Region-wise Rainfall Analysis and Prediction using ML

Bhnau Pratap Rai[1], Ashutosh Kumar Singh[2], Nidhi Shukla[2], Md. Aqib Khan[2]and C. S. Raghuvanshi[1]

[1]Department of Computer Science and Engineering, Rama University, Mandhana, Kanpur, 209217, U.P., India.

[2]Department of Computer Science and Engineering, United College of Engineering Research, Naini, Prayagraj, 211010, U.P., India.

drcsraghuvanshi.fet@ramauniversity.ac.in

Abstract

Rainfall prediction is difficult and indeterminate responsibility that leaves impact on the life of nature, animals and humans. Accurate and up-to-the-time forecasting can be beneficial for human society and economy. Heavy rainfall can cause many disasters like flood, land slides etc. and low rainfall also can cause disaster like drought, damage of corps etc. Climate changes had been become a major concern for last decade. It also affects the weather conditions by which a drastic change in pattern of rainfall had been noticed lately. The rainfall depends on the various factors that are: temperature, pressure, humidity, precipitation and wind speed. Various tools and method are available for predicting rainfall but the achievement of accuracy is still a concern. The other methods which exist becomes less effective whenever a huge and diverse dataset are used. The machine learning algorithm such as: linear regression, Decision Tree, Ridge and LASSO are used for prediction of rainfall. This model intends to provide the easy and accurate technique for the rainfall prediction. It also provides comparative study of papers among the different machine learning techniques for rainfall prediction. The experiment is done by designing the application of framework using real time dataset.

Keywords: Machine Learning, Rainfall, Rainfall prediction, Region wise rainfall, Region wise rainfall prediction

1 Introduction

Region-wise rainfall prediction is based on machine learning models that predict the amount of rainfall over a different region or state of India. The rainfall prediction helps in to predict the state of atmosphere for the specific location. It is one of the difficult and uncertain tasks that have a significant impact on human society. An accurate and well timed weather forecasting can dynamically supports to reduce human and financial loss. Specially in the case of farming which is currently the main financial pillar of our country. This work contains the machine learning techniques based approaches to create a framework that can predict, how much rainfall we can expect from any region, according to the previous data analysis of that region. As related to previous database it will make aware that what will be the rainfall scenario in every month of that particular year so, that if a person plans to go anywhere or going to organize a party on any month, one can aware of rainfall. By this analysis we can easily classify each region into three category of rain pattern that is high, medium, and low. All this information are deployed or provide on web interface (website) i.e. being created using Django framework. So that one can easily access that information on internet and use it on their day to day life and can aware others about the rainfall prediction.

There are various techniques that applied for this purpose such as: classification, regression according to the requirements. For the prediction accuracy, the error between actual and predicted data is calculated. Different techniques produce different results. So, it is important to choose the right algorithm based on the requirements.

- **Regression analysis**: Regression algorithms are used in that place where, if there is a connection between the input variable and output variable. It is generally used for the prediction of continuous variables such as: Market Trends,Weather forecasting etc. The general form of regression is shown in Equation (1)

$$y = \beta 0 + \beta 1 + \epsilon \quad (1)$$

  Where, $\beta 0$ and $\beta 1$ are parameters, and ϵ is a probabilistic error term. Regression analysis is a vital tool for modeling and analyzing information.

- **Classification analysis**: The classification algorithm is based on supervised learning in which by training the data, the classification of new observations into a number of class and groups are identified. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. In this the output variable of classification is a category as compared to regression it is not a value. As the Classification algorithm is supervised learning technique, therefore it takes labeled input data.

- **Linear regression algorithm**: Linear regression algorithm represents a linear relationship between variables.The relationship among a dependent and one or more independent variables are considered. It focus on how the value of one the dependent variable affects the value of an independent variable. For this project dependent variable is annual rainfall and independent variables are sessional data (Jan-Feb, Mar-May, Jun-Sep, Oct-Dec) of that year. Linear regression is categorized into two types of the algorithm: Sample Linear Regression and Multiple Linear Regression.

- **Ridge Regression algorithm**: Ridge Regression is based on model tuning and applicable on multi-collinearity type of data. In this algorithm due to multi-collinearity variance become large and least squares are unbiased so, for the result L2 regularization is applied but the accuracy in result is not up to the mark. The calculation of cost function for ridge regression is mentioned in Equation (2)

$$Min \quad \| Y - X(\Theta) \|^2 + \lambda \, \| (\Theta) \|^2 \quad (2)$$

  Where, λ denotes penalty term. α is controlling the value of λ, if the value of α is large, penalty is bigger hence magnitude of coefficients is reduced.

- **Decision Tree Algorithm**: It use tree-structure to create the regression models. By dividing the data into small subsets, the decision tree steadily grows. After completion it structured as two nodes: decision nodes and leaf node. Decision node must contain at least two branches and leaf node shows decision.

- **Lasso Regression**: Lasso regression enables to shrink or regularize the regression coefficients to avoid over-fitting. It penalizes less valuable features of dataset thus by eliminating respective coefficients it makes it zero. In high dimensonality and correlation lasso regression can be applied.

Rainfall predictions are becoming necessary part for farming and economic areas. The contribution of this paper are as follows:

- We are predicting rainfall status of India.
- Different types of machine learning models are applied and for the accurate result comparison is performed between them.
- The experiment is performed using real time dataset for testing the effectiveness of the work.

- Evaluation matrices RMSE, MAE, MSE and R-Square are used for comparison and also the performance is presented though using graphs.

Remainder paper are structured as follows. Section 2 presents the related work. Section 3 contains the proposed work in which dataset used, evaluation matrices and implementation of the work is mentioned. Results and discussion of the work are shown in section 4. The last section 5 includes the conclusion of the work.

2 Related Work

Authors of [5] used data mining for monthly rainfall prediction. The data used is from locally Regional Meteorological Center contains six years of records of Guwahati, Assam and India. In this paper traditional statistical technique and multiple linear regression is used for prediction.

In [4], author use neural network approach with Lasso regression. It predicts annual rainfall and the rainfall between two months of transition. The author of [10] use ANN with ARIMA technique for the rainfall prediction of Hyderabad. Author of [2] had included two primary parameters for minimum and maximum temperature whereas authors of [1] applied conjugate gradient descendent and levenberg Marquardt algorithms for learning and training but the results of [1, 2] are similar. In [12], used modular artificial neutral network for India and China rainfall forecasting. Authors of [9] used large dataset which contains 115 years of rainfall data of India. The models used in research are multiple linear regression, supple vector regression, and lasson regression. For the season wise rainfall prediction i.e. rabi, karif, and zaid in India, authors of [11], applied LR in which if one season value is given then the other season predictions are made. In is easy to find the other missing season predictions from one season predicted value.

Authors of [3], predicted summer monsoon and post monsoon specifically for Kerala. The models used are based of artificial intelligence which are K-Nearest Neighbor (KNN), Artificial Neutral Network (ANN), and Extreme Learning Machine (ELM). Comparing the prediction performance among these models results that ELM performs better than KNN and ANN. The Focused Time Delay Neural Network (FTDNN) results in highest accuracy when year, biannual, quarter, month wise data is considered [6].

In [8], developed K mean clustering technique using decision tree algorithm. It is used in a river basin for rainfall states generation from large scale atmospheric variables. K-mean clustering is applied on past multi-site rainfall data for rainfall state. The short-term rainfall is predicted by authors of [7] using empirical method. The data of specific region is analyzed for particular month for five years. The elements are grouped by clustering method.

The table 1 presents categorization of different approaches of rainfall prediction.

3 Proposed Work

The proposed model helps to forecast the amount of rainfall of various regions of India by using machine learning algorithms. There are four different type machine learning techniques we have used to predict the rainfall. The dataset collection is performed using the Kaggle dataset. There are four different techniques to predict rainfall, linear regression, ridge, LASSO and Decision tree methods. This model associated with the proposed methods are input of data, pre-processing of the dataset, splitting the dataset into training and testing dataset, testing of dataset comparing all the algorithms, giving the best algorithms which can predict the best possible value of rainfall, in the end getting the more accurate result of rainfall prediction.

3.1 Dataset

We use the rainfall dataset which was collected from Indian Metrological Department and *mousam.imd.gov.in*, from Government of India portal. It contains rainfall data from 1901–2020 for every region in India.

3.2 Implementation

- **Prepare Dataset:** We have prepared our dataset from various datasets taking the required attributes that are useful for our case study. There is a basic understanding of dataset before moving further.

Table 1 Categorization of different approaches of rainfall prediction

Authors	Region	Dataset Time Period	Techniques	Accuracy Measure	Rainfall Predicting Attribute
Dutta et al. [5]	Local	Yearly(2007-2012)	Multiple regression	R-squared	Wind, temperature, precipitation latitude-longitude, sea surface
Dutta et al. [4]	Global	Yearly(1901-2015)	Lasso regression and neural network	MAE,MSE, R-Square and RMSE	January – December Data
V. Somvanshi et al. [10]	Local	Yearly (103 years)	ANN, ARIMA RMSE, MAE	Humidity, Min-Max temperature.	
Chattopadhyay et al.[2]	Global	Monthly	Artificial Neural Network (EBP)	MSE	Temperature, Rainfall.
Chattopadhyay et al. [1]	Global	Monthly	Multilayer perceptron	MSE	Min-Max temperature
Mohammed et al. [9]	Global	Yearly(1901-2015)	Multiple linear regression, supple vector regression, lasso regression	R-square, MAE	January- December
Thirumalai et al. [11]	Global	Linear regression	RMSE	Seasons Data	
Dash et al. [3]	Local	Yearly(1871-2016)	KNN, ANN, ELM	MAE	Humidity, Min-Max temperature
Htike et al. [6]	Global	Yearly, Monthly, Biannually, Quarterly	FTDNN	MAPE	Temperature, solar radiation. evaporation
Kannan at el. [8]	Local	Daily (50 years)	Decision tree, CART, K-mean	MSE	Temperature, pressure, wind speed,rainfall
Kannan at el. [7]	Global	Quarterly-(5years)	Regression	MSE	Min-Max temperature, wind direction, humidity, rainfall pressure

- **Data Pre-processing:** Data Pre-processing is the most vital step while preparing our dataset for model training. Data is often inconsistent, incomplete, and consists of noise or unwanted data. So, prep-rocessing is required. It involves certain steps like handling missing values, handling outliers, encoding techniques, scaling. Removing null values is most important because the presence of null values will disturb the distribution of data, and may lead to false predictions.
  There are different stages of data pre-processing that are as follows: Data Arrangement, the data selected was not in an appropriate format for the models. It is need to be converted in material document for achieving the appropriate format of data. Import dataset and libraries, the libraries used in models are: pandas, seaborn, matplotlib and numpy which are used for reading the imported dataset, manipulation and visualization of data, and for graph plotting. Null values, to remove null values, there are two options one is to remove the entire row which contain null values and another is to calculate the mean value of that row or column. Splitting the dataset, there is needed to split dataset for test and train, select independent variables columns and dependent variable. The dataset splits into two sets: test data and train data into the ratio of 7:3, splitting of dataset depends upon the volume of data.

- **Implementation:** The implementation of the model contains entities such as: user input, pre-process, selection of algorithm (Linear Regression), splitting of dataset, training, testing and prediction at the end. The user input of the data is from local system. Then the pre-process step takes place, and by plotting the dataset into graphs for study. Once the pre-processing is done the algorithms applied for prediction. After that the predicted outcome is generated. The accuracy of all the algorithms are compared for testing the better algorithm.

- **Model Training:** We chose four algorithm to train our model. In Regression, accuracy can be measured by using R2-Score, Mean Squared Error [MSE] and Root Mean Squared Error [RMSE]. The model should be imported from the Sklearn package and then trained. We have to install the Sklearn package and then import it. We have defined a model method to call various models. The model method contains the training and validation statements that will train and test the dataset. MSE value, RMSE value, R2-score are also calculated after training and testing the model. The test dataset will check whether our trained model is efficient for real-time data or not.

Algorithm 1 Pseudo-code for applying the methods for rainfall prediction

1: Select the model: Linear Regression Algorithm, Ridge Algorithms, Decision Tree Algorithms, LASSO Algorithms.
2: Training model by fitting train (X, Y) dataset.
3: Testing Model using predict *X test* dataset
4: Plotting graph to visualize the difference between actual and predicted value.
5: Calculating Mean Absolute Error using metrics mean absolute error(*Y test, Y pred*)

- **Evaluation Metrics:** The error are calculated by using evaluation metrics which are MAE, MSE, and R-square. R-square- It represents the coefficient of determination and proportion of the variance in the dependent variable. The R-square value is always less than one as it is scale-free score. The formula of R-square is shown in Equation (3).

$$R^2 = 1 - \frac{\sum (y_1 - \hat{y})^2}{\sum (y_i - \hat{y})^2} \quad (3)$$

Where, $\hat{y}$ - predicted value of y and y_i – mean value of y.

Mean Absolute Error (MAE)- It calculates the average of absolute difference between actual and predicted values.The formula of MAE is shown in Equation (4).

$$MAE = \frac{1}{N} \left| (y_i - \hat{y}) \right| \quad (4)$$

Where, $\hat{y}$ - predicted value of y and y_i – mean value of y.

Mean Squared Error (MSE) calculates the mean of the squared difference between the original and predicted values and the variance of the residuals. The formula of MSE is shown in Equation (5).

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad (5)$$

Root Mean Squared Error -RMSE calculates the square root of MSE and the standard deviation of residuals. The formula of RMSE is shown in Equation (6).

$$RMSE = \sqrt{MSE} \quad (6)$$

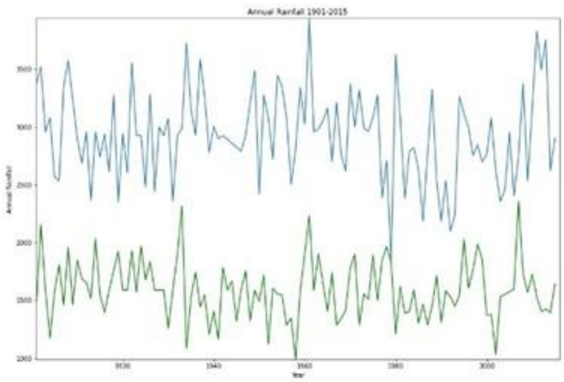

Fig. 1 Difference between rainfall patterns

4 Results and Discussion

The results obtained are visualized which help to understand the numerical more easily by representing numerical data into graphical form. Different types of graphs are generated in execution of process to represent the outcomes of the models.

Fig. 1 is able to take random region and month helps in to understand the difference between rainfall patterns from year 1901 to 2020 (blue line represent "first input" and green line represent "second input"). Fig. 2 helps in to visualize which region and year have received highest rainfall and lowest rainfall in India. For this the group each region with year from 1901 to 2020 sorted in descending order. The Fig. 3 helps in to compare amount of rainfall in a region corresponding to year (1901 to 2020) received in India in form of line graph and Fig. 4 shows the same using bar graph. The Fig. 5 take any two random region and number of month as input fields and helps

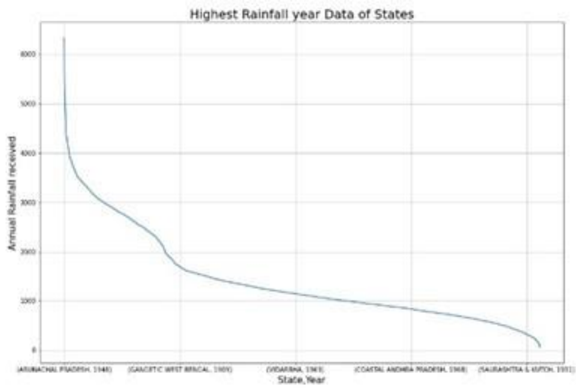

Fig. 2 Highest and lowest rainfall

in to understand the difference between rainfall patterns from year 1901 to 2020 corresponding to given parameter (blue bar represent “first input” and orange bar represent “second input”). The Fig. 6 helps in visualize amount of average rainfall corresponding to each month have received in India using bar graph. The Fig. 7 help to visualize how much average session rainfall corresponding to each region have received in India using bar graph. To calculate average of each session (Jan-Feb, Mar-May, Jun-Sep, Oct-Dec) rainfall data corresponding to region from year 1901 to 2020. A heat map, Fig. 8 is a two-dimensional representation of information with the help of colors. Heat maps can help to visualize simple or complex information. This graph help to visualize Heap map of Indian rainfall pattern and shows how each columns correlated with each other. Fig. 9 help to visualize overall rainfall in India from year 1901 to 2020. Fig. 10 help to visualize box plot of annual rainfall of each region of India from year 1901 to 2020.

5 Conclusion

The large portion of Indian economy is dependent on rainfall as agriculture sector is our primary sector. So, rainfall is considered as primary concern for agricultural enterprise. We used methods: linear regression, Decision Tree, LASSO, ridge for the prediction of rainfall. After comparing, we understand that the accuracy for Linear regression is around 99% and RMSE is around 0.0654892. The results of regression is best among all. The results of analyzing the datasets are shown through graph. There is a need that dataset must be free from all kinds of noise, inconsistency, overfitting and other odds like null values that may affect the performance of the model. We explored different methods in this paper for the prediction of rainfall. In future we will apply more methods to get the best accurate result as possible.

References

[1] Surajit Chattopadhyay and Goutami Chattopadhyay. Comparative study among different neural net learning algorithms applied to rainfall time series. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(2):273–280, 2008.

[2] Surajit Chattopadhyay and Manojit Chattopadhyay. A soft computing technique in rainfall forecasting. *arXiv preprint nlin/0703042*, 2007.

[3] Yajnaseni Dash, Saroj K Mishra, and Bijaya K Panigrahi. Rainfall prediction for the kerala state of india using artificial intelligence approaches. *Computers & Electrical Engineering*, 70:66–73, 2018.

[4] Kaushik Dutta and P Gouthaman. Rainfall prediction using machine learning and neural network. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(1): 1954–1961, 2020.

[5] Pinky Saikia Dutta, Hitesh Tahbilder, et al. Prediction of rainfall using data mining technique over assam. *Indian Journal of Computer Science and Engineering (IJCSE)*, 5(2): 85–90, 2014.

[6] Kyaw Kyaw Htike and Othman O Khalifa. Rainfall forecasting models using focused time-delay neural networks. In *International Conference on Computer and Communication Engineering (ICCCE'10)*, pages 1–6. IEEE, 2010.

[7] M Kannan, S Prabhakaran, and P Ramachandran. Rainfall forecasting using data mining technique. 2010.

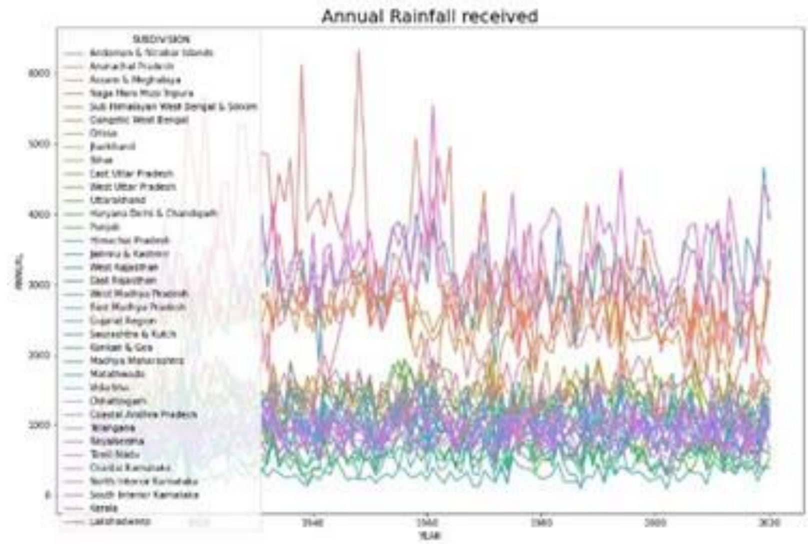

Fig. 3 Amount of rainfall

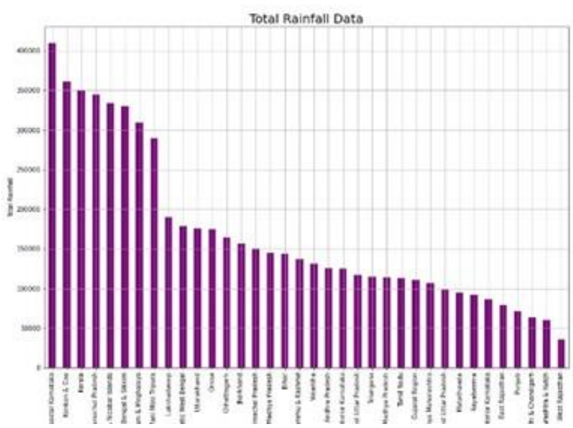

Fig. 4 Amount of rainfall

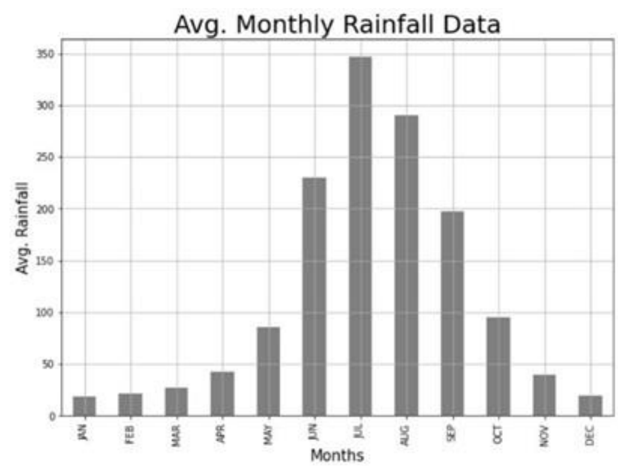

Fig. 6 Amount of average rainfall

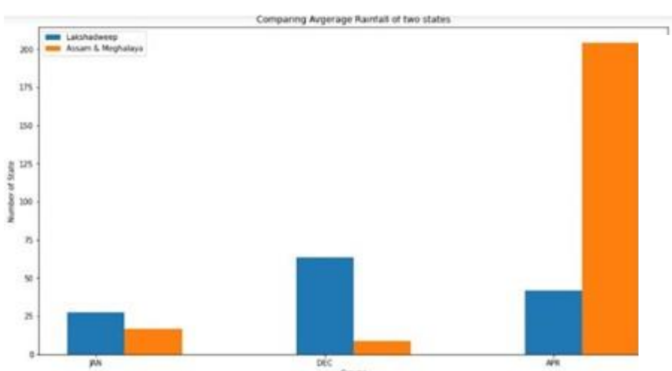

Fig. 5 Difference between the rainfall pattern

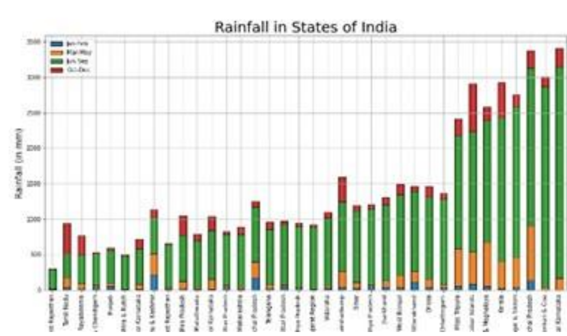

Fig. 7 Amount of average session rainfall

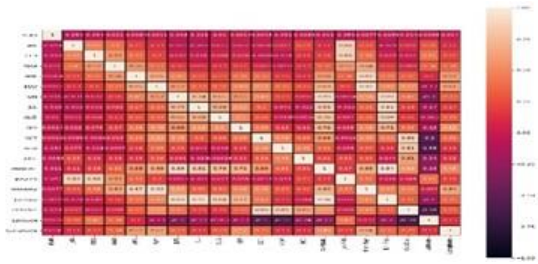

Fig. 8 Heat map

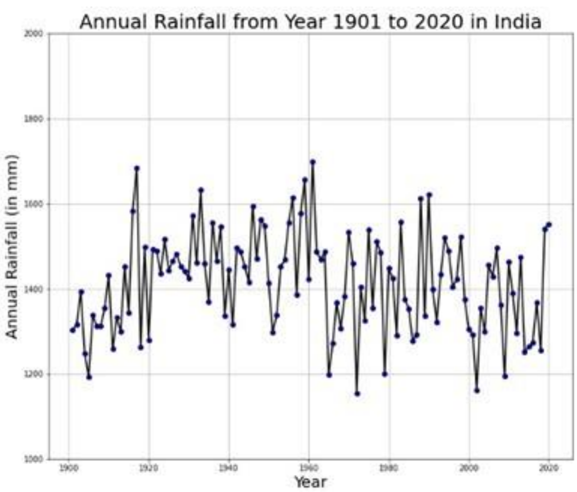

Fig. 9 Overall rainfall

[8] S Kannan and Subimal Ghosh. Prediction of daily rainfall state in a river basin using statistical downscaling from gcm output. *Stochastic Environmental Research and Risk Assessment*, 25(4):457–474, 2011.

[9] Moulana Mohammed, Roshitha Kolapalli, Niharika Golla, and Siva Sai Maturi. Prediction of rainfall using machine learning techniques. *International Journal of Scientific and Technology Research*, 9(01):3236–3240, 2020.

[10] VK Somvanshi, OP Pandey, PK Agrawal, NV Kalanker, M Ravi Prakash, and Ramesh Chand. Modeling and prediction of rainfall using artificial neural network and arima techniques. *J. Ind. Geophys. Union*, 10(2):

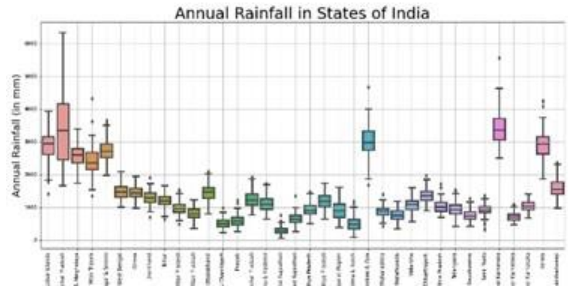

Fig. 10 Annual rainfall

141–151, 2006.

[11] Chandrasegar Thirumalai, K Sri Harsha, M Lakshmi Deepak, and K Chaitanya Krishna. Heuristic prediction of rainfall using machine learning techniques. In *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pages 1114–1117. IEEE, 2017.

[12] CL Wu and Kwok-Wing Chau. Prediction of rainfall time series using modular soft computingmethods. *Engineering applications of artificial intelligence*, 26(3):997–1007, 2013.