Macroscale soybean yield estimation using multi-model approach in Beed district of Maharashtra, India

BHARGAV SONAWANE, UPASANA SINGH, PRIYANKA SHAMRAJ[*] and ASHUTOSH PAWAR

Department of Remote Sensing and GIS, Semantic Technologies and Agritech Services Pvt. Ltd., Pune.
**Corresponding Author: priyanka.shamraj@semantictech.in*

ABSTRACT

This study focuses on estimating soybean crop yield in Beed district of Maharashtra following the methodology outlined by the government norms in insurance aspects. The research addresses significant weather-induced yield losses in the region and targets Revenue Circle (RC) level assessment using a multi-model approach, incorporating various models for precise yield forecasting. The achieved accuracy, measured with root mean square error (RMSE) below ±30% at the RC level, demonstrates the effectiveness of the ensemble approach. The findings highlight the utility of such models in decision-making for agricultural stakeholders, insurance companies, and government policies, especially in rainfed regions facing soybean productivity challenges under diverse climate change scenarios.

Keywords: Remote sensing, GIS, Net primary productivity (NPP), Machine learning, DSSAT, Yield simulation, Revenue circle, Soybean productivity.

INTRODUCTION

Accurate crop yield estimation is crucial for various stakeholders in the agricultural sector. Traditional methods struggle to account for the dynamic nature of modern agriculture with unpredictable weather patterns and environmental challenges. Fortunately, advancements in technologies like remote sensing, GIS, and AI/ML algorithms offer unprecedented capabilities for crop yield estimation (Goodwin & Hungerford, 2009). Accurate estimates enable insurers to assess risks effectively and design targeted products that mitigate financial burdens on farmers during crop failures (Mahlein *et al.*, 2018). Reliable predictions inform commodity markets, trade agreements, and pricing mechanisms, promoting stability and ensuring food security (Goodwin & Hungerford, 2009). Governments leverage accurate estimates to formulate effective policies for subsidy allocation, resource distribution, and strategic interventions during crises. Anticipating potential shortfalls supports proactive food distribution, enhancing access and averting scarcity (Mueller *et al*., 2013). Precise estimates empower farmers to make informed decisions on crop selection, resource allocation, and market participation, ultimately enhancing productivity and livelihoods (Singh *et al.* 2023).

The integration of advanced technologies revolutionizes crop yield estimation. Specialized software streamlines data collection, analysis, and visualization for informed decision-making. Satellite and aerial imagery provide valuable insights on crop health, growth stage, and potential yield based on spectral reflectance (Patel *et al*. 2023). GIS platforms integrate spatial data from various sources like remote sensing and weather stations to create comprehensive yield maps. Machine learning algorithms analyze vast datasets from remote sensing, weather data, and historical yields to predict future yields with high accuracy (Dadhwal and Bhat 2023).

The research paper outlines a case study for Beed district, focusing on: 1) Estimating the area under major *kharif* crops. 2) Crop classification using remote sensing and GIS techniques. 3) Yield estimation through a combination of models including remote sensing, GIS, AI, Google Earth Engine, ground truth data, and DSSAT (Decision Support System for Agrotechnology Transfer) software. This case study exemplifies the practical application of advanced methods for crop yield estimation in a specific region.

MATERIAL AND METHODS

Study area

The study was carried out at Semantic Technologies and Agritech Services, Pvt. Ltd Pune during *kharif* season 2023 for a particular assignment. For this study, all revenue circles (RC) in the districts of Beed of Maharashtra state were used as experimental sites. Field-level data like ground truth, crop cutting experiments were carried out.

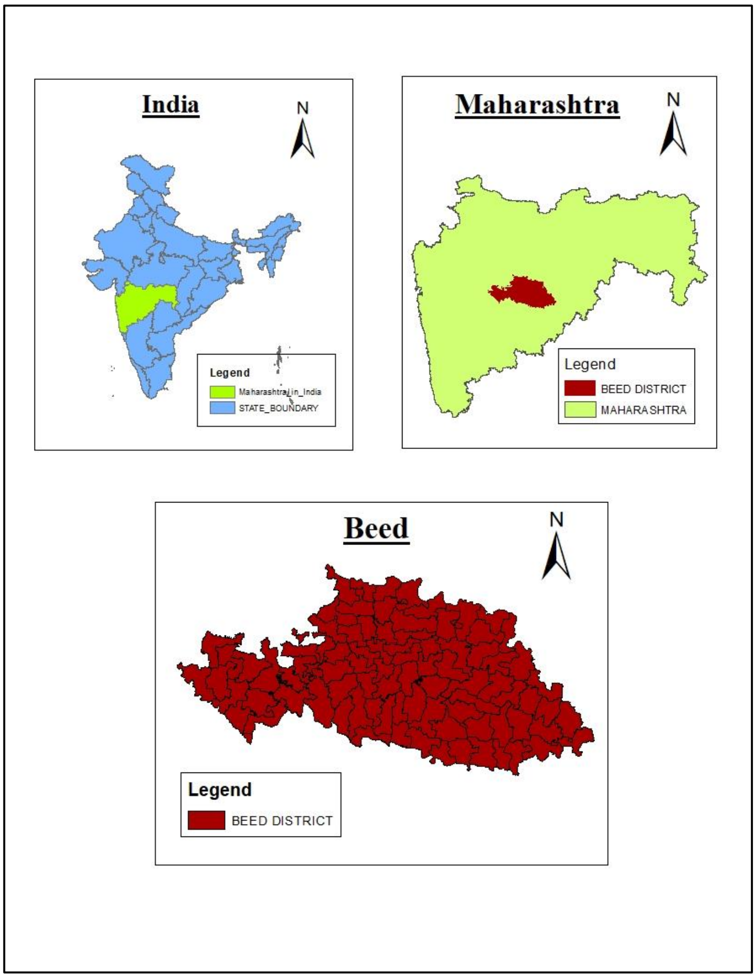


Fig. 1: Study area

Geography and climate of Beed district

Beed district, located in the state of Maharashtra, India, spans an area of approximately 10,693 square kilometres (Fig. 1). Its geographical coordinates are approximately 18.99° N latitude and 75.75° E longitude, with an average elevation of 540 meters above sea level. The district experiences a semi-arid climate with hot summers and cool winters. The annual rainfall typically ranges from 600 to 800 millimetres, primarily occurring during the monsoon season. Temperatures vary widely throughout the year, with average highs peaking around 40°C during the summer months, while winter temperatures can drop to around 10°C in December and January. Humidity levels tend to be relatively lower during the drier months, especially in winter, with higher humidity levels experienced during the monsoon season. The predominant soil types include black soil, red soil, and alluvial soil, supporting the cultivation of crops such as cotton, sorghum, pulses, and soybeans. Beed is bordered by the districts of Ahmednagar, Osmanabad, Aurangabad, and Jalna. The major rivers flowing through the district include the Godavari and the Sindhphana.

Methodology

All methodology was followed by the procedure given by the yield estimation system based on technology (yes-tech) under Pradhan Mantri Fasal Bima Yojana (PMFBY). The methodology used is a multimodal approach for the estimation of crop yield was given below. RC wise yield in t ha^{-1} of soybean crop during *kharif* season 2023 was estimated by three approaches using net primary productivity (NPP), crop simulation model (DSSAT), machine learning and then ensemble model.

Net primary productivity (NPP)

The net primary productivity was computed following the method given by Singh *et al*. (2023). The data and materials used in this study are presented in Table 1. The fraction of absorbed photosynthetically active radiation (FAPAR) data was obtained from Copernicus Land Service (https://land.copernicus.eu/global/index.html). The 10-day composite product with 1 km data was used. The range of FAPAR lies between 0 and 1. The physical values were retrieved from the digital number (DN). The photosynthetically active radiation (PAR) was calculated from daily insolation data. The daily insolation data was converted to 8 - day composite (sum) for the whole period. 50% insolation was considered as PAR. The daily insolation data was collected from MOSDAC from INSAT-3D satellite (www.mosdac.gov.in) for the crop season from 2018 to 2022.

PAR= 8 - day composite * 0.5.

The water stress (Wstress) was calculated from Land Surface Water Index (LSWI). The MODIS time series tool (MODIStsp) was used to download and process the MODIS 8-day composite (MOD09A1) (https://lpdaac.usgs.gov/products/mod09a1v006), and LSWI was calculated for the entire period with the formula

$$LSWI = (\rho \boldsymbol{NIR} - \rho \boldsymbol{SWIR}) / (\rho \boldsymbol{NIR} + \rho \boldsymbol{SWIR})$$

LSWI values range from - 1 to 1, and higher positive values indicate the vegetation and soil water stress. Further, the Wstress is calculated from 8 days of LSWI output –

$$Wstess = (\boldsymbol{1} - \boldsymbol{LSWI}) / (\boldsymbol{1} + \boldsymbol{LSWImax})$$

Table 1: Data used for NPP generation in semi semi-physical model

The

Data	Satellite/Ground	Resolution	Source
Daily insolation/PAR	INSAT-3D	4km resampled to 1km	MOSDAC
10 days composite fAPAR ver. 2	PROBA V and SPOT-VGT	1km	Copernicus Land Service
8 days composite surface reflectance	Terra-MODIS	1km	MODIS Time Series Tool
Paddy Mask	Sentinel 1	5m	USGS Explorer
Temperature	Gridded data from NASA Power website	1km interpolated	NASA Power
Light-use efficiency			Literature
Harvest Index	Ground	CCE	

LSWI max value has been taken from the spatial maximum of a particular crop mask of the entire district. The temperature stress (Tstress) was calculated using daily average temperature data downloaded from NASA power website (https://power.larc.nasa.gov/data-access-viewer.html). It is a gridded data with a resolution of 1°0 * 1°0 latitude and longitude.

$$Tstress = \frac{(T - Tmin)*(T - Tmax)}{[(T - Tmin)*(T - Tmax) - T - Topt)^2]}$$

Where, Tmin = Minimum temperature (°C); Tmax =Maximum temperature (°C); Topt = Optimal temperature (°C); T = Daily mean temperature (°C). Temperature values used for calculation were Tmax, Tmin, and Topt. were 35°C, 10°C and 26°C respectively (Nimje, 2022). On the off chance that air temperature falls beneath Tmin, which is quite a rare chance then Tscalar value will automatically become 0. The light use efficiency (LUE) used for soybean crop was 1.78 for the study (Chavan *et al.,* 2018). The crop mask was derived utilizing Sentinel-1 synthetic aperture radar (SAR) data obtained from the European Space Agency (ESA) Copernicus Hub. Employing the R programming language, we employed the Random Forest algorithm for the generation of the crop mask, implementing hyperparameter tuning techniques and contingency matrix analysis. To compute the final net primary productivity (NPP) and its grain yield, the NPP sum was multiplied by harvest index (0.45) (as per periodic CCE data) to estimate per pixel yield.

NPP *= PAR * FAPAR * Ɛ * Tstress * Wstress* (Monteith, 1972).

Crop simulation model-DSSAT-4.8

We used CROPGRO – for the soybean crop for which weather, soil, crop management data were used to calibrate and validate the model.

Weather data: The input parameter on weather viz. rainfall, solar radiation, maximum temperature and minimum temperature for the last 30 years were collected from NASA Power (https://power.larc.nasa.gov/). A separate weather file for each revenue circle (RC) was generated using the weatherman interface in DSSAT.

Soil data: Soil data was taken from DSSAT website (HC27 data) where Global gridded-soil profile dataset at 10 by 10 km (https://dssat.net/277/) resolution was developed for DSSAT crop simulation models. When we transfer soil data to software files it will automatically create new files with identical IDs of location in the S Build interface of DSSAT software. Importantly, these villages had different rainfall levels, soil types, and elevations.

Crop management data: The crop management data was collected from all revenue circle, where crop-cutting experiments were taken for CropTech Application demonstration and by registered farmers for this app. All basic crop management data were collected by farmers who were registered for the company CropTech Application. Data required like date of sowing, crop and row spacing, plant population, fertilizer applied, variety used, chemical applied etc. The data which is not available at CropTech app was taken from crop management practices given by VNMKV, Parbhani,

Ground data points have been optimized based on criteria, including soil type, rainfall, GIS location and elevation map. X build is the interface in which actual experimental file is present where all data regarding crop management was filled and saved.

Cultivar and Genetic Coefficients**:** The genetic coefficients are the most important parameters that represent the genetic characteristics of the cultivar and on which the crop phenology, biomass production partitioning, and yield potential of the crop depend. However, the actual performance is controlled by the external factors also. In Beed district soybean cultivars JS-335, JS-9305, and TAMS 98- 21 were used mostly by farmers. The genetic coefficients of the soybean varieties JS-335, JS-9305, and TAMS 98-21 were taken from already available in literature published by VNMKV, Parbhani Agriculture University. Genetic coefficients were used to simulate the response of various cultivars to weather and management conditions. The observed experimental data of yield were compared with the model simulation results. The evaluation of the model on an overall basis revealed that the model simulation performance in respect of yield was found to be reliable.

Calibration and Validation of CROPGRO**:** Genetic coefficients were developed for different soybean varieties for DSSAT model validation. Calibration of model was carried out with genetic coefficient file, weather file, soil file, experiment file of all RC level stations. Validation is the comparison of the results of model simulations with observations that were not used for the calibration. The experimental data collected will be used for independent model validation.

Once, all the desired files were created carefully, the model was run for all RCs for soybean crops. Each run of the model created output files containing the yield of a particular location.

Remote sensing approach

We integrated Sentinel-2 and Sentinel-1 imagery from the Copernicus mission to capture both optical and radar data for the designated Area of Interest (AOI). From Sentinel-2 imagery, three vegetation indices were calculated: Normalized Difference Vegetation Index (NDVI), Green Normalized Difference Vegetation Index (GNDVI), and Normalized Difference Red Edge Index (NDRE). Sentinel-1 imagery provided backscatter coefficient values (VV and VH polarization). Pre-processing steps on the optical data included atmospheric and geometric corrections, followed by cloud masking using the Sentinel-2 QA band to minimize cloud influence. Backscatter data from Sentinel-1 was converted from decibels (dB) to natural units and filtered using the Refined Lee filter for speckle reduction.

Crop mask generation**:** A multi-step approach was employed to delineate soybean fields within pre-processed Sentinel-1 SAR data (obtained from ESA Copernicus Hub) using R software. This involved atmospheric and geometric corrections, followed by image enhancement and supervised classification with the Random Forest algorithm. Hyperparameter tuning techniques and contingency matrix analysis ensured optimal model performance. This methodology was applied systematically across all specified crops within the targeted area of interest.

Data extraction from crop mask**:** Following crop mask generation, we extracted relevant data layers specifically for the delineated soybean fields. This included the pre-calculated vegetation indices (NDVI, GNDVI, and NDRE) from Sentinel-2 imagery and backscatter coefficient values (VV and VH polarization) from Sentinel-1 imagery. By focusing on data within the soybean mask, we ensured our analysis targeted the crop of interest and minimized the influence of surrounding land cover types.

Normalization**:** To facilitate standardized comparisons across the study area, all extracted data layers (NDVI, GNDVI, NDRE, VV, and VH) were normalized to a common range of 0 to 1. A custom function calculated minimum and maximum values within each data layer (excluding No Data values) for proportional rescaling. This normalization addressed inherent variations in the original data scales and ensured a consistent basis for evaluating vegetation health indicators and backscatter values across the soybean fields.

Zonal statistics with mean function**:** Zonal statistics techniques within ArcGIS software were employed to calculate the mean for each data layer across user-defined zones (Revenue Circles or RCs). Zonal statistics with the mean function summarized data within these zones, providing a representative value reflecting the average condition of the targeted parameter (e.g., average NDVI) within each RC. This approach allowed us to capture the central tendency of vegetation health indicators and backscatter values across the soybean fields within each RC.

Data export**:** The resulting zonal statistics were exported to a comma-separated values (CSV) format for further analysis and visualization. This CSV file served as the foundation for subsequent steps, enabling the exploration of relationships between vegetation health indicators, backscatter values, and potential influencing factors like soil moisture or agricultural practices across the RCs within the study area.

Machine learning models

The final step of our analysis involved leveraging machine learning techniques for crop yield estimation. This section details the process of model selection, training, evaluation, and prediction.

Data preprocessing and splitting**:** The pre-processed data, containing zonal statistics for vegetation health indicators (NDVI, GNDVI, NDRE) and backscatter coefficient values (VV, VH) across revenue circles (RCs), was loaded into a Python environment using libraries like pandas. The data was then split into training and testing sets using sci-kit-learn's train_test_split function. A common split ratio is 80% for training and 20% for testing. The training set serves to train the machine learning models, while the testing set is used for unbiased evaluation of their performance.

Model selection and training**:** Three common regression models were employed for yield estimation. The linear regression (LR) is a baseline model that establishes a linear relationship between the features (vegetation indices and backscatter values) and the target variable (crop yield data, CCE). Support vector regression (SVR) model can handle non-linear relationships between features and the target variable and is robust to outliers and the random forest regression (RF) is ensemble method that combines multiple decision trees, leading to improved prediction accuracy and robustness compared to a single decision tree. Each model was implemented within a scikit-learn pipeline incorporating a standard scaler for normalization. Normalization ensures all features are on a similar scale, improving model performance.

Model evaluation and selection: The performance of each model was evaluated on the testing set using the coefficient of determination (R-squared) metric. R-squared represents the proportion of variance in the actual yield data explained by the model's predictions. A higher R-squared value indicates a better fit between predicted and actual yield values. The model with the highest R-squared score on the testing set was chosen as the optimal model for yield estimation.

Prediction and output generation: The chosen best-performing model was then used to predict yield (CCE) values for all RCs based on the feature data (NDVI, GNDVI, NDRE, VV, VH) in the testing set. The original dataset was then augmented with a new column containing the predicted yield values from the best-performing model. This allows for easy comparison of actual and predicted yield values for each RC. The best-performing model was saved using a library like Joblib for potential future use or retraining.

Finally, the modified dataset, incorporating the predicted yield values, was exported as a new CSV file for further analysis and visualization. This file serves as a foundation for exploring the spatial distribution of predicted yield across the study area and for relating yield predictions to other relevant factors.

Ensemble model development

The innovative approach integrates Machine Learning (ML), Crop Simulation Models (CSM), and semi-physical models to enhance yield prediction accuracy, tailored for the Beed district's agricultural landscape. It begins with three individual models:

Firstly, the Machine Learning Model undergoes meticulous training using algorithms like linear regression, Random Forest, etc., on Beed-specific datasets. Rigorous validation techniques ensure the model's ability to discern patterns without overfitting.

Secondly, the Crop Simulation Model (DSSAT) is calibrated meticulously using Beed's crop and environmental data, focusing on the Kharif-2023 season. This fine-tuning enables accurate simulation of crop growth and yield dynamics.

Thirdly, a Semi-physical Model leverages the relationship between remotely sensed data and biophysical parameters, trained specifically for Beed. It establishes a predictive link between remote data and actual yield outcomes.

Each model's performance is evaluated using metrics like RMSE and R^2 on Beed-specific hold-out test sets. This comprehensive evaluation informs subsequent ensemble model construction.

Three prominent ensemble techniques are considered: Weighted Averaging, Stacking, and Voting. Weighted Averaging assigns weights based on model performance, ensuring more accurate models contribute more significantly. Stacking creates a meta-model trained on district-specific datasets, optimizing combinations of individual model predictions. Voting determines the ensemble yield based on the most frequent prediction among individual models.

Following ensemble model construction, rigorous validation and quality control measures are implemented. Separate hold-out test sets for Beed district validate the ensemble model's generalizability and accuracy, with metrics like RMSE and R^2 computed.

Quality control ensures reliability and robustness. Normalized RMSE comparisons between observed yield data and predicted yields refine the ensemble model. If the RMSE exceeds a predefined threshold, further refinement may involve adjusting weights, exploring alternative ensemble techniques, or gathering additional data.

Overall, this approach optimizes agricultural decision-making by providing accurate yield predictions tailored for Beed's unique characteristics. By integrating diverse models and leveraging advanced ensemble techniques, it represents a significant advancement in agricultural modeling and decision support systems.

RESULTS AND DISCUSSION

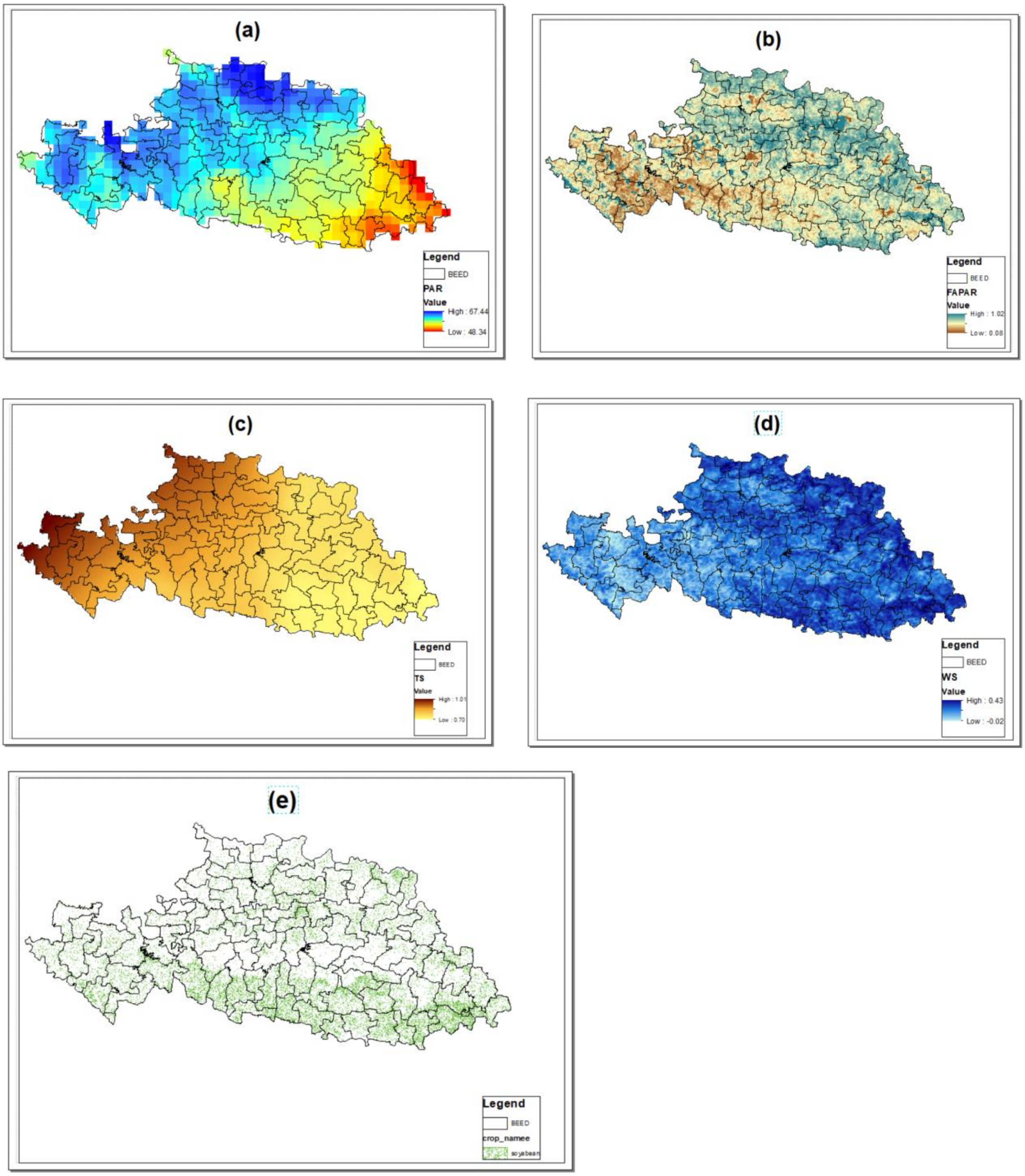


Fig. 2: Spatial distribution of derived parameters in Beed district (a) PAR, (b) FAPAR, (c)Temperature stress, (d)Water stress and (e) Crop mask

In figure A, we have generated the PAR data for Beed district where high indicates high absorption of data for photosynthesis by plants which can cause inhibition if shallow water bodies are present, and low signifies less absorption with chances of equal distribution.

In figure B, FAPAR data has been mapped for Beed District, which signifies good condition of crops of the region with fraction consumption of active radiation from range 0.8 to often close to 1, whereas less values indicate average condition especially if presence of lawns or meadows if there.

In figure C, High-temperature (HT) stress is frequently defined when temperature raises beyond the level of a threshold for a certain period of time and abundantly causes irreversible impairment to the growth and development of plants.

In figure D, Severe water stress may result in the arrest of photosynthesis, disturbance of metabolism and finally the death of plant.

Figure E, is the Crop Mask Map for Soyabean crop for Beed District.

Estimated yield using NPP

Soybean crop yields in Beed district for the year 2023 varied widely across Revenue Circles, with CCE yields ranging from 0.72 to 2.56 t ha-1. The Semi-Physical Yield estimates also showed significant variation, with values ranging from 1.07 to 3.17 t ha-1. Revenue Circles like Jategaon, Gangamasla, and Talwada demonstrated high actual and Semi-Physical Yields, indicating successful soybean cultivation practices. Some areas, such as Chousala, Mahlas Jawala, and Nalwandi, experienced lower actual and Semi-Physical Yields, suggesting potential challenges in soybean production. Overall, there is notable inconsistency between the actual yields and Semi-Physical Yield estimates, indicating that some Revenue Circles may benefit from improvements in soybean cultivation techniques. The data highlights the need for localized strategies and interventions to optimize soybean crop yields in different areas of Beed district. Same results were reported by Xiao, *et al.* (2006) and Yao, *et al.* (2021)

Our results yielded a robust accuracy (Crop Mask) range of 90% to 95% across cultivated crops and various districts, signifying high precision in crop delineation and classification.

RC wise Crop Mask of Soyabean of Beed District 2023 Kharif

Prediction	Soybean	Black Gram	Forest	Roadways	Settlements	Sugarcane	Waterbody	Row Total	User Accuracy
Soybean	**4463**	**12**	**1**	**10**	**39**	**531**	**20**	**5075**	**87.9211**
Black Gram	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**NaN**
Forest	**0**	**0**	**11**	**0**	**4**	**0**	**0**	**15**	**73.33333**
Roadways	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**NaN**
Settlements	**24**	**0**	**37**	**0**	**1083**	**8**	**0**	**1152**	**94.01041**
Sugarcane	**41**	**0**	**0**	**1**	**3**	**134**	**3**	**183**	**73.2240**
Waterbody	**0**	**0**	**0**	**0**	**0**	**0**	**5702**	**5702**	**100**

Kappa statistics = 0.90

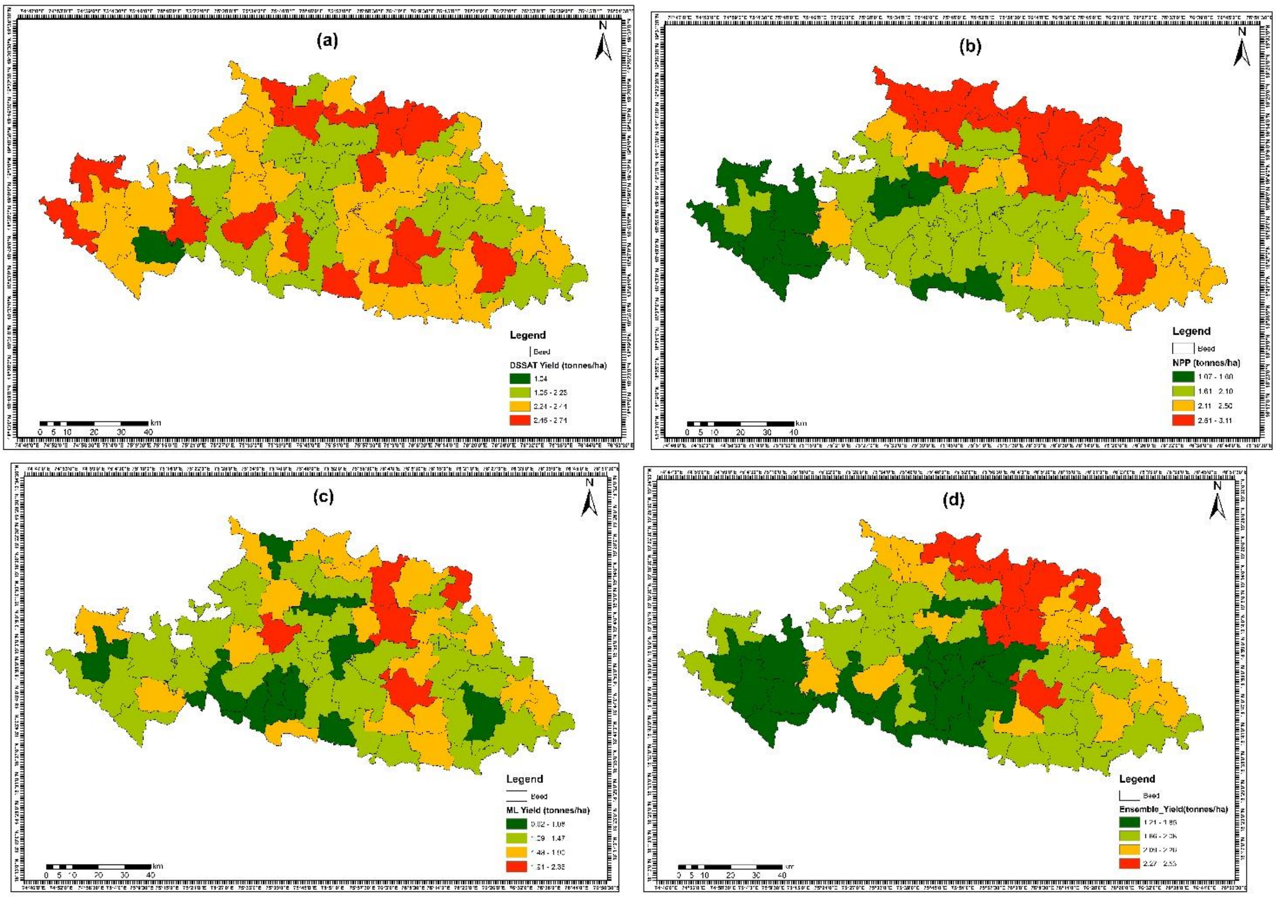
Overall Classification Accuracy = 87.92%

Satellite Data Used – Sentinel -1 (From July to September)

Algorithm used – Random Forest.

Estimated yield using crop simulation model

The District-wise Crop Simulation System (DSSAT) yield estimates were generally higher than the actual yields, indicating potential room for improvement in soybean production practices.



(c) **(d)**

Fig. 3: Soybean yield in Beed district (a) DSSAT, (b) NPP, (c)Machine learning and (d)Ensemble approach

Chinchwan, Ghatsawli, and Pendgaon were among the Revenue Circles with high actual yields, ranging from 2.43 to 2.46 t ha-1. Some areas, such as Ashti and Mahlas Jawala, experienced lower actual yields compared to DSSAT estimates, suggesting challenges or limitations in soybean cultivation in those regions. Overall, the Beed district exhibited a range of soybean crop yields, emphasizing the importance of local factors and agricultural practices in influencing production outcomes. Jadhav, *et al.* (2018), Bhosale, *et al*. (2015) and Deshmukh, *et al.* (2013) also elaborated same results for soybean.

Estimated yield using machine learning

CCE yield and different indices under study showing accuracy 82 % in Machine learning model. By the method (SVR) Support Vector Regression accuracy is showing highest value. The ML (Machine Learning) yield estimates also showed variation, with values ranging from 1.12 to 3.17 t ha-1. Revenue Circles like Jategaon, Dindrud, and Mahlas Jawala demonstrated relatively higher actual and ML yields, suggesting successful soybean cultivation practices in those areas. Conversely, Revenue Circles such as Chousala, Anjandhav, and Nagapur exhibited lower actual and ML yields, indicating potential challenges or limitations in soybean production. Overall, there seems to be a discrepancy between the actual yields and ML yield estimates across various Revenue Circles, indicating the need for further investigation into factors influencing soybean production in Beed district.

The data underscores the importance of implementing localized strategies and agricultural practices to optimize soybean crop yields and address production challenges in different areas of Beed district.

Estimated yield using ensemble model

The Ensemble Yield represents a combination of all above three predictive models or methods to estimate soybean crop yield.

Statistical approach give weightage during *kharif* 2023 as following to different models.

Model Used	**DSSAT Yield**	**Semi-Physical Yield**	**Machine Learning Yield**
Weightages in %	36.72	33.03	30.25

Ensemble Yield estimates also varied significantly, with values ranging from 1.33 to 3.17 t ha-1across different Revenue Circles. Revenue Circles like Jategaon, Dhondrai, and Talkhed demonstrated relatively higher actual and Ensemble Yield, suggesting successful soybean cultivation practices in those areas. Conversely, Revenue Circles such as Chousala, Anjandhav, and Nagapur exhibited lower actual and Ensemble Yield, indicating potential challenges or limitations in soybean production.

Overall, there appears to be a discrepancy between the actual yields and Ensemble Yield estimates across various Revenue Circles, indicating the need for further investigation into factors influencing soybean production in Beed district. Same results were given by Md Didarul Islam, *et.al* (2023), Liujun Xiao, *et.al*. (2022) and Ayan Das, *et.al* (2023) in both Machine learning and ensemble approach.

The yield estimated by various methods and Actual field CCE yield is presented. The yield by the all models with field CCE, presented in table 2 and 3. Average (RMSE) was also presented in tables. As per mentioned in deliverables in YESTECH manual given by Pradhan Mantri Fasal Bima Yojana, the error (nRMSE) between the observed and modeled yield should not be more than ±30%. From table it is cleared that all RMSE values were below 30% range, which indicates that the process adopted for RC wise soybean yield estimation is acceptable in Beed district.

Table No. 2: Yield of soybean crop in t ha^{-1} with DSSAT Models and NPP Model with RMSE for year 2023.

District	**Tehsil**	**RC**	**Field CCE**	**DSSAT Yield (tonnes/ha)**	**(RMSE)**	**NPP (tonnes/ha)**	**(RMSE)**
Beed	Patoda	Amlner	1.51	2.58	1.07	2.19	0.68
Beed	Ambejogai	Ambajogai	1.62	2.66	1.04	2.56	0.94
Beed	Ashti	Ashti	1.69	1.04	0.65	1.07	0.62
Beed	Kaij	Bansarola	1.16	2.28	1.12	1.98	0.82
Beed	Beed	Beed	2.06	2.17	0.11	1.95	0.11
Beed	Georai	Chaklamba	1.26	2.34	1.08	2.17	0.91
Beed	Beed	Chousala	0.72	2.09	1.37	1.49	0.77
Beed	Patoda	Daskhed	1.47	2.21	0.74	1.78	0.31
Beed	Ashti	Daula wadgaon	2.14	2.46	0.32	1.59	0.55
Beed	Ashti	Dhamngaon	1.17	2.34	1.17	1.52	0.35
Beed	Ashti	Dhanora	1.02	2.26	1.24	1.74	0.72
Beed	Parli	Dharmapuri	1.14	2.39	1.25	2.29	1.15
Beed	Dharur	Dharur	1.58	2.57	0.99	2.08	0.5
Beed	Georai	Dhodrai	2.54	2.5	0.04	2.58	0.04
Beed	Manjlegaon	Dindrud	1.35	2.19	0.84	2.49	1.14
Beed	Manjlegaon	Gangamasla	1.37	2.26	0.89	3.11	1.74
Beed	Georai	Georai	1.21	2.51	1.3	2.59	1.38
Beed	Ambejogai	Ghatnandur	1.46	2.06	0.6	2.44	0.98
Beed	Kaij	Hanumant pimpri	1.45	2.39	0.94	2.01	0.56
Beed	Kaij	Hoal	1.41	2.07	0.66	2.04	0.63
Beed	Georai	Jategaon	2.56	2.46	0.1	2.66	0.1
Beed	Ashti	Kada	1.67	2.28	0.61	1.54	0.13
Beed	Kaij	Kaij	1.32	2.56	1.24	2.13	0.81

Beed	Wadwani	Kawadgaon Bu	1.41	2.36	0.95	2.6	1.19
Beed	Manjlegaon	Kitti Adgaon	1.51	2.5	0.99	2.83	1.32
Beed	Beed	Limbaganesh.	2.24	2.39	0.15	2.09	0.15
Beed	Ambejogai	Lokhandi-Sawargaon	1.14	2.39	1.25	2.22	1.08
Beed	Georai	Madalmohi	1.09	2.12	1.03	2.19	1.1
Beed	Beed	Mahlas Jawala	0.85	2.23	1.38	2.5	1.65
Beed	Manjlegaon	Majalgaon	1.29	2.13	0.84	2.94	1.65
Beed	Beed	Manjarsumba	1.13	2.54	1.41	1.91	0.78
Beed	Dharur	Mohkhed	1.38	2.17	0.79	2.42	1.04
Beed	Kaij	Nadurghat	1.24	2.45	1.21	1.31	0.07
Beed	Parli	Nagapur	0.93	2.04	1.11	2.32	1.39
Beed	Beed	Nalwandi	0.91	2.26	1.35	2.1	1.19
Beed	Beed	Neknoor	1.6	2.14	0.54	1.8	0.2
Beed	Manjlegaon	Nithrud	1.25	2.33	1.08	2.54	1.29
Beed	Georai	Pachegaon	1.97	2.15	0.18	1.79	0.18
Beed	Beed	Pali	2.54	2.13	0.41	1.95	0.59
Beed	Parli	Parli	1.02	2.44	1.42	2.25	1.23
Beed	Patoda	Patoda	1.11	2.1	0.99	1.81	0.7
Beed	Ambejogai	Patoda M	1.49	2.32	0.83	2.23	0.74
Beed	Beed	Pendgaon	2.46	2.19	0.27	2.73	0.27
Beed	Parli	Pimpalgaon	1.1	2.09	0.99	2.68	1.58
Beed	Beed	Pimpalner	0.92	2.54	1.62	2.42	1.5
Beed	Ashti	Pimpla	1.49	2.74	1.25	1.58	0.09
Beed	Shirur (Kasar)	Raimoha	1.22	2.35	1.13	1.57	0.35
Beed	Beed	Rajuri (N)	1.29	2.42	1.13	1.6	0.31
Beed	Georai	Revki	1.85	2.21	0.36	2.87	1.02
Beed	Shirur (Kasar)	Shirur Kasar	1.37	2.09	0.72	2.02	0.65
Beed	Georai	Sirasdevi	1.52	2.18	0.66	2.09	0.57
Beed	Parli	Sirsala	1.06	2.42	1.36	2.76	1.7
Beed	Ashti	Takalsing	1.51	2.29	0.78	1.54	0.03
Beed	Manjlegaon	Talkhed	1.88	2.64	0.76	2.67	0.79
Beed	Georai	Talwada	1.1	2.36	1.26	3.08	1.98
Beed	Dharur	Telgaon	0.97	2.21	1.24	2.08	1.11
Beed	Patoda	Therla	0.91	2.64	1.73	2.1	1.19
Beed	Shirur (Kasar)	Tintarwani	1.8	2.32	0.52	1.82	0.02
Beed	Georai	Umapur	1.32	2.26	0.94	2.64	1.32
Beed	Wadwani	Wadwani	1.19	2.25	1.06	1.78	0.59
Beed	Kaij	VIDA	1.43	2.25	0.82	1.87	0.44
Beed	Kaij	Yusufwadgao	1.05	2.32	1.27	1.77	0.72
Average =			**1.43**	**2.30**	**0.92**	**2.15**	**0.80**

Table No. 3: Yield of soybean crop in t ha^{-1} with ML Yield Models and Ensemble Model with RMSE for year 2023.

District	Tehsil	RC	Field CCE	ML Yield (tonnes/ha)	(RMSE)	Ensemble Yield (tonnes/ha)	(RMSE)
Beed	Patoda	Amlner	1.51	1.28	0.23	2.12	0.61
Beed	Ambejogai	Ambajogai	1.62	1.01	0.61	2.21	0.59
Beed	Ashti	Ashti	1.69	1.67	0.02	1.21	0.48
Beed	Kaij	Bansarola	1.16	1.65	0.49	2.01	0.85
Beed	Beed	Beed	2.06	1.2	0.86	1.85	0.21
Beed	Georai	Chaklamba	1.26	1.19	0.07	1.99	0.73
Beed	Beed	Chousala	0.72	1.53	0.81	1.73	1.01
Beed	Patoda	Daskhed	1.47	1.05	0.42	1.76	0.29
Beed	Ashti	Daula wadgaon	2.14	1.9	0.24	2	0.14
Beed	Ashti	Dhamngaon	1.17	1.2	0.03	1.76	0.59
Beed	Ashti	Dhanora	1.02	0.94	0.08	1.74	0.72
Beed	Parli	Dharmapuri	1.14	1.7	0.56	2.18	1.04
Beed	Dharur	Dharur	1.58	2.36	0.78	2.34	0.76
Beed	Georai	Dhodrai	2.54	1.08	1.46	2.18	0.36
Beed	Manjlegaon	Dindrud	1.35	1.4	0.05	2.11	0.76
Beed	Manjlegaon	Gangamasla	1.37	2.1	0.73	2.53	1.16
Beed	Georai	Georai	1.21	1.36	0.15	2.26	1.05
Beed	Ambejogai	Ghatnandur	1.46	1.43	0.03	2.04	0.58
Beed	Kaij	Hanumant pimpri	1.45	1.28	0.17	1.98	0.53
Beed	Kaij	Hoal	1.41	1.83	0.42	2	0.59
Beed	Georai	Jategaon	2.56	1.88	0.68	2.39	0.17
Beed	Ashti	Kada	1.67	1.42	0.25	1.8	0.13
Beed	Kaij	Kaij	1.32	1.55	0.23	2.16	0.84
Beed	Wadwani	Kawadgaon Bu	1.41	2.11	0.7	2.38	0.97
Beed	Manjlegaon	Kitti Adgaon	1.51	1.59	0.08	2.4	0.89
Beed	Beed	Limbaganesh.	2.24	0.97	1.27	1.93	0.31
Beed	Ambejogai	Lokhandi-Sawargaon	1.14	1.28	0.14	2.05	0.91
Beed	Georai	Madalmohi	1.09	1.55	0.46	2	0.91
Beed	Beed	Mahlas Jawala	0.85	1.24	0.39	2.08	1.23
Beed	Manjlegaon	Majalgaon	1.29	1.36	0.07	2.23	0.94
Beed	Beed	Manjarsumba	1.13	0.62	0.51	1.84	0.71
Beed	Dharur	Mohkhed	1.38	1.23	0.15	2.03	0.65
Beed	Kaij	Nadurghat	1.24	1.08	0.16	1.7	0.46
Beed	Parli	Nagapur	0.93	1.32	0.39	1.96	1.03
Beed	Beed	Nalwandi	0.91	0.72	0.19	1.82	0.91
Beed	Beed	Neknoor	1.6	1.16	0.44	1.78	0.18
Beed	Manjlegaon	Nithrud	1.25	1.59	0.34	2.22	0.97
Beed	Georai	Pachegaon	1.97	1.03	0.94	1.74	0.23
Beed	Beed	Pali	2.54	1.27	1.27	1.85	0.69
Beed	Parli	Parli	1.02	1.74	0.72	2.2	1.18
Beed	Patoda	Patoda	1.11	0.94	0.17	1.71	0.6
Beed	Ambejogai	Patoda M	1.49	1.39	0.1	2.06	0.57

Beed	Beed	Pendgaon	2.46	1.32	1.14	2.17	0.29
Beed	Parli	Pimpalgaon	1.1	1.25	0.15	2.1	1
Beed	Beed	Pimpalner	0.92	1.76	0.84	2.3	1.38
Beed	Ashti	Pimpla	1.49	1.34	0.15	1.97	0.48
Beed	Shirur (Kasar)	Raimoha	1.22	1.74	0.52	1.92	0.7
Beed	Beed	Rajuri (N)	1.29	2.18	0.89	2.06	0.77
Beed	Georai	Revki	1.85	1.64	0.21	2.31	0.46
Beed	Shirur (Kasar)	Shirur Kasar	1.37	1.45	0.08	1.91	0.54
Beed	Georai	Sirasdevi	1.52	1.47	0.05	1.97	0.45
Beed	Parli	Sirsala	1.06	1.62	0.56	2.34	1.28
Beed	Ashti	Takalsing	1.51	1.33	0.18	1.78	0.27
Beed	Manjlegaon	Talkhed	1.88	2.15	0.27	2.53	0.65
Beed	Georai	Talwada	1.1	1.57	0.47	2.42	1.32
Beed	Dharur	Telgaon	0.97	1.81	0.84	2.07	1.1
Beed	Patoda	Therla	0.91	1.44	0.53	2.15	1.24
Beed	Shirur (Kasar)	Tintarwani	1.8	1.44	0.36	1.92	0.12
Beed	Georai	Umapur	1.32	1.61	0.29	2.24	0.92
Beed	Wadwani	Wadwani	1.19	1.25	0.06	1.83	0.64
Beed	Kaij	VIDA	1.43	1.14	0.29	1.84	0.41
Beed	Kaij	Yusufwadgao	1.05	1.44	0.39	1.9	0.85
Average =			**1.43**	**1.44**	**0.42**	**2.03**	**0.70**

CONCLUSION

This study investigated the effectiveness of various models for predicting soybean crop yields in Beed, Maharashtra, for the kharif season of 2023. The research compared the performance of three models: the Potential Production (NPP) model, the Decision Support System for Agrotechnology Transfer (DSSAT) model, and a Machine Learning model. The evaluation revealed distinct strengths and limitations in each individual model. While each captured specific aspects of crop growth dynamics, the Machine Learning model demonstrated superior adaptability and predictive accuracy.

To overcome the limitations of individual models and enhance prediction reliability, the study explored an ensemble approach. This approach combined the strengths of all three models, creating a holistic framework that leverages their individual capabilities.

The ensemble model yielded promising results, demonstrating a close alignment with field data. This highlights the potential of such ensemble models to significantly improve the accuracy of crop yield predictions. By minimizing uncertainties associated with individual models, the combined approach provides a more reliable foundation for informed decision-making in the agricultural sector.

In conclusion, this study presents a compelling case for the integration of NPP, DSSAT, and Machine Learning models into an ensemble framework for crop yield prediction. This approach offers a promising avenue for advancing prediction methodologies and ultimately empowers farmers and policymakers with valuable insights to support sustainable agricultural practices in Maharashtra. The findings serve as a foundation for further research and refinement, aiming to continuously improve the accuracy and actionable nature of these predictions.

ACKNOWLEDGMENTS

We express our sincere thanks to Mr. Sairam Iyer for his invaluable guidance, mentorship, and insightful inputs throughout the course of this project. His expertise and encouragement have been instrumental in shaping our research. We extend our appreciation to Mr. Pradip Patil for his valuable contributions and his assistance in navigating various aspects of our project. Lastly, we are grateful to the data provided by MODSAC, MODIS, NASA, DSSAT for software (freely available) and ESA COPERNICUS, as due to them and their data collection, the project went smoothly without any hassle.

Funding: Fundings by Semantic Technologies and Agritech Services, Pvt. Ltd. Pune.

Data availability: Data and software used were freely available on Intermate.

Conflict of interest: Authors declare no conflict issues.

Authors contribution: All authors contributed equally in data collection, analysis, writing and editing the manuscript.

REFERENCES

Anonymous (2023) Yield estimation system based on technology (yes-tech) under PMFBY Manual for Implementation published by Mahalanobis National Crop Forecast Centre Department of Agriculture & Farmers Welfare Ministry of Agriculture & Farmers Welfare Government of India New Delhi -110012. Pp 1-49.

Ayan Das, Mukesh Kumar, Amit Kushwaha, Rucha Dave, Kailash Kamaji Dakhore, Karshan Chaudhari, & Bimal Kumar Bhattacharya. (2023). Machine learning model ensemble for predicting sugarcane yield through synergy of optical and SAR remote sensing. (*RSASE*)., 3(30): 99-105.

Bhosale, A. D., Waskar, D. P., & Shinde, P. B. (2015). Performance of DSSAT model for simulating soybean yield under rainfed condition in Vertisols of central Maharashtra. *IJAE&B.*, 8(3):604-610.

Chavan, K.K., Khobragade, A.M., Kadam, Y.E. and Mane, R.B. (2018) Study the heat unit requirement of soybean (Glycine max) varieties under varied weather conditions at Parbhani. (JPP)., 7(3): 526-530.

Dadhwal, V. K. and Yamini Bhat. (2023). Revisiting statistical spectral-agrometeorological wheat yield models for Punjab using MODIS EVI and NCMRWF re-analysis temperature data. *J. Agrometeorol.*, *25*(1), 10–17. https://doi.org/10.54386/jam.v25i1.2067

Deshmukh, S. D., Waskar, D. P., & Shinde, P. B. (2013). Application of DSSAT model for soybean yield prediction in Vertisols of western Maharashtra. *Int. J. Curr. Microbiol. Appl. Sci*.., 2(8): 555-562.

Goodwin, B. K., & Hungerford, G. (2009). The use of satellite imagery to estimate crop area and production. Precision Agriculture, 10(4): 348-364.

Jadhav, S. D., Waskar, D. P., & Shinde, P. B. (2018). Evaluation of DSSAT model for soybean yield prediction under different sowing dates and irrigation levels in Vidarbha region of Maharashtra. *J. AgriSearch.*, 6(4):37-42.

Liujun Xiao, Guocheng Wang, Hangxin Zhou, Xiao Jin, & Zhongkui Luo. (2022). Coupling agricultural system models with machine learning to facilitate regional predictions of management practices and crop production. *Environ. Res. Lett*., (17):. 124-158

Mahlein, A.-K., Kuske, J., Steiner, U., & Wahlen, S. (2018). Hyperspectral sensors for monitoring and phenotyping crops in the field. European Journal of Remote Sensing, 51(7), 2971-2990. [This citation supports the role of remote sensing in crop health assessment]

Md Didarul Islam, Liping Di, Faisal Mueen Qamer, Sravan Shrestha, Liying Guo, Li Lin, Timothy J. Mayer, & Aparna R. Phalke. (2023). Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning. Remote Sensing, 15(9):2374-2382.

Monteith, J. L. (1972). Solar radiation and productivity in tropical ecosystems. *J. Appl. Ecol.*, 19(3), 657-666.

Mueller, N. D., Gerber, J. F., Johnston, M., Jessup, K. E., & Lobell, D. B. (2013. Closing yield gaps through nutrient and water management in grain crops across the world. Nature, 490(7419), 254-257.

Nimje, P. M. (2022) Soybean Production technology, National Skill Development Corporation AISECT, Agriculture Skill Council of India 2022.

Patel, N. R., Pokhariyal, S., & Singh, R. P. (2023). Advancements in remote sensing based crop yield modelling in India. *J. Agrometeorol.*, 25(3): 343–351. https://doi.org/10.54386/jam.v25i3.2316

Singh, A., Srivastava, S., Verma, L., & Darpe, S. (2018). Crop yield prediction using machine learning models: A review. Journal of Pharmacognosy and Phytochemistry, 7(6): 3211-3218.

Singh, K. K., Kripan Ghosh, S. C. Bhan, Priyanka Singh, Lata Vishnoi, R. Balasubramanian, S. D. Attri, Sheshakumar Goroshi, and R. Singh. (2023). Decision support system for digitally climate informed services to farmers in India. *J. Agrometeorol.*, *25*(2), 205–214. https://doi.org/10.54386/jam.v25i2.2094

Singh, P. (2023). Crop models for assessing impact and adaptation options under climate change. *J. Agrometeorol.*, 25(1), 18–33. https://doi.org/10.54386/jam.v25i1.1969

Upasana Singh, Gargi Gaydhane and Ashutosh Pawar (2023) A Semi - Physical Approach using Remote Sensing based Net Primary Productivity (NPP), Spatial, Spectral & Temporal Paddy Yield Model Development for the State of Assam. *Int. J. of Sc. and Res. (IJSR)*. 12 (8):1175-1785.

VNMKV Diary 2023 Krishi Dainandini 2023 published by Vasantrao Naik Marathwada Krishi Vidyapeeth, Parbhani during 2023.

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., & Moore III, B. (2006). Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *R.S. f En.*., 100(1), 95-113.

Yao, Y., Li, Z., Tian, F., & Tao, F. (2021). Remote Sensing-Based Estimation of Maize Yield Using a Semi-Physical Approach: A Case Study in the North China Plain. *Frontiers in Plt Sci.*, 12: 662-669.