National College of Ireland

Prediction of Suspended Particulate Matter Using Machine Learning

MSc Research Project
Data Analytics

Vinayak Vishnu Kolekar
Student ID: x18185797

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vinayak Vishnu Kolekar
Student ID:	x18185797
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	28/09/2020
Project Title:	Prediction of Suspended Particulate Matter Using Machine Learning
Word Count:	7751
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard speci ed in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	27th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not su cient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator o ce must be placed into the assignment box located outside the o ce.

O ce Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Suspended Particulate Matter Using Machine Learning

Vinayak Vishnu Kolekar
x18185797

Abstract

The availability of freshwater is an essential element for any living being, and its adequate quality is a necessary feature for water-borne disease prevention and improving quality of life. The rapid growth of industrialization and abrupt environmental changes establish emergencies in the control and puri cation of water quality. Therefore drinking water quality monitoring and forecasting is the most important task for water supply management to keep human population healthy. The researchers have contributed to the early identi cation system with the help of machine learning due to the demand for water quality prediction. The proposed study is about the identi cation of suspended particles by forecasting of water quality parameters such as turbidity and pH level along with precipitation as an external physical feature. The impact relationship between them is veri ed with the Johansen Cointegration. The comparative model analysis demonstrated for water parameter forecasting consists of Vector Autoregression (VAR), Vector Error Correction Model (VECM), Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX) and Long Short Term Memory (LSTM). The comparative analysis based on RMSE and MAPE illustrates LSTM performance is better than other autoregressive models.

Keywords— Turbidity, pH Level, ARIMAX, VAR, VECM, LSTM

1 Introduction

The water resources on earth, such as surface water and groundwater, are valuable natural assets and play a signi cant integral part in our society in terms of our health, well-being, economic, etc. The water has a considerable presence and vital need for the body, as it is the main blood component that carries the essential nutrients. Also, it helps to regulate the digestive system, body temperature, and waste disposal in the human body (Jalal and Ezzedine; 2019). Clean water that is free from contaminants has a critical role to play as an ingredient and sanitation in the food industry. Hospitality is also one of the primary sectors where clean water is a de ning part of daily operation and growth. At the same time, the waterborne disease poses a risk to the lifestyle and growth of human beings. Rapidly growing industrialization and natural activities such as acidic rain, type of soil are external factors that contribute to the increase in suspended particles in water. An increase in suspended particles shows the water is probably not acceptable to drink as it is contaminated. Because of this, regular monitoring, quality measuring, and water treatment are necessary for drinking water suppliers to provide safe drinkable water.

According to WHO the better water supply management and sanitation can boost a country's economic growth and helps signi cantly to reduce poverty[1].

The environmental protection agency (EPA) of Ireland circulates drinking water quality guidelines so that drinking water does not a ect human health. These guidelines help water suppliers to take corrective action by identifying substances that increase in water quality measures quantity. Changes in these measures beyond an acceptable range lead to the spread of waterborne disease. Even after successfully monitoring the water quality, there are several boiled water notices placed in Ireland by the water suppliers and suggested to boil the water before consuming it. Due to this, people faced many issues, such as loss in the hospitality domain and su ered from illnesses. The reasons behind putting notices are heavy rainfall and faults in the lter at a water treatment plant (Power; 2019; Carswell; 2019). This means that there is a gap in the monitored data analysis process, and which is not allowing us to identify the exact issue on time. So monitoring is not the only solution to the problems encountered by the increase in suspended particles in water. Water has already distributed in the network until the treatment plant operators react over the alarm of the monitoring system. Therefore, there is a need for an early identi cation system that identi es a change in the quality of water, which makes monitoring systems more robust and able to cope with abrupt events. Such events a change in water quality parameters from the ideal range and can be treated early basis at the treatment plant for sustainable water supply. This early prediction helps to make safely managed drinking water services and save humans from getting ill with waterborne disease and also hospitality and the food industry from loss.

The quality measure mainly consists of physical, chemical, and microbiological parameters such as pH level, Turbidity, Color, Temperature, Dissolved Oxygen (DO), E Coli, etc. (Kale; 2016). The sudden change in natural processes such as heavy rainfall, human activities, and natural parts such as type of soil are the reasons behind the change in pH level and turbidity. Many boiled water notices are issued in di erent counties of Ireland due to heavy rainfall or technical defects at treatment plants. Therefore, This project focused on forecasting of water quality parameters such as pH level and turbidity to early identify the chemical concentration and cloudiness of water. The state of art is amount of rainfall considered as an external factor for forecasting pH level and turbidity. An increase in the level of turbidity from acceptable level (1 NTU) shows problems with the e ciency of the water treatment process. In contrast, uctuation in pH level from ideal range (i.e. 6.5-9.5) states in uences the disinfection process and the concentration of metals in water. Forecasting these quality measures will provide early alerts to take preliminary actions to avoid problems caused by contaminated water.

This research aims to give early identi cation of water quality parameter changes with the help of statistical learning and machine learning. Section 2 provides brief information about researchers' contributions to water quality parameter forecasting with the help of time series analysis. Section 3 and 4 describes the methodology implemented and design speci cations for this research. The implementation part of the research project illustrated in section 5 includes data collection, data pre-processing, and transformation and model building, and evaluation of models. Section 6 is about model comparison and discussion, whereas the conclusion and future work are in section 7.

[1] https://www.who.int/en/news-room/fact-sheets/detail/drinking-water

1.1 Research Question and Objectives

How well can statistical and machine learning models predict suspended particulate matter in water with the help of drinking water monitoring and rainfall data?

Prepare data for model prediction by pre-processing on data loaded from the Environmental Protection Agency of Ireland and New York open data, along with the amount of rainfall in the eld of water supply.

Implementation of statistical tests to verify the e ectiveness of external factors such as rainfall on in the forecasting of water quality parameters.

Implementation of Vector Autoregression (VAR), Vector Error Correction Model (VECM), Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX) and LSTM.

Evaluation of implemented models with the help of root means square error (RMSE), mean square error(MSE), mean absolute error(MAE), and mean absolute percentage error (MAPE).

A comparison of implemented forecasting models on statistical measures such as root means square error (RMSE) and Mean absolute percentage error (MAPE).

1.2 Abbreviations

Parameters	Abbreviation
Biochemical or biological oxygen demand	*BOD*
Chemical oxygen demand	*COD*
Chemical oxygen demand in manganese	*CODmn*
Dissolved oxygen	*DO*
Electrical conductivity	*EC*
Hydrogen iron concentration (pH Level)	*pH*
Total organic carbon	*TOC*
Total phosphorus	*TP*
Trophic level index	*TLI*

Table 1: Abbreviations of water quality parameters

2 Related Work

2.1 Real-time forecasting of water quality parameters

Precise water quality parameter forecasting will promote advanced identi cation of water contamination as well as decision making on control of water supply. Deng et al. (2015)

proposed a novel multi-factor time series hybrid method for water quality prediction with the help of a cloud model and fuzzy forecasting. To reduce the uncertainty in water quality data, the author has used the Heuristic Gaussian cloud transformation algorithm, which extracts the uncertain part of data by calculating the periodicity and minimize noise. The water quality parameters considered for forecasting are Dissolved oxygen, CODmn, water temperature, and EC. The evaluation is done based on Mean square error, mean absolute percentage error, Nash-Sutcli e coe cient of e ciency (CE), and Pearson product-moment coe cient(R). The proposed approach gave more accuracy than traditional time series methods, whereas few studies conducted on cloud model theory and fuzzy time series to deal with uncertainty in data. The e ciency of the model is the limitation and along with decision making support of water supply can be carried out as part of future work.

Disinfection of water at the treatment plant and supply management remains a challenging work for many developing countries, and there is a need for prediction of water eutrophication. The TLI and data of Gaoyou Lake used by Zhang et al. (2020) for water quality prediction. The water parameters used for the TLI are DO, temperature, potassium permanganate index, ammonium, and qualitative evaluation has set. Based on PCA and multiple linear regression are used for early identi cation of eutrophication in water. The increasing trend has found from TLI, and it shows the light-moderate presence of eutrophy. The principle components contribution who's eigenvalues greater than 1 and considered for the linear model and purpose of the study is achieved after getting 75% credibility. The forecasted values over actual values compared and their errors plotted over the time at which samples were taken.

The performance of the statistical model is compared and forecasted DO concentration at several sites of the river by Monteiro and Costa (2018). The period over data used for analysis is from 2002 to 2015. The model used is being compared to are regression model with correlated errors and the state-space model as a calibration model. For short term forecasting, the stet space model is better than the regression model whereas, for long term or h-step ahead forecasting, the regression model with correlated errors performs better. The model performance measures were considered are R-square, Mean square error, mean absolute error, and mean absolute percentage error. The autocorrelation function(ACF) and partial autocorrelation function(PACF) were plotted with several lags. As DO is one of the characteristics of water quality, other quality parameters are used to predict the quality of water in the future.

Precise prediction of time series data encourages researchers to develop innovative models for water supply management systems. The models limitations are based on the linear and nonlinear nature of data, and therefore Faruk (2010) proposed hybrid ARIMA and the neural network model is implemented. At the same time, the strength of the traditional model is combined with a feed-forward, backpropagation structure of the network. The 108-month data is taken for the present study from 1996 to 2004 of river Menderes in Turkey. The hybrid structure model is capable of capturing the nonlinearity from the complex time series. The evaluation measures taken into consideration are RMSE, MAPE, and NSC. The prediction accuracy is compared with the individual ARIMA model and neural network model and gave a better performance than both. The ability to deal with non-linear time series makes hybrid model better recognition time-series patterns.

The time series analysis method and the monitoring data from coastal waters are used by An and Zhao (2017) to forecast the DO from water. The new two days data is predicted with the help of the ARIMA model. The original sequence is subject to the logarithmic rst order, and so it passes the test of stationary. For the tted model, the values of p and d kept it as 3 and 1 whereas q, which is partial autocorrelation value varies from (1,2,4). The average relative error observed is 4.79 percent between predicted values and actual observations. The MAPE found from the implemented model is 0.604 percent. The nonlinear problems can be deal with time series analysis combined with arti cial neural networks in future scope.

The application of water quality parameters forecasting is rarely explored and has challenges of data limitations, extensive computations, and future boundary restrictions. Liang et al. (2020) implemented the LSTM model on a complicated water quality system. The 12 years of data collected from the environmental uid dynamics code (EFDC) to perform LSTM on six types of water quality parameters The LSTM model has tested with combinations of the number of input parameters, hidden layers and its lags. After evaluated with NSE, the results show that the LSTM is good at the prediction of water temperature and TP. It has been observed that the number of hidden layers and the time lag not a ecting much on output and LSTM with a simple structure could show the forecast capacity of EFDC.

2.2 Multivariate time series analysis of water quality data

The degradation of water sources is a global concern, a ecting the survival of human beings, and therefore there is a need for the predictive ability of water quality. Abyaneh (2014) made the multivariate water quality parameter prediction with the help of multivariate linear regression (MLR) and arti cial neural network (ANN). The model performance is evaluated based on the coe cient of correlation (R), root means square error (RMSE), and bias values. The variables taken into consideration for the prediction of BOD and COD are temperature, pH, total suspended solids, and total suspended. The ANN model performance is far better than MLR for BOD forecasting. The estimation of DO percentage, Chloride, alkalinity, and total hardness modeled with the help of input parameters temperature (T), pH, and EC. Salami and Ehteshami (2015) used time-series data of 3000 instances for training with the help of feed-forward algorithm (ANN) the model and evaluation has done by R factor. Among these models, the prediction for DO percentage was more accurate, whose R factor is 0.95. The evaluation measures used in the study are Mean Absolute Error and R factor. These models are useful as the input considered such T, pH, and EC are easily measurable and can be implemented for surface water prediction as well. Water systems are inherently unstable because they contain multiple exposed elements. Perelman et al. (2012) proposed the multivariate time series to detect the faults in the water distribution system. The water parameters consider for the forecasting are pH level, turbidity, temperature, TOC, EC, etc. The arti cial neural networks (ANN) method is used to perform the multivariate time series analysis along with the possible outlier detection. The event probability identi ed after time series forecasting with the help of Bayes rules. The model evaluated with R^2 (R correlation coe cient, Mean squared error, confusion matrix, and True and false-positive rates. The proposed method consists of an alarm system on contamination detection on

single and multiple water quality parameters.

The traditional forecasting methods have to deal with the problems of accuracy, nonlinearity, complex behavior, etc. To predict the pH and water temperature Hu et al. (2019) proposed the Long Short Term-Memory method. The data is prepared before fed to the LSTM model with the help of interpolation, moving average, and smoothing techniques, result in noise reduced and null values overcome with linear values. Pearson's correlation is checked within the pH and temperature of the water and other water quality parameters. Finally, the prepared data were given to LSTM and checked the results in terms of accuracy and time taken to predictions. The short-term predictions are more accurate than long term predictions. The precise water quality prediction is the base of conservation of the water environment. Zhou et al. (2018) proposed water quality prediction with the help of Improved Gray Relation Analysis (IGRA) and LSTM RNN. The multivariate analysis implemented by IGRA, whereas the time series implementation has carried out by LSTM. The relation analysis has done with the correlation between water quality parameters. The optimal number of neurons used in the study are 3,8 and 1, whereas the epoch set to 50, and the data set is divided into 80:20 ratio for training and testing. RMSE evaluates the proposed research, and the minimum RMSE found for the prediction of DO is 0.67 and better match with actual values compared to ARIMA and BP.

The vector time series analysis of water pollution factors such as DO, BOD, and ammonia nitrogen has done by Wu et al. (2012) with the help of Generalized Autoregressive Conditional Heteroscedastic (GARCH) model and Vector Autoregressive Moving Average model (VARMA). The optimal models used after trails are VARMA(1,0,1) and GARCH(1,1) for time series analysis of pollutant factors. The Ljung box test is used before model implementation to check the residuals in regression. Data from the Taiwan watershed are used to establish statistical learning through VARMA-GARCH integration for prediction of water quality. The proposed method can capture the instantaneous changes in time-series data of water quality.

2.3 Anomaly detection in time series water quality

The deep neural network (DNN) and one-class support vector machine (SVM) implemented on the time series data generated from the cyber-physical system (CPS) by Inoue et al. (2017). These applied methods are evaluated against the raw water puri cation treatment plant (SWaT). The DNN is used on time series data as the outlier detector and compared with one-class SVM. The data used for training is log scaled from Swat at normal conditions and then evaluated on the multiple attack scenarios. The comparison based on precision and recall has done, and DNN slightly better than the one-class SVM.In the future, the problem of nding anomalous the improvised neural network architecture can be used.

The anomaly in real-time water quality monitoring data can cause a false alarm, and system-level reliability may decrease. Mitigate this issue, the novel anomaly detection, and anomaly detection and mitigation approaches presented by Zhang et al. (2017). The algorithm used is the autoregressive linear combination model, along with the back-

tracking strategy. The algorithm was tested successfully on three months of data from the water quality monitoring station, and the pH level is considered from the water quality parameters. The proposed anomaly detection algorithm is helpful in signi cantly reduce the false alarm and give accurate forecasting in a dual time window. After forecasting the pH level, the values are veri ed with the prediction interval and classi ed them as a false positive and false negative.

2.4 Impact analysis of time-series objects on each other

The vector autoregression (VAR) model has used by Sethi and Mittal (2020) for air quality prediction. The multivariate time series analysis has been carried out by the researcher and evaluated with statistical measures.Ali et al. (2015) proposed the inter-dependent relationship-based research in between water quality parameters such as DO, BOD, turbidity, and pH level. Johansen's cointegration test describes the relationship between quality parameters in terms of eigenvalues and eigenvectors. The VAR, VECM, and ARIMAX models are implemented in the research to analyze the dynamic and functional behavior between parameters. Whereas Zou (2018) used Vector error correction models (VECM) for analysis of Carbon emission, GDP, and international crude oil prices. The data considered from 1983 to 2013 to determine the impact of oil, GDP, and carbon emissions on each other, through cointegration tests and grangers causality test. The result of this study is obtained where there is a relation between oil price uctuation and carbon emission. There is a short term impact of oil price uctuation on both carbon emission and GDP, but in the long term, the in uence tends to be mild. Similarly, Song et al. (2017) has examined the novel idea of testing cause e ects of wastewater parameters such as COD/NH3-N and economic growth. The results show that there is a long-term bidirectional e ect of wastewater parameters and economic growth.

3 Methodology

The research methodology used for this research project is based on traditional Knowledge discovery in database (KDD) methodology, as shown in gure 1. The research methodology consists of ve main stages, including data collection, data pre-processing, data transformation, model implementation, and results and evaluation.

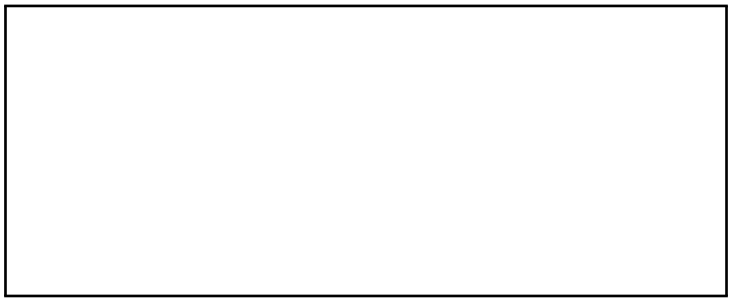

Figure 1: Methodology

3.1 Data collection

In this stage of Data collection, the data is collected from the online repository. As research is the forecasting of pH level and Turbidity, the data collection divided into two parts. For pH level, the data set is collected from the Environmental protection agency (EPA)[2], and Met Eireann[3], Ireland's national meteorological service, provides daily weather information. Whereas for Turbidity, the water quality monitoring data is taken from New York open data(NYC open data)[4], and the data for precipitation collected from the National Climate Data Center[5] formed by NOAA(National Ocean and Atmospheric Administration).

3.2 Data pre-processing

The data exploration is challenging without a proper structure of data handling, exploratory data analysis allow a researcher to explore data statistically, and apply hypothetical cases on it.(Yu; 2016). In this study, the data gathered from multiple sources to infer the water quality parameter. In this section of data pre-processing, to make time-series analysis meaningful, the research concentrate on the observations taken at one place and so the data set is ltered. The identi cation of the minimum value, maximum value, mean, median, rst and third quartile, and missing values in each column has done. Also, the box plot and histogram show the potential outliers and skewness in a feature. After the formation of a new data set, the null values from the data sets are identi ed and lled with linear interpolation. Initially, observations present in data repository are few in every month and after interpolation can be considered the daily, weekly, or monthly time-frequency for analysis.

3.3 Data transformation

The data transformation is one of the critical section in methodology before data apply to the time series model, it transforms data from one form to another form so that it can contribute in satisfactory predictions from the model. In this study, the features which are applied to the model rstly converted in time series data. Also, the series has gone through multiple tests to check whether it is stationary or not stationarity. If it is not then series needs to convert in log transform or rst order di erence has to apply on it to make it stationary. In this research project multivariate time series is implemented and so the e ect of rain has tested on other target variable such as pH level and Turbidity with the help of johansen cointegration test.

[2] http://www.epa.ie/water/dw/
[3] https://www.met.ie/climate/available-data/historical-data
[4] https://opendata.cityofnewyork.us/
[5] https://www.ncdc.noaa.gov/cdo-web/

3.4 Model Implementation

The model implementation is a stage at which the processed data are fed to model to achieve the research objective. In this study, four models are implemented for forecasting of water quality parameters such as pH level and Turbidity. Three models are statistical learning including Vector Autoregression (VAR), Vector Error Correction Model (VECM), and Autoregressive Integrated Moving Average with Explanatory Variables (ARIMAX). Whereas one model of Recurrent Neural Network (RNN) algorithm which is Long Short-Term Memory.

3.4.1 Vector Autoregression (VAR)

The vector autoregression model is generally a univariate model along with the forecasting of the vector of time series. Each variable in the vector represents an equation consist of constants, the lags of the variables present in the structure, and all variables are treated symmetrically along with the in uence of each other variable. The two dimensional VAR with lag one is shown by equations 1 , and 2 ? represents the in uence of1th lag of the variable considered for the analysis on the predicted variable and represents the noise in the process that may be correlated with each other. The computation is in a recursive manner as each new forecast time stamp considers the previous timestamp in the calculation.

$$y_{1;t} = c_1 + ?_{11;1}; y_{1t\ \ 1} + ?_{12;1}; y_{2t\ \ 1} + e_{1;t} \quad (1)$$

$$y_{2;t} = c_2 + ?_{21;1}; y_{1t\ \ 1} + ?_{22;1}; y_{2t\ \ 1} + e_{2;t} \quad (2)$$

VAR forecast each variable in the system up to time T in a recursive manner for h ahead forecast. There are two decision should include while implementing VAR that is the number of variables and the lag. The drawback of this model is the more inclusion of variables, the more system becomes complex, and the more noise will generate. The VAR is useful in many ways, such as one variables usefulness in the forecasting of the variable, response analysis of variables considered in the study, and forecasting of collection of variables where no explicit clari cation is required.(Hyndman; 2018)

3.4.2 Vector Error Correction Model (VECM)

If the set of variables is cointegrated with each other, then one of the suitable estimation models is the Vector error correction model (VECM). The VECM is an autoregressive model along with the cointegration constraints. VECM dose adjustment of the deviation from equilibrium and short term change. The variables present in the VECM model carry the information on the e ect of past values and the current values of variables. The relevant coe cient shows that past equilibrium errors play a role in the evaluation of the present outcome captures the long-run e ect. In the VECM model equation, the error correction model is considered. In the proposed study, the cointegration between water quality parameters and the rainfall is checked, and then VECM is applied to them. (Zou; 2018)

3.4.3 ARIMAX

Autoregressive Integrated Moving Average (ARIMA) is the preferable time series model due to the exibility and forecast accuracy of the model. The autoregression in ARIMA is nothing but the linear regression of variables with past values of itself. The ARIMAX model is an ARIMA with an explanatory variable which is independent variable along with the predictor. In ARIMA, it is represented as ARIMA(p,d,q), whereas p is order of autoregression or lag order, d is the order of di erence, and q is the order of MA. The values of p and q can nd out from the autocorrelation plot (ACF) and partial autocorrelation plot(PACF) (Faruk; 2010). The ARIMAX model adds the external variable with its coe cient and time at t along with the target variable.(Hyndman; 2018)

3.4.4 Long Short-Term Memory (LSTM)

The LSTM is a Recurrent neural network(RNN) algorithm and has the advantage of remembering the old values continuously during the training of the model. The "cell state" processor is the main di erence between traditional RNN and LSTM and enables the bene t of memory storage. The typical RNN disadvantage of information loss in the LSTM neural network overcomes by cell state. The hidden layer of RNN has only one state s whereas, in LSTM, there is the other cell states c. The hidden layer will have three inputs input value, which is current x(t), output value s(t-1) of hidden layer, and third is unite state c(t-1). There two outputs at the end, one from the hidden layer and another from the state c(t). Controlling of c is done by three gates present in LSTM one r1 (forget gate), r2 (input gate), and r3 (output gate). Figure 2 is an internal structure of hidden layers of LSTM, whereas x is input data, S is output from previous neuron layer, C(t-1) is the previous unite state, ct is current unite state. In contrast, the st is the output of the current hidden layer. Also, there are three gates forget gate, input gate, an output gate, and their weights. LSTM has the data back-propagation process as the

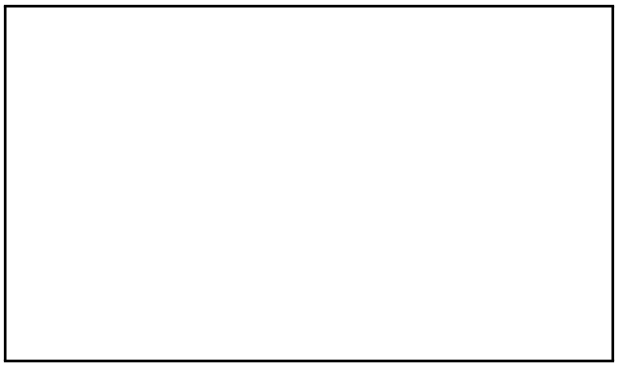

Figure 2: Internal Structure of Hidden Layers of LSTM Netwrok

RNN has the error values carry forward along with the time series data. The gradient update of horizontal and vertical weights and bais and the assignment of the new weight has done by the hidden layer structure. The learning rate controls the error and can be set by .(Hu et al.; 2019)

3.5 Result and Evaluation

The result and evaluation phase of methodology consist of outcomes received from the model implemented in the previous stage. Based on statistical measures such as Mean square error (MSE), Root mean square error (RSME), Mean absolute error (MAE), and Mean absolute percentage error (MAPE), the applied models are evaluated. RSME illustrates the spread of residuals and the most widely used evaluation measure for forecasting analysis. At the same time, the MAPE shows the deviation and the ratio between actual and forecasted values. (Hyndman; 2018)

4 Design Speci cation

The architectural design of the proposed research is shown in gure 3. Stage 1 represents the data collection, exploratory data analysis, and preliminary visualization. Stage 2 consists of model implementation mainly divided into two parts, statistical learnings such as VAR, VECM, and ARIMAX, whereas another part is of machine learning in which LSTM RNN implemented. The 3rd stage is the result and evaluation of applied techniques with the help of statistical measures. The R studio is an environment where stage 1 of the research study is executed. The stage 2 implementation is divided into two parts for statistical learning R studio is used, whereas for Neural network modeling is implemented in Python. The R studio version is 1.3.959, while the implementation of LSTM has carried out in Python 3.7.6.

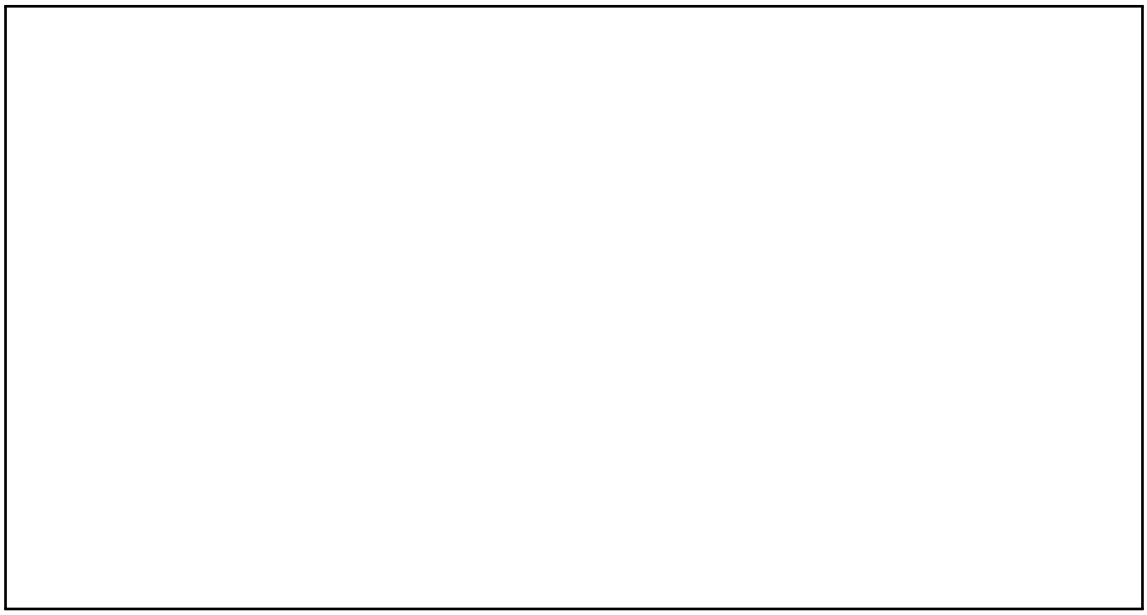

Figure 3: Architectural Design

5 Implementation

5.1 Data collection

The data collected from multiple online repositories clubbed in the data collection section. The water quality data of Wicklow county from the Environmental protection agency (WaterTeam; 2018), and precipitation data from MetEireann, joined together and form the data set for forecasting of pH level. On the other side, the water quality data from New York open data(NYC open data) and the daily precipitation data from National Climate Data Center joined together and form the data set for turbidity forecasting. The data sets concatenated according to the place from where the water quality parameter has taken and the amount of rainfall from the nearest weather station. The concatenation of data is accomplished with the help of the \merge" function from the \dplyr" package, and the factor considered in between two is a \Date" to form a new data set. The New York data set has 80,399 observations of 400 sites from January 2015 to April 2020 and Wicklow data set consists of 2399 observations of 220 schemes from January 2015 to November 2018.

5.2 Exploratory Data Analysis

The exploratory data analysis is part of stage 1 in the proposed research and helps in the methodology of research to explore data statistically. Exploratory data analysis contains two major sections, data pre-processing and data transformation.

5.2.1 Pre-processing

The date column is an important feature and needs to convert in \Date" class as mostly the dates are stored as a character. The date column turned into date class with the help of \as.Date" function. The data set has ltered to make the time series analysis more meaningful the ltration has done on \Scheme Code" and \SampleSite". The size of data is reduced after ltration, but the research outcomes will be signi cant for one particular place. The summary function is used to identify the statistical analysis of columns present in the data set. The desired features, such as Turbidity from New York and pH level from Wicklow, are selected for further processing. The turbidity parameter is also present in Wicklow data with missing values, and the number of observations is less after lter on one site. On the other hand, the number of observations is more for Turbidity in New York data set. As the more number of observations makes the model train more e ciently, and therefore the Turbidity time series from New York data set is considered over the Wicklow data set. The Turbidity and the pH level after ltration on one site, the number of observations noted at the ltered site are 345 in the New York data set and 124 in the Wicklow data set. The boxplot and histogram are used to identify critical outliers from the data set. Figure 4 shows the box plot and histogram for turbidity and precipitation from New York data set and the box plot and histogram for pH and rain from the Wicklow data set.In both cases, it is found that the features have

critical outliers and skewness in data. In pre-processing, these outliers can be removed and replaced with the mean values or linear values. But in this study, outliers are not removed as these features are possible natural observations and can contribute vitally to the research outcome. Observations present in data sets are not in equal intervals,

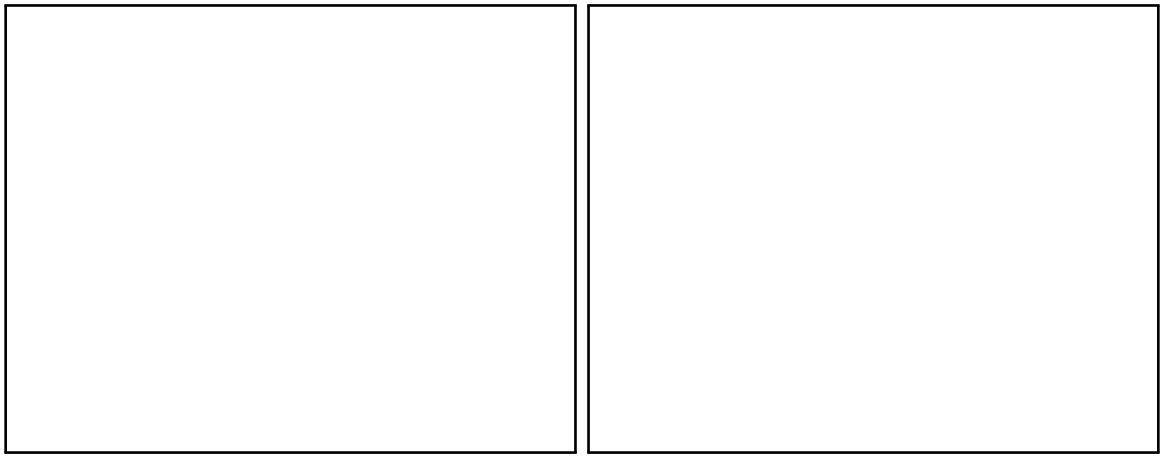

Figure 4: Box plots and histograms

every month few values are noted for water quality parameters in both data set. The identi cation of missing timestamps made by observing the data frame and its plot. The missing dates are lled by generating a regular sequence of dates in the data frame and \fulljoin" to add those in the data frame. The range of observations has obtained by the imputation of the missing data with the help of R package \ imputeTS". There are multiple ways to ll the data, such as interpolation, Moving average, mean, Kalman smoothing, etc. In this study the missing data occupied with linear interpolation and data set becomes larger in number as interpolation lled the missing data between two observations. The resulting data set for turbidity is 1950 and for pH level 1397.

5.2.2 Data transformation

The data transformation transforms data from one form to another to make it suitable for model construction and some preliminary visualizations. The linear interpolated data generate more data, and now it is possible to make it equal interval time stamp such as daily, weekly, monthly, and yearly. Di erent ts plots plotted according to the frequency of timestamp. Therefore with the help of the \ts" function, the series is converted to time series format, as shown in gure 5. The water quality data is observed and noted down along with the date to execute the pattern or consistency of water quality. The \lubridate" package is used to retrieve the month and year from the date format to summarize the data month wise. These summarized data illustrate the seasonal behavior in water quality. Figure 6 shows the seasonal visualization of water quality parameters Turbidity from New York data set and pH level from the Wicklow data set. It is observed that the up and down pattern in water quality level is the same for many years, and the water supplier must be more cautious concerning the quality of water in a speci c time frame. Other seasonal visualizations such as polar plots, can also contribute to analyzing the water quality data in the preliminary stage.

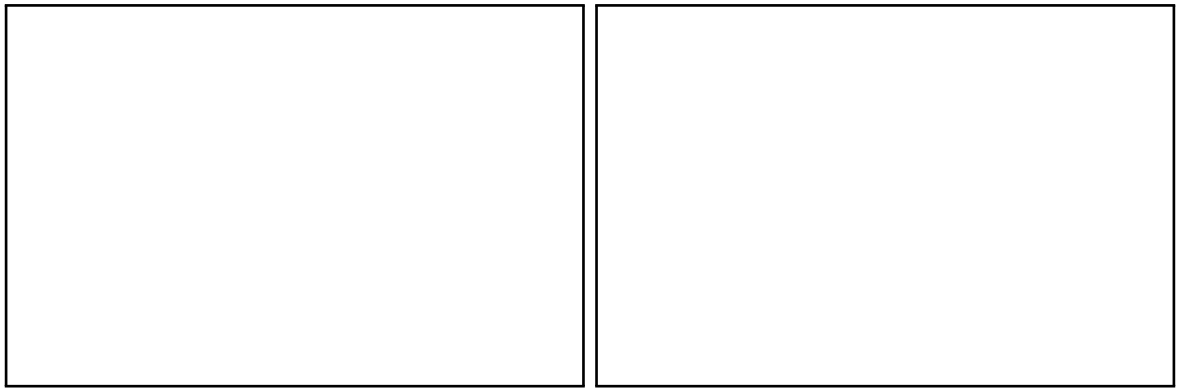

Figure 5: Time series plots of Turbidity and pH level

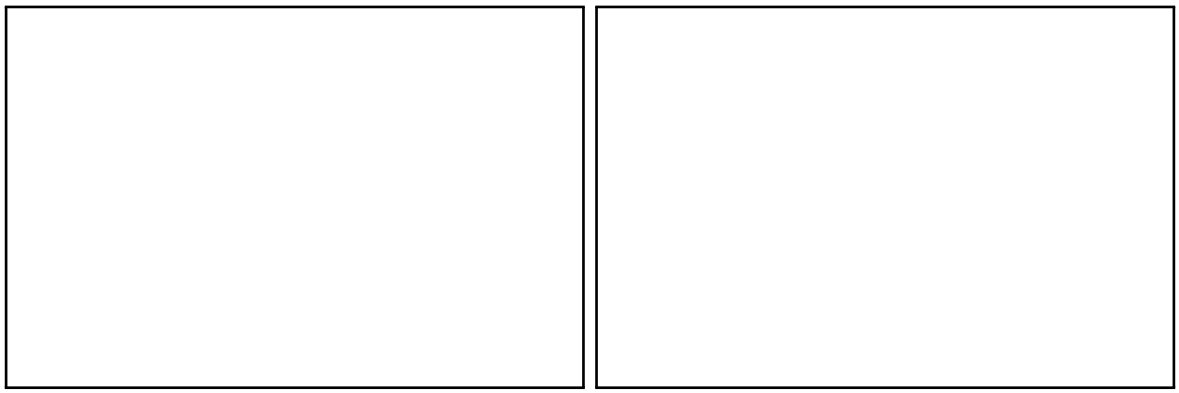

Figure 6: Seasonal plots of Turbidity and pH level

5.3 Modeling and Evaluation

The modeling is the phase where actual data is applied to the models used for forecasting of water quality parameters. The seasonal parameter of water quality and rainfall is adjusted by using the function \seasadj" of R. To apply the time series model on data, it must pass the test of stationarity. The Augmented Dickey-Fuller (ADF) test, also called a unit root test is conducted for each feature in the data set to check its stationary part. The ADF test applied to the rainfall and water quality parameter object of New York as well as the Wicklow data set. The ADF test on all objects gives a p-value of less than 5%, which is less than the critical value. Therefore rejects the null hypothesis, and the alternate hypothesis is stationary. The seasonally adjusted series objects are stationary and ready to apply time series models on it. The correlation between two variables shows the linear relationship in two variables, whereas the autocorrelation gives a relationship between the present value and its lagged values. The ACF plot helps to nd autocorrelation from previous lagged values, whereas partial autocorrelation can be found out from the PACF plot.(Hyndman; 2018)

5.3.1 Vector Autoregression (VAR)

Vector autoregression is used for predicting vectors and considers the e ect of variables on each other. In this research, the impact of rainfall is considered for the prediction of water quality parameters. After pre-processing, transformation, and stationarity check data is combined to form a vector of water quality parameters and rainfall. The function \ts.union" is used to create a vector of Turbidity and rainfall in New York and pH level and rainfall in Wicklow. Furthermore, the data is divided into 80:20 proportion, where 80% of data applied to models for training, and 20% data is reserved for testing. The appropriate combined lagged value is selected from the \VARselect" function from the vars package. The minimum lag has selected based on Akaike Information Criteria (AIC). The lag value for Turbidity and rainfall is 10, whereas for pH level and rainfall is 6. The lag consideration is nothing but past observations in the time series. The model train with the help of the \VAR" command and summary of the model analyzed. The model has forecasted ahead of the values present in the data set for training. The forecasted test values are plotted against actual test values, as shown in gure 7.

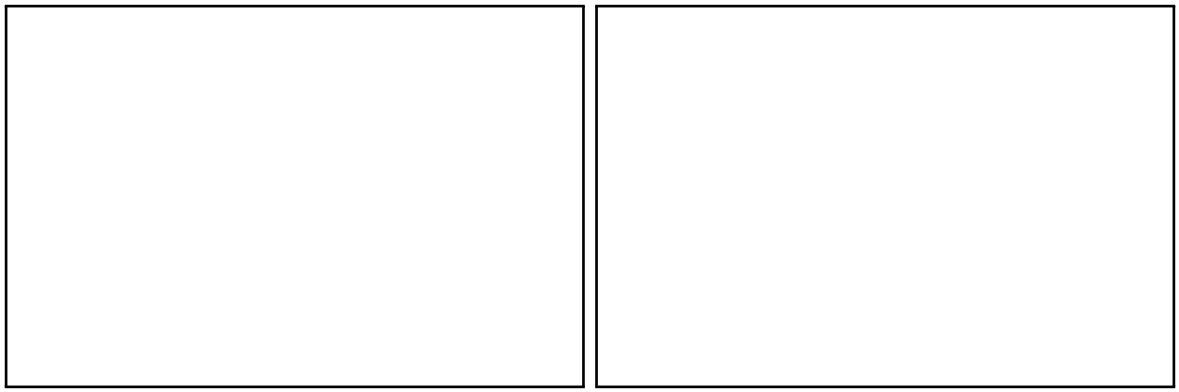

Figure 7: Forecasting of Turbidity and pH level from VAR

Evaluation:- The actual values and forecasted values from the test portion have differences and can be demonstrated in the form of an evaluation matrix. The test portion of the data has reserved aside while training the model. The model forecasted ahead of the train data and compared against the actual values present in the test portion. The evaluation has done with the help of statistical measures such as MSE, RMSE, MAE, and MAPE, as showed in table 2. The MSE and RMSE values counted for Turbidity are 0.0132 and 0.1151, whereas for pH level 0.1774 and 0.4212. Similarly, the MAE and MAPE values for Turbidity are 0.0924 and 0.1371, whereas 0.3110 and 0.0392 for pH level.

Water quality paramter	MSE	RMSE	MAE	MAPE
Turbidity	0.0132	0.1151	0.0924	0.1371
pH Level	0.1774	0.4212	0.3110	0.0392

Table 2: Evaluation matrix VAR

5.3.2 Vector Error Correction Model (VECM)

VECM is an extended version of VAR, and the only di erence is the error correction part has added to it. The impact of rainfall and the water quality parameter has tested with the help of the Johansen Cointegration test. Johansen-Cointegration test is a method to check the linear relationship between two variables. The test has conducted on multivariate vector Turbidity and rainfall as well as pH level and rainfall with the help of \ca.jo" from package \urca". At r = 0, test statistics is 301 more than 17, which is the critical value at signi cance level 5%. The second hypothesis$\prec$= 1 test statistics value is 74 and again more than even 10% critical value and reject the null hypothesis. This means that there is a cointegration relationship between the turbidity and rainfall amount. Similarly, there is also a cointegration relationship between pH level and rainfall amount, as test statistics are 114 at r=0 and 32 at r< = 1 higher than 5% critical value. After satisfying cointegration relation, the divided data is used for model training where 80% of data used for train and 20% data is used for model testing. The same lagged value from \VARselect" is used based on Akaike Information Criteria (AIC) by VECM model 10 for Turbidity and rainfall, whereas for pH level and rainfall is 6. The models trained on train data with the help of the \VECM" command and forecasted. The actual and predicted values of the test part are shown in gure 8 for turbidity and pH level.

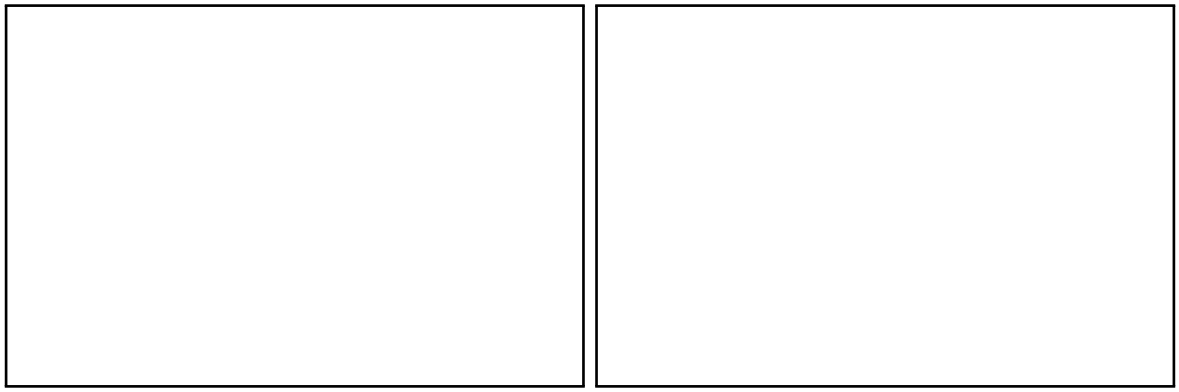

Figure 8: Forecasting of Turbidity and pH level from VECM

Evaluation:- The model forecasted ahead of the train data with the help of \predict" and compared against the actual values present in the test portion. The evaluation has done with the help of statistical measures such as MSE, RMSE, MAE, and MAPE, as showed in table 3. The MSE and RMSE values counted for Turbidity are 0.0273 and 0.1652, whereas for pH level 0.1873 and 0.4328. Similarly, the MAE and MAPE values for Turbidity are 0.1412 and 0.1848, whereas 0.3136 and 0.0394 for pH level.

Water quality paramter	MSE	RMSE	MAE	MAPE
Turbidity	0.0273	0.1652	0.1412	0.1848
pH Level	0.1873	0.4328	0.3136	0.0394

Table 3: Evaluation matrix VECM

5.3.3 ARIMAX

Autoregressive Integrated Moving Average with explanatory variable (ARIMAX) is employed on the same data. However, the individual time series objects which passed the stationarity test are used. The data has divided into train data and test data individually, and the ratio considered for partition is the same as previous, which is 80:20. The model- tting has done with the \Arima" command in R. Requires the details such as (p,d,q), where p is the order of autoregression or lag order, in this case, it is selected 10 for turbidity and 6 for pH. The lag is selected the same from the previous model. The d is the order of di erence, and in this case, it is 0 as there is no rst or second-order di erence has taken for both water quality parameters. The q is the order of the Moving Average (MA), and it is observed from the PACF plot as 10 for turbidity and 4 for pH level. The main di erence in ARIMA and ARIMAX is an external variable, the rainfall amount considered in this case. The model predicted and forecasted plots against actual data are plotted, as shown in Figure 9.

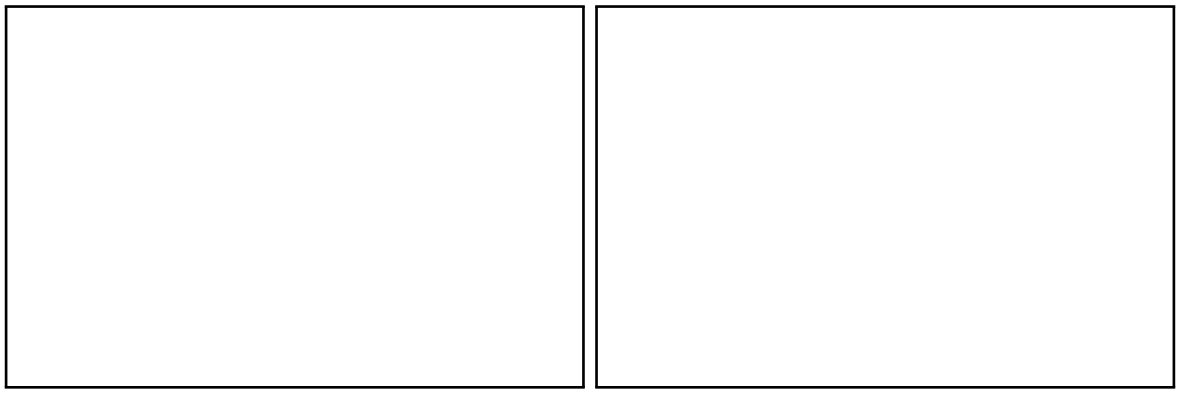

Figure 9: Forecasting of Turbidity and pH level from ARIMAX

Evaluation:- The test portion of the data act as actual values for predicted values from the model, and the evaluation has done by comparing these two. The statistical measures such as MSE, RMSE, MAE, and MAPE are used to evaluate the error between actual values and predicted values. The MSE and RMSE values calculated for Turbidity are 0.0145 and 0.1206, whereas for pH level 0.1814 and 0.4259. Similarly, the MAE and MAPE values for Turbidity are 0.0981 and 0.1451, whereas 0.3078 and 0.0387 for pH level.

Water quality paramter	MSE	RMSE	MAE	MAPE
Turbidity	0.0145	0.1206	0.0981	0.1451
pH Level	0.1814	0.4259	0.3078	0.0387

Table 4: Evaluation matrix ARIMAX

5.3.4 LSTM

The data obtained after data pre-processing and transformation is extracted in the CSV le from R studio and loaded in python for further implementation of LSTM. The machine learning model gives better performance whenever the input features are in the same range, which is called normalization and standardization. The data is scaled with the help of \MinMaxScaler" from \sklearn" pre-processing library. Each feature is scaled and translated in the range from 0 to 1, and the scaling allows input features to be minimum in variations. The supervised learning LSTM has implemented by the de ned function, which considers the lags of input data with the help of the shift function of python, which gives the facility of mapping lagged observation to present observation. In the current experiment lag 1 is considered, which is one day back. Therefore, the multivariate time series data set is formed in such a way that the output of turbidity or pH mapped with its own previous lagged value, and rainfall's lagged value.

After the pre-processing, the data is separated into a train and test. The ratio is the same as the previous 80% data is for training, and 20% data is for testing. The data set has reshaped into 3D manner such a way, that represents [samples, timesteps, features]. For the current experiment, the samples are the number of data present in train or test, the time step is 1, and 2 features considered in implementation. The LSTM sequential model is de ned with the one input layer, one hidden layer, and one output layer with a linear activation function. The neurons consider in the hidden layer are 3 in the current experiment. The mean absolute error is used as a loss function, and Adam optimizer is used to update the weights of neurons after every epoch. The batch size is decided from a mini-batch size gradient descent concept, and after for the current experiment, it is 32(Masters and Luschi; 2018). The number of epochs given is 200 in both cases. The model trained on the above speci cations of LSTM and the loss function plotted at the end after the last epoch. After the prediction of test data, the data is collected and resized to the original shape for inverse scaling. The inverse values will be in actual scale and compare against the true values. The plots of actual and predicted values versus the time step in days are shown in gure 10.

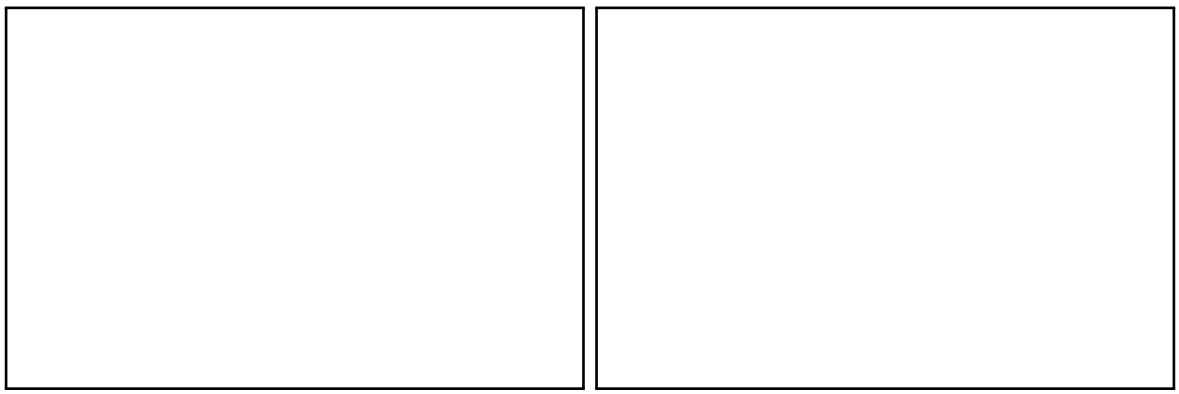

Figure 10: Forecasting of Turbidity and pH level from LSTM

Evaluation:- The evaluation has done by statistical measures such as MSE, RMSE, MAE and MAPE with the help of \sklearn matrics". The MSE and RMSE values calculated for Turbidity are 0.0005 and 0.0241, whereas for pH level 0.0029 and 0.0546. Similarly, the MAE and MAPE values for Turbidity are 0.0169 and 0.0227, whereas 0.0274 and 0.0035 for pH level. The results from LSTM are much better than the autoregressive models.

Water quality paramter	MSE	RMSE	MAE	MAPE
Turbidity	0.0005	0.0241	0.0169	0.0227
pH Level	0.0029	0.0546	0.0274	0.0035

Table 5: Evaluation matrix LSTM

6 Model Comparison and Discussion

Model testing on the same platform is essential to nd the precise model. Among all measures, the RMSE value is much more sensitive to the extreme values, and thus it is useful as the robust evaluation measure of models. Similarly, the MAPE shows the ratio between di erence and actual data. (Hu et al.; 2019). Therefore the implemented turbidity and pH forecasting models are compared based on RMSE and MAPE, as speci ed in table 6. It is observed that the LSTM outperforms other models as its RSME is 0.0241, and the MAPE is just 2%. Similarly, in the case of pH level forecasting again, LSTM gives better performance as its RSME is 0.0546, and MAPE of LSTM is just 0.3%. Overall the implemented statistical autoregressive models have forecasted water quality parameters with nearly similar accuracy. In contrast, the LSTM model forested better as results of RSME and MAPE are far better than other implemented models. The experi-

	Turbidity		pH Level	
Models	RMSE	MAPE	RMSE	MAPE
VAR	0.1151	0.1371	0.4212	0.0392
VECM	0.1652	0.1848	0.4328	0.0394
ARIMAX	0.1206	0.1451	0.4259	0.0387
LSTM	0.0241	0.0227	0.0546	0.0035

Table 6: Model comparison

mental analysis results show that the better prediction of suspended particulate matter from water data along with rainfall amount can be made with the help of LSTM. The statistical analysis, such as cointegration, shows that there is an e ect of rainfall on water quality parameters and from which it states that there can be the inclusion of external physical quantity during the prediction of water quality. The mapping of lag values to future value architecture of LSTM performed in this study gives better performance compared with statistical models applied. There are a few di culties faced while developing the proposed research.

The data considered for the proposed study is from water quality morning repository and the amount of rainfall from the weather station of the water supplier area. The data concatenation also needs to match the position from where the two data are collected, and therefore, the data collection also plays a vital role in meaningful research.

The water quality parameter observations considered in the analysis should be noted at one dedicated place only. After applying the lter on the data set, the number of records available for the research was less, and the results could be more accurate after the addition of more time-series data. The water quality parameter observations noted in the data set are random in a month, and the occurrence was not in the perfect interval. The ideal range makes the study more meaningful, accurate, and requires less e ort to make it perfect for time series analysis.

The time series objects were treated before it fed to the statistical and machine learning models. The test of stationarity, seasonal adjustment, proper lag selection were performed on the data. If the resulting series is non- stationary, then appropriate action such as rst-order di erence or log transform needs to be done to make them stationary. The data scaling and mapping of future and lagged values have done before applying it to the LSTM model. Also, after forecasting the reshaping and inverse scaling has carried out.

7 Conclusion and Future Work

The research explored the possibility of detecting suspended particulates in water by forecasting water quality data. The water quality monitoring data from New York, US, and Wicklow, Ireland, has used for forecasting water quality parameters such as turbidity and pH level. Considering the external physical features such as the amount of rainfall the multivariate analysis has been performed. The Johansen cointegration test executed between water quality parameters and precipitation states that both time series variables could move together. The autoregressive models such as VAR, VECM, and ARIMAX and recurrent neural network LSTM has successfully implemented for water quality parameter forecasting. The experimental comparative analysis states that the LSTM model is more desirable for forecasting of water quality parameters to identify suspended particles from drinking water. In the future, other water quality parameters should be used as a vector, and more external physical features such as the rate of sewage in the area of water supply along with precipitation should include. The performance of LSTM can be observed in the future after consideration of more lagged values of independent features. Also, multi-step forecasting plays a vital role in the domain of water quality management.

8 Acknowledgement

I would like to express sincere gratitude to my supervisor Dr. Christian Horn for guiding throughout the project implementation. I would be grateful to him for his constant guidance and supervision, which made it simpler for the project implementation.

References

Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and arti cial neural networks in prediction of water quality parameters, *Journal of Environmental Health Science and Engineering* **12**(1): 40.

Ali, G. et al. (2015). Cointegration var and vecm and arimax econometric approaches for water quality variates, *Journal of Statistical and Econometric Methods* **4**(1): 1{38.

An, Q. and Zhao, M. (2017). Time series analysis in the prediction of water quality, *7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)*, Atlantis Press.

Carswell, S. (2019). Alerts missed at leixlip water plant that put 600,000 on boil notice, *The IRISH TIMES* .
URL: *https://www.irishtimes.com/news/ireland/irish-news/alerts-missed-at-leixlip-water-plant-that-put-600-000-on-boil-notice-1.4123058*

Deng, W., Wang, G. and Zhang, X. (2015). A novel hybrid water quality time series prediction method based on cloud model and fuzzy forecasting, *Chemometrics and Intelligent Laboratory Systems* **149**: 39{49.

Faruk, D. Ö. (2010). A hybrid neural network and arima model for water quality time series prediction, *Engineering applications of arti cial intelligence* **23**(4): 586{594.

Hu, Z., Zhang, Y., Zhao, Y., Xie, M., Zhong, J., Tu, Z. and Liu, J. (2019). A water quality prediction method based on the deep lstm network considering correlation in smart mariculture, *Sensors* **19**(6): 1420.

Hyndman, R.J.and Athanasopoulos, G. (2018). Forecasting: principles and practice, 2nd edition, otexts: Melbourne, australia, 2018, *Google Scholar* .

Inoue, J., Yamagata, Y., Chen, Y., Poskitt, C. M. and Sun, J. (2017). Anomaly detection for a water treatment system using unsupervised machine learning, *2017 IEEE international conference on data mining workshops (ICDMW)*, IEEE, pp. 1058{1065.

Jalal, D. and Ezzedine, T. (2019). Toward a smart real time monitoring system for drinking water based on machine learning, *2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, IEEE, pp. 1{5.

Kale, V. S. (2016). Consequence of temperature, ph, turbidity and dissolved oxygen water quality parameters, *International Advanced Research Journal in Science, Engineering and Technology* **3**(8): 186{190.

Liang, Z., Zou, R., Chen, X., Ren, T., Su, H. and Liu, Y. (2020). Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach, *Journal of Hydrology* **581**: 124432.

Masters, D. and Luschi, C. (2018). Revisiting small batch training for deep neural networks, *arXiv preprint arXiv:1804.07612* .

Monteiro, M. and Costa, M. (2018). A time series model comparison for monitoring and forecasting water quality variables, *Hydrology* **5**(3): 37.