National College of Ireland

Daily Precipitation Forecasting using Neural Network - A case study of Punjab,India

MSc Research Project
Data Analytics

Punit Lohani
Student ID: x18127339

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Punit Lohani
Student ID:	x18127339
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/12/2019
Project Title:	Daily Precipitation Forecasting using Neural Network - A case study of Punjab, India
Word Count:	XXX
Page Count:	21

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:	12th December 2019

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.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Daily Precipitation Forecasting using Neural Network- A case study of Punjab, India

Punit Lohani
x18127339

Abstract

Precipitation forecasting is one of the most important and crucial task that has gained the attention of the meteorologists around the world. A lot of Changes have been observed in the climatic pattern over a period of time, so this eld has become the area of research for the research communities. Accurately forecasting the amount of precipitation can prove to be very useful in detecting the occurrence of natural calamities like drought and ood in the near future. Due to the lack of proper irrigation facilities in some parts of India, farmers highly rely on precipitation. The traditional methods of forecasting are time consuming and due to non linear data, sometimes results in an inaccurate forecast. Machine learning approach has the ability to overcome the limitations and can present the forecast accurately. This study primarily focuses on the forecasting the daily amount of Precipitation using Seasonal ARIMA model and Long Short Term Memory networks. The objective of the study would be to check the ability of LSTM networks in accurately presenting the daily forecast for the amount of precipitation over Punjab in India.

Keywords: Precipitation forecasting, meteorologists, natural calamities, Seasonal ARIMA, Long Short Term Memory.

1 Introduction

Weather conditions are changing continuously and rapidly and the entire world is su ering from the e ects of the climate change so it is really important for e cient and accurate forecasting. In hydrological circulation, precipitation is considered to be an important variable of meteorology (Den (2012)). In the drought prone areas, people depend on the precipitation for daily survival and planting their crops. The agricultural sector highly relies on the precipitation as this helps the farmers to make an e ective planning to plant and harvest their crops. Sometimes for the farmers of the developing countries, it is not always easy to get the proper irrigation facilities for their elds that result in high dependency on the precipitation. In addition to this, an accurate forecast can help in providing the early warnings of natural calamities that can take place in future.

Wireless sensors, high speed computers, weather radars and meteorological satellites are some of the widely used tools for the collection of data of the weather parameters like precipitation, temperature, humidity (Sawaitul et al.; 2012). The models that are currently used by departments for performing the forecast are highly dependent upon

the physical models that have high complexity and require computer system with high computational power and they are time consuming as well. In spite of using such complex and costly devices the forecast often turns to be incorrect (Jakaria et al. (2018)). Machine learning models on the other hand helps in predicting the weather related variables and they are less complex and less costly as well.

India is an agriculture based country and for higher yielding of crops the Indian farmers always depends upon the precipitation. In India, the summer monsoon rainfall starts in the month of June and lasts till September and this is time that is very crucial for the crops. According to (Swaminathan; 1998)the summer monsoon rainfall in uences around 65 percent of the cultivated land in India, therefore the prior information can help the farmers to take the maximum bene t from it also they can take the appropriate measures in case of less amount of rainfall. Forecasting the summer monsoon has become a challenging task in the meteorological eld.

Thus, the prediction of weather conditions particularly emphasizing on the amount of precipitation becomes really important for the production of crops, management of water resources, replenishment of water table and redistributing water in the water cycle. In the recent years, eld of precipitation forecasting has gained its interest among the research communities around the world and the researchers have carried out several studies by using various statistical, deep learning and neural network techniques.

As a part of this research, the objective is to forecast the daily amount of precipitation for Punjab, India. Punjab is basically located on north-west part of India and is known to be the land of rivers because it is surrounded by ve rivers. The state receives the maximum amount of precipitation from June to September, while the wettest month is July. Although the state is only 1.4 percent out of the total part but produces around 14 percent of crops that plays an important role in the gdp of the country. The state is also among the top three producers of crops like white, rice (Swati; 2018)

Sometimes due to lack of proper irrigation facility, it becomes really di cult for the farmers to take care of their crops because the crops produced by this state requires su cient and good amount of water so that during the time of harvesting the best yield can be achieved. If the accurate forecast can be gives, there can be the possibility of early warnings that can be gives to the people so that they can plan accordingly well in advance. Recently, the state also faced the situation of severe ood due to which many villages were a ected and crops were highly damaged.The incessant precipitation was the main reason behind the severe ood condition. Therefore for the state like Punjab, it becomes really important to forecast the precipitation so as to avoid this kind of situation in the coming future.

In this research, seasonal arima and Long Short Term Memory model have been employed. For evaluating and checking the outcome of the model, root mean square error has been used. In addition to this, for evaluating the performance of LSTM, the persistence model has been taken into consideration as it serves a baseline for the LSTM model. The global weather dataset for conducting this research has been taken from spatial sciences website of Texas A & M University covering in total 28 years of daily observations with longitude and latitude of 75 9375' E and 31 379' N respectively.

Figure 1: Location of study area
(*Philips world atlas*; 2012)

1.1 Research Question and objectives

Now the research question arises that, what can be the best possible accurate forecast can a Long Short Term Memory networks provides as compared to Seasonal ARIMA model for daily amount of precipitation in Punjab, India.

The objective includes, applying the tradition Seasonal ARIMA model , followed by the Long short term memory model in forecasting the time series and checking the forecast ability and accuracy using the root mean square error evaluation metric.

2 Related Work

This section primarily focuses on the state of art and the various techniques that have been adopted and applied by the researchers in the eld of forecasting the amount of precipitation in various regions across the world. In the further sub sections, the works have been categorized as per the di erent techniques so as to get the better understanding.

Clear understanding of the amount of precipitation highly impacts the agricultural countries specially India as the farmers are dependent upon it.The summer monsoon rainfalls have around 65 percent impact on the cultivated land in India (Swaminathan; 1998). In case of less amount of rainfall in the future, appropriate measures can also be taken well in advance that can help in preventing the damage to the crops. In addition to this, if the amount of rainfall exceeds the limit then there can be chances of natural

calamities like ood. With appropriate and e ective forecasting this can also be tackled to a great extent.

In the recent years, forecasting the amount of precipitation has taken attention of the meteorological departments for dealing with the above issues. Researchers have done signi cant amount of work to develop a model for accurate forecast.(Yeshwanth et al.; 2019)in their work presented that, since the nature of the atmosphere is dynamic so sometimes it becomes di cult for the statistical techniques to deliver and present the result accurately, hence machine learning algorithms proves to be useful. Researches are being conducted by taking into consideration di erent algorithms so as to present the accurate result for the target audience.

Forecasting the occurence and amount of precipitation is related very closely with the life of the people around the world. If the prediction can be made in an early phase then it is possible to plan for any upcoming natural calamities like ood and drought.(Parmar et al.; 2017), in their research delivered an indormation non linear nature of the precipitation data and sometimes it becomes really di cult to deal and forecast using statistical techniques which makes machine learning algorithms suitable. This paper presented the works of various researchers using di erent machine learning approach so as to present the accurate forecast for precipitation. The techniques like ARIMA models, SVM have been discussed along with the importance of neural networks and approaches of deep learning.

2.1 Forecasting of Precipitation using Statistical Techniques

Accurate forecasting of rainfall plays a crucial role in day to day life. If the forecast is made correctly then it can become really helpful in providing the information about the occurrence of any future calamities as well. It has been observed that the data mining techniques works well in providing the accurate results of forecasts as compared to the statistical techniques.(Razeef et al.; 2018) conducted a research by comparing the results of some of the data mining algorithms like J48, Logistic regression, Na•ve bayes, Random forest and others etc. For this, they have considered Srinagar, which is the top most part of India. The dataset has been obtained from wundergrounds website for Srinagar consisting of 1 year of data. Out of the 9 attributes, useful 5 attributes have been taken into consideration. The research has been carried on supervised learning approach. Mean absolute error, RMSE, F-measure, accuracy and precision are used for the comparison of result. It has been inferred that that Random forest algorithm performed really well among the other data mining algorithms.

Since India is an agricultural country, therefore accurate prediction of rainfall is very crucial. (DUTTA and TAHBILDER; 2014)presented their work for predicting the amount of rainfall in Assam by using multiple linear regression. For this, they have taken the dataset from meteorological centre in Guwahati covering a period from 2007 to 2012 containing the monsoon months of the state. Variables taken in the analysis includes maximum and minimum temperature, wind speed, pressure and humidity. F test is performed to check whether the variables included in the analysis are su cient for the rainfall prediction. Adjusted R square has been used for evaluating the performance of the proposed model. As a result, acceptable accuracy has been achieved by the technique.

Similar approach have been applied by (Swain et al.; 2017) for forecasting the annual precipitation in Odisha. For this, they have collected the dataset from india water portal website covering the year from 1901 to 2002. In order to evaluate the performance of the proposed model, adjusted R square and coe cient of determination has been used. The result showed that the model performed really well in forecasting the annual amount of precipitation and can be applied for the investigations related to the meteorology.

2.2 Forecasting of Precipitation using Artificial Neural Network

For predicting the total amount of precipitation during the summer season in India, (Sahai et al.; 2000) in their research have made the use of Arti cial Neural Network. The data from the year 1871 to 1994 has been collected from di erent meteorological stations covering the period from June to September. The results of the forecasts proved to be quite promising by achieving the RMSE of 54.2 and it is because of the season taken into consideration as during summers the chances of getting the heavy precipitation are more as compared to the other seasons in India. Also, the dataset that has been taken into consideration helped in e ective training of ANN because of its size, and the capability of ANN to learn and providing the promising forecasting for the precipitation.

Furthermore, for predicting the monthly values of precipitation, the ANN forecast model was developed by (Chantasut et al.; 2014) by taking into consideration the precipitation data from 245 meteorological stations that were situated along the Chao river in Thailand. The dataset covered 48 years starting from 1941 to 1999. The objective was to predict the total amount of precipitation of the next month by taking previous 10 months of precipitation values into account. The Backpropagation neural network has been implemented in this work by the researchers. The overall result from the research provided an excellent accuracy and showed the possibility of predicting the rainfall a year head with an acceptable accuracy.

For predicting the amount of rainfall for the next month , another research has been carried out by (Freiwan and Cigizoglu; 2005) by developing the multiple layer perceptron ANN and training them using the feed forward backpropagation method. For carrying out the research, Jordan has been taken into consideration and taken the dataset from the year 1924 to 2000. As an input the researchers have taken the precipitation total of the last 2 months and additionally considering each months periodic component. The overall research presented the satisfactory result of the prediction with RMSE lying between 25.8 to 33.6 mm and R2 values in the bracket of 0.111 and 0.465 respectively.

For forecasting the total monthly precipitation (Mar and Naing; 2008) used arti cial neural networks. Period of 1970 to 2006 has been taken in carrying out the research. The researchers used the monthly precipitation total as an input data. In their work, the development of various sets of arti cial neural network have been employed di ering in the number of arti cial neurons in them. Depending upon the type of ANN, the prediction of precipitation resulted in the RMSE values ranging between 9.9 to 22.9 mm respectively.

2.3 Forecasting of Precipitation using Deep Learning Approach

Accurate weather forecasting plays an important role in determining the right values for the parameters related to weather and also helps in accurately forecasting the possible future conditions on the basis of these parameters. (Fente and Singh; 2018) in their work presented the technique of forecasting by taking di erent parameters of weather such as wind speed, precipitation, temperature, pressure etc and using Long short term memory. For conducting this research, the authors have taken the dataset from National climate data centre covering the time period from 2007 to 2017. As an input for the neural network , the attributes are taken and training have been performed by LSTM algorithm. As a result, the result of the proposed work proved to be quite promising as compared to the other techniques for weather forecasting.

For processing the large volume of data, speedy algorithms like echo state network and deep echo state network are employed. (Yen et al.; 2019) in their study used these two algorithms for analyzing and predicting the amount of rainfall in southern part of Taiwan. Hourly data from the year 2002 to 2014 has been taken from the meteorological department for conducting this research. The work also examined the important parameters. It was found that humidity, pressure are important factors and in uence the overall performance of the DeepESN model. The result demonstrated that the correlation coe cient of deepesn was better than esn and other neural network algorithms and can be employed in forecasting the climate where high volume of data processing is required.

For accurately predicting the amount of rainfall in Camau, Vietnam, (DUONG et al.; 2018) proposed a novel approach using long short term memory recurrent neural network. Then the performance of this approach has been compared with the ANN and Seasonal ANN. For this research, the dataset has been taken from the meteorological centre in Camau in Vietnam and consists of 39 years of data starting from 1971 to 2010. For evaluating the model's performance, statistical measures like RMSE, MAE and R have been used. The proposed model performed really well as compared to the other two models in terms of these statistical measures and predicted the precipitation accurately. This suggested that the LSTM model can be used for forecasting the amount of precipitation.

(Aswin et al.; 2018) presented the technique for predicting the amount of rainfall using deep learning architecture which is LSTM and Convnet and these are used for the processing of data, training the train data and predicting the rainfall of the test data. Root mean square error and Mean absolute percentage error have been used as an evaluation criteria. The rainfall data used in this research have been taken from the NCEP centre. On a concluding note, the authors suggested to extend the work for a speci c country by using similar techniques for processing and predicting the amount of rainfall with accuracy.

2.4 Forecasting of Precipitation using Neural Networks

The study presented by (Partal et al.; 2015) makes the use of three di erent types of neural network algorithms which are regression neural network ,radial basis function, feed forward back propagation and wavelet transformation for forecasting the daily amount of precipitation. For conducting this research, the dataset has been taken from Turkish meteorological department and covers the period from January 1987 to December

2001. DWT technique has been employed for decomposing the meteorological patterns into the period series. For selecting the components appropriately, correlation between the observed precipitation and wavelet component were evaluated. The input for the hybrid model was provided by summing series that were obtained by wavelet components addition. For the estimation of precipitation, wavelet feed forward back propagation presented the best result in terms of performance. In addition to this, it was inferred that the wavelet feed forward back propagation is more e ective and e cient in successfully predicting the amount of precipitation as compared to wavelet grnn and radial basis neural network.

(Wang and Wu; 2012) presented their work for forecasting the rainfall in Guangxi, china by using hybrid Radial Basis Function Neural Network model and making the use of wavelet support vector machine regression. The proposed work has been conducted in various stages that includes the rst stage where bagging and boosting technique is used for the division of data into the training sets. In the next stage, the training sets served as an input for the RBF-NN model and then producing the RBF-NN predictors by using the principle of diversity. For the appropriate selection of the ensemble members partial least square regression has been used and this techniques helps in the prediction of the dependent variable from a large number of independent variables. Finally, for ensembling of the RBF-NN, Wavelet-SVR has been used. The proposed study has been compared with the existing ensemble techniques by comparing their mean absolute percentage error, normalized mean square error and pearson relative coe cient. The e ectiveness of the forecasting ability of the model can be measured by considering the pearson relative coe cient. Overall, the proposed technique showed promising result in the forecasting of the rainfall.

For improving the learning of Multilayer Perceptron (Beheshti et al.; 2015) employed various metaheuristic algorithms like Imperialist competitive, gravitation search and CAPSO algorithm and then CAPSO-MLP were applied on the rain dataset that was obtained from the Malaysian irrigation and drainage department for forecasting the rainfall for the next 5 years and 10 years respectively. Basically, two modes were employed which means with and without pre processing the data. The method among them were selected on the basis of their performance. The experiment demonstrated that among all, CAPSO-MLP performed better and can be considered best for forecasting with accuracy. Advantage of using CAPSO for training the neural network is that it is easy to implement and there is no need of tuning the speci c parameters. Additionally, CAPSO-MLP provided good result with the test data on coupling it with the data preprocessing technique like SSA.

Forecasting the amount of rainfall plays an important role in making crucial decisions and plans in the countries that are dependent upon the agriculture. The research conducted by (Htike and Khalifa; 2010) focuses on designing, implementing and comparing the rainfall forecast models by using the focused time delay neural network. The dataset has been obtained from the meteorological department in Malaysia covering a period from 1980 to 2009 and then converted into year, bi annual, quarter and month. Each datasets have been used for the training and testing and then the mean absolute percentage error is used for comparing the accuracy. The proposed technique has been applied on each of the datasets containing the level of rainfall in mm. Yearly data exhibited the highest

accuracy with the test data while the forecasted accuracy was found to be the least for the monthly dataset.

For forecasting the weather, speci cally focusing on the monthly amount of rainfall (Hesar et al.; 2012) in their research used Bayesian belief network by taking the dataset from weather stations in Iran covering the period of 1985 to 2011.For the structure learning the e ciency of tabu search algorithm has been analyzed followed by parametric learning of Bayesian belief network by using netica software. K2 algorithm has been taken into consideration for the purpose of structure learning. The actual values and the forecasted values are then compared with each other. As a result, the accuracy of the proposed technique is acceptable and further it can be used for forecasting the weather for other regions as well.

Rainfall is considered to be an important parameter ranging from the researches related to water resources to consideration for the issues like mitigation of disaster, irrigation, ood control etc. For getting the more attention related to rainfall forecasting for any valley, (Leng et al.; 2019) presented their work by extracting the important parameters in uencing the rainfall from MODIS satellite imagery and then they can be combined with the data obtained from the station. The retrieval model is basically based on the BP and GABP neural network. The area covered for conducting this research is North-west part of China. The result from the experiment showed that relation between the rainfall and the extracted factors and MODIS proved to be an e cient and e ective source of data for estimating the amount of rainfall. Further this research can be extended to converging the data from other satellites like GMS data and making the use of MODIS then the possibility of increasing the accuracy is possible. In addition to this, the classi cation of cloud is also possible.

3 Methodology

For the researches related to data mining, usually there are two types of methodologies involved and they are CRISP-DM (Cross-Industry Standard Process for Data Mining) and KDD (Knowledge Discovery and Data Mining). As a part of this research, CRISP-DM methodology has been adopted (Shearer; 2000). Basically, it involves 6 phases:

Business understanding

Data understanding

Data preparation

Modeling

Evaluation

Deployment

3.1 Business Understanding

The first phase of the methodology mainly emphasizes on the understanding of the business for which we are developing the solution. For meeting the requirements, the tasks are performed accordingly so that the business outcomes can be achieved.

The importance of forecasting the amount of precipitation is considered to be very important. As the changes are continuously observed in the climatic conditions so there is a requirement of effectively and accurately forecast the weather parameters. Accurate prediction plays an important part in day to day life and impacts the environment and economy of a nation to a great extent. The meteorological department works continuously in order to forecast the accurate result. Among the weather parameters, precipitation is considered to be crucial because the agricultural countries are highly dependent on it. Farmers highly rely on the amount of precipitation because it helps them in planning and harvesting. The part of the countries that do not have proper irrigation facilities for their crops and fields, the accurate prediction of precipitation can help them to a great extent. Apart from the agriculture, an accurate forecast can also help in minimizing the chances of occurrence of natural calamities like flood and drought.

The main aim of conducting the research is to forecast the daily amount of precipitation. For this, the area that has been taken into consideration is the state of Punjab in India. Punjab is the Indian state that has been chosen for conducting the research. Although the state is just 1.4 percent of the total area of the country but it produces around 14 percent of the cereals and plays an important role in the Gross Domestic Product. The state is among the top 3 producers of crops like wheat, rice, sugarcane, cotton and maize. These crops are the basic necessities and lifeline of the common people and are needed in a sufficient quantity for the survival. Hence the adequate forecast of precipitation becomes an important part for the state like Punjab. (Swati; 2018)

3.2 Data Understanding

Once the objectives of the business are understood, the next step involves understanding and collection of data for carrying out the research. It becomes really important to collect the dataset from the reliable source, checking for the completeness of the data and keeping data integrity in mind.

For performing the analysis and forecasting the amount of precipitation, the global weather forecasting dataset has been taken from spatial sciences website of Texas A&M university[1]. The dataset contains the daily observations of weather parameters covering 28 years of data. For selecting and fetching the data of the specific area, latitude and longitude information were specified. As a part of this research, the data with Longitude as 75.9375 and Latitude as 31.379 has been taken into consideration.

The dataset is publicly available and as a part of the procedure, it can be received by providing the basic details like the file format, specifying the date range along with the geographical information and mentioning the email id on which the data will be delivered.The procedure has been mentioned along with the screenshot in the configuration manual on page no 4 and 5.

[1] https://globalweather.tamu.edu/

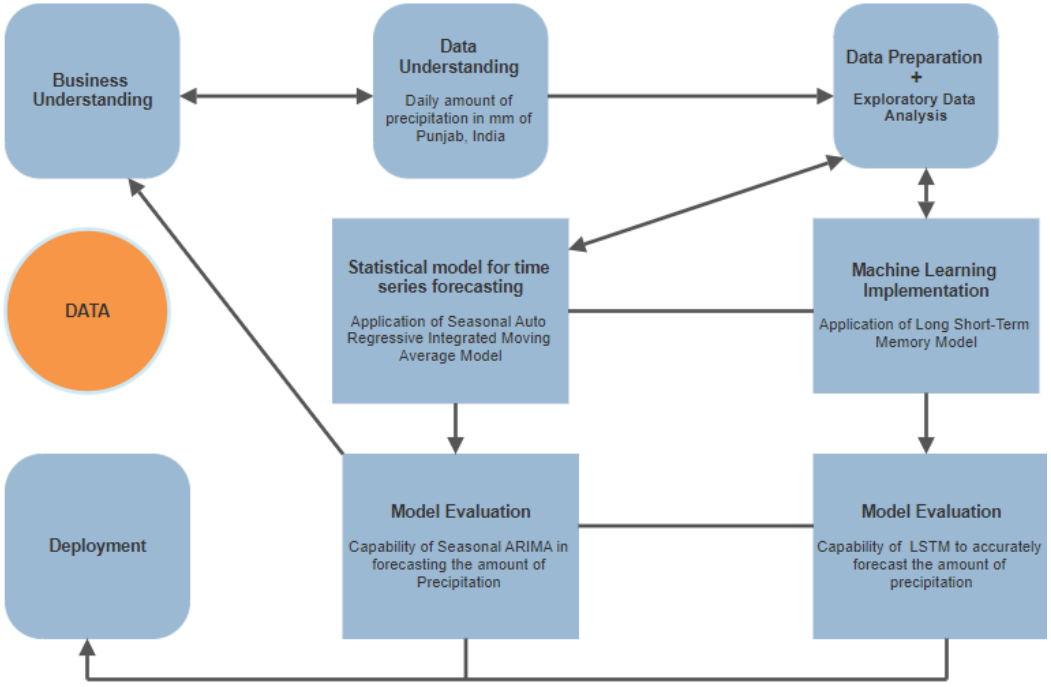

Figure 2: Design ow

3.3 Data Preparation

Data preparation is considered to an important part of the CRISP-DM methodology as in this step the raw data is preprocessed so that it can be made ready to t into the model for getting the accurate output. The daily weather data has been received from the Texas A&M University through link containing the zip le. The zip le has been extracted to get the weather data in the CSV le format. The le consists of the columns like Date, latitude, longitude, elevation, maximum and minimum temperature, precipitation, wind and solar data information. The required columns were selected and the remaining was omitted and the data has been obtained containing the daily precipitation amount along with the time period.

After getting the data, initial check on were made to nd out the occurrence of any missing values and checking if the date is in the correct format. However, the dataset had no missing values but the date was not in the proper format. Hence, the dates were converted into the datetime format using datetime module in python. The nal data has been obtained to proceed further. After this the checks were performed that includes the test for stationarity and then the serial correlation was also tested and taken into consideration.

3.4 Modelling

The next phase after preparing the data is the modeling part that involves the application of appropriate model to the cleaned dataset so that the business objectives can be achieved. The models used in the research are Seasonal ARIMA and LSTM.

Seasonal ARIMA

In this research, the traditional model of forecasting the time series includes SARIMA which is Seasonal Auto Regressive Integrated Moving Average. It is an extension of ARIMA that deals with the seasonal component. Seasonal ARIMA model is obtained and it is represented as:

$$\text{ARIMA (p, d, q) (P, D, Q) m}$$

Where, it represents the non seasonal as well as seasonal part along with the number of periods. The selection of the best model has been done by taking into account the AIC values. The parameters were selected on the basis of lowest AIC values obtained. The data has been forecasted by keeping Box and Jenkins methodology in mind which is identi cation of model, estimating the parameters and diagnostic check. (Nanda et al.; 2013)

Long Short Term Memory (LSTM)

Long Short Term Memory was presented by Hochreiter and Schmidhuber (Hochreiter and Schmidhuber; 1997) to address forecasting and having the capability of memorizing the sequences of data. LSTM is basically a RNN. Recurrent neural network are the special networks that consists of the hidden layer which has the capability of storing the information that has been gathered from the earlier steps and helps in performing the further tasks for the next step. In LSTM, capturing and storing of data streams takes place in the set of cells. These cells have a resemblance with the transport line and connect the modules so as to convey the data from past and gather for the present.

For controlling the states of cell, there are total three types of gates namely; forget, memory and output gate. Because of these gates, disposing, ltering and adding of data in each cell takes place for the next one. In addition to this, LSTM networks helps in solving the long lag relationship in the data related to time series and useful in addressing the problem of vanishing gradient (Gers; 1999).

For carrying the research, this model has also been used in forecasting the daily amount of precipitation.

3.5 Evaluation

After the successful construction and execution of the models, it is essential to perform the check if the outcomes are as per the expectation or not. For this evaluation is considered to be very important as with this the comparison can be made between the forecasted and the expected values.

$$\text{RMSE} = \sqrt{\frac{\left[\sum_{t=1}^{N} \left(y_t^o - y_t^c\right)^2\right]}{N}}$$

Root mean square error (RMSE) has been taken into consideration for evaluating the models. It basically, outlines the result presented by the model by di erence between the actual and the predicted values. (Barnston; 1992)

Additionally, the persistence model has also been used in order to evaluate the performance of LSTM. This model is used to forecast the time series behavior in future by adopting the value of previous time step to forecast the value at next step. Therefore for evaluating the performance of the LSTM model, it can be used as a baseline.

3.6 Deployment

The nal phase of the methodology is the deployment phase which involves the successful deployment of the nal model stating the complete background, problem, objective and its solution. The research report would provide the detailed description about the models would be useful for the audience doing their research in the eld of precipitation forecasting.

4 Implementation

The implementation phase is considered to be an important phase as it involves implementing the proposed model on the dataset. This phase, is describing about the steps that have been involved in implementing the Seasonal Auto Regressive Integrated Moving Average and Long Short Term Memory model. The details have been covered starting from setting up the environment, preparing the dataset and experimental design.

4.1 Setting up the environment

Python 3.7.3 and jupyter notebook 5.7.8 have been used in the implementation of the models. For performing the SARIMA, the statsmodels library have been used while on the other hand for LSTM the use of keras with tensor ow has been made.

4.2 Preparing the dataset

Seasonal ARIMA

The global weather dataset that has been obtained contains the daily observation data covering a period of 28 years starting from January 1985 to December 2013 and having columns like Date, latitude, longitude, elevation, maximum and minimum temperature, precipitation, wind and solar data information. Out of these columns, for carrying our research we have taken Date and Precipitation column.

After getting the data, initial check was performed on the occurrence of any missing values. On further checking it was found that the dataset is not having any missing values. Next, on checking the data type of the date column it was found to be object. Therefore the dates were converted into the datetime format using the datetime module in python. Finally, the Date column has been indexed and the nal data was obtained.

The dataset was then divided into the train and test set. For this, the train set consists of the data containing daily observations from January 1985 to December 2005 and the test set containing the values from January 2006 to December 2013.

Long Short Term Memory (LSTM)

For conducting the research further, LSTM model has been used which takes into account the columns Date and Precipitation. The objective is to forecast the daily amount of precipitation using LSTM model.

The numpy array has been extracted from dataframe and the value is converted into the oating values, it is done to make it suitable during the modeling of a neural network. Normalizing the dataset is important while working with LSTM because of the sensitivity towards the scale of our input data. Rescaling of the data in the range of 0 to 1 is known as the normalizing technique. For this, we have used scikit learn library in python using the minmaxscaler.

While working with time series problem, splitting of dataset into train and test set is an important task. For splitting the data, instead of random splitting, we have taking into consideration the index for the split point and then the data is splitted respectively.The next step is the creation of dataset. Now basically two arguments will be taken by the function i.e. our dataset and look back i.e. for the prediction of next time period how many previous time step that can be used as an input variable.

For LSTM, the input data must be in the array structure containing sample, timestep and features and this can be done by using the numpy.reshape(). Once the above steps are done, the next step is the designing and tting of our model. Here, it is important to specify the number of neurons, batch size and the epoch rates.

4.3 Experimental Design

Checking for Stationarity

Augmented Dickey Fuller test is a statistical test that can be used to nd out whether the time series is stationary or non stationary (Dickey and Fuller; 1981). It is also known as the unit root test. Here, in this test we formulate the null and alternate hypothesis. Null hypothesis suggests that the time series is non stationary or it is having a unit root, while on the other hand an alternate hypothesis states that the time series is stationary and it is not having a unit root.

For interpreting the result, the p value is observed and if the p value is more than 0.05 then the series is non stationary while if it is less than 0.05 then the series is stationary.

Once the test is performed and the value of test statistic is obtained then it is compared with the critical values. The more negative is the value of test statistic, the more chances of rejecting the null hypothesis.

As we can see from the ADF test performed on our time series data, it shows that the p-value is less than 0.05 and the test statistics value is far negative as compared to the critical values. From here it can be inferred that the time series is stationary and the null hypothesis can be rejected and the same has been depicted by the test.

The output for the ADF test can be referred from the con guration manual.

Forecasting time series using Seasonal ARIMA

There are four steps which are considered important to develop ARIMA model. The rst step involves the identi cation of model, that is to make a check for the stationarity. Next step is the estimation of model and it is crucial in determining if the model that ts the data well. Diagnostic check for model is performed in the next step by having a look at the residuals.Finally the forecast is made and the observed and predicted values are compared with each other to get an insight about how well the prediction has been made by the model.

For forecasting the time series data, ARIMA model is considered to be an important forecasting method.The meteorological variable that has been taken into consideration for the research and forecast here is the daily values of Precipitation. When the seasonal terms are added to the ARIMA model then the Seasonal ARIMA model is obtained and it is represented as:

$$\text{ARIMA (p, d, q) (P, D, Q) m}$$

Where, it represents the non seasonal as well as seasonal part along with the number of periods.

By the values of (p, d, q) (P, D, Q) m nalizing of the model takes place and for determining these values, AIC values have been taken into account and the lowest value of AIC will be considered for selecting the model for forecasting. The code uses the SARIMAX function from the statsmodel for tting the SARIMA model. SARIMAX is available in python and it s helpful in forecasting the further points in the time series. In our case, the lowest values of AIC were been obtained for ARIMA(1,1,1)x(0,1,1) followed by ARIMA(1,1,1)x(1,1,1). Therefore the rst value is taken into consideration.

Next step is to apply the diagnostic check and the summary() which in return provides the result of the SARIMAX having useful information. The coef column represents the information about the features and their weight along with the impact it is having on the time series. As we can see the p value of weights is less than 0.05 so all can be retained in the model.After this, Next step involves diagnosing the model.

The model is nally obtained for the time series and we can use this for producing the forecast. The predicted values are compared with the actual values which in return help us in understanding the forecast accuracy that the model has provided. For forecasting our time series, conf int() and get prediction() help in getting the values along with the con dence interval.

Finally, the root mean square value has been calculated to check for the result produced by our model in producing the forecast for the precipitation.

The output for the Seasonal ARIMA model can be referred from the con guration manual.

Forecasting time series using LSTM

Once we get the preprocessed data and splitted into training and testing set, the next task is to t the LSTM model to the training set. The complete network consists of three types of layers and they are input, output and hidden layer. Basically, the hidden layer is useful in transforming an input in such a way that can helpful for the usable for the output layer. For making the output, the weighted sum for the inputs are calculated by the neurons and feed the value further to the activation function.

The optimizer that has been used here is adam and loss function is the mean squared error.In addition to this, the activation function is linear. Adam has been used as it blends the best properties of RmsProp and AdaGrad algorithm. (Kingma and Ba; 2015)

Once the training data has been tted to LSTM, it can be used for making the forecast and once it is done, there is a need to invert the transform so that the value can be brought back to original dimension. It is a necessary step because it helps in calculating the root mean square error in the last which is our evaluation metric.

5 Evaluation

Seasonal ARIMA: For validating the forecast result produced by the Seasonal ARIMA model, we compared the predicted and the actual values in the time series that will be helpful in understanding the forecast accuracy. For time series forecast, the get prediction() and conf int() help us in obtaining the respective values and the con dence interval associated with it. For evaluating the performance of the model that how well the values are forecasted, root mean square error has been used. By taking the value of m as 30 and 7 respectively, the RMSE value of 12.3015 and 11.5734 has been achieved by the model .

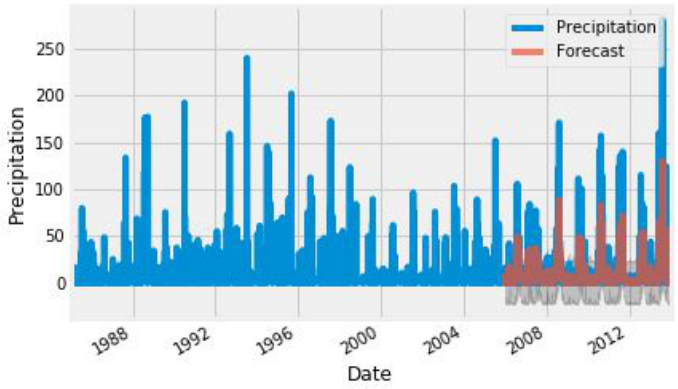

Figure 3: Seasonal ARIMA Forecast

Long Short Term Memory: For validation of the result obtained from LSTM model below are some of the evaluation measures taken which includes:

– **Persistence Model**

While working with the time series related problems, establishment of baseline is essential as it can give us an insight that how other available models can actually work on our problem. For the establishment of the baseline model, in our research we have implemented a persistence model which is also known as the na•ve model. (Brownlee; 2017)

Basically the persistence model, takes the value from the last time step and then predict the outcome that is expected at the next step. For implementing the persistence model, following steps have been taken. The rst and the foremost step are de ning the supervised learning problem. Here, the dataset is loaded and the lagged representation has been created. Say for an instance, the input variable will be column in t-1 and the output variable will be in t+1.

Once the step 1 is completed, the next task is to divide the entire dataset into the training and testing sets. Then as a function, the persistence model will be de ned, which will be returning back the value that has been supplied as an input.Now the evaluation of the model can be performed on the testing set and this is done with the help of method that is known as walk forward validation.

Once, if the predictions are done for every time step from the training set, then the comparison can be made with the expected outcome or the values and then the root mean square error can be calculated. After the successful run of our persistence model, the RMSE values of 16.32 at step t+1, 17.74 at t+2 and 17.96 at t+3 have been achieved.

– **Cross Validation of Time Series**

Instead of splitting the data normally into train and split set, in time series

cross validation is also used. Basically, for a test set series is there that contains the single value and the train set consists of the time series value that is present before the test set in terms of their presence in the original time series. With this, while forecasting no future observation will be used.(Hyndman; 2016). After running the cross validation on our dataset, the RMSE of 14.77 has been achieved.

– **Splitting the data in sequential manner**

For tuning the parameters, the di erent combinations of epoch, batch size and their corresponding rmse values have been analysed so as to derive to the conclusion of choosing the best parameters for tting into our LSTM model.

Sequence of data is really important for forecasting the time series and in LSTM as well depends upon the sequence for learning, therefore the dataset used in the research has been splitted by keeping this in mind. For this the split ratio is taken in such a way that the training set contains the daily observations of precipitation from January 1985 to December 2005 while and the test set contains the values from January 2006 to December 2013.

Epoch	Batch size	Root mean square error
50	10	8.3264
100	10	8.2821
200	10	8.7164

Figure 4: RMSE with di erent Epoch

As it can seen above that for getting the best parameters, rst of all the batch size has been kept as 10 and epochs were increased from 50 to 200. In this way, the main observable column is the root mean square error and its least value has been achieved for the combination of 100 epoch and batch size as 10.

Now, in the second table the epoch are taken as 100 that has been achieved in the earlier step and were executed with the batch size of 20 and 30 respectively. The best combination has been achieved in 100 epoch and 30 batch size with the rmse value of 8.0437.

Epoch	Batch size	Root mean square error
100	20	8.6601
100	30	8.0437

Figure 5: RMSE with di erent Batch size

Finally, these were tted into our LSTM model containing 100 epoch, 20 batch size and 128 neurons. The unit of our daily precipitation data values used in our dataset are in mm and the RMSE value of 8.0292 and mean absolute error

of 2.9635 have been obtained after the successful run of our model. Below gure is showing forecast of precipitation through LSTM model.

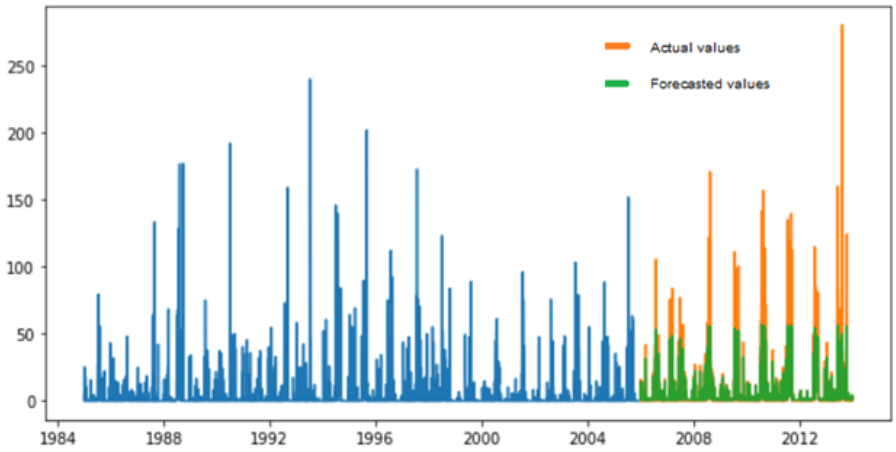

Figure 6: LSTM Model Forecast

6 Discussion

As a part of this research, for forecasting the daily amount of precipitation in Punjab in India, basically the Seasonal ARIMA model and Long Short Term Memory model have been implemented. The model evaluation was done on the basis of root mean square error and checked for the actual and the predicted values. It has been observed that for the Seasonal ARIMA model the RMSE value obtained was 12.3015 and 11.5734 for m as 30 and 7 respectively. In this research, for establishing the baseline, we have implemented the persistence model, also known as na•ve model, therefore it served as the baseline model for evaluating how the other model presented their forecast. The RMSE values of 16.32 at t+1, 17.74 at t+2 and 17.96 at t+3 have been achieved. In addition to this, cross validation of time series has also been performed. On running this, the RMSE value of 14.77 has been achieved.

For tuning the parameters for LSTM model, di erent combinations of batch size and epochs were performed by carefully observing the RMSE value, the best and the least value was obtained for 100 epochs, 30 batch size and thus after this the model has been tted. The forecast of daily precipitation by the LSTM model was made with the RMSE value of 8.0292. In addition to this, for this model the MAE value was also calculated that came around 2.9635. So, for forecasting the daily amount of precipitation this model performed really well with least RMSE values.

7 Conclusion and Future Work

The research started with an objective for forecasting the daily amount of precipitation over Punjab,India. As India is an agriculture country therefore forecasting the amount of precipitation becomes really important. This can help the farmers in making e ective plan by observing the forecasted result and in the meanwhile the early warnings for natural

calamities can also be raised. The work can also help the meteorological centres for forecasting the precipitation of the area that has been covered as a part of this research.

For this, the dataset has been sourced from spatial sciences website of Texas A&M university covering the daily observations of weather variables. Proceeding further, Seasonal ARIMA and Long Short Term Memory model were implemented.The evaluation has been made by taking root mean square error into consideration. Additionally, baseline model like persistence model has been presented and also performed the time series crossvalidation. The performance of LSTM model represented a signi cant result with least value of RMSE. The model performed well in forecasting the daily amount of precipitation.

In addition to this, it has been observed that the deep learning approach of using LSTM can outperform the traditional approaches in presenting the accurate result to a great extent.

As a part of the future work, di erent location and dataset can be taken for conducting the research using the techniques mentioned here and the check can be performed by tuning the parameters and observing the impact on the overall result.

8 Acknowledgement

I would like to devote my thanks to my supervisor Prof. Noel Cosgrave for his continous support and guidance througout my research phase. Also I would like to thanks my Parents and friends for motivating me to deliver the project to the best of my abilities.

References

Aswin, S., Geetha, P. and Vinayakumar, R. (2018). Deep learning models for the prediction of rainfall, *2018 International Conference on Communication and Signal Processing (ICCSP)* .

Barnston, A. G. (1992). Correspondence among the correlation, rmse, and heidke forecast veri cation measures; re nement of the heidke score, *Weather and Forecasting* **7**(4): 699{709.

Beheshti, Z., Firouzi, M., Shamsuddin, S. M., Zibarzani, M. and Yusop, Z. (2015). A new rainfall forecasting model using the capso algorithm and an arti cial neural network, *Neural Computing and Applications* **27**(8): 2551{2565.

Brownlee, J. (2017). *Deep learning with python: develop deep learning models on Theano and TensorFlow using Keras*, Machine Learning Mastery.

Chantasut, N., Charoenjit, C. and Tanprasert, C. (2014). Predictive mining of rainfall predictions using arti cial neural networks for chao phraya river, *4th International Conference of The Asian Federation of Information Technology in Agriculture and The 2nd World Congress on Computers in Agriculture and Natural Resources* p. 117{122.

Den, J. E. (2012). Climate and meteorological information requirements for water management, p. 42.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica* **49**(4): 1057.

DUONG, T. A., BUI, M. D. and RUTSCHMANN, P. (2018). Long short term memory for monthly rainfall prediction in camau, vietnam.

DUTTA, P. S. and TAHBILDER, H. (2014). Prediction of rainfall using data mining techniques over assam, *Indian Journal of Computer Science and Engineering (IJCSE)* **5**(2): 85{90.

Fente, D. N. and Singh, D. K. (2018). Weather forecasting using arti cial neural network, *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* .

Freiwan, M. and Cigizoglu, H. K. (2005). Prediction of total monthly rainfall in jordan using feed forward backpropagation method, *Fresenius Environmental Bulletin* **14**(2): 142{151.

Gers, F. (1999). Learning to forget: continual prediction with lstm, *9th International Conference on Arti cial Neural Networks: ICANN 99* .

Hesar, A. S., Tabatabaee, H. and Jalali, M. (2012). Monthly rainfall forecasting using bayesian belief networks, *International Research Journal of Applied and Basic Sciences* **3**: 2226{2231.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory, *Neural Computation* **9**(8): 1735{1780.

Htike, K. K. and Khalifa, O. O. (2010). Rainfall forecasting models using focused time-delay neural networks, *International Conference on Computer and Communication Engineering (ICCCE10)* .

Hyndman, R. J. (2016). Cross-validation for time series.
URL: *https://robjhyndman.com/hyndsight/tscv/*

Jakaria, A. H. M., Hossain, M. M. and Rahman, M. (2018). Smart weather forecasting using machine learning: A case study in tennessee.

Kingma, D. P. and Ba, J. L. (2015). Adam : A method for stochastic optimization, *ICLR* .

Leng, C., Yi, S. and Xie, W. (2019). Estimation of rainfall based on modis using neural networks, *2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)* .

Mar, K. W. and Naing, T. T. (2008). Optimum neural network architecture for precipitation prediction of myanmar, *World Academy of Science, Engineering and Technology 48* p. 130{134.

Nanda, S. K., Tripathy, D. P., Nayak, S. K. and Mohapatra, S. (2013). Prediction of rainfall in india using arti cial neural network (ann) models, *International Journal of Intelligent Systems and Applications* **5**(12): 1{22.

Parmar, A., Mistree, K. and Sompura, M. (2017). Machine learning techniques for rainfall prediction: A review.

Partal, T., Cigizoglu, H. K. and Kahya, E. (2015). Daily precipitation predictions using three di erent wavelet neural network algorithms by meteorological data, *Stochastic Environmental Research and Risk Assessment* **29**(5): 1317{1329.

Philips world atlas (2012). Philips.

Razeef, M., Butt, M. A. and Baba, M. (2018). Comparative study of rainfall prediction modeling techniques (a case study on srinagar, jk, india), pp. 13{19.

Sahai, A. K., Soman, M. K. and Satyan, V. (2000). All india summer monsoon rainfall prediction using an arti cial neural network, *Climate Dynamics* **16**(4): 291{302.

Sawaitul, S. D., Wagh, K. P. and Chatur, P. N. (2012). Classi cation and prediction of future weather by using back propagation algorithm-an approach, p. 4.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining, *Journal of Data Warehousing* **5**(4).

Swain, S., Patel, P. and Nandi, S. (2017). A multiple linear regression model for precipitation forecasting over cuttack district, odisha, india, *2017 2nd International Conference for Convergence in Technology (I2CT)* .

Swaminathan, M. S. (1998). Climate and sustainable food security, pp. 3{10.

Swati, A. (2018). Agricultural statistics at a glance 2017.

Wang, L. and Wu, J. (2012). Application of hybrid rbf neural network ensemble model based on wavelet support vector machine regression in rainfall time series forecasting, *2012 Fifth International Joint Conference on Computational Sciences and Optimization* .

Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E. and Chen, C.-C. (2019). Application of the deep learning for the prediction of rainfall in southern taiwan, *Scienti c Reports* **9**(1).

Yeshwanth, M., Kumar, P. R. S. and M.e., P. D. G. M. (2019). Comparative study of machine learning algorithms for rainfall prediction, p. 677{681.