2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) | 978-1-7281-9111-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCEEE49695.2021.9429650

Rainfall Prediction using Multiple Linear Regressions Model

Hiyam Abobaker Yousif Ahmed [1] , Sondos W. A. Mohamed [2]
Institute of Space Research and Aerospace (ISRA)
Department of Astronomy and Astrophysics [1], Department of Computer Software [2]
Khartoum,Sudan
hoyamabobakeryousif@gmail.com[1] , ssun344@gmail.com [2]

***Abstract*—Meteorological scientists always try to find means to understand the atmosphere of the Earth, and to develop accurate weather prediction models. Several methods have been used in weather prediction. Recently, machine learning methods are assumed to be accurate techniques and have been widely used as an alternative to classical methods for weather prediction. The rainfall rate is one of the essential phenomena in the weather system, which has a direct influence on the agriculture and biological sectors. This paper aims to develop a multiple linear regression model in order to predict the rate of precipitation (PRCP), i.e., rainfall rate, for Khartoum state. It is based on some weather parameters, such as temperature, wind speed, and dew point. The data used in this research has been provided from the website of the National Climatic Data Center. A Python code using the Pytorch library has been written to develop the model, which applies Artificial Neural Networks. The efficiency of the model has been measured by comparing the average value of the mean square error of the training data with the test data. The obtained results show that the average of the mean square error has been improved by 85% during test time, when the same amount of data is used during the training and test phases. However, it drops to 59% when the amount of data at the test phase exceed the amount of training phase data .**

***Keywords*—Weather Prediction, rainfall, Linear Regression, Machine Learning. Artificial Neural Networks.**

I INTRODUCTION

The process of predicting the state of the atmosphere for a specific location in the future is called weather forecasting [1]. Interest in weather forecasting began from the earliest era, and the forecasting techniques were developed and have been changing with time,Several methods are used to generate weather forecasting, each of which differ in its accuracy and efficiency. There are three important steps that must precede the process of weather forecasting, which are to collect atmospheric data as much as possible, to understand the data and its inter-relation to determine the behavior of the atmosphere, and to use it in numerical models to predict the future state of the atmosphere.Recently, scientists tended to apply machine learning tools for weather prediction, because it does not require a deep and comprehensive understanding of the atmospheric process, thus it represents a good choice for weather forecasting [2].Machine learning (ML) is a process of learning a specific task without any human intervention, which will improve the performance only by the continuous learning process. Learning methods are of three types: supervised learning that is based on labeled data, unsupervised learning, and the reinforcement. The vital process in all machine learning methods is extracting of the features, and then to use these. extracted features for various approaches, like classification and regression [3].Applying machine learning techniques in weather forecasting can compensate complex meteorological physics model. With the availability of metrological data set, the two authors were encouraged to select supervised learning method, which is multiple linear regression, instead of unsupervised learning or reinforcement learning [1]. There are different regression types used in machine learning, such as linear regression, logistic, polynomial regression. The simpler and most frequent method is linear regression, which is used for prediction [4].The aim of this paper is to develop a multiple linear regression model to predict the rainfall rate in Khartoum state, which depends on many variables. The remainder of this paper is organized as follows. Section II provides a brief survey about related work, Section III explains materials and methods, and Section IV shows our Results. Finally, Section V concludes the article.

II LINEAR REGRESSION

Linear regression is one type of the supervised learning techniques to predict a numeric value (dependent variable) form a set of features (predictors). Likewise, it is about finding a function that maps inputs x € R to the corresponding function values f (x) € R [5], It forms a prediction by computing a weighted sum of the input features, plus a constant called the bias (intercept), as shown in the Fig. 1 below.

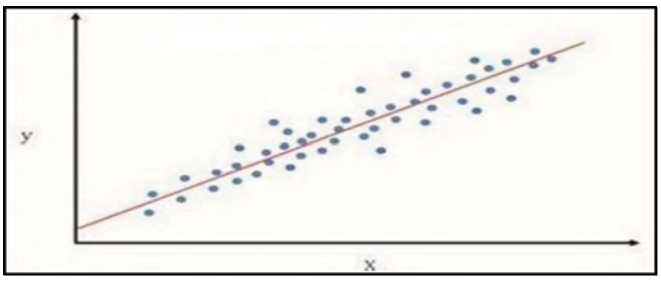

Fig. 1. Show the simple linear regression

When the dependent variable is calculated from one predictor, the regression is called simple regression, as shown in Equation (1) below. $Y = a + bX$ (1)

Where,

Y: dependent variable a: intercept

b: slope

978-1-7281-9111-9/20/$31.00 ©2020 IEEE

X: independent variable

If it is produced from two or more predictors, the regression is called multiple regressions, as shown in Equation (2) below.

$$Y = a + b1\,x1 + b2\,x2 + \ldots bn\,xn \qquad (2)$$

Where,

x1 , x2 , … xn : independent variables.

As other supervised learning methods to develop the linear model, authors have passed through two phases: the training phase also named the learning phase [6] and the testing phase. At the training time, they used a well- defined labeled data to adjust the bias and features weight to obtain multiple linear regression equations. At the testing phase, they used another labeled data so as to verify the validity of their model as a generalized model for prediction.

III RELATE WORK

This section highlights on many work applied machine learning with historical weather data to predict future weather state. these works used neural networks or linear regression to predict temperature ,rainfall sometimes they predict other weather parameters such as Humidity and Dew-point. The following paragraphs show details about each work

E. B. Abrahamsen and O. M. Brastein [1] developed a Python API to read meteorological data, and Artificial Neural Network models using Tensor Flow to study weather and predict temperature. Two weather variables were used in the study for precipitation and temperature. Mark Holmstrom, Dylan Liu, Christopher [2] used a linear regression model and a variation on a functional regression model for predicting the maximum and the minimum temperature for seven days, using data for two past days. Sanyam Gupta, Indumathy K, Govind Singhal [4] applied machine learning algorithms, linear regression model and normal equation optimization method, to predict weather based on few parameters. Folorunsho Olaiya [7] used an Artificial Neural Network and Decision Tree algorithms and meteorological data (2000- 2009) for the city of Ibadan, Nigeria, in forecasting weather variable (maximum temperature, rainfall, and wind speed). S. Prabakaran and others [8] used a modified linear regression model to predict rainfall with less error percentage by adding percentage to the input values. Paras and Sanjay Mathur [9] used the Multiple Linear Regression (MLR) model to predict four weather parameters which are (maximum and minimum temperature, relative humidity, and the category of rainfall).Wanie M. Ridwana,b, Michelle Sapitang et all [10] have applied two methods to predict rainfall forecasting rainfall, which are Autocorrelation Function (ACF) and projected error. Both methods implemented four different regression algorithms (Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression and Neural Network Regression, with different time horizons (daily, weekly, ten-days and monthly),The results showed that Boosted Decision Tree Regression is the best regression developed for M1, with the highest coefficient of determination, but in M2 the overall model performance gives a good result of each category except for 10- days with Boosted Decision Tree Regression and Decision Forest Regression

At this work Authors used multiple correlated weather parameter to predict rainfall rate,authors have selected multiple linear regression instead of linear regression to increase the accuracy of the prediction and the reliability of the model.

IV MATERIAL AND METHODS

A. Data Collection and Selection

The meteorological data used in this study was obtained online from the website of the national climatic data center. The data were provided based on data exchanged under the world metrological organization WMO [11] and hence the data is free for scientific research.

Authors have chosen the data of the Republic of Sudan, Khartoum meteorological station, then divided the data into two parts; the data from (1990- 2005) which was selected to train the model, and the period (2006-2020) which was selected for testing. Selected data set for ten attributes mean temperature (TMP).

Maximum temperature MX, minimum temperature MN, Dew point WP, sea level pressure SLP, station pressure STP, mean visibility VS, and wind speed WSP which represents the dependent variable and rainfall (precipitation) PRCP rate as dependent variable, as shown in Table I. Below

TABLE I. METEOROLOGICAL DATA USED AS INDEPENDENT VARIABLES OF THIS MODEL

Predictor Variable	Abbreviations
mean temperature	TMP
maximum temperature	MX
minimum temperature	MN
Dew point	WP
sea level pressure	SLP
station pressure	STP
mean visibility	VS
wind speed	WSP

B. Data Cleaning and Transformation

The process of data cleaning was done manually using an Excel program. It was done in four steps, which understanding the data set and the correlation between variables, deleting unwanted factors from the data set, dealing with missing data and outliers, and treating data to facilitate handling.

Table II (a,b)below shows a sample from data that will be used in training phase table II (a) shows the first five parameters used in this model as independent variables . and table II (b) shows the remainder parameters .

TABLE II(A). SAMPLES OF METEOROLOGICAL DATA USED AT TRAINING PHASE

	TMP (x1)	WP (x2)	SLP (x3)	STP (x4)	VS (x5)
0	69.8	39.8	1012.8	967.4	2.1
1	69.3	36.4	1014.7	970.0	7.6
2	70.1	33.4	1012.8	968.1	10.3
3	73.5	38.0	1012.6	968.3	9.8
...	...	..	...	...	...

3579	79.4	52.8	1010.3	966.9	10.4
3580	88.7	56.4	1009.7	966.3	11.1
3581	82.6	56.5	1009.8	966.3	6.8

TABLE II(B), SAMPLES OF METEOROLOGICAL DATA USED AT TRAINING PHASE

	WSP (x6)	MXSP (x7)	MX (x8)	MN (X9)
0	14.4	16.9	80	69.5
1	8.9	14.0	77	61.7
2	6.2	12.0	83	61.7
3	6.8	9.9	82	61.2
4	7.7	11.1	82	61.7
...	....	...	...	...
3577	10.4	13.0	87	60.8
3578	10.7	12.0	86	61.7
3579	10.0	12.0	92	64.9
3580	8.1	9.9	93	68.0
3581	10.5	29.9	94	69.4

Summary of Statistical Description of the data is shown in Table III a and b below. Table III(a) shows the Statistical Description of first five parameters used in this model as independent variables. and table II (b) shows the statistical description of remainder parameters.

TABLE III (A) . STATISTICAL DESCRIPTION OF DATA

	TMP (x1)	WP (x2)	SLP (x3)	STP (x4)	VS (x5)
count	3582	3582	3582	3582	3582
mean	86.55	200.6	2438.4	2404.8	12.990
std	8.577	1223.9	3289.8	3307.9	61.875
min	56.80	12.60	999.2	945.30	0.3000
25%	81.20	36.50	1005.5	963.00	7.6000
50%	88.00	47.60	1007.8	964.80	9.3000
75%	93.10	61.30	1012.1	968.00	11.100
max	106.2	999.0	999.9	999.90	999.90

TABLE III (B). STATISTICAL DESCRIPTION OF DATA

	WSP (x6)	MXSp (x7)	MX (x8)	MN (x9)
count	3582	3582	3582	3582
mean	13.70	23.23	101.93	82.60
std	77.60	101.2	165.64	287.3
min	0.000	1.000	59.000	33.80
25%	5.300	9.900	94.100	68.00
50%	7.400	12.00	100.40	76.10
75%	9.800	15.00	105.30	81.50
max	999.9	999.9	999.90	999.9

After the cleaning process, the linear regression hypothesis formalized as shown in the Equation (3) below.

$$PRCP=b_1 TMP + b_2 WP + b_3 SLP + b_4 STP + b_5 VS + b_6 WSP + b_7 MXSP + b_8 MX + b_7 MN \ldots (3)$$

This equation is used to train the model to predict the value of (PRCP),Then, it is used with the new computed data to obtain the difference between the predicted value and actual value of Y (PRCP), which is called error (or the loss) as follows

$$e = y - \bar{y} \qquad (4)$$

The least square error method is used to find the best line fitting the data as follows

$$e^2 = (y - \bar{y}) \qquad (5)$$

In addition, Python language has been used to write the code of linear regression; the pytorch package has been used to develop the ANN, and Adam optimization was used to update the parameters. The least square error method is used to find the best line fitting the data.

V RESULT

Table IV shows sample from actual and predicted values of the rainfall rate during the training phases.

TABLE IV. SHOWS THE ACTUAL AND PREDICTED PRCP VALUES DURING TRAINING PHASE, THE TABLE SHOWS THAT THE DIFFERENCE BETWEEN ACTUAL AND PREDICTED VALUES WAS LARGE ESPECIALLY AT THE BEGINNING OF THE TRAINING.

	Actual	predicted
0	0.0	-202.818268
1	0.0	-184.136261
2	0.0	-161.219086
3	0.0	-137.441727
4	0.0	-118.243645
..	...	...
95	0.0	-139.289917
96	0.0	-134.579208
97	0.0	-131.520737
98	0.0	-129.182846
99	0.0	-125.826492

Fig. 2 shows the learning curve of the model, in which the orange line and the blue line represents the actual and predicted values of the PRCP, respectively.

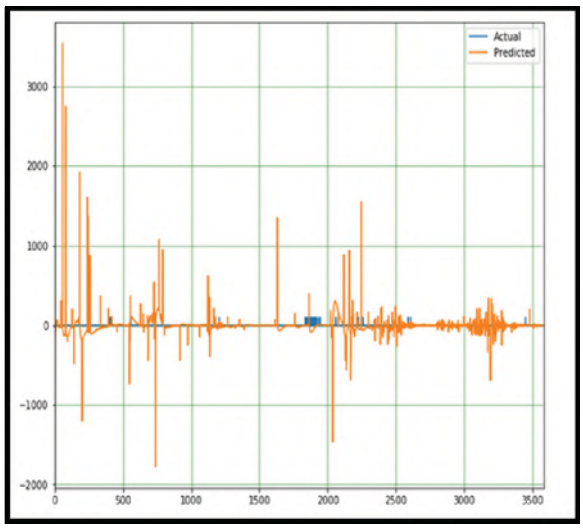

Fig. 2 Learning Curve of the Model the orange line and the blue line represents the actual and predicted values of the PRCP

Table (V) shows sample from actual and predicted values of the PRCP rate during the testing phase.

TABLE V. THE TABLE SHOWS THAT THE DIFFERENCE BETWEEN ACTUAL AND PREDICTED VALUES WAS DECREASED IN THE COMPARISON WITH TABLE IV

	Actual	predicted
0	0.0	0.808235
1	0.0	0.800155
2	0.0	0.246736
3	0.0	0.500647
4	0.0	0.725022
....	....	...
95	0.0	0.649510
96	0.0	0.339417
97	0.0	0.283009
98	0.0	0.366929
99	0.0	-0.092754

Fig. 3 shows the curve of the model during test phases ,in which the orange line and the blue line represents the actual and predicted values of the PRCP, respectively.

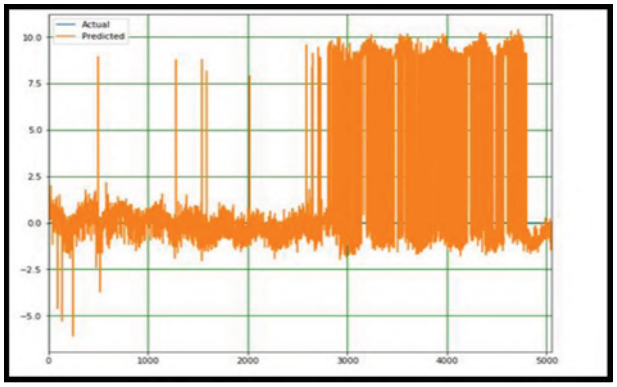

Fig. 3. Curve of the Model during test time the orange line and the blue line represents the actual and predicted values of the PRCP, respectively

TABLE (VI) SHOWS THE COMPARISON BETWEEN MEAN SQUARE ERROR VALUES DURING TRAINING AND TEST TIME, WHICH IT APPEARS THAT THE LOSS HAS A SIGNIFICANTLY DECREASED IN TEST TIME

	mean square error values in training phase	**mean square error values in test phase**
0	88743.984375	0.653243
1	76140.289062	2.433115
2	65288.695312	0.060879
3	53382.593750	0.250656
4	44693.691406	0.525656
..	....	...
1	19910.568359	0.421863
96	19533.806641	0.115204
97	18895.878906	4.695849
98	18209.146484	0.134637
99	17326.076172	-0.008603

The average mean square error values in the training phase equal 27918.9

The average mean square error values in the testing phase equal 324.8

Fig. 4 shows the change in the means square loss between training and testing time, which it appears that the loss has a significant decrease.

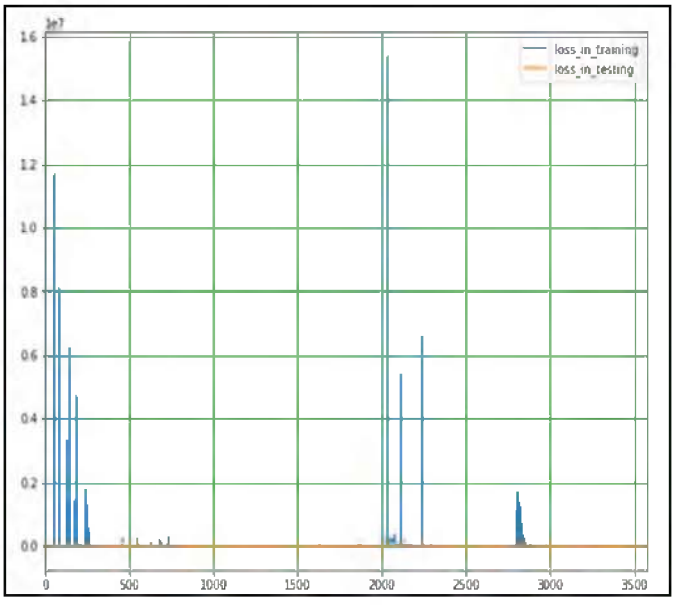

Fig. 4 .the Loss in Training and Test the in which the blue line represents the mean square error values during training phase whereas the orange line represent the mean square error in the phase.

VI. Conclusion

In this paper, authors have used multiple linear regression model to predict the rate of precipitation (i.e., rainfall rate) for Khartoum state, based on some weather parameters taken as the independent variables

Those weather parameters are the mean temperature, maximum temperature, minimum temperature, Dewpoint, sea level pressure, station pressure, mean visibility and wind speed. The average of the mean square error between the actual and predicted value during training and testing phase was calculated.

It was found that obtained results show that the mean square error between actual and predicted values of the rainfall precipitation rate (PRCP) has been significantly decreased during testing time. It has been found to be 85% when the amount of test data equals the amount of training data, and 59% when more test data is used.

Explanation of this reduction needs supplementary research. for example, it may indicate that the model used needs more data in the training phase.

References

1 E. Abrahamsen, O. M. Brastein, and B. Lie, "Machine Learning in Python for Weather Forecast based on Freely Available Weather Data," Proceedings of The 59th Conference on imulation and Modelling (SIMS 59), 26-28 September 2018, Oslo Metropolitan University, Norway, 2018.

2 M. Holmstrom, D. Liu, and C. Vo, "Machine Learning Applied to Weather Forecasting," Dec. 2016.

3 J. Refonaa, M. Lakshmi, R. Abbas, and M. Raziullha, "Rainfall Prediction using Regression Model," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2S3, Jul. 2019.

4 S. Gupta, I. K, and G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques," *International Journal of Computer Science and Information Technologies,* vol. 7, no. 3, pp. 1490–1493, 2016.

5 S. Gupta, I. K, and G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques," International Journal of Computer Science and Information Technologies, vol. 7, no. 3, pp. 1490–1493, 2016.

6 C. Bishop, *Pattern recognition and machine learning*. Springer Verlag, 2006.

7 F. Olaiya and A. B. Adeyemo, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies," *International Journal of Information Engineering and Electronic Business*, vol. 4, no. 1, pp. 51–59, 2012.

8 S. Prabakara, P. N. Kumar, and P. S. M. Tarun, "RAINFALL PREDICTION USING MODIFIED LINEAR REGRESSION," ARPN Journal of Engineering and Applied Sciences, vol. 12, no. 12, Jun. 2017.

9 S. M. Paras, "A Simple Weather Forecasting Model Using Mathematical Regression," *Indian Research Journal of Extension Education*, vol. 12, pp. 161–168, 2016.

10 W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed, and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," Ain Shams Engineering Journal, 2020

11 *Climate Data Online - Select Area*. [Online]. Available: https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd. [Accessed: 21-Jan-2021].