MAUSAM, **75**, 1 (January 2024), 191-196

MAUSAM

DOI : https://doi.org/10.54302/mausam.v75i1.3576
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No. 633.18 (540.27)

Comparative analysis of SMLR, ANN, Elastic net and LASSO based models for rice crop yield prediction in Uttarakhand

PARUL SETIYA, AJEET SINGH NAIN and ANURAG SATPATHI

Department of Agrometeorology, College of Agriculture, G. B. Pant University of Agriculture & Technology, Pantnagar, India

(*Received 1 November 2021 Accepted 12 June 2023*)

e mail : parul.setiya@gmail.com

सार - इस अध्ययन का उद्देश्य चावल की फसल की उपज के लिए उपज पूर्वानुमान मॉडल विकसित करना था। पूर्वानुमान मॉडल बनाने के लिए चार अलग-अलग तकनीकों यानी स्टेपवाइज मल्टीपल लीनियर रिग्रेशन (SMLR), **आर्टिफिशियल न्यूरल नेटवर्क** (ANN), **लीस्ट एब्सोल्यूट श्रिंकेज एंड सिलेक्शन ऑपरेटर** (LASSO) **और इलास्टिक नेट** (ELNET) **का उपयोग किया गया। पूर्वानुमान मॉडल विकसित करने के लिए** 15 **वर्षों के मौसम संबंधी आँकड़े और फसल की उपज आँकड़ों का उपयोग किया गया है। विकसित मॉडलों को तीन वर्षों के आँकडों पर भी वैधीकृत किया गया। विकसित मॉडलों का मूल्यांकन मूल माध्य वर्ग त्रुटि** (RMSE), **सामान्यीकृत मूल माध्य वर्ग त्रुटि** (nRMSE), **माध्य निरपेक्ष त्रुटि** (MAE) **और गुणांक निर्धारण** (R^2) **के आधार पर किया गया। प्रायोगिक विश्लेषण से पता चलता है कि उत्तराखंड के उधम सिंह नगर जिले** (**यूएसएन**) **के लिए चावल की फसल की उपज का पूर्वानुमान के लिए कृत्रिम तंत्रिका नेटवर्क** (R^2 = 0.99, RMSE = 0.07, nRMSE = 2.20, MAE = 0.06) **का प्रदर्शन** SMLR (R^2 = 0.97, RMSE = 0.08, nRMSE = 2.34, MAE = 0.05), (LASSO (R^2 = 0.62, RMSE = 0.26, nRMSE = 7.81, MAE = 0.24) **और** ELNET (R^2 = 0.54, RMSE = 0.38, nRMSE = 11.41, MAE = 0.37) **की तुलना में बेहतर है। इसलिए**, **चावल की उपज के पूर्वानुमान के लिए**, **एएनएन तकनीक का उपयोग उत्तराखंड के उधम सिंह नगर जिले के लिए किया जा सकता है।**

ABSTRACT. The study was aimed to develop the yield forecast model for rice crop yield. Four different techniques, *i.e.*, Stepwise Multiple Linear Regression (SMLR), Artificial Neural Network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ELNET) were used to build the prediction models. Dataset of meteorological data and crop yield data of 15 years have been used to develop the forecast models. The developed models were also validated on the dataset of three years. The assessment of the developed models was done by using root mean square error (RMSE), normalized root mean square error (nRMSE), Mean Absolute Error (MAE) and on the basis of coefficient of determination (R^2). The experimental analysis suggested that the performance for Artificial Neural Network (R^2 = 0.99, RMSE = 0.07, nRMSE = 2.20, MAE = 0.06) is better as compared to SMLR (R^2 = 0.97, RMSE = 0.08, nRMSE = 2.34, MAE = 0.05), LASSO (R^2 = 0.62, RMSE = 0.26, nRMSE = 7.81, MAE = 0.24) and ELNET (R^2 = 0.54, RMSE = 0.38, nRMSE = 11.41, MAE = 0.37) for the prediction of rice crop yield for Udham Singh Nagar (USN) district of Uttarakhand. Therefore, for the prediction of rice yield, ANN technique can be well utilised for Udham Singh Nagar district of Uttarakhand.

Keywords – SMLR, Neural networks, LASSO, ELNET, R^2, RMSE.

1. Introduction

India is a country of agriculture having a variety of food grains. Amongst all the crops rice, wheat, sugarcane, maize, and pulses are the main stable crops for major population living in India. Rice crop contributes more than 40% to the overall crop production (Gandhi *et al.*, 2016). India's top rice suppliers and exporter's states are West Bengal, Uttar Pradesh, Andhra Pradesh, Punjab, Tamil Nadu, Uttarakhand, Orissa, Bihar and Chhattisgarh. Uttarakhand state covers 0.3 million ha area under rice cultivation with average productivity of 2 tons/ha. The share of Udham Singh Nagar in the area and production of Uttarakhand state's rice crop is about 38% and 55% respectively.

Due to various human induced activities and other calamities, there is acceleration in climate change over the

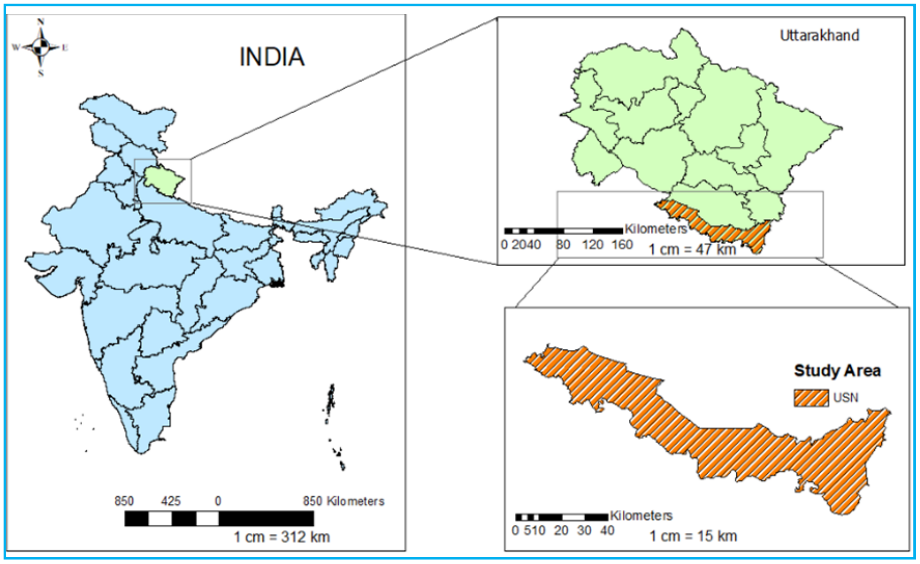

Fig. 1. Study location

last few decades. Besides this, due to the rapid growth of the population, demand of the food is also escalating. Therefore, it is a need to understand the relationships between climatic variability and crop yield. Prediction of crop yield is a very challenging task. The lack of knowledge of experts, negation of personal perception and fatigue, *etc*. can be some issues in prediction of yield. The major concern for agricultural planning purposes is the estimation of exact yield for various crops included in the planning. Such issues can overcome by using the decision tools and models for crop yield prediction. Crop yield prediction is very beneficial for the government in making food policies, market prices and import and export policies. Likewise, industries can benefit from yield prediction by better planning of the logistics of their business. Early prediction of crop yield can be helpful for the farmers in making important decisions related to the crops.

Crop simulation and empirical statistical models were used to give the crop yield forecast. But crop simulation methods do not perform better in case of less amount of data. Thus in this case empirical statistical models can be used as a collective substitute as it requires lesser input data. Therefore empirical statistical models with simple regression techniques including historical yield and weather data are a good alternative of crop simulation model. (Lobell and Burke, 2010; Shi, *et al*., 2013). Useful regression models such as linear regression model to more sophisticated non-linear regression models such as support vector regression, machine learning can be used for yield forecasting purpose. Simple linear regression which is used to model crop yield and climate variables such as temperature and precipitation has a long history (Agrawal, *et al*., 2001, Jaya kumar, *et al*., 2016). Some comparisons among regression models for crop yield prediction have been made, looking for the most accurate technique. Drummond *et al*., (2003) and Fortin *et al*., (2011) compared classical statistical models against Artificial Neural Networks (ANNs). Srivastava *et al*., (2020) also examined the impact of heavy rainfall, rainy days and drought on paddy and soybean crop yield.

In this research apart from stepwise multiple linear regression (SMLR), Artificial neural network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ELNET) technique has been used to give yield prediction of rice crop.

The present research shows the prediction of yield by different methods. Different methods, *viz*., MLR, SVR and ANN, tested and evaluated using appropriate algorithms. A comparison has been made among these methods. Further it is suggested that which method will be ideal to deploy in real world.

2. Data and methodology

For the development of the rice crop yield model meteorological data (rainfall, minimum temperature, maximum temperature, relative humidity and solar radiation) of 18 years, *i.e.*, 2001 - 2018 was taken from the Agrometeorological observatory located at Norman E. Borlaug Crop Research Centre (NEBCRC), G. B. Pant University of Agriculture and Technology, Pantnagar, Udham Singh Nagar (28.96° N, 79.52° E), Uttarakhand,

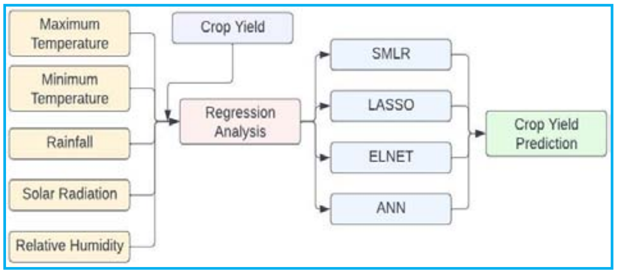

Fig. 2. Flowchart showing the steps involved in model development

India. The historical yield data of the corresponding years has been taken from the Dacnet website. The analysis was performed on the weekly values of the weather parameters.

Four different methods have compared to give the prediction of rice crop yield over the area of study. The models were developed by accumulating 15 years (2001 - 2015) yield and weather data and were validated over the dataset of three years, *i.e.*, 2016 - 2018. For the analysis of dataset, linear regression with step wise technique, artificial neural network with multilayer perceptron (MLP) topology Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ELNET) regression technique have been used. The analysis was performed with the help of R and Statistical Package for Social Sciences (SPSS) software. Significance level of 5% was used to test the developed models. Prior to conducting the regression analysis, the assumptions of normality, homoscedasticity and multicollinearity were tested by using Q-Q scatter plot (Bates *et al.*, 2014; Field, 2013), Residuals scatter plot (Osborne and Walters, 2002) and Durban - Watson test respectively. Figs. 1&2 shows the study location and the steps that involved in model development respectively.

3. Stepwise multiple linear regression

It is the simplest way to develop the statistical models. It provides a way to select the best predictors among all the set of predictors. In this method, at each subsequent step a predictor variable gets added and its significance gets tested. Therefore, the SMLR is generally used when there is need to select the best predictors among the set of predictors. (Singh *et al.*, 2014; Das *et al.*, 2018).

4. Artificial neural network

Artificial Neural Networks are fully connected network that is organized into layers. ANNs usually consist of one input layer, one or multiple hidden layers and one output layer. In the present study three layers have been used, *viz.*, one input, one hidden and one output layer. Neurons of each layer are interconnected with the neuron of the next layer. The number of neurons in input layer is depending upon the predictor variables in the dataset. The major issue in the implementation of artificial neural network is to estimate the optimum number of hidden neurons. To find out the optimum number of hidden neurons, we have implemented the 'nnet' method with 10 - fold cross - validation and used the 'train' function of the 'caret' package in R software (Kuhn 2008, Das *et al.*, 2018).

5. Least absolute shrinkage and selection operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression analysis technique that is used to select the important predictors among a large set of predictors and reduces the coefficients of others predictors to zero. LASSO technique has two parameters, *i.e.*, lambda and alpha that must be tuned to prevent over fitting. The optimal value of lambda was evaluated by minimizing the error using cross-validation (Piaskowski *et al.*, 2016; Das *et al.*, 2020) while the value of alpha was considered as 1.

6. Elastic net (ELNET)

Elastic Net is a regression analysis technique that is used when there is a high correlation among the variables of the dataset. It helps to overcome the weakness of LASSO as well as Ridge regression, *i.e.*, the ELNET technique provides a way to select the best predictors by minimizing the errors. The elastic net consider both L1 AND L2 penalties to give the best prediction (Abbas *et. al.*, 2020).

7. Model performance evaluation

Finally the performance of the developed models was evaluated on the basis of root mean square error (RMSE), normalized root mean square error (nRMSE), coefficient of determination (R^2) and modeling efficiency (EF).

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(A_i - P_i\right)^2}{n}}$$

$$n\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(A_i - P_i\right)^2}{n}} \times \frac{100}{\bar{A}}$$

TABLE 1

Actual and predicted yield of rice by four different techniques

Year	Actual Yield	Predicted Yield				Percentage Error			
		SMLR	ANN	LASSO	ELNET	SMLR	ANN	LASSO	ELNET
2016	3.28	3.25	3.26	3.18	3.02	0.91	0.68	2.91	8.03
2017	3.54	3.53	3.42	3.27	3.04	0.00	3.47	8.20	14.14
2018	3.22	3.09	3.19	2.86	2.88	4.10	0.88	12.39	10.67

$$R^2 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}\left(A_i - \overline{A}\right)\left(P_i - \overline{P}\right)}{\sigma_A \sigma_P} \right)^2$$

$$\text{EF} = \left(1 - \frac{\sum_{i=1}^{n}\left(A_i - P_i\right)^2}{\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2} \right)$$

Here, A_i is the observations corresponding to actual Yield, P_i is the observations corresponding to predicted Yield, $\overline{A}$ is the average of the observations corresponding to actual Yield, $\overline{P}$ is the average of the observations corresponding to predicted Yield, σ_A is the standard deviation of the Actual Yield and σ_p is the standard deviation of the Predicted Yield. The developed model is considered as excellent, good, fair and poor depend upon the values of nRMSE lies in the range of <10%, 10-20%, 20-30% and >30% respectively (Jamieson *et al.*, 1991; Sridhara *et al.*, 2020). On the other hand, the value of EF ranges between -∞ to 1. EF values near to 1 shows good model predictions whereas the value of EF equal to zero indicates that the model does not predict better than the average of the observed values (Therond *et al.*, 2011, Sridhara *et al.*, 2020). The value of RMSE close to 0 and the values of R^2 close to 1 show good model performance.

8. Results and discussion

Linear regression analysis with stepwise technique, Artificial Neural Network (ANN), Elastic Net (ELNET) and Least Absolute Shrinkage and Selection Operator (LASSO) were used to check that the weather parameter (*T*max, *T*min, Rainfall, Relative Humidity and Solar Radiation) significantly predicted the yield of the rice crop of Udham Singh Nagar district.

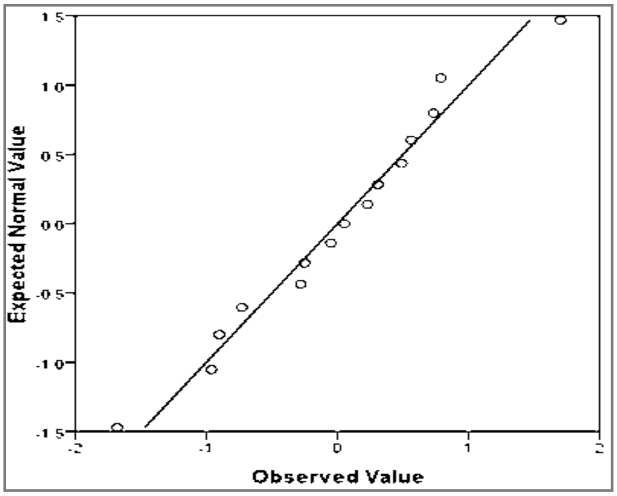

Fig. 3. Q-Q scatterplot testing normality (USN)

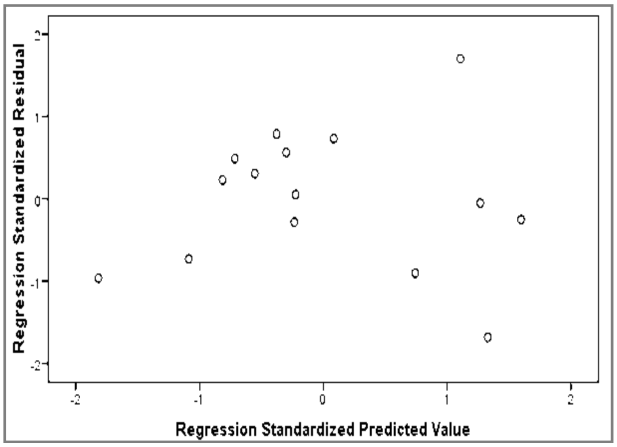

Fig. 4. Residuals scatter plot testing homoscedasticity (USN)

Before conducting the regression analysis, the assumptions of normality of residuals, homoscedasticity of residuals and multicollinearity were examined by using Q-Q scatter plot, Residuals Scatter plot and Durbin-Watson test. It can be observed that the points in the Q-Q scatter plot (Fig. 3) form a straight-line, thus fulfill the

TABLE 2

Model performance at calibration stage

	R^2_{cal}	RMSE$_{cal}$	*n*RMSE$_{cal}$	MAE$_{cal}$	EF
SMLR	0.88	0.09	3.21	0.08	0.86
ANN	1.00	0.02	0.64	0.02	0.99
LASSO	0.99	0.04	1.45	0.03	0.97
ELNET	0.97	0.12	4.03	0.10	0.78

TABLE 3

Model performance at validation stage

	R^2_{cal}	RMSE$_{cal}$	*n*RMSE$_{cal}$	MAE$_{cal}$
SMLR	0.97	0.08	2.34	0.05
ANN	0.99	0.07	2.20	0.06
LASSO	0.62	0.26	7.81	0.24
ELNET	0.54	0.38	11.41	0.37

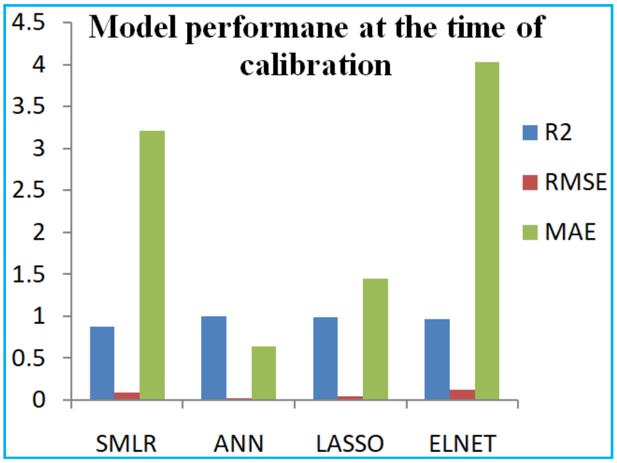

Fig. 5. Performance of calibration of SMLR, ANN, LASSO and ELNET in terms of R^2, RMSE and MAE

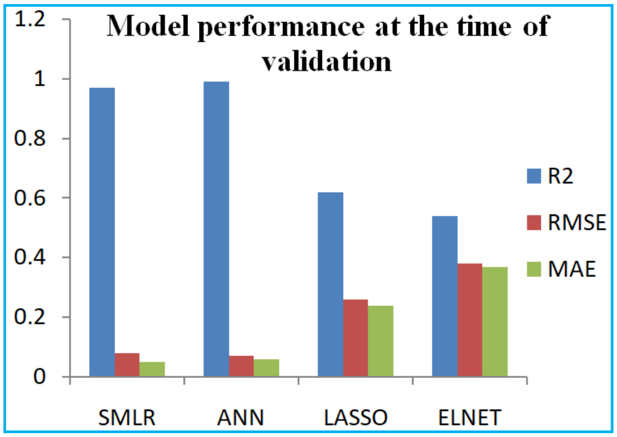

Fig. 6. Performance of validation of SMLR, ANN, LASSO and ELNET in terms of R2, RMSE and MAE

condition of Normality. The points in the Residuals scatter plot (Fig. 4) were randomly distributed around the mean therefore fulfill the condition of homoscedasticity. The calculated value of Durbin-Watson test statistic was for the dataset was 1.60, indicated the absence of multicollinearity.

After checking the assumptions, the model was developed and then validated. For the validation purpose the yield and weather dataset of three years, *i.e.*, 2016, 2017 and 2018 has been used. The predicted yield for these three years from the developed model by using four different methods is shown in Table 1. The performance of the models was measured by using R^2, RMSE, *n*RMSE, MAE and EF (Tables 2&3). All the techniques were performed well to give the crop yield prediction, though the analysis also revealed that ANN technique provided the best results in case of calibration ($R^2_{cal} = 1.00$, $\text{RMSE}_{cal} = 0.02$, $n\text{RMSE}_{cal} = 0.64$, $\text{MAE}_{cal} = 0.02$) as well as at the time of validation ($R^2_{val} = 0.99$, $\text{RMSE}_{val} = 0.07$, $n\text{RMSE}_{val} = 2.20$, $\text{MAE}_{val} = 0.06$) of the model. It can also be seen that

SMLR technique analysis had less value of RMSE, *n*RMSE and MAE in both calibration and validation stage followed by ANN. The performance of ELNET was not found to be worse as compare to the rest of the techniques. Overall, the ranking given to the models based on the different statistical techniques (R^2, RMSE, *n*RMSE and MAE) were ANN > SMLR > LASSO > ELNET. Figs. 5&6 shows the performance of the models at the time of calibration and validation in terms of R^2, RMSE and MAE.

9. Conclusions

In the current study four methods, *viz*., SMLR, ANN, LASSO and ELNET have been used to compare the prediction of rice crop yield at Udham Singh Nagar district. The study concluded that all the four methods performed well though the prediction done by using ANN was more accurate as compared to the rest of the three techniques. Therefore model developed by ANN technique can be very well utilized to give the prediction of rice crop yield of Udham Singh Nagar district of Uttarakhand.

Acknowledgments

The authors acknowledge the India Meteorological Department (IMD) for the financial support given by the organization through FASAL project.

Disclaimer : The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Abbas, F., Afzaal, H., Farooque, A. A. and Tang, S., 2020, "Crop yield prediction through proximal sensing and machine learning algorithms", *Agronomy*, **10**, 7, 1046.

Agrawal, R., Jain, R. C. and Mehta, S. C., 2001, "Yield forecast based on weather variables and agricultural input on agroclimatic zone basis", *Ind. J. Agric. Sci*., **71**, 487-490.

Bates, D., Machler, M., Bolker, B. and Walker, S., 2014, "Fitting linear mixed-effects models using lme4", arXiv preprint arXiv: 1406.5823.

Das, B., Nair, B., Arunachalam, V., Reddy, K. V., Venkatesh, P., Chakraborty, D. and Desai, S., 2020, "Comparative evaluation of linear and nonlinear weather-based models for coconut yield prediction in the west coast of India", *International Journal of Biometeorology*, 1-13.

Das, B., Nair, B., Reddy, V. K. and 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.

Drummond, S. T., Sudduth, K. A., Joshi, A., Birrell, S. J. and Kitchen, N. R., 2003, "Statistical and neural methods for site-specific yield prediction", *Transactions of the ASAE*, **46**, 1, 5.

Field, A., 2013, "Discovering statistics using SPSS (4th ed.). Thousand Oaks, CA: Sage.

Fortin, J. G., Anctil, F., Parent, L. É. and Bolinder, M. A., 2011, "Site-specific early season potato yield forecast by neural network in Eastern Canada", *Precision agriculture*, **12**, 905-923.

Gandhi N., Petkar O. and Armstrong L. J., 2016, "Rice crop yield prediction using artificial neural networks", *IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Chennai, 105-110.

Jamieson, P. D. Porter, J.R. and Wilson, D. R., 1991, "A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand", *Field Crop. Res*., **27**, 337-350.

Jayakumar, M., Rajavel, M. and Surendran, U., 2016, "Climate-based statistical regression models for crop yield forecasting of coffee in humid tropical Kerala, India", *International Journal of Biometeorology*, **60**, 1943-1952.

Kuhn M., 2008, "Building predictive models in R using caret package", *J. Stat. Softw.*, **28**, 1-26.

Lobell D. B. and Burke M. B., 2010, "On the use of statistical models to predict crop yield responses to climate change", *Agric. For Meteorol.*, **150**, 1443-1452.

Osborne, J. and Waters, E., 2002, "Four assumptions of multiple regression that researchers should always test. Practical Assessment", *Research & Evaluation*, **8**, 2, 1-9.

Piaskowski J. L., Brown D. and Campbell K. G., 2016, "Near-infrared calibration of soluble stem carbohydrates for predicting drought tolerance inspring wheat", *Agron. J.*, **108**, 285-293.

Shi, W., Tao, F. and Zhang, Z., 2013, "A review on statistical models for identifying climate contributions to crop yields", *J. Geogr. Sci.*, **23**, 567-576.

Singh, R. S., Patel, C., Yadav, M. K. and Singh, K. K., 2014, "Yield forecasting of rice and wheat crops for eastern Uttar Pradesh", *J. Agrometeorol*., **16**, 199-202.

Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S. D., Sanjivaiah, S. H. and 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.

Srivastava A. K., Yogranjan and Bal L. M., 2020, "Variability of extreme weather events and its impact on crop yield in Bundelkhand Agroclimatic zone of Madhya Pradesh", *MAUSAM*, **71**, 2, 275-284.

Therond, O., Hengsdijk, H., Casellas, E., Wallach, D., Adam, M., Belhouchette, H., Oomen, R., Russell, G., Ewert, F., Bergez, J.E., Janssen S., Wery J. and Ittersum, M. K. V., 2011, "Using a cropping system model at regional scale: Low-data approaches for crop management information and model calibration", *Agric. Ecosyst. Environ*., **142**, 85-94.