Vol. 25, No. 2, pp. 164-174 (2025)
Journal of Agricultural Physics
ISSN 0973-032X
https://doi.org/10.56093/JAP.v25i2.6

Research Article

Wheat Yield Estimation at Different Growth Stage using Multiple Linear, LASSO, Elastic Net and Machine Learning Approach

ANANTA VASHISTH,* ARAVIND K.S., SHRIYA GUPTA, CHANDRAKANT RAJ H., P. KIRSHANAN AND MONIKA KUNDU

Division of Agricultural Physics, ICAR-Indian Agricultural Research Institute, New Delhi-110012

ABSTRACT

Wheat yield estimation was done at different growth stage of the crop. Wheat yield and daily weather data during wheat crop growing period 46^{th} to 15^{th} standard meteorological week were collected from IARI, New Delhi for more than 35 years. Wheat yield estimation was done at tillering, flowering and grain filling stage of the crop by considering weather variables from 46^{th} to 4^{th}, 46^{th} to 8^{th} and 46^{th} to 11^{th} standard meteorological week. Model was developed by stepwise multi linear regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net, support vector regression (SVR), artificial neural network (ANN), PCA-SMLR, PCA-ANN, SMLR-SVR, LASSO-SVR techniques. Analysis was carried out by fixing 70% of the data for calibration and remaining dataset for validation. R statistical software version 3.1.3 was used for developing wheat yield estimation models and making a comparison between the developed models. Results showed that percentage deviation of estimated yield by observed yield was ranged between -7.41 to 4.94, -6.98 to 9.07, -2.37 to 6.66% during tillering, flowering, and grain filling stage respectively. On the basis of percentage deviation and model accuracy LASSO, LASSO-SVR and SVR model was found better having nRMSE value less than 10% hence can be used for district level wheat crop yield estimation at different crop growth stage.

Key words: *S*tepwise multiple linear regression, Least absolute shrinkage and selection operator, Elastic net, Support vector regression, Artificial neural network, Yield estimation

Introduction

Wheat is the most important cereal crop grown during *Rabi* season in north-west part of the India. Weather parameters like maximum temperature, minimum temperature, relative humidity, rainfall, etc. have a great impact on crop yield. Weather variables affect the crop during different stages of development. Thus, the extent of the weather influence on crop yield depends not only on the magnitude of weather variables but also on weather distribution pattern over the full crop season. Hence, estimating crop yield using weather variables is foremost important. Accurate information about history of weather variables and crop yield is useful to take decisions related to agricultural risk management and future predictions.

Yield estimation is crucial for ensuring food security, planning agricultural policies, managing resources like irrigation and fertilizers and supporting farmers with actionable insights. Many techniques have been developed to predict crop production. In traditional methods, crop cutting experiments were widely used for crop yield prediction at different regions. Regularization and feature selection techniques enhance the prediction accuracy and prevent statistical over fitting in a predictive model.

*Corresponding author,
Email: ananta.iari@gmail.com

To overcome the problems of predicting non-linear and non-stationary time series dataset machine learning techniques has been used. Multiple linear regression has the biggest disadvantage of over-fitting when the number of samples is less than the number of variables. Also, another disadvantage is the multi-collinearity when independent predictors are correlated (Verma *et al*., 2016). Kumar et al. (1999) developed stepwise regression technique to predict pigeon pea yield in Varanasi district using different weather variables, appropriate weighted and un-weighted weather indices. SMLR used for pre-harvest wheat crop yield estimation because of its more consistent performance and applicability at zone or state level (Garde *et al*., 2015). Feature selection helps to attain selection of best regression variables and thereby good interpretable results among independent variables (Singh *et al*., 2014). Vashisth *et al*. (2014) reported that percentage deviation of observed yield by estimated yield done at forty-five days before harvest by weather based statistical model was found to be 10.7, 5.7 and 8.53 respectively during the period of 2011-12, 2012-13 and 2013-14. Similarly, the percentage deviation of yield prediction done at 25 days before harvest by weather based statistical model was 9.7, 7.0 and 8.29 respectively.

Azfar (2015) showed the effectiveness of PCA considering all weather indices including interaction indices as regressors was best reliable forecast model for mustard and rapeseed compared to other models. Vashisth and Aravind (2020) reported that Elastic Net, LASSO and SMLR model based on weather parameters can be used for multistage mustard yield estimation and Elastic Net performed best among all the three-model followed by LASSO and SMLR model. White (2009) used correlation analysis or linear regression to study the impact of meteorological factors on crop yields, but these approaches have limitations such as multicollinearity due to the nonlinear nature of biological responses to the environment.

Gupta *et al*. (2024) observed that based on model accuracy parameters RMSE, nRMSE and MSE value, LASSO models were found to be excellent in predicting wheat yield for study areas. The model performance of SVR is increased by the hybrid machine learning approach. The hybrid machine learning LASSO-SVR had more improvement in SVR compared with hybrid machine learning SMLR-SVR. Aravind *et al*. (2022) reported that elastic Net and LASSO was found to be the best model followed by PCA-SMLR, SMLR, PCA-ANN and ANN respectively for wheat yield prediction of different location of north-west India.

In the present study multistage wheat yield estimation model was developed by stepwise multiple linear regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net, support vector regression (SVR), artificial neural network (ANN), and hybrid (PCA-SMLR, PCA-ANN, SMLR-SVR, LASSO-SVR) techniques for improving the accuracy of multi stage wheat yield estimation.

Materials and Methods

Daily weather data (maximum and minimum temperature, morning and evening relative humidity, rainfall) during wheat crop growing period 46^{th} to 15^{th} standard meteorological week (SMW) and wheat yield data for last 35 years were collected from IARI, New Delhi. Weather data were arranged in different standard meteorological weeks for three different stages viz. tillering (46^{th} to 4^{th} SMW), flowering (46^{th} to 8^{th} SMW) and grain filling (46^{th} to 11^{th} SMW) stage separately. Simple and weighted composite weather indices were developed from the combination of weather variables (Vashisth and Aravind, 2020). 70% of data were used for model calibration and remaining 30% were used for the validation of models. Model for estimating the wheat yield estimation was developed at different growth stage by stepwise multiple linear regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net, support vector regression (SVR), artificial neural network (ANN), and hybrid (PCA-SMLR, PCA-ANN, SMLR-SVR, LASSO-SVR) techniques.

Stepwise multi linear regression (SMLR)

The stepwise multilinear regression (SMLR) technique utilizes a combination of forward and backward selection techniques, with variable selection conducted through an automated process.

The significance of all predictor variables in the model is assessed to determine if it meets the specified tolerance level, with nonsignificant variables removed from the model. This method, employs a sequence of automated processes, aids in selecting the most significant predictors from a comprehensive set of predictors (Singh *et al*., 2014; Vashisth and Aravind, 2020; Vashisth and Goyal, 2023). Weather indices developed by combination of different weather variables were used for developing model. For developing wheat yield estimation model at different growth stage, yield was taken as dependent variables and simple and weight weather indices along with time was taken as independent variables in SPSS (version 16.0) by Stepwise multiple linear regression (SMLR) techniques.

The wheat yield (y) is expressed as a linear combination of selected predictors:

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

where x_j are selected weather indices, β_j are regression coefficients, and ε is the random error term. Predictor selection is performed using stepwise forward–backward selection.

Least absolute shrinkage and selection operator (LASSO)

LASSO is a model selection technique. LASSO models are used to overcome the shortcomings of ordinary least square (OLS) and ridge regression. Lasso estimators are used for consistent regression coefficient and automatic variable selection. Continuous shrinkage of some coefficients by imposing L1 penalty and others to zero, hence it helps to reduce multicollinearity and retain some good features of both subset selection and ridge regression. With large number of predictors, smaller subset selection exhibit stronger effect on interpretation of data. Subset selection is discrete and variable process, regressors are either retained or eliminated from the model in order to provide better interpretable model.

LASSO estimates regression coefficients by minimizing:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

where λ is the regularization parameter controlling sparsity.

Elastic Net

Elastic net penalises the size of regression coefficients based on both L1 norm and L2 norm penalty. L1 norm used to generate sparse model, L2 penalty removes the limitation on the number of selected variables, encourage grouping effect, stabilises the L1 regularization path. Alpha and beta are the two model parameters, need to be optimized by minimizing average mean square error in cross validation. Tuning parameter alpha values set in LASSO and Elastic Net were 1 and 0.5. "glmnet" package in R software was used to solve LASSO and Elastic Net.

Elastic Net combines LASSO (L1) and Ridge (L2) penalties:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}$$

where $\alpha \in [0,1]$ controls the balance between L1 and L2 penalties.

Support Vector Regression

Support vector regression (SVR) are widely used machine learning techniques known for their robustness and rapidity. The "kernel trick" enables SVR to capture non-linear relationships in data with great flexibility, as well as their efficiency and robustness. In this method, complex non-linear relationships are transformed into a high-dimensional space where they can be more easily examined by linear methods. The resulting relationships are then projected back into the original low-dimensional space. R software caret package was used for SVR, model implementation. SVR, model were subjected to ten iterations of five-fold cross-validation, each with a different model configuration, to facilitate model optimization and generalizability while minimizing overfitting on the dataset. Hyper-parameters are tuned based on the best cross-validation (CV) results.

SVR minimizes the ε-insensitive loss function:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

subject to:

$$\begin{aligned} y_i - (w \cdot \phi(x_i) + b) \quad &\leq \varepsilon + \xi_i \\ (w \cdot \phi(x_i) + b) - y_i \quad &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \quad &\geq 0 \end{aligned}$$

where $\phi(\cdot)$ is a kernel function, C is the regularization parameter, and defines the insensitive zone.

Artificial Neural Network

Artificial neural networks (ANN) are nonlinear data-driven self-adaptive methods that differ from traditional model-based approaches. ANN is capable of solving complex problems involving nonlinear data, even in the presence of noise and inaccuracies, as they mimic the human learning process. As agricultural data is often complex and nonlinear, hence ANNs are best suited for this type of data. ANN models were analyzed in SPSS Software (version 16.0). Multi-layer perceptron neural network was used for the analysis of ANN.

A multilayer perceptron with one hidden layer is expressed as:

$$y = f\left(\sum_{k=1}^{H} w_k \; g\left(\sum_{j=1}^{p} v_{kj} \, x_j + b_k\right) + b\right)$$

where x_j are inputs, v_{kj} and w_k are weights, b_k and b are biases, $g(\cdot)$ is the hidden-layer activation function, and is the output activation.

Principal component analysis- stepwise multiple linear regression (PCA- SMLR)

It is a combination of feature selection and selection method used for the data analysis. Principle components scores or factors are calculated from the data analysis which is used as an input variable for stepwise multiple linear regression. It is mainly used to reduce the multicollinearity problem arises from the weather variables. Principal component analysis (PCA) primarily deals with explaining variance through linear combination of original variables.

PCA is very effectively used as a multivariate technique for the purpose of data reduction. PCA transforms the original set of correlated variables in to a new set of uncorrelated variables. If the original variables taken into consideration are uncorrelated, then there so significant meaning in PCA analysis. The first principal component can able to explain the major variability in the data in a greater extend, and remaining each succeeding component accounts rest of the variability as possible.

Let xij be the value of jth weather variable/ biophysical parameter (j=1, 2, ..., P) corresponding to ith treatment of experiment (i= 1, 2, ..., n). The principal component analysis for xij's will be carried out. Let PC1, PC2, ...PCK are the principle component obtained from the analysis. First few PC (K< P) principal component will explain variability (maximum variability) about 90 per cent of the total variation in xij's. Using these K principal components as regressor variables and variety yield (yi) as regress and, the following linear multiple regression model for pre-harvest prediction of crop yield has been proposed.

yi = β0 + β1PC1i + β2PC2i+.......... βkPCki+ei (i= 1, 2, ...n)

where yi is the crop yield of ith variety; β0, β1, β2, .., bk are model parameters and ei denote the error term which is assumed to be follow normal distribution with mean zero and variance s2. This technique reduces the number of regressors to be used in the model and hence reasonably precise prediction of wheat yield can be obtained even for small set of observation.

Principal component analysis - artificial neural network (PCA-ANN)

In this techniques data analysis were done through combination of feature selection. Principle components scores or factors are calculated from the data analysis which is used as an input variable for ANN.

Hybrid machine learning approach

For a hybrid machine learning approach, a combination of SMLR with SVR and LASSO with

SVR approach was attempted. In the SMLR-SVR model, SMLR select variables from the data analysis and is used as an input variable for SVR. It is mainly used to reduce the multi-collinearity problem which arises from weather variables. In the LASSO-SVR model, first variables are selected by LASSO techniques and these variables are used as an input variable for SVR. Flow diagram showing methodology for yield prediction of wheat at different growth stage is shown in Fig. 1.

Cross-Validation and Model Calibration

To ensure model robustness and to avoid overfitting, cross-validation procedures were adopted during model development. The complete dataset was first divided into 70% for calibration (training) and 30% for independent validation, following standard practice in agrometeorological yield modeling. For regularized regression models (LASSO and Elastic Net), k-fold cross-validation (k = 5) was implemented during the calibration phase using the *glmnet* package in R to identify optimal regularization parameters (λ and α) by minimizing cross-validated mean squared error. For support vector regression (SVR) and hybrid models (SMLR–SVR and LASSO–SVR), model tuning was performed using five-fold cross-validation repeated ten times, enabling robust optimization of kernel and cost parameters and improving model generalization. For artificial neural network (ANN) and PCA-ANN models, internal validation was carried out using the training subset through iterative learning, while model performance was finally assessed using the independent validation dataset. This cross-validation framework ensured fair comparison among statistical, machine learning, and hybrid models and enhanced the reliability of multi-stage wheat yield estimation.

Accuracy test

Model performance during calibration and validation was observed on the basis of mean square error (MSE) root mean square error (RMSE), normalized mean square error (nRMSE), mean absolute percentage error (MAPE) and percent deviation was used to assess model accuracy, providing an intuitive measure of relative prediction error using the following formula.

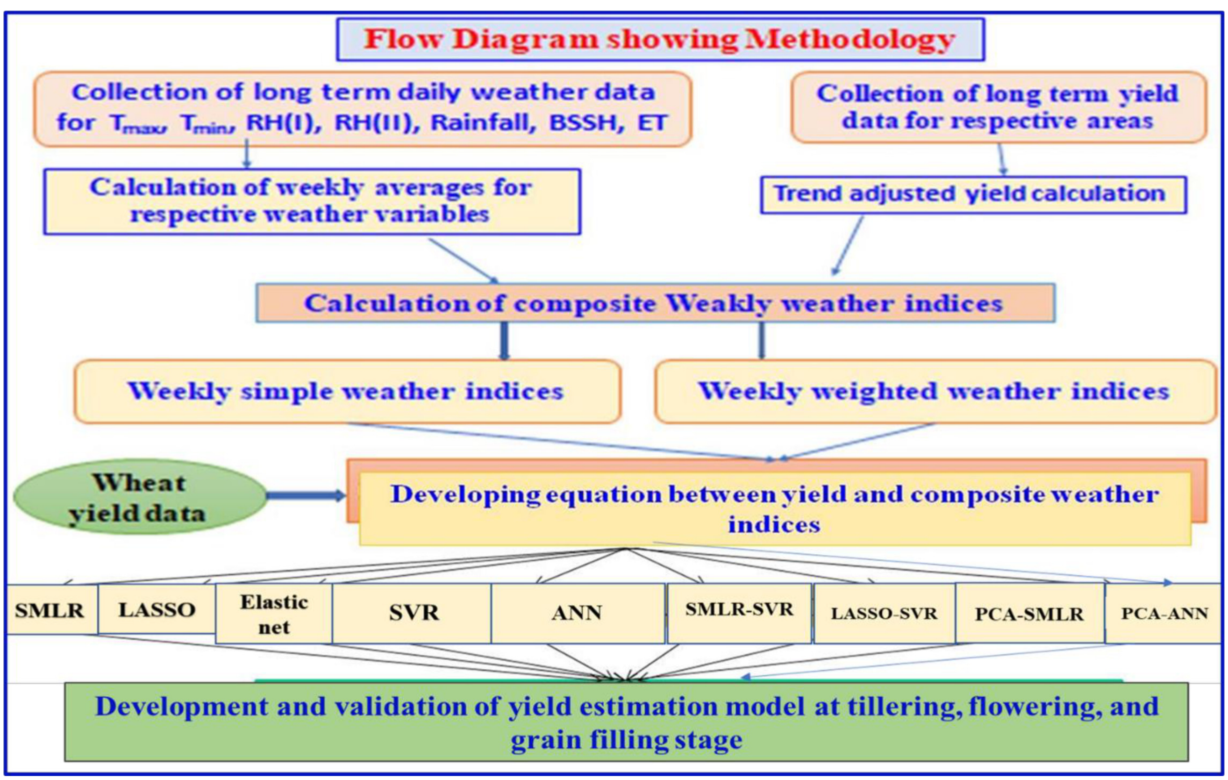

Fig. 1. Development of weather based models for wheat crop yield estimation using different techniques

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2}$$

$$nRMSE = \frac{100}{M} * \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2}$$

$$\mathrm{MAPE}\ (\%) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i - O_i}{O_i}\right| \times 100$$

Percent Deviation **(%)** $= \frac{Pi - Oi}{Oi} * 100$

Where Pi is the estimated value, O_i is the observed value, N is the number of observations, M is the mean of observed value. The estimation is considered excellent with the nRMSE value less than10%, good if 10–20%, fair if 20–30%, poor if more than 30%.

Results and Discussion

Weather based wheat yield estimation models developed by different techniques for IARI, New Delhi

Wheat yield prediction models for IARI, New Delhi was developed using long term crop yield data along with long period daily weather data from 46th to 15th standard meteorological week. The model was developed using stepwise multi linear regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net, support vector regression (SVR), artificial neural network (ANN), and hybrid (PCA-SMLR, PCA-ANN, SMLR-SVR, LASSO-SVR) techniques in R software version 3.1.3.

Accuracy parameters of the developed model during calibration and validation are shown in Table 1.

The RMSE value during calibration was ranged between 40.0 kg/ha for SVR to 161.8 kg/ha for PCA-ANN. During calibration nRMSE value was less than 5% for all the models having lowest value for SVR followed by LASSO-SVR, LASSO, elastic net, SMLR, SMLR-SVR, ANN PCA-SMLR, and PCA-ANN model. MAPE values during validation ranged from 1.12% (LASSO) to 14.92% (ANN), closely matching the trends observed for nRMSE and confirming the superior performance of LASSO, LASSO-SVR, and SVR models. During calibration, MAPE values ranged from 1.21% (SVR) to 4.53% (PCA-ANN), closely following the trends observed for RMSE and nRMSE, indicating stable model fitting and low relative prediction error. Performance during calibration of model developed using different techniques for wheat yield estimation for IARI New Delhi are shown in Fig. 2.

Results showed that model developed by different techniques performed better with value of nRMSE during validation ranged between 1.15 to 15.48%. Minimum value of nRMSE was found for LASSO followed by SMLR-SVR, LASSO-SVR, SVR, SMLR, PCA-SMLR, elastic net, PCA-ANN, and ANN. Performance of during validation of model developed using different techniques for wheat yield

Table 1. Weather based wheat yield estimation models for IARI, New Delhi using different techniques

S. No.	Techniques used for developing model	Modal accuracy parameter during calibration				Modal accuracy parameter during validation			
		MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)
1	SMLR	9254	96.2	2.77	2.68	50969	225.8	5.11	5.02
2	LASSO	5177	72.0	2.08	2.01	2594	50.9	1.15	1.12
3	Elastic Net	6247	79.0	2.33	2.25	123855	351.9	8.33	8.01
4	SVR	1938	44.0	1.26	1.21	11192	105.5	2.39	2.31
5	ANN	15102	122.9	3.57	3.42	431426	656.8	15.48	14.92
6	PCA-SMLR	24847	157.6	4.65	4.48	67923	260.6	6.20	5.94
7	PCA-ANN	26170	161.8	4.70	4.53	382431	618.4	14.58	13.96
8	SMLR-SVR	9650	98.2	2.81	2.72	4659	68.3	1.54	1.47
9	LASSO-SVR	4612	67.9	1.94	1.87	6913	83.1	1.88	1.79

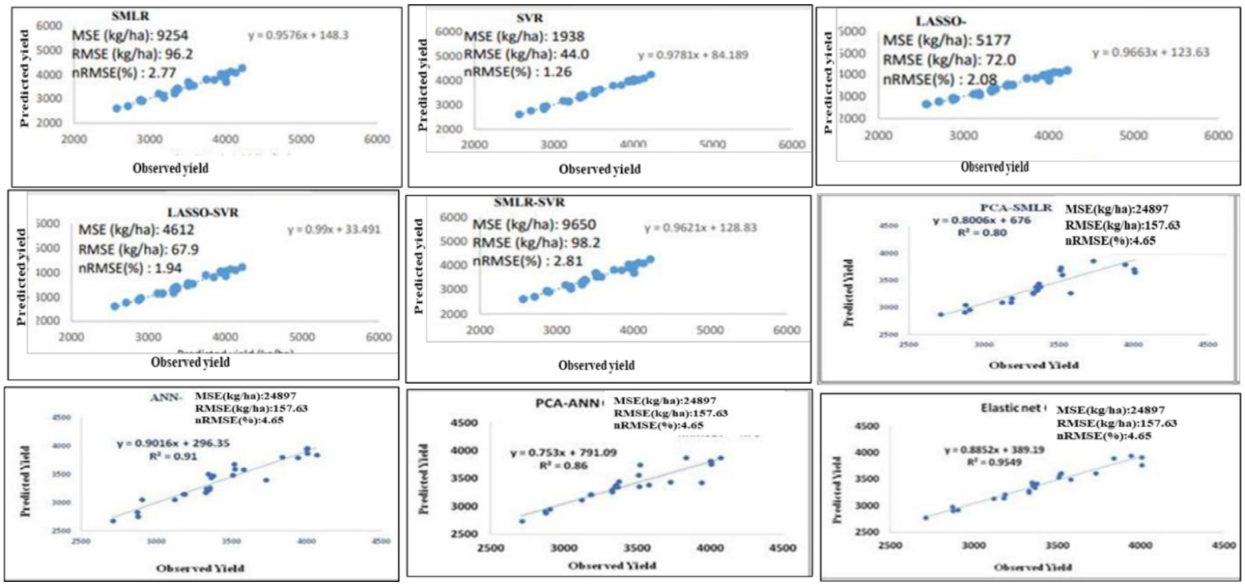

Fig. 2. Performance during calibration of model developed using different techniques for wheat yield estimation IARI New Delhi

estimation IARI New Delhi is shown in Fig 3. Based on nRMSE value during validation, model developed by LASSO, SMLR-SVR, LASSO-SVR and SVR performed excellent having nRMSE value less than 5%. Model developed by SMLR, PCA-SMLR, and elastic net performed excellent having nRMSE value less than 10%. Model developed by PCA-ANN and ANN performed good having nRMSE value less than 15%. RMSE value during validation ranged between 50.9 kg/ha for LASSO to 656.8 kg/ha for ANN.

Wheat yield estimation at tillering stage by model developed by different techniques for IARI, New Delhi

Model for wheat yield at tillering stage for IARI,

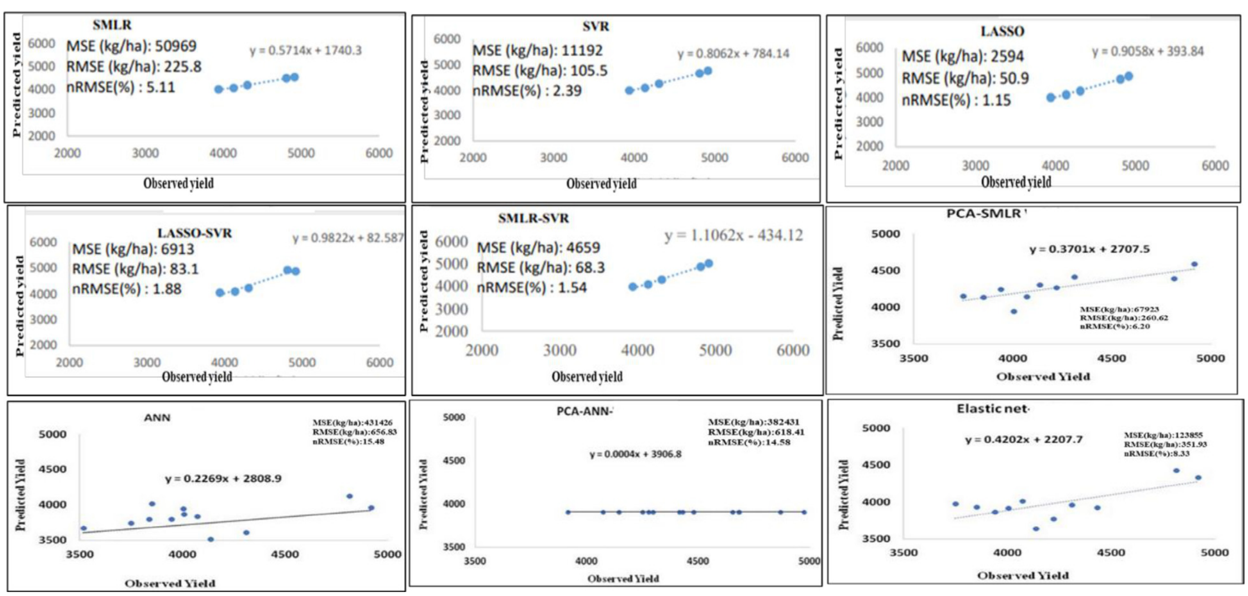

Fig. 3. Performance of during validation of model developed using different techniques for wheat yield estimation IARI New Delhi

New Delhi were developed by different techniques using long term crop yield data as well as long period daily weather data during sowing to vegetative stage (from 46[th] to 4[th] standard meteorological week). Performance of the model during calibration developed for wheat yield estimation at tillering stage using SMLR and PCA-SMLR techniques are shown in Table 2. The RMSE value during calibration was ranged between 42.5 kg/ha for SVR to 144.5 kg/ha for ANN. During calibration nRMSE value was less than 5% for all the models having lowest value for SVR followed by SMLR-SVR, LASSO, elastic net, PCA-SMLR, LASSO-SVR, SMLR, PCA-ANN and ANN model. The coefficient of determination. During calibration, MAPE values ranged from 1.18% (SVR) to 4.03% (ANN), closely following the trends observed for RMSE and nRMSE, indicating stable model fitting and low relative prediction error.

Performance of modal developed for wheat yield estimation at tillering stage using different techniques during validation are shown in Table 2. During validation nRMSE value was ranged between 4.13 to 12.6% having lowest value for LASSO followed by LASSO-SVR, PCA-SMLR, elastic net, SVR, SMLR, SMLR-SVR, ANN and PCA-ANN. The RMSE value during validation was ranged between 174.6 kg/ha for LASSO to 537.2 kg/ha for PCA-ANN. At the tillering stage, validation MAPE ranged from 3.96% (LASSO) to 12.06% (PCA-ANN), corroborating the nRMSE-based ranking and confirming the superior performance of LASSO, LASSO-SVR, SVR and PCA-SMLR models.

Percentage deviation of estimated yield for wheat crop done at tillering stage by observed yield was less than 10% for all the models. The percentage deviation value was lowest for PCA-SMLR, followed by SMLR-SVR, elastic net, LASSO, ANN, LASSO-SVR, PCA-ANN, SMLR, and SVR.

Wheat yield estimation at flowering stage by model developed by different techniques for IARI, New Delhi

Model for wheat yield estimation at flowering stage for IARI, New Delhi was developed by by different models using long term crop yield data as well as long period daily weather data during sowing to flowering stage (from 46[th] to 8[th] standard meteorological week). Model performance during calibration for wheat yield estimation at flowering stage developed by different techniques are shown in Table 3. The RMSE value during calibration was Ranged between 39.8 kg/ha to 188.1 kg/ha having lowest value for LASSO-SVR followed by SVR, SMLR-SVR, LASSO, SMLR, elastic net, PCA-SMLR, ANN and PCA-ANN. The nRMSE value during calibration was between 1.15 to 5.46%. During calibration nRMSE value was less than 10% for all the models having lowest value for LASSO-SVR followed by SVR, SMLR-SVR, LASSO, SMLR, elastic net, PCA-SMLR, ANN and PCA-ANN. During calibration, MAPE values ranged from 1.18% (LASSO-SVR) to 5.25% (PCA-ANN), closely following the trends observed for RMSE and

Table 2. Wheat yield estimation at tillering stage by different modal for IARI, New Delhi

S. No.	Techniques used for developing model	Modal accuracy parameter during calibration				Modal accuracy parameter during validation				Percentage deviation (%)
		MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	
1.	SMLR	18572	136.3	3.95	3.78	64701	254.4	6.00	5.72	-6.62
2.	LASSO	11582	107.6	3.18	3.02	30475	174.6	4.13	3.96	4.01
3.	Elastic Net	12998	114.0	3.36	3.19	42181	205.4	4.86	4.62	3.57
4.	SVR	1807	42.5	1.23	1.18	43512	208.6	4.92	4.68	-7.41
5.	ANN	20872	144.5	4.20	4.03	245847	495.8	11.63	11.18	4.65
6.	PCA-SMLR	13733	117.2	3.45	3.29	38021	195.0	4.62	4.39	-2.19
7.	PCA-ANN	20340	142.6	4.14	3.98	288616	537.2	12.60	12.06	4.94
8.	SMLR-SVR	8304	91.1	2.63	2.51	67872	260.5	6.14	5.86	-3.12
9.	LASSO-SVR	17800	133.4	3.85	3.68	32469	180.2	4.25	4.06	-4.71

Table 3. Wheat yield estimation at flowering stage by different modal for IARI, New Delhi

S. No.	Techniques used for developing model	Modal accuracy parameter during calibration				Modal accuracy parameter during validation				Percentage deviation (%)
		MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	
1.	SMLR	6660	81.6	2.40	2.31	40663	201.7	4.77	4.56	4.46
2.	LASSO	2297	47.9	1.37	1.31	3038	55.1	1.30	1.25	-0.91
3.	Elastic Net	9851	99.3	2.92	2.80	54662	233.8	5.54	5.29	1.63
4.	SVR	1938	44.0	1.27	1.22	24918	157.9	3.72	3.56	-2.05
5.	ANN	15117	123.0	3.57	3.43	212456	460.9	10.81	10.34	-1.26
6.	PCA-SMLR	10658	103.2	3.04	2.92	33368	182.7	4.33	4.14	0.91
7.	PCA-ANN	35378	188.1	5.46	5.25	410573	640.8	15.03	14.41	-6.98
8.	SMLR-SVR	1983	44.5	1.27	1.22	83914	289.7	6.83	6.52	-3.60
9.	LASSO-SVR	1584	39.8	1.15	1.10	17157	131.0	3.09	2.95	-0.85

nRMSE, indicating stable model fitting and low relative prediction error. At the flowering stage, validation MAPE varied from 1.25% (LASSO) to 14.41% (PCA-ANN), closely following the nRMSE-based ranking and reaffirming the superior performance of LASSO, LASSO-SVR, SVR, and PCA-SMLR models

Performance of model developed for wheat yield estimation at flowering stage using different techniques during validation are shown in Table 3. During validation nRMSE value ranged between 1.30 to 15.03%. Minimum value of nRMSE was found for LASSO followed by LASSO-SVR, SVR, PCA-SMLR, SMLR, elastic net, SMLR-SVR, ANN, and PCA-ANN. The RMSE value during validation was between 55.1 to 640.8 kg/ha. Percentage deviation value was less than 10% for all the models. The percentage deviation value was lowest for LASSO-SVR followed by LASSO and PCA-SMLR, ANN, elastic net, SVR, SMLR-SVR, SMLR, and PCA-ANN

Wheat yield estimation at grain filling stage by model developed by different techniques for IARI, New Delhi

Model for wheat yield estimation at grain filling stage for IARI, New Delhi was developed by different models using long term crop yield data as well as long period daily weather data during sowing to grain filling stage (from 40th to 11th standard meteorological week).

Model performance during calibration for wheat yield estimation at grain filling stage developed by different models are shown in Table 4. The RMSE value during calibration was between 40.0 kg/ha for SMLR-SVR and 205.5 kg/ha for SVR. During calibration nRMSE value was between 1.15 to 5.92% having lowest value for SMLR-SVR followed by LASSO, LASSO-SVR, PCA-SMLR, SMLR, PCA-ANN, ANN, Elastic net and SVR. During calibration, MAPE values ranged from 1.10% (SMLR-SVR) to 5.67% (SVR), closely following the trends observed for RMSE and nRMSE, indicating stable model fitting and low relative prediction error. The percentage deviation value was lowest for LASSO followed by SVR, elastic net, LASSO-SVR, SMLR-SVR, PCA-SMLR, SMLR, ANN and PCA-ANN

During validation nRMSE value was between 1.22 to 15.48% having lowest value for LASSO followed by LASSO-SVR, PCA-SMLR, SVR, SMLR, SMLR-SVR, elastic net, PCA-ANN and ANN. The RMSE value during validation was between 51.8 to 659.8 kg/ha. Percentage deviation of estimated yield for wheat crop done at grain filling stage by observed yield was less than 10% for all the models. The percentage deviation value was lowest for LASSO followed by SVR, elastic net, LASSO-SVR, SMLR-SVR, PCA-SMLR, SMLR, ANN and PCA-ANN. At the grain filling stage, validation MAPE ranged from 1.18% (LASSO) to 14.92% (ANN), similar to the trends observed for nRMSE and percentage deviation, and confirming the robustness of LASSO and LASSO-SVR models.

Table 4. Wheat yield estimation at grain filling stage by different modal for IARI, New Delhi

S. No.	Techniques used for developing model	Modal accuracy parameter during calibration				Modal accuracy parameter during validation				Percentage deviation (%)
		MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	MSE (kg/ha)	RMSE (kg/ha)	nRMSE (%)	MAPE (%)	
1.	SMLR	9639	98.2	2.89	2.77	69580	263.8	6.25	5.98	-2.37
2.	LASSO	3674	60.6	1.73	1.66	2685	51.8	1.22	1.18	-0.02
3.	Elastic Net	22204	149.0	4.37	4.18	228111	477.6	10.3	9.86	-0.74
4.	SVR	42230	205.5	5.92	5.67	36133	190.1	4.41	4.21	-0.67
5.	ANN	19249	138.7	4.03	3.86	435310	659.8	15.48	14.92	3.79
6.	PCA-SMLR	9080	95.3	2.81	2.69	33878	184.1	4.36	4.16	-2.29
7.	PCA-ANN	17284	131.5	3.82	3.66	334743	578.6	13.57	13.02	6.66
8.	SMLR-SVR	1596	40.0	1.15	1.10	79052	281.2	6.63	6.34	-2.24
9.	LASSO-SVR	4612	67.9	1.94	1.86	6913	83.1	1.88	1.79	-1.03

Singh *et al.* (2014) reported that statistical models based on weather indices can successfully simulate multi-stage yield forecast of wheat at mid-season and at pre-harvest for Amritsar, Bhatinda and Ludhiana districts. This model is simple, does not require any sophisticated statistical tools, and can be used satisfactorily for district, agro-climatic zone and state level forecasting.

Vashisth *et al*. (2018) reported that percentage deviation of estimated yield by observed yield by weather based statistical model for maize crop done at flowering stage and at grain filling stage was 10.3 and 7.1%. Aravind *et al.* (2025) reported that out of five different models, viz, SMLR, ANN, SVR, RF, and DNN, developed for multi-stage wheat yield estimation using long-term weather data for five distinguished districts of Punjab, RF, SVR, and DNN models performed excellent followed by SMLR and ANN. These three RF, SVR, and DNN models performed best for wheat yield prediction at all three growth stages of the study area as compared to SMLR and ANN models.

Gupta *et al.* (2022) developed the model for multi stage wheat yield prediction by stepwise multi linear regression (SMLR), support vector regression (SVR), least absolute shrinkage and selection operator (LASSO) and hybrid machine learning (LASSO-SVR and SMLR-SVR) techniques. They reported that LASSO performed best having nRMSE value between 1.22% at grain filling stage for IARI, New Delhi to 8.36% for Hisar at flowering stage. The model performance of SVR is increased if a hybrid model in combination with LASSO and SMLR is applied. The hybrid model LASSO-SVR has shown more improvement than SVR model compared with SMLR-SVR. Vashisth and Goyal (2023) reported that out of six models compared for mustard yield prediction for different zones of Rajasthan, PCA-SVM performed the best for the study area followed by SMLR-SVM. Vashisth et al. (2023) reported that on the basis of overall performance of models developed for mustard yield prediction at different growth stage, PCA-SMLR model performed better than SMLR model having lower value of nRMSE, RMSE and percentage deviation of observed yield by predicted yield. Machine learning models have exhibited superior performance compared to statistical models in wheat yield forecasting and have outperformed classic linear equations in predicting wheat and corn yields, based on RMSE and nRMSE evaluation metrics (Setiya *et al*., 2022).

Conclusions

This study demonstrates the effectiveness of integrating statistical, machine learning, and hybrid modeling approaches for multi-stage wheat yield estimation using long-term weather data for IARI, New Delhi. Across tillering, flowering, and grain-filling stages, all developed models produced reliable yield estimates, with percentage deviation between observed and predicted yields remaining within ±10%, indicating strong predictive capability well before harvest. Among the individual approaches, LASSO and SVR consistently outperformed conventional statistical models, while hybrid models, particularly LASSO–SVR, further improved prediction accuracy by combining effective feature selection with nonlinear learning. Model evaluation using RMSE, nRMSE, MAPE, and percentage deviation confirmed that LASSO, LASSO–SVR, PCA–SMLR, and SVR models achieved excellent performance (nRMSE < 10%) across growth stages, with the grain-filling stage providing the highest accuracy.

References

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

Aravind, K.S., Vashisth, Ananta, Krishanan, P., Kundu, M., Prsad, S., Meena, M.C., Lama, A., Das, P., and Das, B. 2025. Development of multistage crop yield estimation model using machine learning and deep learning techniques *International Journal of Biometeorology* **69**(2): 499-515.

Azfar, M., Sisodia, B.V.S., Rai, V.N. and Devi, M. 2015. Pre-harvest forecast models for rapeseed & Maize yield using principal component. *Mausam* **4**: 761–766.

Garde, Y.A., Dhekale, B.S. and Singh, S. 2015. Different approaches on pre harvest forecasting of wheat yield. *Journal of Applied and Natural Science* **7(2)**: 839–843.

Gupta, S., Vashisth, A., Krishnan, P., Lama, A., Prasad, S. and Aravind, K.S. 2022. Multistage wheat yield prediction using hybrid machine learning techniques. *Journal of Agrometeorology* **24**(4): 373-379. https://doi.org/10.54386/jam.v24i4.1835

Gupta, Shreya, Vashisth, Ananta, Krishnan, P., Lama, A. Shivprasad and Aravind, K.S. 2024. Weather based wheat yield prediction using machine learning. *Mausam* **75**(3): 639-648 https://doi.org/10.54302/mausam.v75i3.5606

Kumar, R., Gupta, B.R.D., Athiyaman, B., Singh, K.K. and Shukla, R.K. 1999. Stepwise regression technique to predict pigeon pea yield in Varanasi district, *Journal of Agrometeorology* **1**(2): 183-186.

Setiya, P., Satpathi, A., Nain, A.S. and Das, B. 2022. Comparison of weather-based wheat yield forecasting models for different districts of Uttarakhand using statistical and machine learning techniques. *Journal of Agrometeorology* **24**(3): 255-261. https://doi.org/10.54386/jam.v24i3.1571

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

Singh, A.K., Vashisth, Ananta, Sehgal, V.K., Goyal, A., Pathak, H. and Parihar, S.S. 2014. Development of multi stage district level wheat yield forecast models. *Journal of Agricultural Physics* **14**(2): 189-193.

Verma, U., Piepho, H.P. and Goyal, A. 2016. Role of climatic variables and crop condition term for mustard yield prediction in Haryana. *International Journal of Agricultural and Statistical Science* **12**: 45–51.

Vashisth, Ananta, Singh R. and Choudary, Manu. 2014. Crop yield forecast at different growth stage of wheat crop using statistical model under semi arid region. *Journal of Agroecology and Natural Resource Management*: 1-3.

Vashisth, Ananta, Goyal, Avinash and Roy, Debasish. 2018. Pre harvest maize crop yield forecast at different growth stage using different model under semi arid region of India. *International Journal of Tropical Agriculture* **36**(4): 915-920.

Vashisth, Ananta and Aravind K.S. 2020. Multistage Mustard Yield Estimation Based on Weather Variables using Multiple Linear, LASSO and Elastic Net Models for Semi Arid Region of India. *Journal of Agricultural Physics* **20**(2): 213-223.

Vashisth, Ananta, Aravind, K.S., Beck, M.K., Kundu, Monika and Krishanan, P. 2023. Mustard yield prediction at different growth stage using principal component analysis *Journal of Agricultural Physics* **23**(2): 247-254.

Vashisth, Ananta and Goyal, A. 2023. Prediction of mustard yield using different machine learning techniques: a case study of Rajasthan, India. *International Journal of Biometeorology* **67**(3): 539-551. https://doi.org/10.1007/s00484-023-02434-2

White, J.W. 2009. Comments on a report of regression-based evidence for impact of recent climate change on winter wheat yields. *Agriculture Ecosystem and Environment* **129**(4): 547-548. https://doi.org/10.1016/j.agee.2008.10.025.

Received: 22 July 2025; Accepted: 29 October 2025