MAUSAM, **75**, 3 (July 2024), 639-648

MAUSAM

DOI : https://doi.org/10.54302/mausam.v75i3.5606
Homepage: https://mausamjournal.imd.gov.in/index.php/MAUSAM

UDC No. 633.11 : 004.383.8

Weather based wheat yield prediction using machine learning

SHREYA GUPTA, ANANTA VASHISTH, P.KRISHNAN, ACHAL LAMA*,
SHIVPRASAD and ARAVIND K.S.

ICAR-Indian Agricultural Research Institute, Pusa, New Delhi – 110 012

**ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi – 110 012*

(*Received 21 March 2022, Accepted 11 May 2023*)

e mail :ananta.iari@gmail.com

सार – मौसम के प्राचलों के प्रभाव से गेहूं की फसलें अत्यधिक प्रभावित होती हैं। इसके विश्वसनीय पूर्वानुमान के लिए मशीन लर्निंग का उपयोग करके मौसम-आधारित मॉडल विकसित और मान्य करने की आवश्यकता है। फसल उगाने की अवधि के दौरान गेहूं की उपज और मौसम के आंकड़े आईएआरआई, नई दिल्ली, हिसार, अमृतसर, लुधियाना और पटियाला से एकत्र किए गए। उपज अनुमान मॉडल को चरणबद्ध बहुरैखिक समाश्रयण (एसएमएलआर), सपोर्ट वेक्टर समाश्रयण (एसवीआर), लीस्ट एब्सोल्यूट श्रींकेज & सलेक्शन ऑपरेटर (एलएएसओ) और हाइब्रिड मशीन लर्निंग मॉडल एलएएसएसओ-एसवीआर तथा आर सॉफ्टवेयर में एसएमएलआर-एसवीआर का उपयोग करके विकसित किया गया। अंशांकन के लिए 70% डेटा और सत्यापन के लिए शेष 30% डेटा तय करके विश्लेषण किया गया। अध्ययन क्षेत्रों के लिए गेहूं की उपज अनुमान मॉडल 46वें से 15वें मानक मौसम सप्ताहों के दीर्घकालिक दैनिक मौसम डेटा के साथ-साथ दीर्घकालिक फसल उपज डेटा का उपयोग करके विकसित किए गए। विभिन्न स्थानों के लिए गेहूं की उपज केअनुमान के लिए इन मॉडलों की जाँच करने पर, LASSO ने सबसे अच्छा प्रदर्शन किया, जिसमें nRMSE मान पटियाला के लिए 0.6% से लुधियाना के लिए 4.8% के बीच था। यदि LASSO और SMLR के संयोजन में एक हाइब्रिड मॉडल लागू किया जाता है तो SVR का मॉडल प्रदर्शन बढ़ जाएगा। हाइब्रिड मॉडल LASSO-SVR ने SMLR-SVR की तुलना में SVR मॉडल में अधिक सुधार दिखाया है।

ABSTRACT. Wheat crops are highly affected by the influence of weather parameters. Thus, there is a need to develop and validate weather-based models using machine learning for its reliable prediction. Wheat yield and weather data during the crop growing period were collected from IARI, New Delhi, Hisar, Amritsar, Ludhiana and Patiala. The yield prediction model was developed using stepwise multi linear regression (SMLR), support vector regression (SVR), least absolute shrinkage and selection operator (LASSO) and hybrid machine learning model LASSO-SVR and SMLR-SVR in R software. Analysis was done by fixing 70% of the data for calibration and the remaining 30% data for validation. Wheat yield prediction models for study areas were developed using long term crop yield data along with long period daily weather data from the 46^{th} to 15^{th}standard meteorological weeks. On examining these models for wheat yield prediction for different locations, LASSO performed best having nRMSE value ranged between 0.6 % for Patiala to 4.8% for Ludhiana. 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 in SVR model compared with SMLR-SVR.

Key words– Weather variable, Machine learning model, Support vector regression, Least absolute shrinkage and Selection operator, Stepwise multi linear regression, Yield prediction.

1. Introduction

Wheat (*Triticum aestivum*) is one of the principal crops of the India and wheat cultivation has traditionally been conquered by the northern region of India. Wheat production are significantly influenced and controlled by rainfall, temperature, solar radiation, and relative humidity (Ji *et. al*., 2007, Dutta *et. al*., 2001, Yadav *et. al*., 2015). Weather variability within the crop growing seasons is an intense source of variability in yields. 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, predicting crop yield using weather variables is foremost important (Azfar *et al*., 2015, Pandey *et. al.*, 2014). Crop yield forecast may be done by using biometrical characteristics, weather variables and agricultural inputs. These methodologies can be used individually or in

combination to give a composite model. (Agrawal *et al*., 2001, Jain *et al*., 1980). Therefore, there is a need to develop area specific prediction models based on time series data with the help of machine learning to predict crop yield more accurately.

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 ,Garde *et al*., 2015). To overcome these demerits, least absolute shrinkage and selection operator (LASSO) and machine learning technique can be used. LASSO improves the quality of prediction by shrinking regression coefficient, when compared to prediction models fitted through unpenalized maximum likelihood methods. Tibshirani (1996) proposed LASSO, which can be utilized in the crop yield prediction technique. LASSO minimizes the residual sum of squares subject to the sum of the absolute value of the coefficient being less than a constant. It produces interpretable models like subset selection and exhibits the stability of even the ridge regression. (Aravind *et. al.*, 2022, Kumar*et. al.*, 2019, Vashisth*et. al.* 2018, 2018, 2020)reported that elastic Net and LASSO were found to be the best model for wheat yield prediction of different locations of north-west India. Support vector machine (SVM) are a set of related supervised learning methods used for classification and regression. They belong to a family of generalized linear classifiers, or in other terms it is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fitting to the data. The SVM can be used both for grouping and regression problems and it can be indicated as a two-layered network where the weights are non-linear in the first layer and it is linear in the second layer (Bray and Han, 2004, Parviz *et. al.*, 2018). SVM is applied to construct nonlinear nonparametric forecasting models to be used in Crop yield forecast models for barley, canola and spring wheat grown on the Canadian Prairies developed using vegetation indices derived from satellite data machine learning methods (Johnson *et al.*, 2016). Two hybrid approaches like the ARIMAX-ANN and the ARIMAX-SVM have been used for the rice yield along with weather variables of Aligarh district of Uttar Pradesh. Based on the results obtained, performance of ARIMAX-SVM and ARIMAX-ANN models were close to each other but much superior to the conventional ARIMAX model for the considered data set. Performance of the hybrid ARIMAX model was found to be quite encouraging. (Alam *et al.*, 2018).

To overcome the various challenges in wheat yield prediction, in the present investigation, model was developed by Least absolute shrinkage and selection operator (LASSO), stepwise multiple linear regression (SMLR), support vector machine (SVM), hybrid machine learning (SMLR-SVR, and LASSO-SVR) technique for improving the accuracy of yield prediction model.

2. Materials and methods

2.1. *Data collection and development of heat indices*

Daily weather data during wheat crop growing period of 1971 to 2017 for Amritsar and Patiala were collected from the met centre Chandigarh, 1972 to 2017 for Ludhiana from AMFU Ludhiana, 1985 to 2018 for Hisar and IARI, New Delhi from AMFU Hisar and AMFU New Delhi. Wheat yield data was collected from the Directorate of Economics & Statistics (DES) and the state agricultural department. Different thermal indices, weather indices and evapotranspiration of wheat crop in a given period were developed, with detailed methods described below along with methods used for accuracy assessment and validation. Different thermal indices were calculated from sowing up to harvest of the crop as given by the following equations

$$\text{Growing degree days (GDD)} = \sum \left\{\left[\left(T_{\max} + T_{\min}\right)/2\right] - T_{\text{base}}\right\}$$

where $T_{\max}$ is the daily maximum temperature, $T_{\min}$is the daily minimum temperature and T_{base} is the base temperature. The base temperature varies crop to crop and its value is derived from the growth behaviours of the specific crop. The base temperature is the temperature below which plant growth is zero. Wheat base temperature is taken at 5 °C. The negative value of GDD is taken as zero.

$$\text{Helio-thermal units (HTU)} = \sum\left(\text{GDD} \times \text{SSH}\right)$$

where SSH is the bright sunshine hours

Heat use efficiency = Yield/GDD

Photo thermal index (PTI) = GDD/ crop growing day

2.2. *Estimation of evapotranspiration*

Evapotranspiration (ET) is the combination of two separate processes in which water is lost from the soil surface called evaporation and from the crop by transpiration. Both the processes of evaporation and transpiration occur simultaneously and there is no easy way of distinguishing between them. Evapotranspiration is

TABLE 1

Weather indices used in models using composite weather variables

	Simple weather indices								Weighted weather indices							
	T_{max}	T_{min}	RF	RH I	RH II	SSH	EVP	ET_0	T_{max}	T_{min}	RF	RH I	RH II	SSH	EVP	ET_0
T_{max}	Z10								Z11							
T_{min}	Z120	Z20							Z121	Z21						
Rf	Z130	Z230	Z30						Z131	Z231	Z31					
RH I	Z140	Z240	Z340	Z40					Z141	Z241	Z341	Z41				
RH II	Z150	Z250	Z350	Z450	Z50				Z151	Z251	Z351	Z451	Z51			
SSH	Z160	Z260	Z360	Z460	Z560	Z60			Z161	Z261	Z361	Z461	Z561	Z61		
EVP	Z170	Z270	Z370	Z470	Z570	Z670	Z70		Z171	Z271	Z371	Z471	Z571	Z671	Z71	
ET_0	Z180	Z280	Z380	Z480	Z580	Z680	Z780	Z80	Z181	Z281	Z381	Z481	Z581	Z681	Z781	Z81

normally expressed in millimetres (mm) per unit time. The rate expresses the amount of water lost from a cropped surface in the unit of depth of water. The reference evapotranspiration (ET_0) is the evapotranspiration from the reference surface. The reference surface is a hypothetical grass reference crop with an assumed crop height of 0.12 m, a fixed surface resistance of 70 sm^{-1} and an albedo of 0.23. The reference surface closely resembles an extensive surface of green, well-watered grass, actively growing and completely shading the ground. ET_0 can be calculated from meteorological data using the FAO Penman-Monteith method. This method is recommended as the standard method for the definition and computation of the reference evapotranspiration. It requires radiation, air temperature, air humidity and wind speed data.

ETo is derived from the FAO Penman-Monteith method using the following equation:

$$\mathrm{ET}_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)}$$

where,

ET_0 = reference evapotranspiration [mm day^{-1}],

R_n =net radiation at the crop surface [MJ m^{-2} day^{-1}]

G =soil heat flux density [MJ m^{-2} day^{-1}],

T =mean daily air temperature at 2 m height [°C],

u_2 =wind speed at 2 m height [ms^{-1}],

e_s = saturation vapour pressure [kPa],

e_a = actual vapour pressure [kPa],

e_s-e_a =saturation vapour pressure deficit [kPa],

Δ =slope vapour pressure curve [kPa $°C^{-1}$],

γ =psychrometric constant [kPa $°C^{-1}$]

2.3. *Development of weather indices*

Daily weather data during the crop growing period is used for generating weather indices. Weather indices used for developing the crop yield forecast model are given in Table1.

$$Y = A_0 + \sum_{i=1}^{p}\sum_{j=0}^{1} a_{ij} Z_{ij} + \sum_{i\neq 1=1}^{p}\sum_{j=0}^{1} a_{ii'j} Z_{ii'j} + cT + e$$

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw} \text{ and } Z_{ii'j} = \sum_{w=1}^{m} r_{ii'w}^{j} X_{iw} X_{ii'w}$$

where,

X_w denotes the value of the weather variable under the study in w^{th} week, n is the number of weeks in the crop season and A_o, a_o, a_1 and a_2 are model parameters. These models were extended to study combined effects of weather variables and an additional variate T representing the year time trend. Y is yield; r_{iw}/$r_{ii'w}$ is the correlation coefficient of yield (adjusted for trend effect) with i-th weather variable (X_{iw}) /product of i-th and i'-th weather (X_{iw}/$X_{i'w}$) variables in w-th period; m is the numweek of forecast, p is the number of weather variables used and e is an error term.

In this type of method, for each weather variable, two types of weather indices were developed. The first one being the simple values of weather variables during the crop growing period [un-weighted index -Z_{i0}] and the second one is weighted [weighted index Z_{i1}]. Weights are taken as correlation coefficients between yield and weather variables in respective periods. In the same way, indices were also produced for interaction of weather variables by using weekly products of weather variables taking two at a time. Combinations of a various weather variables for Weather indices were generated and are presented in Table 1. Weather parameters,*viz*., maximum and minimum temperature, morning and evening relative humidity, rainfall, bright sunshine hours, evaporation and evapotranspiration were used for such a model.

2.4. *Development of models*

For development of a yield prediction model, weather indices were developed by weather parameters from 46 to 15th standard meteorological weeks. Thermal and weather indices were used for developing wheat yield prediction model using empirical and machine learning techniques for five different locations. R software was used for developing the multistage wheat yield prediction model, package HDCI was used for LASSO and package e1071 was used for SVR.

Stepwise multiple linear regression was used for developing the model as below:

$$Y = A_0 + a_0 \sum_{w=1}^{n} X_w + a_1 \sum_{w=1}^{n} wX_w + a_2 \sum_{w=1}^{n} w^2 X_w + e$$

where, X_w signifies the value of the weather variable under study in w^{th} week; n is the number of weeks in the crop season and A_0, a_0, a_1 and a_2 are the model parameters. This model was also extended to study the combined effects of weather variables and an additional variate T which is represents the year for considering time trend.

Least absolute shrinkage and selection operator (LASSO) being a model selection technique, is used to overcome the shortcomings of ordinary least square (OLS) and ridge regression. Regressors are either retained or is eliminated from the model in order to provide the better interpretable model. Support vector machine (SVM) is a discriminative classifier defined by a separating hyperplane,*i.e*., given labeled training data, the algorithm outputs an optimal hyperplane which categorizes a new set of examples. Into two dimentional space, this hyperplane is a line dividing a plane in two parts where in each class on either side.It finds a line/ hyper-plane (in multidimensional space that separates outs classes). Support vectors are data points that lie close to the decision surface or hyperplane. The generalized SVMs for time series forecasting have a two-stage neural network architecture. In the first stage a self-organizing feature map (SOM) is used as the clustering algorithm to partition the whole input space into several disjointed zones. A tree-structured architecture is adopted in the partition to avoid the problem of pre-determining the number of partitioned regions. In the second stage, multiple SVMs, also called SVM experts, that best fit the partitioned regions are constructed by finding the most appropriate kernel function along with the optimal free parameters of SVMs. SVMs experts also converge faster and use fewer support vectors. This is established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error. It is shown to be very resistant to the over-fitting problem, ultimately achieving high generalization performance in solving time series forecasting problems. A key property of SVMs is that training SVMs is equivalent to solving a linearly constrained quadratic programming problem so that the solution is always unique and globally optimal.In the modelling of time series, SVM tries to reduce the key problem which are noise and non-stationarity. For each particular region, only the expert that best fits it is used for the final prediction.

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 SVM.

2.5. *Accuracy test*

The performance of statistical models was estimated by calculating root mean square error (RMSE), normalized root mean square error (nRMSE) and mean square error using the following formula.

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

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

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

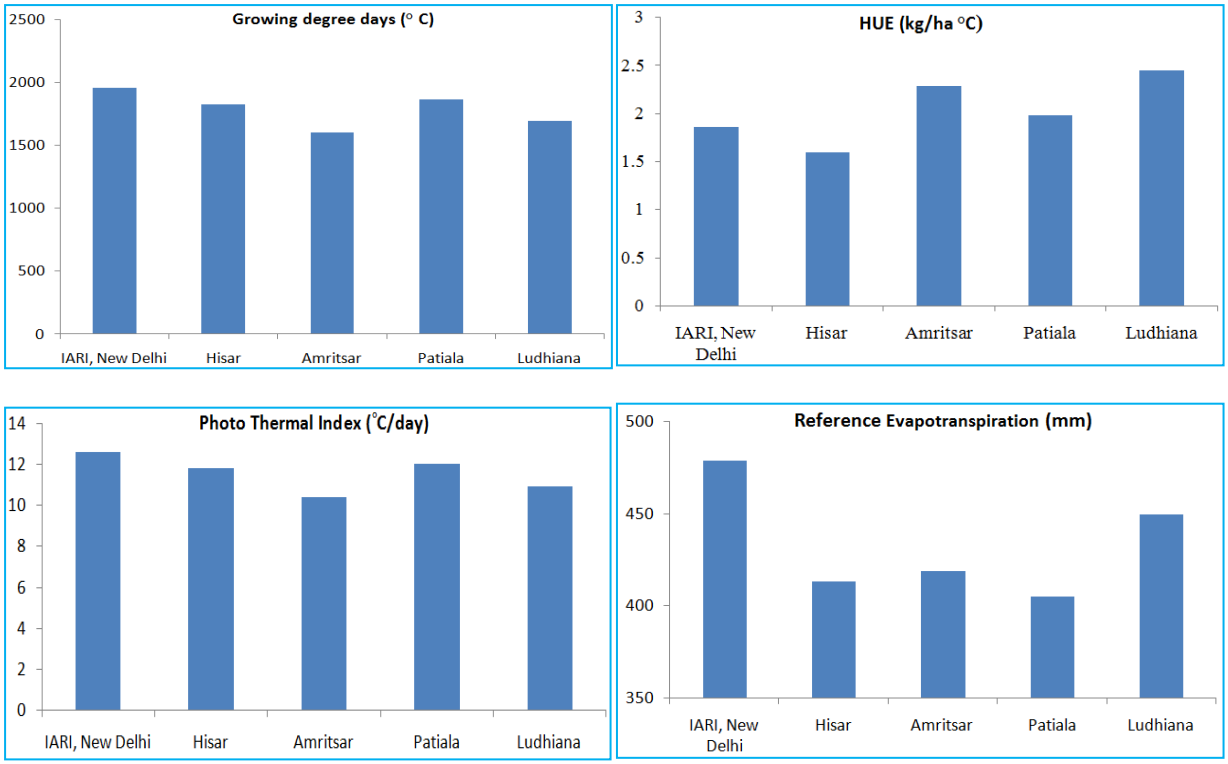

Fig. 1. Growing degree days, Heat use efficiency, average photo thermal index and reference evapotranspiration during wheat crop growing period for different station

where, *Pi*, O_i,*N* and *M* are predicted value, observed value, number of observations and mean of observed value. nRMSE is considered excellent with the nRMSE value less than 10%, good if nRMSE value ranges between 10-20%, fair if value ranged between 20-30% and poor if value is more than 30%.

3. Results and discussion

3.1. *Heat indices and reference evapotranspiration during wheat crop growing period for IARI, New Delhi*

Different heat indices and reference evapotranspiration were calculated during wheat growing period 1984 to 2018 for IARI, New Delhi. Value of growing degree days (GDD) calculated throughout wheat crop growing period was ranged between 1784.8 during 1996 to 2061 °C during 1987. The average value of GDD is 1953.5 °C. Value of helio thermal unit (HTU) calculated for IARI, New Delhi seen to be ranged between 9933.0 during 2010 to 16630.4 °C hour during 1987. Average value of HTU found was 12899.4 °C hour. Heat use efficiency for IARI, New Delhi ranged between 1.27 kg/ha/°C during 1984 to 2.51 kg/ha/°C during 2017. Average value of HUE is 1.85 kg/ha/°C. Value of PTI calculated during crop growing value for IARI, New Delhi ranged between 11.5 °C /day during 1996 to 13.4 °C /day during 2009. Average value of PTI seen is 12.6 °C /day. Cumulative value of reference evapotranspiration calculated during crop growing period for IARI, New Delhi was lowest (404.6 mm) during 2013 and highest (539.8 mm) during 1984. Average value of reference evapotranspiration during crop growing period was 478.8 mm. (Fig. 1)

3.2. *Heat indices and reference evapotranspiration during wheat crop growing period for Hisar, Haryana*

Different heat indices and reference evapotranspiration were calculated during wheat growing season 1985 to 2018 for Hisar. Value of growing degree days (GDD) calculated during wheat crop growing period remained between 1579.7 during 2004 to 3586.6 °C during 2002. The average value of GDD stood 1821.4 °C. Value of helio thermal unit (HTU) calculated for Hisar ranged between 10468.5 during 1997 to 28681.4 °C hour during 2002. Average value of HTU stood 14222.5 °C hour. Heat use efficiency for Hisar ranged between

TABLE 2

Weather based wheat yield prediction models for IARI, New Delhi

S.No.	Model	Modal accuracy parameter during calibration			Modal accuracy parameter during validation		
		MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)	MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)
1.	SMLR	9254	96.2	2.7	50969	225.8	5.1
2.	SVR	1938	44.0	1.2	11192	105.5	2.3
3.	LASSO	5177	72.0	2.0	2594	50.9	1.1
4.	LASSO-SVR	4612	67.9	1.9	6913	83.1	1.8
5.	SMLR-SVR	9650	98.2	2.8	4659	68.3	1.5

0.18 kg/ha/°C during 1971 to 2.96 kg/ha/°C during 2012. Average value of HUE stayed 1.59 kg/ha/°C. Value of PTI calculated during crop growing value for Hisar ranged between 10.7 °C /day during 1981 to 23.2 °C /day during 2002. Average value of PTI stayed 11.8 °C /day. Cumulative value of reference evapotranspiration calculated during crop growing period for Hisar seen to be the lowest (283.9 mm) during 2012 and highest (477.4 mm) during 1975. Average value of reference evapotranspiration during crop growing period seen is 413.2 mm (Fig. 1)

3.3. *Heat indices and reference evapotranspiration during wheat crop growing period for Amritsar, Punjab*

Different heat indices and reference evapotranspiration were calculated during wheat growing season 1971 to 2017 for Amritsar. Value of growing degree days (GDD) calculated during the wheat growing period seen between 1450.9 during 1988 to 1747.0 °C during 1979. The average value of GDD was 1600.2 °C. Heat use efficiency for Amritsar ranged between 1.34 kg/ha/°C during 1975 to 3.97 kg/ha/°C during 2013. Average value of HUE stood 2.28 kg/ha/°C. Value of PTI calculated during crop growing value for Amritsar was between 8.0 °C /day during 2013 to 11.4 °C /day during 2001. Average value of PTI stood 10.4 °C /day. Cumulative value of reference evapotranspiration calculated during crop growing period for Amritsar was lowest (357.7 mm) during 2000 and highest (482.6 mm) during 1977. Average value of reference evapotranspiration during crop growing period was 418.6 mm (Fig. 1)

3.4. *Heat indices and reference evapotranspiration during wheat crop growing period for Ludhiana, Punjab*

Different heat indices and reference evapotranspiration were calculated during wheat growing season 1972 to 2017 for Ludhiana. Value of growing degree days (GDD) calculated during wheat crop growing period ranged between 1512.5 during 1973 to 1878.4 °C during 2010. The average value of GDD was 1691.6 °C. Value of heilo thermal unit (HTU) calculated for Ludhiana ranged between 11930.4 during 1983 to 31752.4 °C hour during 1989. Average value of HTU stayed 21792.6 °C hour. Heat use efficiency for Ludhiana ranged between 1.84 kg/ha/°C during 1975 and 1980 to 3.20 kg/ha/°C during 2011. Average value of HUE was 2.44 kg/ha/°C. Value of PTI calculated during crop growing value for Ludhiana ranged between 9.8 °C /day during 1973 to 12.1 °C /day during 2010. Average value of PTI was 10.9 °C /day. Cumulative value of reference evapotranspiration calculated during crop growing period for Ludhiana was the lowest (354.4 mm) during 2013 and the highest (546.6 mm) during 2010. The average value of reference evapotranspiration during the crop growing period was 449.6 mm (Fig. 1)

3.5. *Heat indices and reference evapotranspiration during wheat crop growing period for Patiala, Punjab*

Different heat indices and reference evapotranspiration were calculated during wheat growing season 1971 to 2017 for Patiala. Value of growing degree days (GDD) calculated during wheat crop growing period ranged between 1665.9 during 1981 to 2079.4 °C during 2003. The average value of GDD was 1866.0 °C. Heat use efficiency for Patiala ranged between 0.66 kg/ha/°C during 1971 to 2.52 kg/ha/°C during 2011. Average value of HUE was 1.98 kg/ha/°C. Value of PTI calculated during crop growing value for Patiala ranged between 10.7 °C /day during 1981 to 13.4 °C/day during 2003 and 2009. Average value of PTI was 12.0 °C/day. Cumulative value of reference evapotranspiration calculated during crop growing period for Patiala was the lowest 321.7 mm during 2004 and 2012 and highest (453.1 mm) during 1976. Average value of reference evapotranspiration during crop growing period was 405.2 mm (Fig. 1).

TABLE 3

Weather based wheat yield prediction models for Hisar, Haryana

S.No.	Model	Modal accuracy parameter during calibration			Modal accuracy parameter duringvalidation		
		MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)	MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)
1.	SMLR	4872	69.8	2.8	155331	394.1	9.0
2.	SVR	20142	141.9	5.7	140440	374.8	8.5
3.	LASSO	12191	110.4	4.4	15158	123.1	2.8
4.	LASSO-SVR	9264	96.3	3.8	21281	145.9	3.3
5.	SMLR-SVR	73826	271.7	10.9	259510	509.4	11.6

3.6. *Weather based wheat yield prediction models for IARI, New Delhi*

Wheat yield prediction models for IARI, New Delhi have been 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), support vector regression (SVR), least absolute shrinkage and selection operator(LASSO), variable section by LASSO and SVR (LASSO-SVR), variable section by SMLR and SVR (SMLR-SVR) techniques in R software version 3.1.3. Performances of the developed model during calibration and validation period are shown in Table 2. Results showed that model developed by different techniques performed better with value of nRMSE ranged between 1.2 to 2.8 %. The nRMSE value during validation ranged between 1.15 to 5.11 %. The lowest value of nRMSE was 1.15 % for the model developed by LASSO followed by SMLR-SVR (1.5 %), LASSO-SVR (1.8 %)), SVR (2.3 %) and SMLR (5.1 %). Based on nRMSE value during validation, model developed for wheat predictions for IARI, New Delhi using different techniques was excellent having nRMSE value less than 5 %. The most important weather parameter identified for wheat crop yield prediction model developed by SMLR techniques for IARI, New Delhi are Z281 (Minimum temperature*evapotranspiration) Z581(minimum relative humidity*evapotranspiration)and Z671 (sunshine hour*evaporation) while the important weather parameter identified for wheat crop yield prediction model developed by LASSO are Z10 (maximum temperature), Z51 (minimum relative humidity), Z141(maximum temperature*morning relative humidity), Z151 (maximum temperature*minimum relative humidity), Z240 (minimum temperature*morning relative humidity), Z260(minimum temperature*bright sunshine hours), Z461(morning relative humidity*bright sunshine hours), Z561(evening relative humidity*bright sunshine hours), Z671(bright sunshine hours*evaporation), Z181(maximum temperature*evapotranspiration), Z581(evening relative humidity *evapotranspiration), GDD, PTI. The Parameters for wheat crop yield prediction model developed by SVR was SVM-Type: eps-regression, SVM-Kernel: linear kernel function with cost as 1, gamma as 0.01282051 and epsilon as 0.1 and Number of Support Vectors are 21. For yield prediction model developed by LASSO-SVR the parameters are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.05882353, epsilon: 0.1 and Number of Support Vectors: 19 and for model developed by SMLR-SVR the parameters recognized are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.2, epsilon: 0.1 and Number of Support Vectors was 17.

3.7. *Weather based wheat yield prediction models for Hisar, Haryana*

Wheat yield prediction models for Hisar have been developed using long term crop yield data along with long period daily weather data from 46th to 15th standard meteorological week. Modal accuracy parameter during calibration and validation are shown in Table 3. The models developed using different techniques have nRMSE value during calibration between 2.8 to 10.9%. The nRMSE for validation ranged between 2.8 to 11.6%. The maximum value of nRMSE was found for the model developed by SMLR-SVR (11.65%), followed by SMLR (9.06%), SVR (8.57%), LASSO-SVR (3.34%) and LASSO (2.82%).Based on the nRMSE value during validation, model developed for wheat predictions for Hisar using all techniques were excellent having nRMSE value less than 10% except for SMLR-SVR techniques having nRMSE value 11.6%. Among the different model developed for wheat crop prediction for Hisar, modal developed by LASSO techniques performed best followed by LASSO-SVR, SVR, SMLR and SMLR-SVR techniques. The important weather parameter identified for wheat yield prediction model developed by SMLR techniques for Hisar are Time, Z21 (minimum

TABLE 4

Weather based wheat yield prediction models for Amritsar, Punjab

S.No.	Model	Modal accuracy parameter during calibration			Modal accuracy parameter duringvalidation		
		MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)	MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)
1.	SMLR	1696	41.2	1.2	42851	207.0	4.6
2.	SVR	4773	69.1	2.0	63399	251.8	5.7
3.	LASSO	3424	58.5	1.7	6431	80.2	1.8
4.	LASSO-SVR	3813	61.8	1.8	26270	162.1	3.6
5.	SMLR-SVR	2052	45.3	1.3	36268	190.4	4.3

TABLE 5

Weather based wheat yield prediction models for Ludhiana, Punjab

S.No.	Model	Modal accuracy parameter during calibration			Modal accuracy parameter duringvalidation		
		MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)	MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)
1.	SMLR	32378	179.9	4.6	252524	502.5	10.5
2.	SVR	14207	119.2	3.0	239146	489.0	10.2
3.	LASSO	61111	247.2	6.3	53080	230.4	4.8
4.	LASSO-SVR	18052	134.4	3.4	221144	470.3	9.8
5.	SMLR-SVR	32941	181.5	4.6	295003	543.1	11.3

temperature*maximum temperature), Z461 (morning relative humidity* bright sunshine hours) and Z581 (evening relative humidity *evapotranspiration) and the weather elements identified by LASSO techniques are time, Z61 (bright sunshine hours* maximum temperature), Z120 (maximum temperature*minimum temperature), Z121 (maximum temperature*minimum temperature), GDD, HUE and PTI.The Parameters for SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.01298701, epsilon: 0.1 and number of Support Vectors: 24. The Parameters for LASSO-SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.1428571, epsilon: 0.1 and number of Support Vectors: 6 while the parameters for SMLR-SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.03333333, epsilon: 0.1 and number of Support Vectors was 30.

3.8. Weather based wheat yield prediction models for Amritsar, Punjab

The models developed for predicting the yield had the value of nRMSE ranged between 1.2 kg/ha to 2.0 kg/ha. The nRMSE for validation ranged between 1.8 to 5.7 %. The maximum value of nRMSE was found for the model developed by SVR (5.7%), followed by SMLR (4.6 %), SMLR-SVR (4.3%), LASSO-SVR (3.6%) and LASSO (1.8%).Based on the value of nRMSE during validation, model developed for wheat predictions for Amritsar using all techniques were excellent having nRMSE value less than 10%. Among the different model developed for wheat crop prediction for Amritsar, modal developed by LASSO performed best followed by LASSO-SVR, SVR, SMLR and SMLR-SVR (Table 4). The most important weather parameter identified by SMLR for Amritsar are Z581 (evening relative humidity*evapotranspiration), HUE and PTI. While the important weather parameter identified by LASSO are time, Z120 (maximum temperature*minimum temperature), Z140 (maximum temperature*morning relative humidity), Z151 (maximum temperature*evening relative humidity), Z241 (minimum temperature*morning relative humidity), Z250 (minimum temperature* evening relative humidity), Z81 (evapotranspiration), Z581 (evening relative humidity*evapotranspiration), GDD, HUE, PTI. The Parameters for SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.02173913, epsilon: 0.1 and number of Support Vectors: 19. ForLASSO-SVR the parameters recognized SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.09090909, epsilon: 0.1 and number of Support Vectors: 10 and for SMLR-SVR the parameters

TABLE 6

Weather based wheat yield prediction models for Patiala, Punjab

S.No.	Model	Modal accuracy parameter during calibration			Modal accuracy parameter duringvalidation		
		MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)	MSE(kg/ha)	RMSE(kg/ha)	nRMSE(%)
1.	SMLR	716	26.8	0.7	919	30.3	0.6
2.	SVR	7164	84.6	2.4	80369	283.5	6.1
3.	LASSO	929	30.5	0.8	893	29.9	0.6
4.	LASSO-SVR	27593	166.1	4.8	177433	421.2	9.0
5.	SMLR-SVR	3517	59.3	1.7	11829	108.8	2.3

recognized are as follows SVM-Type: eps-regression, Kernel: linear, cost: 1, gamma: 0.3333333, epsilon: 0.1 and the number of Support Vectors was 4.

3.9. *Weather based wheat yield prediction models for Ludhiana, Punjab*

The models developed for predicting the yield had the value of nRMSE for calibration ranged between and 3.0% to 6.3%. The nRMSE for validation ranged between 4.8 to 11.3%. The maximum value of nRMSE was found for the model developed by SMLR-SVR (11.3%), followed by SMLR (10.5%), SVR (10.2%), LASSO-SVR (9.8%) and LASSO (4.8%).Based on nRMSE value during validation, model developed for wheat predictions for Ludhiana were excellent for LASSO and LASSO-SVR having nRMSE value less than 10% and good for model developed by SVR, SMLR and SMLR-SVR.Among the different model developed for wheat crop prediction for Ludhiana, modal developed by LASSO performed best followed by LASSO-SVR, SVR, SMLR and SMLR-SVR. (Table 5). The important weather parameter identified by SMLR for Ludhiana are time, Z141(maximum temperature*morning relative humidity) and Z151(maximum temperature*evening relative humidity). While the weather elements identified by LASSO are time, Z41(morning relative humidity*maximum temperature), Z141(maximum temperature*morning relative humidity) and Z151(maximum temperature*evening relative humidity),Z251 (Minimum temperature*evening relative humidity) and Z361 (rainfall*bright sunshine hours). The Parameters for SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.0212766, epsilon: 0.1 and number of Support Vectors: 28. The various parameters for LASSO- SVR are as follows SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.1, epsilon: 0.1, number of Support Vectors: 23 while for SMLR-SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.3333333, epsilon: 0.1 and number of Support Vectors was 24.

3.10. *Weather based wheat yield prediction models for Patiala, Punjab*

The models developed for predicting the yield had the value of nRMSE during calibration values ranged between 0.7 to 4.8%. The nRMSE value during validation, model developed by all techniques for wheat predictions for Patiala performed excellent having nRMSE value less than 10%.Among the different model developed for wheat crop prediction for Patiala, modal developed by LASSO and SMLR performed best followed by SMLR-SVR, SVR and LASSO-SVR techniques. Performances of the developed model during calibration and validation period are shown in Table 6. The various weather parameter identified by SMLR for Patiala are time,Z81 (evapotranspiration *maximum temperature), Z241 (minimum temperature*morning relative humidity), GDD and HUE while the parameters identified by LAASO are time, Z81 (evapotranspiration *maximum temperature), Z140 (maximum temperature* morning relative humidity), Z250 (minimum temperature* evening relative humidity), Z281 (Minimum temperature* evapotranspiration), GDD and HUE. The Parameters for SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.02173913, epsilon: 0.1 and number of Support Vectors: 22; for LASSO-SVR these are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.02173913, epsilon: 0.1 and number of Support Vectors: 22, while for SMLR-SVR are SVM-Type: eps-regression, SVM-Kernel: linear, cost: 1, gamma: 0.2, epsilon: 0.1 and number of Support Vectors was 6.

4. Conclusion

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. From this study, it may be concluded that models developed by weather parameters using machine learning techniques can be used for district level wheat yield prediction.

Acknowledgement

The first author acknowledges PG school, Indian Agricultural Research Institute, New Delhi for providing fellowship for conducting research work. The Authors acknowledge the research facilities extended by Director, ICAR-Indian Agricultural Research Institute, New Delhi.

Disclaimer: The content 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

Agrawal, R., Jain, R. C. and Mehta, S.C., 2001, "Yield forecast based on weather variables and agricultural inputs on agro-climatic zone basis", *Indian J. Agri. Sci.*, **71**, 7,487-490.

Alam, W., Ray, M., Kumar, R. R., Sinha, K., Rathod, S. and Singh K. N., 2018, "Improved ARIMAX modal based on ANN and SVM approaches for forecasting rice yield using weather variables", *Indian J. Agri. Sci.*, **88**, 12, 1909-13.

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", *J. of Agrometeorology*, **24**, 1, 18-25.

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

Bray, M. and Han, D., 2004, "Identification of support vector machines for runoff modeling", *J. Hydro inf.*, **6**, 4, 265-280.

Dutta, S., Patel, N. K. and Srivastava, S. K., 2001, "District wise yield models of rice in Bihar based on water requirement and meteorological data", *J. Indian Soci. Remo. Sens.*, **29**, 3, 175-181.

Garde, Y. A., Dhekale, B. S. and Singh, S. 2015, "Different approaches on pre harvest forecasting of wheat yield", *J. Appli. Nat. Sci.*, **7**, 2, 839-843.

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

Ji, B., Sun, Y., Yang, S. and Wan, J., 2007,"Artificial neural networks for rice yield prediction in mountainous regions", *J. Agri. Sci.*,**145**, 249-261.

Johnson, M. D., Hsieh, W., Cannon, A., Davidson, A. and Bedard, F., 2016,"Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods", *Agri. Fore.Meteorol.*,74-84.

Kumar, S., Attri, S. D. and Singh, K. K., 2019,"Comparison of Lasso and stepwise regression technique for wheat yield predication", *J. of Agrometeorol.*, **21**, 2, 188-192.

Pandey, K. K., Rai, V. N. and Sisodia, B. V. S., 2014,"Weather variables based rice yield forecasting models for Faizabad district of eastern UP," *International J. Agri. Stat. Sci.*, **10**, 2, 381-385.

Parviz, L., 2018, "Assessing accuracy of barley yield forecasting with integration of climate variables and support vector regression", *Anna Biologia*, **73**, 1, 19-30.

Tibshirani, R., 1996,"Regression shrinkage and selection via lasso", *J. Roy. Stat. Soc. B.*, **58**, 267-288.

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", *J. Agroecol. Nat. Res. Manag.*, 1-3.

Vashisth, Ananta, Goyal, A. and Roy, Debasish, 2018,"Pre harvest maize crop yield forecast at different growth stage using different model under semi arid region of India", *Int. J. Trop. Agr.*, **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", *J. Agri. Physi.*, **20**, 2, 213-223.

Verma, U., Piepho, H. P. and Goyal, A., 2016, "Role of climatic variables and crop condition term for mustard yield prediction in Haryana", *Int. J. Agri. Stat. Sci.*,**12**, 45-51.

Yadav, R. R. and Sisodia, B. V. S., 2015, "Predictive models for Pigeon-pea yield using weather variables", *Int. J. Agri. Stat. Sci.*, **11**, 2, 462-472.