Environment and Ecology 44 (2A) 427—431, April—June 2026
Article DOI: https://doi.org/10.60151/envec/HUOL2630
ISSN: 0970-0420 (Print), ISSN: 3049-3625 (Online)

Principal Component Analysis in Developing Statistical Models for Rice Yield Forecasting Using Weather Variables in Varanasi District of Uttar Pradesh, India

Babu Lal, Vishal Mehta, K. K. Pandey, K. K. Mourya, Amit Kumar Yadav

Received 30 April 2026, Accepted 3 June 2026, Published on 23 June 2026

ABSTRACT

In this study, the statistical models for forecasting rice yield using principal component analysis of weather variables were developed. Time series data on rice yield along with weekly meteorological variables, namely maximum temperature (°C), minimum temperature (°C), relative humidity (%), rainfall (mm), wind velocity (km h^{-1}), evaporation (mm) and sunshine hours (hours). for the period 1986–87 to 2022–23 in Varanasi in district of Eastern Uttar Pradesh were used. Four forecasting models were developed using principal component analysis (PCA) as regressors along with time trend (T) and yield as dependent variable. The performance of the models was evaluated using coefficient of determination (R^2), adjusted R^2 and root mean square error (RMSE). The results indicated that Model-2 performed better among all the models with the lowest RMSE and satisfactory explanatory power. The study shows that PCA-based regression models are useful for rice yield forecasting and can be effectively applied for similar studies in other regions.

Keywords Adjusted R^2, Forecasting, Principal component analysis, Rice crop, RMSE.

INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most important cereal crops and serves as a staple food for a large proportion of the world's population. It is mainly grown in tropical and subtropical regions under warm and humid climatic conditions, requiring adequate rainfall and temperature for optimum growth. Rice cultivation is generally suitable for clayey and loamy soils with good water holding capacity, as the crop thrives well under flooded conditions. The growth and yield of rice are highly influenced by weather parameters such as temperature, rainfall, humidity and sunshine, making weather-based analysis essential for accurate yield prediction. Principal Component Analysis (PCA) is an important multivariate statistical technique widely used for dimensionality reduction and extraction of significant information from large datasets. It transforms correlated variables into a smaller number of uncorrelated principal components without much loss of information (Rao 1964, Manly 2000). PCA also helps in reducing multicollinearity and improving model stability.Several studies have contributed to the development and application of PCA in different fields. Jolliffe (1972, 1973) proposed

Babu Lal[1], Vishal Mehta[2]*, K. K. Pandey[3], K. K. Mourya[4], Amit Kumar Yadav[5]

[1,5]PhD Research Scholar,

[2,4]Assistant Professor, [3]Associate Professor, Department of Agricultural Statistics

[1,2,3,4,5]Acharya Narendra Deva University of Agriculture and Technology, Kumarganj, Ayodhya 224229 (UP), India

Email: mehta.vishal@nduat.org
*Corresponding author

methods for variable selection using principal components, while McCabe (1984) introduced the concept of principal variables for selecting optimal subsets of variables. Richman (1986) further explained the theoretical aspects of PCA and compared rotated and unrotated components. In agricultural research, PCA has been widely used for crop yield forecasting and identification of important variables. Jain *et al.* (1984) utilized principal components of biometrical characters for yield forecasting. Verma *et al.* (2015) applied PCA for pre-harvest estimation of cotton yield, demonstrating its usefulness in early prediction. Sharma and Devi (2023) developed wheat yield models using PCA and reported better performance compared to conventional regression models in terms of lower percentage deviation. Further studies have highlighted the efficiency of PCA in reducing redundancy and improving prediction accuracy. Das *et al.* (2017) showed that PCA effectively identifies patterns and removes redundancy in agricultural datasets. Muema *et al.* (2018) used PCA to evaluate irrigation systems in rice cultivation, indicating its applicability in agricultural decision-making. Kumar *et al.* (2019) demonstrated that principal component regression provides better yield prediction with higher R^2 and lower error compared to traditional regression techniques. Recent advancements have also incorporated PCA in combination with modern statistical and machine learning approaches. Abbas *et al.* (2021) applied PCA along with regression techniques to identify key factors affecting wheat yield, while Jiang *et al.* (2022) used PCA to improve prediction models by reducing the dimensionality of meteorological data. Similarly, Pham *et al.* (2022) integrated PCA with machine learning techniques to enhance agricultural yield prediction accuracy. Yadav *et al.* (2014) applied Principal Component Analysis (PCA) to weather variables for developing crop yield forecasting models. The study showed that PCA effectively reduced multicollinearity and improved prediction accuracy, making it a useful approach for weather-based yield forecasting.

MATERIAL S AND METHOD

Sources and description of data

Time series data on rice yield for Varanasi districts of Eastern Plain Zone (EPZ), Uttar Pradesh, for the period 1986-87 to 2022-23 were obtained from the Economics and Statistics Division, Planning Department, Government of Uttar Pradesh (UPDES 2023). Weekly time series data on meteorological variables for the same period (1986-2023) were collected from India Meteorological Department (IMD), Pune. The data covered the rice crop season from 24th to 42nd standard meteorological weeks (SMW), corresponding to the period from 11 June to 21 October. Seven weather variables were considered, namely maximum temperature (°C), minimum temperature (°C), relative humidity (%), rainfall (mm), wind velocity (km h^{-1}), evaporation (mm) and sunshine hours (hours).

Development of weather indices using correlation coefficient as weighted

Weather indices were developed following the procedure suggested by Agrawal *et al.* (1983) using weekly meteorological data of the crop growing period. In this study, weekly data corresponding to 24th to 42nd standard meteorological weeks (SMW) were used.

Seven weather variables, namely maximum temperature, minimum temperature, relative humidity, rainfall, wind velocity, evaporation and sunshine hours were considered for constructing the indices. The weather indices were generated in two forms: unweighted (individual) indices and weighted indices. The unweighted indices were obtained by taking simple averages of the weekly weather variables, while the weighted indices were computed using correlation coefficients between crop yield and the corresponding weekly weather variables as weights.

In addition to individual indices, interaction indices between pairs of weather variables were also developed. In total, 56 indices were obtained, comprising 7 unweighted individual indices, 7 weighted individual indices, 21 unweighted interaction indices and 21 weighted interaction indices.

$$Z_{ij} = \sum_{w=1}^{n} r^{j}_{jw} X_{iw} \Big/ \sum_{w=1}^{n} r^{j}_{iw} \tag{1}$$

$$Z_{ij,j} = \sum_{w=1}^{n} r^{j}_{ii'w} X_{iw} X_{i'w} \Big/ \sum_{w=1}^{n} r^{j}_{ii'w}, \quad j=0.1 \text{ and } i=1,2..., p \tag{2}$$

Where Z_{ij} is unweighted (for j = 0) and weighted (for j=1) weather indices for weather variable and $Z_{ii'j}$ is the unweighted (for j = 0) and weighted (for j = 1) weather indices for interaction between i^{th} and i'^{th} weather variables. X_{iw} is the value of the i^{th} weather variable in w^{th} week, r_{iw} / r_{iiw} is correlation coefficient of yield adjusted for trend effect with i^{th} weather variable/product of i^{th} and i'^{th} weather variable in w^{th} week, n is the number of weeks considered in developing the indices and p is number of weather variables.

Model-1

Model-1 was developed using all weather indices, consisting of 14 individual indices (7 unweighted and 7 weighted) and 42 interaction indices (21 unweighted and 21 weighted). These indices were subjected to Principal Component Analysis (PCA) to reduce dimensionality.

The principal components were arranged based on their eigenvalues, and components having eigenvalues greater than one were retained. A total of 9 principal components were selected, which together explained 97.70% of the total variation in the data. These selected principal components were used as regressors along with time trend (T) for developing the forecasting model.

Model-

$$Y= \beta_0 + \beta_1PC_1 + \beta_2PC_2 + \beta_3PC_3 + \beta_4PC_4 + \beta_5PC_5 + \beta_6PC_6 + \beta_7PC_7 + \beta_8PC_8 + \beta_9PC_9 + \beta_{10}T + e$$

where,
Y = un-trended rice yield
$\beta_0, \beta_1, \beta_2 \ldots \beta_{10}$ are regression parameters
PC_1 to PC_9 are the first nine principal components
T = time trend variable
e = random error term assumed to follow normal distribution with mean 0 and variance σ^2

Model-2

Model-2 was developed using 14 individual weather indices (7 unweighted and 7 weighted). Principal Component Analysis (PCA) was applied to these indices to reduce dimensionality.

Based on the eigenvalue greater than one criterion, the first 5 principal components were retained. These components together explained 84.22% of the total variation in the data. The selected principal components were used as regressors along with time trend (T) in the forecasting model.

$$Y = \beta_0 + \beta_1PC_1 + \beta_2PC_2 + \beta_3PC_3 + \beta_4PC_4 + \beta_5PC_5 + \beta_6T + e$$

where,

Y = un-trended rice yield
$\beta_0, \beta_1, \beta_2 \ldots \beta_6$ are regression parameters
PC_1 to PC_5 are the first five principal components
T = time trend variable
e = random error term assumed to follow normal distribution with mean 0 and variance σ^2

Model-3

Model-3 was developed using 28 unweighted weather indices, consisting of 7 unweighted individual indices and 21 unweighted interaction indices. These indices were subjected to Principal Component Analysis (PCA) to reduce dimensionality.

Based on the eigenvalue greater than one criterion, the first 5 principal components were retained. These components together explained 84.22% of the total variation in the data. The selected principal components were used as regressors along with time trend (T) for developing the forecasting model.

$$Y = \beta_0 + \beta_1PC_1 + \beta_2PC_2 + \beta_3PC_3 + \beta_4PC_4 + \beta_5PC_5 + \beta_6T + e.$$

where,
Y = un-trended rice yield
$\beta_0, \beta_1, \beta_2 \ldots \beta_6$ are regression parameters
PC_1 to PC_5 are the first five principal components
T = time trend variable
e = random error term assumed to follow normal distribution with mean 0 and variance σ^2

Model-4

Model-4 was developed using 28 weighted weather indices, consisting of 7 weighted individual indices and 21 weighted interaction indices. These indices were subjected to Principal Component Analysis (PCA) to reduce dimensionality.

Based on the eigenvalue greater than one criterion, the first 6 principal components were retained. These components together explained 95.63% of the total variation in the data. The selected principal components were used as regressors along with time trend (T) for developing the forecasting model.

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \beta_4 PC_4 + \beta_5 PC_5 + \beta_6 PC_6 + \beta_7 T + e$$

where,

Y = un-trended rice yield
$\beta_0, \beta_1, \beta_2 \ldots \beta_7$ are regression parameters
PC_1 to PC_6 are the first six principal components
T = time trend variable
e = random error term assumed to follow normal distribution with mean 0 and variance σ

RESULTS AND DISCUSSION

The forecast regression models developed using principal components are presented in Table 1 along with their values of R^2 and adjusted R^2. In all the models, the second principal component (PC2) showed a significant effect on rice yield. In Model-3, the time trend (T) was also found to be significant. The value of adjusted R^2 was found to be maximum in Model-4 (46.46%) followed by Model-1 (45.91%), while Model-2 (43.31%) and Model-3 (29.48%) showed relatively lower explanatory power.

Using these models, the forecast values of rice yield for the validation years were obtained and are presented in Table 2. The predicted values were found to be close to the observed yields, and the percentage deviation from the actual yield is given in brackets. The RMSE values for Model-1, Model-2, Model-3 and Model-4 were 2.89, 2.22, 2.39 and 3.24, respectively. Among all the models, Model-2 showed the lowest RMSE value, indicating better prediction accuracy compared to the other models. The performance of the developed models in terms of R^2 and RMSE is also illustrated in Fig. 1 and Fig. 2.

Table 1. Rice yield forecast models based on principal components.

Model	Forecast regression equation	R^2 (%)	Adj. R^2 (%)
Model-1	Yield = 20.21 − 0.228PC1 − 2.505**PC2 + 0.457PC3 + 0.044PC4 − 1.043PC5 + 0.103T	55.75	45.91
Model-2	Yield = 19.43 − 0.733PC1 − 2.407**PC2 − 0.567PC3 − 0.708PC4 + 0.080PC5 + 0.146T	53.61	43.31
Model-3	Yield = 18.84 − 0.911PC1 − 1.869**PC2 + 0.885PC3 + 0.020PC4 + 0.510PC5 + 0.174*T	42.30	29.48
Model-4	Yield = 20.55 − 0.873PC1 − 2.580**PC2 − 0.155PC3 + 0.181PC4 − 0.429PC5 + 0.083T	56.19	46.46

Note: *$P < 0.05$, **$P < 0.01$

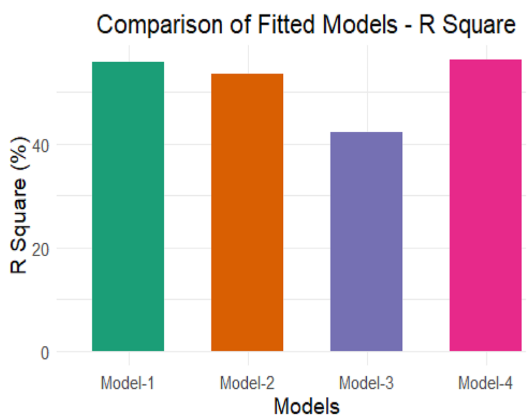

Fig. 1. Comparison of developed models in terms of R^2.

Fig. 2. Comparison of developed models in terms of RMSE.

Table 2. Actual and forecasted rice yield (Q/ha).

Year	Actual yield	Model-1	Model-2	Model-3	Model-4
Year-1	26.29	27.37 (-4.10)	26.78 (-1.87)	25.92 (1.40)	26.26 (0.11)
Year-2	27.41	25.98 (5.22)	26.83 (2.11)	26.50 (3.31)	25.61 (6.57)
Year-3	28.23	25.80 (8.61)	27.00 (4.34)	26.94 (4.56)	25.57 (9.42)
Year-4	30.78	25.86 (15.98)	26.58 (13.64)	26.28 (14.63)	25.15 (18.29)
RMSE		**2.89**	**2.22**	**2.39**	**3.24**

Note: Figures in parentheses indicate percentage deviation of forecast from actual yield.

CONCLUSION

The present study demonstrated that Principal Component Analysis (PCA) based models can be effectively used for rice yield forecasting using weather variables. Among the developed models, Model-2 performed better as it showed the lowest RMSE (2.22) and satisfactory explanatory power. The results indicate that the selected principal components were able to capture the major variability in the data and provide reliable yield predictions.

Therefore, Model-2 can be used for forecasting rice yield for the study area. The approach adopted in this study can also be applied to other crops and regions for improving yield prediction using weather-based models.

ACKNOWLEDGMENT

The authors are grateful to the editor and reviewers for their valuable comments and constructive suggestions, which significantly improved the quality of the revised manuscript.

REFERENCE

Abbas, K., Hussain, Z., Hussain, M., Rahim, F., Ashraf, N., Khan, Q., Raza, G., Ali, A., Khan, D. M., Khalil, U., & Irshad, N. (2021). Statistical modeling for analyzing grain yield of durum wheat under rainfed conditions in Azad Jammu Kashmir, Pakistan. *Brazilian Journal of Biology,* 5(1), 82.

Agrawal, R., Jain, R. C., & Jha, M. P. (1983). Joint effects of weather variables on rice yield. Mausam, 34(2), 189-194.

Das, S., Das, S. S., Chakraborty, I., Roy, N., Nath, M. K., & Sarma, D. (2017). Principal component analysis in plant breeding. J*ournal of Plant Breeding and Crop Science,* 4(6), 87–93.

Jain, R. C., Sridharan, H., & Agrawal, R. (1984). Principal component technique for forecasting of sorghum yield. *Indian Journal of Agricultural Science*, 54(6), 467-70.

Jiang, S., Chen, G., Chen, D., & Chen, T. (2022). Application and evaluation of an improved LSTM model in the soil moisture prediction of Southeast Chinese Tobacco-Producing Areas. Journal of the Indian Society of Remote Sensing, 51(9), 1843–1853. https://doi.org/10.1007/s12524-021-01438-y

Jolliffe, I. T. (1973). Discarding variables in a principal component Analysis. II: real data. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 22(1), 21. https://doi.org/10.2307/2346300.

Kumar, M., Battan, K. R., & Sheoran, O. P. (2019). Pre-harvest .forecast model for rice yield using principal component regression based on biometrical character with R software. *International Journal of Agricultural and Statistical Sciences,* 15(1), 323–326.

Manly, B. F. J. (2000). *Multivariate statistical methods: A primer* (2^{nd} ed.). Chapman and Hall/CRC.

McCabe, G. P. (1984). Principal variables. Technometrics, 26(2), 137–144.

Muema, F. M., Home, P. G., & Raude, J. M. (2018). Application of benchmarking and principal component analysis in measuring performance of public irrigation schemes in Kenya. *Agriculture,* 8(10), 162.

Pham, H. T., Awange, J., Kuhn, M., Nguyen, B. V., & Bui, L. K. (2022). Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices. *Sensors,* 22(3), 719.

Rao, C. R. (1964). The use and interpretation of principal compo nent analysis in applied research. Sankhya: *The Indian Journal of Statistics, Series A*, 26, 329–358.

Richman, M. B. (1986). Rotation of principal components. *International Journal of Climatology*, 6(3), 293–335.

Sharma, A., & Devi, M. (2023). Development of Weather based Yield Models for Wheat Crop in Western Zone of Haryana. International *Journal of Education and Management Studies,* 13(2), 131-136.

Verma, U., Aneja, D. R., & Hooda, B. K. (2015). Principal com ponent technique for pre-harvest estimation of cotton yield based on plant biometrical characters. *Journal of Cotton Research and Development,* 29(2), 339–343.

Yadav, R. R., Sisodia, B. V. S., & Kumar, S. (2014). Application of principal component analysis in developing statistical models to forecast crop yield using weather variables. *Mausam,* 65(3), 357–360.