Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 27 (1) : 63-66 (March - 2025)
https://doi.org/10.54386/jam.v27i1.2807
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Wheat yield prediction of Rajasthan using climatic and satellite data and machine learning techniques

KAVITA JHAJHARIA*

Department of Information Technology, Manipal University Jaipur, India
**Corresponding author email: Kavita.jhajharia@jaipur.manipal.edu*

ABSTRACT

In the present study, three machine learning algorithms viz. support vector regression (SVR), random forest (RF) and XGBoost, one linear regression method, least absolute shrinkage and selection operator regression (LASSO), and one deep learning method, long short-term memory (LSTM), were used to predict the wheat yield during 2008 to 2019 using satellite data derived vegetation indices and climatic parameters. The R^2 obtained with remote sensing data were between 0.56 and 0.67 across all the models and vegetation indices which improved to 0.77 and 0.86 by incorporating climatic data. The results indicated SVR outperformed LSTM in wheat yield prediction.

Keyword: Remote sensing, Crop yield prediction, Machine learning, Deep learning.

Statistical models, biophysical models and field surveys are also being utilised to estimate agricultural production at the field and regional levels (Cao *et al.,* 2021). However, these models and the field surveys face multiple challenges, such as limited scalability, and *an inability* to model complex non-linear interactions effectively (Feng *et al.,* 2020). In recent years, machine learning algorithms have been utilized to address a variety of agricultural challenges, including crop type categorization and crop production estimation (Patel *et al.,* 2023; Sakthipriya and Chandrakumar, 2024; Khan *et al.,* 2023). Remotely sensed data produced by satellite is most suitable choice for crop yield prediction because of its repetitiveness, and multi spectral information (Patel *et al.,* 2023). Many studies have employed satellite-derived vegetation indices (VIs) for crop yield estimation, such as near-infrared reflectance of vegetation (NIRv), normalised difference vegetation index (NDVI), and the enhanced vegetation index (EVI) (Ahmad *et al.,* 2020). Schwalbert *et al.,* (2020) employed NDVI, EVI and land surface temperature (LST) as data with machine learning and deep learning algorithms for soybean yield prediction. Zhang *et al.,* (2024) reported that the deep learning models combined with vegetation indices NDVI and EVI and climate variables proved to be more accurate in predicting wheat yield in comparison to univariate indices. This approach highlighted the need of more efficient data sources to build better models for agricultural forecasting system. Traditional vegetation indicators have in recent times shown to be less responsive to photosynthetic capacity than solar-induced fluorescence (SIF) measured by satellite (Guanter *et al.,* 2014). While satellite SIF data have been used in a few studies to predict crop yields, recent advances, such as the integration of high-resolution SIF with vegetation indices with machine learning and deep learning algorithms, demonstrate promise in improving prediction accuracy. The present study addresses this gap, incorporating high-resolution SIF and advanced ML/DL models to predict wheat yields in Rajasthan's wheat-growing regions.

MATERIAL AND METHODS

The present study concentrated on predicting wheat yield in the Rajasthan region extending from 27° 23' 28.5972'' N and 73° 25' 57.4212'' E, which contribute over 7.49% of the country's wheat production. Wheat is usually sown in the winter (November–December) and harvested in the spring (March–April) (Pathak *et al.,* 2003). The climatic data, satellite data, planting area, and wheat yield data *were* collected from different sources. All the input variables were integrated into monthly temporal and 0.05° spatial resolution. The satellite and climatic data variables were extracted at the district level on monthly basis. For data collecting and pre-processing, authors used the Google Earth Engine platform. The wheat crop yield (t ha^{-1}) data at the district level from 2009 to 2019

Article info - DOI: *https://doi.org/10.54386/jam.v27i1.2807*
Received: 13 November 2024; Accepted: 30 January 2025; Published online : 1 March 2025

Table 1: Correlation coefficients among climate and satellite variables in month of March and April

Variable	Correlation Coefficient						
	SIF (GOME2)	SIF(SCIF)	NDVI	EVI	NIRv	Prec.	Tmax
March							
SIF(SCIF)	0.84						
NDVI	0.72	0.9					
EVI	0.76	0.93	0.96				
NIRv	0.76	0.93	0.94	0.9			
Prec.	0.07	0.14	0.21	0.16	0.16		
Tmax	-0.48	-0.44	-0.49	-0.44	-0.42	-0.58	
Yield	0.74	0.69	0.65	0.69	0.67	-0.01	-0.49
April							
SIF(SCIF)	0.62						
NDVI	0.38	0.73					
EVI	0.38	0.72	0.97				
NIRv	0.34	0.69	0.95	0.98			
Prec.	0.44	0.61	0.36	0.34	0.32		
Tmax	-0.49	-0.67	-0.58	-0.54	-0.5	0.71	
Yield	0.12	0.27	0.55	0.55	0.55	-0.05	-0.28

were acquired from the Area and Production Statistics (https://aps.dac.gov.in). TerraClimate datasets with a high spatial resolution was used to acquire historical climate data on monthly maximum temperature (Tmax) and Precipitation (Prec.).

Satellite data

For this investigation, contiguous solar-induced fluorescence (CSIF) data were used with a 0.05-degree spatial resolution and a four-day temporal window. The Orbiting Carbon Observatory-2 (OCO-2) and the Global Ozone Monitoring Experiment-2 (GOME-2) were the two major satellite sources for SIF data, with OCO-2 providing high-resolution measurements and GOME-2 offering global coverage at a coarser resolution. The CSIF's monthly statistics were aggregated (Zhang *et al.,* 2018). NDVI (Normalized Difference Vegetation Index), NIRv (Near-Infrared Reflectance of Vegetation), and EVI (Enhanced Vegetation Index) were utilized to estimate wheat yield. These indices were derived from MODIS reflectance products. Specifically, we used the MOD13A3 product, which provides a 1-kilometer spatial resolution and monthly temporal resolution, to calculate NDVI and EVI. NIRv was calculated using NDVI and MOD13A3 data, maintaining the same spatial resolution of 1 km (Badgley *et al.,* 2017).

Methodology

Author utilized machine learning and deep learning algorithms to estimate wheat yield and compared the results to identify the best performer. Support vector regression (SVR), Random Forest, extreme gradient boosting (XGBoost), Long short-term memory (LSTM)), and least absolute shrinkage and selection operator regression (LASSO) were implemented. SVR was selected to model the non-linear relationship through kernel functions, which makes it suitable for satellite and climate data. Random Forest is an ensemble learning method, which averages the output of multiple decision trees, ensures the reliability with varying data. XGBoost was selected to enhance the further predictions through gradient boosting techniques. LASSO regression was selected for its feature selection ability which will manage the high dimensional data and reduces less relevant predictors. LSTM was used to capture the temporal dependencies in time series data (satellite and climate).

To evaluate the model's effectiveness, the testing dataset was utilized to determine the root mean square error (RMSE) and coefficient of determination (R^2). The mean predicted R^2 and RMSE for the 5-fold cross-validation procedure was calculated across 70 training– 30 testing splits. The model was trained using multiple climates and remote sensing variables from 2009 to 2018 to evaluate its practical performance at the district level.

RESULTS AND DISCUSSION

Correlation with remote sensing and climate data

Table 1 presents the correlation coefficients between remote sensing data, climate factors, and wheat production, exemplified for the months of March and April. The Table 1 demonstrates that remote sensing variables and wheat yield data are more closely correlated than climate data correlation with wheat yield. SIFGOME2, SIFCSIF, NDVI, EVI, and NIRv are found to have coefficients of 0.72, 0.69, 0.63, 0.67, and 0.67, with wheat crop in March, but among all the SIFGOME2 and SIFCSIF have higher correlation coefficients with wheat yield.

Among all meteorological variables, Tmax in March has the strongest association coefficient with wheat crop yield. Prec., on the other hand, ranks last among all climate factors in March and April between Prec. and wheat yield in April. Other research may only seldom uncover the unfavourable relationship between solar radiation and yield. The finding is influenced by the correlation of climate elements, according to our explanation. In this study, the crop yield and daily Tmax appears to have a negative correlation.

Table 2: The performance of models using remote sensing data

Algorithm	NDVI		EVI		NIRv		SIF (GOME2)		SIF (CSIF)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LASSO	0.57	0.68	0.57	0.67	0.56	0.69	0.60	0.69	0.62	0.63
SVR	0.65	0.61	0.63	0.62	0.63	0.63	0.61	0.63	0.67	0.60
RF	0.63	0.63	0.62	0.63	0.63	0.63	0.61	0.63	0.62	0.63
XGBoost	0.62	0.64	0.61	0.65	0.62	0.62	0.59	0.62	0.62	0.63
LSTM	0.62	0.64	0.61	0.66	0.60	0.67	0.60	0.67	0.64	0.62

Table 3: The performance of models using climate data with remote sensing data

Algorithm	NDVI + climate data		EVI + climate data		NIRv + climate data		SIF (GOME2) + climate data		SIF (CSIF) + climate data	
	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
LASSO	0.79	0.47	0.79	0.46	0.79	0.47	0.77	0.48	0.81	0.45
SVR	0.86	0.37	0.86	0.36	0.86	0.37	0.84	0.40	0.86	0.36
RF	0.83	0.42	0.82	0.43	0.82	0.43	0.81	0.44	0.83	0.42
XGBoost	0.82	0.42	0.83	0.41	0.83	0.41	0.81	0.44	0.83	0.42
LSTM	0.85	0.42	0.85	0.40	0.85	0.40	0.82	0.43	0.85	0.42

Comparison of algorithms performance

Table 2 demonstrates the results with only remote sensing data as predictors and Table 3 demonstrates results with both remote sensing and climatic variables. R^2 for SIFCSIF is between 0.62 and 0.67, and RMSE is between 0.60 and 0.63. In the case of SIFGOME2, R^2 is between 0.59 and 0.61, and RMSE is between 0.64 and 0.67 (Table 2). In addition, NDVI, EVI, and NIRv all had comparable results. For remote sensing data with climate data, the R^2 value ranges from 0.77 to 0.86 and the RMSE value ranges from 0.36 to 0.48, SIFGOME2 scored poorly in comparison to all other remote sensing variables. The performance of SIFCSIF was better with R^2 value ranging from 0.81 to 0.86 and RMSE ranging from 0.36 to 0.45.

SVR performed better than all the implemented methods with R^2 value varying from 0.84 to 0.86 and RMSE value fluctuating from 0.36 to 0.40 t ha^{-1}. Lasso performed poorly using remote sensing data among all the methods with the R^2 value varying from 0.77 to 0.81 and the RMSE value varying from 0.45 to 0.48 t ha^{-1}. The accuracy of LASSO in predicting wheat yield was similar. Although RF outperformed LASSO, but it fell short of the other models. The deep learning model, LSTM, performed better than XGBoost and RF but not better than SVR. In LSTM, the R^2 varying from 0.82 to 0.85 and RMSE varying from 0.40 to 0.43 t ha^{-1}.

The performance of all the models is different. Table 2 and Table 3 shows that RF, SVR, XGBoost and LSTM outperformed LASSO, which is consistent with previous research (Wolanin *et al*., 2020). The reason for this could be that machine learning approaches capture nonlinear correlations between independent factors and wheat production better than classic linear methods. Furthermore, due to SVR's strong generalization capabilities, it outerperformed all the other implemented methods. LSTM compared to SVR, could not perform better due to fewer features in the study. However, by combining more variables and data with fewer time steps, a deep learning method may produce greater outcomes, which should be researched more in the future (Duveiller and Cescatti, 2016). The better performance of SVR in this study could be attributed to its ability to handle non-linear relationships and its kernel-based approach, which is suitable for complex datasets. In contrast, LSTM's lower performance may be due to its sensitivity to the amount of training data, suggesting potential improvement with more extensive datasets. Random Forest's ensemble methodology offers robustness but may lack the accuracy of SVR due to its averaging mechanism.

CONCLUSION

To predict wheat yields using satellite factors and climatic data, one deep learning method, two linear regression methods, and three machine learning methods were employed in this research. Additionally, the ML models and Deep learning models performed better than regression methods, with SVR and XGBoost outperforming RF. Across all years, the model using high-resolution SIFCSIF outperformed the model using coarse-resolution SIFGOME2. The study indicates that employing high-resolution SIF products might not always ensure a good crop yield prediction in comparison to other remote sensing variables. The study will be beneficial for policy formulation for sustainable agriculture.

ACKNOWLEDGEMENT

The author expresses gratitude to Manipal University Jaipur for providing the necessary research infrastructure and resources.

Declaration of interests: The author declares no conflict.

Availability of Data and Materials: The data is available on request.

Funding: None

Authors Contribution: **Kavita**: Conceptualization, Methodology, Data collection, Investigation, Writing.

REFERENCES

Ahmad, I., Singh, A., Fahad, M. and Waqas, M. (2020). Remote sensing-based framework to predict and assess the interannual variability of maize yields in Pakistan using Landsat imagery. *Comput. Electron. Agric.,* 178: 105732. https://doi.org/10.1016/j.compag.2020.105732

Badgley, G., Field, C. B. and Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. *Sci. Adv.,* 3(3): e1602244. https://doi.org/10.1126/sciadv.1602244

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J. and Xie, J. (2021). Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. *Agric. For. Meteorol.,* 297: 108275. https://doi.org/10.1016/j.agrformet.2020.108275

Duveiller, G. and Cescatti, A. (2016). Spatially downscaling sun-induced chlorophyll fluorescence leads to an improved temporal correlation with gross primary productivity. R*emote Sens. Environ.,* 182: 72-89. https://doi.org/10.1016/j.rse.2016.04.027

Feng, P., Wang, B., Liu, D. L., Waters, C., Xiao, D., Shi, L. and Yu, Q. (2020). Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agric. For. Meteorol.,* 285-286: 107922. https://doi.org/10.1016/j.agrformet.2020.107922

Guanter, L., Zhang, Y., Jung, M., Joiner, J., Voigt, M., Berry, J. A., …..., and Griffis, T. J. (2014). Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. *Proc. Natl. Acad. Sci. U.S.A.,* 111(14): E1327-33. https://doi.org/10.1073/pnas.1320008111

Khan, Y., Kumar, V., Setiya, P. and Satpathi, A. (2023). Comparison of phenological weather indices based statistical, machine learning and hybrid models for soybean yield forecasting in Uttarakhand. *J. Agrometeorol.,* 25(3): 425-431 https://doi.org/10.54386/jam.v25i3.2232

Patel, N. R., Pokhariyal, S. and 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

Pathak, H., Ladha, J. K., Aggarwal, P. K., Peng, S., Das, S., Singh, Y., ……. and Gupta, R. K. (2003). Trends of climatic potential and on-farm yields of rice and wheat in the Indo-Gangetic Plains. *Field Crops Res.,* 80(3): 223-234. https://doi.org/10.1016/S0378-4290(02)00194-6

Sakthipriya, D. and Chandrakumar, T. (2024). Weather based paddy yield prediction using machine learning regression algorithms. *J. Agrometeorol.,* 26(3): 344-348. https://doi.org/10.54386/jam.v26i3.2598

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V. and Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.,* 284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Wolanin, A., Mateo-García, G., Camps-Valls, G., Gómez-Chova, L., Meroni, M., Duveiller, G., Liangzhi, Y. and Guanter, L. (2020). Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environ. Res. Lett.,* 15(2), 024019. https://doi.org/10.1088/1748-9326/ab68ac

Zhang, Y., Joiner, J., Hamed Alemohammad, S., Zhou, S. and Gentine, P. (2018). A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences,* 15(19), 5779-5800. https://doi.org/10.5194/bg-15-5779-2018

Zhang, G., Roslan, S. N. A. B., Shafri, H. Z. M., Zhao, Y., Wang, C. and Quan, L. (2024). Predicting wheat yield from 2001 to 2020 in Hebei Province at county and pixel levels based on synthesized time series images of Landsat and MODIS. *Sci. Rep.,* 14(1), 16212. https://doi.org/10.1038/s41598-024-67109-3

Zhou, Y., Zhai, L., Zhou, W., Zhou, J. and Cen, H. (2023). Investigation on data fusion of sun-induced chlorophyll fluorescence and reflectance for photosynthetic capacity of rice (arXiv:2312.00437). arXiv. https://doi.org/10.48550/arXiv.2312.00437