Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 76(1) 2022 23–30

Selection of Best Subset of Weather Input through Step-wise Regression Method for Preharvest Cotton Yield Prediction in Western Agro-climate Zone of Haryana

Aditi, Chetna, Pushpa and Urmil Verma
CCS Haryana Agricultural University, Hisar

Received 25 January 2022; Revised 29 April 2022; Accepted 18 May 2022

SUMMARY

Zonal-yield models incorporating a linear time trend and agro-meteorological (agromet) variables each spanning successive fortnights within the growth period of the cotton crop are developed within the framework of multiple linear regression analysis. These models have been used to predict the cotton yields in four cotton growing districts namely; Hisar, Bhiwani, Sirsa, Fatehabad covering more than 90% of cotton production of the Haryana State. Linear time-trend has been obtained using cotton yield data of the period 1980-81 to 2011-12. The fortnightly weather data along with trend yield have been utilized for the same period for building the zonal weather-yield models. Models have been validated for subsequent years i.e. 2012-13 to 2017-18, not included in the development of the models. The zonal models were fitted by taking DOA yield as dependent variable and fortnightly weather variables along with trend yield/CCT/dummy variables as regressors. The predictive performance(s) of the contending models were observed in terms of average absolute percent deviations of cotton yield forecasts in relation to the observed yield(s) and root mean square error(s). The adequacy of the fitted models was examined through histogram, normal-probability plot for the residuals and residual plot against fitted values for the selected models.

Although, the weather variables were found statistically significant as predictors and gave predictions with reasonably high coefficients of determination (R^2) but the predictions had too high percent deviations to be acceptable and hence were deemed unsuitable for routine crop yield forecasting. To improve the predictive accuracy of the agromet yield models, a dummy regressor variable in the form of Crop Condition Term (CCT), was added to the weather models. The addition of CCT to the weather models significantly improved the accuracies of the district-level yield predictions in the State. The predictive performance of the zonal agromet models was assessed using multiple metrics, including the adjusted R^2, the percent deviations of the forecast yields from the Department of Agriculture (DOA) yield estimates and the root mean square errors (RMSEs).

Keywords: Maximum temperature, Minimum temperature, Rainfall, Sun shine hours and relative humidity, Trend yield, Crop condition term.

1. INTRODUCTION

Cotton is one of the finest natural fibres available to mankind for clothing from time immemorial. Cotton is an important commercial crop of the country and contributing nearly 85% of the total domestic fibre consumption. It is estimated that cotton requirement in India by 2025 will be around 140 lakh bales of lint and the present production is around 123 lakh bales. In India, cotton occupies an area of nearly 124.44 lakh hectares, with a production of 370 lakh bales (2017-18), an estimated area of nearly 120 lakh ha and production 358.70 lakh bales (2018-19), ranking 1st in the world followed by China which accounts for about 29 per cent of the world cotton production. It has also the distinction of having the largest area under cotton cultivation in the world constituting about 36 per cent of the world area under cotton cultivation. The lint productivity of cotton is 524 kg/ ha, which is the lowest and far below that of the world average of 765 kg/ha. Chow (1960) developed a test for testing the equality between sets of coefficients in two linear regressions. Huda *et al.* (1975) explained how weather parameter distribution pattern affects rice yield at different stages of growth. In quantifying the relationship between rice yield and weather variables, the second-degree multiple regression equation can be used profitably.

Corresponding author: Aditi
E-mail address: sangwanaditi1994@gmail.com

The findings showed that the crop responds to weather parameters differently during different stages of growth. Yoshida (1981) observed clear varietal differences for high-temperature injuries at different growth stages. The rice plant appeared to be most sensitive to high temperature at the flowering stage. The most sensitive stage was found about 9 days before flowering.

Verma *et al.* (2014) developed models to predict cotton yields in five cotton growing districts; Hisar, Sirsa, Bhiwani, Rohtak and Jind covering more than 90% of cotton production of Haryana State. The addition of Crop Condition Term to weather models improved the accuracy of yield prediction in the State. Lal *et al.* (2017) developed models through Step wise regression analysis on districts level as well as zonal level, all the models were found highly significant at 0.1% level of significance for Pigeon pea crops.

Jain *et al.* (2019) examined the results of crop weather relationship of different rice varieties grown under different environments. The linear correlation analysis with SPSS-model had been found at tilling stages negative relationship significant at 5% level for maximum and minimum temperature. In 50% flowering stages maximum temperature significant at 1% level and rainfall were significantly correlated at 5% level. and dough stages were negatively correlated (1% level) to rainfall.

Deepankar *et al.* (2020) conducted an analysis for Karnal of Haryana district for a period of 35 years (1980-81 to 2015-16 with yield as dependent variable and weather indices both individually and in combinations, were used as the predictor variables in stepwise multiple regression techniques. The per cent relative deviation ranges from 1.15% to 7.6% and the cumulative effects of maximum temperature and relative humidity morning predicted rice yields that were very similar to real yields.

Time versus cotton yield graphs of all the districts for computation of trend yield

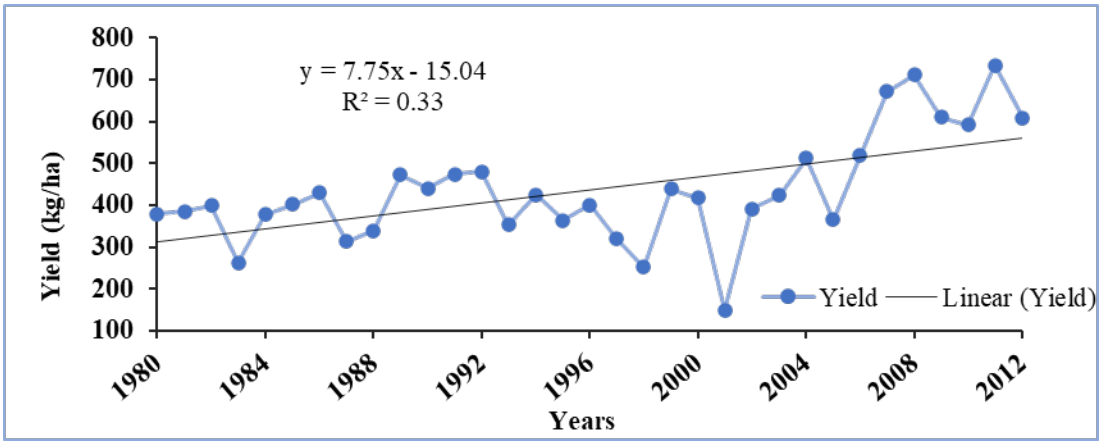

Fig. 1. Annual cotton yield (kg/ha) of Hisar district

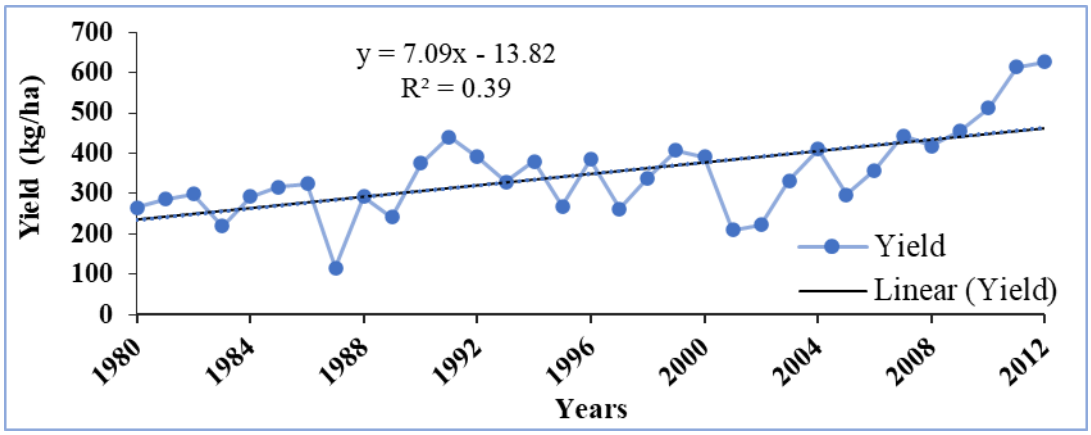

Fig. 2. Annual cotton yield (kg/ha) of Bhiwani district

Table 1. Post-sample period cotton yield estimates based on weather-yield models and their associated percentage deviations of Hisar district

District/ Forecast Years	DOA Yield (kg/ha)	Model-1		Model-2		Model-3		Model-4	
		FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)
2012-13	608.2	600.5	1.3	593.6	2.4	640.8	-5.4	635.0	-4.4
2013-14	501.3	542.9	-8.3	519.2	-3.6	541.1	-7.9	543.7	-8.5
2014-15	368.2	522.9	-42.0	474.4	-28.9	353.3	4.1	360.2	2.2
2015-16	276.0	399.7	-44.8	364.0	-31.9	259.9	5.8	269.3	2.4
2016-17	623.0	597.2	4.1	595.3	4.4	620.9	0.3	630.0	-1.1
2017-18	489	572.8	17.1	594.9	-21.7	449.1	8.2	481.6	1.5
Average abs. RD%		19.08		15.23		5.28		3.35	

Table 2. Post-sample period cotton yield estimates based on weather-yield models and their associated percentage deviations of Bhiwani district

District/ Forecast Years	DOA Yield (kg/ha)	Model-1		Model-2		Model-3		Model-4	
		FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)
2012-13	625.9	524.5	16.2	544.8	13.0	640.8	-2.4	635.0	-1.5
2013-14	523.9	468.2	10.6	442.0	15.6	541.1	-3.3	543.7	-3.8
2014-15	441.1	449.5	-1.9	477.3	-8.2	446.5	-1.2	452.6	-2.6
2015-16	334.0	454.2	-36.0	454.0	-35.9	353.1	-5.7	361.7	-8.3
2016-17	518.0	526.3	-1.6	522.0	-0.8	527.7	-1.9	537.6	-3.8
2017-18	681	534.2	21.6	555.0	18.5	635.5	6.7	666.4	2.1
Average abs. RD%		15.76		16.89		3.53		3.68	

Table 3. Post-sample period cotton yield estimates based on weather-yield models and their associated percentage deviations of Sirsa district

District/ Forecast Years	DOA Yield (kg/ha)	Model-1		Model-2		Model-3		Model-4	
		FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)
2012-13	767.6	726.0	5.4	753.2	1.9	734.0	4.4	727.4	5.2
2013-14	697.1	674.2	3.3	655.0	6.0	634.3	9.0	636.1	8.7
2014-15	619.2	660.1	-6.6	695.1	-12.3	632.9	-2.2	637.4	-2.9
2015-16	295.0	479.4	-62.5	500.8	-69.8	259.9	11.9	269.3	8.7
2016-17	694.0	746.0	-7.5	749.2	-8.0	620.9	10.5	630.0	9.2
2017-18	321	441.8	-37.6	510.3	59.0	355.9	-10.9	363.4	-9.4
Average abs. RD%		21.05		23.40		8.15		8.04	

Table 4. Post-sample period cotton yield estimates based on weather-yield models and their associated percentage deviations of Fatehabad district

District/ Forecast Years	DOA Yield (kg/ha)	Model-1		Model-2		Model-3		Model-4	
		FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)	FittedYield (kg/ha)	RD (%)
2012-13	715.1	744.4	-4.1	772.2	-8.0	734.0	-2.6	727.4	-1.7
2013-14	766.0	692.5	9.6	673.8	12.0	727.5	5.0	728.5	4.9
2014-15	653.5	678.0	-3.8	713.7	-9.2	632.9	3.2	637.4	2.5
2015-16	207.0	373.7	-80.5	368.7	-78.1	259.9	-15.6	269.3	-10.1
2016-17	686.0	763.5	-11.3	767.3	-11.9	620.9	9.5	630.0	8.2
2017-18	454	640.7	-41.1	669.1	-47.4	449.1	1.1	451.2	1.0
Average abs. RD%		26.13		24.21		7.83		8.00	

In the present study, zonal agromet (or weather) models have been developed for pre-harvest cotton yield forecasting in Hisar, Bhiwani, Sirsa and Fatehabad districts (accounting for more than 90% of the crop covered) of Haryana. Step-wise regression analysis was performed to identify the best supported linear/polynomial terms for the time trend in yield and agromet variables representing different fortnights constituting the cotton crop growth period. The analysis was performed using the SPSS software.

2. STUDY AREA AND DATA DESCRIPTION

Average yield data of cotton crop [(1980-81 to 2017-18)] in respect of the four local districts (Figs. 1 to 4) were collected from the Statistical Abstracts of Haryana. The meteorological data on the minimum temperature, maximum temperature, sunshine hours, relative humidity and rainfall for the period 1980-81 to 2017-18 were obtained from meteorological stations in Haryana. Cotton crop is generally sown before the onset of the monsoon (May-June) and harvested during the early part of winter (Oct.-Nov.). Thus, the crop growth period, i.e., 1st May to 31st October was first categorized into 12 successive fortnights. Further, the daily weather data were used to derive the average fortnightly minimum temperature, maximum temperature, sunshine hours, relative humidity and total rainfall across all 37 years. The weather data were exploited in developing the zonal yield models comprised of Hisar, Bhiwani, Sirsa and Fatehabad districts and CCT was included as an additional regressor (dummy) variable in extending the agromet models.

Statistical models and yield forecasts

Average yield data of cotton for Hisar, Bhiwani, Sirsa, and Fatehabad districts were regressed against the cropping year considered as an independent covariate along with the weather variables derived for the purpose. The general linear regression model with

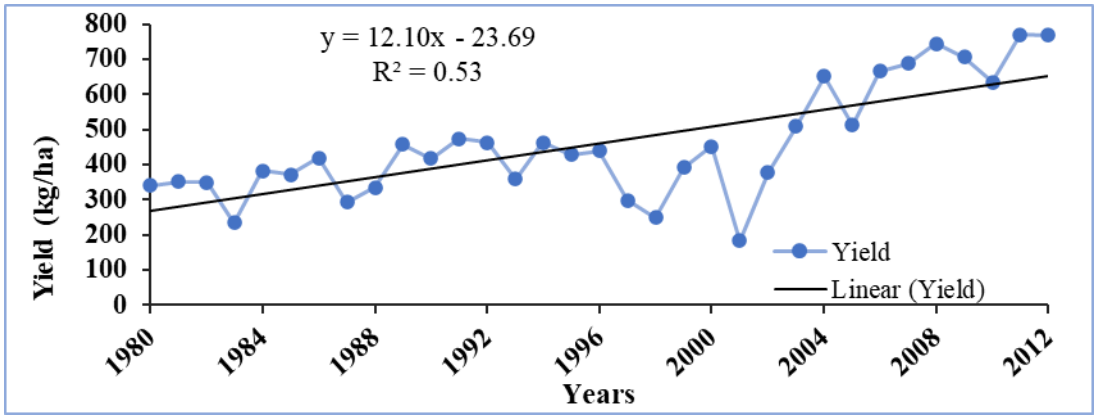

Fig. 3. Annual cotton yield (kg/ha) of Sirsa district

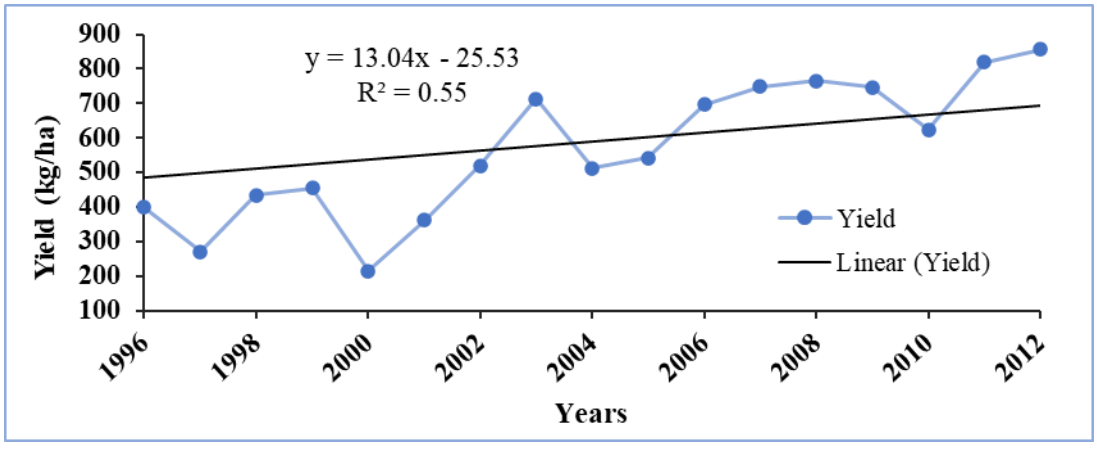

Fig. 4. Annual cotton yield (kg/ha) of Fatehabad district

the weather and time-trend covariates is specified as follows:

$$Y=a_0+,\sum_{i=1}^{12}b_i TMX_i+\sum_{j=1}^{12}b_j TMN_j+\sum_{k=1}^{12}b_k ARF_k+\sum_{l=1}^{12}b_l RH_l+\sum_{m=1}^{12}b_m ASSH_m +cCCT\ /\ dummy+e$$

where,

Y = yield (kg/ha)

a_0 = Overall mean effect

b_i,b_j,b_k,b_l,b_m= Regression coefficients of the weather variables

c = Regression coefficient of dummy variable

TMX_i = ith day maximum temperature

TMN_j = jth day minimum temperature

ARF_K = kth day rainfall

RH_l = lth day relative humidity

SSHm = mth day sunshine hours

i,j,k,l,m = Meteorological fortnights (1,2,3…12

e = The error term with assumption NID (0, σ^2)

Multiple linear regression models were fit by taking DOA yield as the response variable a fortnightly weather parameters along with trend yie CCT/dummy variables as predictor variables.The C is a categorical covariate obtained by dividing the tr predicted yield series into eight non-overlapping clas denoted by 1, 2, 3, 4, 5, 6, 7& 8 and correspondin real-time yields of 100-200, 200-300,300-400, 4 500, 500-600, 600-700, 700-800 &800-900 in kg/ respectively.

The multiple regression models were fitted using various combinations of the time-trend and agromet variables defined to capture all critical phenological periods. The best-supported weather variables were selected using step wise regression programme [Draper and Smith (1981)] in which all variables were first included in the model and eliminated one at a time with decisions at any particular step conditioned by the result of the previous step. The best-supported agromet variables were retained in the model if they had the highest adjusted R^2 and the lowest standard error of estimate (SE) at a given step.

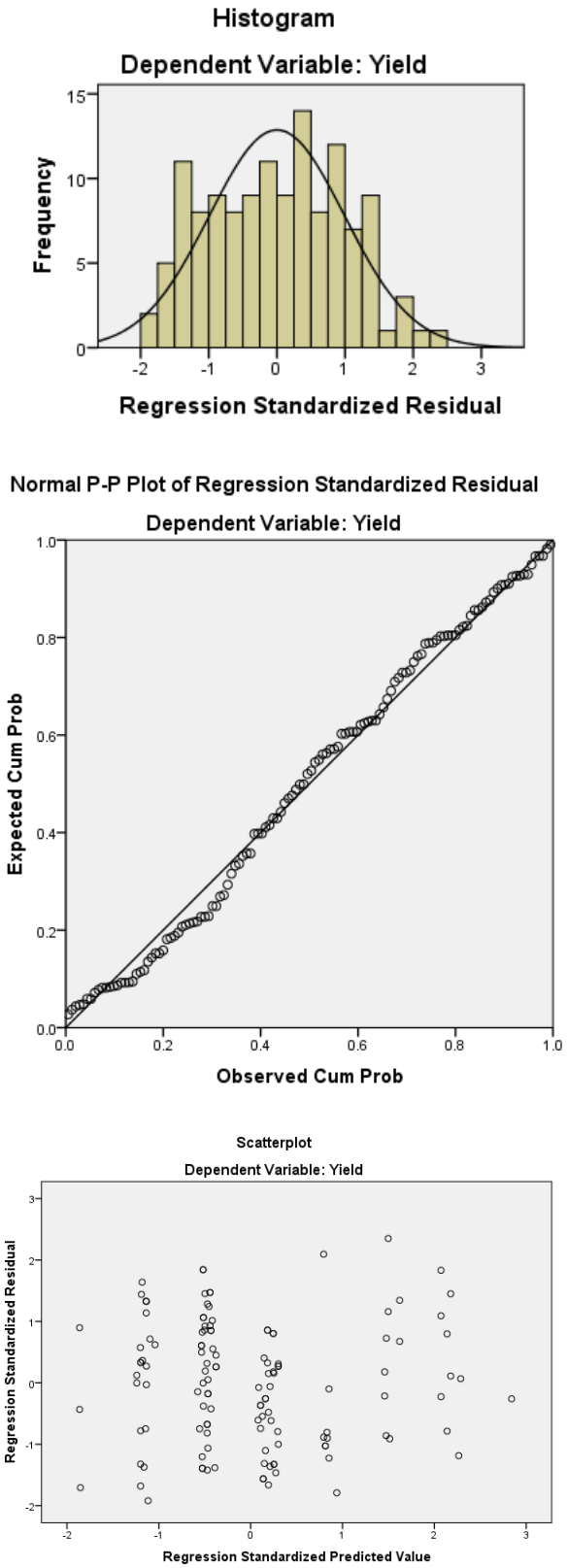

Fig. 5. Regression diagnostics of the selected zonal yield forecast model. (CCT+ weather variables)

Zonal agromet cotton yield models developed

The predictive performance(s) of the zonal yield models were selected based on Adj. R^2 and percent deviations of crop yield estimates from the real-time yield(s) and root mean square errors (RMSEs). The details of the selected yield models are as follows:

Zonal agromet yield models for four local districts with weather parameters

Agromet models- 1 to 4 refer to Hisar, Bhiwani, Sirsa and Fatehabad districts each with

37 years (1980-81 to 2017-18) of yield/weather data for a total sample size of n = 5 × 37.

Model 1: This model contains five weather variables, i.e., maximumand minimum temperatures, rainfall, relative humidity and sun shine hour.

$$Y = -1044.02 + 1.23Tr - 0.73ARF_3 + 30.21TMN_3 + 7.63RH_7 - 0.63RH_8 + 18.4SSH_6 - 11.52TMX_1$$
$$(225.93)\quad (0.09)\quad (0.42)\quad (6.87)\quad (1.59)\quad (0.22)\quad (5.80)\quad (4.44)$$

$R^2 = 0.64$, Adj.$R^2 = 0.63$, SE = 88.3

Model 2: This model contains five weather variables, i.e., maximum and minimum temperatures, relative humidity, sunshine hours and rainfall.

$$Y = -1268.08 + 1.27Tr - 0.70ARF_3 + 30.42TMN_3 + 8.84RH_7 - 0.84RH_8 + 19.4SSH_6 - 10.47TMX_1 - 0.45ARF_5$$
$$(242.48)\quad (0.09)\quad (0.41)\quad (6.96)\quad (1.65)\quad (0.23)\quad (5.72)\quad (4.38)\quad (0.19)$$

$R^2 = 0.66$, Adj.$R^2 = 0.64$, SE = 86.8

Zonal agromet yield model using weather parameters and CCT as dummy variable

Model 3:

$$Y = 192.66 + 93.2CCT - 3.22TMX_8 - 0.24ARF_3$$
$$(54.87)\quad (1.65)\quad (1.52)\quad (0.12)$$

$R^2 = 0.96$, Adj.$R^2 = 0.96$, SE = 26.6

Model 4:

$$Y = 223.68 + 92.39CCT - 4.30TMX_8 + 0.13ARF_4$$
$$(55.88)\quad (1.67)\quad (1.55)\quad (0.06)$$

$R^2 = 0.96$, Adj.$R^2 = 0.96$, SE = 26.5

Zone comprised of Hisar, Bhiwani, Sirsa and Fatehabad districts. Figures in parentheses indicate the standard error and all the regressors are significant at $p \leq 0.05$ in above zonal yield models.

Here; TMX1,...,TMX10, TMN1,….., TMN10,RH1,…...RH10, ARF1,…...ARF10, SSH1,.,SSH10 refer to 10 different fortnights under the crop growth period, CCT is the crop condition term and the other terms are the same as has already been defined.

3. DISCUSSION

Zonal agromet yield models incorporating linear timetrend and fine-resolution weather variables representing successive fortnights within the cotton crop growth period are developed and have been used to predict the cotton yields in four districts of Haryana. Several weather variables were identified as significant predictors of cotton crop yield. Adding linear time-trend to the model, along with the selected weather variables did not significantly improve the accuracy of the yield predictions. The predicted yields based on the agromet models selected by stepwise regression analysis had rather wide per cent deviations from the DOA crop yield estimates, sometimes wider than acceptable limits and thus were found unusable for yield forecasting purposes.

So, we have attempted to improve the predictive accuracy of the zonal yield models by identifying and adding additional covariate to the model, along with the selected weather variables. In particular, adding a crop condition term to the models with the selected weather variables and repeating the stepwise regression analysis substantially improved the predictive accuracy of the models and produced what we consider to be quite satisfactory district-level yield predictions using model 3 (Table 1 to 4). The goodness-of-fit of the model was assessed using residual diagnostics, in particular histograms and normal-probability plots of

residuals as well as the plot of residuals against the fitted values (Fig. 5). Hence, for this empirical study, we recommend using CCT as a dummy variable in addition to the weather variables to enhance the predictive accuracy of the selected zonal yield models.

It is worth emphasizing that the weather variables and linear time trend considered in this study may not suffice as a basis for accurate forecasting of cotton crop yields. It is therefore useful to explore and identify additional suitable agronomic/ biometrical variables that may further enhance the predictive accuracies of the models as we did with CCT for the district-level yield predictions. It would also be worthwhile exploring if other summaries of the weather variables using alternative time windows besides the fortnight summaries we used, may improve the model's performance.

4. CONCLUSIONS

The zonal agromet (or weather) models have been developed for pre-harvest yield forecasts of cotton crop in Hisar, Bhiwani, Sirsa and Fatehabad districts (which contribute > 90% cotton production) of the Haryana State.

The predictive performance(s) of the zonal weather-yield models were compared on the basis of adj-R^2, percent deviation of fitted yield from the real-time yield, mean. The adequacy of the fitted models was examined through histogram, normal-probability plot for the residuals and residual plot against fitted values for the selected models. The regression models with sufficiently good fit, as measured by coefficient of determination (R^2), couldn't provide the satisfactory predictive accuracy. The yield(s) estimated by zonal weather-yield models had sometimes higher percent deviations from the real-time yield(s) *i.e.*too high than considered to be tolerable for reliable yield prediction in the districts under consideration. However, the objective of cotton yield modelling was to assess the predictive accuracy of the developed model(s) for estimating crop yield and how the accuracies are influenced by the adopted procedures.

An attempt was further made to improve the predictive ability of the developed models by adding trend yield based crop condition terms to the zonal weather-yield model and that significantly improved the predictive accuracy of forecast models and produced quite satisfactory district-level cotton yield(s) estimation. The CCT is an indicator variable generated by dividing the trend predicted yield series into different non-overlapping classes. Hence, based on this study, using CCT as a categorical covariate is proposed in addition to the weather parameters to enhance the predictive accuracy of zonal weather–yield models. The results showed that the district-level yield(s) prediction gives good agreement with DOA yield estimates. The average absolute percent deviations of post-sample period forecasts falling between 4-9 percent favour the use of developed models for cotton yield prediction in western zone of Haryana.

The predictive performance(s) of the zonal agromet yield models were assessed using multiple metrics, including the adjusted R^2, the per cent deviations of the forecast yields from the DOA yield estimates.

Seeing the pattern of different statistic(s) and computational ease, model 4 is finally selected for pre-harvest cotton yield prediction in the districts under consideration.

Addition of CCT to the models containing weather variables significantly improved the accuracies of the district-level cotton yield predictions in the State.

This empirical study has produced a model with adequate accuracy for pre-harvest crop yield estimation, however, further work may be explored by identifying additional suitable agronomic/biometrical variables that may further enhance the predictive ability of the model as has been done by incorporating CCT for district-level yield prediction in western agro-climatic zone of Haryana.

ACKNOWLEDGEMENTS

The weather data received from Haryana Space Applications Centre, Hisar and Department of Agrometeorology, CCS HAU, Hisar, India are gratefully acknowledged. The authors would like to thank the learned reviewers for their valuable suggestions and comments to improve the earlier version of this article.

REFERENCES

Deepankar, Aneja, D.R. and Kumar, A. (2020) Impact of weather parameters on rice yield in Karnal district. *Journal of Pharmacognosy and Phytochemistry*, **9**(6), 633-636.

Draper, N.R. and H. Smith (1981). Applied Regression Analysis, 2nd edition (New York: Wiley).

Efroymson, M.A. (1960) Multiple Regression Analysis Mathematical Methods for Digital Computers (Ralston, A. and Wilf, H.S. ed.),1, Wiley, New York.

Gregory, C.C. (1960) Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**(3), 591-605.

Huda, A.K.S., Ghildyal, B.P., Tomar, V.S. and Jain, R.C. (1975) Contribution of climate variables in predicting rice yield. *Agricultural Meteorology*, **15***,* 71-86.

Jain, A., J.L. Chaudhary, M.K., Beck and Love Kumar (2019) Developing regression model to forecast the rice yield at Raipur condition. *Journal of Pharmacognosy and Phytochemistry*, **8**(1), 72-76.

Lal, G. and Pandey, K.K. (2017) Development of pre-harvest forecasting model for Chhattisgarh plain zone on pigeon pea by stepwise regression analysis. *International Journal of Current Microbiology and Applied Sciences*, **6**(11), 9-20.

Verma, U., Piepho, H.P., Ogutu, J.O., Kalubarme, M.H. and Goyal, M. (2014) Development of agromet models for district-level cotton yield forecasts in Haryana State. *International Journal of Agricultural and Statistical Sciences*, **10**(1), 59-65.

Yoshida, S. (1981) Fundamental of rice crop science, International Rice Research Institute, 933, *Manila, Philippines*.