in a time series can be expressed by a suitable linear or nonlinear model; the linear model is widely used in hydrology (Kurunç *et al.* 2005). In addition, most statistical empirical-based model and data-driven model offers simplified nonlinear forms of governing equations. Their ability to simulate feedback inside the physical system that may result from management actions is added advantage of the model (Condon & Maxwell 2013; Devagopal *et al.* 2022). Considering the different WQ inputs chemistry, the future prediction can be modelled using different parsimony methods. Thus, the current study assesses the presence of trends, using a nonparametric (Kendall τ test) to model the process while future prediction for the selected WQ parameters was simulated using T–F regression analysis to serve as a likelihood procedural step in time control for managing the river pollutants.

The rest of the paper is structured as follows: Section 2 illustrates the materials and methods employed for data sampling design, study area description, pre-model analysis, and how the T–F model usage was configured. The model used to extend the historical data of stream flow and weather data in this study is the Markov chain. Section 3 states the data analysis, with the simulated temporal pattern for the selected WQ parameters on a monthly time scale in a catchment. Section 4 connotes the results and discussion for T–F stochastic forecast analysis before summarising the key results finding followed by a conclusion and recommendation.

MATERIALS AND METHODS

Data sampling design and study area description

Aside from the overall cost implication which limit the monitoring agency (DWS) for testing the selected WQ parameters (pH, phosphate, ammonia, conductivity, faecals, and *Escherichia coli*). These parameters had been proven to be a good representation of WQ assessment (Bracmort *et al.* 2006; Faruk 2010; Devagopal *et al.* 2022). This historical repository data dated 2012 to 2020 have been collected at 10 sampling sites due to accessibility. Figure 1 depicts the study area with sampling sites.

The selection of these sampling sites was developed in a phased manner to represent a fair sample representation for the catchment. Prior information on the river's effluent discharge point and a preliminary survey carried out by DWS staff were done to identify the general condition of the environment as well as the best possible sampling point to map the sources of pollutants that could impact the water resources negatively. The collected wastewater samples were made thrice and stored in a cold collection cooler before the collected samples were transported to Talbot and Talbot Laboratory, where they were analysed using established procedures (DWAF 1996; APHA 2017). Each month samples were averaged for the recorded monthly value. A total of 1,728 sample sizes were analysed for the catchment area. The year 2018 dataset was not used due to a lot of missing data aside its corrupt storage compartment for the year.

Pre-model analysis

In achieving the study objective, the collected dataset was subjected to pre-data analysis in correcting the localised errors (standardisation), reliability (outlier correction), and homogeneity (SNHT) to detect if the series is homogeneous over time, or if there is a time at which a change occurs over the long period of recorded parameters of the WQ variables. The consistency in Sen's slope and trend pattern detection was determined by the nonparametric test (Mann–Kendall). This shows that pre-data analyses help to minimise the accumulated errors/bias in getting a good model. Thereafter, the average trend was fitted for comparison, using the mean, standard deviation, coefficient variation, correlation, and bias statistics tools were then employed to measure the goodness-of-fit, before the stochastic T–F model was subsequently used to forecast future likelihood prediction for the observed WQ parameters towards adaptative control measures.

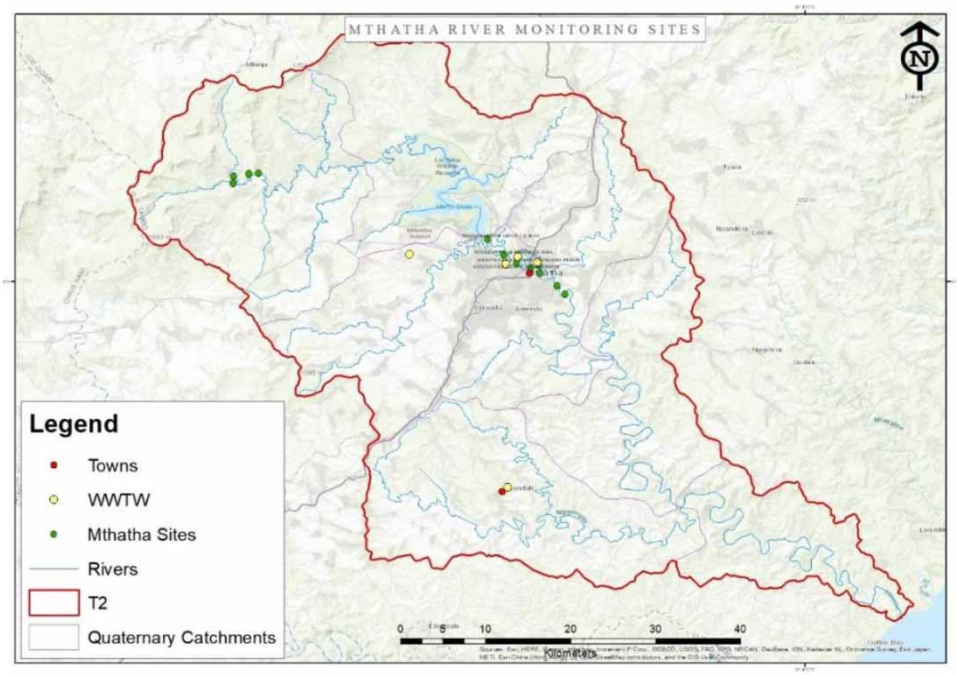

Figure 1 | Map of the study area sampling sites.

The T–F model configuration

The T–F model as a typical stochastic model has been improved through pre-defining-an array previously supposed to be a random-to-random skewed variate prediction of selected WQ constituents and streamflow (Kurunç *et al.* 2005; Cui *et al.* 2016; Sathish & Babu 2017). The Gauss–Markov theorem form the basis assumption in the application of the model. The model was premised to preserve that (1) the datasets are normal distribution in their annual flows; (2) log-normal distribution of their monthly flows; (3) correlation exists between the annual flows; and (4) correlation also exists between the monthly flows. Hence, the model uses the best linear unbiased estimator, which allows the tightest possible sampling distribution of unbiased estimates when compared to other linear estimation methods. Thus, the T–F model uses the Markov process to simulate the monthly WQ variable from historical average monthly measured data (q_i), which are normally distributed with the first-order autoregressive model (Cui *et al.* 2016). According to Thomas & Fiering (2013), the monthly streamflow and WQ data frequently do not appear to be normally distributed. Thus, the annual and monthly streamflow and WQ data are transformed into normal and log-normal distribution, respectively. Also, the model account for the effect of seasonality on the variability of the data by considering month-to-month variation in the average value. Assuming, that there are N years of data available, the calculation of the terms in the T–F model for each month $j = 1, 2, 3, \ldots, 12$ accordingly are as presented in Equations (1)–(6).

$$q_{i+1} = \bar{q}_{j+1} + b_{j,j+1}(q_i - \bar{q}_j) + Z_{i+1}S_{j+1}(1 - r_{j,j+1}^2)^{1/2} \quad i = 1 \text{ to } n \tag{1}$$

$$y_{i+1} = \mu_{i+1} + b_{j,j+1}(y_i - \mu_i) + k_{i+1}\,\sigma_{j+1}\left(\sqrt{(1 - r_{j,j+1}^2)}\right) \quad i = 1 \text{ to } n \tag{2}$$

where n is the period of prediction in month

$$y_i = \ln q_i$$

where q_{i+1} , q_i are the generated flows during $(i+1)$th and ith seasons from the beginning of the synthesised sequences, $\overline{q}_{j+1}$, $\overline{q}_j$ are the mean flows during $(j+1)$th and jth seasons within a repetitive annual cycle of seasons (for monthly period, $1 \leq j \leq 12$), μ_i is the log-transformed of the mean annual historical inflows, y_i , y_{i+1} are the generated log-normal inflows in the month i and $i+1$th, respectively, $b_{j,j+1}$ are the least squares regression

coefficients for estimating $(j+1)$th flow from the jth flow, but $b_{j,j+1}$ can be computed with Equation (3).

$$b_{j,j+1} = r_{j,j+1}\left(\frac{S_{j+1}}{S_j}\right) \tag{3}$$

where k_{i+1} is the normally distributed random number with zero mean and unit variance, S_{j+1}, S_j are the standard deviations of flows during the $(j+1)$th and jth seasons, σ_{j+1} is the standard deviation of annual inflows of the log transform, $r_{j,j+1}$ are the correlation coefficients between flows in jth and $(j+1)$th seasons, Z_{i+1} is the normal random number with zero and variance unity = Normsinv (rand ()), n is the period of prediction in month

If there is N years of available data, the calculation of the terms in the T–F model for each month, $j = 1, 2, 3,\ldots,$ 12 according to McMahon & Mein (1978) are presented in Equations (3)–(6).

- *Mean flow* $(\overline{q}_i)$

$$\overline{q}_i = \sum_{i=1}^{N}\left(\frac{q_{j,i}}{N}\right) \quad (i = j,\ 12+j,\ 24+j,\ \ldots) \tag{4}$$

- *Standard deviation of flow* (S_j)

$$S_j = \sqrt{\frac{\sum_{i=1}^{N}(q_{j,i} - \overline{q}_j)^2}{N-1}} \tag{5}$$

- *Correlation coefficient with the flow in the preceding month* $(r_{j,j+1})$

$$r_{j,j+1} = \frac{\sum_{i=1}^{N}(q_{j,i} - \overline{q}_i)(q_{j+1,i} - \overline{q}_{j+1})}{\sqrt{\left\{\sum_{i=1}^{N}(q_{j,i} - \overline{q}_j)^2 \sum_{i=1}^{N}(q_{j+1,1} - \overline{q}_{j+1})^2\right\}}} \tag{6}$$

Data analysis

Table 1 presents the statistical summary for the selected measured WQ parameters.

Table 1 | Statistical summary for the studied yearly water quality parameters (2012–2020)

Statistic	pH (Unit)	Phosphate (mg/L)	Ammonia (mg/L)	Conductivity (micromhos/m-mS/m)	Faecals (counts/mL)	*E. coli* (MPN/100 mL)
Minimum	6.600	0.002	0.022	4.000	0.000	0.000
Maximum	8.500	16.000	27.600	390.000	83,000.000	48,000.000
First quartile	7.200	0.020	0.110	9.000	63.500	35.000
Median	7.400	0.048	0.220	13.000	360.000	200.000
Third quartile	7.600	0.171	0.800	21.000	2,000.000	800.000
Mean	7.397	0.303	0.857	18.695	1,490.796	868.671
Variance $(n-1)$	0.092	1.883	5.361	765.961	28,347,319.	9,481,174.800
Standard deviation	0.303	1.372	2.315	27.676	5,324.220	3,079.152
Variation coefficient	0.041	4.535	2.702	1.480	3.571	3.545
Skewness (Pearson)	0.270	7.856	9.002	10.489	12.671	13.029
Kurtosis (Pearson)	0.649	66.955	95.717	134.695	178.787	187.210

The skewness and kurtosis values from Table 1 were used to depict the normal distribution check. Skewness measures the degree of asymmetry, the closer the value is to zero, the more skewed the distribution. On the contrary, kurtosis measures the level of sharpness or flatness of a frequency distribution curve. The range between the

minimum and maximum value was used as a validity check and the reliability of such data while the removal of outliers was done by plotting the dataset into boxplots to eliminate all the data points that are outside the limits of the boxplots. Bias occurs if the residual error plot still contains some information that should be accounted for in the model in order to obtain better forecasts. To correct this, the adequacy of fit was judged by the insignificant correlation and the normal distribution of obtained residuals error plot. The autocorrelation function (ACF) plot of the residuals with a large *p*-value greater than 0.05 indicates there is insignificant autocorrelation remaining in the residuals.

RESULTS AND DISCUSSION

Figures 2 and 3 depict the temporal trend pattern observed for the Prison and Efata School discharge locations. A trend exists in a dataset if there is a significant correlation (positive or negative) between the observations and

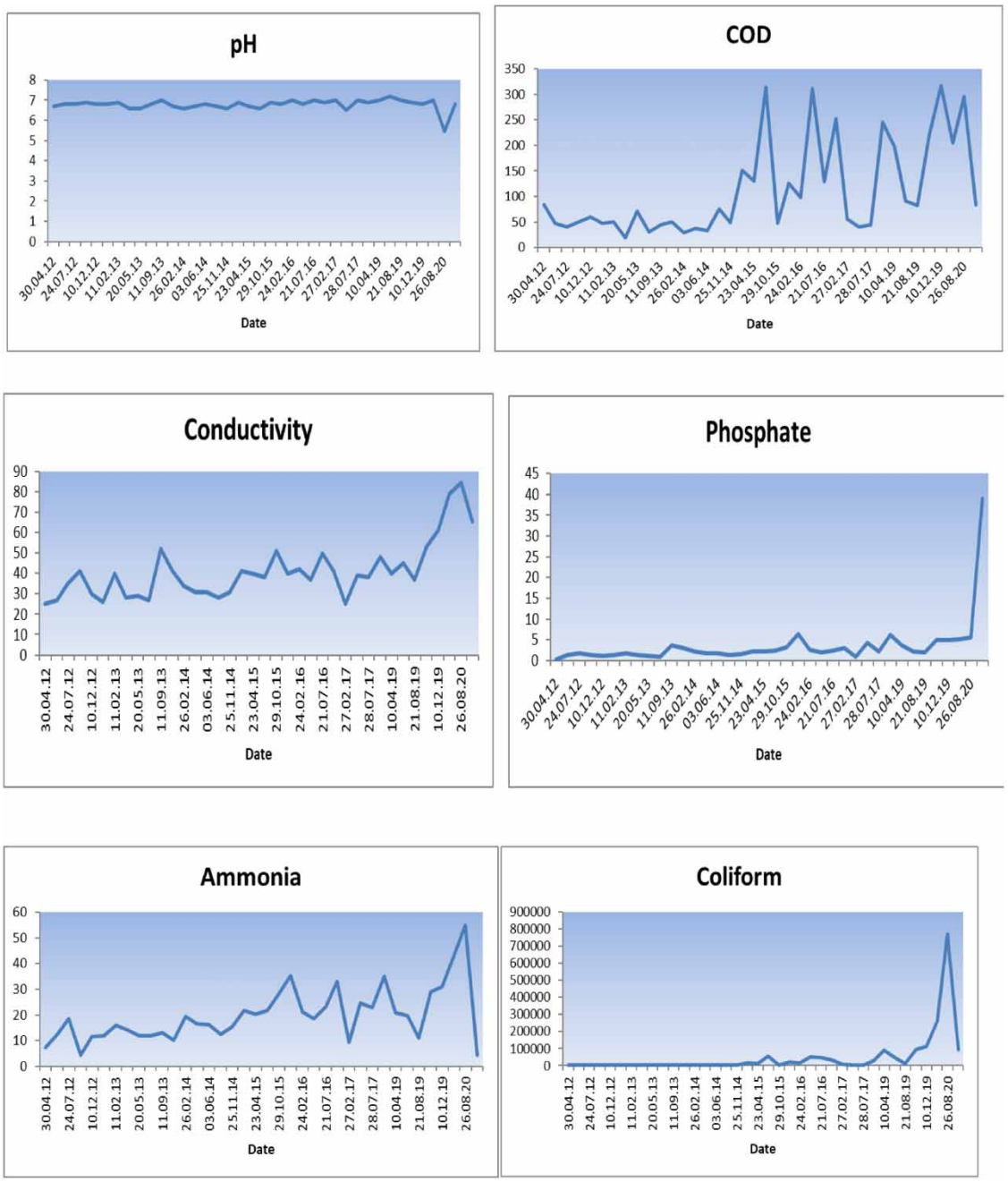

Figure 2 | Mthatha Prison effluent distribution patterns for the study area's water quality parameters.

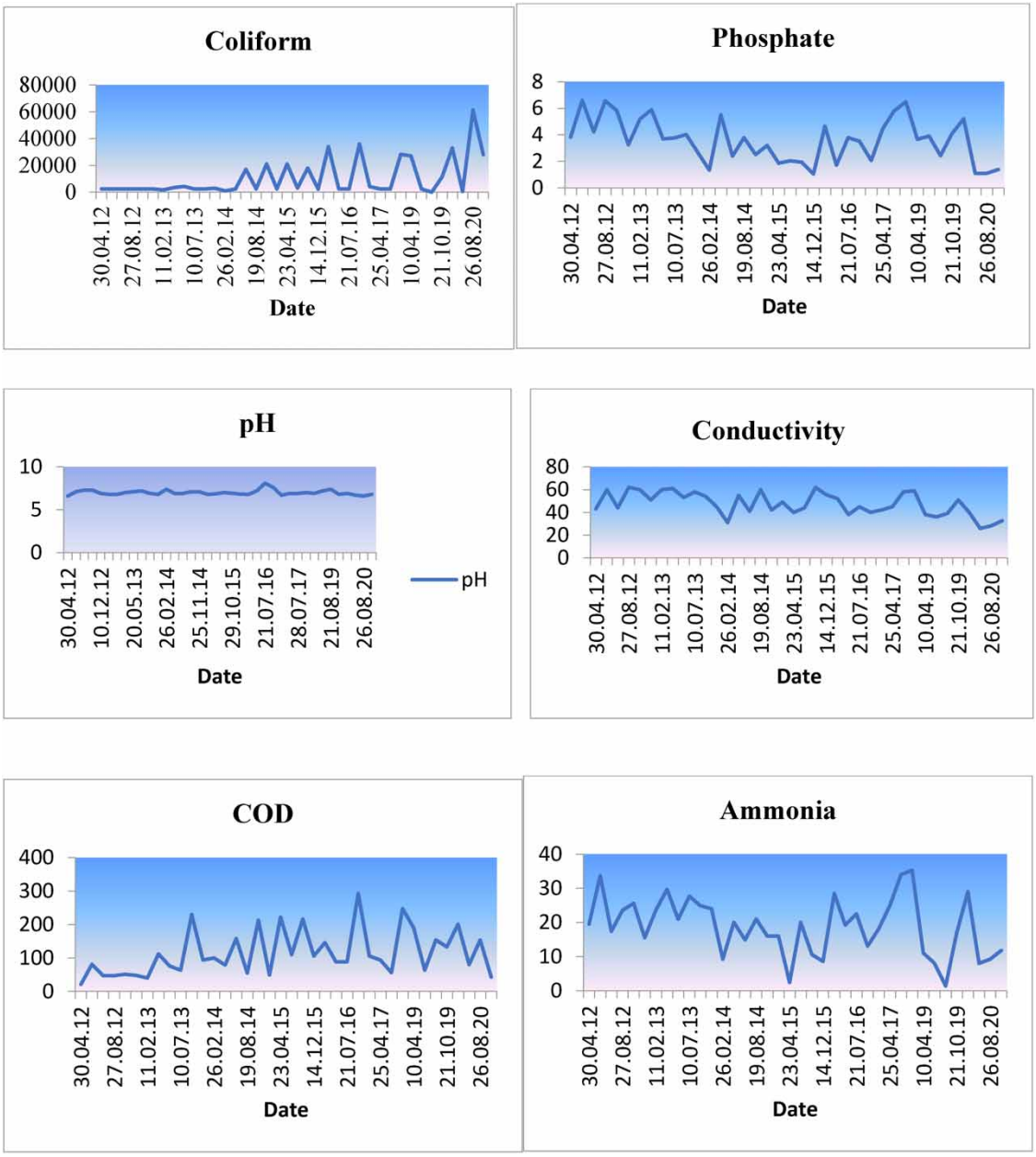

Figure 3 | Efata effluent temporal distribution of water quality parameters from 2012 to 2020.

time. The figure has been divided into six sub-figures labelled as (a) pH, (b) COD, (c) conductivity, (d) phosphate, (e) ammonia, and (f) coliform to represent each of the WQ parameters selected and measured over the period 2012–2020.

The chemical oxygen demand (COD), the amount of dissolved oxygen that must be present in water to oxidise chemical organic materials with conductivity, ammonia, and faecal coliforms showed high peaks, which fluctuated from November to January while there was a gradual decrease in the value of effluent concentration between May and October, as depicted in Figure 4, followed by a gradual increase in faecal coliforms, and *E. coli* representing a group of bacteria found in the water was also very high. These are indicators of poorly treated sewer discharged by the WSPs into the river course. Thus, resulting in the Mthatha River downstream having thick layers of water hyacinth and algae growth.

Figure 3 depicts the Efata School effluent temporal pattern distribution. The recorded values of phosphates were highest at 6 mg/L and occurs in December and April. This may exert more hydraulic pressure onto the river and their presence in the water led to algae growth and eutrophication.

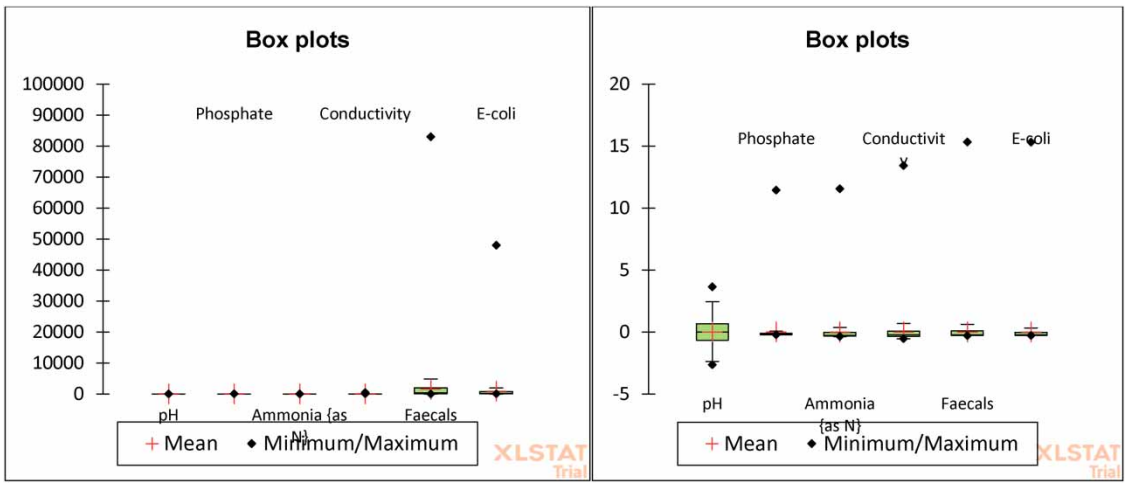

Figure 4 | Normal and standardised representation of water quality parameters' data (2012–2020).

The monthly temporal distribution pattern for coliform, pH, conductivity, COD, and ammonia is shown in Figure 3. In all, there was a decline in peak value during April month (30.04.12) for the year 2012 and the month of December (15.12.2020) for all the years under consideration. This period marks when schools are closed for the holiday and hence, the ponds witness a low varied measured WQ parameters concentration with their time series.

A cursory check on the monthly plotted WQ parameters (coliform, pH, conductivity, COD, and ammonia) dataset as represented in Figure 4, depicts the presence of outliers in all the measured WQ parameters. This implies that the dataset needs to be standardised to show a fair data sample before being used. The outlier analysis was necessary to correct the range weighted average discrepancy noted in the observation, which may be due to the variability in the measurement or may indicate an experimental error in the dataset collection. Also, because of unit disparity and different variances, standardisation of various WQ parameters was done to normalise the datasets for effective statistical analyses and minimisation of bias and accumulation of predicted error from the data seen.

The process of standardisation comprises removing the mean data from the original data and dividing it by the standard deviation (Ikudayisi & Adeyemo 2016). This converts each variable in the dataset into a new variable with a unit standard deviation and a zero-mean value. Figure 4 depicts the normal and standardised version of the WQ parameters.

Since the observed datasets violate the normal distributed test, the nonparametric test such as the Mann–Kendall test (Mann 1945; Pohlert 2016) was applied to assess the statistical significance of trends. Table 2 depicts the results of the Sen's slope and Mann–Kendall test in detecting abrupt change period and variability trend changes among the selected measured WQ parameters at the Mthatha Prison sampling sites and the Efata point samples. Note that Kendall's τ value of zero (0) suggests that there is no trend, while a p-value that was less than the

Table 2 | Mann–Kendall's analysis of water quality parameters at Mthatha River (above and below the prison) sampling sites

Series\Test	Kendall's τ	p-value	Sen's slope
pH	−0.085	0.367	0.000
Conductivity	0.203	**0.029**	0.067
Phosphate	0.180	0.052	0.001
Ammonia	0.182	0.051	0.001
Faecals	0.204	**0.024**	4.842
E. coli	0.077	0.394	0.750

The bolded values correspond to p values that were less than the significance value of 0.05.

significance level of 0.05, means there was a date at which changes were witnessed in the dataset. Thus, a higher *p*-value than the level of significance (0.05) depicts an increasing trend and vice-versa.

From Table 2, Kendall's τ value for all the parameters except the pH is positive suggesting an upward trend, while the negative value for pH suggests a downward trend. The *p*-value for pH and *E. coli* is greater than the significance level of 0.05 implying that the initial assumption that there is no trend was false. Conversely, the conductivity, faecals, and *E. coli* have a *p*-value that was less than the significance level of 0.05, which means there was a date at which changes were witnessed in the dataset. Table 3 depicts the Mann–Kendall analysis for the Mthatha River above and below the Cicira River.

Table 3 | Mann–Kendall's analysis of water quality parameters at the Cicira River (a tributary of the Mthatha River) above and below the Efata School

Series\Test	Kendall's τ	*p*-value	Sen's slope
pH	−0.064	0.499	0
Conductivity	−0.122	0.194	0.011
Phosphate	0.289	**0.002**	0.001
Ammonia	0.191	**0.043**	0.001
Faecals	0.061	0.498	0.556
E. coli	−0.103	0.257	−0.385

The bolded values correspond to *p* values that were less than the significance value of 0.05.

Table 3 shows pH, conductivity, and *E. coli* having negative Kendall's τ values. This indicates a downward slope while the phosphates, ammonia, and faecal values show positive Kendall's τ values, to illustrate an upward slope. The *p*-value signpost significant level of either increasing or decreasing level over time. In all, an increasing trend was witnessed in pH, faecals, and *E. coli* with a downslope for conductivity, phosphate, and ammonia as observed in their *p*-value as shown in Table 3. This indicates irrational behaviour of WQ variables that are not constant over the river waters.

T–F stochastic forecast analysis

Using the T–F model which was limited to log-normally distributed for the WQ variables (k_i) in Equation (5) and the only unknown variable in the model at each step, the pseudo-random normal number is estimated. The k_i values are calculated using the Microsoft Excel functions using the command RAND () as a source of randomness, which is equivalent to $F_z(z)$. The value of $F_z(z)$ estimated by the RAND function is used in the Normsinv function as Normsinv $F_z(z)$ to generate a standard log-normal random number k_i which is the normal part of the stochastic model. The k_i can also be generated directly using NORMSINV (RAND ()) function. The values of k_i generated are multiplied with the random part of the stochastic variable to simulate each monthly WQ variable. Other T–F forecasted model statistics requirements including their means, standard deviation, regression coefficient, and lag one serial correlation of the WQ variable are also essential, and these can be obtained from statistical analysis of the observed monthly collected historical WQ data.

The T–F forecasted model need to solve the problem of identification of key WQ attributes that favour the catchment's historical record characterisation. Thus, the mean monthly logarithmic transformation of the observed historical WQ was used for the simulating future monthly forecasts value for the catchment. The goodness-of-fit test using mean, standard deviation, coefficient of the variate, correlation, and bias value measure had been useful to calibrate and validate the observed measured value in comparison to the simulated forecasted value. Table 4 depicts the arithmetic evaluation for the predicted pH WQ parameter, which was repeated for other variables while Table 5 shows the performance evaluation of the observed and predicted pH (unit) WQ sample.

Each monthly observed and predicted WQ parameter for the pH has been transformed and expressed in a percentage. Table 5 shows the performance evaluation for the monthly observed and predicted pH WQ parameter. Table 5 gives a correlation comparison of 0.79 between the observed and predicted pH WQ parameters which implies that the Markov chain model was able to predict the future pH. The statistical summary shows an average value of 76.82% simulation to 74.24% of the observed value, while both values of standard deviation and coefficient of variation depict viable proximity exist between the observed pH and simulated pH value with a strong –0.03 bias occurrence for the residual values. Similarly, this process was repeated for other WQ parameters

Table 4 | Statistical parameters of natural logarithmic transform of WQ pH variable (2012–2020)

Month	Mean (μ)	Std (σ)	Lag 1 Serial corrected (ri)	Regression coefficient (bj)	$\sigma \times$ SQRT(1–r2)
Jan	5.444	0.078	0.385	0.486	0.072
Feb	4.822	0.098	0.277	0.307	0.095
Mar	3.835	0.109	0.576	0.332	0.089
Apr	3.843	0.063	0.099	0.119	0.063
May	4.245	0.076	0.426	3.723	0.069
June	5.853	0.662	−0.409	−0.173	0.604
July	6.076	0.279	−0.107	−0.059	0.278
August	6.497	0.155	−0.489	−0.268	0.135
September	6.983	0.085	0.164	0.738	0.084
October	6.981	0.383	0.324	0.083	0.362
November	6.895	0.098	−0.886	−2.528	0.045
December	6.031	0.279	−0.607	−0.169	0.222

Table 5 | Performance evaluation of observed and predicted pH water quality parameter

Month	Mean predicted (%)	Mean observed (%)
Jan	74.48	63.78
Feb	73.82	66.00
Mar	71.53	66.33
Apr	70.09	71.22
May	73.97	74.44
June	79.10	80.11
July	80.64	81.56
Aug	80.47	80.33
Sep	80.54	82.11
Oct	80.10	79.22
Nov	78.89	74.56
Dec	78.22	71.22
Average	76.82	74.24
STD	3.81	6.54
Coefficient variation	20.15	11.36
Correlation	0.79	
Bias	−0.03	

including phosphate (mg/L), ammonia (mg/L), conductivity, faecals (counts/mL), and *E. coli* (counts/100 mL) forecasting. Table 6 presents the statistical summary for all the measured WQ parameters.

In summary, Table 6 shows the goodness-of-fit test comparison based on the monthly mean square error for the predictions (mean, standard deviation, coefficient of the variate, correlation, and bias value measure). Table 6 depicts the standard deviation of 3.81 for pH, 2.73 for conductivity, 1.30 for phosphorus, and 2.08 for ammonia (NH_4) parameter value. These values indicate that the various parameter prediction is of less variance and least bias error as the minimum residual estimator lies between –0.01 to +0.06 depicting a satisfactory prediction value. There is insignificant autocorrelation remaining in the residuals going by the negative bias witness −0.03 for pH. This also indicates a small residual bias value depicting a satisfactory prediction value.

Also, there was a close similarity in trend pattern with a performance average range of 20.70–33.44 for central normality tendency while the variability measure (standard deviation) range was 1.35–2.73. This defines the distance between the observations and the average. The coefficient of variance range shows 12.35°–16.84°, with a fair correlation range of 0.57–0.96 strong relation between the mean predicted to mean observed value. This

Table 6 | Performance evaluation of the various water quality parameters between 2012 and 2020

Variable	pH	Conductivity	Phosphorus	Ammonia	Faecals	*E. coli*
Average	76.82	33.75	22.49	33.44	22.41	32.92
STD	3.81	2.73	1.30	2.08	1.35	1.96
Coefficient variation	20.15	12.35	17.32	16.08	16.63	16.84
Correlation	0.79	0.88	0.74	0.95	0.67	0.96
Bias	−0.03	−0.01	0.00	−0.01	0.01	−0.01

implies that the modelling procedure, applied over decades of the dataset was appropriate and a good forecasting tool to measure the varied intra and inter-months nomenclature of wet to normal and dry months. These results were also collaborated by Du Plessis (2015); Du Plessis *et al.* (2014); and Faruk (2010) studies.

The plotted comparison of the mean monthly observed value to the predicted water simulated values for the MRC is depicted in Figures 5–10.

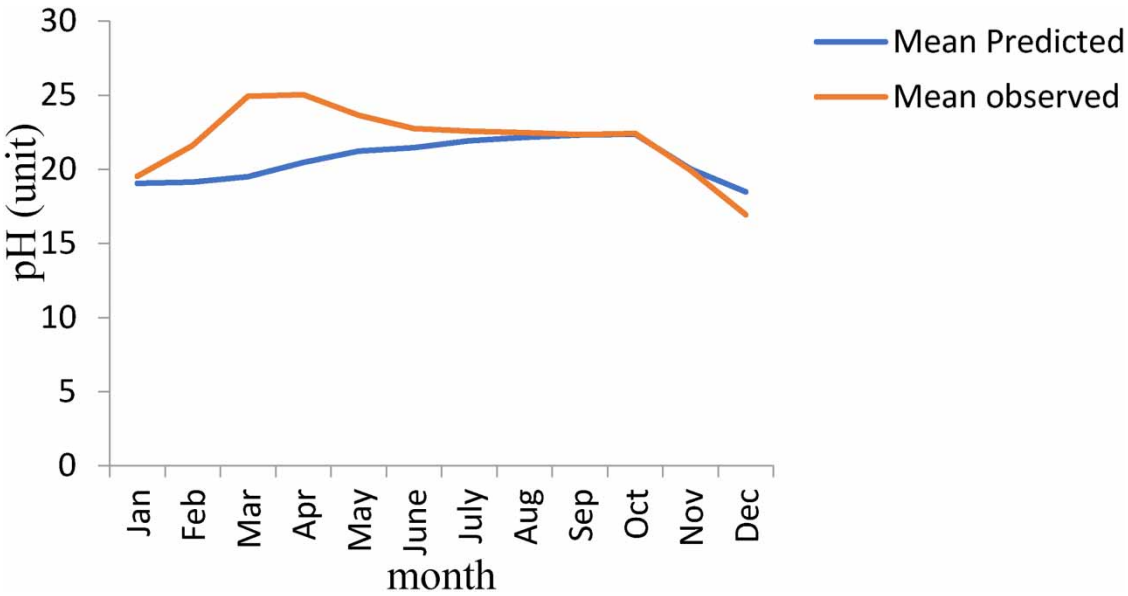

Figure 5 | Mean monthly observed and predicted pH parameters distribution for MRC (2012–2020).

The pH plot in Figure 5 follows a similar pattern with a sharp increase between May and July for observed years while August to November constitute the sharp varied trend for phosphorous and ammonia as shown in Figures 6 and 7. Only the conductivity WQ parameter exhibits a similar normal trend pattern as shown in Figure 8 while erratic/irrational behaviour was witnessed in *E. coli* pattern as depicted in Figures 9 and 10, this may be attributed to its biological modification reaction. In all, the plot of the conductivity, phosphorus, ammonia, faecals, and *E. coli* mean monthly observed value to the predicted simulated values for the MRC shows various discrepancies in their performance coefficient. The forecasted value from the model depicts the future likelihood of the river WQ thereby providing a proactive mechanism to maintain a healthy state assessment of the river and a timely schedule for monitoring plans.

Implication of results

The preliminary results for the reliability, stationary, consistency/homogeneity, and outlier test on the WQ parameters data covering 2012–2020, suggested that necessary preliminary analysis and precautions should be concluded on the available dataset before being used to minimise bias and accuracy of the prediction.

Although the results of the mean monthly logarithmic transformation statistics of observed historical data at the two stations – Mthatha prison and Efafa, indicate their predicted time series (Markov chain model) were

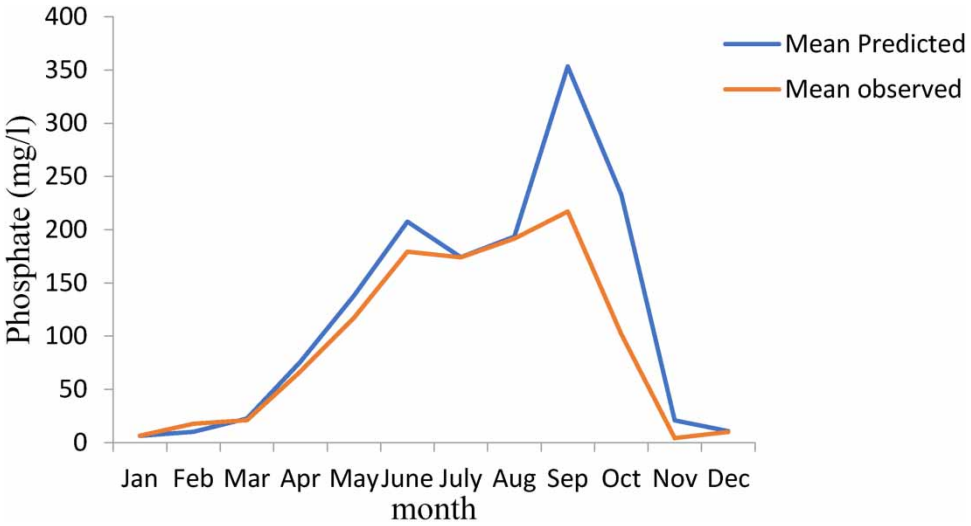

Figure 6 | Mean monthly observed and predicted phosphate parameter for MRC (period).

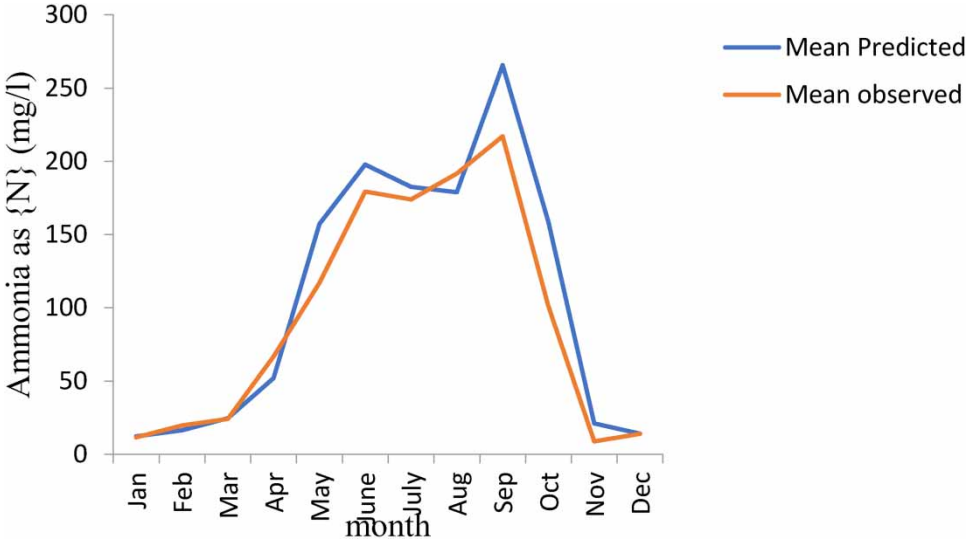

Figure 7 | Mean monthly observed and predicted ammonia parameter for MRC (period).

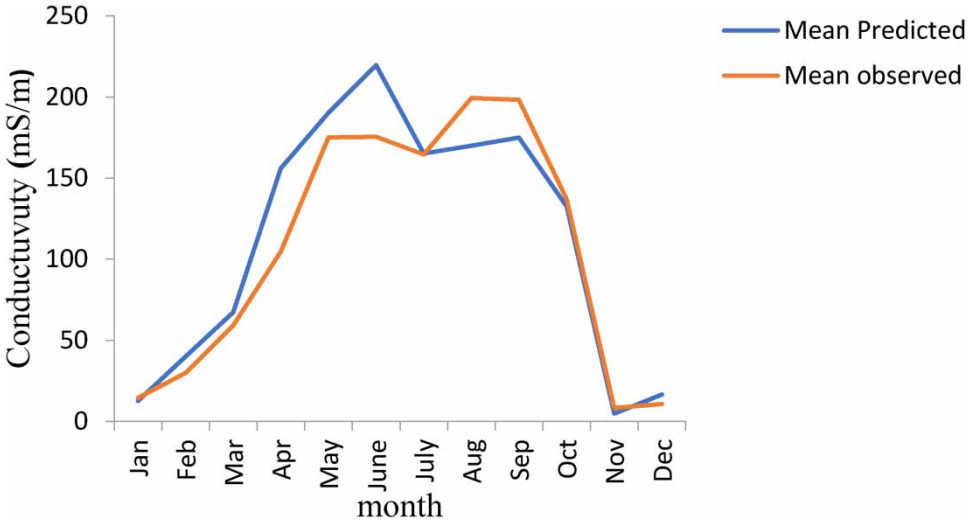

Figure 8 | Mean monthly observed and predicted conductivity for MRC (period).

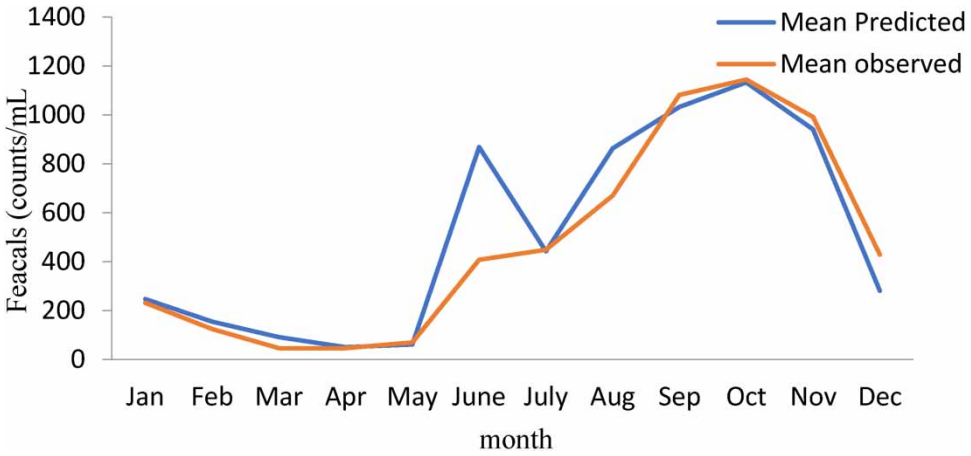

Figure 9 | Mean monthly observed and predicted feacals for MRC (period).

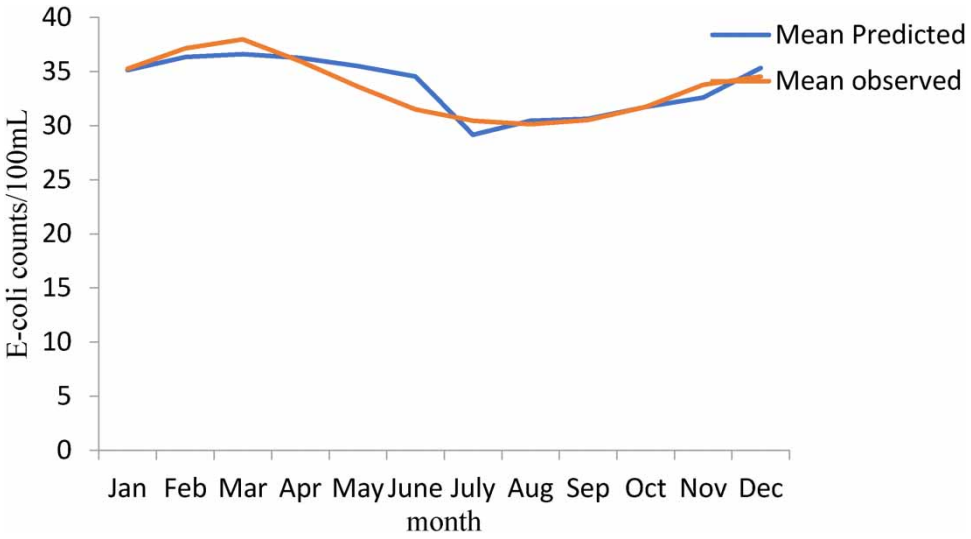

Figure 10 | Mean monthly observed and predicted *Escherichia coli* for MRC (period).

of great similarity, this should not be concluded that the upper and lower reach was of similar hydrological environment. The normal distribution of the gathered datasets determines the form of the distribution. The probability distribution of the simulated parameters with the forecast's mean value and standard deviation indicates that no unusual values for WQ factors exist. However, Kendall's τ test trends show an increase in τ of 0.54 with a 0.67 coefficient of variation at the Mthatha Prison and Efata School for abrupt changes.

According to the findings, there are currently rising trends in pH, NH_4, and *E. coli* WQ parameters (Table 1). This is likely to be continued based on the forested trend (Figures 5, 7, and 10), with observed values of a pH level that has a *Z*-score that is positive depicting high averages of the pH throughout. This may be due as a result of significant increase in the effluent discharge. The mean monthly T–F forecasted model shows an exponential fitted curve with a good correlation value range of 0.79–0.87 for the various predicted parameters as shown in Table 6 while Table 5 suggests that the applied model procedural step gives insight into monthly temporal variations in forecasting WQ parameters.

The application of log-normal in the data distribution has helped visualises the latent power for estimation and predicting value for river's WQ parameters. Hence, the future likelihood of river WQ can be forecasted. Thus, the use of models also makes it possible to consider the challenges, limitations, and opportunities involved in developing a better forecasting model for WQ parameters in a suitably integrated way.

CONCLUSION

This study has presented a monthly time scale distribution for selected measure WQ parameters at Mthatha township and the use of a stochastic T–F forecasting model. The paper provides insights into T–F simulation approach in accessing WQ prediction and forecasting the likely changes that might occur in the future. The study explores the use of Mann–Kendall and line graph models as an intervention in regulating the selected WQ parameters for prompt mitigation measures.

No doubt, the WQ parameters trend exhibits metrics that may be positive or negative over time, but the degree of variability in future prediction is explained by the goodness-of-fit statistics test which invariably was used to calibrate and validate the observed in comparison to the predicted simulated value. As such, the study provides a reliable estimate of WQ constituents and streamflow time series in a river, thereby providing a proactive mechanism to maintain a healthy state assessment of the river. Adaptative catchment managers would find usefulness in the employed stochastic simulation tools in ensuring safe lives for the aquatic animals, and other public users living downstream of the river towards the preservation of their net public benefit values.

Although, previous scholars had advocated and recommended the issuance of noncompliance notices, enforcement of the polluter-pays principle, and seasonal adaptive query for effective management of the WSP that is responsible for the treatment of the effluent discharge in the area. However, all these are rarely implemented and if done, they do lack the political will as per the officer in charge to know the point and when human laxity should not be tolerated. Hence, the study suggests a time interval response for a prompt measure of the River status in terms of certain WQ parameters prediction to complement previous recommendations.

Limitations of the study

This study was limited to modelling the MRC WQ parameter as a whole without considering observation measures at both the upper and lower reaches of the river, thereby assuming a uniform distribution of WQ parameters at the designated sample point. This may not be so, as each upper and lower reach may differ slightly from the simulation of the whole catchment. Also, one could argue that all portions upstream of a river's section have a significant impact on water pollutants downstream, this was not factored into the parameters forecast. Also, the study did not consider the effect of climate change and varied streamflow patterns on the WQ parameters. Similarly, the model's applicability is limited as the analysis ignores mechanisms like self-purification, and pollution retention for example, through sedimentation, and/or dilution due to the inflow of cleaner tributaries.

Although the applied stochastic T–F model may have proved to be effective for prediction, there is a need for further research on its effectiveness compared with other modern trained models like in machine learning (ML), and the use of classification models such as Naive Bayes (NB), Support Vector Machines (SVM), and K Nearest Neighbour (KNN) in forecasting WQ parameter.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support from the Department of Water and Sanitation (DWS), South Africa for providing the data that were used in this study. The financial assistance of the NRF is also acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the DWS.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., Ehteram, M. & Elshafie, A. 2019 Machine learning methods for better water quality prediction. *Journal of Hydrology* **578**, 124084.

Aldhyani, T. H., Al-Yaari, M., Alkahtani, H. & Maashi, M. 2020 Water quality prediction using artificial intelligence algorithms. *Applied Bionics and Biomechanics* **2020**, 1–12.

Alfa, M. I., Ajibike, M. A. & Adie, D. B. 2018 Reliability assessment of Thomas Fiering's method of stream flow prediction. *Nigerian Journal of Technology (NIJOTECH)* **37**(3), 818–823.

American Public Health Association (APHA) 2017 *Standard Methods for Examination of Water and Waste Water*, 25th edn. APHA, Washington, DC.

Bracmort, K. S., Arabi, M., Frankenberger, J., Engel, B. A. & Arnold, J. G. 2006 Modeling long-term water quality impact of structural BMPs. *Transactions of the ASABE* **49**, 367–374.

Brunner, M. I., Bárdossy, A. & Furrer, R. 2019 Stochastic simulation of streamflow time series using phase randomization. *Hydrology and Earth System Sciences* **23**, 3175–3187.

Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H. & Kazakis, N. 2020 Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment* **721**, 137612.

Celeste, A. B., Suzuki, K., Kadota, A. & de Farias, C. A. 2004 Stochastic generation of inflow scenarios to be used by optimal reservoir operation models. *Proceedings of Hydraulic Engineering* **48**, 451–456.

Chen, T., Zhang, H., Sun, C., Li, H. & Gao, Y. 2018 Multivariate statistical approaches to identify the major factors governing groundwater quality. *Applied Water Science* **8**, 1–6.

Condon, L. E. & Maxwell, R. M. 2013 Implementation of a linear optimization water allocation algorithm into a fully integrated physical hydrology model. *Advances in Water Resources* **60**, 135–147.

Cui, Q., Wang, X., Li, C., Cai, Y. & Liang, P. 2016 Improved Thomas–Fiering and wavelet neural network models for cumulative errors reduction in reservoir inflow forecast. *Journal of Hydro-Environment Research* **13**, 134–143.

Devagopal, A., Ashwin, V., Menon, V., Naushad, N. S., Thomas, G. M. & Jyothi, S. 2022 Prediction of water quality parameters of River Periyar using regression models. In: *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, pp. 53–57.

Du Plessis, A. 2015 *Quantifying and Predicting Hydrological Responses of Water Quality Associated with Land Cover Changes Within the Upper Vaal River, South Africa*. University of Johannesburg (South Africa).

Du Plessis, A., Harmse, T. & Ahmed, F. 2014 Quantifying and predicting the water quality associated with land cover change: a case study of the Blesbok Spruit Catchment, South Africa. *Water* **6**, 2946–2968.

DWAF (Department of Water Affairs and Forestry) 1996 *South African Water Quality Guidelines*, 1st edn. Vol. 8. Field Guide, Pretoria, South Africa.

Faruk, D. Ö. 2010 A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence* **23**, 586–594.

Freni, G., Mannina, G. & Viviani, G. 2011 Assessment of the integrated urban water quality model complexity through identifiability analysis. *Water Research* **45**, 37–50.

Frollini, E., Preziosi, E., Calace, N., Guerra, M., Guyennon, N., Marcaccio, M., Menichetti, S., Romano, E. & Ghergo, S. 2021 Groundwater quality trend and trend reversal assessment in the European water framework directive context: an example with nitrates in Italy. *Environmental Science and Pollution Research* **28**, 22092–22104.

Govindarajulu, Z. 1992 Rank Correlation Methods (5th ed.) Technometrics: Vol. 34, No. 1, pp. 108–108.

Hamed, K. H. & Rao, A. R. 1998 A modified Mann–Kendall trend test for autocorrelated data. *Journal of Hydrology* **204**(1–4), 182–196.

Ikudayisi, A. & Adeyemo, J. 2016 Effects of different meteorological variables on reference evapotranspiration modeling: application of principal component analysis. *World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering* **10**, 641–645.

Kamilaris, A. & Ostermann, F. O. 2018 Geospatial analysis and the internet of things. *ISPRS International Journal of Geo-Information* **7**, 269.

Kang, S. & Lin, H. 2007 Wavelet analysis of hydrological and water quality signals in an agricultural watershed. *Journal of Hydrology* **338**, 1–14.

Kumar, K. S. & Rathnam, E. V. 2019 Analysis and prediction of groundwater level trends using four variations of Mann–Kendall tests and ARIMA modelling. *Journal of the Geological Society of India* **94**, 281–289.

Kurunç, A., Yürekli, K. & Cevik, O. 2005 Performance of two stochastic approaches for forecasting water quality and streamflow data from Yeşilurmak River, Turkey. *Environmental Modelling & Software* **20**, 1195–1200.

Loucks, D. P. & Van Beek, E. 2005 *Water Resources Systems Planning and Management: An Introduction to Methods, Model and Applications*. UNESCO, Delft, The Netherlands.

Mann, H. 1945 Non-parametric tests against trend. *Econometrico* **13** (3), 245–259

Manzione, R. L. & Castrignanò, A. 2019 A geostatistical approach for multi-source data fusion to predict water table depth. *Science of The Total Environment* **696**, 133763.

Maroof, L. K., Sule, B. F. & Ogunlela, O. A. 2015 Economic sustainability of integrated hydropower development of Ero-Omola falls, Kwara state, Nigeria. In Decision Making and Knowledge Decision Support Systems (pp. 143–164). Springer, Cham. Gil-Lafuente, A.M. and Zopounidis, C., 2015. In VIII International Conference of RACEF.

McMahon, T. & Mein, R. G. 1978 Reservoir capacity and yield. Elsevier. Amsterdam; 1 Edition, New York: Elsevier Scientific Pub. Co.

Nath, T. K., Tripathy, B. & Das, A. 2018 A study of water quality of river Brahmani, Odisha (India) to assess its potability. *International Journal of Engineering Research & Technology* **7**, 301–311.

Podvezko, V. & Sivilevičius, H. 2013 The use of AHP and rank correlation methods for determining the significance of the interaction between the elements of a transport system having a strong influence on traffic safety. *Transport* **28**, 389–403.

Pohlert, T. 2016 Non-parametric trend tests and change-point detection. *CC BY-ND*, *4*.

Rao, C. M., Modi, P. & Jhajharia, D. 2022 Water Quality Analysis at Mancherial, Jagdalpur and Konta Using Non-parametric Methods. In Advanced Modelling and Innovations in Water Resources Engineering (pp. 609–619). New Delhi, Rao, C.M., Patra, K.C., Jhajharia, D. and Kumari, S. eds., 2022. Advanced Modelling and Innovations in Water Resources Engineering: Select Proceedings of AMIWRE 2021. Singapore: Springer Singapore.

Sathish, S. & Babu, S. K. 2017 November. Stochastic time series analysis of hydrology data for water resources. In IOP Conference Series: New Delhi, Materials Science and Engineering (Vol. 263, No. 4, p. 042140). IOP Publishing. Liu, J., Xiang, Z., Huang, R.H., Wang, D.H. and Ju, Y.Z., 2017. IOP conference series: materials science and engineering.

Sen, P. 1968 Estimated of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association* **39**, 1379–1389.

Thomas, J. A. & Fiering, M. B. 2013 12. Mathematical Synthesis of Streamflow Sequences for the Analysis of River Basins by Simulation. In Design of water-resource systems (pp. 459–493). Harvard University Press. Mass, A., Hufschmidt, M., Dorfman, R., Thomas, H.A., Marglin, S.A. and Fair, G.M., 1962. Design of Water-Resource Systems, Harvard University Press. Cambridge, Mass.

Tripathi, M. & Singal, S. K. 2019 Use of principal component analysis for parameter selection for development of a novel water quality index: a case study of river Ganga India. *Ecological Indicators* **96**, 430–436.

Wang, Q., Tang, J., Zeng, J., Leng, S. & Shui, W. 2019 Regional detection of multiple change points and workable application for precipitation by maximum likelihood approach. *Arabian Journal of Geosciences* **12**, 1–16.

Yang, K., Yu, Z., Luo, Y., Yang, Y., Zhao, L. & Zhou, X. 2018 Spatial and temporal variations in the relationship between lake water surface temperatures and water quality – a case study of Dianchi Lake. *Science of the Total Environment* **624**, 859–871.

First received 29 August 2022; accepted in revised form 4 January 2023. Available online 10 January 2023