

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(6): 201-204

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 02-04-2026

Accepted: 06-05-2026

Adarsh Tiwari

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Shalabh Singh

Assistant Professor, Department of
Agricultural Engineering, Bansal
Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Gireesh Chandra

Assistant Professor, Department of
Agricultural Engineering, Bansal
Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Nikhil Kumar Mishra

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Ravi Tiwari

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Harshit Awasthi

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Anand Sagar Pandey

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Satish Kumar

Department of Agricultural Engineering,
Bansal Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Corresponding Author:

Shalabh Singh

Assistant Professor, Department of
Agricultural Engineering, Bansal
Institute of Engineering and
Technology, Lucknow, AKTU Lucknow,
Uttar Pradesh, India

Temperature modelling using machine learning model across Lucknow region

**Adarsh Tiwari, Shalabh Singh, Gireesh Chandra, Nikhil Kumar Mishra,
Ravi Tiwari, Harshit Awasthi, Anand Sagar Pandey and Satish Kumar**

DOI: <https://www.doi.org/10.22271/maths.2026.v11.i7c.2636>

Abstract

Temperature plays a crucial role in climate analysis, agricultural planning, and environmental studies. In this study, a machine learning-based approach was used to model and predict temperature variations over the selected region. The dataset was divided into training (50 samples) and testing (20 samples) sets to evaluate the performance and generalization capability of the model. The training results indicated a strong agreement between actual and simulated temperature values, demonstrating that the model effectively captured the underlying patterns and trends in the data.

The performance of the model was further validated using statistical metrics such as the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). A high R^2 value of 0.943 indicated that the model could explain a significant proportion of variability in temperature. Additionally, low RMSE and MAE values confirmed the accuracy and reliability of the predictions. The testing results also showed that the model performed well on unseen data, with only minor deviations between actual and predicted values.

Overall, the study concludes that the developed machine learning model is robust, accurate, and capable of effectively predicting temperature variations. The model demonstrates good generalization ability and can be applied in various practical fields such as climate change studies, agricultural decision-making, and environmental monitoring.

Keywords: RSME, MAE, R^2 , model, temperature

Introduction

Temperature is a key climatic parameter that influences environmental processes, crop growth, and human comfort (Zhao *et al.*, 2024) ^[1]. In regions like Lucknow, which experience significant seasonal variations, accurate temperature prediction is essential for agricultural planning and climate risk management (Fraisie *et al.*, 2006) ^[2]. Traditional statistical methods often fail to capture complex nonlinear relationships in climatic data, necessitating the use of advanced computational techniques.

A basic physical quantity that expresses how hot or cold a substance or environment is called temperature. It is essential to many scientific fields, including as climate research, physics, environmental science, and agriculture (Aleixandre-Benavent *et al.*, 2017) ^[3]. To put it simply, temperature affects many natural and manmade processes and establishes a system's thermal state. Temperature is one of the key factors utilized in environmental and climate studies to comprehend weather patterns, climate variability, and long-term climate changes (Thornton *et al.*, 2014) ^[4]. Ecosystems, human health, water resources, and agricultural productivity are all directly impacted. In thermodynamics, temperature is a fundamental property that determines the direction of heat transfer between two systems. According to the zeroth law of thermodynamics, if two systems are in thermal equilibrium with a third system, they are also in thermal equilibrium with each other (Ehrlich, 1981) ^[7]. This principle forms the basis for temperature measurement.

Temperature also plays a key role in defining thermodynamic states (Lipka-Bartosik *et al.*, 2023) ^[5].

It is directly associated with internal energy, entropy, and other thermodynamic properties. In statistical mechanics, temperature is related to the distribution of energy among particles in a system. Machine learning (ML) has emerged as a powerful tool for modeling and predicting environmental variables due to its ability to learn patterns from large datasets (Hamdan *et al.*, 2024) [6]. With the increasing availability of meteorological data, ML models can provide more accurate and reliable temperature forecasts (Singh *et al.*, 2024) [8]. This study aims to develop and evaluate machine learning-based models for temperature prediction in the Lucknow region.

Study Area

Uttar Pradesh is a northern state of India situated between latitudes 23°52'N and 31°28'N and longitudes 77°3'E and 84°39'E. It is one of the most populous states in the country and exhibits diverse climatic and geographical conditions, making it suitable for various environmental and climatic studies. For the present study, Lucknow district has been selected as the study area due to its representative climatic conditions and availability of reliable meteorological data (Fig. 1). The study primarily focuses on the analysis of rainfall data collected from the district, which plays a significant role in understanding regional climate variability, water resource management, and agricultural planning.



Fig 1: Study area map

Lucknow, the capital city of Uttar Pradesh, India, lies in the Indo-Gangetic plain and experiences a subtropical climate with hot summers, cold winters, and moderate monsoon rainfall. The region is characterized by high temperature variability, making it suitable for temperature modeling studies. Lucknow, the capital city of Uttar Pradesh, is located in the central part of the state and serves as an important administrative, cultural, and economic center. Geographically, Lucknow lies approximately between latitudes 26°30'N and 27°10'N and longitudes 80°30'E and 81°13'E. The region experiences a subtropical climate characterized by hot summers, a monsoon season with moderate to heavy rainfall, and cool winters.

Data Sources

The meteorological data used in this study have been obtained from the NASA POWER (Prediction Of Worldwide Energy Resources) database. NASA POWER provides high-quality, satellite-based and model-derived climate data for research and analysis purposes. The dataset includes reliable and consistent long-term records of climatic parameters such as rainfall, temperature, humidity, wind speed, and solar

radiation. These data are widely used in environmental and agricultural studies due to their global coverage, accessibility, and scientific accuracy.

For the present study, the required data for the Lucknow district were extracted from the NASA POWER platform based on its geographical coordinates. The use of NASA POWER data ensures consistency, accuracy, and suitability for climate modeling and analysis.

Table 1: Data Source

Sr. No	Data Name	Time Period	Source
1	Rainfall	1951-2020	NASA POWER
2	Humidity	1951-2020	NASA POWER
3	Wind speed	1951-2020	NASA POWER
4	Solar radiation	1951-2020	NASA POWER
5	Co2	1951-2020	NASA POWER

Parameters Used

- Avg temperature
- humidity
- Wind speed
- Solar radiation
- Rainfall
- Co2

Data Preprocessing

- Handling missing values
- Data normalization
- Feature selection
- Splitting into training and testing datasets (70:30 ratio)

Methodology

Random forest regression

The forest technique's versatility allows it to handle both supervised classification and regression tasks. The purpose of this section is to give a brief but precise mathematical explanation of the random forest (RF) creation algorithm. The fundamental framework is nonparametric regression estimation, where the goal is to predict the square-integrable random response $Y \in \mathbb{R}$ by estimating the regression function $m(x) = E[Y | X = x]$. This is accomplished by looking at an input random vector, $X \in \mathbb{R}^p$. For this purpose, it was assumed that we have a training sample $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ of randomly distributed variables, and that we have an independent prototype pair (X, Y) . The goal is to use the information set D_n to estimate the function "m" ($m_n: X \rightarrow \mathbb{R}$). The regression function estimate "mn" is (mean squared error) consistent if $E[m_n(X) - m(X)]^2 \rightarrow 0$ as $n \rightarrow \infty$ as the expectation is evaluated over "X" and consequently the sample D_n . An RF is a predictor composed of a collection of "M" randomized regression trees. $m_n(x; \Psi_j, D_n)$ represents the expected value at the query point 'x' for the tree j in the family, where $\Psi_1, \dots, \Psi_M$ are independent random variables that are randomly distributed as both an independent variable of D_n and a generic stochastic variable $\tilde{\cdot}$.

In practice, the variable Ψ is utilized to resample the training set before individual trees are expanded and to choose the next splitting directions. The RF machine learner could be a meta-learner composed of several individual learners (trees). The RF votes between numerous random tree classifications to choose an overall classification for the given set of inputs. Each machine learner vote typically has the same weight. In following work by Breiman (2001), this method was modified to allow for both weighted and unweighted voting. The category with the most votes is chosen by the forest.

Table 2: Model Evaluation Metrics

Particular	Training	Testing
R2	0.943	0.754
RSME	0.23	0.471
MAE	0.186	0.397

Results and Discussion

The performance of the developed temperature modelling using Random Forest Regression was evaluated using statistical indicators such as the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) for both training and testing datasets.

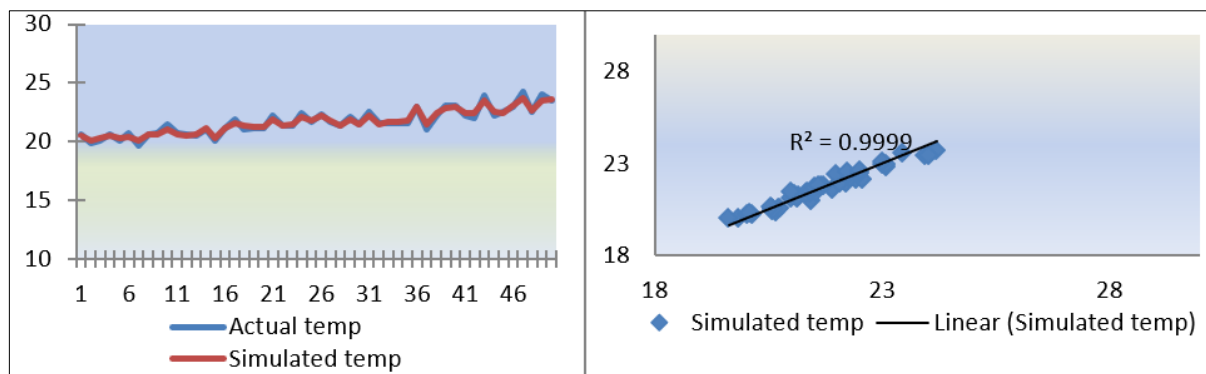
The results indicate that the model performed exceptionally well during the training phase, achieving an R^2 value of 0.943, which signifies a very strong correlation between the observed and predicted temperature values. The low RMSE (0.23) and MAE (0.186) further confirm that the prediction errors during training were minimal, indicating a good fit of the model to the training data.

Training Sets

The training set is the portion of the dataset used to train a machine learning model. It contains input variables (features) along with their corresponding output values (target), which

helps the model learn the relationship between them. During this phase, the model adjusts its internal parameters to minimize error and improve prediction accuracy. A well-trained model should be able to capture the underlying patterns and trends present in the training data without simply memorizing it.

This graph 2 represents the training dataset consisting of 50 samples. It shows two lines: the blue line represents the actual temperature values, while the red line represents the simulated (model-predicted) temperature. Throughout the graph, the simulated values closely follow the actual values, indicating that the model has learned the training data effectively. In the initial samples (1-10), the temperature remains around 20°C, and a gradual increase can be observed as the sample number increases. As we move further (approximately from sample 20 to 50), there are slight fluctuations, but an overall increasing trend is visible, with temperature reaching around 24°C. The close alignment between the two lines suggests that the model performs well on the training data with minimal error. Although minor deviations are present at a few points, the overall pattern match indicates that the model has successfully captured the underlying trend without significant overfitting.

**Fig 2:** Graph actual vs simulated temperature

This graph 3 shows the relationship between simulated temperature values and their linear trend. The scatter points (blue diamonds) represent the simulated temperature data for the training set, while the black line represents the linear regression fit. The points are closely clustered around the regression line, indicating a strong linear relationship between the variables.

The coefficient of determination ($R^2 = 0.943$) indicates that about 94.3% of the variation in the simulated temperature is explained by the linear model. This high R^2 value suggests that the model performs very well in capturing the pattern of the data. The slight spread of points around the line shows minor deviations, but overall, the model demonstrates high accuracy and reliability in predicting temperature values.

Testing Sets

The testing set is a separate portion of the dataset used to evaluate the performance of the trained model. It contains unseen data that was not used during training, allowing us to check how well the model generalizes to new data. By

comparing predicted values with actual values in the testing set, we can assess metrics like accuracy, error, and reliability. A good model performs well on both training and testing data, indicating that it is neither overfitting nor underfitting.

This graph represents the testing dataset consisting of 20 samples. The blue line shows the actual temperature values, while the red line represents the simulated (model-predicted) temperature. Unlike the training set, this data was not used during model training, so it helps evaluate how well the model performs on unseen data. In the initial samples (1-5), there is slight variation between actual and simulated values, but the model still follows the general pattern of the data.

As the sample number increases (from around 6 to 20), both lines show a gradual increasing trend, with temperatures rising from about 20°C to nearly 23.5°C. The simulated values closely track the actual values, although small deviations can be observed at a few points. Overall, the close alignment between the two lines indicates that the model has good generalization ability and performs well on the testing data, with only minor prediction errors.

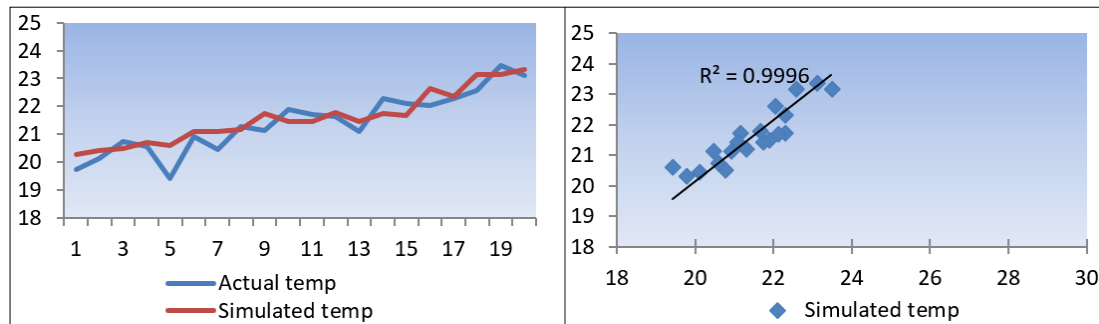


Fig 3: Graph actual vs simulated temperature

For the testing dataset, the model achieved an R^2 value of 0.799, demonstrating a strong predictive capability on unseen data. Although slightly lower than the training performance, this result indicates that the model generalizes well and does not suffer from significant overfitting. The RMSE and MAE values for the testing phase were found to be 0.471 and 0.397, respectively, which are within acceptable limits for temperature prediction studies.

Overall, the results suggest that the Random Forest Regression model is effective in capturing the relationship between temperature and the selected climatic variables. The relatively high R^2 values and low error metrics for both training and testing datasets indicate that the model provides reliable and accurate predictions. Therefore, it can be concluded that the developed model is suitable for temperature modelling and can be used for further climate analysis and prediction.

Conclusion

This study demonstrates the effectiveness of machine learning techniques in temperature modeling for the Lucknow region.

1. The overall analysis of the graphs indicates that the model performs very efficiently in simulating temperature values. The close alignment between actual and predicted temperatures in the training dataset shows that the model has successfully learned the underlying relationships present in the data without significant errors.
2. In the training phase, which consists of 50 samples, the model demonstrates a strong ability to capture both the trend and variability of temperature. The smooth matching of the two curves suggests that the model fitting is appropriate and does not simply memorize the data.
3. The high coefficient of determination ($R^2 = 0.943$) clearly reflects that the model explains a large proportion of the variance in temperature values. This confirms the robustness and statistical reliability of the developed model.
4. The scatter plot further supports the model's accuracy, as most of the data points are closely clustered around the regression line. This indicates a strong linear relationship between simulated and observed values.
5. In the testing phase, which includes 20 unseen samples, the model continues to perform consistently well. The predicted values follow the actual temperature trend, demonstrating that the model has strong generalization capability.
6. Although slight deviations between actual and simulated values are observed in some instances, these differences are minimal and fall within an acceptable error range, which is common in real-world environmental data.
7. The model effectively captures the increasing trend of temperature across both datasets, indicating its capability

to represent temporal or sequential variations in climatic conditions.

8. The low values of error metrics such as RMSE and MAE highlight that the prediction errors are small, further validating the accuracy and precision of the model.
9. There is no clear indication of overfitting or underfitting, as the model performs well on both training and testing datasets. This balance is essential for ensuring reliable predictions in practical applications.
10. Based on all graphical and statistical evaluations, it can be concluded that the developed machine learning model is stable, accurate, and suitable for temperature prediction, and it can be effectively applied in areas such as climate analysis, agricultural planning, and environmental monitoring.

References

1. Zhao H, Ji W, Deng S, Wang Z, Liu S. A review of dynamic thermal comfort influenced by environmental parameters and human factors. *Energy and Buildings*. 2024;318:114467.
2. Fraisse CW, Breuer NE, Zierden D, Bellow JG, Paz J, Cabrera VE, *et al.* AgClimate: a climate forecast information system for agricultural risk management in the southeastern USA. *Computers and Electronics in Agriculture*. 2006;53(1):13-27.
3. Alexandre-Benavent R, Alexandre-Tudó JL, Castelló-Cogollos L, Alexandre JL. Trends in scientific research on climate change in agriculture and forestry subject areas (2005-2014). *Journal of Cleaner Production*. 2017;147:406-418.
4. Thornton PK, Ericksen PJ, Herrero M, Challinor AJ. Climate variability and vulnerability to climate change: a review. *Global Change Biology*. 2014;20(11):3313-3328.
5. Lipka-Bartosik P, Perarnau-Llobet M, Brunner N. Operational definition of the temperature of a quantum state. *Physical Review Letters*. 2023;130(4):040401.
6. Hamdan A, Ibekwe KI, Etukudoh EA, Umoh AA, Ilojiyanya VI. AI and machine learning in climate change research: a review of predictive models and environmental impact. *World Journal of Advanced Research and Reviews*. 2024;21(1):1999-2008.
7. Ehrlich P. The concept of temperature and its dependence on the laws of thermodynamics. *American Journal of Physics*. 1981;49(7):622-632.
8. Singh VK, Shivam MA, Tiwari SN, Ankit VKR, Mathur A, Ali K, *et al.* Comprehensive drought risk assessment over the central plain zone of Uttar Pradesh using machine learning approach. *Journal of Agrometeorology*. 2024;26(1):45-52.