Predictive Water Quality Modelling Using ARIMA and Water Parameter Forecasting Model (WPFM) for Godavari River, Maharashtra, India

Sucheta Sable/Kakde* and Rajesh Kherde†
D.Y. Patil University, Ambi, Pune, INDIA

Abstract

The Godavari River in Maharashtra, India, is used as an example in this paper to demonstrate the recital use of machine learning methods, including auto regression with mean average and random forest regression. Water data are gathered for modelling from the Hydrological Data Users Group in Nasik, Maharashtra. The procedure used in this article demonstrated tension during the creation of Python code and its transfer to the operation data to predict output. Comparing the two models, ARIMA is used to forecast the places' subsequent six progressive values. After comparing the models' outputs, the RSME of Water Parameter Forecasting Model had a score of 0.90, while the ARIMA received less than 0.5, it was declared to be a reliable model for forecasting the following six values. In order to compare the predicted and generated results, samples from the relevant study sites are gathered and tested in the lab concurrently with the machine learning process.

Key Words: ARIMA, water parameter forecasting model, Jupyter notebook, Godavari river.

1 Introduction

The majority of nations recognize surface water quality as a delicate and important problem. The effect of surface and groundwater quality on human health, aquatic life concerns, and related factors is significant. [12] Water has a major impact on life sustainability which is determined as a key element in our environment. Water found in sea and land plays an important part in day to day life activities such as drinking, agriculture, industrial and other uses [14]. Ground and surface water quality is declining as a result of human-made activities like industrial refuse, agri-/aquaculture, and discharges from other uses [2]. Hydro chemical property analysis is a key component in identifying the quality of water for domestic, commercial, and irrigation uses. Remote sensing, GIS, and statistical analytical tools were used to identify the variables. It regulates the flow system at various Wheeler Lake Basin sites in Northern Alabama. The research evaluated ground and surface water after human consumption by using the water quality index (WQI) method. In various locations around the globe, numerous investigations are carried out to address quality standards and issues. With the aid of engineering studies on rivers, a number of investigations into earth science is the evaluation of water quality parameters. Water found in sea was used to evaluate the status of concerned locations. The most frequent occurrence when it comes to earth science is the evaluation of water quality parameters, water quality, the transport of sediment, and the transmission of pollutants. Human socio-economic growth depends upon readiness of quality water. [11] A rapid increasing in population and thus expansion in agriculture and industries has shown us how that quality water is becoming difficult to achieve. [5] Additionally, measuring is divided into two categories: components of water quality and the spread of pollutants and their mechanisms. Physicochemical analysis was used to classify the water quality parameters as a result of environmental engineering breakthroughs. BOD, COD, pH, temperature, K, Mg, Na, TDS, and other assays are among them [1]. Departments of the federal, state, and local governments as well as the Hydrological Data Users Group are mandated to routinely assess the water quality parameters.

Additionally, the station points will contain fundamental data for creating conservation initiatives. The availability of water quality parameters made it possible to review with the aid of an earlier study [13]. The benefit of using soft computing methods and having access to time series analysis is that the majority of engineering researchers today have conducted the necessary research to make accurate predictions of future quality parameters using mathematical formulas [6]. The MLP model has accurate findings in predicting the parameters when using adaptive neuro-fuzzy inference system (ANFIS), radial basis network (RBF), and multilayer perception (MLP) methods to forecast the water quality parameters [3]. It was suggested to use a model to plan the spread of classified surface water in accordance with the quality of Iran's locations using probabilistic support vector machines (PSVMs) and GIS techniques [10]. Numerous case studies were used to predict the water quality parameters in a different research [4]. The prediction of quality factors with internal relationships used time series analysis and statistical models. Numerous research papers make use of the quality evaluation of water and predictive analysis for the planning of projects related to water conservation. The Godavari River in Maharashtra, which has been identified as the main contributor to the area's surrounding canals, was the subject of the literature reviews used in the development of the current paper's water quality model. Both the ARIMA and the Water Parameter Forecasting Model are used for parameter prediction within the chosen region. Additionally, Jupyter notebook and scripts are created and included to manage the study's data.

* PhD Research Scholar, Department of Civil Engineering.
† Professor, Department of Civil Engineering.

1.1 Proposed Methods and Materials

This paper uses machine learning methods to suggest the distinctive structured code in the Jupyter notebook. There are two program models included: ARIMA and Water Parameter Forecasting Model. The introduction to the study area is followed by a presentation of the periodic water quality statistics and their ranges. The overview of the used time series models is then presented, and the obtained findings are contrasted.

2 Study Area and Data Collection

The Godavari waterway, which has its source in Trimbakeswar, Nasik, Maharashtra, India, is the second-largest waterway in that country. It flows through the regions of Andhra Pradesh, Madhya Pradesh, Karnataka, and Orissa. The waterway, which flows through Nasik City, is 82% domestically and 18% industrially polluted. The study spans a distance of 350km along the river, beginning in Kushawart Trimbakeswar and ending in Saikheda Village, where the river joins the city. The river's water was sampled from ten different places, and the samples were then examined in a lab run by the Hydrological Data Users Group for indicators of water quality.

The information on water quality is gathered from the Hydrological Data Users Group, Nasik (HDUG) database in India, which is maintained by the authorities who are in charge of overseeing the various aspects of water quality. From 2011 to 2021, a total of 10 years are used. The statistics for the monthly and yearly averages were only available in a few places. The four chosen sites are shown in Figure 1 along with the parameters for the water quality. Regarding the Hydrological Data Users Group, Table 1 presents an overview of the elements that make up water quality. BOD, COD, pH, temperature, K, Mg, Na, TDS, and DO are the parameters this research takes into account.

3 Predictive Modelling

Making predictions based on historical data can undoubtedly assist in resource management and raise the standard of the water supplied to the community.

The dependent variable's future value can be predicted using a unique set of techniques and methods that are accessible in the machine learning domain [15]. The ideas of the ARIMA model, which capture the essence of time series analysis, are used in this paper. The back end date given must be checked with the various types of components, such as trend, noise, and seasonality, in order to run the ARIMA model. The time periods are listed and depicted in Figure 2. The data must be stationary according to the statistical model for a time series collection. The following list includes the codes that are necessary to complete the time series analysis and statistical model.

A specified "excel" file containing the data has been imported into the Jupyter notebook, and the command window appears as shown in Figure 2. Making a good forecast model requires making sure the time series data is stationary.

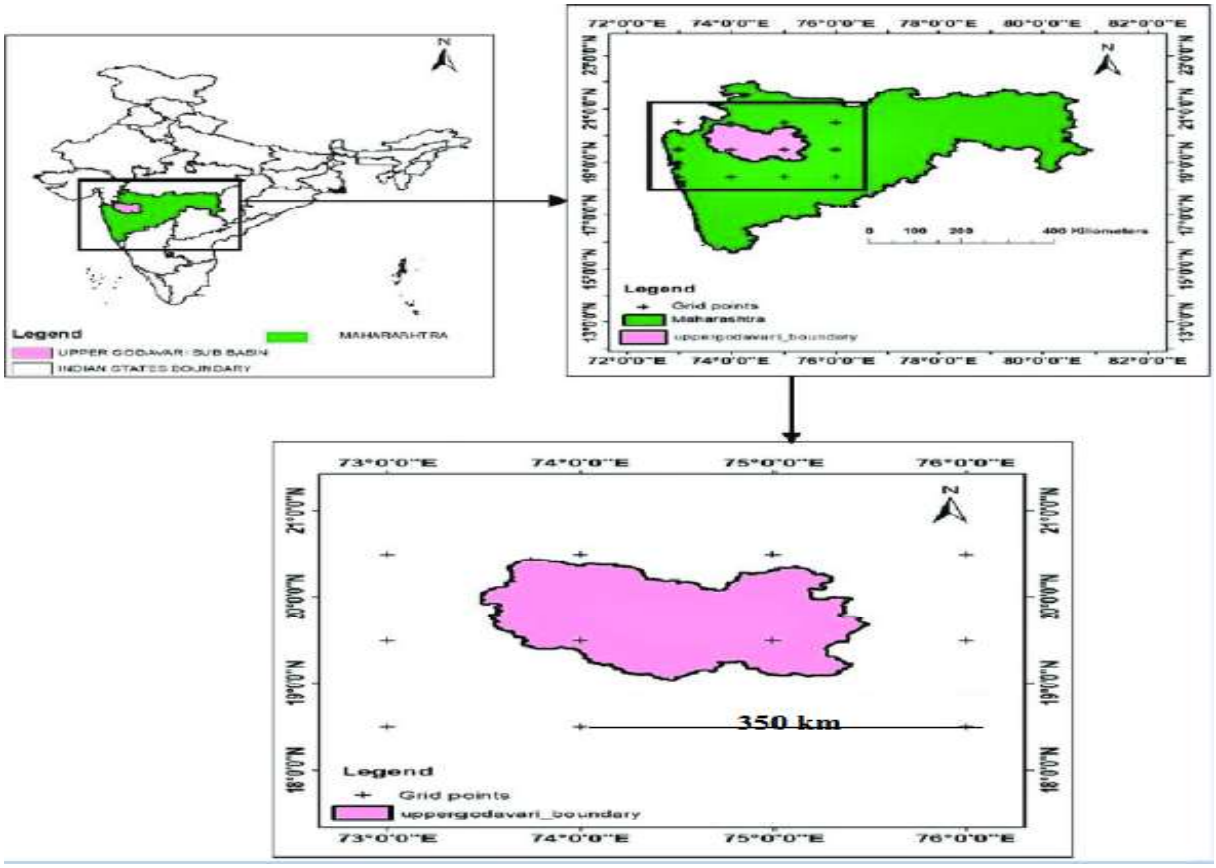


Figure 1: Location of the study area

Table 1: The summary of water quality parameters by HDUG at Godavari River

Annual Average Values	BOD	COD	DO	Ph	Temp	SS	K	TDS
Date	Average							
2012	18.9	24.4	1.8	7.28	25.7	0.65	2.774286	344.6
2013	21.1	21.6	2.5	7.24	25.7	0.59	2.685714	333.6
2014	17.4	23.4	3.3	7.25	25.7	0.54	2.315714	337.9
2015	16	25.4	3.3	7.24	25.8	0.5	2.128571	340.6
2016	17.6	28.1	3.9	7.23	25.8	0.47	1.957143	345.6
2017	19.1	29.3	4.3	7.25	25.7	0.51	1.471429	340.3
2018	19.8	30.3	3.8	7.26	25.7	0.49	1.071429	319.3
2019	20	29.9	3.7	7.27	25.6	0.48	1.342857	323.2

The given time sequence's future values, after being evaluated by its past and current statistical values, can be predicted using the ARIMA model. The ARIMA model can be divided into three categories, with "AR" standing for autoregressive, which takes prior values into account when a variable is evolving. While "I" reverses the integration part of data values evolved under the process of differentiation between the values of present and past, "MA" showed the linear combination of errors and values with a moving average part of regression at various times of past values. In the end, the ARIMA model's features can help the data match the range shown in Tables 1 and 2.

3.1 Model Determination

Mushtaq analyzed the data using the rolling statistics and augmented the Dick fuller test (ADF) [9]. Further, non-stationary data, an autocorrelation plot with decay shall be viewed in the window. Analysis of the excel data of different parameters was done to understand the pattern present between data.

The following algorithm is suggested for second-order differencing based on the outcome produced following the first-order differencing of non-stationery data:

3.2 Parameter Estimation and Analysis Using Model

The regression model is created by averaging the parameter data and using the date "x" as an independent variable and the parameter "y" as a dependent variable.

The Autoregressive Integrated Moving Average model, or ARIMA model, is a well-liked time series forecasting technique that incorporates moving average, autoregressive, and differencing components.

The general formula for an ARIMA(p,d,q) model is:

$$Y_t = c + \phi_1(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) - \theta_1\varepsilon_{t-1} - \ldots - \theta_q\varepsilon_{t-q} + \varepsilon_t$$

Where:

- Y_t is the value of the time series at time t
- c is a constant term (the intercept)
- ϕ_1, ϕ_p are the autoregressive coefficients (AR terms) for lags 1 to p
- Y_{t-1}, Y_{t-p} are the lagged values of the time series
- μ is the mean of the time series
- $\theta_1, \ldots, \theta_q$ are the moving average coefficients (MA terms) for lags 1 to q
- $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are the lagged errors or residuals
- ε_t is the error term or residual at time t
- is the degree of differencing, which is the number of times the series is differenced to achieve stationary.

3.3 ARIMA Determination Step

The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series forecasting technique used to analyse and predict data that exhibits a trend, seasonality, or other patterns. The process of determining the ARIMA model for a given time series involves the following steps:

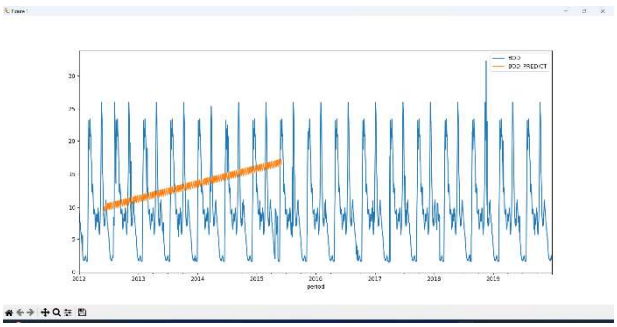
1. Visualize the data: Plot the time series and examine its behavior. Look for patterns, trends, and seasonality.
2. Stationarity check: A stationary time series has constant mean and variance over time and is necessary for ARIMA modelling. Check for stationarity using statistical tests like Augmented Dickey-Fuller (ADF) or KPSS test. If the series is not stationary, apply transformations like differencing or seasonal differencing.
3. Identify the order of differencing: If the data is not stationary, apply differencing until it becomes stationary. The order of differencing is the number of times the data is differenced. The order of differencing can be identified visually or using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.
4. Determine the order of Autoregressive (AR) term: The AR term is the number of lagged values of the time series that are used to predict the current value. The order of the AR term can be identified using the PACF plot.
5. Determine the order of Moving Average (MA) term: The MA term is the number of lagged forecast errors used to predict the current value. The order of the MA term can be identified using the ACF plot.
6. Model selection: Combine the identified orders of differencing, AR, and MA terms to form a candidate ARIMA model. Select the best model based on statistical measures such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), or

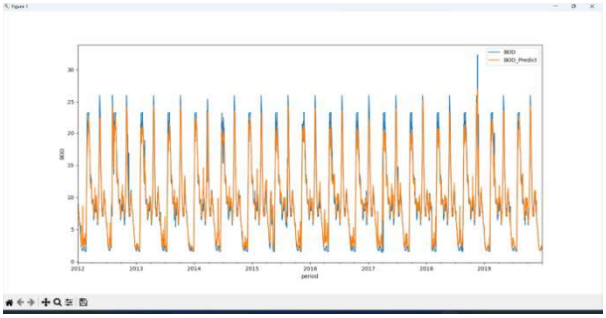
MSE (Mean Squared Error) on a validation set [7].

7. Model validation: Validate the selected model using statistical tests and plots, such as residual analysis, Q-Q plots, and Ljung-Box tests.
8. Forecasting: Use the selected and validated ARIMA model to make predictions on future values of the time series.

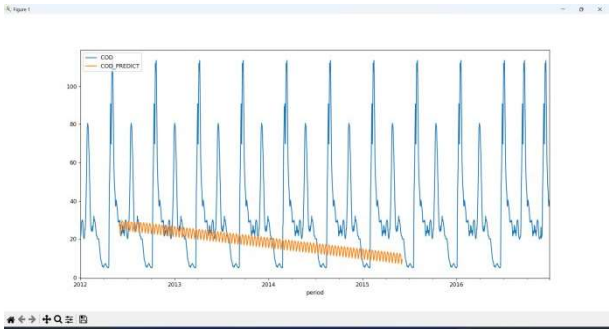
ARIMA **WPFM**

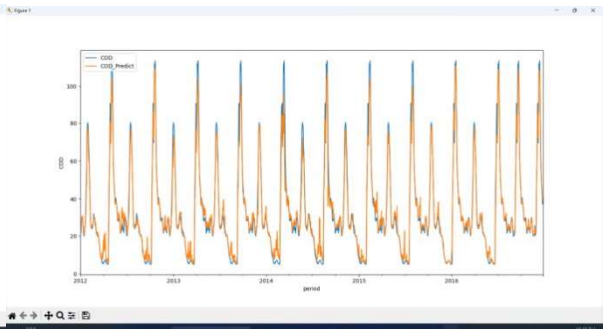
BOD



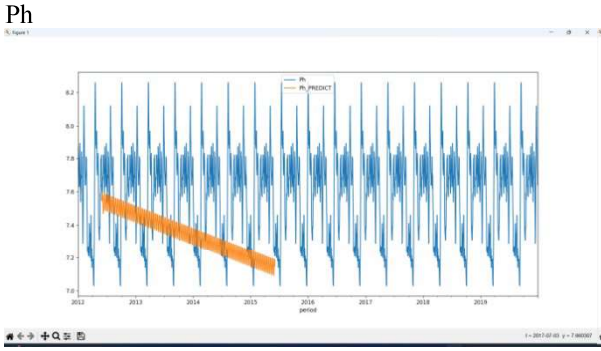


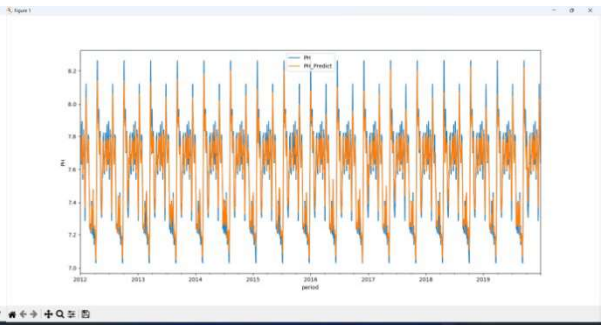
COD



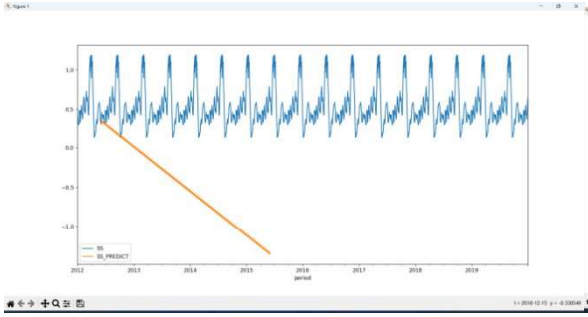


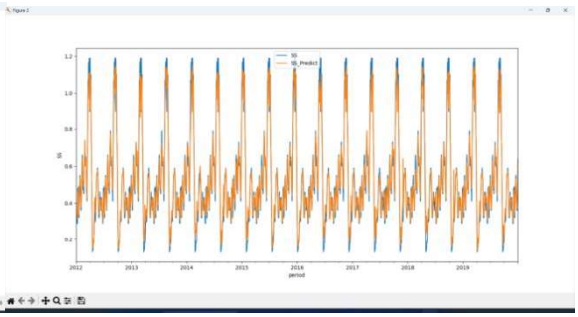
Ph



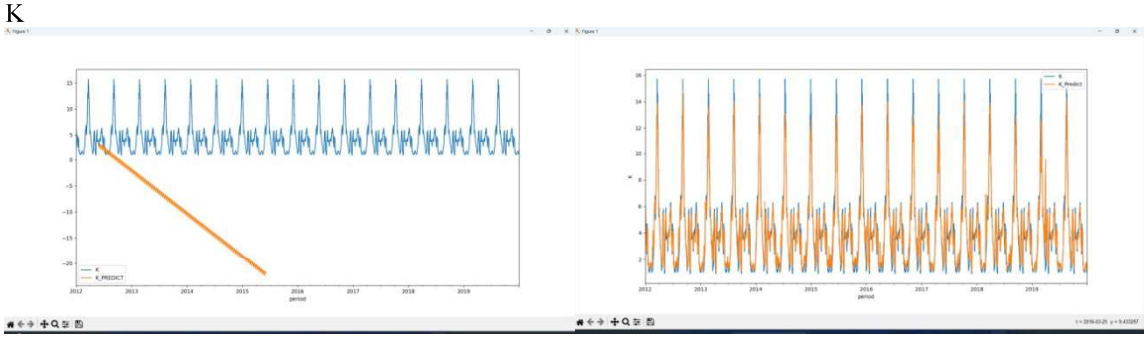


SS

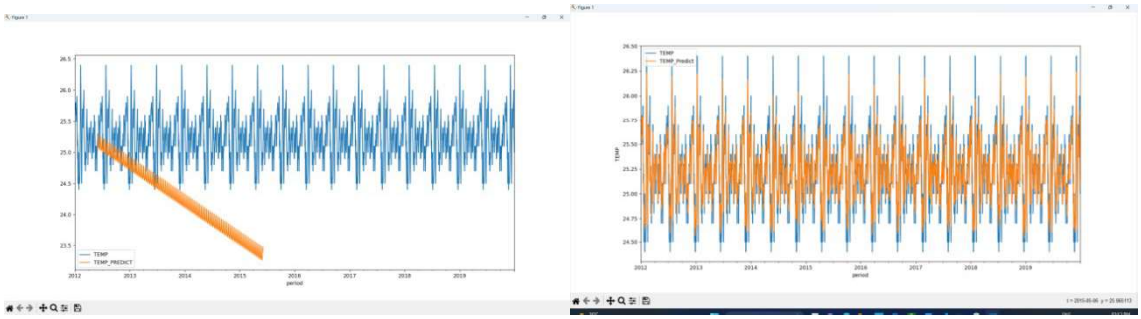




K



TEMP



DO

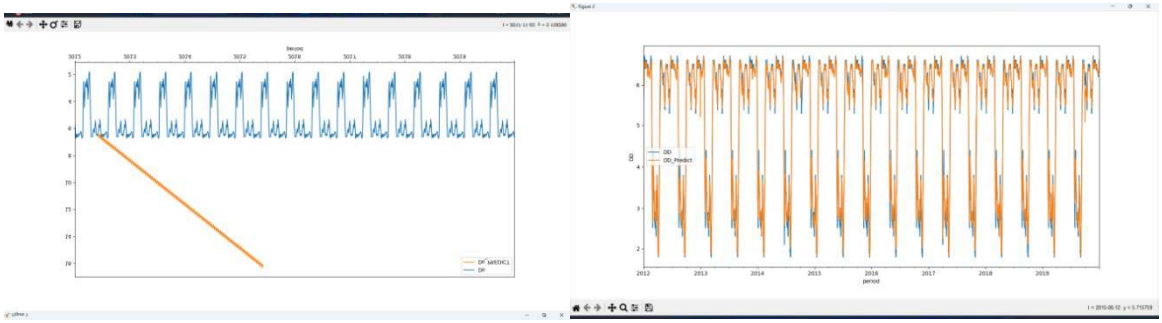


Figure 2: Predictive modelling using water parameter forecasting model

3.4 Water Parameter Forecasting Model (WPFM)

WPF model is used as a random regression method, which is based on the generate tree structure to estimate the parameter prediction. The process of random forest regression entails building several decision trees, which are then combined to produce a model that is more reliable and precise. A random subset of the accessible data and features is chosen to create each decision tree [8]. The random pick contributes to improving the model's generalizability and decreasing over fitting. The regression model is used in statistical analysis to forecast the parameters "*y*". It uses past time series data to forecast the output. Random forest regression model it is an ensemble technique to take the sum of the all the estimated output.

3.5 Model Determination

$$X_{newt} = X_t + X_{t+7} \quad (1)$$

To find the output by Random Forest Regression

$$\text{Standard Deviation} = \sum (X\text{-}X_{MEAN})/N \quad (2)$$

Find the probability of each attribute

$$\text{Probability Distribution} = \sum P(X) \times \text{Standard Deviation}(X) \quad (3)$$

3.6 Main Library:

From Sklearn.ensemble.Random Forest.
Pandas
numpy
Sklearn.model-selection
matplotlib.

3.7 Regression Tree Generation using Machine Learning (ML) Involves the Following Steps

1. Data preparation: Collect and pre-process the data, which involves cleaning, transforming, and normalizing

the data. Split the data into training and testing sets.
2. Feature selection: Select the most important features that are relevant to the prediction task. Feature selection can be done using statistical tests, domain knowledge, or ML algorithms.
3. Tree building: The tree building process starts with the selection of the root node. The root node is selected based on the feature that maximizes the difference between the target variable and the feature values. The data is then partitioned into two subsets based on the value of the selected feature. This process is recursively applied to each subset until a stopping criterion is met.
4. Stopping criterion: The stopping criterion determines when to stop partitioning the data into subsets. The stopping criterion can be based on the maximum depth of the tree, the minimum number of samples in a leaf node, or the minimum reduction in the variance of the target variable.
5. Tree pruning: Tree pruning is a process of removing unnecessary branches from the tree to prevent over fitting. Over fitting occurs when the model is too complex and captures noise in the training data.
6. Model validation: Validate the model on the testing set using statistical measures such as Mean Squared Error (MSE), R-squared, or Root Mean Squared Error (RMSE).
7. Hyperparameter tuning: Hyperparameters are parameters that are not learned from the data but are set before the training process. Examples of hyperparameters include the maximum depth of the tree, the minimum number of samples in a leaf node, and the learning rate. Hyperparameter tuning involves selecting the best hyperparameters that optimize the performance of the model.
8. Prediction: Once the model is trained and validated, use it to make predictions on new data.

3.8 Model Development

The first stage in putting machine learning models into practice is the preparation of the data set. The obtained data collection should now be divided into two categories: training the input data and testing. The validation and evaluation of models that have been applied use training and trial data sets, respectively. The process of allocating a subgroup to each category differs depending on the parameter values for time series modelling. In time series, it is best to assume that the history of data gathering has been modelled, and shuffling the data set is not appropriate, though it is acceptable when a feature is present.

Usually in both situations, approximately 75-85% of the data are set aside for confirmation and the resulting 20-30% for confirmation. The implementation of machine learning models, including the tested and best-chosen ARIMA and Water Parameter Forecasting Model, is the following step. The network's capacity needs to be adjusted in the following phase to improve the consistency of the Predictive Water Quality Modelling Using ARIMA and Water Parameter Forecasting Model for Locations of Godavari River. In order to accomplish this, more neurons or secret layers will be added. This method's final two stages also cover the creation of the Initiate software architecture.

4 Results and Discussion

The ARIMA and Water Parameter Forecasting Model are both described in this paper along with their use in predicting the outcomes of water tests for a few water components. An optimal model created in this paper has experienced analysis, and in the later stages, predictions have been made using various time series analysis methods. Two-time series analysis models were modified from separate models, and new code was generated to fit the ARIMA (autoregressive integrated moving average) and Water Parameter Forecasting Model. These two models' various parameters, including temperature, pH, NA, K, COD, BOD, dissolved oxygen, and TDS, are chosen based on historical data from government organizations. In the same way that the predictive values for the year 2030 were predicted, the total data values from the years 2012 to 2019 have been gathered and evaluated. The time series deflection for the predictive values could be seen for both models, which had both done well.

Following the completion of these operations, a number of observations are made to assume the predicted values can be discovered, and the best model that was predicted is then chosen. Personal interest can be used to determine the number of observations, but more backend data can actually help the algorithm perform better. Given that there will be 100 observations in this specific value, the data is split into two sets: one for training and the other for testing. Based on test data values from the two models, ARIMA and Water Parameter Forecasting Model, the projected values will be obtained after the test data; we obtain the predicted values after the test data. We will then plot the graphs for the test data and the predicted numbers after making our predictions.

In Tables 2 and 3, the Water Parameter Forecasting Model achieved 95% accuracy and the RSME of the predicted data for both models was met at 85% accuracy for the ARIMA-model.

Table 2: ARIMA

Performance Measure	BOD	COD	Do	Ph	Temp	SS	TDS	K
RMSE	3.4	20.4	1.34	2.23	11.3	5.65	12.3	5.22
R2_Score	.49	.5	.56	.55	.44	.25	.55	.5

Table 3: Water parameter forecasting model

Performance Measure	BOD	COD	Do	Ph	Temp	SS	TDS	K
RMSE	0.89	1.85	.5	.65	.33	.45	3.233	2.1
R2_Score	.92	.89	.925	.94	.88	.85	.89	.88

5 Conclusion

The model's outcomes in this article were based on a mathematical approach. Utilizing Jupyuter Note, statistical formulas, the code necessary for parameter analysis is imported from the data library, and both the data and the code are changed in accordance with the time series. In the latter part, the predicted values had also met the R2 values, which range from 0.92 to 0.97. As we look at the graphs, the initial values are shown in blue, and predicted values are coded in orange. The seasonality of data sets is thoroughly examined in detail. Data from the last 8 years' worth of time series are seen to be nonstationary. Thus, it appears that the graph generated by the Python-based Jupyter notebook is also nonstationary.

Water Parameter Forecasting Model forecasts values that are closer to the original values than the ARIMA model, when the two models are compared. Compared to the ARIMA model of RSME values, the RSME & R2 values in the Water Parameter Forecasting Model are more comparable. The values discovered through data analysis are shown in the chart below for both the ARIMA and Water Parameter Forecasting Model.

References

[1] M. B. Assad and R. Kiczales, "Deep Biomedical Image Classification using Diagonal Bilinear Interpolation and Residual Network," *Int. J. Intell. Netw.* 1:148-156, 2020.

[2] Ajay D. Chavhan, M. P. Sharma and Renu Bhargava, "Water Quality Assessment of the Godavari River," *Hydro, Nepal*, 5:31- 35, 2009.

[3] S. Emamgholizadeh, H. Kashi, I. Marofpoor, and E. Zalaghi, "Prediction of Water Quality Parameters of Karoon River (Iran) by Artificial Intelligence-Based Models," *Int. J. Environ. Sci.Technol*. 11(3):645-656, 2013.

[4] S. Heddam, "Generalized Regression Neural Network Based Approach as a New Tool for Predicting Total Dissolved Gas (TDG) Downstream of Spillways of Dams: A Case Study of Columbia River Basin Dams, *USA, Environ. Process*, 4(1):235-253, 2016.

[5] C. T. Hunsaker and D. A. Levine, "Hierarchical Approaches to the Study of Water Quality in Rivers—Spatial Scale and Terrestrial Processes are Important in Developing Models to Translate Research Results to Management Practices," *Bio Science*, 45(3):193-203, 1995.

[6] N. S. Jaddi and S. Abdullah, "A Cooperative-Competitive Master-Slave Global-Best Harmony Search for ANN Optimization and Water-Quality Prediction, *Appl. Soft Compute*, 51:209-224, 2017.

[7] L. Kilian, "Small-Sample Confidence Intervals for Impulse Response Functions," The Review of Economics and Statistics, *MIT Press*, 80(2):218-230, May 1998.

[8] M. Maleki and S. M. Kashefpour, "Application of Numerical Modelling for Solution of Flow Equations and Estimation of Water Quality Pollutants in Rivers (Case study: Karkheh River)," *Civil Environ. Eng.*, 42.3(68):51-60, 2012.

[9] R. Mushtaq, "Augmented Dickey Fuller Test, [online] papers.ssrn.com. Available at: https:// ssrn.com/abstract=1911068, 2011.

[10] M. R. Nikoo, N. Mahjouri, "Water Quality Zoning using Probabilistic Support Vector Machines and Self-Organizing Maps, *Water Resour. Manag.*, 27(7):2577-2594, 2013.

[11] C. C. Obropta, M. Niazi, and J. S. Kardos, "Application of an Environmental Decision Support System to a Water Quality Trading Program Affected by Surface Water Diversions," *Environmental Management*, 42:946-956, 2008.

[12] Y. Ouyang, "Evaluation of River Water Quality Monitoring Stations by Principal Component Analysis," *Water Res*., 39(12):2621-2635, 2005.

[13] A. Parsaie, A. H. Haghiabi, M. Saneie, and H. Torabi, "Applications of Soft Computing Techniques for Prediction of Energy Dissipation on Stepped Spillways," *Neural Compute. Appl*., 29:1393-1409, 2018.

[14] Sucheta Sable/Kakde, Rajesh Kherde, and Gauri Patil, "A Review for Water Quality Modelling for a River Basin, The Seybold Report, 17(9):1410-1420, ISSN No-1533-9211, DOI 10.5281/zenodo.7115974, 2022.

[15] Purushottam R. Sarda and Parag Sadgir "Water Quality Modelling of Godavari River, India using Q2kw Soft Tool," Conference: India Water Week 2015: At: Delhi, India, January 2015.

Sucheta Sable/Kakde completed B.E.in civil engineering and ME in water resources engineering. Currently Pursuing Ph.d in the water resources engineering. I am currently working as an assistant professor in Engineering Collage and I am passionate about my job. Email: suchetasable17@gmail.com. Phone no.-+917338308858

Rajesh Kherde (photo not available) is the Principal at D. Y. Patil School of Engineering & Technology, Ambi, Pune. He has a Ph.D. in Civil Engineering with specialization in Water Recourses Management. Prof. Dr. Rajesh Kherde is an academician with Ph.D. in Civil Engineering from Mumbai University and a Gold Medallist of Gujrat University in a Master's study. He has 23 years of experience in the capacity of Principal, Professor and Head of the department for several years. He has worked as BOS member of civil engineering course in faculty of engineering at Sandip University, Nashik. He has successfully handled several responsibilities like NAAC coordinator, IQAC coordinator, Admission in charge and vice chancellor nominee as an observer for admission process.