**Development of rice yield forecasting model using linear regression for Imphal west district, Manipur, India**

**Abstract:** Rice is a staple food crop and India's principal food grain. It is generally grown under completely flooded conditions and any changes in weather parameters might affect the rice productivity thereby impacting the food security of the ever-increasing population Prevailing weather conditions during the crop growth period determine the yield of Rice. Hence, the crop yield forecasting models based on weather parameters will be an appropriate option for policymakers and researchers to develop sustainable cropping strategies. The present study examines the application of stepwise multiple linear regression for rice yield prediction using long-term weather data. Analysis was carried out by fixing data from 1998-99 to 2016-17 for calibration and the remaining 2017-18 to 2020-21 data for validation. The accuracy of these models was estimated by $R^2$ (coefficient of determination) and the performance by Mean Square error (MSE), Root mean square error (RMSE), and Normalised root mean square error (NRMSE). The $R^2$ of the developed models ranged from 0.27 – 0.95. The best-performing model was the $5^{th}$ model with $R^2$ (0.95) with MSE (0.03%), RMSE (0.17%), and NRMSE (0.05%) during validation.

**Keywords:** Rice yield prediction, weather parameters, statistical models, stepwise multiple linear regression,



## Introduction

"Rice (*Oriza sativa*) is India's principal food grain crop which occupies about 43.50 m ha area with the production of 104.32 million tons (Government of India, Ministry of Agriculture and Farmers Welfare: Department of Agriculture Cooperation, and Welfare 2016). As per the Directorate of Rice Development, Patna, Government of India, Manipur ranked 8th in the country with 2369 kg/ha in 2006-07 in rice yield. In the past few decades, rice production and productivity in India have shown remarkable growth. In this regard, crop yield forecasting is essential for proper planning and policy-making to manage excess produce" (Dutta et al. 2001). "Crop yield Prediction is important for agricultural planning and resource distribution decision-making. Efficient models were developed, to help reduce the error. Achieving maximum crop yield at minimum cost is one of the goals of agricultural production. Early detection and management of problems associated with crop yield indicators can help increase yield and subsequent profit. There are mainly two types of approaches to forecasting crop yield: crop simulation and empirical statistical models" (Bocca and Rodrigues 2016). "Crop simulation models are process-based and input-data-intensive. Though crop simulation models are precise, hardly these models can be applied to large spatiotemporal scales due to the unavailability of sufficient input data. On the other hand, empirical statistical models are simple and require less input data. So, statistical models using crop yield and weather data using simple regression techniques have been broadly used as a common alternative to process-based models" (Lobell and Burke 2010; Shi et al. 2013). "Multiple linear regression (MLR) is the standard and simplest approach for the development of calibration models. Feature selection in the form of stepwise MLR (SMLR) gives good results over large datasets. A stepwise regression procedure was adopted for the selection of the best regression variable among many independent variables and found that models were able to explain 51 to 79% variability for rice yield" (Singh et al. 2014). The study determined

the predominance of various meteorological data on the yield of the crop. Crop yield is mostly affected by technological changes and weather variability. It can be assumed that the technological factors will increase yield smoothly through time and therefore, year or other parameters of time can be used to study the overall effect of technology on crop yield. Weather variables affect the crop differently during different stages of development. Thus, there is a requirement to quantify the relationship between crop yield and weather variables to predict the regional yield, so that it may be useful for the farmers and policymakers. In this present study, rice forecasting models were developed using SMLR for the Imphal west district of Manipur, India.

## Materials and Method

Manipur, the state of India, is located in the northeastern part of the country. It is bordered by the Indian states of Nagaland to the north, Assam to the west, Mizoram to the southwest, and Myanmar (Burma) to the south and east. It is located at a longitude of 93°03'E to 94°78'E and a latitude of 23°56'N to 25°68'N. Manipur has a total geographical area of 22, 327 sq. km., out of which ninety percent (20, 089 sq. km.) is covered under hill districts and the remaining (2, 238 sq. km.) under valley districts. Temperature varies from sub 0 to 36□ and its annual temperature falls between 20□ to 25□. The climate is temperate in the valley and cold in the hills. In summer the average high temperature is in the low 90s F (about 32–34 °C), while in the winter temperatures can drop into the mid-30s F (about 1–2 °C). Rainfall is abundant, with about 65 inches (1,650 mm) of precipitation occurring annually. The study was based on the Imphal West district of Manipur, India. (Fig. 1).

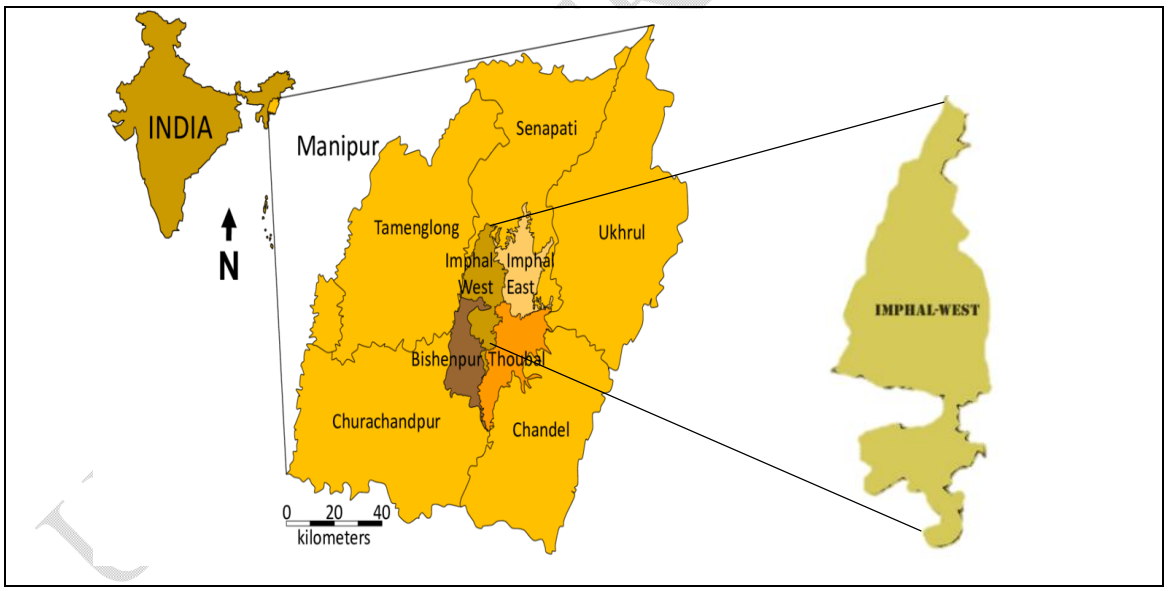


**Figure 1. Study map**

***Development of rice yield forecasting model using Multiple Linear Regression.***

Daily weather data viz maximum temperature, minimum temperature, and rainfall were obtained from NASA's Prediction of Worldwide Energy Resources (NASA/POWER; power.larc.nasa.gov) and morning and evening relative humidity (0830 & 1730 hrs) were collected from Regional Meteorological Centre, Guwahati. The data on five weather variables namely maximum and minimum temperature (Tmax and Tmin, °C), rainfall (RF mm) morning relative humidity (RH1 %), and evening relative humidity (RH2 %), for 23 weeks of crop growing period, which includes 23rd standard meteorological week (SMW) to

45th SMW had been used in the study. Daily data of Tmax, Tmin, RF, RH1, and RH2 had been converted into their weekly average values with the help of Weather cock. Out of the 23-year data, 19-year data were used for model calibration while the remaining 4 years data were used for model validation. Stepwise multiple linear regression (SMLR) was used to develop the model with 115th weather variables, and five (5) models were developed with the help of SPSS (Statistical Package for the Social Science).

***Model performance***

The accuracy of the forecasting models developed was estimated by the value of $R^2$ and the performance of the developed models was estimated by calculating MSE, RMSE, and NRMSE, the formula for calculating the statistical methods are as follows:

R-squared ($R^2$) which is the coefficient of determination represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model. Adjusted $R^2$ identifies the percentage of variance in the target field that is explained by the input or inputs.

$$\mathbf{R^2 = [(n\,(\sum xy) - \sum x \sum y / \sqrt{\{n^*\,(\sum x^2 - (\sum x)^2)] * [n^*\,(\sum y^2 - (\sum y)^2)\}]^2},}$$

where, n = number in the given dataset, x = first variable in the context, y = second variable.

Mean Square Error (MSE) measures the amount of error in statistical models.

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n},$$

where $y_i$ is the $i^{th}$ observed value, $\hat{y}_i$ is the corresponding predicted value, and n = the number of observations.

Root Mean Square Error (RMSE) is used as a measure of comparing different models.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)^2},$$

where $O_i$ is the observed value and $M_i$ is the estimated value.

Normalized root mean square error (NRMSE) relates the RMSE to the observed range of the variable.

$$\mathrm{nRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)} \times \frac{100}{\bar{O}},$$

where, $\bar{O}$ is the average observation value.

"$R^2$ values close to 1 and RMSE close to 0 indicated better model performance. According to nRMSE, the model was considered excellent, good, fair, and poor when the values ranged < 10%, 10–20%, 20–30%, and > 30%, respectively" (Jamieson et al. 1991).

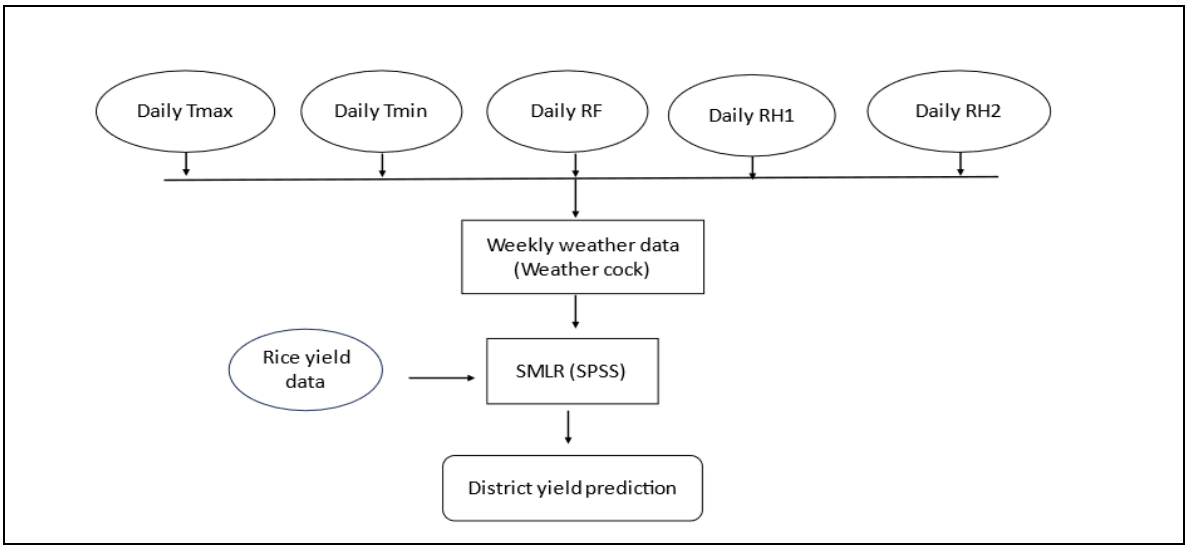


**Figure 2. Flowchart representing model preparation**

## Results and Discussion

Yield prediction models for rice crops have been developed using long-term crop yield data as well as long-period weekly weather data during the crop growing period (23rd to 45th standard meteorological week) of (1998 – 2016) as a regressor and yield data from (1998 – 2016) as dependent variable. Here, five yield prediction models were developed for rice yield prediction in Imphal west district, Manipur. The coefficient of determination of these multivariate models has been presented in Table 1. The highest $R^2$ value was of (Model 5-0.95), while the lowest was Model-1 with just one parameter i.e., rainfall of 43rd week after sowing. These models were developed using Stepwise multiple linear regression (SMLR) and the performance of the models was categorized based on the value of MSE, RMSE, and NRMSE during calibration and validation are presented in Table 1 and Table 2 respectively.

**Table 1. Developed yield forecasting models for the Imphal West district.**

| Models | Equation | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| Model 1 | Y= 3.013 + 0.007 * X1 | 0.27 | 0.22 |
| Model 2 | Y= 6.721 + 0.010* X1 – 0.130* X2 | 0.69 | 0.65 |
| Model 3 | Y= 6909+ 0.010* X1-0.131 * X2 -0.006 * X3 | 0.86 | 0.83 |
| Model 4 | Y= 7.963+ 0.011 * X1 – 0.128 * X2 -0.007 * X3 – 0.041 * X4 | 0.92 | 0.90 |
| Model 5 | Y= 6.690 + 0.011 * X1 – 0.139 * X2 – 0.004 * X3 – 0.063 * X4 + 0.076 * X5 | 0.95 | 0.93 |

where, X1= rainfall of 43rd week after sowing, X2= maximum temperature of 23rd week after sowing, X3= rainfall of 35th week after sowing, X4= maximum temperature of 33rd week after sowing, X5=maximum temperature of 35th week after sowing, $R^2$ = Coefficient of determination.

**Table 2. Performance of the developed models during calibration and validation time**.

| Models | During calibration | | | During validation | | |
|---|---|---|---|---|---|---|
| | MSE | RMSE | NRMSE | MSE | RMSE | NRMSE |
| Model 1 | 0.003 | 0.06 | 0.02 | 0.11 | 0.32 | 0.09 |
| Model 2 | 0.002 | 0.05 | 0.02 | 0.07 | 0.27 | 0.07 |
| Model 3 | 0.003 | 0.05 | 0.02 | 0.07 | 0.27 | 0.07 |
| Model 4 | 0.003 | 0.05 | 0.02 | 0.09 | 0.30 | 0.08 |
| Model 5 | 0.001 | 0.04 | 0.01 | 0.03 | 0.17 | 0.05 |

The weather variables which were identified to be important through stepwise multiple linear regression were rainfall of the 43$^{rd}$ week, the maximum temperature of the 23$^{rd}$ week, rainfall of the 35$^{th}$ week, the maximum temperature of the 33$^{rd}$ week, and maximum temperature of the 35$^{th}$ week after sowing. The most accurate model out of all the developed models was Model-5 with $R^2$ (0.95) and adjusted $R^2$ (0.93), with five weather variables (mentioned above), during calibration, MSE (0.001%), RMSE (0.04%), and NRMSE (0.01%), and during validation was MSE = 0.09%, RMSE = 0.17% and NRMSE = 0.05%. The least accurate model was Model-1 with only one weather parameter which is a rainfall of 43$^{rd}$ week after sowing, and its $R^2$ value (0.27) and adjusted $R^2$ (0.22). Model 2 and 3 have $R^2$ (0.69; 0.86) respectively and the performance was quite the same with MSE = 0.07%, RMSE= 0.27%, and NRMSE = 0.07% approximately during validation and during calibration MSE (0.003%), RMSE (0.05%) and NRMSE (0.02%). Model-4 with $R^2$ (0.92) also had the same values of MSE, RMSE, and NRMSE (as Model-3) during calibration but during validation, it had MSE =0.09%, RMSE= 0.3%, and NRMSE=0.08%. The graphical representation of the comparison of different models is in Figure 3. During the validation period (2017-18 to 2020-21), the estimated yield of the most accurate model developed, obtained from the multivariate equations from Table 1 were 2017-18 (3.24t/ha), 2018-19 (3.14 t/ha), 2019-20 (3.69 t/ha) and 2020-21 (3.44 t/ha), Table 3.

**Table 3. Rice yields estimated by different models during validation for Imphal West.**

| Year | Actual yield | Estimated yield 1 | Estimated yield 2 | Estimated yield 3 | Estimated yield 4 | Estimated yield 5 |
|---|---|---|---|---|---|---|
| 2017-18 | 3.30 | 3.32 | 3.53 | 3.42 | 3.44 | 3.24 |
| 2018-19 | 3.19 | 3.08 | 3.31 | 3.20 | 3.12 | 3.14 |
| 2019-20 | 3.01 | 4.25 | 3.99 | 4.03 | 4.13 | 3.69 |
| 2020-21 | 3.47 | 3.84 | 3.85 | 3.79 | 3.83 | 3.44 |

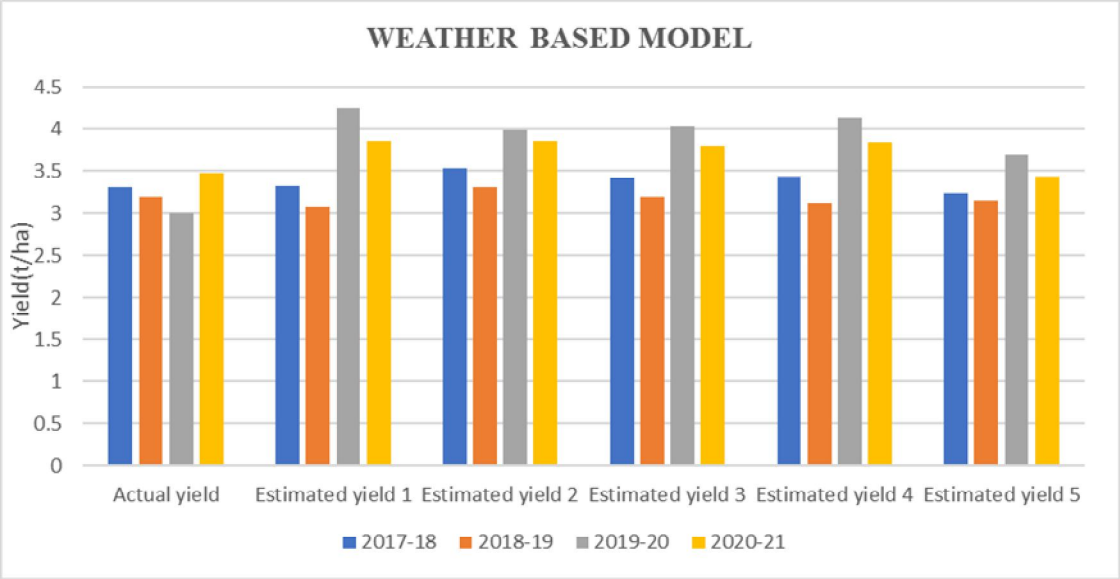


**Figure 3. Comparison of different developed models for Imphal West district**.

## Conclusion

In the present study, five models were developed by using Stepwise multiple linear regression in SPSS software with long-term weather data and yield data for 23 years (1998-99 to 2020-21) for Imphal West district, Manipur. The models were compared and the most accurate model was chosen by $R^2$ value nearest to 1, which was Model-5 which has an $R^2$ value (0.95), the performance of the model during validation was MSE = 0.11%, RMSE= 0.33%, and NRMSE = 0.09%. Hence, SMLR can be used for developing a forecasting model for future rice yield prediction.

## REFERENCES

Balabin, R. M., Lomakina, E. I., & Safieva, R. Z. (2011). Neural network (ANN) approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol, and water contents using near infrared (NIR) spectroscopy. *Fuel*, *90*(5), 2007-2015.

Bocca, F. F., & Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modeling. *Computers and electronics in agriculture*, *128*, 67-76.

Bhattacharyya, B., Biswas, R., Sujatha, K., & Chiphang, D. Y. (2021). Linear Regression Model to Study the Effects of Weather Variables on Crop Yield in Manipur State. *Int. J. Agricult. Stat. Sci*, *17*(1), 317-320.

Cai, Q., Wang, W. and Wang, S. (2015). Multiple regression model based on weather factors for predicting the heat load of a district heating system in Dalian, China a case study. *The Open Cybernetics & Systemics Journal*, 9, 2755-2773.

Das, B., Nair, B., Reddy, V. K., & Venkatesh, P. (2018). Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. *International journal of biometeorology*, *62*(10), 1809-1822.

Dutta, S., Patel, N. K., & Srivastava, S. K. (2001). District-wise yield models of rice in Bihar based on water requirement and meteorological data. *Journal of the Indian Society of Remote Sensing*, *29*, 175-182.

Han X, Chang L, Wang N, Kong W, Wang C (2023). Effects of Meteorological Factors on Apple Yield Based on Multilinear Regression Analysis: A Case Study of Yantai Area, China. *Atmosphere*. 2023; 14(1):183. https://doi.org/10.3390/atmos14010183

Jain, R. C., Agrawal, R., & Jha, M. P. (1980). Effect of climatic variables on rice yield and its forecast. *Mausam*, *31*(4), 591-596.

Jamieson, P. D., Porter, J. R., & Wilson, D. R. (1991). A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. *Field crops research*, *27*(4), 337-350.

K.S. Aravind, Ananta Vashisth, Krishnan and B. Das (2022). Wheat yield prediction based on weather parameters using multiple linear, neural network and penalized regression model. *Journal of Agrometeorology*., (2022) 24(1): 18-25.

Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and forest meteorology*, *150*(11), 1443-1452.

Mallick, K., Mukherjee, J., Bal, S. K., Bhalla, S. S., & Hundal, S. S. (2007). Real time rice yield forecasting over Central Punjab region using crop weather regression model. *Journal of Agrometeorology*, *9*(2), 158-166.

Matsumura, K., Gaitan, C. F., Sugimoto, K., Cannon, A. J., & Hsieh, W. W. (2015). Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. *The Journal of Agricultural Science*, *153*(3), 399-410.

Nain, G., Bhardwaj, N., Jaslam, P. M., & Dagar, C. S. (2021). Rice yield forecasting using agro-meteorological variables: A multivariate approach. *Journal of Agrometeorology*, *23*(1), 100-105.

Parekh, F. P., & Suryanarayana, T. M. V. (2012). Impact of climatological parameters on yield of wheat using neural network fitting. *International Journal of Modern Engineering Research*, *2*(5), 3534-3537

Piekutowska M, Niedbała G, Piskier T, Lenartowicz T, Pilarski K, Wojciechowski T, Pilarska AA, Czechowska-Kosacka A. (2021). The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy*. 2021; 11(5):885

Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S. D., Sanjivaiah, S. H., ... & Elansary, H. O. (2020). Weather-based neural network, stepwise linear and sparse regression approach for rabi sorghum yield forecasting of Karnataka, India. *Agronomy*, *10*(11), 1645.

Shi, W., Liu, C., Shu, Y., Feng, C., Lei, Z., & Zhang, Z. (2013). Synergistic effect of rice husk addition on hydrothermal treatment of sewage sludge: fate and environmental risk of heavy metals. *Bioresource Technology*, *149*, 496-502.

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta□Espino, J., Bhavani, S., Barbier, H., ... & Sorrells, M. E. (2014). Genomic selection for quantitative adult plant stem rust resistance in wheat. *The plant genome*, *7*(3), plantgenome2014-02.

Vanitha, G., Kennedy, J. S., Prabhu, R., & Rajkishore, S. K. (2021). Trained neural network to predict paddy yield for various input parameters in Tamil Nadu, India. *Journal of Applied and Natural Science, 13*(SI), 135-141.