Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February, 2025; pp 167-176

Predictive Modeling of Maize Yield in Jammu Subtropical Zone using Weather Data and Penalized Regression

Manish Sharma[1], Amandeep Verma[1], Nishant Jasrotia[1], Sushil Sharma[2] and Divyam Sharma[1]

[1]*Division of Statistics and Computer Science, FBSc, SKUAST-Jammu*
[2]*Faculty of Agricultural Engineering, SKUAST-Jammu*

Received: 16 June 2025; Revised: 19 August 2025; Accepted: 21 August 2025

Abstract

The present study employed regression techniques such as Ordinary Least Squares (OLS), Ridge, and Lasso regression to analyse maize production data in relation to weather parameters from the subtropical region of Jammu. The models were evaluated across various training and testing splits based on performance metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Among the models, the Lasso regression demonstrated the best overall performance with the lowest AIC (47.37) and BIC (50.76) values, indicating superior model fit and simplicity. The optimal regularisation parameter ($\alpha = 0.126$) and having minimum MSE, thus ensuring a well-balanced trade-off between model complexity and predictive accuracy. The Lasso regression model successfully identified key predictors influencing maize production, with maximum temperature and area being the most influential variables, followed by sunshine hours and relative humidity in the evening. Rainfall and minimum temperature were found to have minimal or no impact. Therefore, the proposed Lasso regression model, with its optimal alpha value and refined feature selection, serves as a robust and interpretable tool for predicting maize production in the subtropical region of Jammu.

Key words: Maize; Regularizations; Penalized regression; Weather; MSE.

1. Introduction

Statistical modelling tools can sometimes produce suitable models quite fast. The researchers used statistical models as baseline predictive models to assess the performance of advanced methods, even for situations where more adaptable machine-learning techniques (such regularization techniques and neural networks) can ultimately produce better results. The term regression was first introduced by a British anthropologist and meteorologist Sir Francis Galton (1886) in a paper entitled "Regression towards Mediocrity in Hereditary Stature". Since the regression analysis has emerged as a powerful statistical tool. With

Corresponding Author: Manish Sharma
Email: manshstat@gmail.com

the enhanced application of statistics in economics, industry, agriculture, social sciences, biology, medical sciences, psychology and education. According to Tibshirani (1996), the ordinary least square (OLS) estimates of the regression parameter in general have low bias but large variance and often have poor performance in both prediction and interpretation when the assumptions violated such as residual normality, homoscedasticity, independence, and linearity and the estimates may become skewed or undependable, especially when there are outliers and many correlations between the predictor variables. The prediction accuracy can sometimes be improved either by shrinking of some coefficient towards zero or by allowing a little bias to reduce the variance of the parameter estimates and predicted value. The outcomes of a linear regression can be severely distorted by outliers, which can result in incorrect interpretations and poor prediction accuracy. High correlation across predictor variables can lead to multicollinearity, which alters the variance of coefficient estimates and makes it challenging to isolate the impact of each predictor on the response variable. These problems may make yield projections less reliable and consider for exploring different modeling approaches. Penalized regression models like lasso and ridge regression have become effective alternatives for ordinary linear regression. By include a penalty term in the loss function, these models introduce regularization approaches that assist reduce the impact of outliers and multicollinearity. Through the efficient selection of important characteristics and the reduction of overfitting, this method not only increases model stability but also improves predictive performance.

One of the most adaptable developing crops, maize (*Zea mays L.*), has a wide range of modification under various agro-climatic situations. Maize has the largest genetic yield potential of all the cereals, it is referred to as the ”Queen of cereals” worldwide. It is grown on roughly 150 million ha in about 160 nations, where there is a greater variety of soil, climate, biodiversity, and management techniques. This increases global grain production by 36 percent (782 million tonnes) Kiran *et al.* (2018).The maize crop is susceptible to the fluctuations of rainfall distribution as a whole. In Jammu and Kashmir, maize is widely planted in the kandi, karewa, and plain regions. It does well in loamy to sandy loam soils. Additionally, maize varieties that thrive in colder hilly and mountainous regions have been produced. In all such areas where the summer is long enough to support its cultivation and where frost does not arrive too early, it can be grown. When it is growing and developing, it needs a temperature of around 30°C, and when it is ripening, it needs a temperature of at least 20°C. A fertile, deeply tilled soil is necessary for maize. The soil is prepared in advance of the sowing season, which is April to May on the Jammu plain and May to June in the Kashmir valley, the kandi, and the state’s mountainous regions. There are ten districts in the Jammu area of the UT of Jammu and Kashmir, maize is grown in almost all the districts of the Jammu region. In terms of production, the districts of Rajouri (110.20 thousand MT), Udhampur (70.11 thousand MT), and Poonch (69.59 thousand MT) have the largest concentration of maize. (Digest of Statistics, 2023-24). Jorvekar *et al.* (2024) conducted a study to evaluate and compare the performance of different regression models for agriculture crop yield prediction on the basis historical crop yield data, weather parameters and pesticides data features from various agricultural regions. Various regression models, including Linear Regression (LR), K Nearest neighbor Regression (KNR), Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Gradient Boosting Regression (GBR),Linear Model Lasso Regression, Elastic net Regression, Ridge Regression to predict crop yields for various crops. This study involved evaluating the performance of these regression models based on several performance.Krishnadoss and Ramasmy (2024) used various machine learning models for crop yield prediction to make dynamic pre monsoon decisions. The input variables precipitation, temperature, evaporation, wind speed, and chemical use influence crop yield estimations. Jasrotia *et al.* forecast the production of walnut for Jammu and Kashmir using Time series models. Holt linear exponential Smoothing and Autoregressive Integrated Moving Average (ARIMA) model have been applied and it shows that ARIMA(1,2,1) is appropriate model for forecasting on the basis of minimum value of information criterion as compared to other models. Based on the forecast provided by the proposed model, there is a projected 56.69 percent increase in walnut production for the year 2035 with respect to 2022. Gupta *et al.* compared classification techniques through statistical as well as artificial neural network models for the primary data related to 140 rice genotypes from the trial laid in SKUAST, Jammu on the basis of maturity. And, the characters like yield per plant, number of days for 50 percent flowering, number of days for full flowering, plant height, number of effective tillers per plant, panicle length, grain length, grain width and ratio of grain length and grain width acts as supporting variables for classification.

2. Material and methods

The study was undertaken based on secondary data related to area and production of maize from three decades with effect from 1992 to 2023, 31 years of subtropical region of Jammu.To assess the performance of the model, different proportions of training and testing data were utilized related to 31 years of data of different parameters. The data pertains to weather-based components such as; Maximum temperature, minimum temperature, Rainfall, Relative humidity morning, Relative humidity evening and sunshine (hrs.) were also collected from digest of statistics published by Directorate of Economics and Statistics, UT Administration of Jammu and Kashmir and Agro-metrological unit of SKUAST-Jammu. In order to estimate the relation and prediction of maize production of Jammu division through weather parameters the following statistical models are applied for handling the problem of overfitting and challenges of influential observations in the data.

2.1. The least square method

The least square method used in regression is relatively straightforward. Imagine a scatterplot of data points that form a linear trend. An OLS linear regression procedure builds a line of best fit that would serve as the most accurate way of depicting the spread of the data points with a single line. The least squares property states that the line fit in the OLS method will have the smallest value of the summed squared deviations of each data point from the line.

2.2. Penalized regression models

The two most common techniques are ridge regression given by Hoerl and Kennard (1970). and lasso regression Tibshirani (1996). The predictor variables are all kept in the model through penalized regression techniques, but the regression coefficients are regularized by reducing them to zero or a value equal to zero. Shrinkage or regularization methods are other names for penalized regression.

(i) Ridge regression model

Ridge regression shrinks the regression coefficients, so that variables, with a minor contribution to the outcome, have their coefficients close to zero. The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called L2-norm, which is the sum of the squared coefficients. Mathematical form of ridge regression model is:

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij}\right)^2+\alpha 2\|\beta_j\|\right\}$$

where y is the dependent variable, x is the covariate, β is the corresponding coefficient, and α_2 represents the L2 norm penalty. The amount of the penalty can be fine-tuned using a constant called Alpha (α). Selecting a good value for α is critical. When $\alpha = 0$, the penalty term has no effect, and ridge regression will produce the classical least square coefficients. However, as α increases to infinity, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get close to zero.

(ii) Least absolute shrinkage selection operator (LASSO)

It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called L1-norm, which is the sum of the absolute coefficients. In the case of lasso regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero. This means that lasso can be also seen as an alternative to the subset selection methods for performing variable selection in order to reduce the complexity of the model. As in ridge regression, selecting a good value of α for the lasso is critical. One obvious advantage of lasso regression over ridge regression is that it produces simpler and more interpretable models that incorporate only a reduced set of predictors. The mathematical model of the of the LASSO Regression is

$$\min\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{i=1}^{p}\beta_j x_{ij}\right)^2+\alpha_1\|\beta\|\right\}$$

where y is the dependent variable, x is the covariate, β is the corresponding coefficient, and α_1 represents the L1 norm penalty. However, neither ridge regression nor the lasso will universally dominate the other. Generally, lasso might perform better in a situation where some of the predictors have large coefficients, and the remaining predictors have very small coefficients. The penalized model is trained using a different subset of the data observations, called the training set and rest is used as testing set. The performance of the proposed model *i.e.* OLS, RR and Lasso checked through the MSE, RMSE, AIC and BIC.

(a) Mean squared error (MSE)

The Mean Squared Error (MSE) measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. The Mean Squared Error is calculated as:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

where, n is sample size, actual is the actual data value and forecast the predicted data value

(b) Root mean squared error (RMSE)

RMSE is employed while assessing models. If we have a sample of n observations $y(Yi, i = 1, 2..., n)$, and n matching model predictions y, then RMSE is defined as the square root of the mean squared error (MSE), which is provided by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(c) Akaike's information criterion (AIC)

An important statistic for identifying and assessing statistical models is AIC. The likelihood function, L, and the number of hyperparameters estimated from the model, n, are used to determine this criterion, which was proposed by Akaike in 1979. Its is calculated as $AIC = -2logL + 2n$ where, L is the value of the likelihood and n is the number of estimated parameters.

(d) Bayesian information criterion (BIC)

The criterion was proposed by Schwarz (1978) using Bayesian likelihood maximization. Schwarz further demonstrated the BIC's validity by showing that it is independent of prior distribution. $SBIC = -2logL + nlogT$, where T is the total number of observations, is the formula for the BIC. The fitted model performs better when these data have a lower value.

3. Results and discussion

In this study, regression techniques like Ordinary Least Squares (OLS), Ridge, and Lasso regression were applied to the data of maize production with whether parameters from the subtropical region of Jammu. The subtropical region of Jammu shows moderate variability in maize cultivation as shown in table 1 with an average area of 21.15 thousand hectares and production of 38.74 thousand metric tonnes. Climatic conditions remain fairly stable, with average minimum and maximum temperatures of 16.32°C and 28.96°C, respectively. Rainfall displays notable variation, averaging 1240.30 mm with a high coefficient of variation (16.16%).

The normality tests conducted for the Jammu (subtropical) region revealed mixed results across variables as shown in table 2. Area and minimum temperature show slight deviations from normality as per Shapiro-Wilk and Anderson-Darling tests. Sunshine hours significantly deviate from normality across all three tests, indicating strong non-normal distribution. Production, rainfall, and relative humidity (morning) largely follow a normal distribution with high p-value. These findings suggest that while most climatic variables are normally distributed, sunshine data requires transformation or non-parametric handling in statistical modeling.

To evaluate the model performance, the dataset was divided into different training

Table 1: Descriptive statistics for subtropical region of Jammu

Variable	Minimum	Maximum	Average	SD	CV (%)
Area (000 hectares)	15.71	25.03	21.15	2.63	12.43
Production (000 MT)	23.70	48.24	38.74	5.32	13.73
Minimum Temperature (°C)	15.15	17.22	16.32	0.57	3.51
Maximum Temperature (°C)	28.00	29.96	28.96	0.58	2.00
Rainfall (mm)	769.14	1730.12	1240.30	200.48	16.16
Sunshine (hrs.)	4.36	6.79	6.14	0.40	6.54
Relative Humidity Morning (%)	74.98	87.74	80.76	2.48	3.08
Relative Humidity Evening (%)	43.05	55.04	49.98	2.81	5.54

Table 2: Test of normality for area, production, and different weather parameters for subtropical region of Jammu using Shapiro-Wilk (S-W), Kolmogorov-Smirnov (K-S), and Anderson-Darling (A-D) tests

Variable	Shapiro-Wilk (S-W)		Kolmogorov-Smirnov (K-S)		Anderson-Darling (A-D)	
	Statistic	p-value	Statistic	p-value	Statistic	p-value
Area (000 hectare)	0.93	0.05	0.76	0.60	0.78*	0.04
Production (000 MT)	0.96	0.41	0.54	0.93	0.31	0.52
Minimum Temperature	0.93*	0.04	0.89	0.41	0.86*	0.02
Maximum Temperature	0.96	0.25	0.77	0.59	0.35	0.46
Total Rainfall (mm)	0.99	0.99	0.36	1.00	0.16	0.94
Relative Humidity (Morning)	0.97	0.54	0.65	0.79	0.37	0.41
Relative Humidity (Evening)	0.93	0.08	1.13	0.16	0.85*	0.02
Sunshine (hrs.)	0.69**	0.00	1.56**	0.01	2.85**	0.00
Sunshine (hrs.)	0.70**	0.00	1.56*	0.01	2.86**	0.00

and testing ratios: 80:20, 70:30, and 60:40. Here, the first value shows the percentage of data used for training the model, and the second value indicates the percentage used for testing. The basis for choosing the optimal ratio is obtaining the lowest MSE and RMSE for testing datasets. These evaluation criteria are essential because they provide insight on the model's accuracy and precision and capacity for generalization

The examination of various training and testing data splits for the OLS regression model determined that the 60:40 ratio is the optimal option for the subtropical region of Jammu, as shown in table 3. This ratio showed the best performance of the model on unknown data, yielding the lowest testing MSE (15.72) and RMSE (3.96) as compared to training MSE (16.36) and RMSE (4.04). The 80:20 ratio had the greatest testing MSE (21.43) and RMSE (4.63) than training MSE (13.87) and RMSE (3.72), indicating overfitting even though it displayed the lowest training error. When it came to testing MSE and RMSE, the 60:40 ratio performed better than the 70:30 ratio having testing MSE(17.45) and RMSE (4.18). Therefore, the 60:40 ratio is selected for the model and optimal option for the RR model in the subtropical area of Jammu, according to the criterion of choosing the ratio with the smallest testing MSE and RMSE. Its training MSE (16.38) and RMSE (4.05) and testing MSE (15.61) and RMSE (3.95) were the lowest, demonstrating the best model performance and generalization. The 80:20 ratio exhibited the highest testing MSE (19.29) and RMSE (4.39), indicating overfitting, despite having the lowest training MSE (14.54) and RMSE (3.81). The 70:30 ratios show overfitting having lower training MSE (14.75), RMSE (3.84) and higher testing MSE (16.84) and RMSE (4.10). The selected ratio is the best option for reducing errors on the testing dataset and promising improved generalization because

Table 3: The MSE, RMSE, AIC and BIC of regression models for subtropical region of Jammu w.r.t different ratios for training and testing datasets

Model	Statistics	60:40		70:30		80:20	
		Training	Testing	Training	Testing	Training	Testing
OLS	MSE	16.36	15.72	14.58	17.45	13.87	21.43
	RMSE	4.04	3.96	3.82	4.18	3.72	4.63
	AIC	64.31	49.81	70.27	42.60	77.11	35.45
	BIC	70.54	53.77	77.58	44.71	85.36	35.08
Ridge regression	MSE	16.38	15.61	14.75	16.84	14.54	19.29
	RMSE	4.05	3.95	3.84	4.10	3.81	4.39
	AIC	66.34	51.72	72.52	44.23	80.25	36.72
	BIC	73.46	56.24	80.77	46.66	89.67	36.28
Lasso regression	MSE	16.67	15.19	15.09	16.54	14.57	18.20
	RMSE	4.08	3.90	3.89	4.07	3.82	4.27
	AIC	62.65	47.37	70.99	42.06	76.30	32.31
	BIC	67.99	50.76	78.30	44.18	83.37	31.98

it shows a good trade-off between training and testing errors. The selcted ratio is the best option for the lasso regression model based on the criterion of choosing the ratio with the smallest testing MSE and RMSE for subtropical region of Jammu. Its testing MSE (15.19) and RMSE (3.90) were the lowest, in comparison to training MSE (16.67) and RMSE (4.08) which demonstrating the best model performance and generalization. The testing MSE (18.20) and RMSE (4.27) of the 80:20 ratio was reasonably high than training MSE (14.57) and RMSE (3.82), and showing overfitting. The 70:30 ratio performed poorer than the 60:40 ratio as it also shows overfitting having lower training MSE (15.09) and RMSE (3.89) and higher testing MSE (16.54) and RMSE (4.07). As a result, the common ratio selected was the best option for minimizing testing MSE and RMSE, and indicating strong generalisation with reduced overfitting for all the models.

AIC, BIC, and coefficient values based on the OLS, RR, and LR models analysis utilizing the selected ratio for the subtropical region of Jammu are shown in table 4, With an AIC (49.81) and a BIC (53.77) for OLS regression, 51.72 and a BIC (56.24) for RR, and the lowest AIC (47.37) and a BIC (50.76) for lasso regression, these results were obtained. Lasso showed the smallest testing AIC and BIC, suggesting superior model performance and generalization, the LR model is chosen as the best-performing model for predicting maize production in the subtropical region.

The variation in the coefficients of several meteorological variables with different log(alpha) values in a Lasso regression model is illustrated in fig. 1. The coefficients are displayed on the y-axis, and log(alpha) values are represented on the x-axis. The significance of each meteorological variable in the model is indicated by the line that corresponds to it. The optimal alpha value (0.126), is selected to minimize the mean squared error, which can be observed by the dashed line. The coefficients for every variable are visible at this ideal alpha, emphasizing their relative significance. The most important variables are found using this optimum alpha value in the Lasso regression, producing a more precise and understandable model.

The relationship between the log of the regularization parameter (alpha) and the mean

Table 4: Regression coefficient estimates of different parameters for subtropical region of Jammu with respect to OLS, ridge and lasso regression models

Parameters	OLS	Ridge regression	Lasso regression
Intercept	-143.53	-90.48	-51.66
Area (X1)	1.02*	1.01*	1.03*
Max temp. (X2)	3.92*	3.46*	2.58*
Min temp. (X3)	0.18	0.03	0
Rainfall (X4)	0.00	0.00	0.0001*
RH (morning) (X5)	-0.24	-0.30	-0.35
RH (evening) (X6)	0.36*	0.35*	0.29*
Sunshine (hrs.) (X7)	1.70*	1.35*	0.30*
R^2	0.70		
AIC	49.81	51.72	47.37
BIC	53.77	56.24	50.76

squared error (MSE) in a lasso regression model is depicted in fig. 2. The x-axis represents log(alpha) values, and the y-axis shows the MSE. The dotted line indicates how MSE varies with different log(alpha) values, and the red dot marks the best alpha value (0.126), which minimizes the MSE. At this optimal alpha, the model achieves the best balance between bias and variance, resulting in the lowest prediction error. Thus, model is proposed at best alpha (0.126) and Lasso model describes the effect of various features on the target variable is shown in the fig. 3. The coefficients for each attribute are shown by the bars, and larger values denote greater significance. Maximum temperature (2.58*) is the most significant predictor, with the largest positive coefficient; area (1.03*) comes in second. While relative humidity morning (0.35) has a slight negative influence, features like sunshine (0.30*) and relative humidity evening (0.29*) have moderately good effects. The coefficients for rainfall (0.0001*) are almost zero, suggesting that they have little effect. whereas, coefficient for minimum temperature is reduced to zero indicating lesser effect on production than other variables. The optimum alpha value, which minimizes prediction error while setting some coefficients to zero, strikes a compromise between model complexity and accuracy when determining these coefficients.

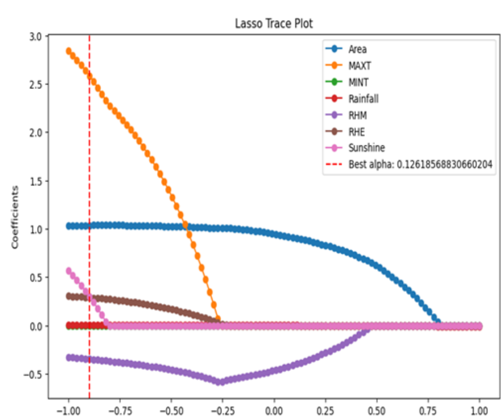

Figure 1: The curve for lasso trace of coefficients w.r.t log alpha

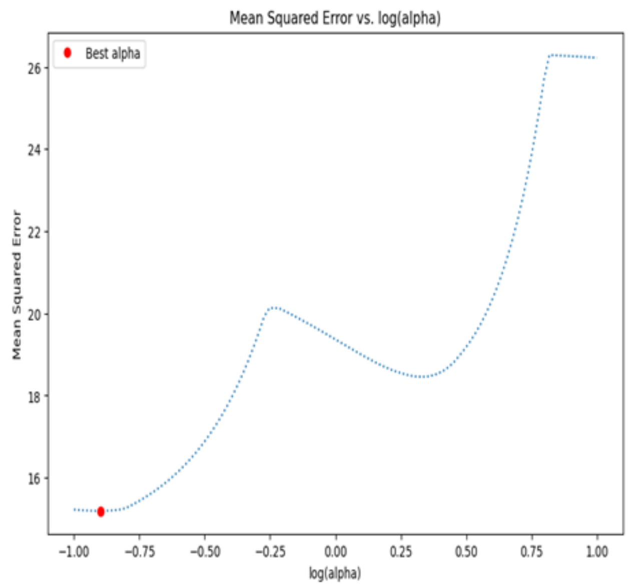

Figure 2: Plot between mean squared error and log alpha

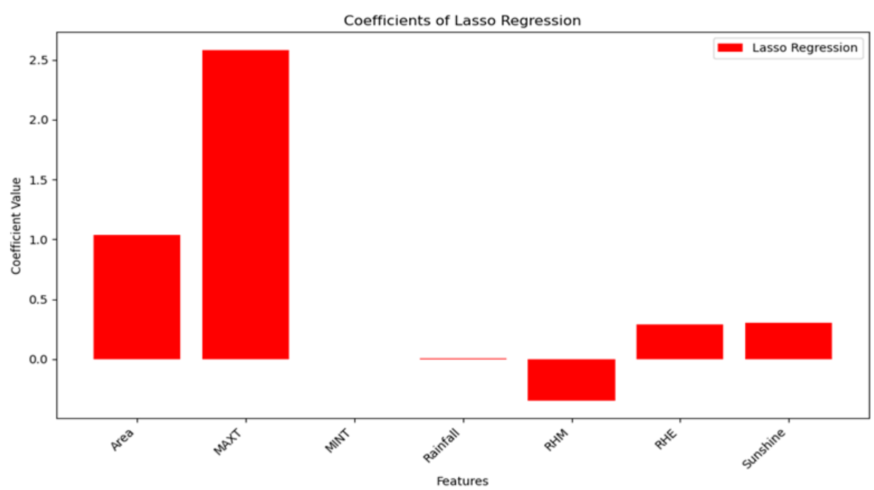

Figure 3: Plot for coefficients of lasso regression

4. Conclusion

Lasso regression outperformed OLS and Ridge models, offering the best predictive accuracy and simplicity, with the lowest AIC (47.37) and BIC (50.76) values. Key predictors identified by the Lasso model were maximum temperature and area, followed by sunshine hours and evening relative humidity; rainfall and minimum temperature had negligible influence. The model offers a reliable and interpretable framework for forecasting maize production, supporting data-driven agricultural planning in the subtropical region of Jammu.

Thus, proposed penalized (Lasso) regression model for maize production prediction is:

$$y = -51.66 + 1.03 \cdot \text{Area} + 2.58 \cdot \text{MAXT} + 0.29 \cdot \text{RH(Evening)} + 0.30 \cdot \text{Sunshine (hrs)}. \quad (1)$$

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Akaike, H. (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, **66**, 237-242.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246-263.

Gupta, S., Sharma, M., Jasrotia, N. and Mahajan, S. (2025). Comparative analysis for classification of rice genotype using statistical and artificial neural network models on the basis of maturity. *Model Assisted Statistics and Applications*, 19, 342-349.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, **12**, 55-67.

Jasrotia, N., Sharma, M., Bhat, A., Mahajan, S., & Gupta, S. (2024). Forecasting models for the production of walnut in Jammu and Kashmir - A comparative study. *Special Proceedings of the 26th Annual Conference, Society of statistics and Computer Application*, 51–60.

Jorvekar., P. P., Wagh., S. K., and Prasad., J. R. (2024). Predictive modeling of crop yields: a comparative analysis of regression techniques for agricultural yield prediction. *Agricultural Engineering International: CIGR Journal*, **26**, 125-140.

Krishnadoss. N. and Ramasamy. L. K (2024). Crop yield prediction with environmental and chemical variables using optimized ensemble predictive model in machine learning. *Environmental Research Communications*, **6**, 1-11.

Kiran, A. S. Shashi, Umesh, K. B., and Shankara, M. H. (2018). Growth and instability in agriculture - A case of maize production in India. In *Conference, July 28-August 2, 2018, Vancouver, British Columbia 277535, International Association of Agricultural Economists*.

Kumar, S., Attri, S. D., and Singh, K. K. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *Journal of Agrometeorology*, **21**, 188-192.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**, 267-288.