Plant Archives Vol. 25, Special Issue (ICTPAIRS-JAU, Junagadh) Jan. 2025 pp. 109-116

e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org

DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2025.v25.SP.ICTPAIRS-019

COMPARING OPEN-SOURCE RAINFALL DATA PRODUCTS FOR PREDICTING PADDY YIELD IN MIDDLE GUJARAT: A MACHINE LEARNING APPROACH

S.H. Bhojani[1*], P.A. Pandya[2], M.K. Tiwari[3] and Ram Bhavin[4]

[1]Directorate of IT, Anand Agricultural University, Anand, Gujarat, India.
[2]Research, Testing and Training Centre, Junagadh Agricultural University, Junagadh, Gujarat, India.
[3]Dept. of Soil and Water Engineering, Anand Agricultural University, Godhra, Gujarat, India.
[4]Office of the Vice Chancellor, Anand Agricultural University, Anand, Gujarat, India.
*Corresponding author Email: shitalbhojani@aau.in

ABSTRACT

Rainfall plays a critical role in determining crop yield, particularly in regions dependent on monsoons, such as Middle Gujarat, India. Accurate and reliable rainfall data are essential for predicting crop productivity, but the limited availability of observed station data often poses a challenge. This study aims to address this gap by evaluating the effectiveness of open-source rainfall data products, such as ERA5, NASA Power, and IMD Gridded data, in predicting the yield of paddy crops in three districts of middle Gujarat: Ahmedabad, Dahod, and Panchmahal. The statistical analysis of monthly rainfall across Ahmedabad, Dahod, and Panchmahal using ERA5, IMD Gridded, and NASA Power reveals significant trends in rainfall averages, standard deviations, and coefficients of variation during the monsoon season (June to September). NASA Power consistently reports the highest rainfall averages, especially in June, while ERA5 often captures higher values in July and August but shows greater variability. Overall, both NASA Power and ERA5 offer robust datasets for predicting paddy yields, with IMD Gridded providing lower rainfall estimates across the regions. We utilized monthly rainfall data from these sources and applied two predictive models: Linear Regression (LR) and Artificial Neural Network - Multilayer Perceptron (ANN-MLP). The results reveal significant variations in the performance of the models depending on the data source. For Ahmedabad, the Coefficient of Determination (R^2) values were highest for ERA5 (0.58) using the MLP model, compared to IMD Gridded (0.30) and NASA Power (0.52). In Dahod, NASA Power showed superior performance with an R^2 of 0.67 using MLP, while ERA5 and IMD Gridded had lower predictive accuracy (0.20 and 0.43, respectively). For Panchmahal, NASA Power again performed best with an R^2 of 0.67 using MLP, whereas IMD Gridded achieved an R^2 of 0.63, and ERA5 scored 0.45. The Mean Absolute Error (MAE) percentages further confirm the varying levels of accuracy across the models and data sources. For instance, in Ahmedabad, the MLP model using ERA5 data had the lowest MAE (13.49%), while in Panchmahal, the NASA Power dataset with the MLP model had the lowest MAE (13.6%). These findings underscore the importance of selecting appropriate rainfall data products for crop yield prediction, as the choice of data source and modeling approach significantly affects predictive accuracy. This study contributes to improved decision-making for sustainable agriculture by demonstrating the potential of open-source data products in regions with sparse observational networks.

Key words: Artificial Neural Network, Crop yield, ERA5, IMD Gridded data, NASA Power, Paddy

Introduction

Rainfall is a critical factor in rain fed agriculture, particularly in monsoon-dependent regions like India. Crops such as paddy, which are water-intensive, are highly sensitive to the amount and timing of rainfall, affecting yields and food security (Pai *et al.*, 2014). Accurate rainfall prediction is vital for optimizing agricultural production. However, the sparse distribution of meteorological stations, especially in regions like Middle Gujarat, makes localized rainfall data collection challenging.

Open-source rainfall data from satellite observations and reanalysis models offer promising alternatives, providing continuous rainfall estimates at high resolutions (Hersbach *et al.*, 2020). Still, their accuracy varies by region and application, highlighting the need to evaluate these datasets for specific agricultural purposes. D.A. Hughes emphasizes the importance of validating satellite rainfall data against ground-based observations. This study aims to address this challenge by comparing the effectiveness of three widely used open-source rainfall data products-ERA5, NASA Power, and IMD Gridded data-for predicting paddy yield in three districts of Middle Gujarat: Ahmedabad, Dahod, and Panchmahal. These districts represent diverse agro-climatic conditions, with varying rainfall patterns and levels of agricultural productivity. By focusing on paddy, a water-intensive crop, this research highlights the importance of accurate rainfall data for yield forecasting in water-stressed environments. The findings of this study have the potential to improve decision-making for sustainable agriculture, particularly in regions with sparse observational networks. Djavan De Clercq *et al.*, 2024 investigates the potential of machine learning techniques for predicting rice yields at the district level in India by utilizing climate reanalysis (ERA5) and remote sensing data. Saicharan, V. *et al.*, (2023) addresses the critical issue of rainfall data accuracy in India by comparing multiple datasets against IMD gridded data. Its insights are valuable for researchers, policymakers, and practitioners involved in hydrological and climate studies, contributing to improved decision-making in the context of water resource management.

In regions like Middle Gujarat, where monsoon rainfall is the primary source of water for agriculture, accurate rainfall data is crucial for effective crop yield prediction and agricultural planning. Timely and reliable rainfall forecasts allow farmers to make informed decisions regarding sowing, irrigation, fertilization, and harvesting. However, the lack of dense meteorological station networks in many parts of India, including Middle Gujarat, presents a major obstacle to obtaining the high-quality rainfall data needed for accurate yield forecasting. To address this challenge, researchers and policymakers are increasingly turning to open-source rainfall data products, which offer an alternative to traditional in-situ observations.

Some of the most commonly used open-source rainfall data products includeERA5, Produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 is a global reanalysis dataset that combines observations from various sources with model outputs to generate comprehensive estimates of atmospheric variables, including rainfall (Hersbach *et al.*, 2020). ERA5 offers high spatial and temporal resolution, making it one of the most widely used datasets for climate and weather research. The NASA Prediction of Worldwide Energy Resources (POWER) project provides meteorological and solar energy datasets derived from satellite observations and reanalysis models. NASA Power data is available at a daily temporal resolution and is commonly used in agricultural and environmental applications due to its global coverage and ease of access (Stackhouse *et al.*, 2019). The India Meteorological Department (IMD) provides gridded rainfall data for India, generated through the interpolation of observations from its network of meteorological stations. IMD Gridded data is widely used for climate studies and agricultural planning within India, as it offers region-specific rainfall estimates that account for the unique monsoon patterns of the subcontinent (Pai *et al.*, 2014). P. A. Pandya *et al.*, (2020) addresses the critical issue of rainfall's impact on cotton productivity. It offers important insights that can inform agricultural practices and policy decisions in the context of climate change. Parthsarthi Pandya *et al.*, (2023) addresses the pressing issue of agricultural drought through innovative methodologies. It offers important insights that can inform both policy and practice, ultimately contributing to more resilient agricultural systems in the face of climate variability. Yesilkoy, S. *et al.*, (2024) presents a valuable exploration of crop yield prediction using advanced data integration techniques having reanalysis and crop phenology data across different agro climatic zones. Anil Kumar Singha *et al.*, provides a comparative analysis of various satellite-derived rainfall products against IMD gridded data during the ISM.

In recent years, machine learning has emerged as a powerful tool for predicting crop yields based on climatic and environmental variables. Machine learning algorithms can analyse large datasets, identify complex patterns, and generate accurate predictions, making them particularly well suited for agricultural applications. Among the various machine learning techniques, two approaches have shown particular promise for crop yield prediction: Linear Regression (LR) and Artificial Neural Networks (ANNs). Rachidi, S. *et al.*, (2023) evaluates various satellite-based rainfall products, validating their accuracy through hydrological modelling using Artificial Neural Networks (ANN) in a semi-arid zone.

In this study, both LR and ANN-MLP models are applied to predict paddy yield in the three districts of Middle

Gujarat, using rainfall data from ERA5, NASA Power, and IMD Gridded datasets. By comparing the performance of these models across different rainfall data sources, this research aims to identify the most effective combinations of data and modelling techniques for crop yield prediction in this region. Bhojani, S.H. *et al.*, (2020) explores the application of novel activation functions in neural networks to predict wheat crop yields. Bhojani, S.H. *et al.*, (2021) also investigates the impact of different activation functions on the performance of machine learning models used for predicting wheat crop yields. It aims to determine which activation functions enhance the predictive accuracy of models.

Materials and Methods

Study Area and Climate

The study was conducted in three districts of Middle Gujarat: Ahmedabad, Dahod, and Panchmahal, which represent distinct agro-climatic conditions within the region.

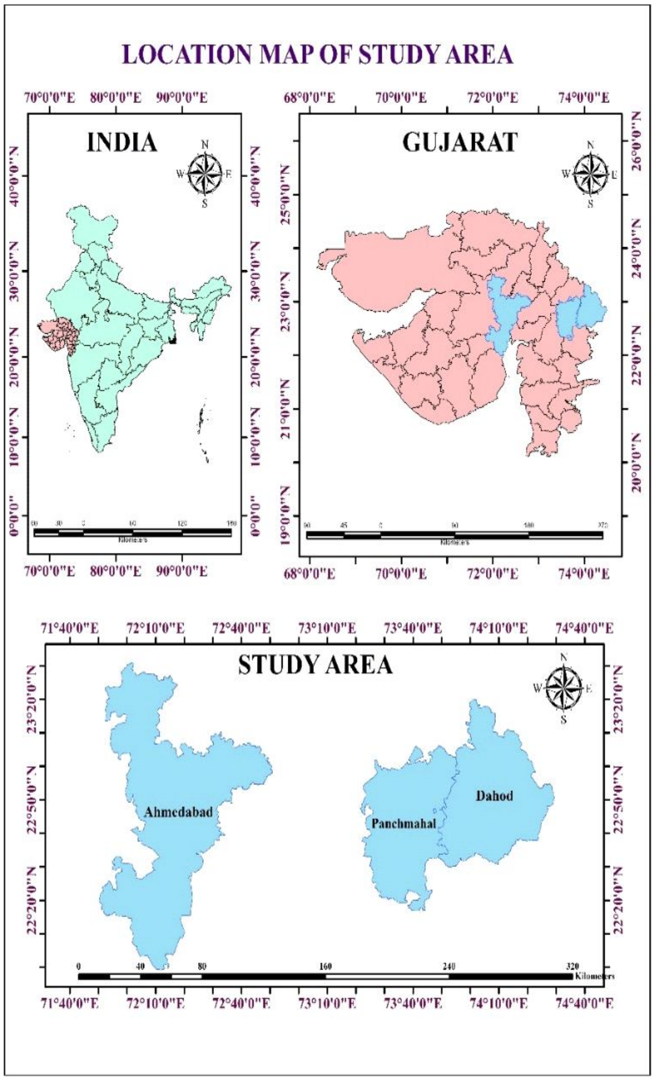

Fig. 1: Location of Study Area.

These districts were selected due to their diverse rainfall patterns and agricultural activities, particularly the cultivation of paddy, a water-intensive crop heavily reliant on monsoon rainfall. Ahmedabad is located in the western part of Middle Gujarat and experiences a semi-arid to sub-humid climate. The region receives moderate rainfall, primarily during the southwest monsoon season, from June to September. The average annual rainfall in Ahmedabad ranges between 600 to 800 mm, but internal variability is common. Dahod is located in the eastern part of Gujarat, characterized by a more humid climate with higher annual rainfall compared to Ahmedabad. The region receives between 900 to 1,200 mm of rainfall during the monsoon season. Rainfed agriculture is predominant in Dahod, with paddy being one of the major crops. Panchmahal located adjacent to Dahod, also experiences a humid climate with monsoon rainfall patterns similar to Dahod. The annual rainfall ranges between 900 to 1,100 mm, and paddy cultivation is widespread in the district. Variability in rainfall, however, poses significant challenges to agricultural productivity.

Data Used

The paddy yield data for this study were obtained from the Directorate of Agriculture, Government of Gujarat (GoG). The dataset includes historical annual paddy yield values for two districts over the period 1998 to 2022 for Ahmedabad and Panchmahal, and from 2003 to 2022 for Dahod. The data was collected at the district level and represents the average paddy yield in kilograms per hectare (kg/ha) for each year. These data were used as the dependent variable in the crop yield prediction models. The temporal range of the yield data corresponds to the availability of open-source rainfall data products, ensuring a comprehensive assessment of the relationship between rainfall and yield.

To assess the influence of rainfall on paddy yield, monthly rainfall data for the monsoon season (June to September) were collected from three open-source rainfall data products:-

ERA5: The European Centre for Medium-Range Weather Forecasts (ECMWF) provides the ERA5 reanalysis dataset, which includes rainfall estimates based on a combination of observations and numerical weather models.

IMD Gridded Data: The India Meteorological Department (IMD) provides gridded rainfall data for India, created by interpolating observations from meteorological stations across the country. NASA Power: The NASA Prediction of Worldwide Energy Resources (POWER) project provides satellite-derived meteorological data, including rainfall estimates.

Statistical Analysis of Rainfall Data

The monthly rainfall data for each district (Ahmedabad, Dahod, and Panchmahal) from the three sources (ERA5, IMD Gridded, and NASA Power) were subjected to statistical analysis to summarize their central tendency and variability through Arithmetic Mean, Standard Deviation and CV%. Additionally, the Z-score of annual rainfall was calculated to standardize the rainfall data and facilitate comparison across the three datasets. The Z-score is a statistical measure that indicates how many standard deviations an observation is from the mean. For each district and dataset, the Z-score was computed for annual rainfall totals (June to September) to visualize trends and anomalies in the rainfall pattern over time.

$$\text{Z score}= \frac{x-\mu}{\sigma} \qquad (1)$$

Where x is rainfall of a particular month, μ is its arithmetic mean, and σ is its standard deviation.

Crop Yield Prediction Models

Two crop yield prediction models were used in this study to assess the relationship between monsoon rainfall and paddy yield: Multiple Linear Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (ANN-MLP). The MLR model was used to assess how variations in rainfall during different months of the monsoon season influenced paddy yield in each district. The model can be expressed as:

$$Y = b_0 + b_1X_1+b_2X_2+b_3X_3 \qquad (2)$$

Where,

Y = dependent variable (*i.e.* Paddy yield)

$X_1, X_2... X_n$ = independent variables (Month Rainfall of June to September)

b_0, b_1, b_2 … bn = regression coefficients

n = number of independent variables.

ANN-MLP is a type of machine learning model that is capable of capturing complex, nonlinear relationships between variables. The MLP model consists of an input layer, one or more hidden layers, and an output layer. In this study, the input layer consisted of four neurons, representing the monthly rainfall for June, July, August, and September. The output layer had one neuron, representing the predicted paddy yield. The hidden layers allowed the model to capture nonlinear interactions between the rainfall variables. The model was trained using a back propagation algorithm to minimize the prediction error. A hyper parameter tuning process was employed to optimize the number of hidden layers, neurons, and learning rate for the MLP model. To assess the effectiveness of the rainfall data from ERA5, IMD Gridded, and NASA Power in predicting paddy yield, the performance of the MLR and ANN-MLP models was evaluated for each district and each dataset based on Coefficient of Determination (R²) and Root Mean Square Error (RMSE).

Results and Discussion

Statistical Analysis of Monthly Rainfall

The Statistical analysis of monthly rainfall in terms of Arithmetic Mean, Standard Deviation and CV% is given in Fig. 2 to 4.

The statistical analysis of monthly rainfall in Ahmedabad, Dahod, and Panchmahal using three open-source rainfall data products-ERA5, IMD Gridded, and NASA Power (NASAP)-reveals interesting trends in terms of rainfall averages, standard deviations (STDV), and coefficient of variation (CV%). This section compares the data from each product across the months of the monsoon season (June to September) for the three districts.

For Ahmedabad, NASA Power reports the highest June rainfall (121 mm), followed by IMD Gridded (95 mm) and ERA5 (87 mm). ERA5 shows the greatest

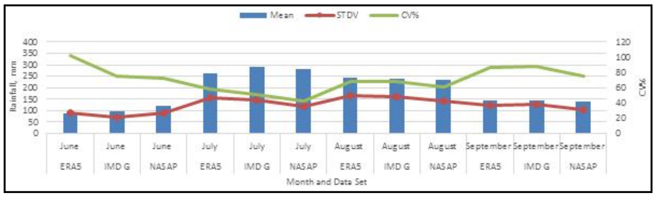

Fig. 2: Arithmetic Mean, Standard Deviation and CV% of Monthly Rainfall based on various data sets for Ahmedabad.

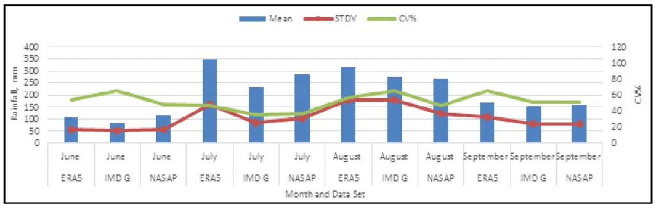

Fig. 3: Arithmetic Mean, Standard Deviation and CV% of Monthly Rainfall based on various data sets for Dahod.

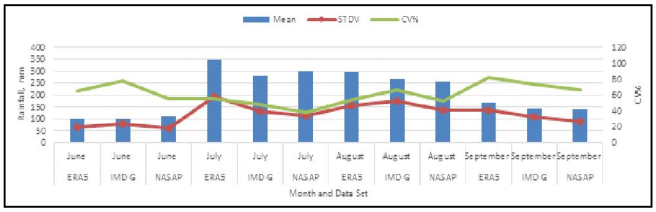

Fig. 4: Arithmetic Mean, Standard Deviation and CV% of Monthly Rainfall based on various data sets for Panchmahal.

variability, with a standard deviation (STDV) of 89 mm, compared to IMD Gridded (71 mm) and NASA Power (88 mm). Both NASA Power and ERA5 indicate higher rainfall in June, which is critical for early paddy growth. In July, all datasets show increased rainfall compared to June, with IMD Gridded reporting the highest (289 mm), closely followed by NASA Power (283 mm), while ERA5 reports the lowest (264 mm). ERA5 also has the highest STDV (155 mm), suggesting greater year-to-year fluctuation. August averages are closer across the datasets, with ERA5 reporting 243 mm, IMD Gridded 238 mm, and NASA Power 233 mm. ERA5 continues to show the highest variability (164 mm), while NASA Power shows the least (144 mm). In September, the average rainfall decreases, with ERA5 and IMD Gridded reporting nearly identical values (~144 mm), and NASA Power slightly lower (138 mm). The STDV is highest for ERA5 (125 mm), indicating more fluctuation, while NASA Power has the lowest variability (104 mm).

In Dahod, NASA Power again reports the highest June rainfall (114 mm), followed by ERA5 (104 mm) and IMD Gridded (83 mm). The standard deviation across all products is similar, ranging from 54 mm to 56 mm. In July, ERA5 shows the highest rainfall (349 mm), well above NASA Power (284 mm) and IMD Gridded (236 mm), with the highest variability again in ERA5 (163 mm). For August, ERA5 reports the most rainfall (316 mm), while NASA Power (266 mm) and IMD Gridded (275 mm) show lower values. The standard deviation is highest for ERA5 (178 mm), while NASA Power has the lowest variability (124 mm). September rainfall averages are similar across the datasets, with NASA Power (156 mm) and IMD Gridded (155 mm) closely aligned, while ERA5 reports slightly higher rainfall (168 mm).

In Panchmahal, NASA Power reports the highest June rainfall (112 mm), while ERA5 and IMD Gridded show equal values (102 mm). NASA Power has the lowest STDV (61 mm), while ERA5 and IMD Gridded show slightly higher variability (67 mm and 80 mm). In July, ERA5 reports the highest rainfall (348 mm), followed by NASA Power (302 mm) and IMD Gridded (280 mm). August follows a similar trend, with ERA5 showing the highest rainfall (293 mm), and September rainfall averages are comparable across the datasets (~145 mm). The rainfall data for Ahmedabad, Dahod, and Panchmahal across the months of June to September show that NASA Power often reports the highest or near-highest average rainfall, particularly in June. ERA5 tends to report higher rainfall values in July and August, especially in Dahod and Panchmahal, but also shows greater variability compared to NASA Power. IMD Gridded data, while useful, consistently reports lower rainfall totals than the other two products, particularly in Dahod. Overall, NASA Power and ERA5 provide robust data that can be effectively used for paddy yield predictions using machine learning models.

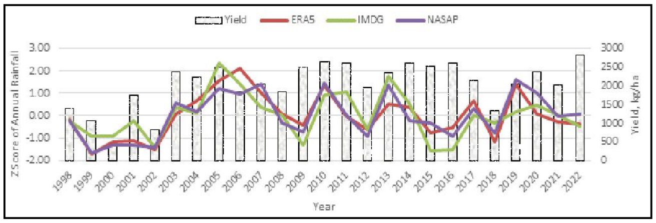

Fig. 5: Historic Paddy yield and corresponding Z Score of annual rainfall for based on ERA5, IMD Gridded and NASA Power data for Ahmedabad District.

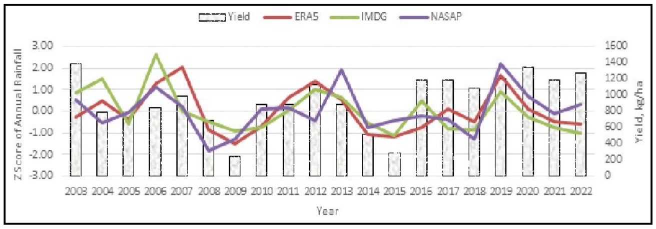

Fig. 6: Historic Paddy yield and corresponding Z Score of annual rainfall for based on ERA5, IMD Gridded and NASA Power data for Dahod District.

Historic Paddy yield and Annual Rainfall

The Historic Paddy yield and corresponding Z score of annual rainfall for based on ERA5, IMD Gridded and NASA Power data for various districts is given in Fig. 5 to 7.

The analysis of historic paddy yields and corresponding Z-scores of annual rainfall from ERA5, IMD Gridded, and NASA Power data reveals key insights into how rainfall anomalies correlate with crop productivity. Both ERA5 and NASA Power datasets emerge as valuable tools, each demonstrating reliable results for assessing rainfall impacts on crop yields.

In high-yield years, both datasets generally provide a good representation of favorable rainfall conditions. For example, in Ahmedabad during 2005, when the paddy yield was 2480 kg/ha, ERA5 reported a Z-score of 1.56, indicating favorable rainfall, while NASA Power also aligned well with a Z-score of 1.23. Although IMD

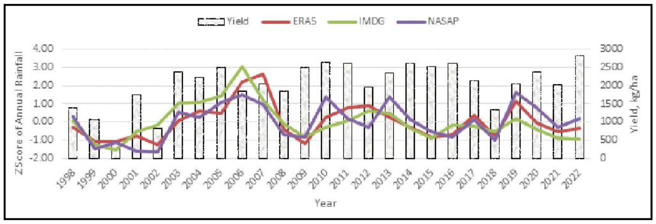

Fig. 7: Historic Paddy yield and corresponding Z Score of annual rainfall for based on ERA5, IMD Gridded and NASA Power data for Panchmahal District.

Gridded overestimated the favorable conditions with a Z-score of 2.33, both ERA5 and NASA Power gave more reasonable estimates, closely reflecting the favorable rainfall experienced that year. Similarly, in 2010, when yields were high at 2640 kg/ha, NASA Power gave a Z-score of 1.45, suggesting abundant rainfall, while ERA5 followed closely with a Z-score of 1.37. Both datasets aligned well with the actual yield, highlighting their reliability in favorable conditions.

In Dahod, during the high-yield year of 2019, where the yield reached 1200 kg/ha, both ERA5 and NASA Power gave reasonable Z-scores that accurately reflected favorable rainfall conditions. ERA5 reported a Z-score of 1.67, closely aligning with the high yields, while NASA Power showed a slightly higher Z-score of 2.20. Although NASA Power indicated more rainfall than ERA5, both datasets accurately captured the overall trend of abundant rainfall that contributed to the high yields.

For low-yield years, both NASA Power and ERA5 continue to perform well, each reflecting rainfall deficits that led to lower crop productivity. For instance, in Ahmedabad during 1999, when the yield was only 1050 kg/ha, ERA5 reported a Z-score of -1.72, indicating severe drought conditions, while NASA Power produced a similarly accurate Z-score of -1.67. Both datasets effectively captured the significant rainfall deficit that impacted yields, demonstrating their usefulness in tracking adverse weather conditions. Similarly, in 2000, when the yield was just 450 kg/ha, NASA Power reported a Z-score of -1.34, slightly more negative than ERA5's -1.18, but both datasets clearly reflected the severe drought that year.

In Dahod during 2009, when yields dropped to 250 kg/ha, ERA5 recorded a Z-score of -1.48, indicating the presence of a major rainfall deficit, while NASA Power closely followed with a Z-score of -1.30. Both datasets showed reliable results in identifying the drought conditions that led to low crop productivity. In Panchmahal during 1999, when yields dropped to 1050 kg/ha, NASA Power recorded a Z-score of -1.47, and ERA5 provided a similar result with a Z-score of -1.10, once again showing the reliability of both datasets in assessing rainfall deficits.

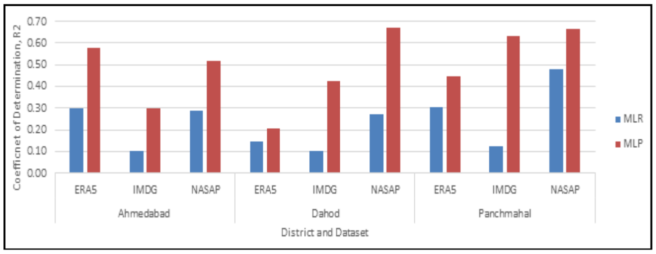

Fig. 8: Coefficient of determination (R^2) for Paddy yield prediction models based Multiple Liner Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (MLP) using monthly rainfall of ERA5, IMD Gridded and NASA Power data.

In summary, both ERA5 and NASA Power prove to be effective and reliable in reflecting rainfall anomalies that correspond to paddy yield outcomes. ERA5 tends to provide slightly more balanced results, particularly in both high and low-yield scenarios, but NASA Power also shows strong consistency, particularly in capturing the extremes of rainfall variability. Both datasets serve as robust tools for assessing rainfall impacts on crop productivity, with their results aligning well with the observed yield outcomes in various years.

Paddy yield Prediction Models

The assessment of paddy yield prediction models through the Coefficient of Determination (R^2) and Root Mean Square Error (RMSE) offers insights into the effectiveness of various rainfall datasets-ERA5, IMD Gridded (IMDG), and NASA Power (NASAP)-when applied to Multiple Linear Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (MLP) methodologies. The Coefficient of determination (R^2) for Paddy yield prediction models based Multiple Liner Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (MLP) using monthly rainfall of ERA5, IMD Gridded and NASA Power data is given in Fig. 8.

In Ahmedabad, the results show that the MLR model using ERA5 data has an R^2 of 0.30, indicating that the model can explain approximately 30% of the variability in paddy yield. The MLP model improves this significantly, achieving an R^2 of 0.58, meaning that it can explain nearly 58% of the variability, suggesting a stronger relationship between the rainfall data and paddy yield. In contrast, the IMDG dataset shows significantly lower R^2 values, with 0.10 for MLR and 0.30 for MLP, indicating that this dataset is not capturing the relationship effectively. NASAP performs moderately, with R^2 values of 0.29 for MLR and 0.52 for MLP, falling between ERA5 and

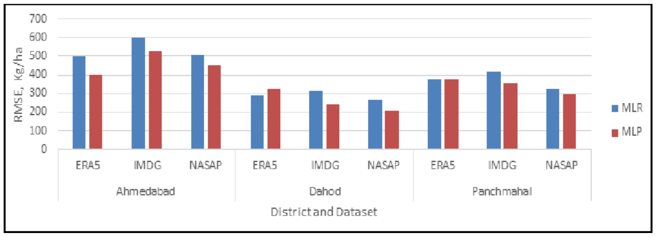

Fig. 9: Root Mean Square Error (RMSE) for Paddy yield prediction models based Multiple Liner Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (MLP) using monthly rainfall of ERA5, IMD Gridded and NASA Power data.

IMDG but still highlighting the limitations of the IMDG dataset.

The findings for Dahod reveal a more nuanced picture. The R² for the ERA5 MLR model is 0.15, indicating a weak relationship, but the MLP model increases this to 0.20, suggesting some level of predictive ability. IMDG data again shows low performance, with 0.10 for MLR, though it improves to 0.43 for MLP, indicating better predictive power with the non-linear model. Notably, the NASAP dataset excels in this region, achieving an R² of 0.27 for MLR and 0.67 for MLP, underscoring its strength in capturing the dynamics between rainfall and yield.

For Panchmahal, the ERA5 dataset shows an R² of 0.31 for MLR and 0.45 for MLP, illustrating moderate effectiveness in predicting paddy yields. The IMDG data continues to lag, with values of 0.13 for MLR and 0.63 for MLP. However, NASAP outshines the others in this district, with R² values of 0.48 for MLR and 0.67 for MLP, indicating a strong capacity for predictive modelling.

The Root Mean Square Error (RMSE) for Paddy yield prediction models based Multiple Liner Regression (MLR) and Artificial Neural Network - Multilayer Perceptron (MLP) using monthly rainfall of ERA5, IMD Gridded and NASA Power data is given in Fig. 9. The RMSE results for Ahmedabad show that the ERA5 dataset has lower errors, with 502 kg/ha for MLR and 399 kg/ha for MLP, indicating better predictive accuracy. IMDG, however, has higher RMSE values, reflecting poor model fit, at 600 kg/ha for MLR and 527 kg/ha for MLP. NASAP presents a competitive RMSE of 506 kg/ha for MLR and 452 kg/ha for MLP, indicating that while it is effective, ERA5 slightly outperforms it in predictive accuracy.

In Dahod, the ERA5 dataset achieves RMSE values of 291 kg/ha for MLR and 328 kg/ha for MLP, reflecting relatively low predictive errors. In contrast, the IMDG data yields higher RMSE values of 315 kg/ha for MLR and 244 kg/ha for MLP. Notably, NASAP demonstrates the lowest RMSE values, at 269 kg/ha for MLR and 206 kg/ha for MLP, suggesting strong predictive capability.

For Panchmahal, the RMSE values for ERA5 are 374 kg/ha for MLR and 376 kg/ha for MLP, indicating that the predictions are reasonably accurate. The IMDG dataset, again, has higher RMSE values of 420 kg/ha for MLR and 356 kg/ha for MLP, which reflects its limitations. NASAP shows a compelling performance, with RMSE values of 324 kg/ha for MLR and 297 kg/ha for MLP, reinforcing its effectiveness in paddy yield prediction.

Overall, the analysis of both R² and RMSE reveals that ERA5 and NASAP datasets are effective for predicting paddy yields, with ERA5 showing slightly better performance in most instances. The consistent higher values of R² and lower RMSE values across both datasets indicate their reliability and robustness in capturing the relationship between rainfall and paddy yield. On the other hand, the IMDG dataset tends to underperform in both metrics, suggesting it may not be the best choice for accurate agricultural predictions.

While NASAP shows strong predictive power, the latency in data availability makes ERA5 the preferred choice for operational applications in paddy yield forecasting. This is crucial for stakeholders in agriculture who rely on timely data for effective decision-making and resource management. The findings highlight the importance of selecting appropriate datasets for modelling agricultural outputs, as they can significantly influence both the accuracy of predictions and the operational effectiveness of agricultural practices.

The ability to accurately predict crop yields based on climatic data is crucial for ensuring food security and optimizing agricultural productivity, particularly in regions like Middle Gujarat that are highly dependent on monsoon rainfall. Open-source rainfall data products, such as ERA5, NASA Power, and IMD Gridded data, offer valuable alternatives to traditional ground-based observations, but their effectiveness for yield prediction must be carefully evaluated. By comparing the performance of these datasets in predicting paddy yield across different districts of Middle Gujarat, this study contributes to the growing body of knowledge on the use of remote sensing and reanalysis data in agricultural decision-making. Furthermore, the application of machine learning models, such as mLR and ANN-MLP, highlights the potential of advanced computational techniques to improve the accuracy of yield forecasts, ultimately supporting more sustainable and resilient agricultural practices.

Conclusion

The statistical analysis of rainfall revealed that NASA Power frequently records the highest or near-highest average rainfall, especially in June, while ERA5 provides substantial totals in July and August but with greater variability. IMD Gridded consistently reports lower totals across the three districts. The robust datasets from NASA Power and ERA5 indicate their potential for effective use in paddy yield predictions using machine learning models. The analysis of historic paddy yields and Z-scores of annual rainfall from ERA5, IMD Gridded, and NASA Power demonstrates that ERA5 and NASA Power effectively reflect how rainfall anomalies correlate with crop productivity in both high and low-yield years. While ERA5 often offers balanced results across various scenarios, NASA Power consistently captures the extremes of rainfall variability, making both datasets reliable tools for assessing rainfall impacts on paddy yields. The assessment reveals that both ERA5 and NASA Power datasets are effective for paddy yield prediction, with ERA5 generally demonstrating slightly better performance in terms of R² and RMSE. However, due to the latency in data availability, ERA5 is recommended for operational applications in agricultural forecasting.

Acknowledgement

The authors gratefully acknowledge the financial support provided by the National Agricultural Science Fund (NASF), Indian Council of Agricultural Research (ICAR), for conducting the study and in facilitating necessary resources.

References

Bhojani, S.H. and Bhatt N. (2021). Performance Analysis of Activation Functions for Wheat Crop Yield Prediction. In IOP Conference Series: *Materials Science and Engineering*, **1042(1)**, 012015. IOP Publishing.

Bhojani, S.H. and Bhatt N. (2020). Wheat crop yield prediction using new activation functions in neural network. *Neural Comput & Applic,* **32**, 13941-13951. https://doi.org/10.1007/s00521-020-04797-8.

Clercq Djavan De and Mahdi Adam (2024). Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. *ELSEVIER journal: Agricultural Systems,* **220**, 104099.

Hughes D.A. (2006). Comparison of satellite rainfall data with observations from gauging station networks. *Journal of Hydrology,* **327**, 399-410.

Hersbach, H., Bell B. and Berrisford P. *et al.,* (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society,* **146(730)**, 1999-2049.

Kumar Ajit and Singh Vivekanand (2023). Comparative Analysis of Gridded Rainfall Datasets Over the Bagmati River Basin, India. *Water Practice & Technology,* **18(12)**, 3141.

Pandya, P.A., Ghosiya S.M., Pithiya V.H. and Dudhatra S.P. (2020). Effect of Rainfall on Productivity of Cotton. *Emer Life Sci Res*, **6(1)**, 56-63 *(Emergent Life Sciences Research).*

Pai, D.S., Sridhar L. and Ramesh K.V. *et al.,* (2014). Development of a new high spatial resolution (0.25° × 0.25°) long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *MAUSAM*, **65(1)**, 1-18.

Pandya, Parthsarthi and Gontia Kumar Narendra (2023). Early crop yield prediction for agricultural drought monitoring using drought indices, remote sensing, and machine learning techniques. *Journal of Water and Climate Change*, **14(12)**, 4729.

Rachidi, S., El-Mazoudi E.H., El-Alami J., Jadoud M. and Er-Raki S. (2023). Assessment and Comparison of Satellite-Based Rainfall Products: Validation by Hydrological Modeling Using ANN in a Semi-Arid Zone. *Water*, **15**, 1997.

Saicharan, V. and Rangaswamy S.H. (2023). A Comparison and Ranking Study of Monthly Average Rainfall Datasets with IMD Gridded Data in India. *Sustainability*, **15**, 5758.

Stackhouse Jr., P.W., Gupta S.K. and Cox S.J. *et al.,* (2019). The NASA POWER project: Improving data accessibility and process efficiency for global assessment and applications. *Asia-Pacific Solar Research Conference*.

Sun, Q., Miao C., Duan Q., Ashouri H., Sorooshian S. and Hsu K.L. (2018). A review of global precipitation data sets: Data sources, estimation, and inter comparisons. *Reviews of Geophysics*, **56**, 79-107.

Singha Kumar Anil, Tripathib B.J.N., Singha K.K., Singha Virendra and Sateesh M. (2019). Comparison of different satellite-derived rainfall products with IMD gridded data over Indian meteorological subdivisions during Indian Summer Monsoon (ISM) 2016 at weekly temporal resolution. *Journal of Hydrology,* **575**, 1371-1379.

Yesilköy, S. and Demir I. (2024). Crop yield prediction based on reanalysis and crop phenology data in the agro climatic zones. *Theor Appl Climatol,* **155**, 7035-7048.