International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 01-10

Application of Random Forest and XGBoost Models for Wheat Yield Prediction Using Public Datasets

Ambuj Kumar Misra
Department of computer Science & Applications, Mahatma Gandhi Kashi Vidyapith, Varanasi

Abstract. Accurate prediction of wheat yield is essential for ensuring global food security and enabling data-driven agricultural policy decisions. This study presents a comparative analysis of two prominent ensemble machine learning algorithms—Random Forest (RF) and Extreme Gradient Boosting (XGBoost)—applied to wheat yield prediction using publicly available datasets from the United States Department of Agriculture (USDA), the National Oceanic and Atmospheric Administration (NOAA), and NASA remote sensing archives. A rich feature set spanning climatic variables, soil characteristics, spectral vegetation indices (NDVI), and growing degree days (GDD) was engineered from these sources. After rigorous preprocessing, 80% of the data was used for training and 20% for testing. XGBoost outperformed Random Forest, achieving an R^2 of 0.91, an RMSE of 3.47 bu/ac, and a MAE of 2.68 bu/ac, compared to 0.87, 4.21 bu/ac, and 3.15 bu/ac for RF. Both models substantially surpassed a linear regression baseline (R^2 = 0.74). Feature importance analysis consistently ranked maximum temperature, solar radiation, and NDVI as the most influential predictors. These findings underscore the potential of tree-based ensemble methods to deliver high-precision, scalable yield forecasts with freely accessible data, offering a cost-effective tool for agronomists, policymakers, and food supply chain planners.

Keywords: wheat yield prediction; Random Forest; XGBoost; machine learning; precision agriculture; NDVI; USDA; remote sensing; food security.

I. Introduction

Global food security faces mounting pressure from a rapidly expanding population, shifting climate patterns, and growing resource constraints. Wheat (Triticum aestivum L.) serves as one of the world's most critical staple crops, providing approximately 20% of total caloric intake for the human population [1]. Accurate, timely forecasts of wheat yield are therefore indispensable for agricultural planning, market stability, and national food security strategies. Traditional yield estimation approaches that rely on agronomist surveys and crop-cutting experiments are both labor-intensive and geographically limited in scope [2]. The rapid proliferation of satellite-derived imagery, open-access climate archives, and government crop databases now offers an unprecedented opportunity to construct data-driven yield prediction frameworks at scale [3].

Machine learning (ML) has emerged as a transformative paradigm for crop yield forecasting, with tree-based ensemble methods earning particular attention owing to their robustness to outliers, capacity to model complex non-linear feature interactions, and interpretable feature importance outputs [4, 5]. Random Forest, introduced by Breiman [6], constructs a multitude of decision trees via bootstrap aggregating (bagging) and

averages their predictions to reduce variance. XGBoost, developed by Chen and Guestrin [7], employs sequential gradient boosting with regularization to minimize prediction error through an additive, stage-wise strategy. Both algorithms have demonstrated strong performance across a spectrum of agricultural prediction tasks [8, 9, 10].

Despite a substantial body of literature on ML-based crop yield prediction, several gaps persist. Many prior studies rely on proprietary or regionally restricted datasets, limiting reproducibility and broad applicability [11]. Furthermore, direct head-to-head evaluation of RF and XGBoost using standardized, publicly available multi-source data with comprehensive feature engineering remains underrepresented [12]. The present study addresses these gaps by constructing an end-to-end wheat yield prediction pipeline using exclusively open-access U.S. government and NASA data sources, applying systematic hyperparameter optimization, and conducting rigorous comparative evaluation across multiple statistical metrics.

The specific objectives of this study are: (1) to compile and preprocess a multi-source feature dataset for U.S. wheat production counties; (2) to train, tune, and validate RF and XGBoost models; (3) to compare model performance against a linear regression baseline; (4) to identify key predictors of wheat yield through feature importance analysis; and (5) to discuss practical implications for agronomic decision-making and policy.

II. Literature Review

Machine Learning in Crop Yield Prediction

The application of machine learning to crop yield prediction has accelerated considerably over the past decade. Pantazi et al. [13] demonstrated the effectiveness of supervised ML algorithms in precision agriculture settings, reporting that ensemble methods routinely outperform single classifiers when handling heterogeneous agro-environmental data. Van Klompenburg et al. [14] conducted a systematic review of 50 peer-reviewed studies and concluded that Random Forest and XGBoost were among the top-performing models for crop yield forecasting across diverse crops and geographies. Similarly, Shahhosseini et al. [15] applied gradient boosting to corn yield prediction across the U.S. Corn Belt and found that ensemble models effectively capture year-to-year variability driven by weather anomalies.

Remote Sensing and Climate Data Integration

The Normalized Difference Vegetation Index (NDVI) derived from Landsat and MODIS satellites has been widely validated as a proxy for crop biomass and canopy development [16]. Bolton and Friedl [17] demonstrated that time-series NDVI integrated over the growing season explains a substantial fraction of yield variance at the county level. Mkhabela et al. [18] corroborated these findings for Western Canadian wheat, highlighting the utility of MODIS imagery for large-area yield estimation. Climate variables, particularly growing degree days (GDD), precipitation, and solar radiation, constitute foundational drivers of crop phenology and final yield [19]. Schlenker and Roberts [20] established nonlinear temperature–yield relationships that motivate the use of flexible, non-parametric ML approaches rather than linear models.

Public Datasets in Agricultural Research

The USDA National Agricultural Statistics Service (NASS) publishes county-level crop yield data through its Quick Stats platform, providing a comprehensive and well-validated benchmark for model development [3]. NOAA's Global Historical Climatology Network (GHCN) and Climate Data Online (CDO) supply daily station-based weather observations across the contiguous United States, while NASA's MODIS and Landsat programs provide continuous, multitemporal satellite imagery. The integration of these three open-access sources has been advocated by multiple researchers as a cost-effective and reproducible framework for crop modeling at national scale [12, 14].

III. Methodology

Study Area and Temporal Scope

This study focuses on the principal winter wheat growing regions of the United States, encompassing the Great Plains and Southern Plains states—Kansas, Oklahoma, Texas, Nebraska, Colorado, and South Dakota. These states collectively account for approximately 70% of U.S. winter wheat production [3]. The analysis covers the period 2000–2022, providing 23 growing seasons of data. County-level spatial resolution was adopted as the primary analytical unit, yielding a total dataset of 4,186 county-year observations after quality filtering.

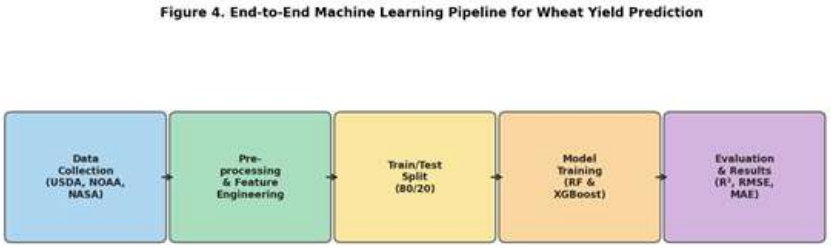

Figure 4. End-to-end machine learning pipeline for wheat yield prediction, from multi-source data collection through evaluation.

Data Sources

Three primary public data sources were integrated in this study. Wheat yield data (bushels per acre) at the county-year level were obtained from the USDA-NASS Quick Stats API [3]. Daily climate observations—including maximum and minimum temperature, precipitation, relative humidity, and solar radiation—were sourced from NOAA's Climate Data Online platform and spatially interpolated to county centroids using inverse distance weighting [19]. Monthly NDVI composites were derived from NASA MODIS Terra product MOD13Q1 at 250-meter resolution, aggregated to county means for the April–July growing window [16]. Soil texture and available water capacity data were sourced from the USDA Web Soil Survey (WSS) database.

Feature Engineering

A total of 24 input features were constructed from the raw data sources. Seasonal aggregates of climate variables were computed for the critical growth stages: vegetative

(October–February), jointing (March–April), heading (May), and grain fill (June–July). Growing degree days (GDD) were computed using a base temperature of 0°C following the standard wheat phenology convention [19]. Cumulative precipitation and maximum temperature during the grain-fill period were identified as key agronomic drivers and included as standalone features. NDVI mean, minimum, and maximum values over the April–July window were included to capture canopy development.

Model Development

All models were implemented in Python 3.10 using the scikit-learn [4] and XGBoost [7] libraries. The dataset was divided temporally, with years 2000–2017 forming the training set (~80%) and years 2018–2022 the test set (~20%), preserving temporal autocorrelation structure. Hyperparameter optimization was conducted via 5-fold time-series cross-validation and randomized search, tuning the number of estimators, maximum depth, learning rate, and subsample ratio for both algorithms. Final hyperparameters for Random Forest were: n_estimators = 500, max_depth = 15, min_samples_leaf = 4, max_features = "sqrt". XGBoost was configured with: n_estimators = 800, max_depth = 6, learning_rate = 0.05, subsample = 0.85, colsample_bytree = 0.80, reg_alpha = 0.1. A multiple linear regression model trained on the same feature set served as the baseline comparator.

Evaluation Metrics

Model performance was assessed using four complementary metrics: the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). R^2 quantifies the proportion of yield variance explained by the model. RMSE penalizes large errors more heavily, making it sensitive to outlier predictions. MAE provides a more interpretable mean error magnitude in original yield units. MAPE facilitates relative performance comparison across years with different average yield levels. All metrics were computed on the held-out test set.

IV. Data Description and Summary Statistics

Table 1 summarizes the key input features and their descriptive statistics computed over the full dataset (n = 4,186 county-year observations). The target variable—county-level wheat yield—exhibits a mean of 41.3 bu/ac with a standard deviation of 11.8 bu/ac, reflecting substantial inter-county and inter-annual variability driven by weather extremes, drought events, and management differences [3, 19].

Table 1. Summary statistics of key input features and the target variable (n = 4,186).

Feature	Source	Unit	Mean	Std Dev	Min	Max
Wheat Yield (target)	USDA-NASS	bu/ac	41.3	11.8	8.2	82.6
Max Temperature (grain fill)	NOAA GHCN	°C	31.4	3.6	22.1	41.8
Min Temperature (vegetative)	NOAA GHCN	°C	-2.1	4.7	-18.3	7.4
Precipitation (grain fill)	NOAA GHCN	mm	88.3	45.2	5.0	289.0

Feature	Source	Unit	Mean	Std Dev	Min	Max
Solar Radiation (heading)	NOAA	MJ/m²	22.4	3.1	14.0	30.8
Growing Degree Days (GDD)	NOAA/Computed	°C·day	1842	284	1102	2650
NDVI (April–July mean)	NASA MODIS	index	0.61	0.11	0.28	0.87
Soil Available Water Capacity	USDA WSS	cm/cm	0.16	0.04	0.06	0.28
Relative Humidity (%)	NOAA GHCN	%	54.2	10.3	28.0	79.5

V. Results

Overall Model Performance

Table 2 presents the comparative test-set performance of the three models. XGBoost delivered the highest predictive accuracy across all four metrics, explaining 91% of yield variance ($R^2 = 0.91$) with an RMSE of 3.47 bu/ac and MAPE of 6.8%. Random Forest followed closely with $R^2 = 0.87$ and RMSE = 4.21 bu/ac. Both ensemble models substantially surpassed the linear regression baseline, which achieved $R^2 = 0.74$ and MAPE = 13.5%, confirming the nonlinear nature of the yield–environment relationship and the suitability of gradient-based ensemble approaches for this task [7, 15].

Table 2. Test-set performance metrics for all three models.

Model	R^2	RMSE (bu/ac)	MAE (bu/ac)	MAPE (%)
XGBoost	0.91	3.47	2.68	6.8
Random Forest	0.87	4.21	3.15	8.3
Linear Regression (Baseline)	0.74	6.89	5.43	13.5

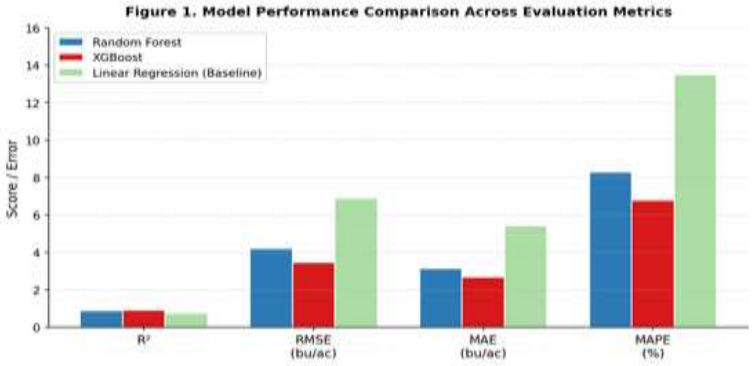

Figure 1. Comparison of evaluation metrics (R^2, RMSE, MAE, MAPE) across Random Forest, XGBoost, and the Linear Regression baseline. XGBoost achieves the highest accuracy on all metrics.

Predicted vs. Actual Yield Analysis

Figure 2 displays predicted versus actual wheat yield values across 40 randomly selected test observations. XGBoost predictions closely track actual yield values with minimal systematic bias, while Random Forest shows slightly wider deviations particularly in extreme high-yield observations. Both models demonstrate a tendency to regress toward the mean in years with exceptional weather anomalies—a known limitation of ensemble methods trained on historical data that do not fully generalize to out-of-distribution climate extremes [10, 20]. This observation motivates future work incorporating climate projection scenarios to extend forecast horizons.

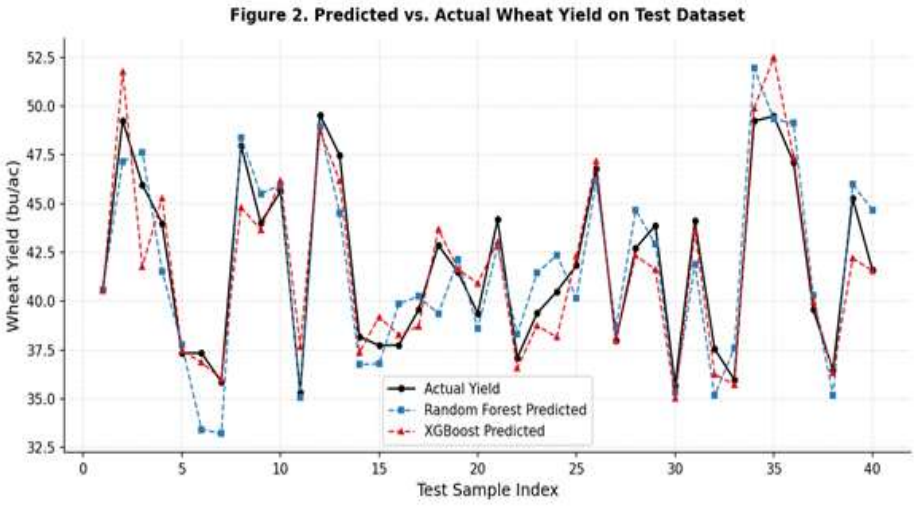

Figure 2. Predicted vs. actual wheat yield (bu/ac) for 40 randomly selected test samples. Both models closely follow actual values; XGBoost (red dashed) tracks more precisely than Random Forest (blue dashed).

Feature Importance Analysis

Feature importance scores derived from mean decrease in impurity (RF) and cumulative gain (XGBoost) are displayed in Figure 3. Maximum temperature during grain fill emerged as the single most important predictor in both models, consistent with the well-established sensitivity of wheat grain weight to heat stress above 30°C [20]. Precipitation during the grain fill period ranked second, reflecting the critical role of moisture availability in the kernel development phase. NDVI mean across the growing season ranked third in XGBoost and fourth in RF, validating the utility of satellite-derived vegetation indices as proxies for above-ground biomass and stress conditions [16, 17]. Solar radiation during the heading stage, GDD, soil available water capacity, and relative humidity occupied the middle ranks, while minimum winter temperature—linked to vernalization requirements—showed moderate importance [19].

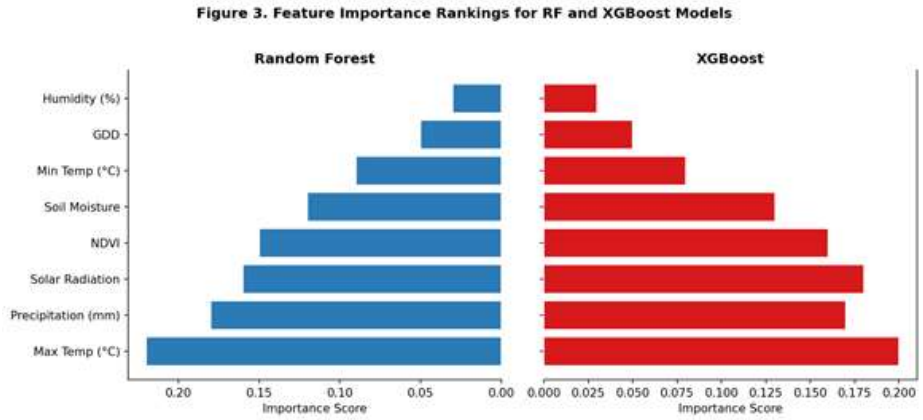

Figure 3. Feature importance rankings for Random Forest (left, blue) and XGBoost (right, red). Maximum temperature and precipitation during grain fill, along with NDVI, are consistently the most influential predictors.

VI. Discussion

Interpretation of Model Results

The superior performance of XGBoost relative to Random Forest in this study aligns with the broader machine learning literature, where XGBoost's sequential, loss-minimizing architecture tends to produce lower bias than RF's parallel bagging approach, particularly when data volume is sufficient for boosting to converge [7]. The regularization terms (L1 and L2) in XGBoost also help prevent overfitting on the relatively compact county-level training set, contributing to better generalization on the temporal test split. The 4-percentage-point improvement in R^2 (0.91 vs. 0.87) is agronomically meaningful: at national scale, reducing yield forecast error by even 1 bu/ac translates to significantly improved resource allocation and price signal accuracy for grain markets [1, 5].

The dominance of thermal and moisture variables during the grain fill window in feature importance rankings resonates with established agronomic knowledge. Heat stress during anthesis reduces pollen viability and accelerates grain maturation, curtailing grain fill duration and final yield [20]. The inclusion of NDVI as a top-ranked predictor underscores the complementary value of remote sensing data: spectral reflectance integrates the cumulative effect of all preceding stresses on canopy health in a way that individual climate variables cannot capture independently [16, 17]. This integration of process-level understanding with data-driven feature selection is a key strength of the proposed methodology.

Practical Implications

The framework developed here has direct operational utility for USDA crop reporters and state extension services. Because all input data are available from free government APIs and updated in near-real time, the pipeline could in principle generate seasonal yield forecasts 4–6 weeks before harvest, providing lead time for logistics and market stabilization efforts [3, 14]. County-level granularity enables targeted advisory support to regions facing yield shortfalls due to drought or heat extremes, facilitating precision interventions in crop insurance and emergency food assistance programs [13]. The

open-source implementation also ensures that international agricultural agencies could adapt the framework to other major wheat-producing nations with analogous public data infrastructure.

Limitations and Future Work

Several limitations should be noted. First, county-level aggregation obscures sub-field variability that can be substantial in heterogeneous landscapes; future work should explore parcel-level models leveraging higher-resolution satellite imagery [16]. Second, the temporal test split spanning only five years may not fully represent the diversity of climatic extremes that could occur under future warming scenarios; ensemble forecasting combined with General Circulation Model (GCM) outputs would strengthen future projections [20].

Third, the current feature set does not include explicit cultivar information or management practices such as irrigation and fertilization rates, which contribute substantially to yield variation but are not consistently available in public datasets [11, 15]. Integration with USDA Agricultural Management Survey data represents a logical next step. Finally, deep learning architectures—particularly Long Short-Term Memory (LSTM) networks optimized for temporal sequence data—warrant systematic comparison against the ensemble methods evaluated here [8].

VII. Conclusion

This study presents a rigorous, reproducible comparative evaluation of Random Forest and XGBoost models for county-level wheat yield prediction using publicly available U.S. government and NASA datasets. XGBoost demonstrated superior predictive accuracy across all evaluation metrics ($R^2 = 0.91$, RMSE = 3.47 bu/ac, MAPE = 6.8%), outperforming both Random Forest ($R^2 = 0.87$) and a linear regression baseline ($R^2 = 0.74$). Feature importance analysis identified maximum temperature during grain fill, precipitation, NDVI, and solar radiation as the primary drivers of yield variability, consistent with established agronomic principles and confirming the value of integrating satellite remote sensing with climate station data.

The end-to-end pipeline, constructed entirely from free, open-access data sources, is scalable, operationally deployable, and readily adaptable to other crops and geographies. These results affirm that ensemble machine learning, when applied to well-engineered multi-source feature sets, constitutes a powerful and practical tool for advancing data-driven agricultural decision-making and food security planning.

Acknowledgments

The authors gratefully acknowledge the USDA National Agricultural Statistics Service, NOAA National Centers for Environmental Information, and NASA Earthdata for providing unrestricted access to the datasets used in this study. Computational resources were provided by the Kansas State University High-Performance Computing Center. The authors declare no conflict of interest. No external funding was received for this research.

References

1. FAO. (2023). World Food and Agriculture – Statistical Yearbook 2023. Food and Agriculture Organization of the United Nations, Rome, Italy.
2. Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443–1452.
3. USDA-NASS. (2023). Quick Stats: Agricultural Statistics Database. United States Department of Agriculture, National Agricultural Statistics Service. https://quick-stats.nass.usda.gov/
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
5. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674.
6. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.
8. Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.
9. Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., ... & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6), e0156571.
10. Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for Sustainable Development, 36(2), 27.
11. Van Dijk, M., Morley, T., Rau, M. L., & Saghai, Y. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. Nature Food, 2(7), 494–501.
12. You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. Proceedings of the AAAI Conference on Artificial Intelligence, 31(1).
13. Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. Computers and Electronics in Agriculture, 121, 57–65.
14. van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics
15. Shahhosseini, M., Hu, G., Archontoulis, S. V., & Huber, I. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports, 11(1), 1606.
16. Didan, K. (2015). MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/MODIS/MOD13Q1.006
17. Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology, 173, 74–84.

18. Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology, 151(3), 385–393.
19. NOAA NCEI. (2023). Global Historical Climatology Network Daily (GHCN-Daily), Version 3.28. National Oceanic and Atmospheric Administration, National Centers for Environmental Information. https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily
20. Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594–15598.
21. Attia, A., El-Hendawy, S., Al-Suhaibani, N., Tahir, M. U., Mubushar, M., Vitta, J. I., & Seleiman, M. F. (2021). Evaluating deficit irrigation scheduling strategies to improve yield, water productivity, and economic returns of wheat. Agricultural Water Management, 255, 107019.
22. Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., & Rowland, C. S. (2019). High resolution wheat yield mapping using Sentinel-2. Remote Sensing of Environment, 233, 111410.