ISSN:0976-0172

Dr.B. Sai Venkata Krishna, JBio sci Tech, Vol 13(2),2025, 01-07

Journal of Bioscience And Technology
www.jbstonline.com

Predicting Precipitation with the Use of Multiple Linear Regressions

[1] Dr.B. Sai Venkata Krishna, [2] P. Shirisha,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

Abstract

Accurate weather prediction models are the work of meteorologists who are always on the lookout for new methods to learn about Earth's atmosphere. The art of weather prediction has made use of a number of techniques. Machine learning approaches have now largely replaced more traditional methods of weather prediction due to widespread belief in their accuracy. The rate of rainfall is a crucial meteorological phenomenon that impacts the agricultural and biological industries directly. In order to forecast the rate of precipitation (PRCP), also known as the rainfall rate, for the state of Khartoum, this research intends to construct a multiple linear regression model. The dew point, wind speed, and temperature are some of the meteorological factors that form its basis. The National Climatic Data Center website provides the data used in this study. The model was developed using Python code that makes use of the Pytorch framework and uses Artificial Neural Networks. We compared the training and test sets of data using the average value of the mean square error to see how well the model performed. When the data used for training and testing is same, the findings demonstrate an 85% improvement in the average of the mean square error during testing. It falls to 59%, however, when the data from the test phase is more than the data from the training phase.

Keywords

—Linear regression, machine learning, rainfall, and weather prediction. Deep Learning Systems.

I. Introduction

The term "weather forecasting" refers to the practice of making future predictions about the condition of the atmosphere at a particular area [1]. Interest in weather forecasting dates back to prehistoric times, and forecasting methods have evolved and progressed over that period. There are a number of approaches to producing weather forecasts, and their relative merits are questionable. In order to make accurate weather predictions, three critical steps must be completed first: gathering extensive atmospheric data, analyzing that data to determine the atmosphere's behavior, and finally, using numerical models to predict the atmosphere's future state. Because it does not need an in-depth and thorough comprehension of the atmospheric process, machine learning has recently become a popular method for weather prediction among scientists [2]. What we call "machine learning" (ML) is really a series of steps that computers take to figure out how to do a certain job better over time, with little to no help from humans. Supervised learning, which makes use of labels in data, unsupervised learning, and reinforcement learning are the three main categories of learning techniques. Features are essential in machine learning approaches, and the first step is to extract them. used characteristics for a variety of

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

methods, including regression and classification [3]. A complicated model of meteorological physics may be compensated by using machine learning approaches to weather forecasting. Rather of using unsupervised or reinforcement learning, the two writers were advised to use supervised learning, specifically multiple linear regression, with the availability of the metrological data set [1]. Machine learning makes use of a variety of regression types, including logistic, linear, and polynomial regression. Linear regression is the most common and easiest way to make predictions [4]. The several factors that affect the rate of rainfall in Khartoum state are the focus of this paper's multiple linear regression model development. What follows is an outline of the rest of the paper. The second section gives a synopsis of relevant literature, the third describes the study's methodology and materials, and the fourth displays the findings. Section V serves as the article's last section.

II. Linear Regression

One kind of supervised learning is linear regression, which uses a collection of characteristics (predictors) to make predictions about a numerical value (dependent variable). Similar to how it's shown in Fig. 1 below, it's all about discovering a function that maps inputs x € R to the associated function values f (x) € R [5]. A prediction is then formed by calculating a weighted sum of the input characteristics plus a constant termed the bias (intercept).

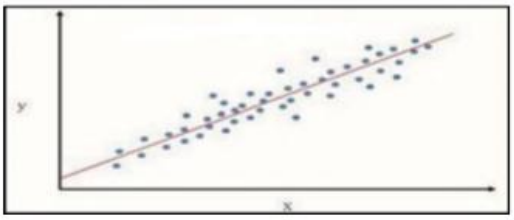

Fig. 1. Show the simple linear regression

Regression is referred to as simple regression when the dependent variable is computed from a single predictor, as seen in Equation (1) below. The formula for Y is Y = a + bX (1). In cases when,

Y: dependent variable a: intercept

b: slope

X: independent variable

As seen in Equation (2) below, a regression is referred to as multiple regressions if it is generated from two or more predictors.

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

Where,

x1 , x2 , … xn : independent variables.

There are two stages to developing a linear model using supervised learning techniques; the first is training, often called learning [6]; and the second is testing. They trained using clearly labeled data, which they utilized to tweak the bias and feature weight in order to derive numerous linear regression equations. In order to ensure that their model could be used for generalized prediction, they used additional labelled data throughout the testing phase.

III RELATE WORK

This section focuses on several studies that used machine learning in conjunction with weather history data to forecast the weather in the future. These studies have used neural networks or linear regression to forecast various meteorological characteristics, including temperature, rainfall, humidity, and dew point. Each piece of art is described in depth in the paragraphs that follow. The Python API for reading meteorological data was created by E. B. Abrahamsen and O. M. Brastein [1]. To research weather and make temperature predictions, they used Tensor Flow to train Artificial Neural Network models. The research used temperature and precipitation as its two meteorological variables. By combining data from the two days before and present, Mark Holmstrom, Dylan Liu, and Christopher [2] were able to forecast the high and low temperatures for the next seven days using a linear regression model and a variant on a

functional regression model. Machine learning techniques, a linear regression model, and the normal equation optimization approach were used to provide weather predictions using a small number of parameters by Sanyam Gupta, Indumathy K., and Govind Singhal [4]. To predict weather variables (highest temperature, rainfall, and wind speed) for the Nigerian city of Ibadan, Folorunsho Olaiya [7] utilized meteorological data from 2000 to 2009 in conjunction with Artificial Neural Network and Decision Tree methods. To reduce the proportion of errors in rainfall predictions, S. Prabakaran and colleagues [8] adjusted a linear regression model by adding percentages to the input data. Four meteorological parameters—high and low temperatures, relative humidity, and the kind of precipitation—were predicted by Paras and Sanjay Mathur [9] using the Multiple Linear Regression (MLR) model. Two approaches have been used to forecast rainfall by Wanie M. Ridwana,b, Michelle Sapitang et al. [10]: the Autocorrelation Function (ACF) and projected error. With varying time periods (daily, weekly, ten-days, and monthly), both techniques used four distinct regression algorithms: Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, and Neural Network Regression. While Boosted Decision Tree Regression produced the best results across the board in M1 (with the highest coefficient of determination), Decision Forest Regression and Boosted Decision Tree Regression both performed admirably in M2 (with the exception of 10-day performance). In this study, the scientists opted using multiple linear regression rather than linear regression to forecast the rate of rainfall based on a number of interrelated meteorological parameters.boost the model's dependability and forecast accuracy.

IV. Material And Methods

Gathering Information and Making Choices The weather information used in this research was sourced from the National Climatic Data Center's website. The data is freely available for scientific inquiry since it is based on data transferred under the World Meteorological Organization (WMO) [11]. The data from the Republic of Sudan's Khartoum meteorological station was picked by the authors, who then split it into two sets: one set for training the model, covering the years 1990–2005, and another set for testing, covering the years 2006–2020. Choosen dataset for 10 characteristics, average temperature (TMP). Rainfall (precipitation), mean visibility (VS), dew point (WP), sea level pressure (SLP), station pressure (STP), maximum temperature (MX), and minimum temperature (MN) are all dependent variables. The PRCP

Table I. Meteorological Data Used as in d e pe n d e n t v a r ia b l e s 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

The Cleaning and Transformation of Data (B) We used an Excel software to manually clean the data. There were four stages to this process:(1) gaining familiarity with the dataset and its relationships;(2)

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

eliminating irrelevant variables and factors;(3) addressing missing data and outliers; and(4) processing the data to make it easier to work with. This training phase will make use of the data sampled in Table II (a, b) below. The first five parameters used as independent variables in this model are shown in table II (a). the parameters that make up the remaining are shown in table II (b).

Table II(A). Sa mpl e s of Meteorological Data used at t r a in in g ph a s e

	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), Sa mpl es of Meteorological Data used at t r a ining pha se

	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

Tables Ill a and b following provide a statistical summary of the data. The statistical description of the first five parameters used as independent variables in this model may be seen in Table 111(a). and the statistical description of the rest parameters is shown in table II (b).

Table III (a) . St a t ist ica l 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). St a t ist ica l 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

Equation (3) below shows the formalization of the linear regression hypothesis after the cleaning step.

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

The model is trained using this equation to predict the value of (PRCP). Then, using the newly computed

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

data, the error (or loss) is calculated as the difference between the predicted and actual values of Y (PRCP).

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

To choose the best line fitting the data, the least square error approach is used in the following way:

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

Furthermore, the linear regression algorithm was written in Python, the ANN was developed using the pytorch package, and the parameters were updated using Adam optimization. To choose the optimal line fitting the data, the least square error approach is used.

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.

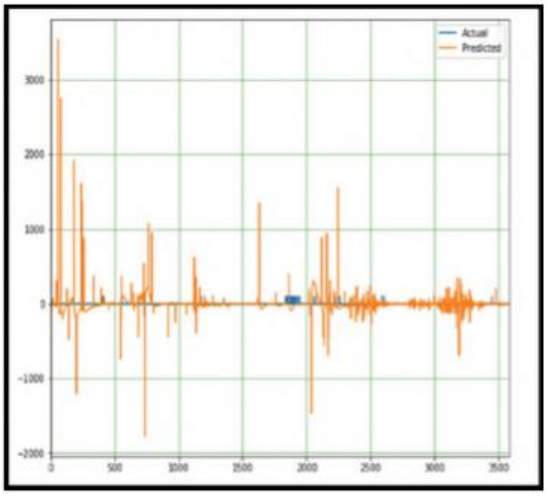

Fig. 2 Learning Curve of the Model the orange line and the blue line represents the actual and predicted values of the PRCP

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

Table V. the table shows that the d if f e r e n c e between actual and pr e d ic t e d values was decreased in the c o m pa r is o n w it h 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

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

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.

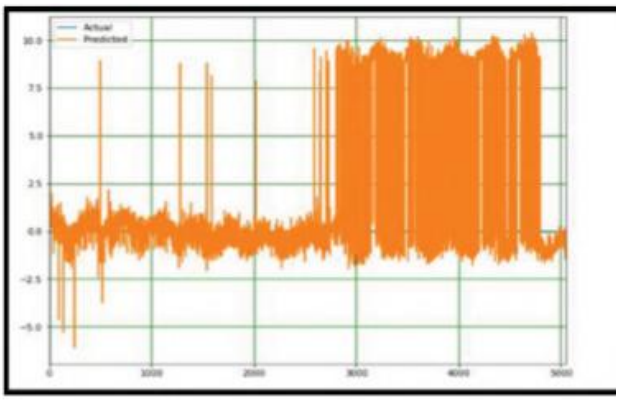

Fig. 3. Curve of the Model during test time the orange line and the blue line represents the actual and predicted values of the PRCP, respectively

Table (VI) shows the c o mpa r is o n 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.

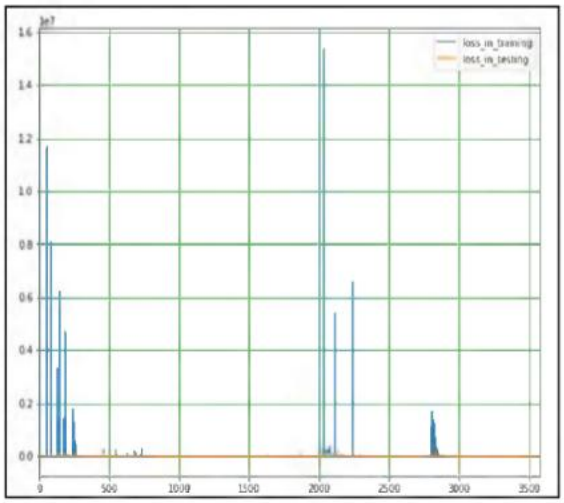

Fig. 4 .the Loss in Training and Test the in which the blue line represents the mean square error values during training phase whereas the orange line represent the mean square error in the phase.

VI. Conclusion

The authors of this work have used a multiple linear regression model to forecast the annual rainfall rate in Khartoum state, using a number of meteorological indicators as its independent variables. The following meteorological variables are taken into account: average, maximum, and lowest temperatures; dewpoint; pressure at sea level; pressure at the weather station; average visibility; and wind speed. When comparing the actual and projected values throughout the testing and training phases, the average of the mean square errors was determined. During the testing period, it was discovered that there was a considerable reduction in the mean square error between the anticipated and actual values of the precipitation rate (PRCP). Results show that it's 85% when training and test data quantities are comparable, and 59% when test data quantities are increased. Additional research is necessary to explain this

decrease. For instance, it may suggest that more data is needed for the training phase of the model.

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-Jan2021].