JARS Journal of Applied Remote Sensing

RESEARCH PAPER

Harnessing data-driven approaches to predicting corn yields across different weather zones

Weiwei Xie,[a] Yanbo Huang,[a,*] Qingmin Meng,[b] Bo Tang,[c] and Ying Ouyang[d]

[a]USDA Agricultural Research Service, Mississippi State, Mississippi, United States
[b]Mississippi State University, Department of Geosciences, Mississippi State, Mississippi, United States
[c]Worcester Polytechnic Institute, Department of Electrical and Computer Engineering, Worcester, Massachusetts, United States
[d]USDA Forest Service, Southern Research Station, Center for Bottomland Hardwoods Research, Mississippi State, Mississippi, United States

ABSTRACT. Forecasting corn yields with precision across diverse weather zones is vital to supporting food supply resilience and sustainable agriculture. We introduce a data-integrated modeling framework applied across the Greater Mississippi River Basin—one of the largest corn-producing regions in the United States. Using MODIS-based NDVI time series, historical meteorological data, and spatial crop coverage, we investigate the relationships between vegetation health indicators and corn yields under varying climatic conditions. A new vegetative metric is introduced to enhance correlation with yield outcomes. Both statistical regressions and advanced machine learning models—including support vector regression, XGBoost, and Gaussian processes—are employed. Among these, XGBoost demonstrated the highest predictive accuracy. The results underscore the benefits of combining multisource data with nonlinear models for regionally robust and scalable corn yield estimation.

 [DOI: 10.1117/1.JRS.20.024501]

Keywords: corn yield prediction; crop growth modeling; weather zones; remote sensing; machine learning

Paper 250603G received Sep. 5, 2025; revised Jan. 12, 2026; accepted Mar. 24, 2026; published Apr. 10, 2026.

1 Introduction

Agricultural productivity plays a central role in societal well-being, providing the foundation for food systems and economic growth.[1] Although global yields have increased over recent decades, food insecurity remains a challenge due to uneven access and environmental volatility.[2] As climate patterns continue to shift, there is a growing need to understand how these changes influence staple crop outputs such as corn.

Modern tools such as satellite imagery and data-driven analytics offer new capabilities for yield monitoring and forecasting.[3–12] Among these, vegetation indices such as NDVI[6–9] and EVI,[10–12] derived from remote sensing platforms, have proven valuable for tracking crop development over time. When combined with weather indicators such as precipitation and temperature, these indices can offer powerful insights into crop performance at different spatial and temporal scales.

Numerous studies have demonstrated that blending physiological signals from remote sensing with meteorological conditions improves predictive accuracy. However, variability due

*Address all correspondence to Yanbo Huang, yanbo.huang@usda.gov

1931-3195/2026/$28.00

to regional weather zones, crop types, and model configurations introduces uncertainty.[4,13–18] Therefore, a regionally adaptive and model-flexible approach is essential to advancing yield prediction across heterogeneous agricultural landscapes.

Several modeling methods have been designed to further refine the estimation of crop yields. For instance, Beeri and Peled[19] integrated real-time remote sensing observations for precision crop monitoring. Huang et al.[20] proposed an optimized data assimilation framework targeting key phenological stages to improve the prediction of winter wheat yields. Similarly, incorporating leaf area index and soil moisture into crop simulation models has shown potential for enhancing maize yield estimation.[13] The advent of machine learning has also propelled yield prediction capabilities forward, enabling the modeling of complex relationships between environmental variables and agricultural outputs.[21]

Advanced machine learning techniques such as deep Gaussian processes,[22] support vector machines,[23] and random forest regression[24] have demonstrated considerable success in crop yield prediction across different regions and crops. Furthermore, spiking neural networks have been utilized for the spatiotemporal modelling of NDVI time series,[25] whereas deep learning approaches—including convolutional neural networks (CNNs)[26] and hybrid CNN-recurrent neural network (RNN) architectures[27]—have been applied to predict soybean and corn yields with high accuracy. Srivastava et al.[28] further illustrated the potential of CNNs to predict winter wheat yields by integrating phenological and environmental variables.

Building on this body of work, the present study investigates corn yield prediction across the Greater Mississippi River Basin, one of the United States' largest and most critical agricultural regions, from 2008 to 2021. Using NDVI measurements and high-resolution weather datasets, we assess the relationship between vegetation dynamics, weather factors (including minimum temperature, maximum temperature, and mean precipitation), and corn yield across four distinct weather zones. We also propose a novel vegetation growth metric (VGMmc) that maximizes correlation with yield outcomes. Unlike traditional studies that primarily employ static NDVI averages, this work introduces a dynamic, data-driven approach for vegetation quantification. Furthermore, by integrating DOE-defined weather zone stratification, this study provides regionally adaptive modelling insights.

Recent advances in deep learning and temporal hybrid models (e.g., Srivastava et al., 2022; Khaki and Archontoulis, 2023; Xie et al., 2025; Zhang et al., 2024) demonstrate that combining multisource data with temporal dependencies can substantially enhance crop yield forecasting. Motivated by these developments, our objectives are threefold: (1) to design a multisource, data-driven modeling framework for corn yield prediction across diverse weather zones; (2) to introduce and validate a new vegetative growth metric optimized for maximum correlation with yield; and (3) to compare linear and nonlinear machine learning models to identify robust predictive strategies applicable under varying climatic conditions.

2 Study Area

The geographical focus of this study is the Greater Mississippi River Basin, a vital agricultural corridor encompassing much of the central United States. Spanning over 3 million square kilometers and covering parts of 31 U.S. states and two Canadian provinces, the basin supports extensive corn cultivation due to its fertile soils and varied climate zones. The region exhibits diverse weather patterns, making it ideal for studying environmental influences on crop yield. Based on climate classification criteria established by the U.S. Department of Energy, the basin is categorized into four major zones:[29] very cold, cold, mixed humid, and hot humid. These zones differ markedly in seasonal temperature ranges and precipitation profiles, which in turn affect crop growth cycles and productivity potential. This climatological diversity allows for a comprehensive analysis of weather-yield dynamics.

Corn cultivation across these diverse zones must contend with significant weather variability, impacting planting schedules, growth dynamics, and final yields. Understanding how corn responds to these environmental gradients is critical for developing robust, region-specific yield prediction models. Figure 1 illustrates the spatial extent of the study area and highlights the mean annual runoff flows across the different weather regions, providing important context for the basin's environmental heterogeneity and its implications for corn production.

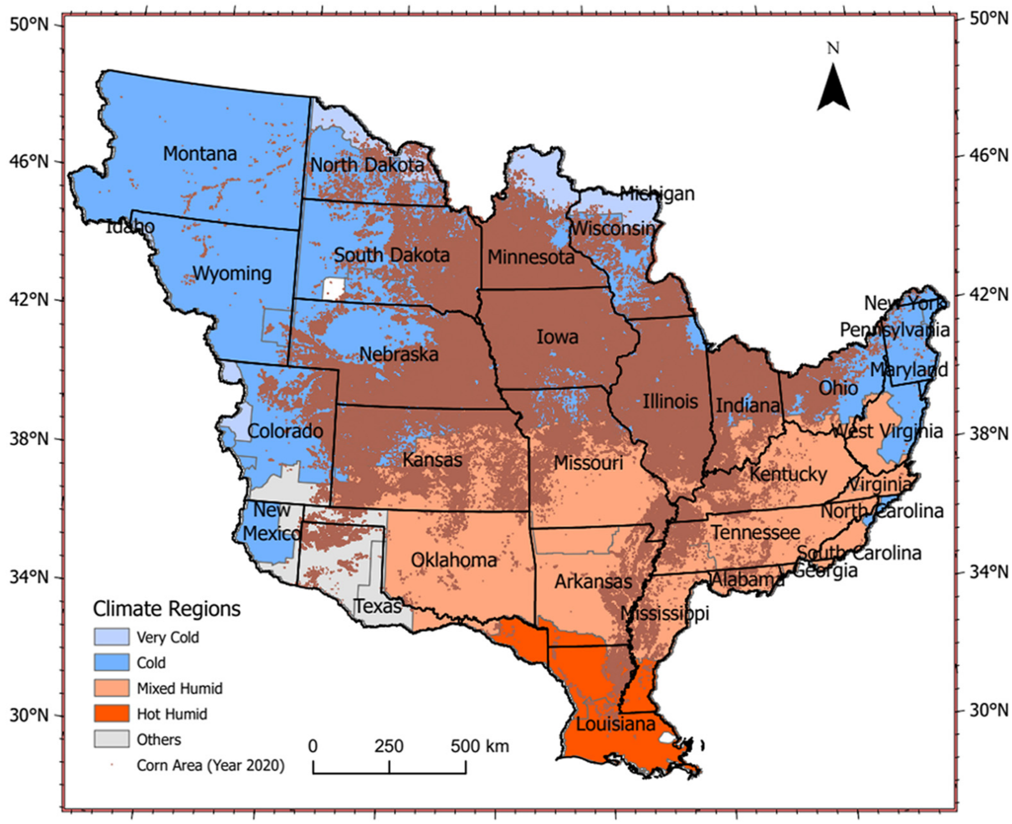

Fig. 1 Study area of the Greater Mississippi River Basin for corn yield modeling.

3 Materials and Methods

3.1 Data Collection

To develop the yield prediction framework, we compiled five publicly available datasets spanning from 2008 to 2021. These include the following:

1. Corn Cropland Layers from the USDA NASS program, providing high-resolution spatial boundaries (30 m) of corn-growing areas, used for crop-specific masking.
2. County-Level Yield Statistics, also from USDA NASS, offering yearly yield values (in bushels/acre) across U.S. counties, which serve as reference outputs for training and validation.
3. MODIS NDVI (MOD13Q1) time series at 250 m resolution and 16-day intervals, which were scaled and temporally aligned to capture in-season vegetative trends.
4. Climatic Data from the WorldClim database, including monthly averages of precipitation and both minimum and maximum temperatures.
5. Weather Zone Definitions, based on DOE's Building America Program, are used to stratify the study area into four major climate zones for zonal analysis. All spatial layers were reprojected to a unified coordinate reference system, and the data were processed to align temporally with the April–October corn growing season.

All spatial layers were reprojected to a unified coordinate reference system, and the data were processed to align temporally with the April–October corn growing season.

3.2 Geographic Data Processing

The geographic data processing involved a series of steps to prepare the datasets for analysis and modeling, ensuring that all data were properly aligned and structured for further analysis. Figure 2 illustrates the processing workflow, including:

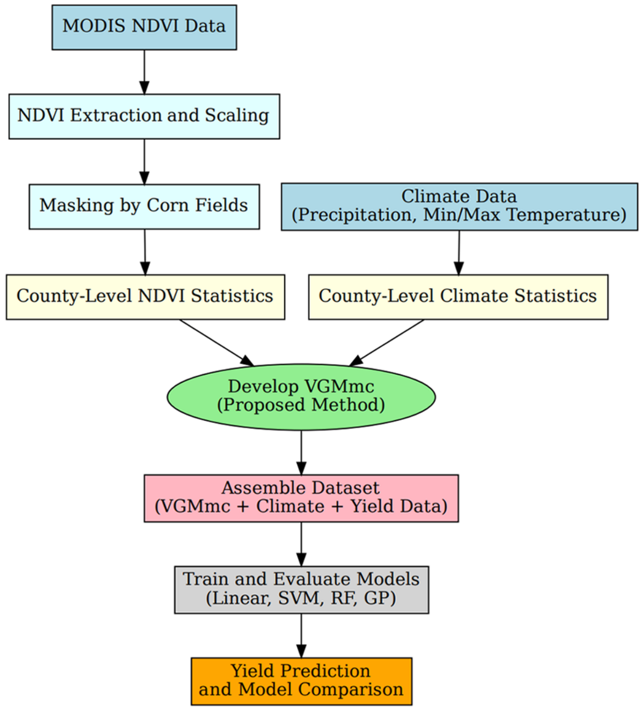

Fig. 2 Schematic diagram for geographic data processing and yield modeling.

i). *NDVI Preprocessing:* The first step in the data processing pipeline involved extracting the NDVI time series from the MODIS product. MODIS provides 16-day composite NDVI images, but raw digital numbers (DN) need to be converted into NDVI values that range from −1 to +1. We applied the MODIS scale factor to convert the DNs to NDVI values, which are then used to quantify vegetation health. NDVI values provide critical insights into crop growth dynamics, as they reflect the amount of photosynthetically active vegetation. We retained this 16-day interval to capture key vegetative phases during the corn growing season, from emergence to maturity.

ii). *Field Masking by Corn Croplands*: Since the spatial resolution of the MODIS NDVI data (250 m) is coarser than that of the CDL (30 m), the next step involved downscaling the corn field data to match the MODIS resolution. We created a mask based on the CDL dataset, identifying all MODIS pixels whose footprint completely overlapped with corn fields. By masking the NDVI data with this corn-specific mask, we ensured that only the NDVI values corresponding to corn fields were used in subsequent analyses. This step is essential to focus on the crop-specific vegetation indices, which are critical for modeling corn yield.

iii). *County-Level Aggregation:* For each 16-day NDVI composite, summary statistics (mean, min, max, and standard deviation) were computed at the county level.

iv). *Weather Variable Summarization:* Climatic data were aggregated seasonally per county to calculate average precipitation and temperature metrics from April to October. These weather summaries capture the essential environmental variables that influence corn growth, such as temperature extremes and water availability.

v). *Dataset Assembly*: In the final step, we integrated the processed NDVI statistics, weather summaries, and county-level yield data into a single, comprehensive dataset. This dataset was organized by weather zone, enabling us to assess how corn yield is influenced by both vegetation growth (as captured by NDVI) and weather conditions (temperature and precipitation) across different weather regions. The dataset was prepared for modeling, where we applied correlation analysis and various predictive models to explore the relationships between the independent variables (NDVI and weather) and corn yield. These steps were implemented using a combination of Python packages such as rasterio, geopandas, and pandas, allowing for efficient handling and processing of large geospatial datasets. The preprocessing pipeline ensured that the data were ready for subsequent analysis, providing a robust foundation for the predictive models.

3.3 Correlation Analysis

To explore how vegetation health and climatic variables relate to corn yield, Pearson correlation coefficients were calculated across multiple spatial and temporal scales. Specifically, NDVI values sampled at 16-day intervals were analyzed for their strength of association with yield metrics at the county level. Similarly, aggregated monthly weather variables (precipitation, $T_{\min}, T_{\max}$) were evaluated for their independent contributions to yield variation. This analysis helped identify the most influential periods and features for each weather region, informing the design of region-specific predictive models. Mathematically, for any two predictors, e.g., x and y, with a total of n measurements, Pearson's correlation coefficient between them can be computed using the formula:

$$\rho = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \sqrt{n \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

where $\mathrm{cov}(x, y)$ is the covariance between two variables: x and y, and σ_x and σ_y are their standard deviation, respectively.

3.4 Vegetative Growth Metrics with Maximum Correlations

An effective vegetative growth metric can enhance the yield modeling and prediction performance, and recent studies have explored various metrics of NDVI, such as VGM85 (i.e., mean NDVI over 85 days from the emergence), VGM120 (i.e., mean NDVI over 120 days from the emergence), $\mathrm{VGM}_{\mathrm{mean}}$ (i.e., mean NDVI over the growth season), and $\mathrm{VGM}_{\mathrm{max}}$ (i.e., maximum NDVI over the growth season), etc.[18] It has been shown that these vegetative growth metrics achieve different yield prediction performance depending on the weather regions, whereas $\mathrm{VGM}_{\mathrm{max}}$ is identified as one of the most significant predictors. Motivated by the fact that Pearson's coefficient is essentially the standardized slope of a simple linear regression line, i.e., a larger Pearson's coefficient indicating a stronger strength of the linear relationship between NDVI measurements and the yield, we propose a data-driven accumulated vegetative growth metric that aims to maximize the linear correlation coefficient. This metric is the sum of NDVI measurements over several selected days from days 1, 17, 33, …, 353, i.e., the total greenness of vegetation over a period. However, the exhaustive search for the best subset of these days is known as a NP hard problem, if one examines each possible combination. Here, we leverage a greedy approach to select the best subset of NDVIs by maximizing the correlations, as detailed below:

- Step 1: The correlation between mean NDVIs on each individual day and corn yield is measured, as described in Sec. 3.3.
- Step 2: Then, it ranks these 16-day interval NDVIs in descending order. The largest correlation indicates that the NDVI of this day is strongly correlated with the yield, which typically also has the highest greenness (i.e., largest NDVI), as shown in our results. Therefore, the NDVI with the largest correlation, denoted as S_1, is selected at the end of this step.
- Step 3: Next, the NDVI with the second largest correlation, denoted as S_2, is selected for consideration. The sum of S_1 and S_2 indicates the total greenness of these 2 days, denoted as $S = S_1 + S_2$, and we measure its correlation with the yield. If this new correlation coefficient increases compared with the one with S_1 only, we will repeat this procedure by considering the one with the third largest correlation and continue until the new correlation coefficient is reduced.

In the end, this approach will produce a sum of M NDVIs, $\{S_1, S_2, \ldots, S_M\}$, that achieves the largest correlation coefficient with the yield. We take its average, i.e., $1/M \sum_{i=1}^{M} S_i$, which ranges from 0 and 1, and call this maximum correlation-based vegetative growth metric as VGM_{mc}.

3.5 Yield Prediction Modelling

For crop yield prediction, we applied linear regression and a range of machine learning algorithms for crop yield prediction, including both linear and nonlinear regression models that have been trained with NDVI-based vegetation metrics and weather factors.

- Linear regression: A basic linear regression model was used to capture linear relationships between explanatory variables and response variables. We first considered only vegetation metrics as explanatory variables and then added weather factors into the model to illustrate the contribution of weather factors for corn yield prediction. The linear regression models are summarized below:

$$y = + \sum_{i=1}^{k} a_i x_i^{\mathrm{NDVI}}$$

$$y = b + \sum_{i=1}^{k} a_i x_i^{\mathrm{NDVI}} + \sum_{i=1}^{l} c_i x_i^{c}$$

  where y denotes crop yield, known as the response variable; b is the intercept for the estimated line; x_i^{NDVI} is the i'th NDVI vegetation growth metric, and x_j^c is the j'th climatic factor, known as an explanatory variable. To solve the multicollinearity issue that commonly exists in multidimensional regression models, stepwise regression is performed to find the subset of parameters for crop yield prediction by automatically removing or adding variables in a step-by-step manner.
- Nonlinear quadratic regression: Compared with the linear regression model, the nonlinear regression model includes additional squared terms for each explanatory variable and all products of pairs of distinct explanatory variables. Mathematically, assuming we have m explanatory variables, the nonlinear regression model predicts the following:

$$y = b + \sum_{i=1}^{m} a_i x_i^{\mathrm{NDVI}} + \sum_{i=1}^{m} c_i (x_i^{\mathrm{NDVI}})^2 + \sum_{i \neq j} d_{ij} x_i^{\mathrm{NDVI}} x_j^{\mathrm{NDVI}}$$

$$y = b + \sum_{i=1}^{m} a_i x_i^{\mathrm{NDVI}} + \sum_{i=1}^{m} c_i (x_i^{\mathrm{NDVI}})^2 + \sum_{i \neq j} d_{ij} x_i^{\mathrm{NDVI}} x_j^{\mathrm{NDVI}} + \sum_{i=1}^{m} a_i x_i^{c} + \sum_{i=1}^{m} c_i (x_i^{c})^2 + \sum_{i \neq j} d_{ij} x_i^{c} x_j^{c}$$

  where c_i and d_{ij} are coefficients for squared terms and their products of pairs, respectively. Note that the above regression model is still linear in its coefficients, and we can still apply the stepwise regression technique to find the best model for crop yield prediction.
- Support vector machine (SVM) regression: The SVM algorithm was originally proposed for classification, which is able to find a linear hyperplane to separate data from different classes (Cortes and Vapnik, 1995), and it was lately extended for prediction modeling, known as support vector regression (SVR) (Drucker et al., 1996). In our SVR, a Gaussian kernel is used to transform the input data to a higher-dimensional space, and it finds the hyperplane by minimizing:

$$|y_{\mathrm{combined}} - \langle w, x \rangle - b| \leq \epsilon$$

  where x is the input data, w is the normal vector to the hyperplane, and b is the intercept of the hyperplane. After training, we use the inner product $\langle w, x \rangle$ and intercept to predict the outcome within the range of ϵ.
- Gaussian process (GP) regression: GP regression is a well-known nonparametric Bayesian method for regression tasks. A GP is defined as a collection of random variables, any finite subset of which follows a joint Gaussian distribution. Formally, a GP model is computed by a mean function $m(x)$ and a covariance function $k(x, x')$, written as

$$f(x) \sim \mathrm{GP}(m(x), k(x, x')),$$

  where $m(x) = E[f(x)]$ and $k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$. These functions define the behavior of the GP, with the covariance function $k(x, x')$ controlling the smoothness and correlation of the outputs. The flexibility of the GP framework comes from its ability to model complex functions using different choices of the covariance function, such as the squared exponential kernel. GPR provides not only point predictions but also uncertainty estimates, making it particularly useful for tasks requiring confidence-aware prediction.

- XGBoost regression: Among various regression methods used in practice, the gradient boosted regression tree, especially the XGBoost regression approach[30] has shown state-of-the-art prediction results in many complex and challenging regression problems and competitions. The XGBoost regression method trains a set of K decision trees and makes the final prediction based on the sum of the predictions from each tree. It starts with a simple prediction, often the mean of the target values, and iteratively builds and adds new trees. Each new tree is learned to correct the errors (or residuals) made by previous trees. Mathematically, it aims to minimize the mean-squared error (MSE) with a regularization term that prevents overfitting, i.e., the model fitting the training data too much and making it unable to generalize to unseen data. Suppose at iteration t with the current prediction $\hat{y}_i^{(t)}$, XGBoost approximates the loss using a second-order Taylor expansion

$$l(y_i, \hat{y}_i^{(t)} + f_t(x_i)) \approx l(y_i, \hat{y}_i^{(t)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2$$

  where $l(.)$ is the differentiable loss function, g_i and h_i are the first and second derivatives, respectively. The goal is then to build a new tree $f_t(x_i)$ to fit the negative gradient for each training data. After K trees are trained, the final prediction is simply the sum of the outputs from all the trees:

$$\hat{y}(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

To ensure reproducibility and methodological clarity, the dataset was divided into training and testing subsets using a 90/10 split, where 90% of the data were randomly selected for model calibration and 10% reserved for independent testing. In addition, we employed fivefold cross-validation within the training set to assess model robustness and mitigate overfitting. Each cross-validation iteration used a different random seed to guarantee stability of results.

3.6 Performance Metrics

To assess the predictive capabilities of the various models developed in this study, we employed a set of commonly accepted evaluation metrics that quantify both accuracy and generalization:

1. Adjusted R-squared (adjusted R^2): This metric reflects the proportion of variance in corn yield explained by the model while adjusting for the number of predictors used. It penalizes overfitting, offering a more reliable comparison between models with differing complexities. A higher adjusted R^2 indicates a better fit without unnecessary variables.
2. Root mean-squared error (RMSE): RMSE measures the average prediction error. It is computed as the difference between values predicted by a model and the values actually observed from the thing being modeled. This metric is sensitive to large errors, making it useful for identifying model robustness.
3. Normalized root mean-squared error (NRMSE): To enable fair comparison across different scales or datasets, RMSE is normalized by the mean of the observed yield. The NRMSE expresses model error as a percentage of the average yield, offering an interpretable measure of relative accuracy.

Mathematically, the above performance metrics are written as follows:

$$\text{Adjusted } R^2 = 1 - \frac{(N-1)\text{SSE}}{(N-P)\text{SST}}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

$$\text{NRMSE} = \frac{\text{RMSE}}{1/N \sum_{i=1}^{N} \hat{y}_i}$$

where N is the number of data observations and P is the number of predictors. SSE denotes the sum of squared residuals, and SST denotes the total sum of squares.

4 Results and Discussions

4.1 Temporal Variable Statistics

We first summarize the statistics of our selected variables used in this research with time, including corn yields, vegetative growth metrics, and weather factors.

4.1.1 *Corn yield*

Figure 3 shows the average yield (kg/ha) of the corn in each weather area over different years. The results indicate that, overall, the cold region consistently produced higher corn yields compared with the other weather zones—except during the 2012 to 2014 period, when the hot humid region experienced a notable surge in yield. Yield trends in the hot humid and mixed humid zones were largely comparable, whereas the very cold region consistently recorded the lowest average yields throughout the study period.

4.1.2 *Vegetative growth metrics and weathers*

NDVI values offer a quantitative indicator of vegetation and biomass by capturing reflectance differences in the red and near-infrared spectral bands, which change dynamically over the course of the growing season. In parallel, weather conditions exert a significant influence on corn development, which typically begins with planting in early April and concludes with harvest by late October, as outlined by the USDA National Agricultural Statistics Service (USDA-NASS, 2010). For each weather region, we use the proposed vegetative growth metric, called VGM_{mc}, that maximizes the linear correlation with the yield over its selected days, as shown in Table 1. It can be shown that Pearson's correlation coefficient of VGM_{mc} is increased compared to other individual days.

Throughout the corn growing season, the vegetative growth metric $\mathrm{VGM}_{\mathrm{mc}}$ was evaluated for each weather zone. Figures 4(a)–4(d) present the temporal patterns of $\mathrm{VGM}_{\mathrm{mc}}$ alongside key weather factors, i.e., mean precipitation, minimum temperature, and maximum temperature,

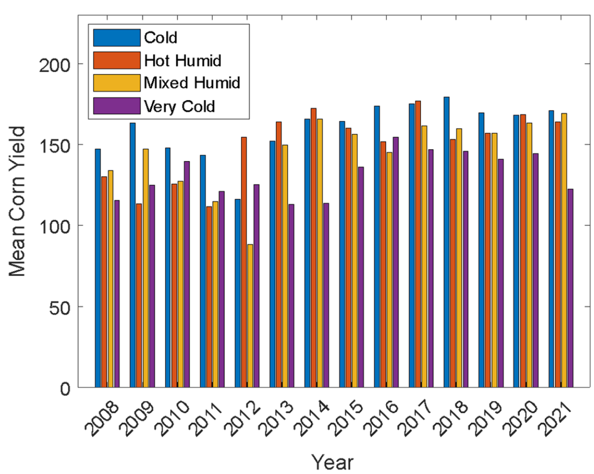

Fig. 3 Average corn yield of four weather regions over the years.

Table 1 Days selected by $\mathrm{VGM}_{\mathrm{mc}}$ maximizing Pearson's correlation coefficient for different weather regions.

	Day with peak correlation (Pearson's coefficient)	Days with maximum correlation (Pearson's coefficient)
Cold region	Day 209 (0.764)	Day 177, day 193, day 209, day 225, and day 241 (0.788)
Hot humid region	Day 177 (0. 572)	Day 161 and day 177 (0. 590)
Mixed humid region	Day 193 (0. 625)	Day 177, day 193, day 209, and day 225 (0.679)
Very cold region	Day 209 (0.685)	Day 177, day 193, day 209, day 225, and day 241 (0.727)

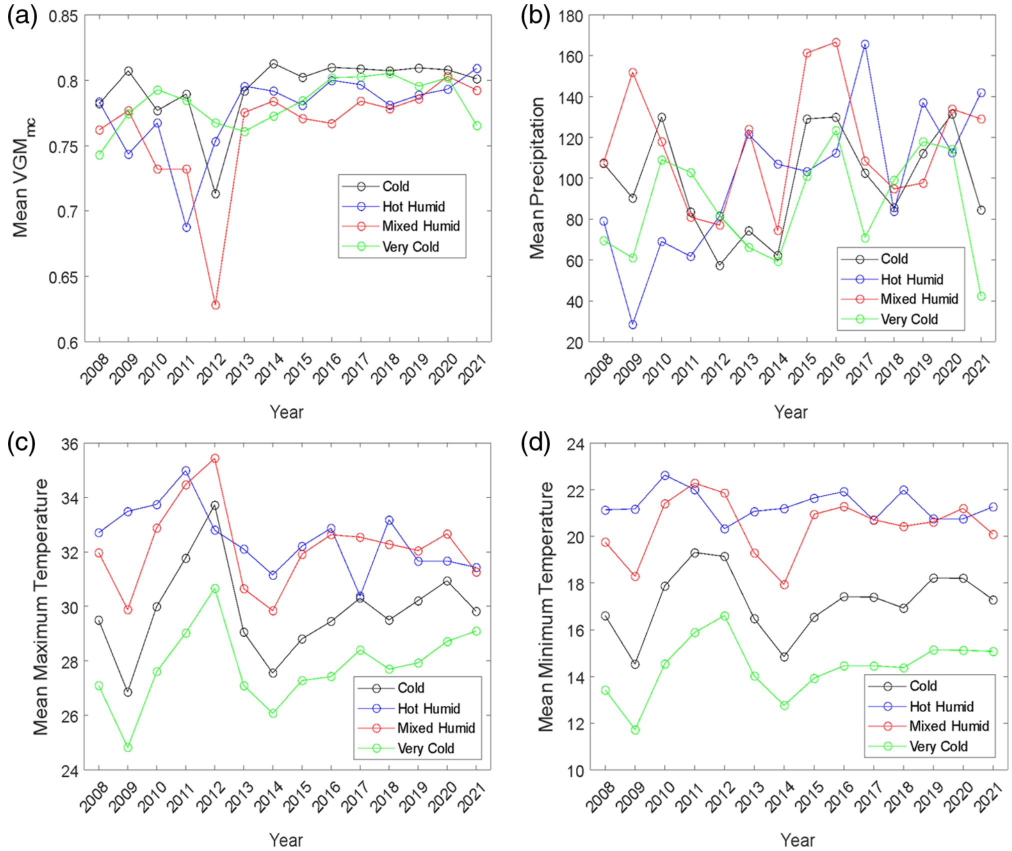

Fig. 4 Temporal curves of four predictors: (a) VGM$_{mc}$, (b) mean precipitation, (c) minimum temperature, and (d) maximum temperature, over the selected days that produce VGM$_{mc}$.

across the four climate regions. As shown in Fig. 4(a), the cold zone consistently recorded the highest VGM$_{mc}$ values, whereas the mixed humid region displayed the lowest. Notably, both the cold and mixed humid regions experienced sharp declines in VGM$_{mc}$ during the 2012 season. Figure 4(b) highlights that mean precipitation followed a relatively stable pattern across regions and years. In contrast, Fig. 4(c) illustrates clear regional differences in minimum temperatures, with the hot humid region showing the warmest night-time temperatures and the very cold region the lowest. Interestingly, 2012 saw a marked rise in minimum temperatures in both the cold and mixed humid zones, which may be linked to the observed reductions in yield and VGM$_{mc}$ that year. Similarly, Fig. 4(d) demonstrates notable variation in maximum temperatures among the four zones, further underscoring the diverse climatic influences across the study area.

4.2 Correlation Analysis

Figure 5 presents the correlation analysis between corn yield and key predictors—VGM$_{mc}$, mean precipitation, maximum temperature, and minimum temperature—evaluated at the county level across each weather region. The results demonstrate a consistently strong positive association between VGM$_{mc}$ and yield in all four regions. Among the weather-related variables, mean precipitation also shows a generally positive correlation with yield across the board. By contrast, maximum temperature exhibits a negative correlation in all regions, suggesting that lower daytime temperatures may be favourable for corn productivity. In addition, minimum temperature is positively correlated with yield in the cold and very cold zones, indicating that warmer nighttime conditions may enhance crop performance in cooler climates.

To further explore these relationships over time, we examined the temporal alignment between VGM$_{mc}$ and yield at the regional level. As shown in Fig. 6, both variables were normalized to the range [0, 1] to enable more intuitive visual comparisons. Across all regions, VGM$_{mc}$ and yield display a strong temporal correlation. Quantitatively, Pearson's correlation coefficients were calculated as follows: 0.92 for the cold region, 0.77 for the hot humid region,

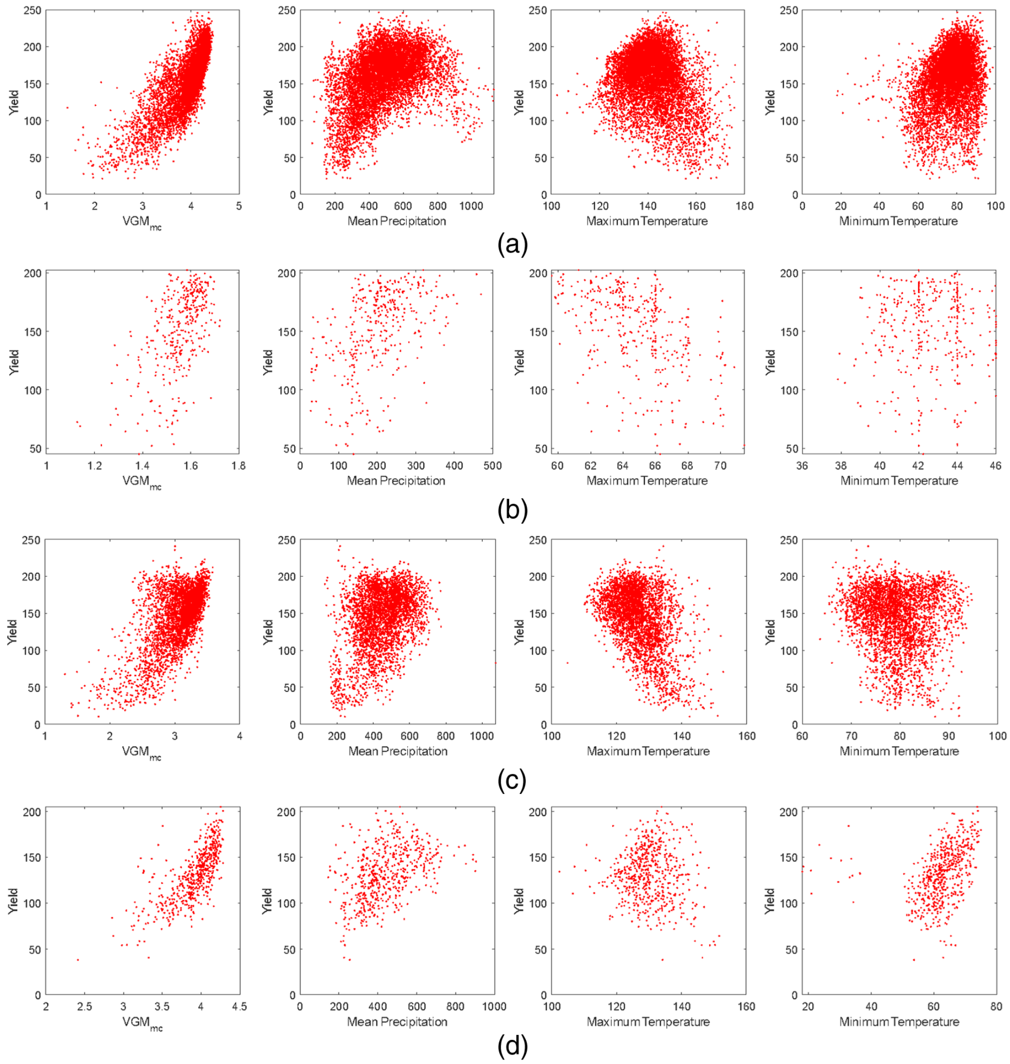

Fig. 5 Relationship between the corn yield and predictors at the county level in (a) cold, (b) hot humid, (c) mixed humid, and (d) very cold regions.

0.94 for the mixed humid region, and 0.89 for the very cold region. These values affirm the robustness of VGM_{mc} as a predictive indicator of yield under varying weather conditions.

4.3 Yield Modeling and Prediction

This section presents the performance of different modelling strategies using both NDVI-based and climatic predictors. All models were evaluated using the 90/10 split and verified through fivefold cross-validation. For each weather zone, we developed two types of linear regression models: y_{NDVI}, which relies solely on the vegetative growth metric VGM_{mc} as the predictor, and y_{combined}, which incorporates both VGM_{mc} and three weather factors—mean precipitation, maximum temperature, and minimum temperature. By comparing the predictive performance of these models, we assess the added value of including climatic factors in yield forecasting. Table 2 presents the optimal linear models identified for each weather region through a stepwise regression approach, guided by the Bayesian information criterion (BIC) for model selection. The analysis shows that VGM_{mc} consistently contributes positively to yield prediction across all regions. Furthermore, weather variables exhibit region-specific effects: in the cold and very cold zones, both precipitation and maximum temperature are positively associated with yield, whereas in other regions, different trends emerge, including some negative associations.

Table 3 lists the performance of the best estimated linear regression models, SVR, Gaussian process regression, and XGBoost regression methods. These results show that XGBoost can

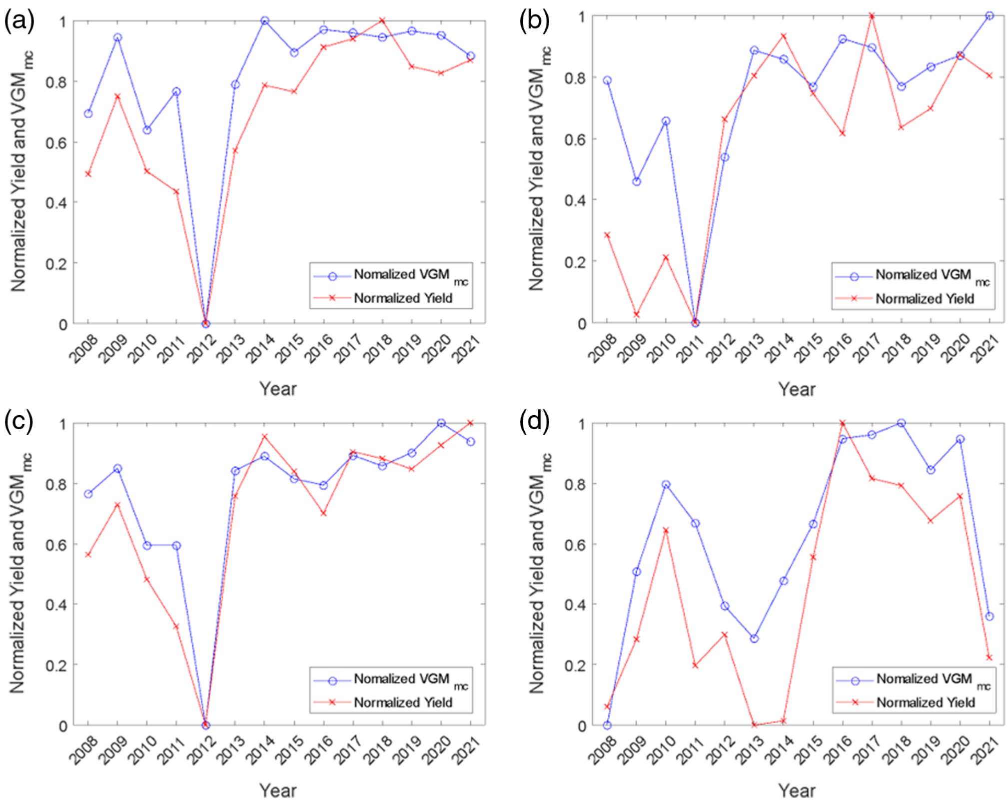

Fig. 6 Relationship between the normalized VGMmc and corn yield over time, both normalized, in (a) cold, (b) hot humid, (c) mixed humid, and (d) very cold.

Table 2 Best linear prediction models for different weather regions.

Weather region	Best linear models
Cold	$y_{\text{combined}} = 94.08 \times \text{VGM}_{\text{mc}} + 0.02 \times \text{Precipitation} + 1.50 \times \text{Max } T - 0.85 \times \text{Min } T - 371.96$
Hot humid	$y_{\text{combined}} = 144.41 \times \text{VGM}_{\text{mc}} - 6.49 \times \text{Max } T + 6.32 \times \text{Min } T + 72.12$
Mixed humid	$y_{\text{combined}} = 72.32 \times \text{VGM}_{\text{mc}} - 2.06 \times \text{Max } T + 2.80 \times \text{Min } T - 26.94$
Very cold	$y_{\text{combined}} = 92.88 \times \text{VGM}_{\text{mc}} + 0.02 \times \text{Preciitation} + 0.71 \times \text{Max } T - 0.95 \times \text{Min } T - 264.44$

All models have a p value less than 0.00001.

achieve the best prediction performance in terms of Adj R2, RMSE, and NRMSE for all weather regions.

Although nonlinear regression models often provide a closer fit to the training data compared with linear models, they are also more susceptible to overfitting, which can limit their effectiveness in forecasting future crop yields. To evaluate model performance in a more realistic setting, we applied cross-validation to both linear and nonlinear approaches. In this process, 90% of the dataset was randomly selected for training, whereas the remaining 10% was reserved exclusively for testing—ensuring that the evaluation reflects the model's ability to generalize to unseen data. Table 4 summarizes the cross-validated model prediction performance between linear and nonlinear models for each weather region. Compared with the results reported in Table 3, the performances of all nonlinear models drop due to the potential overfitting issue for their high model complexity. However, this does not occur for linear regression models. Again, the XGBoost regression method shows the state-of-the-art prediction performance for all weather regions.

Table 3 Comparison of prediction performance between linear regression and nonlinear machine learning models for each weather region for corn crop, where the best Adj R^2 is highlighted as underlined and bold, and the second best one is highlighted as bold.

Models	Performance	Cold	Hot humid	Mixed humid	Very cold
Linear regression	Adj R^2	0.62	0.44	0.45	0.55
	RMSE	21.13	26.05	28.82	19.45
	NRMSE	0.08	0.17	0.20	0.15
Nonlinear quadratic regression	Adj R^2	0.64	0.40	0.49	0.56
	RMSE	20.67	27.42	27.12	19.11
	NRMSE	0.07	0.18	0.18	0.15
SVM regression	Adj R^2	0.66	0.34	0.42	0.39
	RMSE	21.29	28.27	29.75	22.52
	NRMSE	0.13	0.18	0.21	0.17
Gaussian process	Adj R^2	**0.75**	0.30	**0.84**	**0.65**
	RMSE	18.05	29.14	15.72	17.03
	NRMSE	0.11	0.19	0.11	0.13
XGBoost	Adj R^2	0.92	0.99	0.93	0.99
	RMSE	12.50	0.30	12.08	0.73
	NRMSE	0.05	0.01	0.06	0.01

Table 4 Comparison of cross-validated prediction performance between linear and nonlinear models, where the best Adj R^2 is highlighted as underlined and bold, and the second best one is highlighted as bold.

Models	Performance	Cold	Hot humid	Mixed humid	Very cold
Linear regression	Adj R^2	0.69	**0.56**	0.40	0.60
	RMSE	20.72	23.95	28.28	21.02
	NRMSE	0.13	0.16	0.19	0.16
Nonlinear quadratic regression	Adj R^2	0.71	0.55	0.43	0.61
	RMSE	20.01	24.10	27.83	21.21
	NRMSE	0.13	0.16	0.18	0.16
SVM regression	Adj R^2	0.68	0.35	0.38	0.54
	RMSE	19.63	28.89	28.63	21.43
	NRMSE	0.12	0.19	0.20	0.17
Gaussian process	Adj R^2	**0.78**	0.27	**0.56**	**0.66**
	RMSE	17.22	30.65	24.17	18.36
	NRMSE	0.11	0.20	0.16	0.15
XGBoost	Adj R^2	0.81	0.58	0.62	0.71
	RMSE	17.10	22.86	23.28	16.42
	NRMSE	0.10	0.15	0.14	0.13

5 Discussions

This study presents a comprehensive data-driven framework for predicting corn yields across diverse weather regions within the Greater Mississippi River Basin. By integrating vegetation dynamics and weather factors, we developed and compared both linear and machine learning models for spatially robust yield estimation. Our findings not only validate the strong predictive capacity of NDVI-based vegetative growth metrics but also advance the understanding of how weather variables differentially impact corn yield across distinct environmental settings.

First, the novel VGM_{mc} metric developed in this study, based on maximizing correlation with yield outcomes, offers a significant enhancement over traditional static NDVI metrics. Consistent with prior research,[5,18] which emphasized the utility of NDVI-derived indices for crop yield forecasting, our results reaffirm that vegetation dynamics are central to predicting corn productivity. However, by optimizing the selection of critical NDVI observations through a data-driven approach, this study improves the sensitivity and robustness of vegetation-based predictors, particularly across heterogeneous weather zones.

Although NDVI is a strong in-season indicator of crop health, its retrospective nature limits direct application for forward-looking yield forecasting. To address this constraint, we discuss several potential strategies: (1) lag-based NDVI models, which use early-to-mid-season NDVI composites to estimate end-of-season yield; (2) integration with seasonal weather forecasts, enabling the substitution of future NDVI with predicted climate-driven vegetation states; and (3) hybrid frameworks combining crop simulation outputs (e.g., DSSAT or APSIM) with machine learning to project yields before harvest. These approaches can mitigate the unavailability of future NDVI data and extend model applicability for operational forecasting.

In addition to vegetation indices, the study emphasizes the substantial influence of weather variables, including mean precipitation, minimum temperature, and maximum temperature, on corn yield prediction. Similar to findings in Ref. 18, our correlation analysis revealed clear and region-specific temporal relationships between weather factors and corn yield. Specifically, positive correlations between precipitation and yield were observed across all weather zones, whereas the impacts of temperature varied regionally: warmer minimum temperatures benefited yield in the cold and very cold regions, whereas cooler maximum temperatures favored yield in the hot humid and mixed humid regions. These insights align with previously reported geographic variations in weather-crop interactions[18,31–33] and further reinforce the need for weather zone-specific modeling approaches.

Notably, compared with previous studies such as Ref. 18, this work adopted the U.S. Department of Energy's Building America weather classifications—widely recognized across disciplines—to systematically segment the study area. This stratification allowed for a more granular assessment of environmental impacts on yield variability and revealed nuanced weather-yield relationships that could be critical for future agricultural adaptation strategies under weather scenarios.

In line with prior modeling efforts,[6,12,15,20,34] our study underscores the critical role of integrating phenological information (NDVI) with environmental variables for improving model accuracy. Although prior works have shown the promise of remote sensing data for yield prediction, this study contributes further by demonstrating that dynamic, optimized vegetative growth metrics, when combined with weather variables, can achieve stronger and more regionally stable predictive performance. Model comparison results illustrate that while nonlinear machine learning models, such as XGBoost, achieved the highest predictive performance with Adj R^2 values as high as 0.99 in some regions (Table 3), these models are also more susceptible to overfitting, particularly when tested on unseen data. Cross-validation experiments (Table 4) revealed a noticeable performance drop for nonlinear models, similar to observations in Refs. 8, 18, and 35, whereas linear regression models demonstrated greater stability and generalizability, even if at a slightly lower accuracy. This trade-off between model complexity and robustness is critical for real-world applications where unseen variability is inevitable. Compared with previous studies such as Ref. 15, our linear corn yield prediction models achieved competitive performance, with cross-validated adjusted R^2 values ranging from 0.56 to 0.69 and RMSE values generally lower than those reported for similar corn types and prediction tasks. These results demonstrate that carefully constructed linear models, when enriched with well-selected vegetative and weather predictors, can provide reliable and interpretable forecasts suitable for operational agricultural decision-making.

The regional differences in weather-yield relationships identified in this study offer valuable guidance for anticipating the impacts of weather variability and extremes on corn production. For instance, increases in minimum temperatures in colder regions could have beneficial effects on yields, whereas rising maximum temperatures in already hot, humid regions could exacerbate yield losses. Such differentiated impacts emphasize the urgency for region-specific adaptation measures and resilient agricultural practices.

The current study employed static NDVI and aggregated weather metrics to characterize seasonal crop conditions. However, recent developments highlight that temporal deep learning models, such as long short-term memory (LSTM) networks, temporal convolutional networks (TCNs), and hybrid CNN-RNN architectures, can capture intraseason dynamics more effectively. Incorporating such time-series frameworks could allow for real-time yield forecasting that evolves as new NDVI or climate observations become available. Recent studies[30,36,37] demonstrate substantial accuracy gains when temporal dependencies are modeled explicitly. Future work will extend our framework by integrating temporal regression and sequence modeling to achieve continuous yield forecasting capability. Moreover, motivated by the study using an ensemble method,[38] the combination of multiple regression methods could significantly boost the performance for yield predictions.

In conclusion, this study not only enhances methodological approaches for crop yield forecasting but also provides important empirical evidence to inform weather resilience strategies in agriculture. Future work should continue expanding the integration of remote sensing, high-resolution weather data, and machine learning to further strengthen predictive capabilities under evolving weather conditions.

6 Conclusion

This study developed a comprehensive, data-driven framework for predicting corn yields across the Greater Mississippi River Basin, integrating satellite-derived vegetation metrics and weather variables under varying environmental conditions. By proposing a vegetative growth metric (VGMmc) that optimizes the correlation with yield outcomes, we demonstrated that remote sensing data, when carefully selected and aggregated, can significantly enhance the accuracy of yield forecasting.

Our findings highlight the critical role of weather factors—particularly precipitation, minimum temperature, and maximum temperature—in shaping corn yields across different weather zones. The regional analysis revealed important spatial variability in how weather impacts corn production, providing new insights that can inform targeted adaptation strategies. Positive yield responses to increased minimum temperatures in colder regions and negative responses to maximum temperatures in hotter, more humid areas emphasize the differentiated vulnerabilities and opportunities facing U.S. agriculture under different weather conditions. Through extensive model evaluation, we showed that advanced nonlinear machine learning methods, such as XGBoost, offer superior predictive performance when modeling complex relationships among vegetation, weather, and yield. However, we also demonstrated that simpler linear regression models maintain better generalizability under cross-validation, suggesting that model selection should carefully balance accuracy and robustness depending on the intended application.

Compared with prior studies, this work makes several significant contributions: (i) it leverages an optimized, region-specific vegetation metric rather than static NDVI averages, (ii) it systematically incorporates weather zone classifications to account for environmental heterogeneity, and (iii) it provides rigorous cross-validation assessments to ensure model reliability. These advancements collectively strengthen the potential of data-driven approaches for operational corn yield forecasting and long-term agricultural planning. Future research could extend this framework by incorporating additional variables such as soil properties, crop management practices, and socioeconomic factors. Furthermore, linking yield models with weather projection data would provide valuable foresight into future yield scenarios and help guide weather resilience strategies for the agricultural sector. Finally, although the current framework is primarily designed for corn yield estimation, the methodological foundation can be extended to predictive forecasting by incorporating lag-based NDVI metrics, seasonal climate forecasts, and time-series modeling architectures. These directions will form the next phase of this research to bridge the gap between remote sensing-based assessment and forward-looking agricultural forecasting.

Acknowledgments

This research was supported in part by a postdoctoral research fellow appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DESC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, DOE, or ORAU/ORISE.

References

1. Food and Agriculture Organization (FAO), *The State of Food Insecurity in the World 2024* (2024).
2. H. Godfray et al., "Food security: the challenge of feeding 9 billion people," *Science* **327**(5967), 812–818 (2010).
3. T. R. Tenreiro et al., "Using NDVI for the assessment of canopy cover in agricultural crops within modelling research," *Comput. Electron. Agric.* **182**, 106038 (2021).
4. F. Waldner et al., "High temporal resolution of leaf area data improves empirical estimation of grain yield," *Sci. Rep.* **9**(1), 15714 (2019).
5. S. A. Shammi and Q. Meng, "Use time series NDVI and EVI to develop dynamic crop growth metrics for yield modeling," *Ecol. Indic.* **121**, 107124 (2021).
6. P. C. Doraiswamy et al., "Crop condition and yield simulations using Landsat and MODIS," *Remote Sens. Environ.* **92**(4), 548–559 (2004).
7. L. Wall, D. Larocque, and P.-M. Léger, "The early explanatory power of NDVI in crop yield modelling," *Int. J. Remote Sens.* **29**(8), 2211–2225 (2008).
8. M. S. Mkhabela et al., "Crop yield forecasting on the Canadian Prairies using MODIS NDVI data," *Agric. Forest Meteorol.* **151**(3), 385–393 (2011).
9. B. Seo et al., "Improving remotely-sensed crop monitoring by NDVI-based crop phenology estimators for corn and soybeans in Iowa and Illinois, USA," *Field Crops Res.* **238**, 113–128 (2019).
10. D. K. Bolton and M. A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agric. Forest Meteorol.* **173**, 74–84 (2013).
11. K. Yu et al., "Crop growth condition monitoring and analyzing in county scale by time series MODIS medium-resolution data," in *Second Int. Conf. on Agro-Geoinformatics (Agro-Geoinformatics)*, IEEE (2013).
12. X. Zhang and Q. Zhang, "Monitoring interannual variation in global crop yield using long-term AVHRR and MODIS observations," *ISPRS J. Photogramm. Remote Sens.* **114**, 191–205 (2016).
13. A. V. M. Ines et al., "Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction," *Remote Sens. Environ.* **138**, 149–164 (2013).
14. W. Xie, Y. Huang, and Q. Meng, "Soybean yield modeling and analysis with weather dynamics in the greater Mississippi River Basin," *Climate* **13**(2), 33 (2025).
15. D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States," *Remote Sens. Environ.* **141**, 116–128 (2014).
16. K. Frieler et al., "Understanding the weather signal in national crop-yield variability," *Earth's Future* **5**(6), 605–616 (2017).
17. J. Van Wart, P. Grassini, and K. G. Cassman, "Impact of derived global weather data on simulated crop yields," *Glob. Change Biol.* **19**(12), 3822–3834 (2013).
18. S. A. Shammi and Q. Meng, "Modeling crop yield using NDVI-derived VGM metrics across different climatic regions in the USA," *Int. J. Biometeorol.* **67**(6), 1051–1062 (2023).
19. O. Beeri and A. Peled, "Geographical model for precise agriculture monitoring with real-time remote sensing," *ISPRS J. Photogramm. Remote Sens.* **64**(1), 47–54 (2009).
20. J. Huang et al., "Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model," *Agric. Forest Meteorol.* **204**, 106–121 (2015).
21. T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: a systematic literature review," *Comput. Electron. Agric.* **177**, 105709 (2020).
22. J. You et al., "Deep Gaussian process for crop yield prediction based on remote sensing data," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 31 (2017).

23. N. Gandhi et al., "Rice crop yield prediction in India using support vector machines," in *13th Int. Joint Conf. on Comput. Sci. and Softw. Eng. (JCSSE)*, IEEE (2016).
24. Y. Everingham et al., "Accurate prediction of sugarcane yield using a random forest algorithm," *Agron. Sustain. Dev.* **36**(2), 27 (2016).
25. P. Bose et al., "Spiking neural networks for crop yield estimation based on spatiotemporal analysis of image time series," *IEEE Trans. Geosci. Remote Sens.* **54**(11), 6563–6573 (2016).
26. A. X. Wang et al., "Deep transfer learning for crop yield prediction with remote sensing data," in *Proc. 1st ACM SIGCAS Conf. Comput. and Sustain. Soc.*, pp. 1–5 (2018).
27. S. Khaki, L. Wang, and S. V. Archontoulis, "A CNN-RNN framework for crop yield prediction," *Front. Plant Sci.* **10**, 1750 (2020).
28. A. K. Srivastava et al., "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Sci. Rep.* **12**(1), 3215 (2022).
29. M. C. Baechler, *High-Performance Home Technologies: Guide to Determining Weather Regions by County*, U.S. Department of Energy, Energy Efficiency & Renewable Energy, Building Technologies Office (2015).
30. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowl. Discov. and Data Mining*, pp. 785–794 (2016).
31. D. B. Lobell and C. B. Field, "Global scale weather–crop yield relationships and the impacts of recent warming," *Environ. Res. Lett.* **2**(1), 014002 (2007).
32. T. Iizumi et al., "Responses of crop yield growth to global temperature and socioeconomic changes," *Sci. Rep.* **7**(1), 7800 (2017).
33. D. K. Ray et al., "Weather variation explains a third of global crop yield variability," *Nat. Commun.* **6**(1), 5989 (2015).
34. P. C. Doraiswamy et al., "Crop yield assessment from remote sensing," *Photogramm. Eng. Remote Sens.* **69**(6), 665–674 (2003).
35. L. Kouadio et al., "Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale," *Remote Sens.* **6**(10), 10193–10214 (2014).
36. Y. Ren et al., "Analysis of corn yield prediction potential at various growth phases using a process-based model and deep learning," *Plants* **12**(3), 446 (2023).
37. B. K. Osibo et al., "TCNT: a temporal convolutional network-transformer framework for advanced crop yield prediction," *J. Appl. Remote Sens.* **18**(4), 044513 (2024).
38. B. K. Osibo et al., "Integrating remote sensing data and efficiently weighted ensemble method for simplified crop yield prediction," *J. Appl. Remote Sens.* **18**(4), 044520 (2024).

Biographies of the authors are not available.