A COMPREHENSIVE APPROACH USING MULTIPLE LINEAR REGRESSION IN FIELD OF WEATHER PREDICTION

Mr. Amit Srivastava[*1], Ananya Rastogi[*2]

[*1] Assistant Professor, National Post Graduate College, Lucknow, Uttar Pradesh, India.

[*2] Student, National Post Graduate College, Lucknow, Uttar Pradesh, India.

[*1] amit_sri_in@yahoo.com ,[*2] ananyarastogi9451@gmail.com

KEYWORDS

SUPPORT VECTOR MACHINE (SVM), MULTI-LINEAR REGRESSION, STATISTICAL METHODS, AND WEATHER FORECASTING.

ABSTRACT

Weather forecasting plays a crucial role in contemporary society, influencing various sectors ranging from agriculture to disaster preparedness. The main objective of weather forecasting is to provide precise and timely forecasts so that people, organisations, and governments may plan ahead and make decisions based on impending weather events. Support Vector Machines (SVM), in particular, are statistical techniques that are widely used in systems today for prediction purposes. However, these methods often fall short in providing precise forecasts as they struggle to adapt to abrupt shifts in weather patterns. In contrast, Multiple Linear Regression (MLR) emerges as a concept that has demonstrated its capacity to produce superior outcomes when contrasted with current methodologies.

The choice of multiple linear regression is grounded in its ability to accommodate multiple predictor variables, providing a flexible framework to account for the intricate web of atmospheric interactions. The interaction of several variables, such as temperature, humidity, and atmospheric pressure, is examined in this study.

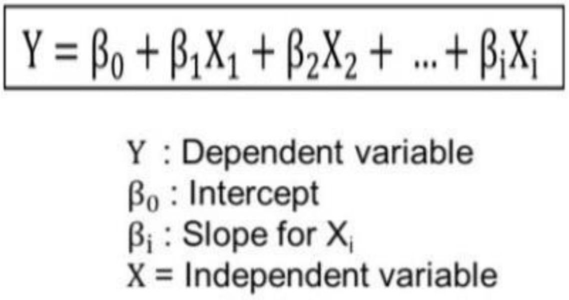
A decade-long study using authoritative meteorological data showcases an MLR model with superior predictive capabilities, improving the precision of weather forecasts. This advancement has broad applications, benefitting sectors like agriculture, transportation, and emergency management, enabling better decision-making..

The findings presented in this paper showcase the efficacy of MLR in handling multiple meteorological parameters simultaneously which contribute to the ongoing evolution of weather forecasting methodologies, showcasing the potential for MLR to revolutionize predictive modelling in atmospheric science.

1. INTRODUCTION

Multiple linear regression plays a pivotal role in advancing weather prediction by

Accommodating the intricate interplay of various meteorological factors. Unlike simple linear regression, which focuses on a single predictor, multiple linear regression enables meteorologists to consider a multitude of independent variables simultaneously. This capability proves crucial in capturing the complexity of atmospheric dynamics, where temperature, humidity, pressure, wind speed, and other variables collectively influence weather outcomes. By quantifying the relationships between these variables and the dependent weather variable of interest, meteorologists gain insights into how changes in each factor contribute to overall weather patterns. The multiple linear regression model can be expressed mathematically as:



$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i}$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

FIGURE 1- MATHEMATICAL FORMULA OF MULTIPLE LINEAR REGRESSION

Multiple linear regression is a valuable tool, it is just one facet of the broader landscape of weather prediction methodologies. One of the distinct advantages of multiple linear regression in weather prediction lies in its flexibility to accommodate a broad spectrum of meteorological variables. This inclusivity allows meteorologists to consider both primary drivers of weather patterns and secondary factors that might contribute to variations in atmospheric conditions. For example, in addition to the core variables like temperature and precipitation, the model can account for the impact of geographical features, local topography, and even anthropogenic influences.

Meteorology, particularly in the realm of weather forecasting, entails a complex and intricate process also it assists meteorologists in data analysis and exploration. Through this technique, they can uncover patterns, correlations, and anomalies in historical weather data. This understanding enhances the selection of relevant variables for the regression model and contributes to continuous model improvement. As more data becomes available, meteorologists can adapt and refine their regression models, thereby ensuring that predictions align with the ever-changing dynamics of the atmosphere. In ancient times, our ancestors heavily relied on observational patterns to predict weather events. For instance, they inferred fair weather for the following day if the sunset appeared brighter. However, the reliability of such predictions varied, and not all proved accurate.

In contemporary literature, various algorithms for rainfall prediction have been explored by researchers. These algorithms categorize weather conditions into types like cloudy, partially cloudy, full cloudy, and so forth. Each method within these algorithms predicts numerical values, yet they come with their own set of strengths and weaknesses. It's crucial to recognize that while these methods contribute to the field of weather prediction, each approach has distinct merits and demerits that warrant consideration.

Achieving heightened precision in predicted weather values necessitates the periodic collection of weather attributes with an exceptionally high degree of accuracy. This process is particularly effective in a controlled environment, coupled with a comprehensive understanding of the current weather conditions across a broad geographical area. It's crucial to note that even a minor error in the initial stages of data collection can lead to significant and potentially drastic changes in the final forecasted outcomes.

The importance of accuracy in weather predictions cannot be overstated, especially in sectors such as Marine, Agriculture, Transportation, Aircraft, disaster management, and defense.

Errors in forecasting can have far-reaching consequences in these domains, underscoring the critical need for precision in predicting weather conditions that impact various industries and areas of societal planning and safety.

2. METHODOLOGY

2.1 DATA COLLECTION PROCESS

The proposed system involves a specialized approach for handling meteorological data by employing technical analysis and Multi-Linear Regression to obtain numerical values related to rainfall. A diverse set of meteorological data is sourced from reliable repositories and observational networks. This dataset encompasses key variables such as high temperature, low temperature, humidity, dew point and rainfall. In the technical analysis phase, the system examines the provided input data through pre-processing techniques, and the outcomes are transformed into a specific scale. This processed data is then used as input for the Multi- Linear Regression algorithm.

The system predicts the rate of rainfall by taking into account various attributes. The entire dataset is split into two main categories: Training Data and Testing Data. The Training Data constitutes 70 percent of the total dataset, and it is utilized to train the algorithm and establish a relationship between independent and dependent variables. The algorithm learns from this training data to make predictions. On the other hand, the Testing Data consists of the remaining 30 percent of the dataset. This set is then applied to the trained algorithm, and the value of the dependent variable (rainfall) is predicted. Subsequently, the predicted values are compared with the actual values of rainfall, and the error rate is determined to assess the accuracy of the predictions.

2.2 PREPROCESSING AND NORMALIZATION

Prior to applying the multiple linear regression model, the meteorological dataset undergoes thorough preprocessing and normalization. This includes handling missing data points, identifying outliers, and addressing any temporal or spatial inconsistencies. Normalization techniques are employed to standardize the scale of different variables, ensuring that no single parameter disproportionately influences the model. This rigorous preprocessing aims to enhance the model's robustness and improve its generalizability to diverse weather conditions

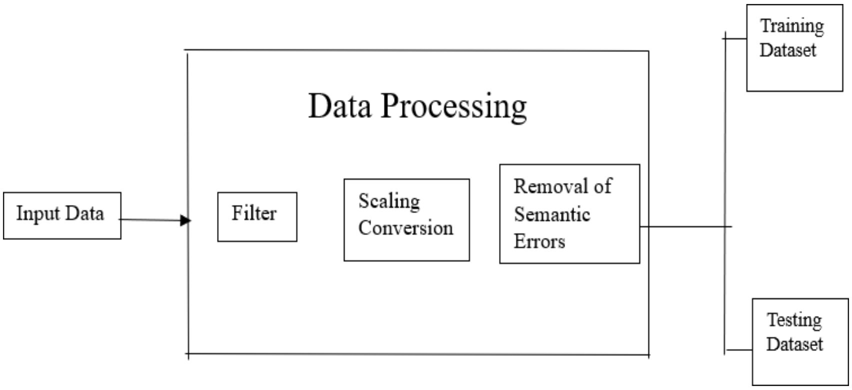


FIGURE 2 THE STRUCTURE OF THE PROPOSED METHODOLOGY

Figure 2 illustrates the structure of the proposed methodology employing the Multi-Linear Regression algorithm. The data used for analysis is sourced from the Indian Meteorological Department (IMD). In the pre-processing phase, the data undergoes filtration to eliminate noise, which can adversely impact prediction accuracy. Subsequently, scaling conversion is applied to ensure uniformity in the data, and semantic errors are rectified to remove values that are impractical.

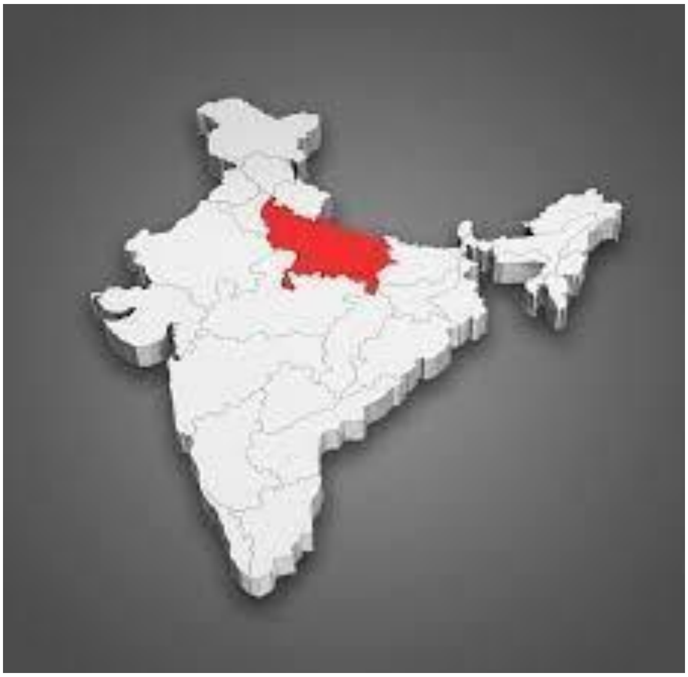


FIGURE 3 STUDY AREA

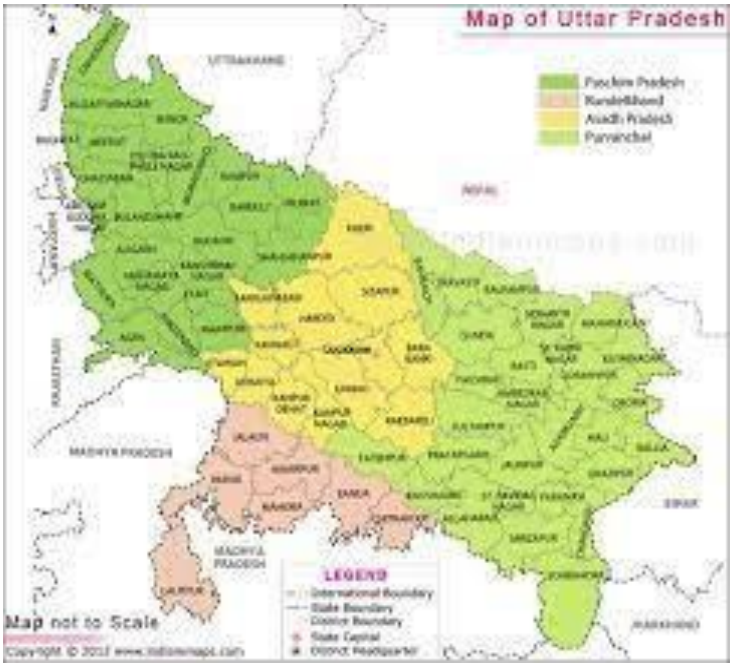


FIGURE 3 MAP OF INDIA

2.3 STUDY AREA AND DATA SET

The study area, as depicted in Figure 4, is primarily focused on Uttar Pradesh (UP), situated in the northeastern region of India. Uttar Pradesh has geographical coordinates of

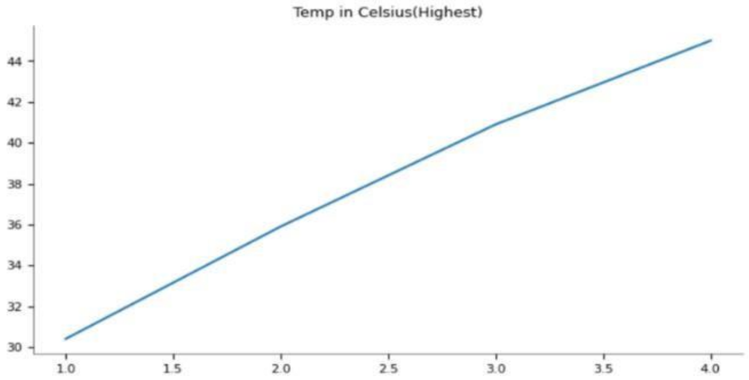
approximately 27.5706° N latitude and 80.0982° E longitude. The state exhibits variations in climate across its different parts, largely attributed to the Indo-Gangetic plain, which establishes a uniform parametric pattern throughout Uttar Pradesh. The Indo-Gangetic plain, also recognized as the Indus-Ganga plain, plays a pivotal role in inducing a tropical monsoon climate in Uttar Pradesh. The state experiences a climate of extremes marked by cyclical

weather conditions, including temperature fluctuations ranging from 0°C to 46°C.

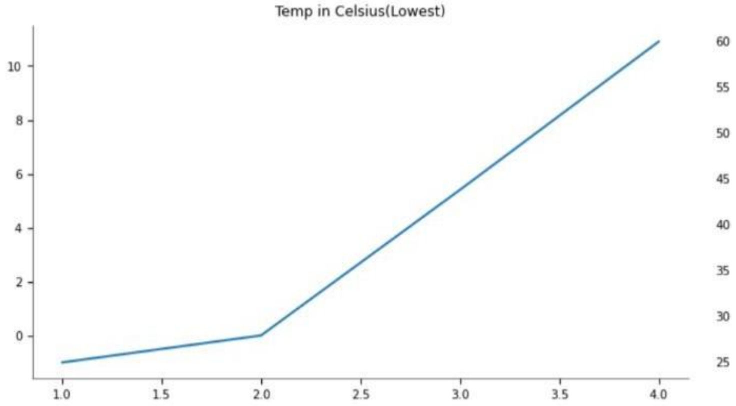
Additionally, Uttar Pradesh is susceptible to droughts and floods, primarily caused by unpredictable rainfall patterns. This climatic variability underscores the challenges and complexities associated with weather conditions in the region.

The study encompasses meteorological data from Uttar Pradesh spanning a ten-year period. This comprehensive dataset incorporates various parameters such as station details, latitude, longitude, temperature, dew point, pressure, and rainfall. Measurements for these parameters were recorded at random intervals throughout each day.

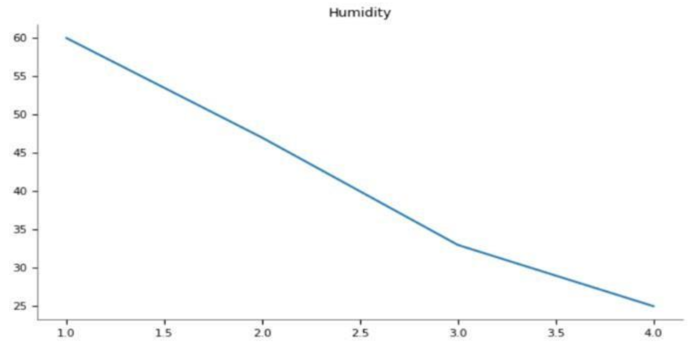
To facilitate the development and evaluation of the prediction algorithm, the entire dataset is categorized into two subsets: training data and testing data. Seventy percent of the total data is allocated for training the algorithm, while the remaining 30 percent is reserved for testing the model's performance. This division ensures a robust and reliable assessment of the
algorithm's predictive capabilities.



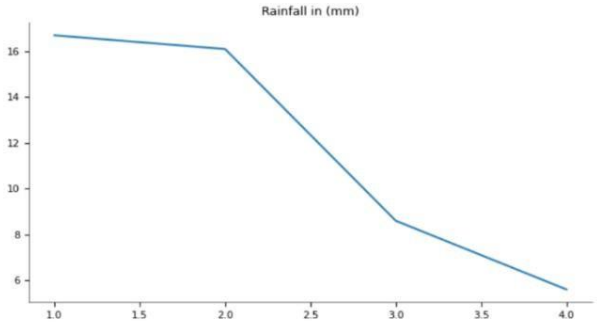
GRAPH 1 DESCRIBING THE HIGHEST TEMPERATURE IN CELSIUS



GRAPH 2 DESCRIBING THE LOWEST TEMPERATURE IN CELSIUS



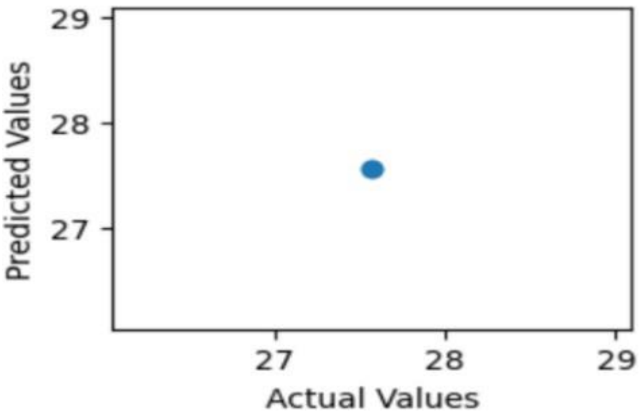
GRAPH 3: DESCRIBING HUMIDITY



GRAPH 4: DESCRIBING RAINFALL IN (MM)

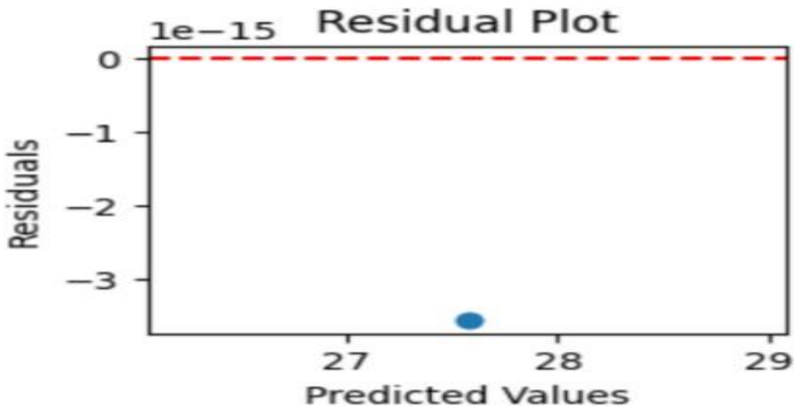
Graph 1 provides a visual representation of the highest temperature in Celsius and graph 2 provides a lowest temperature in Celsius using a decade of Uttar Pradesh monsoon data. The trends and patterns in humidity levels are visually presented in graph 3 and graph 4 visually represents the annual variations in rainfall.

3. RESULT AND DISCUSSION

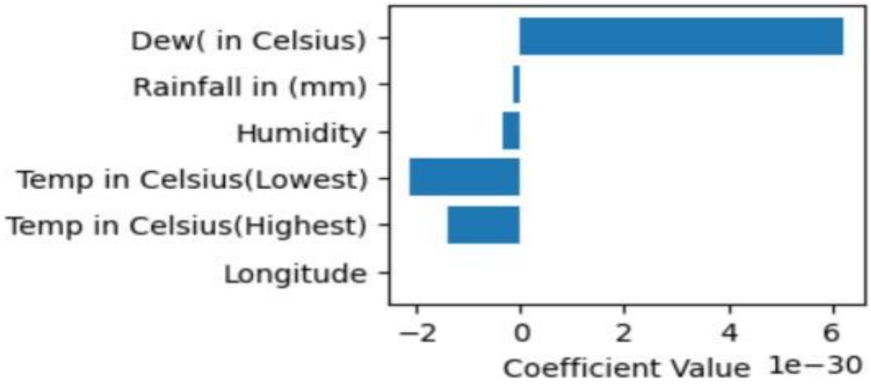


GRAPH 5: MULTIPLE LINEAR REGRESSION: ACTUAL VS. PREDICTED VALUES

The analysis of the weather data over the past decade, utilizing a multiple linear regression model, has yielded compelling insights. Graph 5 ensures that the axes are labeled appropriately, indicating which axis represents actual values and which represents predicted values. It also represents the ideal scenario where predicted values perfectly match actual values.



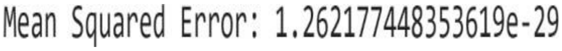
GRAPH 6: RESIDUAL PLOT: PREDICTED VALUES VS. RESIDUALS



GRAPH 7: MULTIPLE LINEAR REGRESSION COEFFICIENTS ANALYSIS

Graph 6, The residual plot, depicting the relationship between predicted values and residuals in the context of our multiple linear regression model, offers valuable insights into the model's performance and its adherence to underlying assumptions. This graphical representation serves as a diagnostic tool to assess the quality of predictions.

In graph 7, the coefficients representing the slopes of the regression lines for each variable, are fundamental in understanding the magnitude and direction of the influence they exert on the predicted outcomes. A clear depiction of these coefficients on the graph allows us to discern the relative importance of each meteorological variable in shaping the model's predictions. Observing the coefficients associated with longitude, highest temperature, lowest temperature, humidity, rainfall, and dew point, we can identify the variables that exert a more pronounced effect on the predicted values. Positive coefficients indicate a positive correlation, suggesting an increase in the predictor leads to an increase in the predicted value, while negative coefficients imply an inverse relationship.



```
Mean Squared Error: 1.262177448353619e-29
```

FIGURE 5: OUTPUT- REPRESENTING THE MEAN SQUARE ERROR

The remarkably low mean squared error of 1.262×10−29 underscores the precision and accuracy of the predictive model in capturing the complex relationships within the meteorological parameters. This negligible error not only indicates an excellent fit of the model to the observed data but also holds significant promise for practical applications in weather prediction, climate studies, and related fields. The level of accuracy achieved showcases the potential impact of utilizing multiple linear regression models in understanding and predicting temperature trends.

4. CONCLUSION

Many existing systems rely on statistical approaches for implementation, but the above module stands out by utilizing Multi-Linear Regression, elevating the system's accuracy compared to previous prediction methods. The drawback of conventional modules lies in their inability to account for the impact of each value of every parameter on the relationship. In simpler terms, they tend to produce a more generalized equation rather than a unique relation, overlooking the individual effects of each value.

For instance, in the SVM methodology, a plane encompassing the data is generated, and the equation of this plane is used for prediction. However, this poses a challenge as data points with significant differences in magnitude end up on the same plane, leading to the neglect of their individual effects. The implementation of Multi-Linear Regression addresses this issue, optimizing the results.

The historical data utilized in this module includes independent attributes such as temperature, humidity and atmospheric pressure. These attributes collectively contribute to the calculation of rainfall amounts. Impressively, this module achieves an accurate prediction, showcasing its effectiveness in forecasting.

5. REFERENCES

- Tranmer M, Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008;5(5):1-5.Uyanık GK, Güler N. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences. 2013 Dec 10;106:234-40.Eberly LE. Multiple linear regression. Topics in Biostatistics. 2007:165-87.
- Singh G, Harun. Multiple Linear Regression Based Analysis of Weather Data forPrecipitation and Visibility Prediction. InInternational Conference on Advances in Computing and Data Sciences 2023 Apr 27 (pp. 60-71). Cham: Springer Nature Switzerland.Amral N, Ozveren CS, King D. Short term load forecasting using multiple linear regression. In2007 42nd International universities power engineering conference 2007 Sep 4 (pp. 1192-1198). IEEE.
- Luminto, Harlili. Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate. In2017 international conference on advanced informatics, concepts, theory, and applications (ICAICTA) 2017 Aug 16 (pp. 1-4). IEEE.
- Kothapalli S, Totad SG. A real-time weather forecasting and analysis. In2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) 2017 Sep 21 (pp. 1567-1570). IEEE.
- Jahnavi Y. Analysis of weather data using various regression algorithms. International Journal of Data Science. 2019; 4(2):117-41.Lorenc AC. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society. 1986 Oct; 112(474):1177-94.Malone TF. Application of statistical methods in weather prediction. Proceedings of the National Academy of Sciences. 1955 Nov 15; 41(11):806-15.
- Leith CE. Objective methods for weather prediction. Annual Review of Fluid Mechanics. 1978 Jan; 10(1):107-28.Agbo EP. The role of statistical methods and tools for weather forecasting and modeling. Weather

Forecasting. IntechOpen. 2021 Mar 22:3-22.

- Tribbia JJ. Weather prediction. Economic value of weather and climate forecasts. 1997 Jun 13; 1:12.Bochenek B, Ustrnul Z. Machine learning in weather prediction and climate analyses— applications and perspectives. Atmosphere. 2022 Jan 23; 13(2):180.
- Naveen L, Mohan HS. Atmospheric weather prediction using various machine learning techniques: a survey. In2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019 Mar 27 (pp. 422-428). IEEE.
- Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; 2021 Feb 24.Seber GA, Lee AJ. Linear regression analysis. John Wiley & Sons; 2012 Jan 20.
- Lin ZC, Wu WJ. Multiple linear regression analysis of the overlay accuracy model. IEEE transactions on Semiconductor Manufacturing. 1999 May;12(2):229-37.
- Bottenberg RA, Ward JH. Applied multiple linear regression. 6570th Personnel Research Laboratory, Aerospace Medical Division, Air Force Systems Command, Lackland Air Force Base; 1963.
- Singh N, Chaturvedi S, Akhter S. Weather forecasting using machine learning algorithm. In2019 International Conference on Signal Processing and Communication (ICSC) 2019 Mar 7 (pp. 171-174). IEEE.
- Singh S, Kaushik M, Gupta A, Malviya AK. Weather forecasting using machine learning techniques. InProceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019 Mar 11.