



# Prediction of dissolved oxygen, biochemical oxygen demand, and chemical oxygen demand using hydrometeorological variables: case study of Selangor River, Malaysia

Sinan Q. Salih<sup>1</sup> · Intisar Alakili<sup>2</sup> · Ufuk Beyaztas<sup>3</sup> · Shamsuddin Shahid<sup>4</sup> · Zaher Mundher Yaseen<sup>5</sup>

Received: 17 October 2019 / Accepted: 3 August 2020 / Published online: 8 August 2020  
© Springer Nature B.V. 2020

## Abstract

In this research, three water quality (WQ) indexes, namely dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD), in Selangor River of peninsular Malaysia were simulated using a stochastic model based on vector auto-regression (VAR). The simulation was adopted based on three modeling scenarios of inputs as predictor: (i) related WQ parameters, (ii) WQ parameters and river flow data, and (iii) WQ parameters and rainfall data. The WQ parameters as input were determined based on the correlation analysis. The numerical analyses revealed that the prediction accuracy of VAR model substantially increases with the increase in input number. The model provided better accuracy in predictions of WQ indexes (root mean square error  $\approx 0.11$  and mean absolute error  $\approx 0.26$ ) when all environmental, hydrological, and climatological variables were considered. Further improvement in model performance (root mean square error  $\approx 0.0248$  and mean absolute error  $\approx 0.1259$ ) can be achieved if physiochemical parameters like suspended solid material and the turbidity are used as additional inputs.

**Keywords** Water quality parameters · Prediction · Hydrometeorological variables · Tropical environment

---

✉ Zaher Mundher Yaseen  
yaseen@tdtu.edu.vn

<sup>1</sup> Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam

<sup>2</sup> Chemistry Department, Faculty of Science, University of Benghazi, Benghazi, Libya

<sup>3</sup> Department of Mathematics, Bartın University, Bartın, Turkey

<sup>4</sup> School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310, Johor Bahru, Malaysia

<sup>5</sup> Sustainable Developments in Civil Engineering Research Group, Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

# 1 Introduction

## 1.1 Research Background

Several aspects of human life such as hygiene, drinking, and domestic activities need clean water supply (Khalil et al. 2019; Liu & Lu 2014). Unsustainable human activities have contributed immensely to the pollution of water bodies and caused high stress on freshwater resources (Dizaji et al. 2020a; Zhao et al. 2019). The deterioration of water quality (WQ) has severely affected humans well-being and the environment across the globe (Dizaji et al. 2020b; Ouyang 2005). The frequent water pollution episodes due to intense development activities in recent years has attracted the environmental management experts in assessment and prediction of WQ parameters (Martin & McCutcheon 2018; Sharafati et al. 2020).

As one of the fastest developing countries in Asia, Malaysia experienced a rapid land use changes in last few decades (Wan Mohtar et al. 2019). These caused an increase in water pollution in most of the rivers of the country. This situation is further worsened by the environmental, climatic, and hydrological changes experienced in the country in recent years. Due to its location in the tropical zone, natural factors also play a significant role to water pollution (Lee et al. 2017; Dada et al. 2012; Shuhaimi-Othman et al. 2007). For example, heavy rainfall-driven flash flood often sweeps the catchment area to load the river with all kinds of pollutants and worsens river WQ. Soil erosion due to extreme rainfall reduces river water oxygen contents and increases the population of harmful algae. Extensive soil erosion causes sediment deposition in riverbed which significantly influences the river flow conditions. These caused WQ of many rivers in Malaysia deteriorate beyond the recommended level for domestic, agricultural, and industrial purposes (Mukate et al. 2019). Continuous monitoring and prediction of WQ in developing countries like Malaysia are necessary for adoption of necessary measures for reduction in pollution and mitigation of its impacts on society (Naubi et al. 2016).

## 1.2 Research significant and problem statement

The water quality index (WQI) is used for the evaluation suitability of a water body for various water activities (Abba et al. 2020). Its determination is a highly complicated process that involves several water quality variables which often lead to inaccurate determination. Several WQIs have been implemented in different countries, such as Brazil, India, the USA, Korea, and Portugal (Abrahão et al. 2007; Bordalo et al. 2006; Cude 2001; Sargaonkar & Deshpande 2003; Song & Kim 2009). Different WQ parameters are used for the development of WQI in different regions. Among the WQ parameters, dissolved oxygen (DO), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) are considered as the most important components of WQI whose accurate determination can significantly affect the pollution control measures. These WQ parameter values are usually determined through laboratory test. However, the laboratory analysis is a tedious and time-consuming process. A reasonable and accurate prediction of these parameters through analytical processes can save cost, time, and energy. Consequently, researchers have been motivated to develop reliable BOD, BO, and COD prediction models based on other available WQ data. Such models are extremely important for

regions experience frequent water pollutions but have low budgets for environmental quality assessment and monitoring, and control of contamination level.

### 1.3 Soft computing literature review

The need for an accurate, non-vulnerable, and dependable simulation models has been driven by the acknowledgment of the fact that surface water pollution has become a growing concern (Yaseen et al. 2018a, b). Soft computing (SC) methods can be used for reliable modeling of WQ (Yaseen et al. 2018a, b). The main advantage of SC methods lies in their capability to handle the highly nonlinear and complicated relationships between input and output compared to traditional statistical models which assume a linear input–output relationship (Tiyasha et al. 2020). Despite the wide use of SC in the modeling surface WQ, they are still prone to several challenges, especially regarding the model parameters tuning, time inefficiency, lack of generalization, and human–model interactions. Hence, efforts are ongoing toward exploring novel and strong mathematical models that are highly flexible in solving complicated environmental problems (Danandeh Mehr et al. 2018; Tiyasha et al. 2020).

Among different mathematical models applied for solving regression problems, the VAR proposed by Sims (1980) is one of the frequently used statistical models to analyze multivariate time series (Abrigo & Love 2016; Dowell & Pinson 2016; Karlsson 2013; B. Xu & Lin 2015). Basically, it is a natural extension of the univariate time series model to a multivariate data where each variable in the multi-equation system is considered as endogenous (Karlsson 2013). The VAR model has a flexible structure and successfully been applied in many scientific and engineering fields to predict multivariate time series data (Baumeister & Kilian 2012; Fresoli et al. 2015; Kilian & Vigfusson 2013). However, it is yet to be explored for the application of environmental engineering problems such as WQ prediction.

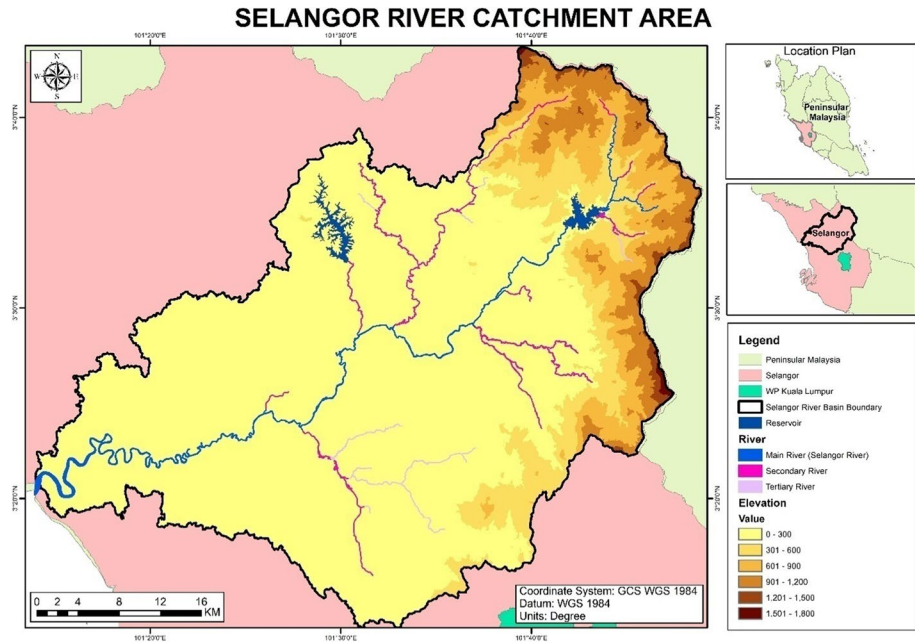
### 1.4 Research objectives

The study involves the statistical analysis of different WQ parameters and establishment of predictive models. The specific objectives of the research are presented as follow:

1. Development of a robust statistical model based on VAR for the prediction of WQ parameters.
2. Prediction of three WQ parameters, i.e., DO, BOD, and COD, using various attributes as predictors including environmental (e.g., related WQ variables), hydrological (e.g., river flow), and climatology (e.g., rainfall) variables.
3. Formulation of a prediction matrix (predictors/predictand) based on correlation statistics.
4. Prediction of different WQ indices from related WQ variables.
5. Prediction of the standard WQI using related water quality indices as predictors.

## 2 Case study and data description

The Selangor River basin is selected in this study due to its significance as a major source of water supply for the capital of Malaysia and the highly populated surrounding regions. It accounts 25% of the total area (2200 km<sup>2</sup>) of peninsular Malaysia (Fig. 1). The size of this



**Fig. 1** Location of Selangor River catchment in the map of peninsular Malaysia

area demands more attention to its environment. The origin of Selangor River is at the Selangor–Pahang border at an elevation of 1700 m. It flows 110 km in southwest direction before emptying into the Straits of Malacca in Kuala Selangor. Several other rivers serve as its main tributaries. Half of the Selangor River basin is covered by natural forest, while a small portion is used for various agricultural purposes. Pollution source of this river comprises of both point source and nonpoint source. There is no readily available data to quantify the level of pollution of the river accurately. Because of the need to study WQ data separately through expert knowledge and experiences, it is difficult to determine the impact of the elemental content in water bodies on both humans and the environment.

River WQ can be assessed in three different ways: (i) considering physicochemical and biological water qualities; (ii) using a physical quality evaluation system which considers the level of manmade changes on the channel margins, river banks, and main channel; and (iii) using a biological WQ evaluation system which considers the state of the living organisms in the water body. Monthly data of several water quality variables (i.e., physical, chemical, and biological) were obtained from the Department of Environment (DoE), Malaysia, for the period 2000–2012. The river flow and rainfall data for the same period and scale were obtained from the Department of Irrigation and Drainage (DID), Malaysia. The statistical properties of data are presented in Table 1.

**Table 1** Statistics and *p*-values of Jarque-Bera normality and augmented Dickey-Fuller (ADF) stationarity tests of different variables used as inputs for the development of prediction model, estimated for the period 2000–2014

Cases	Input variables	Mean	Sd	Skewness	Kurtosis	JB	ADF
Case 1	DO	6.79	1.26	2.48	0.40	0.01	<0.01
	BOD	2.48	1.74	5.35	7.22	<0.01	0.05
	COD	20.67	9.94	4.15	3.44	<0.01	<0.01
	DO per sat	85.73	14.59	2.16	1.12	<0.01	<0.01
	Temp	27.77	1.52	2.78	1.13	<0.01	<0.01
	TS	106.08	107.26	5.22	6.47	<0.01	<0.01
	CA	2.95	1.89	4.31	1.40	<0.01	0.03
	DO, BOD, COD, DO per sat, Temp, TS, CA			3072.37	80.25	<0.01	
Case 2	DO	6.79	1.26	2.48	0.40	0.01	<0.01
	BOD	2.48	1.74	5.35	7.22	<0.01	0.05
	COD	20.67	9.94	4.15	3.44	<0.01	<0.01
	DO per sat	85.73	14.59	2.16	1.12	<0.01	<0.01
	Temp	27.77	1.52	2.78	1.13	<0.01	<0.01
	TS	106.08	107.26	5.22	6.47	<0.01	<0.01
	CA	2.95	1.89	4.31	1.40	<0.01	0.03
	River flow	56.61	31.49	4.53	3.11	<0.01	<0.01
DO, BOD, COD, DO per sat, Temp, TS, CA, river flow			3171.77	70.47	<0.01		

**Table 1** (continued)

Cases	Input variables	Mean	Sd	Skewness	Kurtosis	JB	ADF
Case 3	DO	6.79	1.26	2.48	0.40	0.01	<0.01
	BOD	2.48	1.74	5.35	7.22	<0.01	0.05
	COD	20.67	9.94	4.15	3.44	<0.01	<0.01
	DO per sat	85.73	14.59	2.16	1.12	<0.01	<0.01
	Temp	27.77	1.52	2.78	1.13	<0.01	<0.01
	TS	106.08	107.26	5.22	6.47	<0.01	<0.01
	CA	2.95	1.89	4.31	1.40	<0.01	0.03
	River flow	56.61	31.49	4.53	3.11	<0.01	<0.01
	Rainfall	8.30	3.87	3.44	0.19	0.05	<0.01
	DO, BOD, COD, DO per sat, Temp, TS, CA, river flow, rainfall			3265.76	63.28	<0.01	<0.01
Case 4	DO index	90.48	13.09	1.72	8.16	<0.01	<0.01
	BOD index	90.04	6.86	1.91	5.68	<0.01	0.04
	COD index	73.48	10.58	2.50	1.07	<0.01	<0.01
	ANN Index	84.02	13.92	2.62	-1.30	<0.01	<0.01
	SS index	73.13	18.99	2.50	-0.64	0.01	<0.01
	pH index	95.29	11.56	-2.08	26.42	<0.01	<0.01
	DO per sat	85.73	14.59	2.16	1.12	<0.01	<0.01
	TUR	76.96	97.89	4.23	5.70	<0.01	<0.01
	SS	89.34	108.83	4.31	7.10	<0.01	<0.01
	TS	106.08	107.26	5.22	6.47	<0.01	<0.01
DO index, BOD index, COD index, ANN index, SS index, pH index, DO per sat, TUR, SS, TS			3244.96	57.99	<0.01	<0.01	

**Table 1** (continued)

Cases	Input variables	Mean	Sd	Skewness	Kurtosis	JB	ADF
Case 5	DO index	90.48	13.09	1.72	8.16	<0.01	<0.01
	BOD index	90.04	6.86	1.91	5.68	<0.01	0.04
	COD index	73.48	10.58	2.50	1.07	<0.01	<0.01
	ANN index	84.02	13.92	2.62	- 1.30	<0.01	<0.01
	SS index	73.13	18.99	2.50	- 0.64	0.01	<0.01
	pH index	95.29	11.56	- 2.08	26.42	<0.01	<0.01
	WQI	84.51	7.90	2.40	- 0.48	<0.01	<0.01
	DO index, BOD index, COD index, ANN index, SS index, pH index, WQI			1290.30	30.41	<0.01	

*DO* Dissolved oxygen, *BOD* Biological oxygen demand, *COD* Chemical oxygen Demand, *DO per sat* Dissolved oxygen per saturation, *Temp* Temperature, *TS* Total solids, *TUR* Turbidity, *CA* Calcium, *ANN* Ammoniacal nitrogen, *SS* Suspended solids

### 3 Methodology

#### 3.1 Vector auto-regression (VAR) model

The VAR is a multivariate statistical analysis model frequently used to explain relationship between different attributes to form a dynamic system of linear equations. While doing so, it incorporates interdependence between different time series by expressing them as linear combinations of the series. In other words, it is used to describe; “How a time series depends on other related time series and their past lags?”. The VAR model is used in this study to describe "How DO, BOD, and COD depend on the related WQ, hydrological, and climatological variables and their past values?".

If  $Y_t = \{Y_{1t}, Y_{2t}, \dots, Y_{nt}\}$  denote a  $n$  dimensional time series of size  $T$ , the VAR process with  $n$  endogenous variables and  $p$  lags, VAR ( $p$ ), is represented as follows:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \dots + \varphi_p Y_{t-p} + \varepsilon_t, t = 1, 2, \dots, T, \tag{1}$$

where  $Y_t = (Y_{1t}, Y_{2t}, \dots, Y_{nt})^T \in R^n$  denotes a  $n$  dimensional multivariate variable,  $\Phi = (\varphi_0, \varphi_1, \varphi_2, \dots, \varphi_p)$  is the vector of unknown model parameters, where each  $\varphi_i, i = 1, 2, \dots, p$ , represents an  $n \times n$  coefficient matrix and  $\varphi_0 = (\varphi_{01}, \varphi_{02}, \dots, \varphi_{0n})^T$  is an  $n$  vector of intercept term,  $\varepsilon_t$  follows a multivariate normal distribution with mean zero and variance–covariance matrix  $\Sigma, \varepsilon_t \sim N(0, \Sigma)$ . For  $k = 1, 2, \dots, n, Y_{(k)} = (Y_{k1}, Y_{k2}, \dots, Y_{kT})^T$  for the  $k$ th time series data,

$$Z = \begin{bmatrix} 1 & z_0^T \\ \vdots & \vdots \\ 1 & z_{T-1}^T \end{bmatrix}, \tag{2}$$

with  $z_{t-1} = (Y_{t-1}^T, Y_{t-2}^T, \dots, Y_{t-p}^T)$ . The vector of coefficients  $\Phi = (\varphi_0, \varphi_1, \varphi_2, \dots, \varphi_p)^T$  can be estimated by the ordinary least squares method solving the system of linear equations  $Y = Z\Phi + \varepsilon$  as follows:

$$\hat{\Phi} = (Z^T Z)^{-1} Z^T Y. \tag{3}$$

Note that in this study, the VAR parameters were calculated by considering the time series stationary (Kilian 1998).

The VAR model provides a flexible way to predict the unobservable realization of the multivariate time series data. Once the parameter vector  $\hat{\Phi} = (\hat{\varphi}_0, \hat{\varphi}_1, \hat{\varphi}_2, \dots, \hat{\varphi}_p)^T$  is estimated, the point predictor of  $h = 1, 2, \dots$  step-ahead future value conditionally on the available data set,  $Y_{T+h}$ , is obtained using the recursion formula, as follows (Xu and Moon 2011):

$$\hat{Y}_{T+hVT} = \hat{\varphi}_0 + \hat{\varphi}_1 Y_{T+h-1VT} + \hat{\varphi}_2 Y_{T+h-2VT} + \dots + \hat{\varphi}_p Y_{T+h-pVT}, \tag{4}$$

where  $\hat{Y}_{T+kVT} = 0$  for  $k \leq 0$ . However, obtaining point forecasts may not be enough to assess forecasting accuracy since the uncertainties associated with point forecasts are unknown in practice (Chatfield 1993). To overcome this issue, the prediction interval, which takes into account the uncertainty associated with each point forecast, can be used to make a reliable inference about the future values. For VAR model, the forecast densities are obtained in the form of ellipsoid, and the marginal prediction intervals can be obtained accordingly (Lütkepohl and Poskitt 1991). The mean square error matrix of  $\hat{Y}_{T+jVT}$  is



denoted as  $\Sigma(h) = \sum_{j=0}^{h-1} \psi_j \psi_j^T$ , where  $\psi_j = \sum_{k=1}^{p-1} \psi_{j-k} \varphi_k$  and  $\psi_0 = I_n$  and  $\varphi_k = 0$  for  $j > p$ . Under the assumption of normal distributed errors, the  $h$ -step-ahead forecast density is estimated as follows:

$$Y_{T+h} \sim N(\hat{Y}_{T+h|T}, \Sigma(h)). \tag{5}$$

The prediction interval with coverage probability  $100(1 - \alpha)\%$  for the  $k$ th time series is then obtained as follows (Xu and Moon 2011):

$$\{Y_{T+h} \vee Y_{T+h} \in [\hat{Y}_{kT} \pm z_{\alpha/2n} \hat{\sigma}_k(h)]\}, \tag{6}$$

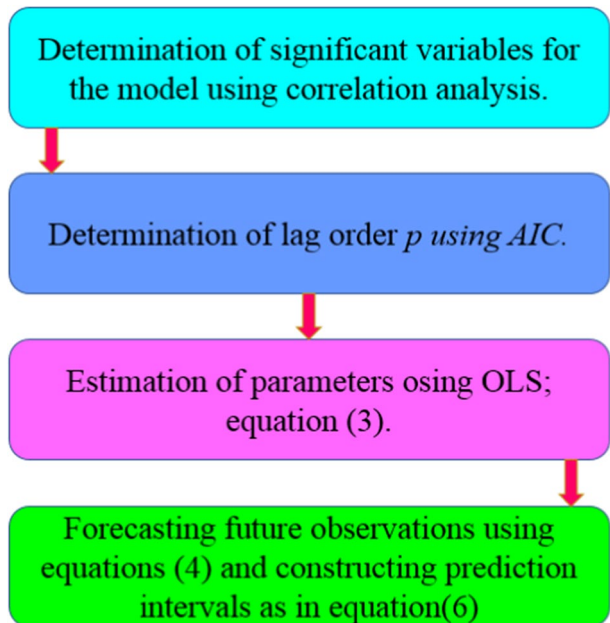
where  $\hat{\sigma}_k(h)$  is the  $k$ th element of  $\hat{\Sigma}(h)$  and  $z_{\alpha/2n}$  is the  $\alpha/2n$ th, say 0.025, quantile of the standard normal distribution.

For the sake of clarity, a flowchart is presented in Fig. 2 to show how the proposed method was used in this study to obtain the experimental results.

### 3.2 Model development

Five different models were built based on three different input scenarios. For each case, the inputs were determined by taking into account their correlations with the targeted variable. Only the variables which have moderate or high correlation ( $>0.4$ ) were selected. The correlation matrix between the predictors and predictand is illustrated in Fig. 3. The statistical characteristics of each investigated variable considered in the present study are given in Table 1. The sample statistics and  $p$  values of Jarque–Bera (JB) normality and augmented Dickey–Fuller (ADF) stationarity tests for each variable individually and jointly are given in Table 1 (Jarque 2011; Mushtaq 2011). The obtained results can be interpreted as follows:

Fig. 2 Flowchart of the proposed VAR model



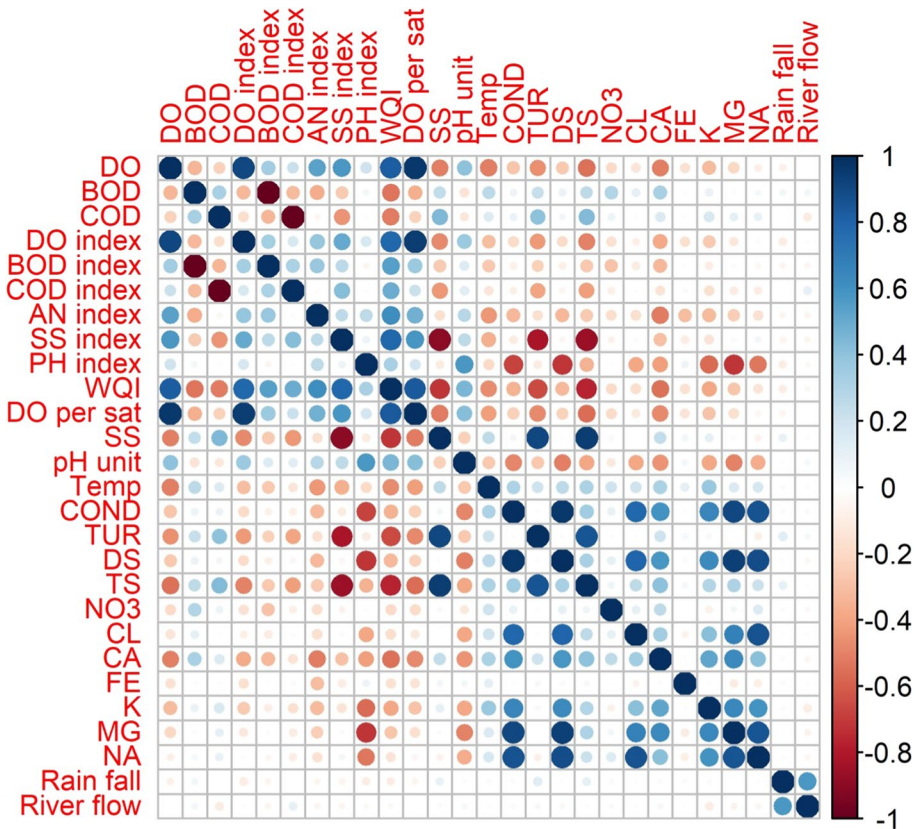


Fig. 3 Correlation matrix between the predictand and predictor water quality variables

(i) the JB test results indicate that the distribution of input variables is Gaussian, and (ii) small  $p$  values of ADF test indicate that all the time series are stationary.

The Ljung–Box (LB) test was performed for each case to test the autocorrelation structure in the original and squared series. The results suggested that there was a dynamic dependence in the conditional mean for each case. Obtained LB test statistics of the squared series were relatively smaller than those of original series. This indicates that the dependency of the second-order moments was not significant than that obtained from the conditional mean dependency. Overall, the results of explanatory data analysis suggested suitability of VAR to model each objective. For each case, the optimal lag order was determined by Akaike information criterion (AIC) by equating the maximum lag order to 14 (Kilian 1998; Marcellino et al. 2006). The results are given in Table 2, which shows that VAR(5) (Objective 1), VAR(14) (Objective 2-3-5), and VAR(13) (Objective 4) models were optimal.

### 3.3 Prediction performance metrics

The predictability of the developed model for the research objectives was evaluated and validated using various statistical indicators including root mean squared errors

**Table 2** Results of Akaike information criterion for each designed objective

Order	Objective 1	Objective 2	Objective 3	Objective 4	Objective 5
1	18.48	25.23	27.45	52.45	21.41
2	17.99	24.79	27.09	52.32	21.03
3	17.82	24.62	26.99	52.16	21.09
4	17.95	24.80	27.15	52.30	21.28
<b>5</b>	<b>16.81</b>	23.74	26.07	52.85	21.47
6	17.12	24.08	26.54	52.95	21.68
7	17.31	24.30	26.56	53.36	21.79
8	17.42	24.30	26.25	53.43	21.76
9	17.49	24.36	26.23	53.22	21.49
10	17.57	24.36	26.02	52.72	21.64
11	17.39	23.78	24.90	51.73	21.71
12	17.35	23.24	23.68	50.20	21.55
13	17.02	22.03	20.85	<b>36.35</b>	21.41
14	16.93	<b>20.36</b>	<b>16.76</b>	38.45	<b>20.47</b>

Bold indicates the best values of the Akaike information criterion for each designed objective.

(RMSE), mean absolute errors (MAE), determination coefficient ( $R^2$ ), and correlation coefficient ( $R$ ). The mathematical expression of these indicators are given in Eqs. (7)–(10) (Tao et al. 2018a, b; Tao et al. 2018a, b):

$$RMSE = \sqrt{\frac{\sum_{t=1}^T (\hat{Y}_t - Y_t)^2}{T}}, \tag{7}$$

$$MAE = \frac{\sum_{t=1}^T |\hat{Y}_t - Y_t|}{T}, \tag{8}$$

$$R^2 = \left( \frac{\sum_{t=1}^T (Y_t - \bar{Y}_t) \sum_{t=1}^T (\hat{Y}_t - \bar{Y}_t)}{\sqrt{\sum_{t=1}^T (Y_t - \bar{Y}_t)^2} \sqrt{\sum_{t=1}^T (\hat{Y}_t - \bar{Y}_t)^2}} \right)^2, \tag{9}$$

$$R = \left( \frac{\sum_{t=1}^T (Y_t - \bar{Y}_t) \sum_{t=1}^T (\hat{Y}_t - \bar{Y}_t)}{\sqrt{\sum_{t=1}^T (Y_t - \bar{Y}_t)^2} \sqrt{\sum_{t=1}^T (\hat{Y}_t - \bar{Y}_t)^2}} \right), \tag{10}$$

where  $\hat{Y}_t$  and  $Y_t$  denote the predicted and observed time series, respectively.  $\bar{\hat{Y}}_t$  and  $\bar{Y}_t$  are the mean of predicted and observed time series, respectively.  $T$  is the sample number.

## 4 Results and discussion

The aim of this study was to develop a reliable mathematical model based on VAR for prediction of DO, BOD, COD, and WQI in river water. The concentration of WQ parameters are highly correlated with various environmental, hydrological, and climatological variables. Hence, establishing a comprehensive prediction model through incorporation of all those variables is essential for reliable prediction of WQ parameters. In the current research, three different modeling scenarios were investigated to predict DO, COD, BOD, and WQI: (i) environmental, (ii) environmental and hydrological, and (iii) environmental, hydrological, and climatological variables.

Table 3 reports the results obtained for DO, COD, and BOD for the first three designed cases (see Table 1). The DO was predicted with minimal RMSE  $\approx 0.92$  and MAE  $\approx 0.76$  for the first case where only the related water quality variables were used as predictors. The DO value was predicted with RMSE  $\approx 0.20$  and MAE  $\approx 0.36$  for Case 2, where river flow magnitude was included as hydrological attribute. In Case 3, where river flow and rainfall information were also incorporated as hydrological and climatological attributes, the prediction accuracy of DO was enhanced much (RMSE  $\approx 0.11$  and MAE  $\approx 0.26$ ).

The model performance in predicting COD was found similar to that observed for DO. For Case 1, COD was predicted with minimal RMSE  $\approx 1.84$  and MAE  $\approx 0.099$ . Case 2 showed a slight prediction enhancement with RMSE  $\approx 0.66$  and MAE  $\approx 0.63$ , whereas a noticeable improvement in prediction accuracy was observed for Case 3 after incorporation of hydrological and climatological attributes (RMSE  $\approx 0.49$  and MAE  $\approx 0.54$ ). The results obtained for prediction of BOD were as follows: Case 1: RMSE  $\approx 66.19$  and MAE  $\approx 6.10$ ; Case 2: RMSE  $\approx 19.83$  and MAE  $\approx 3.42$ ; and Case 3: RMSE  $\approx 9.72$  and MAE  $\approx 2.46$ . The model performance estimated using RMSE, MAE,  $R^2$ , and  $R$  presented in Table 3 revealed that the applied VAR predictive model provided a confirmatory evidence that incorporation of different environmental, hydrological, and

**Table 3** Performance of the models in prediction of water quality parameters for Cases 1 to 3

Performance indicators	Variables		
	DO	COD	BOD
Case-1			
RMSE	0.9236	1.8456	66.1975
MAE	0.7646	0.9959	6.1020
$R^2$	0.3285	0.4014	0.3415
$R$	0.5735	0.6336	0.5844
Case-2			
RMSE	0.2063	0.6600	17.8312
MAE	0.3600	0.6328	3.4278
$R^2$	0.8487	0.7969	0.8239
$R$	0.9212	0.8926	0.9076
Case-3			
RMSE	0.1120	0.4914	9.7250
MAE	0.2636	0.5409	2.4609
$R^2$	0.9178	0.8487	0.9039
$R$	0.9580	0.9212	0.9507

climatological variables as predictors improves the accuracy of the prediction model. The model performance in predicting water quality indices and standard WQI for the fourth and fifth cases are presented in Table 4.

Taylor diagram was generated for better evaluation of model for the first three inspected cases. Taylor diagram is a presentation of three different statistics including correlation, standard deviation, and RMSE (Taylor 2001). Figure 4a–c presents the Taylor diagrams for DO, COD, and BOD. The trend of the coordinates for the three inspected cases was almost the same, where the closest to the observed data (black circle) with a correlation coefficient of above 0.92 was observed when all predictors (i.e., environmental, hydrological, and climatological) were taken into consideration.

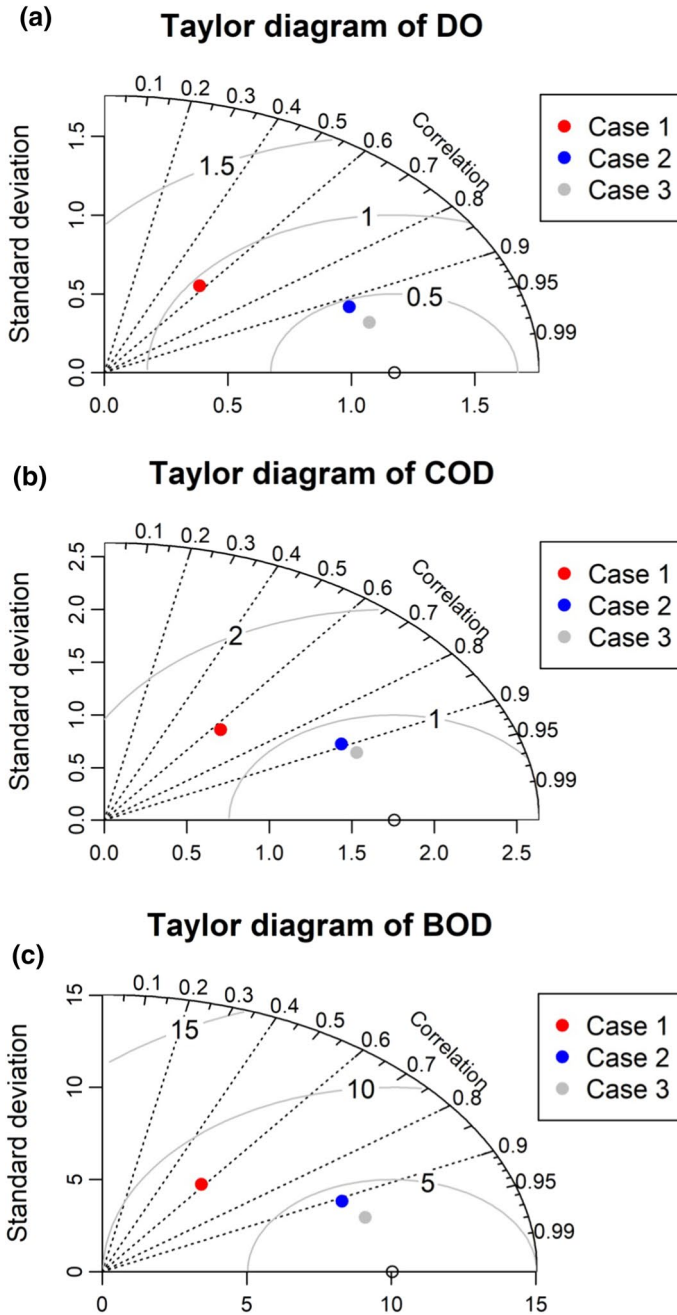
Figure 5a–c illustrates the time series of the observed and predicted values. It is very obvious that prediction model performed best in prediction of DO, COD, and BOD for Case 3 (see Fig. 4a). The VAR predicted values exhibited a remarkable match with the observed WQ values. Figure 5c reports the model performance in prediction of standard WQI from the indices presented in Fig. 5b. In a general overview, an acceptable match was noticed between the observed and predicted time series.

Figures 6, 7, and 8 present  $h=6$ -step-ahead data forecasts and asymptotic prediction intervals produced by VAR models for Case 3 (as the best model) and Cases (4–5). Each variable considered in this study was consisted of 156 observations in total. To obtain the forecasts and prediction intervals, first 150 observations of each variable were used to train the VAR model. Subsequently, the point forecasts and prediction intervals were obtained using Eqs. (4) and (6), respectively. The results indicated that the VAR model produced reasonable predictions for Cases 3 and 5 (Figs. 6 and 8). The forecasted data were relatively close to the observed data points, and almost all the future values were covered by the prediction intervals. For Case 4, the VAR model was failed to produce satisfactory forecasts (see Fig. 7). This is due to the fact that the variables in this case were ranged in a large scale and thus a larger error compared to other two cases. Another possible reason could be the non-Gaussian forecast errors produced by the VAR model. Alternatively, a nonparametric resampling technique such as bootstrap can be used to improve the prediction performance of the VAR model.

It was observed that an increment in the number of the input attributes substantially augmented the prediction accuracy of the model. Hence, it was very interesting to explore the influence of suspended solid material and the turbidity on DO, COD, and BOD prediction for first three cases (Table 1). The obtained results are given in Table 5. The results showed a better prediction in comparison with that presented in Table 3. The SS and TUR exhibited a noticeable influence on DO and COD prediction, while no influence on BOD. This is justifiable as BOD is mainly influenced by biological properties of water instead

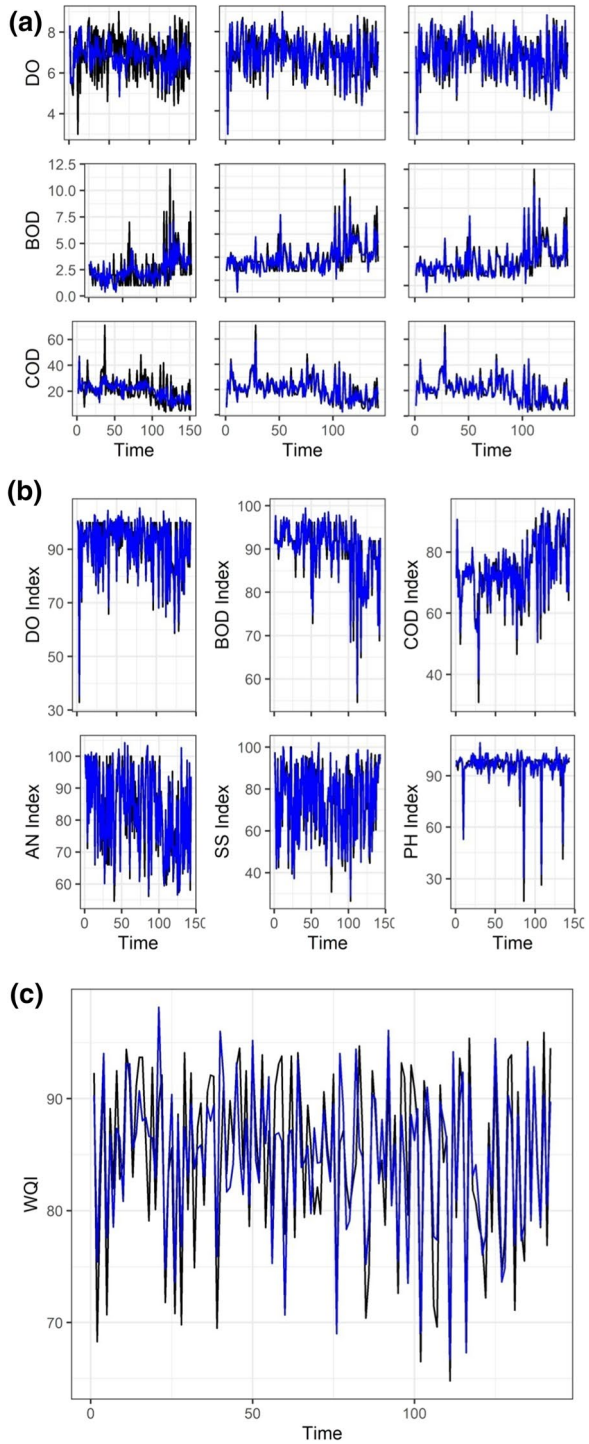
**Table 4** Performance of the models in prediction of different water quality indices and general WQI for the fourth and fifth cases

Performance indicators	DO index	BOD index	COD index	AN index	SS index	pH index	WQI
Case-1							
RMSE	5.5966	3.3304	9.0097	16.8250	19.6633	16.0593	19.2687
MAE	1.9519	1.3848	2.3557	3.2032	3.4633	3.1063	3.6137
$R^2$	0.9574	0.9337	0.9213	0.9130	0.9407	0.8686	0.6748
$R$	0.9784	0.9663	0.9598	0.9555	0.9699	0.9320	0.8214

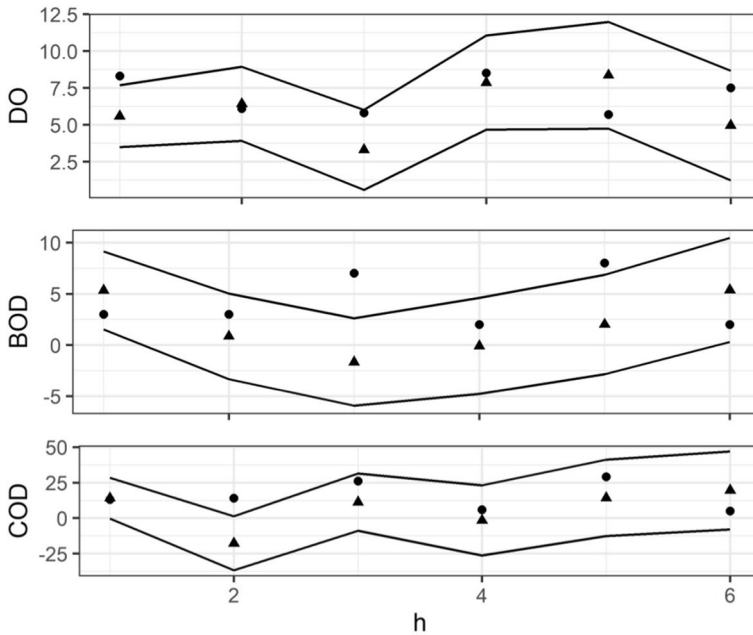


**Fig. 4** Taylor diagram of the observed and predicted water quality parameters, **a** DO, **b** COD, and **c** BOD for the first three Cases

**Fig. 5** **a** Observed (black lines) versus fitted (blue lines) time series for Case 1 (first row), Case 2 (second row), and Case 3 (third row), **b** Observed (black lines) versus fitted (blue lines) time series for Case 4, and **c** Observed (black lines) versus fitted (blue lines) time series for Case 5







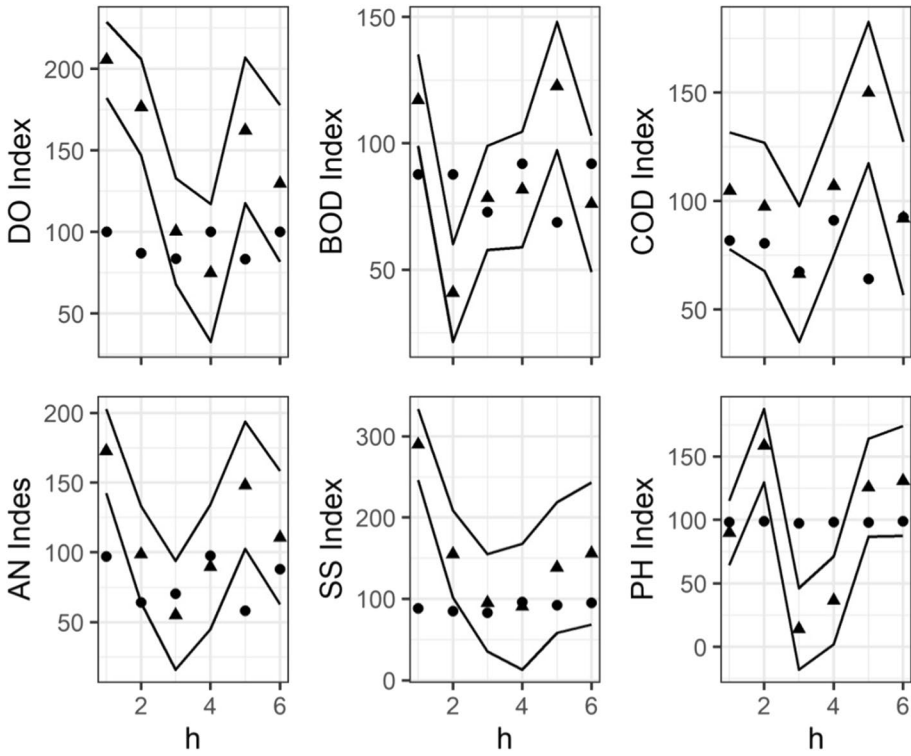
**Fig. 6** Six-step-ahead unobserved data (dot points) forecasts (triangle points) and 95% prediction intervals (solid lines) for Case 3

of physical properties. The results tabulated in Table 5 exhibits a high complexity in modeling due to an increase in tuning parameters of VAR model. Overall, the current investigation confirmed significant influence of hydrometeorological variables on modeling water quality. The model performance was noticed to improve significantly after incorporation of climatic and hydrological information.

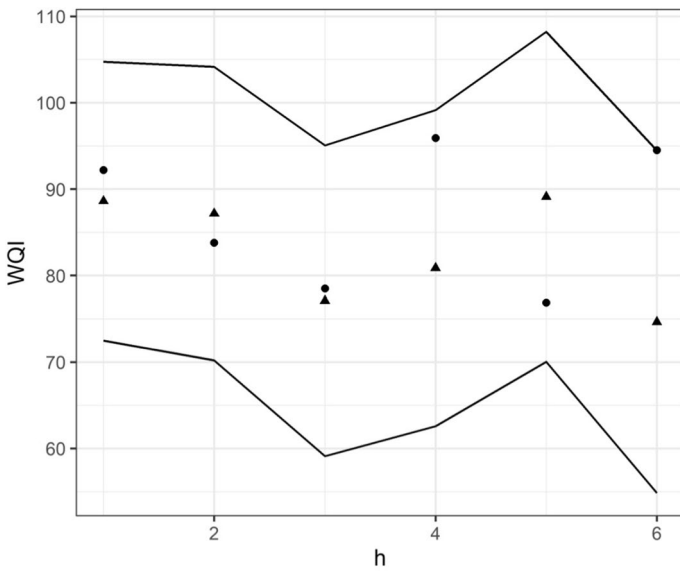
## 5 Conclusions

Modeling spatial and temporal variation of river WQ parameters is important for several consumptive uses such as drinking, hygiene, and irrigation. The river WQ data are characterized by high redundancy and stochasticity due to its dependence on several environmental, climatological, and hydrological parameters. A stochastic model based on VAR is proposed in this study for modeling of river WQ parameters. The results revealed that VAR was successful to address this complex environmental engineering problem efficiently. The VAR model also exhibited its robustness in prediction of WQ variables six steps ahead. Assessment of model performance of different input scenarios revealed that incorporation of river flow and rainfall information as an external climatology and hydrological attributes significantly improves the prediction capability of VAR model. Besides, the physiochemical parameters like suspended solid material and the turbidity can further improve the model performance. The results indicate the potential of the proposed VAR model to be integrated with an expert system to provide a reliable and dependable source





**Fig. 7** Six-step-ahead unobserved data (dot points) forecasts (triangle points) and 95% prediction intervals (solid lines) for Case 4



**Fig. 8** Six-step-ahead unobserved data (dot points) forecasts (triangle points) and 95% prediction intervals (solid lines) for Case 5

**Table 5** Computed performance indicators metrics in prediction DO, COD, and BOD for the first three cases through the involvement of SS and TUR as external attributes

Performance metrics	Variables		
	DO	COD	BOD
Case-1			
RMSE	0.0960	0.3011	12.4550
MAE	0.2407	0.4371	2.9070
$R^2$	0.9295	0.9073	0.8770
$R$	0.9641	0.9525	0.9364
Case-2			
RMSE	0.0521	0.2746	7.8509
MAE	0.1840	0.4102	2.2248
$R^2$	0.9617	0.9149	0.9219
$R$	0.9806	0.9565	0.9601
Case-3			
RMSE	0.0248	0.2902	10.2937
MAE	0.1259	0.4220	2.5664
$R^2$	0.9817	0.9095	0.8969
$R$	0.9908	0.9536	0.9470

of information to the decision makers for taking actions toward enhancement of sustainability of river systems.

**Author contributions** SS was involved in conceptualization; ZY, IA, and SS were involved in formal analysis; ZY and UB were involved in the investigation; UB performed the methodology; SS was involved in project administration; SS was involved in resources; UB was involved in software; SS was involved in supervision; ZY was involved in validation; UB was involved in visualization; ZY, IA, SS, and SS were involved in writing—original draft; ZY, IA, and SS were involved in writing—review and editing.

**Funding** This research received no external funding.

## Compliance with ethical standards

**Conflict of interest** The authors declare no conflict of interest.

## References

- Abba, S. I., Hadi, S. J., Sammen, S. S., Salih, S. Q., Abdulkadir, R. A., Pham, Q. B., et al. (2020). Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *Journal of Hydrology*, *587*, 124974. <https://doi.org/10.1016/j.jhydrol.2020.124974>.
- Abrahão, R., Carvalho, M., Da Silva, W. R., Machado, T. T. V., Gadelha, C. L. M., & Hernandez, M. I. M. (2007). Use of index analysis to evaluate the water quality of a stream receiving industrial effluents. *Water SA*, *33*(4), 459–465. <https://doi.org/10.4314/wsa.v33i4.52940>.
- Abrigo, M. R. M., & Love, I. (2016). Estimation of panel vector autoregression in Stata. *The Stata Journal*, *16*(3), 778–804.
- Baumeister, C., & Kilian, L. (2012). Real-time forecasts of the real price of oil. *Journal of Business & Economic Statistics*, *30*(2), 326–336.

- Bordalo, A. A., Teixeira, R., & Wiebe, W. J. (2006). A water quality index applied to an international shared river basin: The case of the Douro River. *Environmental Management*, 38(6), 910–920. <https://doi.org/10.1007/s00267-004-0037-6>.
- Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2), 121–135.
- Cude, C. G. (2001). Oregon water quality index: A tool for evaluating water quality management effectiveness. *Journal of the American Water Resources Association*, 37, 125–137.
- Dada, A. C., Asmat, A., Gires, U., Heng, L. Y., & Deborah, B. O. (2012). Bacteriological monitoring and sustainable management of beach waterquality in Malaysia: problems and prospects. *Global Journal of Health Science*, 4(3), 126.
- Danandeh Mehr, A., Nourani, V., Kahya, E., Hrnjica, B., Sattar, A. M. A., & Yaseen, Z. M. (2018). Genetic programming in water resources engineering: A state-of-the-art review. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2018.09.043>.
- Dizaji, A. R., Hosseini, S. A., Rezaverdinejad, V., & Sharafati, A. (2020a). Groundwater contamination vulnerability assessment using DRASTIC method, GSA, and uncertainty analysis. *Arabian Journal of Geosciences*, 13(14), 1–15.
- Dizaji, A. R., Hosseini, S. A., Rezaverdinejad, V., & Sharafati, A. (2020b). Assessing pollution risk in Arda-bil aquifer groundwater of Iran with arsenic and nitrate using the SINTACS Model. *Polish Journal of Environmental Studies*, 29(4) (in press).
- Dowell, J., & Pinson, P. (2016). Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Transactions on Smart Grid*, 7(2), 763–770.
- Fresoli, D., Ruiz, E., & Pascual, L. (2015). Bootstrap multi-step forecasts of non-Gaussian VAR models. *International Journal of Forecasting*, 31(3), 834–848.
- Jarque, C. M. (2011). Jarque–Bera test. *International Encyclopedia of Statistical Science*, pp. 701–702.
- Karlsson, S. (2013). Forecasting with Bayesian vector autoregression. In *Handbook of economic forecasting* (Vol. 2, pp. 791–897). Elsevier.
- Khalil, B., Adamowski, J., Abdin, A., & Elsaadi, A. (2019). A statistical approach for the estimation of water quality characteristics of ungauged streams/watersheds under stationary conditions. *Journal of Hydrology*, 569, 106–116.
- Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of economics and statistics*, 80(2), 218–230.
- Kilian, L., & Vigfusson, R. J. (2013). Do oil prices help forecast US real GDP? The role of nonlinearities and asymmetries. *Journal of Business & Economic Statistics*, 31(1), 78–93.
- Lee, I., Hwang, H., Lee, J., Yu, N., Yun, J., & Kim, H. (2017). Modeling approach to evaluation of environmental impacts on river water quality: A case study with Galing River, Kuantan, Pahang, Malaysia. *Ecological Modelling*, 353, 167–173.
- Liu, M., & Lu, J. (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-014-3046-x>.
- Lütkepohl, H., & Poskitt, D. S. (1991). Estimating orthogonal impulse responses via vector autoregressive models. *Econometric Theory*, 7(4), 487–496.
- Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1–2), 499–526.
- Martin, J. L., & McCutcheon, S. C. (2018). *Hydrodynamics and transport for water quality modeling*. Boca Raton: CRC Press.
- Mukate, S., Wagh, V., Panaskar, D., Jacobs, J. A., & Sawant, A. (2019). Development of new integrated water quality index (IWQI) model to evaluate the drinking suitability of water. *Ecological Indicators*, 101, 348–354.
- Mushtaq, R. (2011). Augmented dickey fuller test.
- Naubi, I., Zardari, N. H., Shirazi, S. M., Ibrahim, N. F. B., & Baloo, L. (2016). Effectiveness of water quality index for monitoring Malaysian river water quality. *Polish Journal of Environmental Studies*, 25, 231–239. <https://doi.org/10.15244/pjoes/60109>.
- Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. *Water Research*, 39(12), 2621–2635. <https://doi.org/10.1016/j.watres.2005.04.024>.
- Sargaonkar, A., & Deshpande, V. (2003). Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. *Environmental Monitoring and Assessment*, 89(1), 43–67. <https://doi.org/10.1023/A:1025886025137>.
- Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection*.

- Shuhaimi-Othman, M., Lim, E. C., & Mushrifah, I. (2007). Water quality changes in Chini lake, Pahang, west Malaysia. *Environmental monitoring and assessment*, 131(1–3), 279–292.
- Sims, C. A. (1980). Vector autoregressions and reality. *Econometrica*, 48, 1–48.
- Song, T., & Kim, K. (2009). Development of a water quality loading index based on water quality modeling. *Journal of Environmental Management*, 90(3), 1534–1543. <https://doi.org/10.1016/j.jenvm.2008.11.008>.
- Tao, H., Bobaker, A. M., Ramal, M. M., Yaseen, Z. M., Hossain, M. S., & Shahid, S. (2018). Determination of biochemical oxygen demand and dissolved oxygen for semi-arid river environment: Application of soft computing models. *Environmental Science and Pollution Research*. <https://doi.org/10.1007/s11356-018-3663-x>.
- Tao, H., Diop, L., Bodian, A., Djaman, K., Ndiaye, P. M., & Yaseen, Z. M. (2018). Reference evapotranspiration prediction using hybridized fuzzy model with firefly algorithm: Regional case study in Burkina Faso. *Agricultural Water Management*, 208, 140–151.
- Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7), 7183–7192. <https://doi.org/10.1029/2000JD900719>.
- Tiyasha, Tung, T. M., & Yaseen, Z. M. (2020). A survey on river water quality modelling using artificial intelligence models: 2000–2020. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2020.124670>.
- Wan Mohtar, W. H. M., Abdul Maulud, K. N., Muhammad, N. S., Sharil, S., & Yaseen, Z. M. (2019). Spatial and temporal risk quotient based river assessment for water resources management. *Environmental Pollution*, 248, 133–144. <https://doi.org/10.1016/j.envpol.2019.02.011>.
- Xu, B., & Lin, B. (2015). Carbon dioxide emissions reduction in China's transport sector: A dynamic VAR (vector autoregression) approach. *Energy*, 83, 486–495.
- Xu, J., & Moon, S. (2011). Stochastic forecast of construction cost index using a cointegrated vector autoregression model. *Journal of Management in Engineering*, 29(1), 10–18.
- Yaseen, Z. M., Ehteram, M., Sharafati, A., Shahid, S., Al-Ansari, N., & El-Shafie, A. (2018). The integration of nature-inspired algorithms with least square support vector regression models: application to modeling river dissolved oxygen concentration. *Water*, 10(9), 1124. <https://doi.org/10.3390/w10091124>.
- Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K.-W. (2018). An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology*, 569, 387–408. <https://doi.org/10.1016/j.jhydrol.2018.11.069>.
- Zhao, C. S., Yang, Y., Yang, S. T., Xiang, H., Wang, F., Chen, X., et al. (2019). Impact of spatial variations in water quality and hydrological factors on the food-web structure in urban aquatic environments. *Water Research*, 153, 121–133.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.