Thimmegowda *et al.*
Journal of Cotton Research (2025) 8:6
https://doi.org/10.1186/s42397-024-00208-8

Journal of Cotton Research

RESEARCH

Open Access

Comparative analysis of machine learning and statistical models for cotton yield prediction in major growing districts of Karnataka, India

THIMMEGOWDA M. N.[1], MANJUNATHA M. H.[1], LINGARAJ H.[1], SOUMYA D. V.[1*], JAYARAMAIAH R.[1], SATHISHA G. S.[1] and NAGESHA L.[1]

Abstract

Background Cotton is one of the most important commercial crops after food crops, especially in countries like India, where it's grown extensively under rainfed conditions. Because of its usage in multiple industries, such as textile, medicine, and automobile industries, it has greater commercial importance. The crop's performance is greatly influenced by prevailing weather dynamics. As climate changes, assessing how weather changes affect crop performance is essential. Among various techniques that are available, crop models are the most effective and widely used tools for predicting yields.

Results This study compares statistical and machine learning models to assess their ability to predict cotton yield across major producing districts of Karnataka, India, utilizing a long-term dataset spanning from 1990 to 2023 that includes yield and weather factors. The artificial neural networks (ANNs) performed superiorly with acceptable yield deviations ranging within ± 10% during both vegetative stage (F1) and mid stage (F2) for cotton. The model evaluation metrics such as root mean square error (RMSE), normalized root mean square error (nRMSE), and modelling efficiency (EF) were also within the acceptance limits in most districts. Furthermore, the tested ANN model was used to assess the importance of the dominant weather factors influencing crop yield in each district. Specifically, the use of morning relative humidity as an individual parameter and its interaction with maximum and minimum temperature had a major influence on cotton yield in most of the yield predicted districts. These differences highlighted the differential interactions of weather factors in each district for cotton yield formation, highlighting individual response of each weather factor under different soils and management conditions over the major cotton growing districts of Karnataka.

Conclusions Compared with statistical models, machine learning models such as ANNs proved higher efficiency in forecasting the cotton yield due to their ability to consider the interactive effects of weather factors on yield formation at different growth stages. This highlights the best suitability of ANNs for yield forecasting in rainfed conditions and for the study on relative impacts of weather factors on yield. Thus, the study aims to provide valuable insights to support stakeholders in planning effective crop management strategies and formulating relevant policies.

Keywords Cotton, Machine learning models, Statistical models, Yield forecast, Artificial neural network, Weather variables

*Correspondence:
Soumya D. V.
dvsoumya@gmail.com
Full list of author information is available at the end of the article

Introduction

Cotton (*Gossypium hirsutum* L.) is one of the most important and widely produced commercial crop in the world (Aslam et al. 2020) and is cultivated mainly in tropical regions under rainfed conditions. With respect to acreage, India ranks first in the world (13.06 million hectares) and second in production, with an annual production of 34.34 million bales (170 kg per bale in India). This crop is produced as a raw material for the textile industry and has important uses in medicine and automobile industries. As cotton is cultivated mainly in rainfed regions, its production dynamics rely on the dynamics of weather factors, which prevail during the crop growth period. Although the final yield of the crop is dependent on the interaction between the genotype and environment, the dynamics of weather and resource availability (water and nutrients) play a pivotal role in the yield potentials. The crop has shown a positive response to solar radiation and temperature under optimum availability of other resources like soil moisture and nutrients (Mao et al. 2019). Biomass formation is also closely related to the accumulation of resources such as photosynthetically active radiation (PAR), effective temperature accumulation, and soil water content (SWC) during the crop growth period (Wu et al. 2022). In addition to the variability in above ground microclimate, variations in soil temperature and moisture are known to impart root growth, which in turn affects above ground biomass formation (Tang et al. 2010). Many studies have been conducted to prove the intimate association of weather factors with the metabolism and physiological activation of cotton phenology (Wang et al. 2019). There is a projection of resource scarcity in the future because of changing climate, so improving the utilization of light, temperature and water resources is highly important for sustainable production of cotton (Howden 2008). A key for this purpose is to advance the prediction of crop performance and estimate the major factors impacting growth and yield using theoretical and applied techniques.

Most yield forecasting studies have focused on food crops, such as wheat (Kogan et al. 2013), and rice (Wang et al. 2010), but there has also been considerable interest in forecasting important fibre crops, such as cotton (Baigorria et al. 2010). As these crops are cultivated under natural/rainfed conditions, weather-based crop yield forecasting is essential for shaping policies related to supply, trade, and production exchange (Dharmaraja et al. 2020). The reliability of such weather-based crop yield forecasting depends on the choice of model, its input requirements (Hara 2021; Chipanshi 2015), and the objective evaluation of model performance. Many techniques have been developed to forecast growth and to make the best forecast of cotton area, production, and yield in different cultivation conditions of India, but their suitability depends on the ability of the model to describe the observed data. Crop simulation models and statistical models are two broad approaches to yield forecasting (Bocca et al. 2016). Crop simulation models offer detailed insights into crop biology through their reliance on extensive data such as soil, plant, and weather data. However, these models often face challenges due to limited data availability. In response to these difficulties, statistical models based on weather parameters have been developed to provide reliable crop acreage estimation and yield predictions (Sharma et al. 2018). Although statistical models provide forecasts with reasonable precision, the calibration and testing of those models using historical datasets are crucial. Multiple linear regressions (MLRs) are commonly used statistical crop yield prediction models (Rai et al. 2013; Dhekale et al. 2014; Kumar et al. 2014). Vashisth et al. (2018) conducted studies on maize at the flowering stage and the grain filling stage with weather-based statistical model. However, there are chances of model over-fitting when the number of samples is lesser than the number of predictors and the existence of multicollinearity among independent factors (Verma et al. 2016). To overcome such discrepancies, feature selection techniques such as stepwise multiple linear regression (SMLR), least absolute shrinkage and selection operator (LASSO), elastic net (ENet) or feature extraction such as principal component analysis (PCA) statistical techniques are used (Das et al. 2017) in forecasting yields in many crops (Paswan et al. 2013; Das et al. 2018; Bali et al. 2021), showcasing their enhanced effectiveness in yield forecast. Statistical models have the potential to expand the scope of advance yield estimation and examine more crop types, particularly for those where established process-based models are lacking due to a scarcity of crop-specific parameters. Going beyond major crops to include more will provide a better picture of future global food availability under climate change (Hu et al. 2024).

Considering the above background, this study was planned based on long term weather and yield datasets spanning from 1990 to 2021 with a motive of identifying the most suitable forecasting technique for predicting cotton yield in major production districts of Karnataka, India, estimating the range of variability in cotton yield as predicted by different models, the factors that limit the ability of a model to predict the yield, options to overcome these limitations, etc. The weather-induced production variability impacts regional food security, thus it is necessary to study the major weather factors behind crop production.

Materials and methods

Study districts

Among the top cotton producing states in India, Karnataka stands the fourth, with an area accounting for 7% of

the area under cotton in the country and the production accounting for 4%, with a productivity of 653 kg·hm^{-2}. The top ten districts of Karnataka were selected for the study, including Ballari, Belagavi, Chitradurga, Dharwad, Haveri, Kalaburagi, Koppal, Mysuru, Raichur, and Vijayapura, which collectively contribute to approximately 70% of the state's cotton area and production. The rainfall and rainy days of ten districts are presented in Supplementary Table 1. In 2021, Kalaburagi emerged as a significant contributor to cotton cultivation, leading to an area of 67 065 hm^2, producing 205 930 bales and achieving a higher productivity of 522 kg·hm^{-2}. Raichur, with an extensive cultivation area of 169 518 hm^2, also played a substantial role, producing 427 785 bales of cotton at a productivity of 429 kg·hm^{-2} (Table 1). In 2023, a notable change in the cotton area was observed across these districts. Raichur contributed the most to the cotton cultivation area (179 701 hm^{-2}), followed by Kalaburagi. These districts were the focus of the study for forecasting cotton yield using different models, considering their significant contribution to the state's cotton production.

The cotton yield distribution across ten districts from 1990 to 2021 is shown in Fig. 1, highlighting notable variations in yield patterns. Raichur has the highest variability whereas Mysuru, Dharwad, Chitradurga, and Belagavi display moderate variability in yield. Koppal exhibits relatively low variability in yield, suggesting a more stable and consistent yield pattern. The median of yield is positioned towards the upper end of the range, indicating a tendency for higher yields. In Kalaburagi, the median leans towards the upper end, signifying wide variability in yield indicating a mix of higher yields but with notable variability. Vijayapura showes a wide interquartile range, indicating significant variability in yield. The median yield is toward the lower end, suggesting that the majority of yields are below the median.

Datasets sources

Long term (from 1990 to 2021) dataset on planted area, production, and productivity of cotton during kharif (crop sown during south-west monsoon season grown under rainfed conditions) season in major growing districts of Karnataka was sourced from the Directorate of Economics and Statistics, Government of Karnataka (https://des.karnataka.gov.in). The datasets were checked for the presence of outliers, *i.e.*, extremes, and were detrended based on their regression with time factor. After each step of detrending, the significance was checked and if there was no significant yield change with time, again the same process was continued until there was an observed significant change in yield with time. A dataset pertaining to daily weather parameters, i.e., maximum and minimum temperature, morning and evening relative humidity, and rainfall, pertaining to the study years, was sourced from the India Meteorological Department (https://mausam.imd.gov.in) using the inverse distance weightage method.

Calculation of weather indices

To formulate a composite model that considers the individual and interactive impacts of weather variables, a set of independent factors such as weather variables and weather indices were calculated. These factors can be classified into two categories: unweighted and weighted weather variables. Unweighted weather variables representing direct observation and weighted weather variables were calculated to account for the interactive impact of weather factors on crop performance. To account for yield variability due to both sole and interactive effects of weather factors forecasting models that depend on both unweighted/individual and unweighted/interactive weather factors are often employed for predicting yields in crops like rice, wheat, sugarcane and potato (Manideep 2022; Mehta

Table 1 The planted area, production, and productivity of cotton in top ten districts of Karnataka

District	2021			2023
	Planted area /(hm^2)	Production/bales	Yield / (kg·hm^{-2})	Planted area /(hm^2)
Ballari	32 812	98 823	512	26 926
Belagavi	22 435	54 768	415	16 918
Chitradurga	108 22	23 363	367	8 715
Dharwad	53 479	101 609	323	30 573
Haveri	25 572	60 771	404	19 298
Kalaburagi	67 065	205 930	522	104 525
Koppal	13 081	28 778	374	15 870
Mysuru	25 570	33 993	226	36 312
Raichur	169 518	427 785	429	179 701
Vijayapura	35 046	99 159	481	53 676

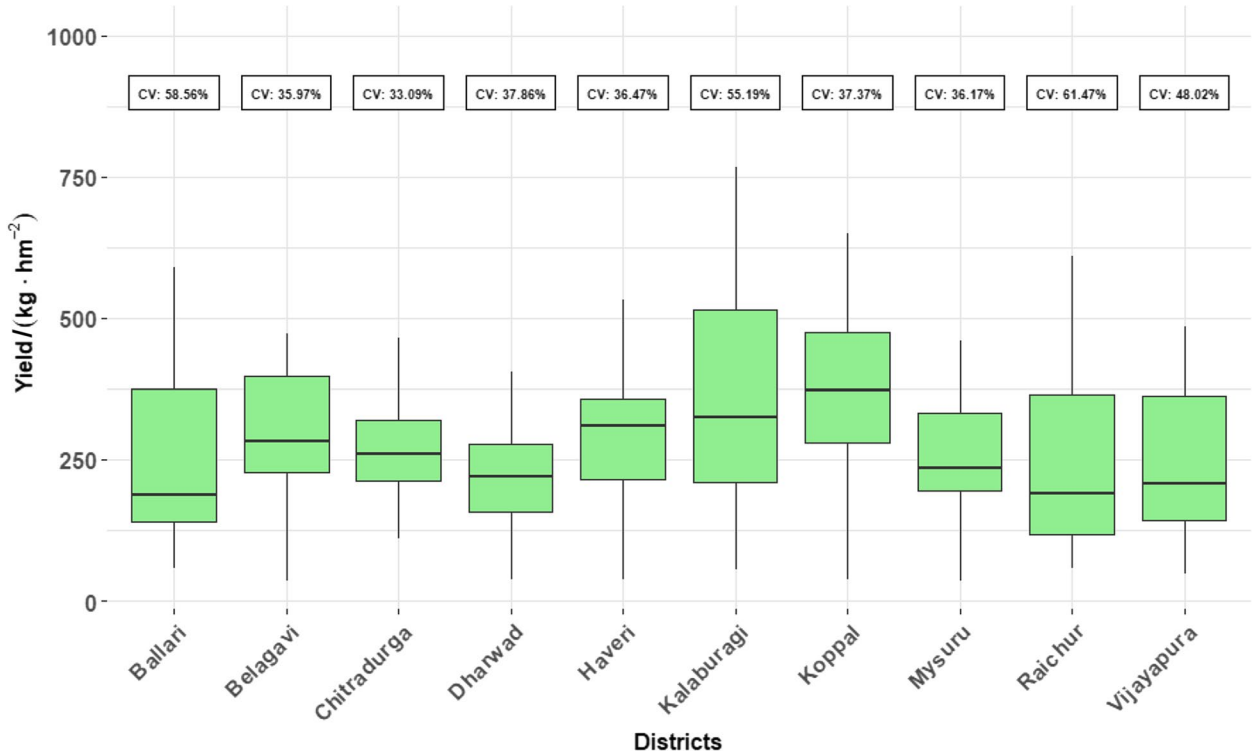

Fig. 1 Box plot representing district wise cotton yield distribution during 1990–2021. The central box in each plot represents the interquartile range, with the median line inside the box. The whiskers extend to the minimum and maximum values

et al. 2010). To generate these weather indices, two distinct methodologies were used. The unweighted weather indices were computed by aggregating weekly weather variables encountered throughout the crop period. On the other hand, the weighted indices were established by summing the product of the correlation coefficient and the value of the corresponding weekly weather variable. The formulas for computing unweighted and weighted weather indices are summarized below. By doing this, a time series dataset comprising 32 (from 1990 to 2021) weather variables (Table 2) and yield was generated.

Unweighted weather indices:

$$Z_{ij} = \sum_{w=1}^{m} x_{iw}$$

$$Z_{ii'j} = \sum_{w=1}^{m} x_{iw} x_{i'w}$$

Weighted weather indices:

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} x_{iw}$$

Table 2 Derived unweighted and weighted indices of composite weather parameters for model analysis

Parameter	Unweighted weather indices	Weighted weather indices
T_{max}	*Z*10	*Z*11
T_{min}	*Z*20	*Z*21
Rf	*Z*30	*Z*31
Rh I	*Z*40	*Z*41
Rh II	*Z*50	*Z*51
T_{max}*T_{min}	*Z*120	*Z*121
T_{max}**Rf*	*Z*130	*Z*131
T_{max}**Rh* I	*Z*140	*Z*141
T_{max}**Rh* II	*Z*150	*Z*151
T_{min}**Rf*	*Z*230	*Z*231
T_{min}**Rh* I	*Z*240	*Z*241
T_{min}**Rh* II	*Z*250	*Z*251
*Rf***Rh* I	*Z*340	*Z*341
*Rf***Rh* II	*Z*350	*Z*351
Rh I**Rh* II	*Z*450	*Z*451

Note: T_{max} represents maximum temperature, T_{min} represents minimum temperature, *Rf* represents rainfall, *Rh* I represent relative humidity in the morning, *Rh* II represents relative humidity in the evening, * represents multiplicative interaction between the two variables, Z_{ik} represents weather index generated by one weather variables or combination of two weather variables

$$Z_{ii'j} = \sum_{w=1}^{m} r^{j}_{ii'w} x_{iw} x_{i'w}$$

Where, x_{iw} and $X_{i'w}$ are values of two distinct weather variables (i^{th}/i'^{th}) for the same time period (the w^{th} week), r^{j}_{iw} represents the correlation coefficient between the de-trended yield and the i^{th} weather variable during the w^{th} week of the j^{th} time period, and $r^{j}_{ii'w}$ represents the correlation coefficient related to the interaction between i^{th} and i'^{th} weather variables, and the detrended yield during the w^{th} week in the j^{th} time period.

Brief background of multivariate models used in the study

The details of multivariate models used in this study to develop kharif cotton yield prediction are described below and the structured framework for yield forecasting models is illustrated in Fig. 2.

Stepwise multiple linear regression

MLR is the standard and simplest approach for developing calibration models. However, its application to datasets with more independent variables and a greater sample size is not always successful (Balabin 2011). Feature selection in the form of SMLR gives good results on large datasets. A stepwise regression procedure was adopted for the selection of the best regression variable among many independent variables (Singh 2014). ICAR - Indian Agricultural Statistical Research Institute developed models to express the effect of weather variables on crop yield. Yield is considered the dependent variable and weekly weather variables are considered the independent variables. Weekly weather variables are generated from daily data by averaging daily maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, and rainfall summing up. Two weather indices (unweighted and weighted) are developed for each weather variable, and

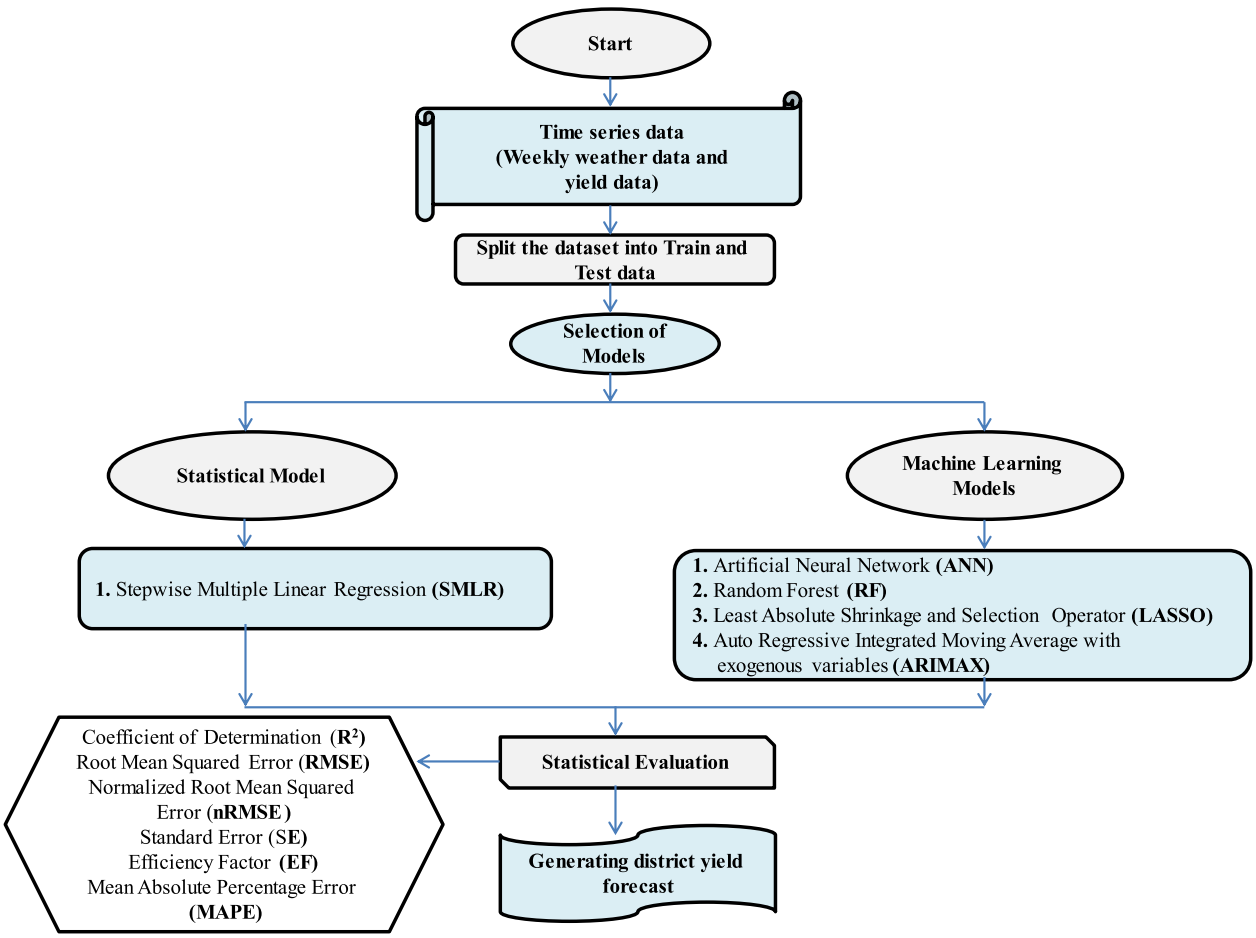

Fig. 2 Framework representing different stages in model for yield forecasting of cotton

indices are also generated for the interaction of weather variables. A combination of weather indices generated from weather variables (Table 2). Regression analysis is used to fit equations; weighting coefficients in the equations are obtained empirically using standard statistical procedures such as multivariable regression analysis using SPSS software. It appears that the study focuses on understanding the relationship between weather variables and crop yield and the use of regression models, including SMLR, to analyze and predict these relationships. The weighting coefficients are determined through empirical methods to enhance the accuracy of the model.

Artificial neural networks

These artificial intelligence (AI) methodologies provide a more effective means of tackling complexities within natural systems characterized by a multitude of inputs. ANNs are nonlinear and non-statistical models that mimic the learning process of the human brain (Starks 2019; Lawrence 1994) and no assumption of normality of the data is implied. Achieving optimal crop yield at minimal cost is a primary objective in agricultural production. The timely identification and management of issues associated with crop yield indicators play a pivotal role in amplifying overall productivity. The recent application of AI, encompassing technologies like ANNs, fuzzy systems, and genetic algorithms, has showcased enhanced efficiency in addressing challenges linked to agricultural yield. In the current study, a three-layered feed-forward artificial neural network comprising input, hidden, and output layers was proposed. The neurons or nodes in each layer are interconnected, with the number of nodes in the input and output layers predetermined by the dataset. The number of nodes in the input and output layers is fixed by the dataset used. There is a need to take care to choose the optimum number of hidden layers while implementing the ANN for yield forecasting, by using the 'train' function of the 'caret' package, using the method 'nnet' with 10-fold cross-validation in R software (Kuhn 2008). The ANN model is iteratively trained and evaluated until its predictive accuracy is maximized (Yang 2017). The analysis involved allocating 80% of the dataset for calibration (training) purposes and, the remaining 20% for validation (testing). A comprehensive set of 32 weather indices was utilized as inputs with yield serving as the dependent variable, and other factors acting as independent variables (Fig. 3).

Least absolute shrinkage and selection operator

LASSO and ENet methods are two shrinkage regression methods used for handling multicollinearity by penalizing the magnitude of regression coefficients (Piaskowski et al. 2016). LASSO reduces the number of predictors in a regression model and identifies important predictors. By shrinking the coefficients of less useful predictors to zero, LASSO can automatically choose an important variable and reject the rest from the model. By adopting a regularization technique, the variance of the estimated regression coefficients is minimized, and thus, the estimators are more stable.

Random forest (RF)

The RF model is a supervised technique for both classification regression and non-linear problems. This method uses the ensemble learning method for regression and is a bagging technique because it combines individual decision trees to yield better results. The advantage of the RF model is that it handles the missing values and maintains accuracy (Fang et al. 2021). A RF is an ensemble machine learning technique that constructs multiple trees while training data and gives class labels for classification problems or mean/average prediction for regression. It can also be used in both univariate and multivariate time series forecasts by manually creating lag and seasonal component variables. According to the nature of the data, different algorithms react differently.

Autoregressive integrated moving average (ARIMAX)

The ARIMAX model is an extended version of the ARIMA model. The ARIMAX model is linear in nature and hence does not explain the nonlinearity components. Here, we have tried to improve the performance of the ARIMAX model by explaining residuals through machine learning approaches such as ANN and support vector machines (Zhang 2003).

Model performance evaluation

Model performance was tested using different statistical model performance evaluation measures. The use of more than one measures helps us to evaluate a single model's performance and compare multiple models. In this study, the coefficient of determination (R^2), root mean square error (RMSE), normalized root mean square error (nRMSE), modelling efficiency (EF), and mean absolute percentage error (MAPE) were calculated.

The R^2 is important for measuring the effectiveness of the models (Shaikh et al. 2021; Ağbulut et al. 2020), ranging from 0 to 1. This approach provides insight into how well the trend of the model result is able to track the trends of observed data (Ağbulut et al. 2021). A value closer to 1 indicates that the model is more

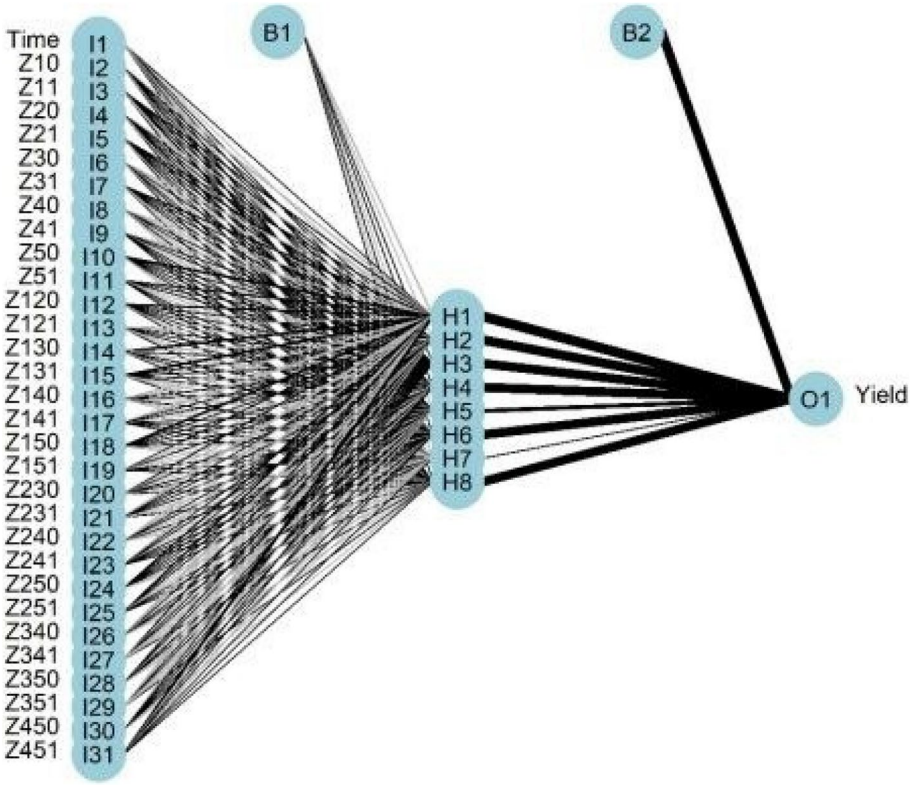

Fig. 3 Graphical image of the established artificial neural network for cotton yield forecasting (I1, I2, I3……. I31 represent the input layer consisting of independent weather indices, B1 represents the hidden layer accounting for the intermediate effect of the combination of weather indices as eight neurons *viz*., H1, H2…H8, B2 indicates the second hidden layer addressing the impact of the first hidden layer and at the end, O1 indicates the output *i.e.* yield.)

accurate. RMSE measures the average magnitude of the error and is related to the deviation from the actual value. An RMSE value of 0 indicates that the model has a perfect fit. The lower the RMSE is, the better the model and its predictions. The nRMSE expresses the spread around the measurements used for the classification of model performance into distinct groups (excellent, good, fair, or poor when the values are in the range of < 10%, 10–20%, 20–30%, or > 30%, respectively). The modelling efficiency indicates whether the model describes the data better than simply the average of the predictions. The optimal values are those that are near 1 (Thimmegowda et al. 2023). The MAPE was defined as the sum of the percentage to mean absolute error (MAE) (Kumar et al. 2020). For a good model, a smaller MAPE value is desirable. The MAPE less than 5% is considered as an indication that the forecast is acceptably accurate. MAPE greater than 10% but less than 25% indicates low accuracy, but a MAPE greater than 25% indicates very low accuracy. The model with a lower MAPE is preferred for forecasting purposes.

Stage specific evaluation of cotton yield prediction models using weather data

The quantification of weather impacts on crop growth is a cumbersome task, as weather factors impart yield through their direct and interactive effects. In our study, kharif cotton yield was forecasted at the vegetative stage (leaf development, stem growth, and root expansion, *i.e.*, 40 to 60 days after sowing) and mid stage (flowering and fruit development, *i.e.* 80 to 100 days after sowing) using five different models (SMLR, LASSO, ANN, RF, and ARIMAX). The crop duration in cotton is around 160–190 days, depending on the variety and growing conditions. Here the models were calibrated (1990–2018) and validated (2019–2021) using the historical dataset of weather variables and crop yield datasets, and the yield in 2023 was forecasted. Previous studies have reported that data from a couple of months prior to harvest can be used for short range crop predictions using statistical regression models (Chipanshi 2015; Mkhabela 2011; Seiler 2000). A similar methodology was used to analyse yield prediction with 16 standard meteorological weeks (SMWs) corresponding to the vegetative stage, and 20 SMWs corresponding to the mid-stage of crop growth. This

Table 3 District-wise deviation percent of kharif cotton yield at F1 and F2 stages validated in 2020 and 2021 using SMRL model

District	F1 stage						F2 stage					
	2020			2021			2020			2021		
	A	*P*	*D*/%	*A*	*P*	*D*/%	*A*	*P*	*D*/%	*A*	*P*	*D*/%
Ballari	579	357	38	512	357	30	579	443	24	512	452	12
Belaagavi	414	298	28	415	301	27	414	259	37	415	302	27
Chitradurga	406	276	32	367	300	18	406	337	17	367	322	12
Dharwad	315	322	–2	323	325	0	315	316	0	323	318	2
Haveri	341	444	–30	404	420	–4	341	408	–20	404	402	0
Kalaburagi	405	717	–77	522	743	–42	405	669	–65	522	677	–30
Koppal	382	434	–14	374	254	32	382	362	5	374	200	47
Mysuru	286	202	29	226	132	41	286	143	50	226	297	–31
Raichur	530	548	–3	429	558	–30	530	545	–3	429	556	–29
Vijayapura	360	511	–42	481	433	10	360	509	–4	481	520	–8

A actual yield (kg·hm^{-2}), *P* predicted yield (kg·hm^{-2}), *D* deviation

Table 4 District-wise deviation percent of kharif cotton yield at F1 and F2 stages validated in 2020 and 2021 using ARIMAX model

District	RMSE	MAPE	F1 stage						RMSE	MAPE	F2 stage					
			2020			2021					2020			2021		
			A	*P*	*D*/%	*A*	*P*	*D*/%			*A*	*P*	*D*/%	*A*	*P*	*D*/%
Ballari	48	17	579	525	9	512	582	–14	55	19	579	316	45	512	305	40
Belagavi	52	16	414	479	–16	415	586	–41	79	33	414	312	25	415	356	14
Chitradurga	41	15	406	272	33	367	296	19	47	17	406	316	22	367	336	9
Dharwad	32	11	315	347	–10	323	316	2	34	15	315	306	3	323	350	–8
Haveri	52	16	341	478	–40	404	562	–39	32	9	341	267	22	404	300	26
Kalaburagi	81	23	405	756	–87	522	776	–49	77	22	405	765	–89	522	848	–62
Koppal	87	28	382	271	29	374	418	–12	81	23	382	356	7	374	248	34
Mysuru	51	16	286	276	3	226	235	–4	43	15	286	179	38	226	322	–43
Raichur	43	21	530	505	5	429	509	–19	41	22	530	544	–3	429	547	–28
Vijayapura	41	16	360	511	–42	481	212	56	39	16	360	509	–42	481	515	–7

A actual yield (kg·hm^{-2}), *P* predicted yield (kg·hm^{-2}), *D* deviation

approach ensured the stage-specific weather influences were accurately captured and integrated into the prediction model.

Results

Cotton yield prediction using SMLR model

The Kharif cotton yield was validated in 2020 and 2021 at the F1 and F2 stages using SMLR across ten districts (Table 3). The model's prediction accuracy by displaying the actual yield, predicted yield, and percent deviation among them. The results showed that prediction accuracy was good with low deviations in few districts, while in other districts it showed large deviation in other districts, indicating varying performance of the SMLR model across different districts and stages among the districts. The negative deviation indicated that the model has overestimated the yield, and positive deviations indicated under-estimations. The yield forecasted in Dharwad district at both stages in 2020 and 2021 exhibited better results in comparison with other districts, while the forecast results for Kalaburagi were worse with overestimation reaching −77% at the F1 stage and −30% at the F2 stage followed by Raichur district with the same trend.

Cotton yield prediction using ARIMAX model

The predicted cotton production for ten districts deviated from the actual cotton production using the ARIMAX model (Table 4). The RMSE ranged between 87 (Koppal) and 41 (Chitradurga and Vijayapura) at the F1

stage and between 81 (Koppal) and 32 (Haveri) at the F2 stage. Except for the Koppal at the F1 stage and Belagavi district at the F2 stage, all districts had a MAPE value of less than 25% at both the F1 and F2 stages, indicating lower but acceptable accuracy. The yield was overstimated and ranged from −16% to −87% in 2020 and −49% to −4% in 2021 at the F1 stage. Furthermore, cotton production was underestimated in the remaining districts (for example, Ballari, Chitradurga, Koppal, Mysuru, and Raichur in 2020 and Chitradurga, Dharwad, Vijayapura in 2021). Similar results were also observed at the F2 stage. The model consistently performed better in districts like Dharwad and Chitradurga with lesser deviation and lower RMSE. On the other hand, Kalaburagi and Koppal districts showed large deviations with higher RMSE suggesting the model need improvement or external factors are influencing cotton yield in these areas.

Cotton yield prediction using LASSO model

The yields forecasted in 2020 and 2021 for cotton at the F1 and F2 stages using LASSO model were calibrated and validated against the actual yields (Table 5). The RMSE ranged from 33 (Dharwad) to 91 (Koppal) at the F1 stage and from 31 (Haveri) to 85 (Koppal) at the F2 stage. At the F1 stage, except for the Koppal district, all other districts had MAPE values less than 25%, indicating lower but acceptable accuracy, and at the F2 stage, the MAPE values for most districts were less than 25%, except for Belagavi and Koppal district. The LASSO model demonstrated variable performances across districts and years. Cotton production is overstimated across the districts, yield deviating as −18% (Belagavi), −4% (Dharwad), −19% (Haveri), −79% (Kalaburagi), and −35% (Vijayapura) in 2020 at the F1 stage. At the F2 stage of 2021, the percent deviation of −33% in Belagavi, −13% in Haveri, −41% in Kalaburagi, −79% in Kalaburagi, −17% in Raichur district and other districts showed understimates. Underestimates and overestimates were observed in different districts and different years, suggesting the need for refinement. Further analysis and refinement of the LASSO model may be necessary to improve accuracy, especially in districts where significant deviations were observed.

Cotton yield prediction using RF model

The cotton yield prediction was validated in 2020 and 2021 at the F1 and F2 stages using random forest (Table 6). The model calibrated and tested for RMSE and MAPE ranged from 29 (Raichur) to 105 (Koppal) and from 13 (Raichur) to 39 (Koppal) at the F1 stage, respectively. Similarly, at the F2 stage, the calibrated yields for the RMSE and MAPE ranged from 28 (Mysuru) to 89 (Kalaburagi) and from 10 (Mysuru) to 30 (Belagavi), respectively. Dharwad and Vijayapura districts showed smaller deviations, indicating more accurate predictions among the other districts. Whereas, districts like Haveri, Kalaburagi, and Koppal showed higher deviations, particularly at the F1 stage. However, the validated results at the F1 and F2 stages in 2020 and 2021 showed mixed results of underestimation and overestimation by the model; in a few districts, the deviation percentage was within the acceptable limit, and in other districts, the predicted yield tended to vary.

Cotton yield prediction using ANN model

The percentage difference between the forecast and actual yield was validated for the period 2020 and 2021 to determine the accuracy of the ANN model (Table 7).

Table 5 District-wise deviation percent of kharif cotton yield at F1 and F2 stages validated 2020 and 2021 using LASSO model

District	RMSE	MAPE	F1 stage						RMSE	MAPE	F2 stage					
			2020			2021					2020			2021		
			A	*P*	*D*/%	*A*	*P*	*D*/%			*A*	*P*	*D*/%	*A*	*P*	*D*/%
Ballari	62	18	579	451	22	512	457	11	58	18	579	334	42	512	322	37
Belagavi	48	16	414	487	−18	415	552	−33	82	32	414	290	30	415	313	25
Chitradurga	38	13	406	241	41	367	245	33	40	14	406	263	35	367	270	26
Dharwad	33	14	315	328	−4	323	301	7	35	16	315	287	9	323	325	−1
Haveri	48	16	341	405	−19	404	456	−13	31	10	341	295	14	404	333	18
Kalaburagi	83	22	405	725	−79	522	736	−41	79	20	405	740	−83	522	808	−55
Koppal	91	33	382	302	21	374	425	−14	85	30	382	366	4	374	285	24
Mysuru	49	16	286	278	3	226	224	1	44	14	286	174	39	226	249	−10
Raichur	41	20	530	489	8	429	501	−17	40	19	530	516	3	429	540	−26
Vijayapura	40	14	360	486	−35	481	440	9	38	14	360	494	−37	481	697	−45

A actual yield (kg·hm^{-2}), *P* predicted yield (kg·hm^{-2}), *D* deviation

Table 6 District-wise deviation percent of kharif cotton at F1 and F2 stages validated in 2020 and 2021 using random forest model

District	RMSE	MAPE	F1 stage						RMSE	MAPE	F2 stage					
			2020			2021					2020			2021		
			A	*P*	*D*/%	*A*	*P*	*D*/%			*A*	*P*	*D*/%	*A*	*P*	*D*/%
Ballari	69	25	579	475	18	512	482	6	62	23	579	418	28	512	412	20
Belagavi	57	18	414	479	−16	415	571	−38	73	30	414	406	2	415	369	11
Chitradurga	39	13	406	201	51	367	223	39	53	19	406	273	33	367	330	10
Dharwad	40	19	315	330	−5	323	286	12	38	17	315	281	11	323	326	−1
Haveri	57	18	341	506	−48	404	515	−28	46	15	341	455	−34	404	436	−8
Kalaburagi	85	23	405	680	−68	522	664	−27	89	22	405	662	−64	522	756	−45
Koppal	105	39	382	305	20	374	374	0	76	24	382	541	−42	374	439	−17
Mysuru	53	18	286	245	14	226	234	−4	28	10	286	92	68	226	359	−59
Raichur	29	13	530	514	3	429	468	−9	37	16	530	439	17	429	491	−15
Vijayapura	60	22	360	371	−3	481	398	17	63	21	360	393	−9	481	410	15

A actual yield (kg·hm^{-2}), *P* predicted yield (kg·hm^{-2}), *D* deviation

Table 7 District-wise deviation percent of kharif cotton yield at F1 and F2 stages validated in 2020 and 2021 using ANN model

District	2020					2021				
	Actual yield/ (kg·hm^{-2})	F1		F2		Actual yield/ (kg·hm^{-2})	F1		F2	
		Forecast yield/ (kg·hm^{-2})	Deviation/%	Forecast yield/ (kg·hm^{-2})	Deviation/%		Forecast yield/ (kg·hm^{-2})	Deviation/%	Forecast yield/ (kg·hm^{-2})	Deviation/%
Ballari	579	579	0.0	581	−0.4	512	503	2.0	526	−2.8
Belagavi	414	400	3.5	377	8.9	415	400	4.0	377	9.1
Chitradurga	406	397	2.2	374	7.8	367	344	6.0	374	−2.0
Dharwad	315	313	0.6	309	1.9	323	309	4.0	322	0.3
Haveri	341	339	0.7	339	0.7	404	404	0.0	402	0.5
Kalaburagi	405	460	−13.6	399	1.4	522	460	12.0	525	−0.6
Koppal	382	402	−5.2	352	7.9	374	372	1.0	349	6.6
Mysuru	286	288	−0.6	286	0.0	226	227	−1.0	233	−3.1
Raichur	530	541	−2.1	524	1.2	429	429	0.0	426	0.8
Vijayapura	360	362	−0.4	394	−9.3	481	486	−1.0	480	0.3

In 2020, at the F1 stage, Koppal (−5.2%), Kalaburagi (−13.6%), Mysuru (−0.6%), Raichur (−2.1%), and Vijayapura (−0.4%) exhibited overestimation, and other districts exhibited underestimation, ranging from 0.6 to 3.5%. At the F2 stage, Ballari and Vijayapura yield validated were overestimation with −0.4% and −9.3%, respectively.

Similarly, in the 2021 F1 stage, two (Mysuru and Vijayapura) districts overstimated the cotton yield, with a −1% deviation each; however, for other eight districts, the forecasted yields were underestimated; at the F2 stage, the yield was understimated for six districts out of ten districts, with a deviation percent ranging from 0.3% to 9.1% and rest of the districts overstimated the yield. The results revealed an excellent agreement between the actual and forecasted yields. The errors calculated by this model were within the acceptable limits *i.e.*, ± 10%, for most of the districts except for Kalaburagi at the F1 stages in both 2020 and 2021; hence, this can be best used for yield predicting.

The performance of the calibrated kharif cotton yield prediction model using ANN was evaluated across various districts (Table 8). A model with smaller RMSE, nRMSE, and higher EF values is considered to be better. The ANN models were used to used to evaluate for the F1 and F2 stages, with RMSE values ranging from 1.30 to 49.0 for the F1 stage and 1.8 to 60.1 for the F2 stage. The nRMSE values ranged from 0.4 to 16.9 for the F1

Table 8 Statistical evaluation of validated kharif cotton yield using ANN model

District	F1 stage				F2 stage			
	RMSE	nRMSE	nRMSE class	EF	RMSE	nRMSE	nRMSE class	EF
Ballari	15.5	6.5	Excellent	1.0	17.3	7.2	Excellent	1.0
Belagavi	49.0	16.9	Good	0.8	57.1	19.7	Good	0.7
Chitradurga	22.1	8.7	Excellent	0.9	17.5	6.9	Excellent	1.0
Dharwad	19.8	9.4	Excellent	0.9	22.5	10.7	Good	0.9
Haveri	1.3	0.4	Excellent	1.0	1.8	0.6	Excellent	1.0
Kalaburagi	22.5	6.3	Excellent	1.0	60.1	16.9	Good	0.9
Koppal	32.7	8.4	Excellent	1.0	25.9	6.6	Excellent	1.0
Mysuru	5.2	2.0	Excellent	1.0	9.5	3.7	Excellent	1.0
Raichur	33.1	14.1	Good	1.0	14.7	6.2	Excellent	1.0
Vijayapura	16.4	6.9	Excellent	1.0	13.5	5.7	Excellent	1.0

nRMSE classes are as follows: <10%=excellent, 10%–20%=good, 20%–30%=fair, and >30%=poor

Table 9 Kharif cotton yield forecast in 2023 at the F1 stage using SMLR

District	Equation	R^2	F-value	SE	Forecast yield (kg·hm^{-2})
Ballari	Y=–48.35 + 6.764*Time + 0.317*$Z51$ + 0.013*$Z351$	0.79	33.31	71.44	316
Belagavi	Y=–2.66 + 0.05* $Z131$ + 0.08*$Z151$	0.52	15.53	83.61	329
Chitradurga	Y = 239.17–40.53*$Z21$–0.66*$Z50$ + 0.39*$Z141$ + 0.07*$Z451$	0.74	17.81	52.13	181
Dharwad	Y = 35.24 + 3.18*Time + 0.04*$Z131$	0.78	50.01	44.16	262
Haveri	Y=–4.91 + 0.05*$Z131$ + 0.007*$Z150$	0.75	30.57	66.62	446
Kalaburagi	Y = 18.39 + 15.22*Time–1.64*$Z41$ + 0.04*$Z451$	0.76	27.87	112.16	734
Koppal	Y = + 3.94*$Z20$ + 101.24*$Z21$ + 0.02*$Z341$ + 0.05*$Z451$	0.77	15.14	88.56	682
Mysuru	Y = 16.23–2.69*Time + 0.02*$Z151$–0.01*$Z230$ + 0.02*$Z341$	0.84	27.98	44.51	166
Raichur	Y=–15.44 + 10.31*Time–0.07*$Z131$ + 0.04*$Z341$	0.87	59.88	58.18	423
Vijayapura	Y=–9.26 + 11.02*Time–0.24*$Z131$ + 0.09*$Z351$	0.76	28.90	62.14	404

The weather parameters in the formula are shown in Table 2

SE standard error

stage and 0.6 to 16.9 for the F2 stage, while the EF values ranged from 0.9 to 1.0 for both stages. Among the districts yield predicted, at the F1 stage, lower values of RMSE (1.30), nRMSE (0.4) and the highest EF (1.00) was found in Haveri district and higher value was observed in Belagavi district, with 49.0, 16.9, and 0.80 of RMSE, nRMSE, and EF, respectively. Similarly, at the F2 stage, lower value of RMSE (1.8), nRMSE (0.6) and EF (1.00) was found in Haveri district and higher RMSE was observed in Kalaburagi and Belagavi districts, with the highest nRMSE of 19.7. Overall, the model performed excellently, with an nRMSE value less than 10% categorized as excellent for eight out of ten districts in the F1 stage and for seven districts as excellent in the F2 stage. Moreover, the nRMSE value was categorized as good in two districts in the F1 stage and in three districts during the F2 stage.

Inter comparison of models for their yield predictability

The kharif cotton yield was forecasted in 2023 at the F1 stage using SMLR for ten districts. The model performance was evaluated using R^2, F value, and standard error (*SE*) of the estimates resulted from different weather variables (Table 9). The R^2 value in the model ranges from 0.52 to 0.87. The model generally performs well across the districts, with R^2 values above 0.7 for nine districts. Dharwad district had a lower *SE* (44.16), indicating relatively accurate predictions, and a higher R^2 (0.78). The Kalaburagi district has a higher *SE* (112.16), suggesting less accurate predictions, but still with a moderate R^2 (0.76). The R^2 value of the Belagavi district is less than 0.6, indicating a moderate fit and suggesting that the model may not fit the data in that region. While R^2 provides an overall measure of goodness-of-fit, it is essential to consider the specific context of each district and the agricultural factors that might influence the predictions.

Similarly, the yield was also forecasted in 2023 at the F2 stage using SMLR for ten districts, and the regression equations and weather variables influencing the equation and the model performance were evaluated (Table 10). The model generally performs well across all districts, with consistently higher R^2 values. A lower *SE* value suggests that the model provides accurate estimates for most districts. Dharwad district had a lower *SE* (41.91), indicating relatively accurate predictions, and the highest R^2 (0.89). The Kalaburagi district has a higher *SE* (95.26), suggesting less accurate prediction, but still with a good R^2 of 0.84. All districts have an R^2 value above 0.7, indicating a stronger fit for predicting cotton yield at the F2 stage compared with the F1 stage.

The kharif cotton yield forecasted in 2023 at the F1 and F2 stages for the ten districts using ARIMAX, LASSO, RF, and ANN (Tables 11 and 12). The estimated yield at F1 stage ranged from 206 kg·hm^{-2} (Vijayapura) to 916 kg·hm^{-2} (Kalaburagi), 196 kg·hm^{-2} (Chitradurga) to 860 kg·hm^{-2} (Kalaburagi), 170 kg·hm^{-2} (Chitradurga) to 792 kg·hm^{-2} (Kalaburagi), and 145 kg·hm^{-2} (Chitradurga) to 486 kg·hm^{-2} (Vijayapura) using ARIMAX, LASSO, RF, and ANN, respectively.

Similarly, at the F2 stage, the estimated yields ranged from 201 kg·hm^{-2} (Mysuru) to 883 kg·hm^{-2} (Kalaburagi), 144 kg·hm^{-2} (Mysuru) to 839 kg·hm^{-2} (Kalaburagi), 144 kg·hm^{-2} (Koppal) to 797 kg·hm^{-2} (Kalaburagi), and 193 kg·hm^{-2} (Mysuru) to 819 kg·hm^{-2} (Ballari) using ARIMAX, LASSO, RF, and ANN, respectively.

The district average yields forecasted at the F1 stage in 2023 were found to be 401, 414, 408, and 380 kg·hm^{-2} using ARIMAX, LASSO, RF, and ANN, respectively. Similarly, the average yield predicted at the F2 stage in

Table 10 Kharif cotton yield forecast in 2023 at the F2 stage using SMLR

District	Equation	R^2	F-value	SE	Forecast yield (kg·hm^{-2})
Ballari	Y=−28.86 + 8.90*Time + 0.30*Z51 + 0.10*Z241	0.80	36.12	69.17	404
Belagavi	Y=−7.75 + 0.41*Z10 + 0.05*Z131	0.71	14.90	84.51	334
Chitradurga	Y=−1.10–0.02*Z150 + 0.06*Z451	0.80	24.48	58.69	208
Dharwad	Y = 31.94 + 3.36*Time + 0.04*Z131	0.89	57.08	41.91	267
Haveri	Y=−10.09 + 0.02*Z351 + 0.12*Z151	0.78	35.57	62.86	420
Kalaburagi	Y = 2.37 + 7.48*Time–0.82*Z40 + 0.91*Z121 + 0.05* Z131 + 0.04*Z451	0.84	25.58	95.26	897
Koppal	Y=−216.47 + 2.45*Z20 + 33.39*Z21 + 0.02*Z341 + 0.03*Z451	0.78	16.85	84.93	478
Mysuru	Y=−59.17 + 4.88*Z21 + 2.78*Z41–0.10*Z140 + 0.34* Z141 + 0.04*Z341–0.04*Z351	0.91	32.83	34.96	200
Raichur	Y=−21.01 + 10.42*Time–2.92*Z31 + 0.04*Z341	0.93	60.21	58.05	424
Vijayapura	Y = 2.52 + 10.65*Time–0.22*Z131 + 0.09*Z351	0.87	29.29	61.81	406

The weather parameters in the formula are shown in Table 2

SE standard error

2023 was 424, 408, 383, and 395 kg·hm^{-2} using respective models. The predicted mean yield in 2023 using different models was found to be higher than the average yield (1990–2021) of 289 kg·hm^{-2} (Fig. 4).

Assessment of major weather factors imparting cotton yield

As the performance of the ANN in comparison to other models was statistically good, the tested model was further used to assess variables of importance. Assessment of variable importance is a statistical methodology commonly used for identifying top variables having greater contribution over the dependent parameter and is dependent on the 'weights' assigned by the ANN during the formulation of the model. Significant weather variables that strongly influenced cotton production in all districts in the present study were identified (Figs. 5 and 6). Firstly, ANN and SMLR differed with respect to the identification of important variables, as there was a deficiency of SMLR to consider a large number of variables except input variables. In the case of ANN, it considers the interaction between two variables as a new variable and assigns 'weights' to the particular variable after iterating with all other variable combinations. Secondly, there was a district wise difference in the type of important variable identified by ANN because of relative variability of interaction of weather factors in each district (Das 2018). The districts differed with respect to variable importance for cotton yield formation during the vegetative stage (F1), however, similarities in the most influencing factor for cotton yield *i.e.* interaction between $T_{\max}$ and *Rh* II (Z151) were identified in districts like Dharwad and Vijayapura districts. In districts like Kalaburagi, Belagavi, and Mysuru, the interaction between T_{min} and *Rf* (Z230) was found highly influential on yield. In remaining districts, there was a mixed occurrence of highly influential variables of importance, signifying a differential role of weather variables in different districts on yield formation.

During mid-stage (F2) there was differential influence of weather factors identified, in Mysuru and Raichur districts the unweighted interaction of $T_{\max}$ and *Rf* (Z130) was responsible for yield formation, whereas, in Dharwad and Ballari, the weighted interaction of $T_{\min}$ and *Rf* (Z231) were responsible. Except these, in the remaining six districts a mixed influence of weather factors was observed. The influence of critical weather variables varied notably among districts and among the growth stages. The variability is largely due to the crop's specific weather requirements for optimal growth and yield. For example, in dry farming districts like Kalaburagi, Ballari, there is a limited availability of soil moisture (rainfall) and a high scope for evapotranspiration hence the yield is most likely to be imparted through these. These variables not only impact crop yield but also influence pest and disease epidemiology, which in turn impacts cotton yield (Madasamy et al. 2020). Previously, correlation studies have revealed a positive correlation between morning and evening relative humidity on the population of sucking insect pests and a negative correlation between maximum and minimum temperature on the population of sucking insect pests (Shivaray Navi et al. 2021; Krishna et al. 2020).

Discussion

Crop yield, being a complex function of different factors like edaphic/soil, climate, and management, relied on the variabilities brought in among them. The edaphic factors are relatively stable, and the management is constant, the

Table 11 Kharif cotton yield forecast in 2023 at the F1 stage using different machine learning mode

District	Average yield / (kg·hm^{-2})	ARIMAX				LASSO				RF				ANN			
		R^2	SE	F	Predicted yield /(kg·hm^{-2})	R^2	SE	F	Predicted yield /(kg·hm^{-2})	R^2	SE	F	Predicted yield /(kg·hm^{-2})	R^2	SE	F	Predicted yield /(kg·hm^{-2})
Ballari	270	0.26	220.64	0.68	211	0.76	149.37	0.01	281	0.30	128.53	0.23	278	0.81	69.46	3.32	481
Belagavi	301	0.72	53.77	1.88	519	0.07	97.89	0.01	510	0.81	54.29	1.24	512	0.85	60.81	0.65	400
Chitradurga	265	0.47	61.03	0.58	223	0.34	77.79	0.00	196	0.50	59.97	0.26	170	0.64	48.91	2.20	145
Dharwad	222	0.35	90.38	0.72	311	0.15	78.24	0.01	294	0.44	58.15	0.80	261	0.63	48.16	1.14	270
Haveri	306	0.89	133.92	1.06	516	0.61	123.01	0.04	420	0.45	102.14	0.26	500	0.84	72.52	0.63	339
Kalaburagi	373	0.54	127.24	0.43	916	0.08	171.68	0.00	860	0.25	156.25	0.33	792	0.60	134.91	1.17	461
Koppal	385	0.11	137.27	0.24	376	0.44	145.52	0.01	368	0.64	119.74	0.18	371	0.05	183.44	0.26	391
Mysuru	257	0.39	46.85	0.36	259	0.40	60.22	0.09	271	0.48	43.26	1.19	274	0.37	77.88	1.43	415
Raichur	257	0.74	207.38	0.79	476	0.80	159.9	0.00	470	0.50	117.09	0.38	526	0.96	31.88	21.13	414
Vijayapura	254	0.36	116.99	0.31	206	0.73	120.78	0.00	469	0.22	106.81	0.16	400	0.63	79.86	0.64	486
Average	**289**				**401**				**414**				**408**				**380**

The average yield was the average of the production data of each district from 1991 to 2021

Table 12 Kharif cotton yield forecast in 2023 at the F2 stage using different machine learning model

District	Average yield /(kg·hm^{-2})	ARIMAX				LASSO				RF				ANN			
		R^2	SE	F	Predicted yield /(kg·hm^{-2})	R^2	SE	F	Predicted yield /(kg·hm^{-2})	R^2	SE	F	Predicted yield / (kg·hm^{-2})	R^2	SE	F	Predicted yield / (kg·hm^{-2})
Ballari	270	0.59	95.89	1.08	335	0.49	149.37	0.01	295	0.41	127.78	0.17	271	0.83	69.33	3.43	819
Belagavi	301	0.61	61.95	1.62	382	0.02	97.89	0.01	321	0.88	61.49	0.50	443	0.85	56.52	0.87	321
Chitradurga	265	0.48	58.17	0.62	305	0.1	77.79	0.00	238	0.41	64.22	0.16	288	0.63	55.54	1.70	248
Dharwad	222	0.48	67.98	0.69	292	0.76	78.24	0.01	271	0.39	60.84	0.54	275	0.74	42.16	1.73	203
Haveri	306	0.95	119.69	1.19	404	0.46	123.01	0.04	368	0.99	90.22	0.21	433	0.95	80.07	0.27	281
Kalaburagi	373	0.6	113.93	0.90	883	0.12	171.68	0.00	839	0.19	153.53	0.25	797	0.60	126.01	1.08	476
Koppal	385	0.36	118.72	0.41	350	0.64	147.47	0.04	356	0.21	133.73	0.13	144	0.02	154.53	0.23	341
Mysuru	257	0.79	34.29	1.01	201	0.10	63.35	0.06	144	0.84	29.39	1.89	230	0.54	58.36	1.52	193
Raichur	257	0.75	206.97	0.79	481	0.65	159.9	0.00	478	0.54	111.95	0.61	500	0.96	34.37	25.64	586
Vijayapura	254	0.62	88.74	0.35	602	0.24	120.7	0.00	765	0.29	106.79	0.11	450	0.67	77.9	0.65	480
Average	**289**				**424**				**408**				**383**				**395**

The average yield was the average of the production data of each district from 1991 to 2021

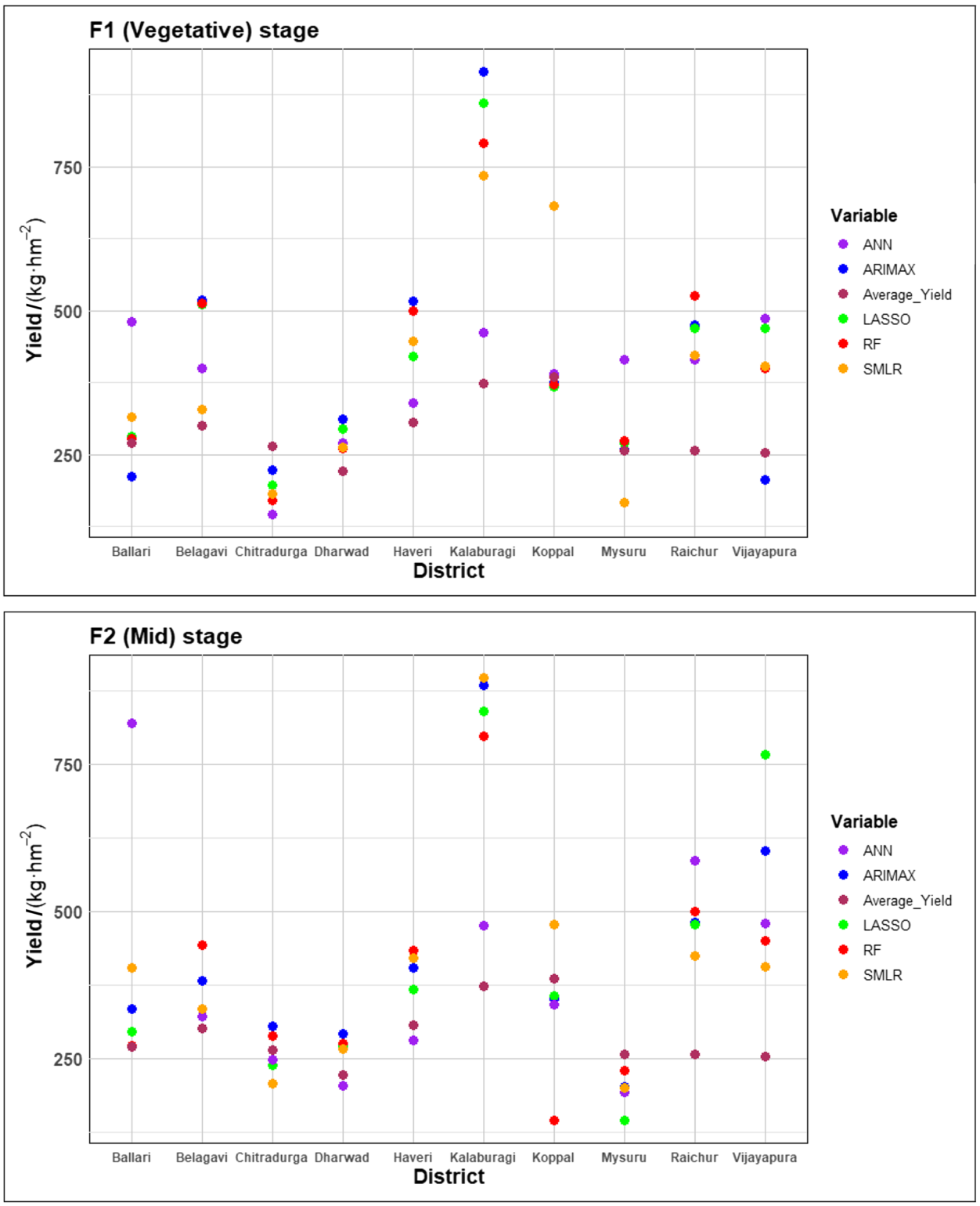

Fig. 4 Inter comparison of different multivariate models for their kharif cotton yield predictability during vegetative (F1) and mid (F2) stages in major cotton growing districts of Karnataka

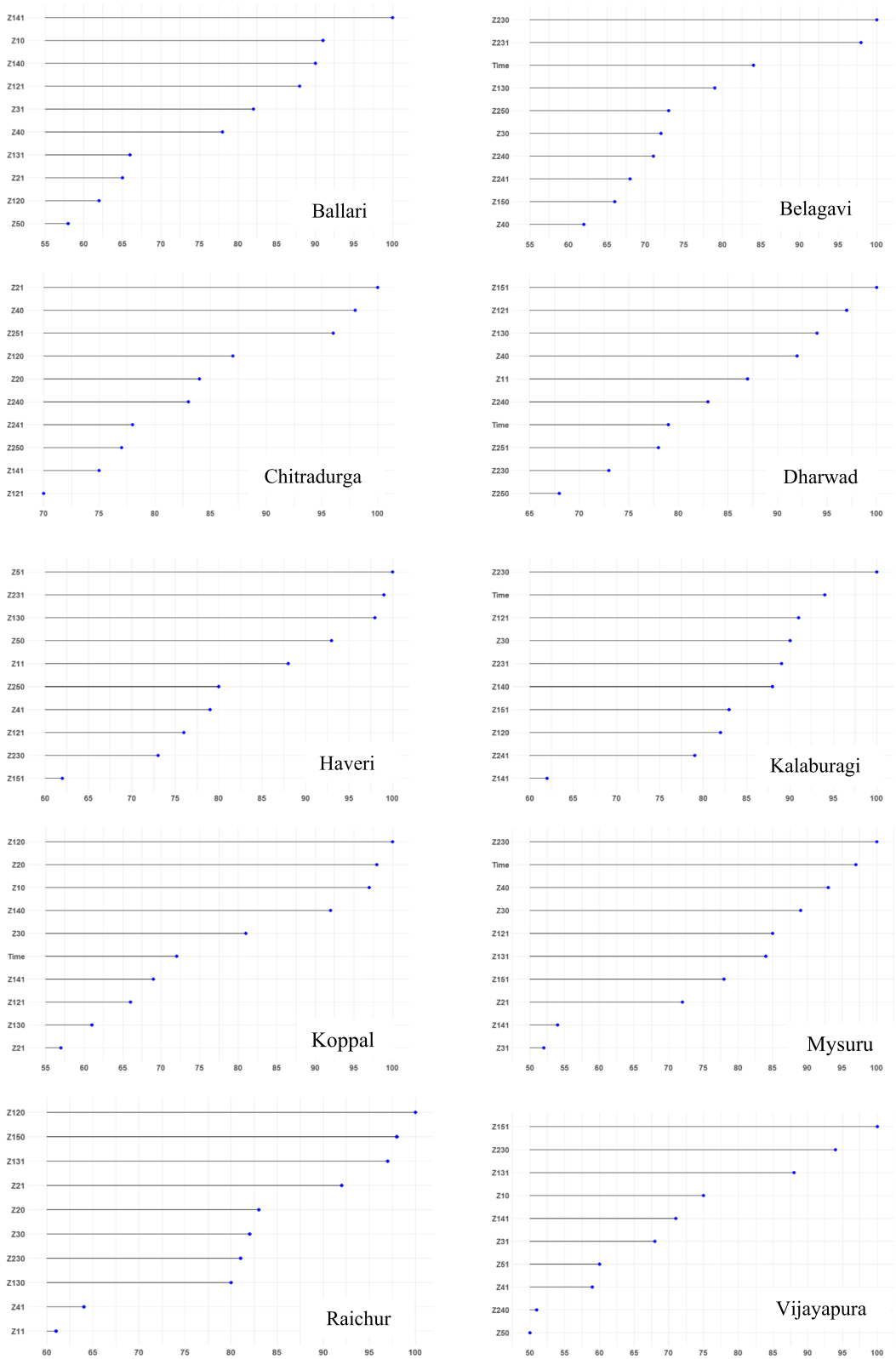

Fig. 5 Importance of top 10 weather indices in predicting cotton yield using the ANN model at the F1 stage. The y-axis indicates the weather indices, and x-axis indicates the importance of the particular feature in predicting the yield

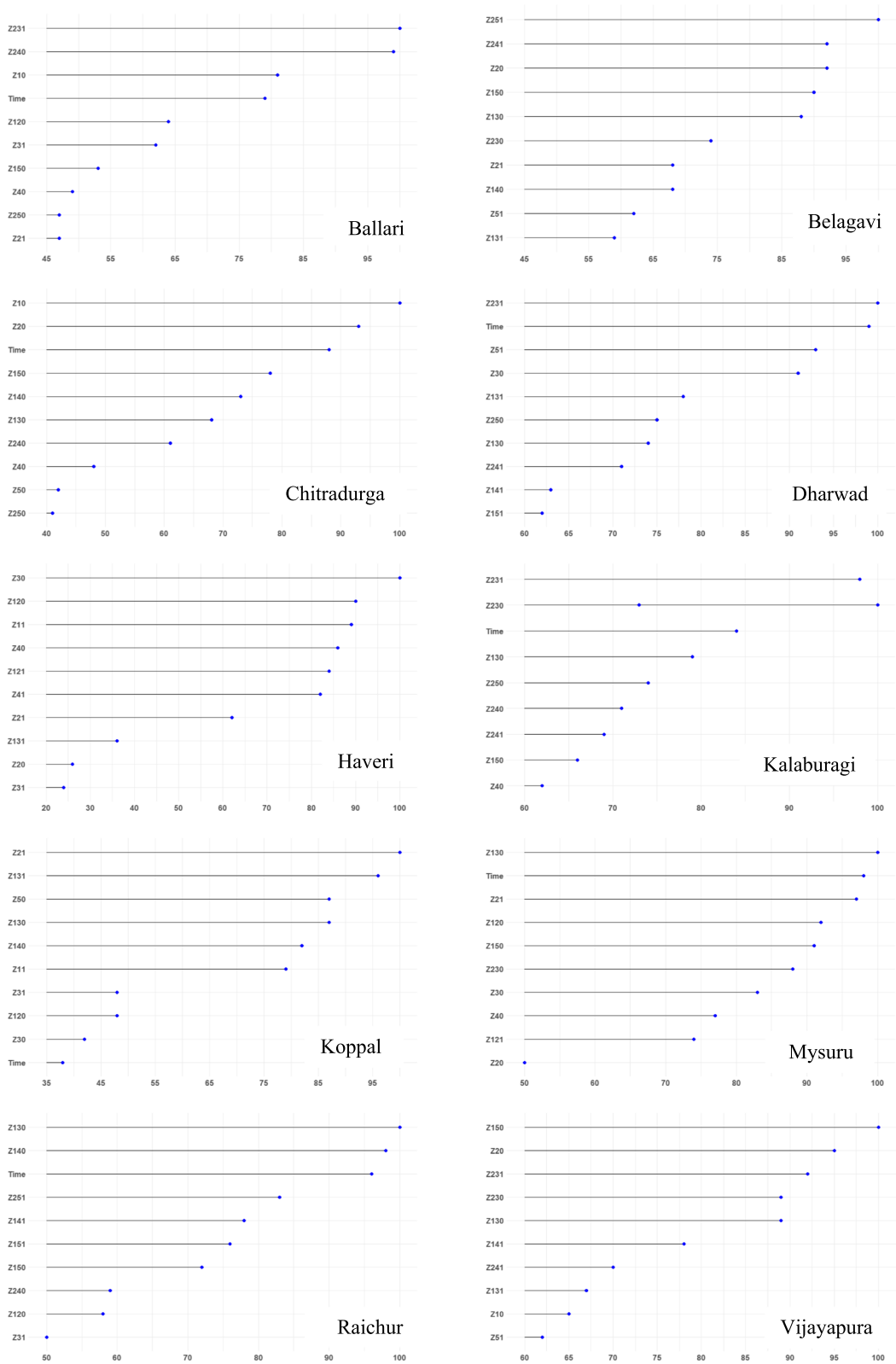

Fig. 6 Importance of top 10 weather indices in predicting cotton yield using the ANN model at the F2 stage. The y-axis indicates the weather indices, and x-axis indicates the importance of the particular feature in predicting the yield

yield is defined majorly by climatic factors. These being dynamic with time and space, impart a variable impact on the crop. Though efforts have been made to have an estimate of crop yield due to climate variability, the traditional techniques such as SMLR fail to capture complex interactive effects of climatic parameters thus necessitating the application of machine learning models. These models also differ in their ability to capture the influence of weather factors, models like ANNs are successful to a maximum extent in predicting the yield (Khaki 2021; Alvarez 2009; Li et al. 2007).

For evaluating the effectiveness of ANN in cotton yield prediction, a comparison of different methodologies based on RMSE and MAPE values during model calibration was conducted. The results showed that the ANN approach outperformed other methods as evidenced by lower error values highlighting the superiority of ANN model in predicting the yield across different districts of Karnataka (Table 8). The superiority of the ANN approach over the conventional empirical model to predict the yield of maize (Uno et al. 2005), rice (Paswan et al. 2013) and other food crops (Behroozi-Khazaei et al. 2017; Basir et al. 2021). The performance of the ANN approach was based on nRMSE during model calibration in the F1 and F2 stages, except in Belagavi and Raichur, all other districts exhibited excellent results (F1 stage); and at the F2 stage, the performance was good in Dharwad, Belagavi, and Kalaburagi districts, while the remaining districts exhibited excellent results. This might be due to the ability of ANN to consider the collinearity between weather variables for yield prediction (Haghverdi 2018; Abrouguia 2019). Variations in average weather patterns and extreme weather conditions have posed major risks to crop production worldwide. The use of machine learning algorithms is a reliable method for yield forecasting with lower error. Proper tuning of model parameters and inclusion of large datasets for model calibration and validation is the key to successful prediction. A study on the effects of remote sensing and data size and climate on cotton yield prediction, cotton yield is affected by many factors that can be largely categorized as genetics, environment, and management practices (Sawan 2017; Bakhsh 2005; Chaudhry 2009; Haghverdi 2018; Pokhrel 2018; Niedbala 2019). Therefore, there is a need for more studies to determine how ANN models can be used to determine the effects of these factors on cotton yield (Yildirim et al. 2022). The use of machine learning tools such as ANN, LASSO, NNet, etc., paves a promising approach for precision yield forecasting in other rainfed crops such as sorghum, rice, etc. where there's observed variability in yield which can mainly be attributed for variations in weather conditions during the crop growth period. In turn, the outcomes of such studies aid in having an idea of advanced estimates of crop yields based on weather conditions during initial crop growth periods, especially for decision making in future crop management and planning of policies. Furthermore, the inclusion of more features related to soil and crop growth parameters in the future can help improve the accuracy of machine learning models. Observed differences in the model performances and these can be minimized by combining factors such as edaphic variables (e.g., soil moisture, nutrient availability) and management practices (e.g., irrigation, nutrient application). These factors may be gathered through physical observations or remote sensing by measuring plant vigor using the normalized difference vegetation (NDVI).

Further, the machine learning models are not devoid of critical limitations, for example, ANNs may perform well, but they often function as black-box models, lacking interpretability and failing to reveal underlying relationships, such limitations have to be taken into account while the intention is to have an idea of underlying relationships besides yield estimation only (Hu et al. 2023). The result provides an understanding of the model's performance across different districts and years, shedding light on both success and areas where improvements or further exploration may be beneficial. The interplay of factors influencing agricultural yield is complex, and these analyses serve as valuable guides for refining predictive models and agricultural strategies.

Conclusions

The study emphasizes the comparison of different statistical and machine learning techniques for forecasting kharif cotton yield in the growing regions of major cotton producing states in India. To account for the individual and interactive impacts of weather factors, weighted and unweighted weather indices were calculated and used as independent factors. One statistical model (SMLR) and four machine learning models (ANN, LASSO, ARIMAX, and RF) were tested and compared for their performance in cotton yield forecasting in two growth stages (F1, vegetative stage and F2, mid-stage). The ANN model outperformed all other models, as demonstrated by the satisfactory ranges of the model performance evaluated by RMSE, nRMSE, and EF values. Furthermore, the tested ANN model was used to identify the top ten variables of importance impacting kharif cotton yield in each district, which indicated the difference in the set of variables in different districts because of variability in weather factors and their interaction in each district under study. Morning relative humidity, along with its interactions with maximum and minimum temperatures significantly affects cotton yield in most of the predicted districts. Necessitating development of appropriate planning mitigate the negative impacts of weather variables on agricultural policies.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s42397-024-00208-8.

Supplementary Material 1. Table S1. District wise rainfall (mm) and rainy days of the study area.

Acknowledgements
Authors are grateful to acknowledge the FASAL-India Meteorological Department, New Delhi and the Directorate of Economics and Statistics, Bangalore for providing the weather and yield data.

Authors' contributions
Thimmegowda MN and Manjunatha MH: resources, conceptualization, validation. Lingaraj H and Soumya DV: analysis, investigation, original draft preparation. Satish GS and Nagesha L: data curation, original draft preparation. Jayaramaiah R: review and editing, visualization, supervision. All authors have read and agreed to the published version of the manuscript.

Funding
This study was funded through India Meteorological Department, New Delhi, India under the Forecasting Agricultural output using Space, Agrometeorology and Land based observations (FASAL) project and fund number: No. ASC/FASAL/KT-11/01/HQ-2010).

Data availability
The data that supports the findings of this study can be provided upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]All India Co-ordinated Research Project on Agrometeorology, University of Agricultural Sciences, Bengaluru, Karnataka 560065, India.

Received: 28 February 2024 Accepted: 27 November 2024
Published online: 24 February 2025

References

Abrouguia K, Gabsib K, Mercatorisc B, et al. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil Tillage Res. 2019;190:202–8. https://doi.org/10.1016/j.still.2019.01.011.

Ağbulut Ü, Gürel AE, Ergün A, et al. Performance assessment of a V-Trough photovoltaic system and prediction of power output with different machine learning algorithms. J Clean Prod. 2020;268:122269. https://doi.org/10.1016/j.jclepro.2020.122269.

Ağbulut Ü, Gürel AE, Sarıdemir S. Experimental investigation and prediction of performance and emission responses of a CI engine fuelled with different metal-oxide based nanoparticles–diesel blends using different machine learning algorithms. Energy. 2021;215: 119076. https://doi.org/10.1016/j.energy.2020.119076.

Alvarez R. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. Eur J Agron. 2009;30(2):70–7. https://doi.org/10.1016/j.eja.2008.07.005.

Aslam S, Khan SH, Ahmed A, et al. The tale of cotton plant: from wild type to domestication, leading to its improvement by genetic transformation. Am J Mol Biol. 2020;10:91–127. https://doi.org/10.4236/ajmb.2020.102008.

Baigorria GA, Chelliah M, Mo KC, et al. Forecasting cotton yield in the southeastern United States using coupled global circulation models. Agron J. 2010;102:187–96.

Bakhsh K, Hassan I, Maqbool A. Factors affecting cotton yield: a case study of Sargodha (Pakistan). J Agric Soc Sci. 2005;1:332–4.

Balabin RM, Lomakina EI, Safieva RZ. Neural network (ANN) approach to biodiesel analysis: analysis of biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. Fuel. 2011;90:2007–15. https://doi.org/10.1016/j.fuel.2010.11.038.

Bali N, Singla A. Deep learning based wheat crop yield prediction model in Punjab region of North India. Appl Artif Intell. 2021;35(15):1304. https://doi.org/10.1080/08839514.2021.1976091.

Basir MS, Chowdhury M, Islam MN, et al. Artificial neural network model in predicting yield of mechanically transplanted rice from transplanting parameters in Bangladesh. J Agric Food Res. 2021;5:100186. https://doi.org/10.1016/j.jafr.2021.100186.

Behroozi-Khazaei N, Nasirahmadi A. A neural network-based model to analyze rice parboiling process with small dataset. J Food Sci Technol. 2017;54(8):2562–9. https://doi.org/10.1007/s13197-017-2701-x.

Bocca FF, Rodrigues LHA. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Comput Electron Agric. 2016;128:67–76. https://doi.org/10.1016/j.compag.2016.08.015.

Chaudhry IS, Khan MB. Factors affecting cotton production in Pakistan: empirical evidence from Multan District. J Qual Technol Manag. 2009;5:91–100.

Chipanshi A, Zhang Y, Kouadio L, et al. Evaluation of the integrated Canadian crop yield forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric for Meteorol. 2015;206:137–50. https://doi.org/10.1016/j.agrformet.2015.03.007.

Das B, Sahoo RN, Pargal S, et al. Comparison of different uni-and multi-variate techniques for monitoring leaf water status as an indicator of water-deficit stress in wheat through spectroscopy. Biosyst Eng. 2017;160:69–83. https://doi.org/10.1016/j.biosystemseng.2017.05.

Das B, Nair B, Reddy VK, et al. Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. Int J Biometeorol. 2018;62(10):1809–22. https://doi.org/10.1007/s00484-018-1583-6.

Dharmaraja S, Jain V, Anjoy P, et al. Empirical analysis for crop yield forecasting in India. Agricultural Res. 2020;9:132–8.

Dhekale BS, Sawant PK, Upadhye T. Weather based pre-harvest forecasting of rice at Kolhapur (Maharashtra). Trends Biosci. 2014;7:39–41.

Fang X, Li X, Zhang Y, et al. Random forest-based understanding and predicting of the impacts of anthropogenic nutrient inputs on the water quality of a tropical lagoon. Environ Res Lett. 2021;16(5):055003. https://doi.org/10.1088/1748-9326/abf395.

Haghverdi A, Washington-Allen RA, Leib BG. Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. Comput Electron Agric. 2018;152:186–97. https://doi.org/10.1016/j.compag.2018.07.021.

Hara P, Piekutowska M, Niedbała G. Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. Land. 2021;10(6):609. https://doi.org/10.3390/land10060609.

Howden M. Climate change and its implications for cotton production. In: Proceedings of 14th Australian Cotton Conference. Queensland. 2008. p. 12–14.

Hu T, Zhang X, et al. Climate change impacts on crop yields: a review of empirical findings, statistical crop models, and machine learning methods. Environ Model Softw. 2024;179:106119. https://doi.org/10.1016/j.envsoft.2024.106119.

Khaki S, Pham H, Wang L. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Sci Rep. 2021;11(1):11132. https://doi.org/10.1038/s41598-021-89779-z.

Kogan F, Kussul NN, Adamenko TI, et al. Winter wheat yield forecasting: a comparative analysis of results of regression and biophysical models. J Autom Inf Sci. 2013;45:68–81.

Krishna MS, Reddy YR, Chandrayudu E. Impact of weather parameters on seasonal incidence of insect pests in Bt and non bt cotton. J Pharmacogn

Phytochem. 2020;9(6):696–701. https://doi.org/10.22271/phyto.2020.v9.i6j.13023.
Kuhn M. Building predictive models in R using caret package. J Stat Softw. 2008;28:1–26.
Kumar N, Pisal RR, Shukla SP, et al. Regression technique for South Gujarat. MAUSAM. 2014;65:361–4.
Kumar R, Kumar P, Kumar Y. Time series data prediction using IoT and machine learning technique. Procedia Comp Sci. 2020;167:373–81. https://doi.org/10.1016/j.procs.2020.03.24.
Lawrence J. Introduction to neural networks: design, theory, and applications. 6th ed. Nevada City, CA, USA: California Scientific Software; 1994.
Li A, Liang S, Wang A, et al. Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photogramm Eng Remote Sens. 2007;73:1149–57.
Manideep APS, Kharb SA. Comparative analysis of machine learning prediction techniques for crop yield prediction in India. Turk J Comput Math Educ. 2022;13:120–33.
Mao LL, Guo WJ, Yuan YC, et al. Cotton stubble effects on yield and nutrient assimilation in coastal saline soil. Field Crops Res. 2019;239:71–81. https://doi.org/10.1016/j.fcr.2019.05.004.
Mehta SC, Pal S, Kumar V. Weather based models for forecasting potato yield in Uttar Pradesh. New Delhi, India: Indian Agricultural Statistics Research Institute; 2010.
Mkhabela MS, Bullock P, Raj S, et al. Crop yield forecasting on the Canadian prairies using MODIS NDVI data. Agric Meteorol. 2011;151:385–93.
Niedbala G. Application of artificial neural networks for multi-criteria yield prediction of winter rapeseed. Sustainability. 2019;11:533. https://doi.org/10.3390/su11020533.
Paswan RP, Begum SA. Regression and neural networks models for prediction of crop production. Int J Sci Eng Res. 2013;4(9):98–108.
Piaskowski JL, Brown D, Campbell KG. Near-infrared calibration of soluble stem carbohydrates for predicting drought tolerance in spring wheat. Agron J. 2016;108:285–93. https://doi.org/10.2134/agronj2015.0173.
Pokhrel BK, Paudel KP, Segarra E. Factors affecting the choice, intensity, and allocation of irrigation technologies by U.S. cotton farmers. Water. 2018;10:706. https://doi.org/10.3390/w10060706.
Rai KK, Bharti NPV. Pre-harvest forecast models based on weather variable and weather indices for Eastern UP. Adv Biores. 2013;4:118–22.
Sawan ZM. Cotton production and climatic factors: studying the nature of its relationship by different statistical methods. Cogent Biol. 2017;3:1292882.
Seiler RA, Kogan F, Wei G. Monitoring weather impact and crop yield from NOAA AVHRR data in Argentina. Adv Space Res. 2000;26:1177–85.
Shaikh S, Gala J, Jain A, et al. Analysis and prediction of covid-19 using regression models and time series forecasting. In: 2021 11th international conference on cloud computing, data science & engineering (Confernce). New Jersey: IEEE; 2021. p. 989–95.
Sharma SK, Bhagat DV, Ranjeet PD, et al. Soybean and wheat crop yield forecasting based on statistical model in Malwa agroclimatic zone. Int J Chemic Stud. 2018;6(4):1070–3.
Shivaray Navi S, Kumar C, Somu G, et al. Population dynamics of insect pests of cotton in southern dry zone of Karnataka. J Entomol Zool Stud. 2021;9(1):1402–5.
Singh RS, Patel C, Yadav MK, et al. Yield forecasting of rice and wheat crops for eastern Uttar Pradesh. J Agrometeorol. 2014;16:199–202.
Starks PJ, Steiner JL, Neel JPS, et al. Assessment of the standardized precipitation and evaporation index (SPEI) as a potential management tool for grasslands. Agronomy. 2019;9:235. https://doi.org/10.3390/agronomy9050235.
Tang LS, Li Y, Zhang JH. Partial root zone irrigation increases water use efficiency, maintains yield and enhances economic profit of cotton in arid area. Agr Water Manage. 2010;97(10):1527–33. https://doi.org/10.1016/j.agwat.2010.05.006.
Uno Y, Prasher SO, Lacroix R, et al. Artificial neural networks to predict corn yield from compact airborne spectrographic imager data. Comput Electron Agric. 2005;47(2):149–61.
Vashisth A, Goyal A, Roy D. Pre harvest maize crop yield forecast at different growth stage using different model under semi-arid region of India. Int J Tropi Agric. 2018;36(4):915–20.
Verma U, Piepho HP, Goyal A, et al. Role of climatic variables and crop condition term for mustard yield prediction in Haryana. Int J Agric Stat Sci. 2016;12:45–51.
Wang YP, Chang KW, Chen RK, et al. Large-area rice yield forecasting using satellite imageries. Int J Appl Earth Obs Geoinf. 2010;12:27–35.
Wang M, Tu L, Yuan D, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat Genet. 2019;51:224–9. https://doi.org/10.1038/s41588-018-0282-x.
Wu F, Qiu Y, Huang W, et al. Water and heat resource utilization of cotton under different cropping patterns and their effects on crop biomass and yield formation. Agric for Meteorol. 2022;323:109091. https://doi.org/10.1016/j.agrformet.2022.109091.
Yang T, Asanjan AA, Welles E, et al. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour Res. 2017;53:2786–812. https://doi.org/10.1002/2017WR020482.
Yildirim T, Moriasi DN, Starks PJ, et al. Using artificial neural network (ANN) for short-range prediction of cotton yield in data-scarce regions. Agronomy. 2022;12: 828. https://doi.org/10.3390/agronomy12040828.
Zhang GP. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 2003;50:159–75.