A Comparative Analysis of Multivariate Statistical Time-series Models for Water Quality Forecasting of the River Ganga

Mogarala Tejoyadav[1, 2][0000-0002-0780-9887], Rashmiranjan Nayak[1,3][0000-0002-1380-701X], and Umesh Chandra Pati[1,4][0000-0001-9805-2543]

[1]National Institute of Technology Rourkela, Odisha-769008, India
[2]tejoyadav413@gmail.com, [3]rashmiranjan.et@gmail.com, [4]ucpati@nitrkl.ac.in

Abstract. Water plays an important role in the livelihood of mankind. Hence, water that is used for agriculture, marine culture, human consumption, etc., should be in good condition to minimize the hazardous effect of water pollution on human health. Rapid unsustainable industrialization, improper huge waste disposal, excess amount fertilizer usage, etc., are responsible for the rapid deterioration of the water quality in rivers and other freshwater bodies. Manual continuous water quality measurement is risky, expensive, and time-consuming. Hence, it is essential to forecast the water quality using statistical time-series models. In this paper, three widely used statistical multivariate techniques such as Vector Moving Average (VMA), Vector Auto Regression (VAR), and Vector Auto Regression Moving Average (VARMA), are investigated to forecast water quality parameters like Fecal Coliform (FC), Total Coliform (TC), Biological Oxygen Demand (BOD), Dissolved Oxygen (DO), and the associated Water Quality Index (WQI) of the Ganga river. Most of the previous methods worked on forecasting the future values based on past values of individual parameters without considering the interdependency among the water quality parameters. Here, correlation among each parameter is estimated. Subsequently, the future values of a parameter are estimated based on its previous values and the previous values of its correlated parameters. The proposed research work can help properly manage the water quality of the river Ganga by utilizing the forecasted results for the planning of the pollution control strategies. Finally, it helps improve the quality of human beings by minimizing the health issues caused by water pollution.

Keywords: VAR, VMA, VARMA, River Ganga, Multivariate time series forecasting, water quality index.

1 Introduction

Water is the primary need of every living being on the earth for the continuation of life. Hence, the quality of water plays the main role in any water bodies like rivers, ponds, lakes, reservoirs, etc. Water covers 70.9% of this earth's surface, all that is in oceans and seas, which is not useful for the livelihood of mankind. Only a small portion of the

water present on the ground surface is useful for mankind. Rivers are considered as one of the most important sources of water for irrigation, industrial needs, and other uses. Due to the dynamic nature of the river structures and also of its smooth access in disposing of waste, the river structures are becoming most liable for the unfavorable results of environmental pollution [1]. Subsequently, humankind has been suffering from water pollution-related health issues such as typhoid, dysentery, cholera, etc. Further, not only humans but also agricultural crops and marine culture are getting damaged. Therefore, it is necessary to take necessary steps to manage and control the water quality of the rivers. Water quality may be defined as the biological, chemical, and physical condition or state of water [1]. The river Ganga is one of the important and largest rivers in India. However, its water quality has been continuously deteriorating. Hence, this work aims at implementing the time-series forecasting techniques to forecast the water quality parameters of Ganga like Fecal Coliform (FC), Total Coliform (TC), Biological Oxygen Demand (BOD), Dissolved Oxygen (DO), and the associated Water Quality Index (WQI) of the Ganga river. Manual monitoring of the water quality by accessing all the water pollutants from various water sources is a complex, tedious, time-consuming, and risky job. Further, it is difficult to explain the overall quality of the water by using different water pollutants simultaneously as per the standards. Hence, the WQI measurement method has been preferred, which gives information about total water quality in a single value[2]. Moreover, forecasting techniques that use the historical water pollution data are beneficial in predicting future water quality. Recently, various time-series forecasting techniques using statistical approaches [3] and machine learning approaches [4], [5] have been proposed for the river Ganga. However, statistical approaches are lightweight and accurate for the structured water pollution data. Hence, this work intends to use statistical approaches. Recently, the statistical time series methods like Prophet, Seasonal ARIMA (SARIMA), and Auto-Regressive Integrated Moving Average (ARIMA) have been implemented to forecast the parameters like DO and BOD in river Ganga [3]. However, the interdependency of the various water quality parameters is not considered by these univariate time series forecasting techniques. Practically, different water quality parameters are highly correlated. For example, future values of DO depend on the past values of the DO and that of TC. The time series models which use the past values of other parameters in the prediction of one parameter are called multivariate models. Some of the multivariate models are Vector Auto Regression (VAR) [6], Vector Moving Average (VMA) [7], Vector Auto-Regressive and Moving Average (VARMA) [8]. VAR is one of the famous statistical models utilized to get the forecasted results of time series data. It can easily grasp the non-linear trend and seasonality of any time series data. This model is best suitable for multivariate time series data [6]. All the water quality variables are considered endogenous variables, but this model also can include the other exogenous factors of the input time series.

In this paper, we are discussing the implementation of three Multivariate models, namely VAR, VMA, and VARMA, to forecast the future pollutants in a portion of the Ganga river that flows in Uttar Pradesh (UP), India.

2 Problem Statement

Recently, the quality of water in the Ganga River has decreased because of large pollution in its basin. In this work, three widely used multivariate time-series statistical forecasting techniques such as VAR, VMA, and VARMA to forecast four crucial water quality parameters such as DO, BOD, FC, TC, and their associated WQI of the river Ganga. The forecasting water quality may be used by the concerned authorities to plan and execute various water pollution prevention as well as control strategies. Subsequently, this research work will help in controlling multiple polluted water-related diseases.

3 Proposed Methodology

Following essential steps are followed to develop three efficient multivariate statistical models like VAR, VMA, and VARMA for water quality forecasting of the river Ganga.

3.1 Data Collection and Preprocessing

Firstly, the water quality data in the Ganga river have been collected from the UP Pollution Control Board (UPPCB) [9] for experimental purposes. Four parameters: DO, BOD, TC, and FC, are collected from nine stations of UP, India, as indicated in Fig. 1. In order to make the data suitable for statistical models, the data is pre-processed before applying it in the model.

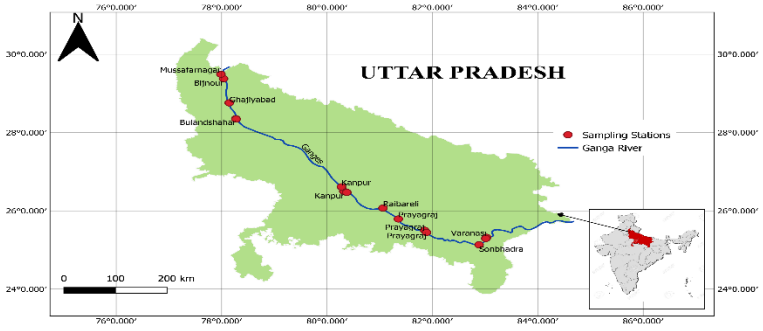


Fig. 1. The area under study.

3.2 Calculation of WQI

WQI is a single index that provides total information about water quality in a unique value. It can be determined using the parameters that are considered as a true value for the quality of that water body [10]. Here, four parameters, namely DO, BOD, FC, and

TC are used to calculate the value of WQI done using Eq. 1 [11]. The Q value in the range of 0 to 100 gives the normalized value of an individual parameter.

$$WQI = \sum_{j=1}^{M} W_j \times Q_j \tag{1}$$

Table 1. Parameter Weights for WQI calculation [12]

Parameters	**Moderate range**	**Weight Factor**
Dissolved Oxygen (mg/L)	4.5-5	0.17
BOD (mg/L)	3-5	0.11
Total coliform (mpn/100ml)	500-5000	0.18
Fecal coliform (mpn/100ml)	500-2500	0.16

Here, M = Total number of considered water pollutants, Q_j= sub-index of j^{th} pollutant, W_j=j^{th} pollutant's weight factor. The weight factors of each parameter that are used in the calculation of WQI are represented in Table 1.

3.3 Testing for Stationarity

Time series whose trend, seasonality, mean, and variance will not change with time is known as stationary. Further, the time series must be stationary for effective statistical modeling. Usually, an augmented dicky fuller test is used to check the stationarity of each parameter. The difference method is used to convert the nonstationary series into stationary.

3.4 Granger Causality Test

This test should be performed to find whether one-time series is useful in the prediction of other time series. The relationships among multiple variables during water quality forecasting, weather forecasting, stock analysis, anomaly detection, human action classification, etc., can be determined using the Granger causality [13].

3.5 Development of Multivariate Models

Time series analysis is an area of data analysis that works on processing, describing, and forecasting datasets that are time ordered. This time series analysis is broadly divided into univariate and multivariate analysis. The univariate models, in which the future values of individual parameters only depend on their own past values, are already discussed in [3]. In multivariate models, the future values of individual parameters depend on their own past values and the past values of their correlated parameters. The general flow diagram of multivariate statistical models is shown in Fig. 2. In this research work, three multivariate statistical models, namely VAR, VMA, and VARMA, have been implemented.

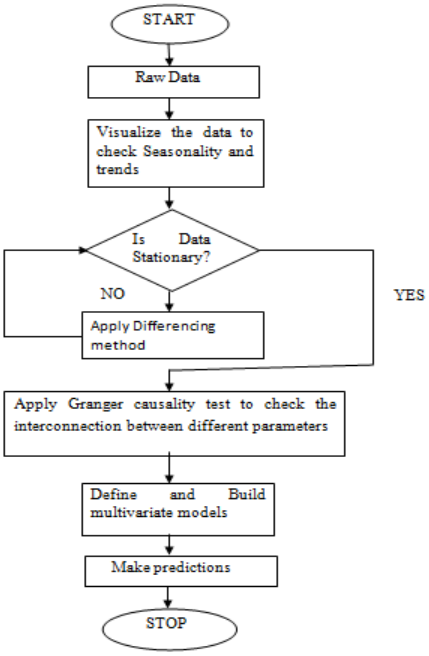
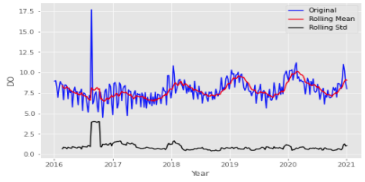


Fig. 2. General Flow chart multivariate models.

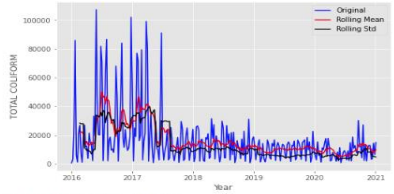
	BOD	DO	TOTAL COLIFORM	FECAL COLIFORM
count	1827.000000	1827.000000	1827.000000	1827.000000
mean	5.677701	7.849661	16439.609743	9996.061850
std	12.490381	2.418478	36876.924280	23572.841184
min	0.600000	0.200000	110.000000	70.000000
25%	2.490000	7.200000	2200.000000	1300.000000
50%	3.200000	7.900000	4700.000000	2700.000000
75%	4.100000	8.700000	17000.000000	10500.000000
max	268.000000	81.000000	350000.000000	220000.000000

Fig. 3. Exploratory data analysis.




```
ADF results for DO:
1. ADF: -3.2854175617050716
2. p-value: 0.015543027030050699
3. No. of lags: 11
4. No. of observations used for ADF regression: 250
5. critical values :
        1% :  -3.456780859712
        5% :  -2.873171506600003
        10% :  -2.572968544
```

Fig. 4. DO results for Dicky-Fuller test.




```
ADF results for TOTAL COLIFORM:
1. ADF: -1.819363073051007
2. p-value: 0.37089171820286465
3. No. of lags: 12
4. No. of observations used for ADF regression: 249
5. critical values :
        1% :  -3.4568881317725864
        5% :  -2.873218513016057
        10% :  -2.5729936189738876
```

Fig. 5. TC results for Dicky-Fuller test.

	BOD_X	DO_X	TOTAL COLIFORM_X	FECAL COLIFORM_X
BOD_y	1.0	0.0199	0.0000	0.0000
DO_y	0.0	1.0000	0.0001	0.0001
TOTAL COLIFORM_y	0.0	0.0130	1.0000	0.0086
FECAL COLIFORM_y	0.0	0.0235	0.0069	1.0000

Fig. 6. Granger causality test results.

VAR is an algorithm used to forecast the future time series values when two or more parameters influence each other [6] [8] [14]. Let z_1 and z_2 be two-time series that are correlated with each other, which are under study. Then to forecast z_1 and z_2 values at time t, the VAR algorithm uses past values of both z_1 and z_2. The equations for the VAR model are given in Eq. 2 and Eq. 3 [6].

$$z_{1,t} = \alpha_1 + \theta_{11,1} z_{1,t-1} + \theta_{12,1} z_{2,t-1} + \theta_{11,2} z_{1,t-2} + \theta_{12,2} z_{2,t-2} + \varepsilon_1 \quad (2)$$

$$z_{2,t} = \alpha_2 + \theta_{21,1} z_{1,t-1} + \theta_{22,1} z_{2,t-1} + \theta_{21,2} z_{1,t-2} + \theta_{22,2} z_{2,t-2} + \varepsilon_2 \quad (3)$$

The above equations show the prediction of z_1 and z_2 for two lags, i.e., $p = 2$ (where p is the number of lags selected). So, it is like VAR (2). The general form of *VAR(p)* for z_t of m stochastic time series like $z_t = [z_{1t}, z_{2t}, ..., z_{mt}]$ is given inEq.4.

$$(I_m + \theta_1 L + ... + \theta_p L^p) z_t = \varepsilon_t \quad (4)$$

Here, θ_i is the matrix of parameters, I_m is an identity matrix, L is a lag operator, and ε_t is a column vector.

VMA is similar to the moving average model used to forecast multivariate time series. In this model, the next step in the sequence is calculated by using the linear function of past residual errors [7]. Hence, VMA is also called as the model of residual errors.

$$z_{1,t} = \beta_1 + \theta_{11,1} \varepsilon_{1,t-1} + \theta_{12,1} \varepsilon_{2,t-1} + \theta_{11,2} \varepsilon_{1,t-2} + \theta_{12,2} \varepsilon_{1,t-2} \quad (5)$$

$$z_{2,t} = \beta_2 + \theta_{21,1} \varepsilon_{1,t-1} + \theta_{22,1} \varepsilon_{2,t-1} + \theta_{21,2} \varepsilon_{1,t-2} + \theta_{22,2} \varepsilon_{2,t-2} \quad (6)$$

Here, ε is the residual error. The equations Eq. 5 and Eq. 6 [7] are the VMA equations for two correlated time series z_1 and z_2, up to two moving average trends, i.e., $q = 2$. The ideal value of q for any data will be determined from the PAC plot of the individual series.

VARMA is the combination of VAR and VMA models [8]. Also, it is a generalized version of the ARMA model, which predicts forthcoming values of multivariate time series. This model takes values of 'p' and 'q' and is also able to work as a VAR by making 'q' as 0 and as VMA by making 'p' as 0. It has an advancement over individual VAR and VMA models. The general representation for *VARMA(p, q)* for z_t of m stochastic time series like $z_t = [z_{1t}, z_{2t}, ..., z_{mt}]$ is given in Eq.7 [8].

$$(I_m + C_1 L + ... + C_p L^p) z_t = (I_m + D_1 L + ... + D_q L^q) \varepsilon_t \quad (7)$$

where, I_m is an identity matrix with order m, L is a lag operator, ε_t is a column vector, and C, D are matrices of parameters [8].

3.6 Performance Measures

Here, the performance metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are taken to evaluate the accuracy of the multivariate models, as defined in Eq. 8 and Eq. 9 [15]. MAE is the average value calculated from the absolute differences between predicted and original values. In contrast, MSE is calculated as the

mean of squares of deviations from original to predicted values, and RMSE is the root value of MSE [8]. Here, error e_k is difference between predicted and original values for $k = 0,1,2,3...m$ [12]. Models having less values of RMSE and MAE are considered as best for the prediction.

$$MAE = \frac{1}{m}\sum_{k=1}^{m}|e_k| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{m}(\sum_{k=1}^{m}e_k^2)} \tag{9}$$

4 Results and Discussion

The results of the implemented multivariate time series models are briefly presented in this section.

4.1 Exploratory Data Analysis

It is a preprocessing step that is implemented to get the information of the data like standard deviation, mean, total count, minimum and maximum values of each parameter in the data, as shown in Fig. 3. Also, the backward filling method is used to fill the null values in the data.

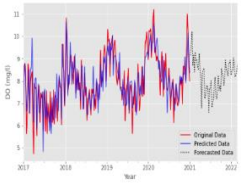
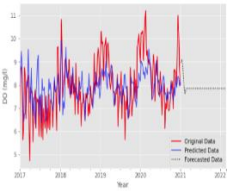
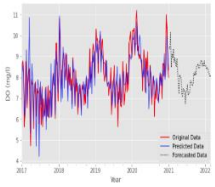
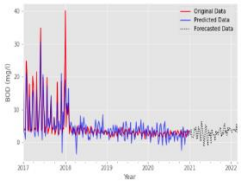
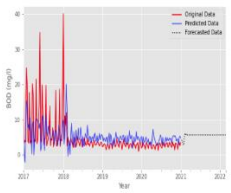
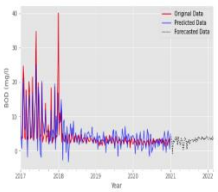
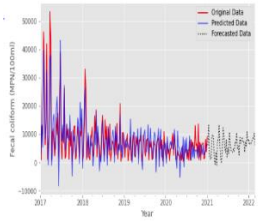
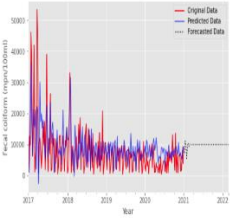
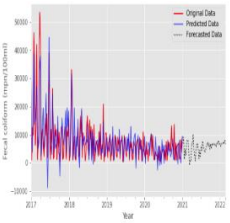
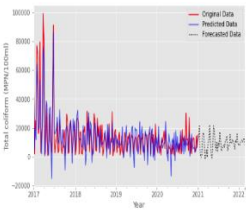
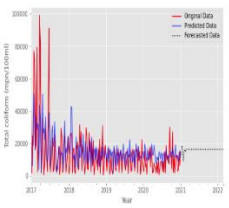
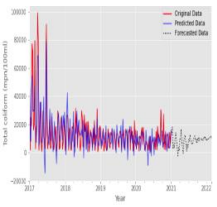
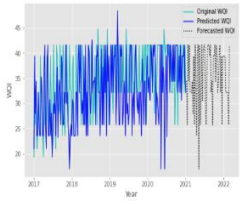
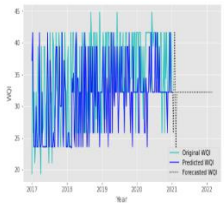
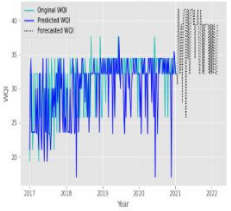
4.2 Stationary Test

Test for stationarity is conducted on the pre-processed data with the help of an augmented dicky fuller test. Also, the rolling mean and standard deviation are plotted as shown in Figs.4 and 5. According to the test results, only the DO series is stationary, with a *p*-value less than 0.05. But the time series corresponding with other parameters are not stationary as their p values are greater than 0.05. The nonstationary series are converted into stationary by the difference method.

4.3 Granger Causality Test

It is performed to know the correlation between the water quality parameters as shown in Fig. 6. Here, the columns are the predictors, and the rows are the responses with corresponding displayed *p* values selected in the test. If the *p* value is less than 0.05, then we can say that the corresponding row element granger causes the respective column element. For example, the value in the second row and third column, which is 0.0001<0.05, shows that the prediction of DO depends on the past values of TC. Hence, from Fig.6, it can be observed that all the parameters are granger-causing each other interchangeably.

Table 2. Prediction results of implemented multivariate time series models for various water quality parameters.

	VAR	VMA	VARMA
DO			
BOD			
FC			
TC			
WQI			

4.4 Model Recognition

Akaike Information Criteria (AIC) is used to select the lag order of p for the VAR model. Similarly, PAC plots are used to determine the order q value for VMA.

4.5 Forecasting Results of Multivariate Models

After completing the stationary test and the Granger causality test, the multivariate models like VAR, VMA, and VARMA are trained on weekly sampled Ganga river data set from 2016 to 2020. Then fitting of the models is performed from 2017 to 2020. Finally, forecasting of each pollutant is performed from January 2021 to February 2022. The forecasting results for the four pollutants and the corresponding WQIs are as shown in Table 2.

5 Comparative Analysis

In this section, a comparative analysis is carried out among the four implemented multivariate statistical time series models such as VAR, VMA, and VARMA as presented in Table 3. VAR and VARMA models are able to forecast for a longer duration. In contrast, the VMA model is only able to forecast for a shorter duration that is only two months ahead of the dataset.

Table 3. Comparison of the model performances.

Model	Perf. metric	DO	BOD	FC	TC	WQI
VAR	**MAE**	0.48	1.85	3502.17	6423.64	5.59
	RMSE	0.62	2.73	4452.3	8196.26	7.67
VMA	**MAE**	0.77	2.82	5230.62	8950.49	6.59
	RMSE	0.98	4.55	7195.89	9615.14	8.10
VARMA	**MAE**	0.53	2.16	3900.36	6842.40	3.12
	RMSE	0.68	3.22	5228.27	9399.45	4.70

6 Conclusion

In this study, three multivariate statistical models like VAR, VMA, and VARMA have been implemented to forecast the pollutants of river Ganga like DO, BOD, FC, and TC. It can be observed from Table 2 that the VAR model performs better in predicting DO, BOD, FC, and TC. But, VARMA model is performing best in the prediction of WQI values. The VAR and VARMA models are able to forecast for a longer duration, whereas VMA is only able to forecast for a short duration. This research work can be further augmented using deep learning and hybrid models. This work is helpful in reducing the number of polluted water-related diseases of humanity. Hence, it may help in reducing the burden of the health care system of the country.

References

1. A. N. Ahmed et al., “Machine learning methods for better water quality prediction,” J. Hydrol., vol. 578, p. 124084, 2019.
2. I. Ahmad and S. Chaurasia, “Water quality index of Ganga river at Kanpur (UP),” Themat. J Geogr, vol. 8, no. 11, pp. 66–77, 2019.
3. A. P. Kogekar, R. Nayak, and U. C. Pati, “Forecasting of Water Quality for the River Ganga using Univariate Time-series Models,” in 2021 8th International Conference on Smart Computing and Communications (ICSCC), 2021, pp. 52–57.
4. A. P. Kogekar, R. Nayak, and U. C. Pati, “A CNN-BiLSTM-SVR based Deep Hybrid Model for Water Quality Forecasting of the River Ganga,” in 2021 IEEE 18th India Council International Conference (INDICON), 2021, pp. 1–6.
5. A. P. Kogekar, R. Nayak, and U. C. Pati, “A CNN-GRU-SVR based Deep Hybrid Model for Water Quality Forecasting of the River Ganga,” in 2021 International Conference on Artificial Intelligence and Machine Vision (AIMV), 2021, pp. 1–6.
6. C. Y. Keng, F. P. Shan, K. Shimizu, T. Imoto, H. Lateh, and K. S. Peng, “Application of vector autoregressive model for rainfall and groundwater level analysis,” in AIP Conference Proceedings, 2017, vol. 1870, no. 1, p. 60013.
7. K. Hua and D. A. Simovici, “Long-lead term precipitation forecasting by hierarchical clustering-based bayesian structural vector autoregression,” in 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC), 2016, pp. 1–6.
8. S. S. Izquierdo, C. Hernández, and J. del Hoyo, “Forecasting VARMA processes using VAR models and subspace-based state space models,” 2006.
9. E. Center, “Water Quality Database.” Nov-2020.
10. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. Garc\’\ia-Nieto, “Efficient water quality prediction using supervised Machine Learning,” Water, vol. 11, no. 11, pp. 1–14, 2019.
11. S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, “Water quality assessment in terms of water quality index,” Am. J. water Resour., vol. 1, no. 3, pp. 34–38, 2013.
12. M. Kachroud, F. Trolard, M. Kefi, S. Jebari, and G. Bourrié, “Water quality indices: Challenges and application limits in the literature,” Water, vol. 11, no. 2, pp. 361–387, 2019.
13. D. Yang, H. Chen, Y. Song, and Z. Gong, “Granger causality for multivariate time series classification,” in 2017 IEEE international conference on big knowledge (ICBK), 2017, pp. 103–110.
14. C. M. Thasnimol and R. Rajathy, “Vector Error Correction Model for Distribution Dynamic State Estimation,” in Control Applications in Modern Power System, Springer, 2021, pp. 15–27.
15. K. K. R. Samal, K. S. Babu, S. K. Das, and A. Acharaya, “Time Series based Air Pollution Forecasting using SARIMA and Prophet Model,” in Proc. Int. Conf. on Information Technology and Computer Communications, 2019, pp. 80–85.