Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 77(1) 2023 59–69

Weather based Models for Pre-harvest Crop Yield Forecasting

Ranjana Agrawal[1], K.N. Singh[1], Achal Lama[1], Bishal Gurung[1], Md. Ashraful Haque[1], K.K. Singh[2] and Priyanka Singh[2]

[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*India Meteorological Department, New Delhi*

Received 06 March 2023; Revised 20 April 2023; Accepted 23 April 2023

SUMMARY

This article presents the state of art review in the domain of weather based pre-harvest crop yield forecasting models. To begin with, it discusses the various approaches evolved over time for construction of weather indices from weather variables. Owning to complex relationship between crop yield and weather variables, various models have been proposed in literature. It has been documented starting with widely accepted weather based regression model followed by its improvements in variable selection and relaxation of assumptions. LASSO, ANN and random forest have been proposed as an improved method for variable selection, whereas Bayesian framework provides a way out when the data fails to satisfy the usual regression assumptions. The other developments such as use of complex polynomial through GMDH, principal component regression, discriminant function analysis and water balance technique have also been discussed. Further, we have briefly documented a webtool named Weather Indices based Automated Yield Forecasting System (WIAYFS) which has been developed for ease of implementation and reaching out to more researchers and users.

Keywords: Yield forecast models, weather variables, LASSO, ANN, Bayesian regression, WIAYFS.

1. INTRODUCTION

Crop yield is affected by technological change and weather variability for a given location. It can be assumed that the technological factors will increase yield smoothly through time and, therefore, year or some other parameter of time can be used to study the overall effect of technology on yield. Weather variability both within and between seasons is the second and the only uncontrollable source of variability in yields. Weather variables affect the crop differently during different stages of development. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into fine intervals. This will increase number of variables in the model and in turn a large number of constants will have to be evaluated from the data. This will require a long series of data for precise estimation of the constants which may not be available in practice.

1.1 Model Development

Fisher (1924) and Hendricks and Scholl (1943) have done pioneering work in crop weather relationship. They have given models which require small number of parameters to be estimated while taking care of distribution pattern of weather over the crop season.

Fisher assumed that the effect of change in weather variable in successive periods would not be an abrupt or erratic change but an orderly one that follows some mathematical law. He assumed that these effects are composed of the terms of a polynomial function of time. Further, the value of weather variable in w-th week, X_w was also expressed in terms of orthogonal functions of time.

$$A_w = a_0[f_0(w)] + a_1[f_1(w)] + \ldots\ldots + a_k[f_k(w)] \quad (1)$$

Corresponding author: Ranjana Agrawal
E-mail address: ranjana201@gmail.com

$$X_w = p_0 [f_0(w)] + p_1 [f_1(w)] + \ldots\ldots + p_k [f_k(w)] \quad (2)$$

Where p_i's are distribution constants.

Substituting (1) and (2) in usual regression equation

$$Y = A_0 + A_1X_1 + A_2X_2 + \ldots.. + A_nX_n + e \ldots \quad (3)$$

(Here Y denoted yield and X_w rainfall in wth week, w = 1,2,....,n) and utilising the properties of orthogonal and normalized functions, he obtained

$$Y = A_0 + a_0\rho_0 + a_1 \rho_1 + a_2\rho_2 + \ldots.. + a_k\rho_k + e \quad (4)$$

where $A_0,a_0,a_1,a_2,\ldots.a_k$ are constants to be determined and ρ_i (i=1,…k) are distribution constants of X_w in (4). Fisher has suggested to use k = 5 for most of the practical situations. In fitting this equation for k = 5, the number of constants to be evaluated will remain 7, no matter how finely growing season is divided. This model was used by Fisher for studying the influence of rainfall on the yield of wheat.

Hendricks and Scholl (1943) have modified Fisher's technique. They divided the crop season into n weekly intervals and have assumed that a second degree polynomial in week number would be sufficiently flexible to express the effect of weather on yield in successive periods. Further, they used values of weather variables as such. Mathematically

$$A_w = a_0 + a_1w + a_2w^2$$

In particular, $A_1 = a_0 + 1.a_1 + 1^2.a_2$

$$A_2 = a_0 + 2.a_1 + 2^2.a_2$$

$$A_n = a_0 + n.a_1 + n^2.a_2$$

Substituting the expression for A_w in regression equation (4), the model was obtained as

$$Y = A_0 + a_0 \sum_w X_w + a_1 \sum_w w X_w + a_2 \sum_w w^2X_w + e \quad (5)$$

In this model number of constants to be determined reduces to 4, irrespective of n.

This model was extended for two weather variables to study joint effects.

The model obtained was

$$Y = A_0 + a_0 \sum_w X_{1w} + a_1 \sum_w w X_{1w} + a_2 \sum_w w^2 X_{1w} + b_0 \sum_w X_{2w} + b_1 \sum_w w X_{2w} + b_2 \sum_w w^2 X_{2w} + c_0 \sum_w X_{1w} X_{2w} + c_1 \sum_w w X_{1w} X_{2w} + c_2 \sum_w w^2 X_{1w} X_{2w} + e \quad (6)$$

Since the data for such studies extended over a long period of years, an additional variate T representing the year was included to make allowance for time trend.

Another important contribution in this field is by Baier (1977). He has classified the crop weather models in three basic types.

1. Crop growth simulation models
2. Crop weather Analysis models
3. Empirical statistical models

1.2 Crop-Growth simulation models

A crop growth simulation model may be defined as a simplified representation of the physical, chemical and physiological mechanisms underlying plant growth processes. If the basic plant processes production and distribution of dry matter and water relations are properly understood and modelled, the entire response of the plant to the environmental conditions can be simulated. Therefore, there is no need to differentiate between climatic regions, since the simulation model itself will show the limiting factors for growth. In humid climates with low temperature and radiation levels, the model will generally show the greatest response of yields to increase in total radiation received. In an arid and hot climate, it will show the greatest response to the distribution and total amount of precipitation.

Various time intervals can be introduced in simulation models, for example, in view of the daily cycle of many plant processes, hourly intervals are most practical. It is then assumed that the rate calculated for a particular moment does not change appreciably over a period of one hour. It is possible to evaluate thereby specific processes such as photosynthesis, transpiration or respiration for an hour and then accumulate the hourly rates over the day and the daily rates over the growing season in order to arrive at the total seasonal dry matter production or yield of economic products. Simulation can be most useful if the model accounts for most relevant phenomena and contains no false assumptions. Simulation provides an insight into cropweather relationships, explains why some factors are more important for yield than others, suggests factors likely to have statistical significance and provides the basis for new experiments on processes which appear to be important but are not yet sufficiently understood. Thus, the simulation approach does not replace the statistical approach, but is complementary to it.

1.3 Crop-weather analysis models

Crop weather analysis models are defined as the product of two or more factors, each representing the (simplified) functional relationship between a particular plant response (e.g. yield) and the variations in selected variables at different plant developmental phases. The overall effects, as expressed by the numerical values of the factors modify each other but are not additive as in the case of a multivariate linear regression equation. Such models do not require a formulated hypothesis of the basic plant and environmental process; thus, the input requirements are less stringent but the output information is more dependent on the input data and less detailed than in the case of simulation models. Therefore, cropweather analysis models are a practical research tool for the analysis of crop responses to weather and climate variations when only climatological data are available. Conventional statistical procedures are used in such models to evaluate the coefficients relating crop responses to climatological or derived agrometeorological data. A convenient time interval is one day, but in practice shorter or longer periods can also be used, provided the response characteristics of the crops do not change appreciably over the selected period in relation to the variable taken into consideration.

1.4 Empirical statistical models

In the empirical approach, one or several variables (representing weather or climate, soil characteristics or a time trend) are related to crop responses such as yield. The weighting coefficients in these equations are by necessity obtained in an empirical manner using standard statistical procedures, such as multivariable regression analysis. This statistical approach does not easily lead to an explanation of the cause and effect relationships but it is a very practical approach for the assessment or prediction of yields. The coefficients in such empirical models and the validity of the estimates depend to a large extent on the design of the model, as well as on the representativeness of the input data. If the soil and climate conditions and the cropping practices are fairly homogeneous over the area represented by the input data, or if soil and geography are properly weighted in the equations, then it can be expected that the estimates have practical significance for the assessment of the crop conditions or prediction of yields for the specific area in question.

Several Empirical Statistical models were developed all over the world. The independent variables included weather variables, agrometeorological variables, soil characteristics or some suitably derived indices of these variables. Water Requirement Satisfaction Index (WRSI), Thermal Interception Rate Index (TIR), Growing Degree Days (GDD) are some agroclimatic indices used in models. Southern Oscillation Index (SOI) has also been used with other weather variables to forecast crop yield (Ramakrishna *et al.* 2003). To account for the technological changes year variable or some suitable function of time trend was used in the model. Some workers have also used two time trends. Moving averages of yield were also used to depict the technological changes.

In contrast to empirical regression models, the Joint Agricultural Weather Information Centre employs the crop weather analysis models that simulate accumulated crop responses to selected agrometeorological variables as a function of crop phenology. Observed weather data and derived agrometeorological variables are used as input data.

M Frere and G.F. Popov (1979) used the method which utilises actual rainfall and climatological information for the calculation of water requirement of crops and in turn crop water balance. The method is based on a cumulative water balance established over the whole growing season for the given crop and for successive periods of 10 days or a week. The water balance is difference between precipitation received by the crop and the water lost by the crop and the soil. Based on water surplus and deficit they have calculated index. Initially the index is taken to be 100 and is modified in successive periods depending on the water surplus or deficits. This index has been shown to be directly related to yield and can give a very satisfactory and early qualitative estimation of yields in rainfed crops. It may be possible to derive quantitative estimations of yields also but these estimates will have to be based on the potential yield of crops which will depend on local environmental conditions and will vary from place to place. It may also be mentioned that the method is intended mainly for utilisation in developing countries, which mainly practice rainfed agriculture where inadequate availability of water to crop is the main constraint. Therefore, the method does not directly involve the temperature which conditions the growth of the crop. However, the temperature intervenes indirectly in three ways in the method of crop water balance assessment. Firstly, the effect of air temperature may be noticed in the length of the growing cycle which is generally directly dependent on temperature. Further air temperature intervenes directly in the calculation of potential evapotranspiration and in this respect influences the whole water balance. Finally, the external temperatures may be important in some climatic zones, particularly as regards frosts.

In India, major organisations involved in developing methodology for forecasting crop yield based on weather are IMD and ICAR-IASRI. The methodology adopted by IMD involves identification of significant correlations between yield and weather factors during successive overlapping periods of 7 to 60 days of the crop growing season. By analysing the correlation coefficients for statistical and phenological significance, the critical periods when the weather variables have significant effect on yield are identified. The weather variables in critical periods are used through multiple regression analysis to obtain forecast equations. Using this methodology models were developed for principal crops on meteorological subdivisions basis. Data from various locations are averaged to get the figures for meteorological subdivisions and these are utilized along with time trend to develop the forecast model. Monthly forecasts are issued from these models by taking the actual data upto time of forecast and normal for the remaining period. In some models yield Moisture Index, Generalised Monsoon Index, Moisture Stress, aridity anomaly Index are also used (Sarwade, 1988; Sarkar, 2002). Under Forecasting Agricultural output using Space, Agro-meteorology and Land based observations (FASAL) programme crop growth simulation models are also being attempted.

2. IASRI APPROACHES

2.1 Weather indices based models

At IASRI, the model suggested by Hendricks and Scholl has been modified (Agrawal et al 1980; 1983; Jain et al 1980) by expressing effects of changes in weather variables on yield in the successive periods as second degree polynomial in respective correlation coefficients between yield and weather variables. This will explain the relationship in a better way as it gives appropriate weightage to different periods. Under this assumption, the models were developed for studying the effects of weather variables on yield using complete crop season data whereas forecast model utilised partial crop season data. These models were found to be better than the one suggested by Hendricks and Scholl.

These models were further modified (Agrawal et al 1986) by expressing the effects of changes in weather variables on yield in successive periods as a linear function of respective correlation coefficients between yield and weather variables. As trend effect on yield was found to be significant, its effect was removed from yield while calculating correlation coefficients of yield with weather variables to be used as weights. Effects of second degree terms of weather variables were also studied. The results indicated that (i) the models using correlation based on yield adjusted for trend effect were better than the ones using simple correlations, (ii) inclusion of quadratic terms of weather variables and also the second power of correlation coefficients did not improve the model. This suggests that the following model can be used to study effects of weather on yield.

$$Y = A_0 + a_0 z_0 + a_1 z_1 + cT + e \tag{7}$$

Where

$$Z_j = \sum_{w=1}^{n} r_w^j X_w \quad ; \quad j = 0, 1$$

Here Y is yield, r_w is correlation coefficient between the weather variable in wth period(X_w) and yield (adjusted for trend effect) and e is error term in equation 7. The models were further extended for studying joint effects. Azfar *et al.* 2015 studied the effect of changes in climatic variables on rice production in Faizabad district, UP, India using above approach.

The forecast model has been developed using partial crop season data considering all weather variables simultaneously. The model finally recommended was of the form

$$Y = A_0 + \sum_{i=1}^{p} \sum_{j=0}^{1} a_{ij} Z_{ij} + \sum_{i \neq i'=1}^{p} \sum_{j=0}^{1} a_{ii'j} Z_{ii'j} + cT + e \tag{8}$$

where

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^j X_{iw} \text{ and } Z_{ii'j} = \sum_{w=1}^{m} r_{ii'w}^j X_{iw} X_{i'w} \tag{9}$$

$r_{iw}/r_{ii'w}$ is correlation coefficient of Y with ith weather variable/product of ith and i'th weather variable in wth period. m is period of forecast and p is number of weather variables used.

In this approach, for each weather variable, two types of indices were developed, one as simple total of weather values in different periods [un-weighted index -Z_{i0}] and the other one as weighted total [weighted index Z_{i1}] weights being correlation coefficients between yield /de-trended yield (if trend is present) and weather variable in respective periods. On similar lines, for studying joint effects, un-weighted and weighted indices for interactions were computed with products of weather variables (taken two at a time). Stepwise regression technique was used to select important indices in the model.

These models were used to forecast yield of rice and wheat in different situations, viz (i) rainfed area having deficient rainfall (rice), (ii) rainfed area having adequate rainfall (rice) and (iii) irrigated area (wheat). The results revealed that reliable forecasts can be obtained using this approach when the crops are 10-12 weeks old. This approach was also used to develop forecast model for sugarcane (Mehta, *et al.* 2000). However, these studies were carried out at district level and required a long series data of 25-30 years which are not available for most of the locations. Therefore, the study has been undertaken to develop the model on agro-climatic zone basis for rice and wheat by combining the data of various districts within the zone so that a long series could be obtained in a relatively shorter time. Previous year's yield, moving averages of yield and agricultural inputs were taken as the variables taking care of variation between districts within the zone. Year was included to take care of technological changes. Different strategies for pooling district level data for the zone were adopted. Results revealed that reliable forecasts can be obtained using this methodology at 12 weeks after sowing i.e. about 2 months before harvest. The data requirement reduced to 10-15 years as against 25-30 years for district level models. The study also revealed that forecast model will be appropriate to forecast the yield of zone even if data for some districts within the zone are not available at model development stage or at forecasting stage (Agrawal *et al.* 2001). The approach has been successfully used for forecasting yields of rice, wheat, sugarcane and potato for Uttar Pradesh. (Agrawal, *et al.* 2005, Mehta, *et al.* 2010). The methodology is being implemented by IMD under FASAL programme.

2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO technique has been used to select significant weather indices affecting the yield of wheat crop. (Singh, *et al.*, 2019).

LASSO is a regularization technique which reduces the number of predictors in a regression model and identifies important predictors. Lasso is a shrinkage estimator with potentially lower predictive errors than ordinary least squares and it also includes a penalty term that constrains the size of the estimated coefficients. Therefore, it resembles ridge regression. It generates coefficient estimates that are biased but the bias is small. Nevertheless, a lasso estimator can have smaller mean squared error than an ordinary least-squares estimator when you apply it to new data. Unlike ridge regression, as the penalty term increases, lasso sets more coefficients to zero. This means that the lasso estimator is a smaller model, with fewer predictors. As such, lasso is an alternative to stepwise regression and other model selection and dimensionality reduction techniques.

2.3 Feature Selection Approach

As crop yield prediction is a complex phenomenon and has many underline nonlinear patterns. Such, datasets are difficult to deal with stringent assumptions of the statistical models. Hence, machine learning (ML) techniques which have very few prior assumptions and are data driven provide great deal of flexibility for modelling and forecasting the crop yield.

Varma, *et al.* 2022 studied variable selection algorithms also called feature selection algorithms such as Forward Selection, Backward Selection, Random Forest Feature Selection, Least Absolute Shrinkage and Selection Operator (LASSO) and Correlation Based Feature Selection (CBFS) and have applied to three different datasets for crop yield. Regression forecasting models have been developed with selected features for all the algorithms. Machine learning techniques such as Random Forest Regression and Support Vector Regression are also applied. Also, Random Forest Regression and Support Vector Regression have been compared with the stepwise regression method. The Forecasting Performance of the proposed model was also evaluated using statistical measures such as Root Mean Square Error (RMSE), Mean Absolute Prediction Error (MAPE) and Mean Absolute Deviation (MAD).

A comparison has been made between all the feature selection algorithms and between all the machine learning algorithms. Finally, the machine learning technique has been compared with the regression model coupled with feature selection with the help of Random Forest. The CBFS has been found to be the best feature selection algorithm for prediction with a regression model. Random Forest regression was observed as the best machine learning technique. The machine learning techniques were found to be more efficient as compared to feature selection with the regression framework.

2.4 Complex Polynomial through Group Method of Data Handling (GMDH) technique

This methodology has been successfully applied by Mustafi and Chaudhuri (1981) for forecasting monthly tea crop production. At ICAR-IASRI use of this technique was explored for forecasting potato yield in Uttar Pradesh (Mehta, *et al.* 2010).

The main feature of this technique is that it itself selects the structure of the model without using a prior information about relationship of dependent variable (y) with independent variables (x_1, x_2, …….,x_p).

The fitted polynomial is of the form

$$y = a + \sum_{i=1}^{p} b_i x_i + \sum_{i,j=1}^{p} c_{ij} x_i x_j + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} d_{ijk} x_i x_j x_k + \dots \tag{10}$$

The technique involves fitting of quadratic equations for all pairs of independent variables and identifying a few best performers in terms of predictive ability (using appropriate statistic); converting entire set of independent variables (called zero generation variables) to new variables (first generation variables) which are obtained as predicted values from these selected quadratic equations (of zero generation variables). The process of fitting and identifying best quadratic equations is repeated using first generation variables and second generation variables are obtained. The whole process is repeated with every new generation of variables till appropriate model is obtained (using certain criteria). At final stage, one best quadratic equation is selected as the final model.

Two approaches are followed for identifying few best performers. In the first approach, available data set is divided into two non-overlapping sets - training set and checking set. From training set, quadratic equations are fitted and checking set is used to test the predictability of different quadratic equations using root mean square error. In the other approach, PRESS statistic (predicted sum of square) is used wherein the whole data set is used as fit and check set. For evaluation of PRESS, each data point (one by one) is taken as a testing set (of size 1) and is predicted from the quadratic equation fitted from the remaining n-1 data points. PRESS is calculated as

$$PRESS = \sum_{s=1}^{n} \left[y_s - \hat{y}_s^{(s)} \right]^2 \tag{11}$$

where y_s is the value of the dependent variable for s^{th} observation and $\hat{y}_s^{(s)}$ is the predicted value of y_s computed from an equation based on remaining n -1 data points.

Theoretically the generation of variables is continued till the decreasing trend of PRESS statistic ceases i.e. PRESS statistic starts increasing. Some times even after number of generations the decreasing trend continues. This increases complexity of the model. In practical situations when there is abrupt decrease in value of PRESS statistic and at the same time the coefficient of determination is quite high, the procedure is terminated.

This approach was used to obtain district as well as agroclimatic zone level models for potato in Uttar Pradesh using weather indices (unweighted and weighted) as explanatory variables. The performance of this model was found to be better than indices based regression approach for Bareilly district and north eastern zone. For remaining districts and zones the performance was worse or at par with the indices based regression approach.

2.5 Discriminant Function Technique

Discriminant function technique is a linear / quadratic function that discriminates different groups the best. Use of this technology has been explored for obtaining quantitative forecast of yields for rice in Raipur district. This methodology involved grouping the long series of years into three groups - congenial, normal and adverse with respect to crop yields (adjusted for trend effect, if any). Using weather data of these groups, linear / quadratic discriminant functions were obtained using phase-wise weather data. Weather scores for each year at different phases of crop growth obtained through these discriminant functions were used along with inputs and time trend as regressors in model development through stepwise regression. Quadratic discriminant function was found appropriate and the methodology could provide reliable forecast two months before harvest. (Rai and Chandrahas, 2000). The methodology was further modified using weekly weather data. Various strategies were proposed to solve the problem of number of variables more than number of observations. The study was carried out to forecast wheat yield in Kanpur district. The finally recommended strategy involved following steps; Discriminant functions were developed using weather data of first week. These discriminant functions were used to compute scores for each year. Taking data on weather variables in the second week and discriminant scores computed from the first week, discriminant function analysis was carried out which provided scores for each year based on data upto second week. The process was repeated for successive weeks' data till the time of forecast and finally discriminant scores based on entire data were obtained for each year which along with trend were used to develop the model through stepwise regression technique. In contrast to earlier model by Rai *et al.* (2000) this model was based on complete data upto the time of forecast and relative importance of weather variables in different weeks (Agrawal *et al.* 2012). A detailed study using the recommended strategy has been carried out for forecasting rice, wheat and sugarcane in Uttar Pradesh. Methodology was found to be successful for obtaining district / agroclimatic zone/state level forecasts. Performance was found to be better than weather indices based regression models for some cases whereas in some cases reverse trend was found (Chandrahas *et al.* 2010). However, this approach requires larger data base as compared to weather indices based regression approach. Use of discriminant function analysis has also been explored for pre-harvest forecasting of wheat yield (Sisodia *et al.* 2014) and pigeon pea yield (Yadav *et al.* 2016) in Faizabad district of Eastern UP India where above models were employed and also some variations in the models were proposed.

2.6 Ordinal Logistic Regression Model

Forecast model has been developed taking probabilities of different classes (generated through ordinal logistic model) as explanatory variables. Methodology was similar to that of discriminant function based model. This approach has been demonstrated through forecasting wheat in Kanpur. Results revealed that Ordinal Logistic Model performed better than Discriminant Function Approach (Kumari *et al.* 2016).

2.7 Principal Component Regression

Principal component regression is a well known technique to reduce number of explanatory variables in the model. The technique involves conversion of explanatory variables into a set of uncorrelated variables with variances in descending order (known as principal components). The whole variation of the system explained by explanatory variables is explained by first few principal components which are used as regressors in the model in place of original variables. Besides solving the problem of number of explanatory variables more than number of observations, the technique also solves the problem of multicollinearity. The approach has been attempted for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh but the approach was not found to be successful (Chandrahas *et al.* 2010). Further, studies were conducted for forecasting wheat and rapeseed & mustard yield in Faizabad district of Uttar Pradesh, India in which principal components of weather indices were used as regressors in the model. (Yadav *et al.* 2014, Afzar *et al.* 2015).

2.8 Bayesian Regression Model

The most extensively used technique for forecasting of crop yield is regression analysis. Significance of parameters is one of the major problems of regression analysis. Non-significant parameters lead to absurd forecast values and are not reliable. In such cases, models need to be improved. To improve the models prior knowledge can be incorporated through Bayesian technique (Yeasin, *et al.*, 2021). Bayesian technique is one of the most powerful methodologies in modern era of statistics. They have discussed different types of prior (informative, non-informative and conjugate priors). The Markov Chain Monte Carlo (MCMC) methodology has been briefly discussed for estimation of parameters under Bayesian framework. To illustrate these models, production data of banana, mango and wheat yield are taken. They have compared the traditional regression model with the Bayesian regression model and conclusively infered that the models estimated under Bayesian framework provide superior results as compared to the models estimated under classical approach. The main advantage of this technique of estimation is that it uses the prior information. By incorporating prior information, we get a posterior distribution. The posterior distribution contains more information due to incorporation of extra information in the form of prior distribution.

Prior distribution is simply a probability distribution which expresses the possible uncertainty before examination of current data. Main problem of Bayesian estimation is to find a suitable prior with its parameters. Irrelevant prior misleads the researcher as it gives spurious results. So, one should be careful at the time of selection of priors. There are various types of prior found in literature. These are informative or non-informative, conjugate or non-conjugate priors. Our interest lies on the conjugate priors because the form of posterior distribution remains same as its prior distribution, only hyper-parameters are updated. Exponential family of distribution is commonly used conjugate priors. In this study, most common distribution of exponential family i.e. normal distribution is taken as a prior distribution. Steps of Bayesian approach of regression estimation:

I. Model specification

II. Selection of prior distribution

III. Find the likelihood function

IV. Apply Bayes theorem and generate posterior distribution by the help of Markov Chain Monte Carlo (MCMC) Method

V. Find the expectation of the posterior distribution

2.9 Artificial Neural Network Technique

In contrast to regression approach, Artificial Neural Network (ANN) technique has been explored for forecasting yields of rice, wheat and sugarcane in Uttar Pradesh. (Kumar *et al.* 2010).

This is an attractive tool under machine learning techniques for forecasting and classification purposes. ANNs are data driven self-adaptive methods in that there are few apriori assumptions about the models for problems under study. These learn from examples and capture subtle functional relationships among the data even if the underlying relationships are unknown or hard to describe. After learning from the available data, ANNs can often correctly infer the unseen part of a population even if data contains noisy information. As forecasting is performed via prediction of future behaviour (unseen part) from examples of past behaviour, it is an ideal application area for ANNs, at least in principle. (Dewolf *et al.* 1997, 2000). However, the technique requires a large data base.

2.10 Water Balance Technique

After the introduction of the concept of potential evapotranspiration (PET) by Thornthwaite and Penman in 1948 use of agrometeorological variables in yield assessment models became popular. Performance of models using agrometeorological variables was found to be better than the models using only meteorological variables (Baier and Robertson 1968). This is due to the fact that agrometeorological models use variables like soil moisture, actual and potential evapotranspiration, crop water requirement, effective rainfall etc. which take into account the soil properties and crop characteristics in addition to meteorological variables. These variables are estimated using Water Balance technique. This technique is useful for rainfed crops.

The water balance technique or model is a simple equation keeping an account of receipt and expenditure of water from the soil reserves. Mathematically

$$S_i = S_{i-1} + R_i - E_i \quad \dots (12)$$

where S_i, R_i and E_i represent soil moisture, water received and expenditure respectively at the end of i^{th} period. The receipt (R) is in the form of rainfall and expenditure (E) is in the form of evapotranspiration from soil and crop cover. The stress depends on the demand and availability. When demand is more than the availability the stress occurs. Degree of stress depends on the gap between demand and availability. The variation in the model used by different research workers is in the form of variables used for indicating receipt and expenditure and the limitation imposed on these variables. For example in this equation, i represents the period after which balancing is done. It can be a day, week, fortnight or month. The number of periods depend on the length of the crop season, starting from the time of sowing. The depth of root zone can be kept constant for the entire season or can vary with the age of the crop. Receipt can be in the form of rainfall assuming that it is absorbed by the soil at a constant rate irrespective of the intensity of rain and antecedent moisture condition of soil. It can be made more realistic by using the actual infiltration rate of water for a given soil under a given intensity of rain and antecedent moisture condition. Similarly, expenditure which depends on demand can be estimated in different ways. It can simply be PET, the amount that can potentially evapotranspirate under a given climatic condition or a fraction of it i.e. PET/2 or PET/4, irrespective of crop, and its stage of growth. This form of expenditure/demand is used when the objective is crop planning, estimation of drought proneness or agroclimatic classification of an area. For the purpose of monitoring and assessing yield of a crop, actual water requirement (WR) of the crop is taken as demand. The water requirement (at different time points) is estimated by multiplying the crop coefficients k (representing the crop characteristic and stage of the crop growth) and the value of evaporation (representing the loss of water to climatic condition such as temperature, wind velocity, humidity and sunshine at that stage). Estimates of evaporation are obtained either directly from an evaporimeter or from different formulae developed by Thornthwaite (1948), Penman (1948) and Christiansen (1966). For estimates obtained from different formulae different values of k are used. The values of k for different crops or groups of crops under given climatic conditions are developed by agronomists and water technology scientists by conducting field experiments.

If the receipt is more than the expenditure, the excess water is stored in the soil depending upon the water holding capacity and water already stored in the soil. If excess water is more than the retention capacity of the soil, it goes waste as runoff. Thus, in the process of estimating soil moisture an estimate of runoff is also obtained. The amount of rain water used by plants and stored in the soil is taken as effective rainfall (ER), when rainfall is not enough to meet the crop water requirement, the requirement is met from the soil moisture stored in the rootzone. When rain and soil moisture together are not enough to meet the requirement the amount actually available is extracted by the plants and part of requirement remains unfulfilled. In this manner an estimate of actual evapotranspiration (AE) is obtained. When AE is less than WR a stress to the crop occurs. 1-(AE/WR) is used as estimate of stress to the crop. The ratio AE/PE is also sometimes used as variable for yield assessment. Since the final yield is the outcome of the aggregate of water/moisture availability or nonavailability through its life cycle an accumulated stress or satisfaction index is prepared. As the deficiency at critical stages causes greater damage, weights are assigned to stress at different stages according to their importance. In this manner an accumulated weighted stress index (SI) or water requirement satisfaction Index (WRSI) is obtained for each season. The index is related to yield through a regression model. The accumulated index helps in crop monitoring right from the date of sowing and provides preharvest estimate of yield during any time of the season. Therefore, prerequisites for a seemingly simple model are estimates of many parameters of soil like depth, water holding capacity, wilting point, field capacity, infiltration rate under different antecedent moisture condition and crop parameters like crop coefficients, rooting pattern, and water extraction pattern of roots. Also a knowledge of critical stages of growth, an insight into effect of stress at different stages and an estimate of initial soil moisture are necessary. The success of the model depends on the accuracy of estimates of these parameters.

In a study conducted at IASRI a water balance model of the following form was used

$$S_i = S_{i-1} + ER_i - WR_i \tag{13}$$

for estimating moisture stress to the pearlmillet crop of IARI, New Delhi. Balancing was done at the end of each day of the crop season. Depth of root zone varied with the age of the crop. A new method was developed to calculate effective rainfall (ER) on the basis of amount of rain and antecedent moisture condition. ER was used in the model in place of rainfall. Estimates of moisture stress and moisture surplus were obtained from the model. Weights were assigned to stress at different stages. Detrimental effect of excess water in the rootzone was also taken into account and weights were assigned to surplus water also depending upon the time of its occurrence. An accumulated stress and surplus moisture index (SI) was prepared. SI was related to yield through a regression equation. The fitted equation explained 91% variation in pearlmillet yield. A reduction of 42.7 kg/ha was expected due to per unit of stress in the potential yield of 3000 kg/ha. The error in predicted yield of two years were was 3.2% and 0.5% respectively (Saksena and Bhargava 1995).

In another study, models were developed for rainfed sorghum, maize and rice using agrometeorological indices. Water balance was carried out at weekly intervals. Weighted stress index was prepared phase−wise by applying weights to surplus as well as deficit moisture depending upon the stage at which it occurred. Stress index of phase 2 i.e. 4 to 7th weeks after sowing played an important role in determining the yield of sorghum both at Delhi and Parbhani (Maharashtra) district. Models with phase 2 Index and trend variables as regressors could forecast yield 6 weeks before harvest. Similarly, for maize crop at Delhi model with trend, index of surplus moisture at phase 1 and 2 and deficit moisture at phase 3 and 4 as predictor variables could forecast yield 4 weeks before harvest. Model for rice in Raipur district included trend and accumulated index for the five phases up to maturity as explanatory variables (Saksena *et al.* 2001).

3. WEATHER INDICES BASED AUTOMATED YIELD FORECASTING SYSTEM (WIAYFS)

For ease of implementation and reaching out to more researchers and users, WIAYFS (Weather Indices based Automated Yield Forecasting System), a webtool has also been developed at ICAR-IASRI. In the webtool stepwise regression model based on weather indices along with other models such as ARIMAX, LASSO regression, Bayesian Regression model and Random Forest technique are included. The process of calculation of weather indices from raw weather data along with fitting of various models has been automized. WIAYFS is being implemented by IMD (India Meteorological Department), New Delhi.

REFERENCES

Agrawal, R., Chandrahas, and Aditya, K. (2012). Use of discriminant function analysis for forecasting crop yield, Mausam, **63**(3), 455-458.

Azfar, M., Sisodia, B.V.S., Rai, V.N. and Kumar, S. (2015). A Study on Impact of Changes in Climatic variables on Rice Production in Faizabad district of Uttar Pradesh, Mausam, **66**(22), 281-286.

Azfar, M., Sisodia, B.V.S., Rai, V.N. and Devi, M. (2015). Pre-harvest forecast models for rapeseed & mustard yield using principal component analysis of weather variables. Mausam, **66**(4), 761-766.

Agrawal, R., Jain, R.C. and Jha, M.P. (1983). Joint effects of weather variables on rice yield, Mausam, **34**(2), 18994.

Agrawal, R., Jain, R.C. and Jha, M.P. (1986). Models for studying rice cropweather relationship, Mausam, **37**(1), 6770.

Agrawal, R., Jain, R.C. and Mehta, S.C. (2001). Yield forecast based on weather variables and agricultural inputs on agroclimatic zone basis. Indian Journal of Agricultural Sciences, **71**(7), 487-490.

Agrawal, R., Jain, R.C. Jha, M.P. and Singh, D. (1980). Forecasting of rice yield using climatic variables. Indian Journal of Agricultural Sciences, **50**(9), 68084.

Agrawal, R., Ramakrishna, Y.S. Kesava Rao, A.V.R. Kumar, A, Bhar, L.M., Madan M. and Saksena, A. (2005). Modeling for forecasting of crop yield using weather parameter and agricultural inputs. (AP Cess Fund project report).

Baier, W. (1977). Crop weather models and their use in yield assessments. Tech. note no. 151, WMO, Geneva, 48 pp.

Baier, W., Robertson, G.W. (1968). The performance of soil moisture estimates as compared with direct use of climatological data for estimating crop yields. *Agric. Meteorol.*, **5**, 731.

Chandrahas, Agrawal, R. and Walia, S.S. (2010). Use of discriminant function and principal component techniques for weather based crop yield forecasts. (IASRI, New Delhi).

Christiansen, J.E. (1966). Estimating panevaporation and evapotranspiration from climatic data. *Irrigation and drainage conference. Las Vegas, Nev. Nov. 24.*

Dewolf, E.D. and Francl, L.J. (1997). Neural network that distinguish in period of wheat tan spot in an outdoor environment. Phytopathalogy, **87**(1), 83-87.

Dewolf, E.D. and Francl, L.J. (2000). Neural network classification of tan spot and stagonespore blotch infection period in wheat field environment. Phytopathalogy, **20**(2), 108-13.

Fisher, R.A. (1924). The influence of rainfall on the yield of wheat at Rothamsted. Roy. Soc. (London), Phil. Trans. Ser.B., 213 : 89142.

Frere, M. and Popov, G.F. (1979). Agrometeorological crop monitoring and forecasting. FAO plant production and protection paper, plant production and protection division, FAO, Rome.

Hendrick, W.A. and Scholl, J.C. (1943). Technique in measuring joint relationship. The joint effects of temperature and precipitation on crop yield. N. Carolina Agric. Exp. Stat. Tech. Bull. 74.

Jain, R.C., Agrawal, Ranjana and Jha, M.P. (1980). Effects of climatic variables on rice yield and its forecast, Mausam **31**(4), 59196.

Kumar, Amrender, Ramasubramanian, V. and Agrawal, Ranjana (2010). Neural network based forecast modeling crops. (IASRI Publication).

Kumari, Vandita, Agrawal, Ranjana and Kumar Amrender(2016). Use of ordinal logistic regression in crop yield forecasting, Mausam, **67**(4). 13-18.

Mehta, S.C., Agrawal, Ranjana and Singh V.P.N. (2000): Strategies for composite forecast. JISAS, **53**(3).

Mehta, S.C., Satya Pal, Kumar, Vinod (2010). Weather based models for forecasting potato yield in Uttar Pradesh. (IASRI Publication).

Mustafi, A. and Chaudhary, A.S. (1981), Use of Multilayer Group Method of Data Handling for prediction of tea crop production, Jour. Ind. Soc. Agril. Stat, **33**(1), 89-101.

Rai, T. and Chandrahas (2000). Use of discriminant function of weather parameters for developing forecast model for rice crop. (IASRI Publication).

Ramakrishna, Y.S.; Singh, H.P. and Rao, G. Nageswara (2003). Weather based indices for forecasting foodgrain production in India. Journal of Agrometeorology **5**(1), 1-11.

Saksena, A., Jain, R.C., Yadav, R.L. (2001). Development of early warning and yield assessment model for rainfed crops based on agrometeorological indices. (IASRI publication).

Sarkar, J. (2002). Forecasting rice and wheat yield over different meteorological sub–divisions of India using statistical models. 56th Annual Conference of ISAS.

Sarwade, G.S. (1988). Meteorological yield models. IRS-UP/SAC/CYM/TN/17/58, SAC, Ahmedabad.

Singh, K.N., Singh, K.K., Kumar, S., Panwar, S. and Gurung, B. (2019). Forecasting crop yield through weather indices through LASSO. *Indian Journal of Agricultural Sciences*. **89**, 540-544.

Sisodia, B.V.S., Yadav, R.R., Kumar, S. and Sharma, M.K. (2014). Forecasting of pre-harvest crop yield using discriminant function analysis of agro meteorological parameters. Journal of Agro-Meteorology, **16**(01), 121-125.

Thornthwaite, C.W. (1948). An approach towards rational classification of climate. Geographical Review. **38**, 5594.

Varma, M., Singh, K.N. and Lama, A. (2022). Exploring the suitability of machine learning algorithms for crop yield forecasting using weather variables. *Journal of Crop and Weed*, **18**, 210-214.

Yadav, R.R., Sisodia, B.V.S. and Kumar, S. (2014). Application of principal component analysis in developing statistical models to forecast crop yield using weather variables. Mausam, **65**(3), 357-360.

Yadav, R.R., Sisodia, B.V.S. and Kumar, S. (2016). Pre–harvest forecast of Pigeon –pea yield using discriminant function analysis of weather variables. Mausam, **67**(3), 577-582.

Yeasin, M., Singh, K.N., Lama, A. and Gurung, B (2021). Improved weather indices based Bayesian regression model for forecasting crop yield. *Mausam*, **72**(4), 879-886.