Scientific Reports

https://doi.org/10.1038/s41598-026-38724-z

Article in Press

Harvesting insights: interpretable machine learning to understand environmental drivers of U.S. maize and soybean yield

Received: 11 August 2025

Accepted: 30 January 2026

Published online: 13 February 2026

Cite this article as: Smith H.W., Heffernan C.J., Ashworth A.J. *et al.* Harvesting insights: interpretable machine learning to understand environmental drivers of U.S. maize and soybean yield. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38724-z

Harrison W. Smith, Christopher J. Heffernan, Amanda J. Ashworth, L. Lanier Nalley, David S. Bullock, Jason Tullis & Phillip R. Owens

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Submission ID: 598d859c-a6ee-4bc7-9e6a-09db7db76f89

Harvesting insights: Interpretable machine learning to understand environmental drivers of U.S. maize and soybean yield

Harrison W. Smith[a], Christopher J. Heffernan[b], Amanda J. Ashworth[c], L. Lanier Nalley[d], David S. Bullock[e], Jason Tullis[f], Phillip R. Owens[g]

Corresponding author: Harrison W. Smith, Email: hws001@uark.edu

[a]Environmental Dynamics Program, University of Arkansas, Fayetteville, AR
[b]Department of Computer Science and Computer Engineering, University of Arkansas, Fayetteville, AR
[c]USDA-ARS Poultry Production and Product Safety Research Unit, Fayetteville, AR
[d]Agricultural Economics and Agribusiness Department, University of Arkansas, Fayetteville, AR
[e]Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL
[f]Department of Geosciences, University of Arkansas, Fayetteville, AR
[g]USDA-ARS, Dale Bumpers Small Farms Research Center, Booneville, AR

Abstract

Accurate crop yield prediction is crucial for enhancing food security and agricultural sustainability; however, existing models frequently struggle to capture the intricate relationships between environmental drivers and crop performance. Here we leveraged a large, spatially explicit yield monitor dataset of U.S. commercial maize (*Zea mays*) and soybean (*Glycine max*) fields (134 unique crop-site-years). Machine learning models were trained to predict yield with high accuracy ($R^2 > 0.87$, RMSE < 1.13 Mg ha^{-1}), and Shapley Additive Explanations were used to quantify how weather, soil, and terrain properties predict yield variability. Our results highlight the potential of machine learning to disentangle environmental constraints on crop production, thereby providing actionable insights for more resilient U.S. food systems. The results presented here represent a novel approach to identifying maize and soybean yield constraints that can inform the next generation of crop breeding and precision management strategies.

Introduction

Maize and soybeans are the backbone of the modern United States (U.S.) agricultural economy and a critical part of global food security. In 2023, these two crops accounted for 57% of all U.S. agricultural area and half of total U.S. crop cash receipts, equivalent to $132 billion [1,2]. In 2023 alone, over $40 billion was generated through the export of these crops to global trading partners [3]. Global demand for maize and soybean is increasing, driven by factors such as increases in livestock production, biofuels, and global population growth [4–6]. However, several challenges constrain crop yields and hinder efforts to meet growing demand. These challenges include both biotic constraints (e.g., pest pressure, disease, weeds), as well as abiotic constraints (e.g., drought, temperature extremes, soil fertility)[7]. Management-related factors, including planting date, nutrient management, tillage, and cover cropping, also play critical roles in affecting crop yield[8].

Yield can be understood as a complex phenotypic expression resulting from genotype, environment, management, and their interactions[9]. Human influence on maize and soybean genotypes has expanded significantly over the last several decades, particularly in the era of advanced crop breeding and genetics[10,11]. Hybridization, advanced breeding, and genetic modification have further increased yields in maize and soybean, but this process has also led to genetic bottlenecks in both species, narrowing the genetic basis of U.S. maize and soybean cultivars[12,13]. This may increase potential for yield losses if environmental conditions deviate too far from the norm[14]. Advancements in management techniques like crop rotation, irrigation, fertilizer use, pest deterrence, and weed and disease control have led to additional increases in yield throughout the U.S.[15]. More recently, precision agriculture has emerged as a potential avenue for increasing yields through the incorporation of technology and data-driven agricultural decision making[16].

Despite historically increasing yields, environmental conditions continue to pose a unique challenge for producers and agricultural scientists. If environmental conditions shift too drastically, management strategies alone may not be enough to protect yields. Moreover, the interaction between heterogeneous environmental patterns and genotype shapes yield in ways that are not always well understood [17]. Understanding how these factors impact crop performance is essential to predicting yield and meeting future increases in demand. However, previous yield prediction efforts have often been limited by the availability of high-resolution, spatially explicit data. Large-scale yield prediction studies in the U.S. often rely on county-level data, which reflect broad regional patterns but do not accurately capture between-field patterns in yield[18].

This study aims to address this gap by utilizing a multi-year yield monitor dataset from farms across nine U.S. states (Nebraska, South Dakota, Illinois, Iowa, Indiana, Ohio, Oklahoma, Arkansas, and Pennsylvania). Yield monitor data provide spatially explicit point observations, which can be directly linked with other geospatial data. Here we bring together a large dataset spanning multiple states and years, including yield observations for rainfed maize and soybean fields (Figure 1). This dataset was paired with geospatial weather, soil, and terrain data to enable quantification of associations between environmental conditions and yield.

Because of the size of modern geospatial data collection, the number of relevant features, and a high degree of collinearity in environmental data, frequentist statistical methods are often challenging to use in yield prediction[19]. Machine learning (ML) is recognized as a strong alternative given its predictive power and ability to handle high-dimensional datasets with correlated features, though the limited interpretability of ML models constrains their ability to reveal the causal drivers of yield needed for actionable decision-making[20,21]. However, recent developments have improved interpretability in ML, offering a novel opportunity to explore heterogeneity and complex non-linear effects of the environment on yield[22].

We focused on three primary environmental conditions influencing yield: weather, terrain, and soil. Previous studies have demonstrated shortwave radiation, temperature, and precipitation are primary drivers of yield variability[20]. Terrain (shape and changes in elevation) also plays a vital role in crop development because it affects the movement of water and soil formation processes[23]. Soil properties such as texture, soil organic matter, water storage capacity, fertility, and many others also directly influence crop growth and yield [24]. Below, we present a framework for rigorously developed ML models trained to predict maize and soybean yield based on publicly available environmental data.

The overall objectives of this study were to (i) quantify how weather, soil, and terrain conditions influence maize and soybean yield across major U.S. production regions using yield monitor observations; (ii) evaluate the predictive performance and generalizability of ML models trained on publicly available environmental covariates; and (iii) identify the most influential environmental predictors using model-agnostic interpretability methods. We hypothesize that weather variables, especially temperature, solar radiation, and precipitation, would dominate maize yield predictions, while soil water-related attributes would play a greater role for soybean yield. We also expect non-linear and threshold-like responses for key variables, especially for high temperatures in maize. Overall, we anticipate that integrating yield monitor data with interpretable ML will reveal generalizable environmental patterns affecting yield across diverse U.S. growing conditions.

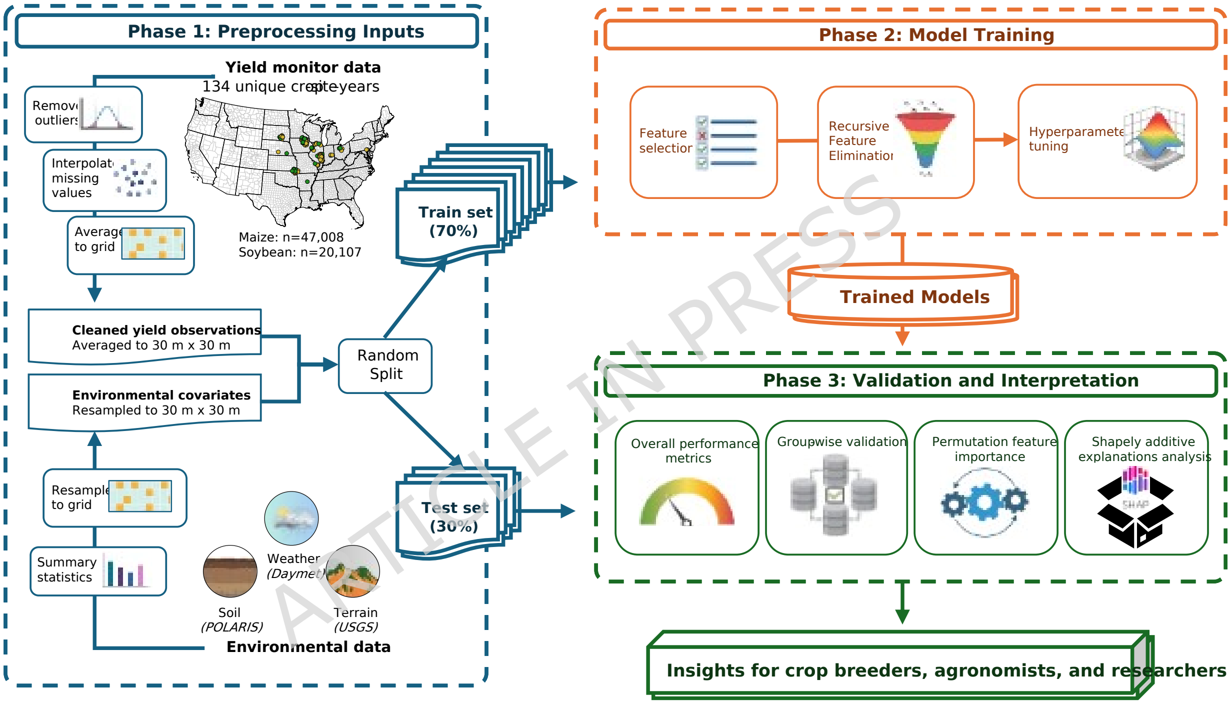

Figure 1. Workflow of the analysis used in this study to predict yield based on environmental covariates and yield monitor data collected via grain harvester.

Results

Machine learning models are highly predictive of yield

Several ML base models were evaluated for performance, including CatBoost, XGBoost, LightGBM, Extremely Randomized Trees, Neural Networks, Random

Forest, and K-Nearest Neighbors regression (Supplemental Table 1)[25]. Final trained models exhibited high accuracy in predicting yield for maize (R^2=0.87, RMSE=1.12 Mg ha^{-1}) and soybean (R^2=0.90, RMSE=0.46 Mg ha^{-1}). The top-performing maize model was a tuned LightGBM model, while the top soybean model was an Extremely Randomized Trees model [26,27]. Pearson's correlation coefficient was above 0.9 for both models, and a regression of observed vs. predicted values reveals a slight trend of overprediction when actual yield values are low, and a corresponding underprediction at higher yield values (Figure 2).

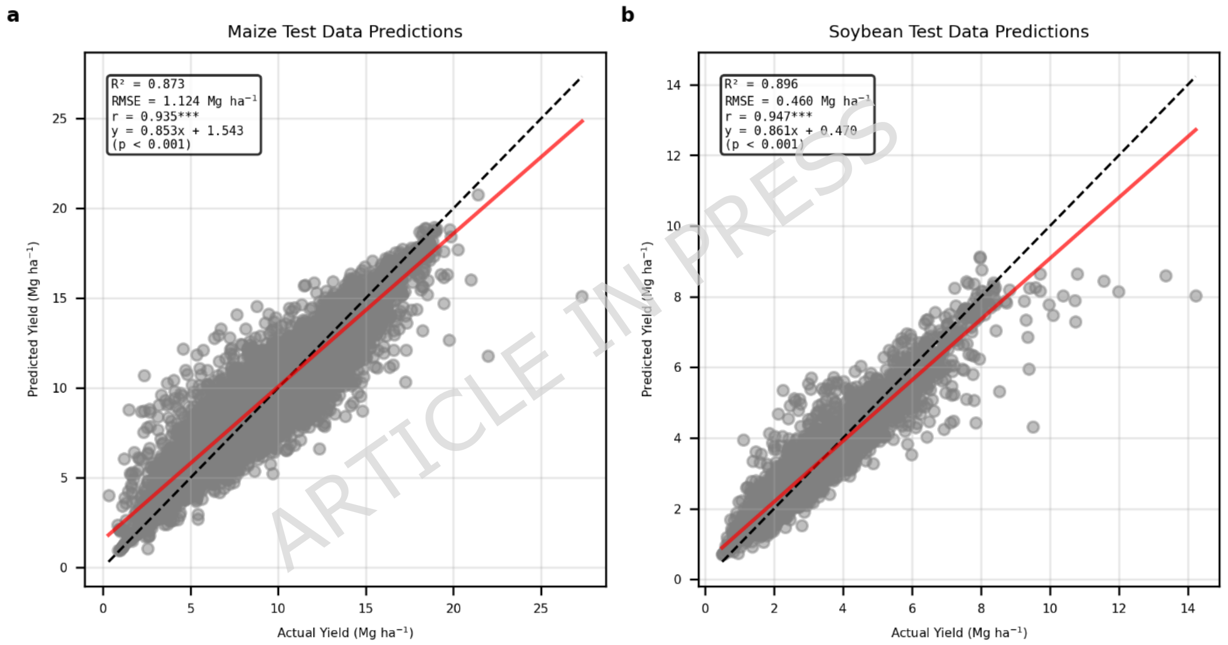

Figure 2. Comparison of model predicted yield values and actual yield values from testing datasets. A linear regression trend line is shown in red, while the black dashed line represents a perfect 1-to1 relationship.

Model generalizability and spatial independence of residuals

We found that the R^2 for individual years ranged from 0.45 to 0.93 in maize and 0.57 to 0.92 in soybean. R^2 averages across years were only slightly lower than the overall model (R^2=0.78 for maize and R^2=0.76 for soybean) [28,29]. Over the years, RMSE ranged from 0.57 to 1.42 Mg ha^{-1} for maize and 0.18 to 0.86 Mg ha^{-1} for soybeans. Group-wise validation across states revealed a mean R^2=0.77 in maize (ranging 0.54 to 0.91) and R^2=0.79 in soybean (ranging 0.67 to 0.88), while the mean RMSE across states was 1.13 Mg ha^{-1} in maize (ranging 0.35 to 1.76 Mg ha^{-1}) and 0.39 Mg ha^{-1} in soybean (ranging 0.25 to 0.69 Mg ha^{-1}). Validation metrics across all sites and years are included in Supplemental Figures 1 and 2.

To assess whether any unmodeled spatial structure remained in the predictions, we evaluated the spatial autocorrelation of model residuals across multiple spatial scales using multi-distance Moran's I. Residuals showed no detectable autocorrelation at 500 m or 50 km indicating that between-field and regional-scale patterns were appropriately captured by environmental covariates. However, a small but significant positive Moran's I was observed at 50 m (I=0.19, p=0.01), suggesting within-field spatial dependence remained below the 30 m grid resolution used for model training.

Feature importance highlights the key role of weather for maize yield

Feature elimination narrowed models from 128 initial features (see Supplemental Table 2 for complete list of initial variables) down to twenty features in the final maize model (Supplemental Table 3) and thirteen in the soybean model (Supplemental Table 4). Two methods were used to evaluate feature importance in models: permutation feature importance and Shapley Additive Explanations (SHAP) importance (Figure 3). The top five features in the maize model were the same for both methods, and weather variables accounted for four out of the five top features. Maximum daily temperature during the growing season (approximated as April 1 –

September 30 for all sites) was most important in maize yield predictions but was absent from the final set of soybean model features. Minimum daily shortwave radiation and standard deviation of precipitation were also important for predicting maize yield. Both importance methods ranked similar features highly in the soybean model: slope, June precipitation, and elevation were the top three. June precipitation was the only weather variable included in the soybean model. Most of the final features in the soybean model were terrain attributes, with the remaining two being soil parameters: the Brooks-Corey pore size distribution index (λ) at 30-60 cm depth, and residual soil water content (θ_r) at 0-5 cm depth.

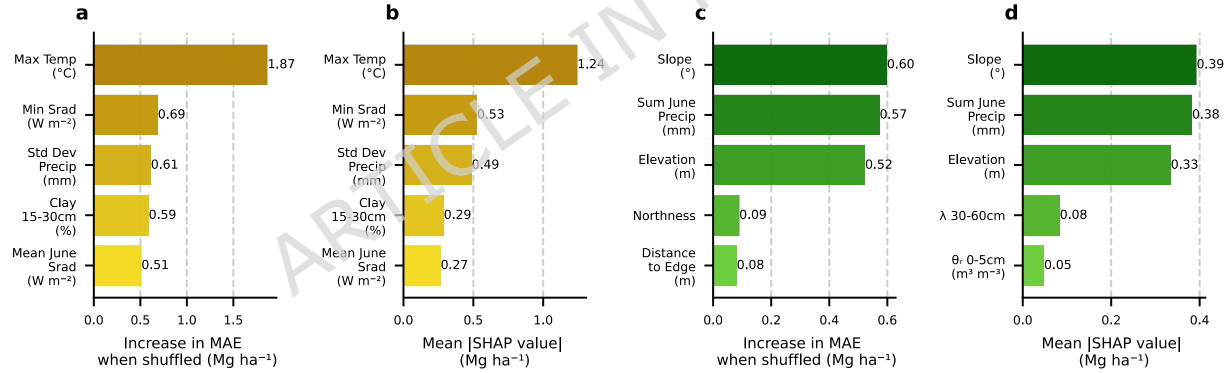

Figure 3. Feature importance of top variables in final maize (a-b) and soybean (c-d) models. Permutation feature importance (a, c) measures the decrease in overall model performance when the values of a given feature are randomly shuffled. SHAP importance (b, d) ranks features based on the mean absolute SHAP value (change in model output from the average model prediction when a feature is added to the model). Unless otherwise specified, weather variables are summarized with the listed statistic from daily observations April 1 – September 30. Elevation is meters

above sea level and Distance to Edge is measured from field boundaries. Northness refers to the degree of north-facing aspect, λ refers to the Brooks-Corey pore size distribution index (unitless), and θ_r is an estimate of residual soil water content (m^3 water per m^3 soil) [30].

Interpreting model results with SHAP

While feature importance gives important indications about the features that most affect model output, it does not clarify why those features are important or in what direction they shift model predictions. To understand this, we calculated SHAP values, which quantify the difference between a model's prediction for a specific observation versus the average model prediction across all observations. These individual differences are then additively combined, enabling global interpretation based on the aggregation of SHAP values (Figure 4). The SHAP value can be interpreted as the impact of that feature on the model prediction in units of the predicted variable, in this case crop yield (Mg ha^{-1}).

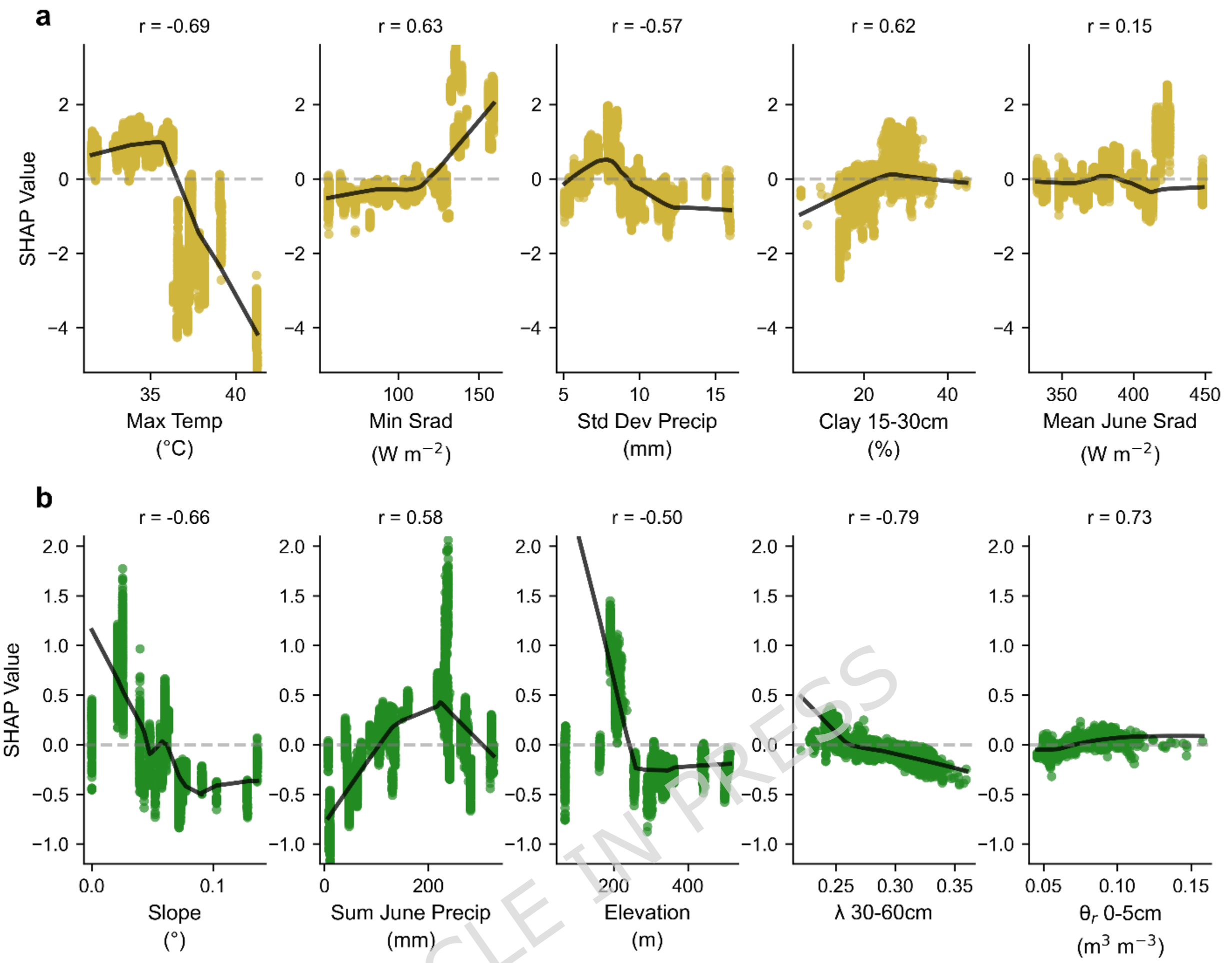

Figure 4. Dependence plots show the relationship between top variables and SHAP values for maize (a) and soybean (b) models. Points represent raw SHAP values. Pearson product-moment correlation coefficients between each feature and its SHAP values are shown above each plot. A smoothed trend line fitted using locally estimated scatterplot smoothing (LOESS) is also included in black.

From a dependence plot of SHAP values versus each feature, a response curve can be estimated from the model. Complex, non-linear patterns emerge from these plots, providing some indication of the correlation between environmental variables and yield. Results show higher maximum temperatures (both overall maximum daily temperature and mean maximum July temperature) were associated with lower maize yields, as was higher variability in precipitation (standard deviation of April 1 – September 30 daily precipitation). Increased solar radiation values were associated with increased maize yield predictions, especially in June. For soybeans, grid pixels with mean slope above 0.05° saw lower yield predictions, while lower elevation sites (especially below 200 m) tended to have higher yield predictions. We also saw that greater June precipitation was associated with higher soybean yields. Soil parameters λ and θ_r showed opposite effects on yield predictions. As λ at 30-60 cm increased past 0.25, yield predictions decreased, while increasing θ_r tended to increase yield predictions once values exceeded 0.07 m^3 water per m^3 soil or above.

Because SHAP values are consistent at both local (observation-level) and global (overall) scales, we can also aggregate them spatially and temporally to show how unique combinations of factors shape yield. For example, it was found that maximum temperature had a strong negative impact on predicted maize yield in years 2012 and 2018 when temperatures were above 37 °C across sites, even exceeding 40 °C in Illinois. SHAP analysis identified a possible threshold transition in maize yield predictions when maximum daily temperatures exceed 36–38 °C, beyond which yield predictions declined, a threshold that is supported by experimental studies on heat stress in maize[31]. This suggests that this method could be beneficial for detecting non-linear thresholds in environmental stressors affecting crop growth across landscapes. However, in most other years the effect of maximum temperature was positive. In soybeans, the impact of slope and June precipitation showed diverging direction and different magnitudes depending on the year (Figure 5).

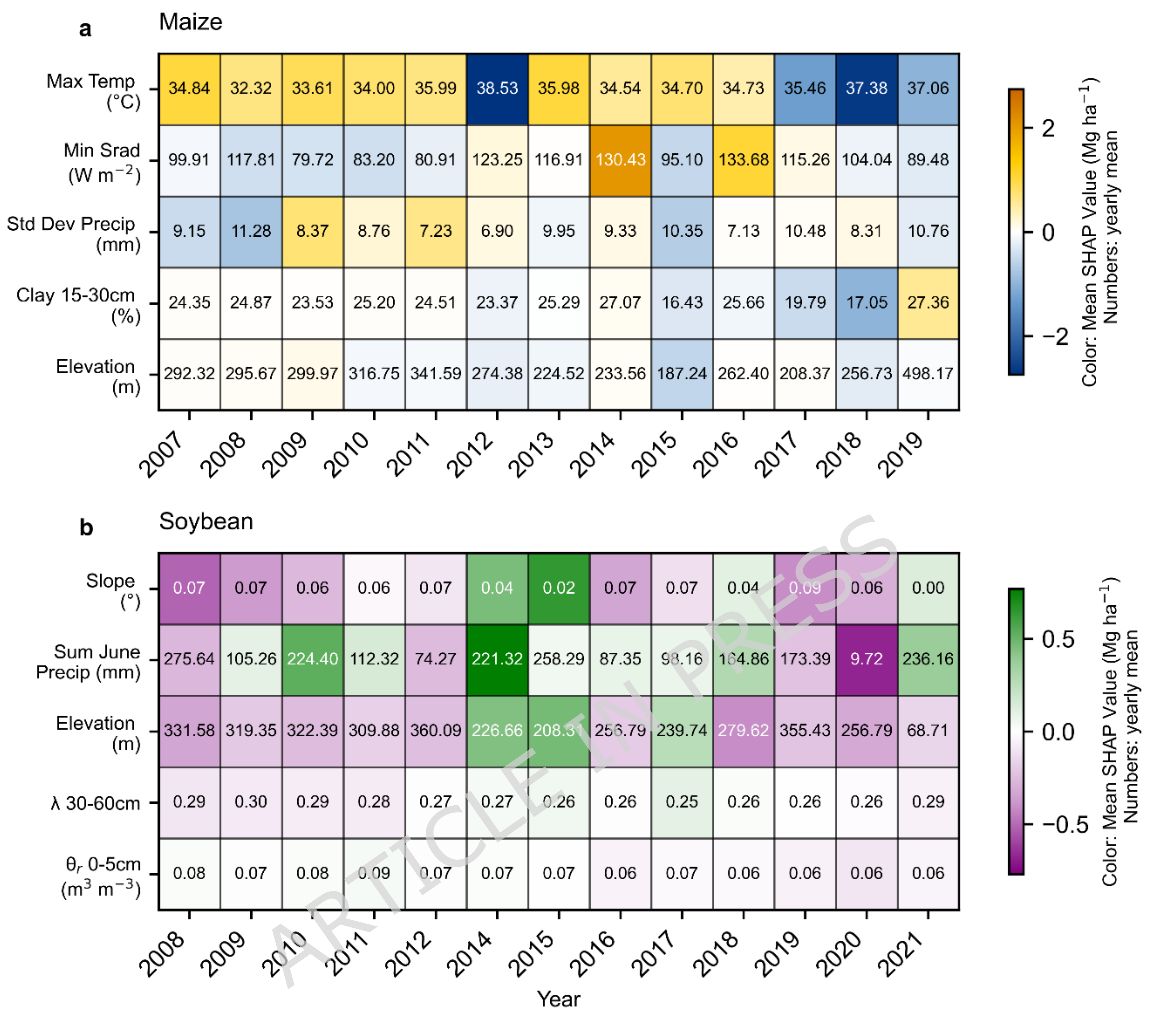

Figure 5. Mean effect for the top five variables in each model, grouped by year. Colors represent SHAP values, indicating how much that feature shifts the predicted yield value from the average model prediction. SHAP values have the same units as the outcome variable of the model (yield in Mg ha^{-1}). Numbers in each cell are the annual average of each environmental variable and are expressed in the variable's native units.

Discussion

Weather drivers of yield dominate in maize

These results demonstrate that pairing yield monitor data with public geospatial datasets enables ML models to predict crop yields with high accuracy while also generalizing well across geographic regions and growing seasons in the U.S. A key finding is that weather variables are the dominant drivers of yield, a result supported by previous work demonstrating weather conditions often explain the majority of spatiotemporal variability in maize yield [32,33]. Khaki & Wang (2019) found that environmental factors explained crop yield better than genotype alone, noting precipitation, solar radiation and temperature were most important for predicting maize yield variation[20]. Their results are consistent with our findings which demonstrate that very high maximum temperatures reduce maize yields, most likely via heat stress during sensitive growth stages. Similarly, Shahhosseini et al. (2020) found weather variables of precipitation, vapor pressure, temperature, and solar radiation were among the top predictors in an ensemble ML framework using county level corn yield in the U.S.[21]. Their findings also reinforce the notion that timing of heat and rainfall matter, consistent with our observation that extreme heat (maximum temperature) can harm yields during sensitive growth stages depending on crop maturity group[34]. Bhattarai et al. (2025) demonstrated environmental variability, especially maximum temperature, evapotranspiration, and precipitation, were highly influential of maize yield, underscoring maize’s sensitivity to water availability and heat stress[35]. In addition, Hoffman et al. (2020) observed threshold-like responses to high temperatures in maize, similar to what we found here[32]. These findings highlight the need to anticipate and adapt production (through breeding, irrigation, or other management) to potential environmental shocks, especially extreme heat and variable precipitation, that pose significant risks to maize yield stability.

Soil and topographic effects in soybean

Results from the soybean model point to the greater importance of terrain and soil attributes in predicting yield, though early summer precipitation was also of high importance. Given that the top terrain (slope, elevation) and soil (λ and θ_r) features strongly influence water movement and storage in the soil, it is likely that these attributes serve as a proxy for water availability or soil conditions that are conducive to soybean yields. An earlier precision agriculture study by Kaspar et al. (2003) found that spatial yield patterns are strongly related to terrain attributes like slope, elevation, and curvature[23]. We observed that soybean yield was similarly associated with slope and elevation (Figure 5). We observed negative associations between predicted soybean yields and the terrain properties of slope and elevation, consistent with findings from previous work that quantified the relationship between topographic properties and soybean yield[36]. However, we note the observed effect of elevation on soybean yield is complicated by the limited representation of high and low elevations within the same year. Given the clustered nature of elevation values, elevation may also serve as a partial proxy for location-specific effects, capturing other underlying factors. Dhillon et al. (2024) found that precipitation in the month of August was the most dominant weather variable in soybean yield, and Hoffman et al. (2020) note that yields increased with total growing-season precipitation up to an optimal point, beyond which additional rain gave diminishing or no returns, a finding consistent with the results of this study (Figure 4)[28,32]. These findings support our assertion that water availability is one of the most important factors for soybean yield prediction[28].

Integrating environmental drivers into crop breeding

The results presented here suggest important associations between weather, terrain, soil, and yield variability. By identifying the environmental factors that most strongly influence yield, our results could guide future crop breeding efforts to enhance yield in the face of weather variability. For example, SHAP-based analysis of U.S. maize yield revealed a sharp tipping point when maximum daily temperatures exceed about 36-38°C, aligning with field studies that show maize grain set and pollination are highly sensitive to heat stress in that range[31]. Such specific knowledge of key environmental yield drivers can inform selection environments and trait priorities for crop breeding. For maize, this has led to commercial drought-tolerant hybrids that outperform under rainfed stress while maintaining competitive yields in normal conditions, mirroring international efforts like CIMMYT’s heat and drought screening pipelines for tropical maize[37]. Soybean breeders are similarly targeting drought and waterlogging tolerance, using traits like deep rooting, stomatal control, and anaerobic stress resilience to improve performance in variable soil moisture environments [38]. These strategies are increasingly supported by interpretable models and G×E-informed genomic selection, which integrate environmental covariates and high-throughput phenotyping to enhance prediction and accelerate adaptation to environment-related yield constraints.

Limitations of this approach

Despite the strengths of this approach, some limitations must be addressed to improve its applicability and reliability. One major challenge is the presence of confounding factors, including differences in crop varieties, management practices, economic conditions (the ratio of output to input prices as maximizing yield does not relate to maximizing profitability), and policy environments. Additionally, multicollinearity among environmental predictors reduces interpretability, though this study attempted to mitigate this issue by applying robust feature selection techniques to refine the model's input variables. Another key challenge is scale mismatch—while the model is trained on nationwide data, translating these insights into locally relevant recommendations for field-level management remains a complex task [39].

While broad scale spatial patterns were well captured in our modeling approach, a small but significant positive autocorrelation was observed at 50 m, suggesting remaining fine-scale spatial dependence at within-field distances. This reflects sub-pixel heterogeneity in soil properties, micro-topography, or management practices that were not resolved in the environmental covariates used here. Although this dependence does not strongly influence model performance, it highlights the importance of scale alignment and indicates finer-resolution environmental, or management covariates could further reduce unexplained within-field spatial variability. Further research is needed to bridge this gap by integrating such site-specific data without losing the model's broader generalizability.

While this approach provides valuable predictive power, it remains correlative. To inform policy and management decisions more effectively, future efforts should incorporate causal inference methods to better distinguish the true drivers of yield from spurious correlations. The spatial and temporal representation of training data remains a significant limitation; expanding the dataset across more locations and growing seasons may improve model robustness. One promising avenue for addressing this issue is the incorporation of satellite-based yield estimates from time series of moderate to high spatial resolution sensors [18,40]. By continuously integrating remotely sensed yield data, future iterations of similar models could benefit from a larger and more temporally continuous dataset, enhancing their predictive accuracy and real-world applicability.

The increasing availability of yield monitor data presents a promising new frontier in precision agriculture, offering opportunities to enhance yield prediction and inform data-driven management decisions for cereal grains, legumes, oilseeds, and forage crops. However, while yield monitors may be effective in grain crops like maize and soybeans, their applicability is limited in many vegetable, tuber, fruit, and fiber crops where harvesting is done by hand or requires specialized machinery.

Future directions for ML and yield monitor

Models following a similar approach could enhance yield forecasting by identifying the most influential environmental factors over the growing season[21]. By determining the key drivers of yield, these models provide a strong foundation for early-season predictions, extending high predictive accuracy further into the early stages of crop development[21]. This advancement would enhance yield forecasts, allowing for more proactive decision-making in irrigation scheduling, fertilizer management, and crop insurance. Earlier and more precise predictions could help producers optimize resource allocation, reduce input waste, and mitigate financial risk, ultimately contributing to a more resilient and stable agricultural system[41]. By integrating these advances, ML could play an increasingly central role in optimizing agricultural productivity and resilience in the face of environmental stochasticity.

Conclusion

By integrating yield monitor data with public geospatial datasets and leveraging ML techniques, our approach demonstrates the potential to predict crop yields with high accuracy, even across varying geographic regions and growing seasons. Key environmental drivers of yield were identified, including temperature, solar radiation, and precipitation, enabling a deeper understanding of tipping points that can guide future crop breeding efforts and precision agriculture practices. These results can directly inform crop breeding strategies by highlighting thresholds in environmental tolerance between varieties and refining management practices to mitigate environmental shocks, providing valuable insights for improving agricultural productivity and resilience across changing environmental conditions.

Materials and Methods

Historical yield monitor data

Historical yield monitor data were provided by authors with the Data-Intensive Farm Management Program (DIFM), a National Resources Conservation Service (NRCS) initiative based at the University of Illinois – Urbana-Champaign [42]. These data represent typical, uniformly managed farm sites, with no active research trials. Historical yield was collected from 60 unique rainfed farms across the United States, ranging in size from 7.4 ha to 158.3 ha in size (46.3 ha field size on average). In total, the dataset includes observations from 52 maize fields and 25 soybean fields collected between 2007 to 2021, resulting in 134 unique site-year combinations. States represented in this dataset include Illinois, South Dakota, Iowa, Oklahoma, Nebraska, Ohio, Pennsylvania, Arkansas, and Indiana. With observations from 1-6 years per farm, the dataset reflects 134 unique field-year-crop observations. Historical yield data were cleaned and preprocessed prior to analysis. Raw maize and soybean yield values were first converted to dry yield at standard moisture (15% moisture). Global outlier removal included the following: any yield values of zero were removed, and values greater than three standard deviations from the mean for each field were also removed [43]. Values within 30 m of field edges were removed to reduce bias from edge effects. Spatial outliers, which are points that are significantly different from surrounding values, were identified using spatial autocorrelation based on Local Moran's I, and observations with negative values were removed from the dataset [43]. A 30 x 30 m grid cell was overlayed, and all removed data points (except for field edges) were imputed using inverse distance weighted interpolation based on values within each cell. Finally, observations from each cell were averaged to calculate yield in a grid with 30 x 30 m resolution.

Environmental Data

Soil data were derived from POLARIS, which has a spatial resolution of 30 m. The POLARIS soils dataset provides probabilistic estimates of soil properties at varying depths, derived from the USDA Soil Survey Geographic Database [30]. Weather data were extracted from Daymet version 4 [44]. Daymet provides daily meteorological data, including temperature, precipitation, and incident shortwave radiation variables. These data were processed to calculate summary statistics for an approximation of each year's growing season (April 1–September 30), since planting and harvesting dates were not always available. Summary statistics for each month were also included as covariates. All weather statistics used in the models are included in Supplemental Table 2.

Terrain data were derived from the U.S. Geological Survey (USGS) 3D Elevation Program (3DEP), a lidar-derived elevation model with a native resolution of 1 m [45]. These data were processed using the “tagee” package for terrain analysis in Google Earth Engine. From tagee, the following terrain parameters were calculated: slope, aspect, hillshade, northness, eastness, horizontal curvature, vertical curvature, mean curvature, minimal curvature, maximal curvature, gaussian curvature, and shape index. Detailed descriptions of these metrics are provided by Safanelli et al. (2020). In addition to these metrics, a topographic wetness index was also calculated using the 15 arc-second HydroSHEDS flow accumulation dataset from the World Wildlife Fund and the USGS watershed boundary dataset of basins dataset at the basin hydrologic unit level[47] . Terrain attributes were included because they represent stable, within-field sources of hydrological and pedological variation that can be reliably mapped at 30 m resolution, and have been shown to influence yield through runoff, erosion, and topsoil distribution[23,46]. After calculating terrain metrics, all environmental data were resampled to a 30 m resolution pixel and aligned with the soil and yield data. All environmental data were processed in Google Earth Engine [48].

We note that some variables known to influence yield, such as evapotranspiration (ET), soil nutrients (N, P, K, SOM), and field-level management practices, were not included. These variables were excluded because they were not consistently available across the multi-state dataset and would introduce potential confounding effects. Additionally, ET would potentially introduce redundancy and collinearity and could obscure direct relationships between fundamental environmental drivers and yield, which are the focus of this interpretable ML analysis. We note these omissions as an important limitation and interpret results accordingly.

Model selection and training

For this study, our initial dataset included 128 environmental features. A brief description, units, and data source is provided for each of the initial features in Supplemental Table 2. To reduce dimensionality and collinearity, any features that had low variance (less than 0.1 after scaling features between 0 and 1) or that were highly correlated (r > 0.9) were removed from the dataset, resulting in 80 features remaining for initial model training. The dataset was then split using a random 70-30 training-testing split.

Several automated ML platforms have recently emerged, enabling systematic evaluation of candidate ML models. Here we used the AutoGluon package in Python for automated ML training, enabling a comparison of ML models in terms of predictive accuracy and error rates [25]. All models were optimized to reduce the root mean square error (RMSE). The top-performing model from the first round of training was selected based on the lowest RMSE. A recursive feature elimination with five-fold cross-validation (RFECV) was then implemented to eliminate additional features from the top performing model and to reduce the feature space [49]. The optimal number of features for each model was automatically selected during the RFECV process. Reduced, optimal-features datasets (20 features for maize, 13 for soybeans) were used in a second round of model training in AutoGluon. Hyperparameter tuning was then performed on the top-performing model using a random search space, and the final trained model was saved for analysis. All figures and results reported in this manuscript were derived from these crop-specific final models.

Model validation and interpretation

Validation was conducted by withholding 30% of pixel values from the dataset for testing purposes. Performance was evaluated for those unseen observations after completing model training, feature elimination, and hyperparameter tuning. Permutation feature importance and SHAP, both model-agnostic methods, were used to interpret model predictions and identify features of high relevance to model outputs.

Permutation importance quantifies how much each feature contributes to overall model performance by measuring the increase in RMSE after shuffling its values; larger performance degradation indicates greater importance. SHAP values, conversely, measure the average magnitude of a features' effect on model predictions and are expressed in the units of the outcome variable (Mg ha^{-1}) [50]. Features with larger mean absolute SHAP values therefore exert greater influence on predicted yield. Because SHAP has a rigorous theoretical foundation and can provide explanations that are both locally and globally consistent, SHAP has become one of the most widely used methods for ML interpretation and is considered by many to be the current standard for interpretable ML [51]. Including both metrics enables evaluation of feature contributions both in terms of predictive performance (permutation importance) and in terms of direction, magnitude, and non-linear structure (SHAP).

High overall performance may obscure poor performance in particular sites or years, reflecting the potential for overfitting in ML models [52]. To address this, we grouped testing data by growing year and site to perform an additional group-wise validation. This approach assesses how models perform across subsets of the testing dataset, indicating both temporal and spatial generalizability. We also performed an analysis of spatial autocorrelation of model residuals to test how well our model captures spatial patterns in the data. To accomplish this, we calculated multi-scale Moran's I statistics at distance intervals of 50m, 500m and 50km, to evaluate within-field, between-field, and between-region spatial clustering of residuals, respectively.

References

1. USDA ERS. *Farming and Farm Income: U.S. Farm Sector Cash Receipts*. https://www.ers.usda.gov/data-products/farm-income-and-wealth-statistics (2023).
2. USDA NASS. *Crop Production Annual Summary, 2023*. https://usda.library.cornell.edu/concern/publications/k3569432s (2024).
3. USDA FAS. *Global Agricultural Trade System (GATS)*. https://apps.fas.usda.gov/gats (2024).
4. Godfray, H. C. J. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327**, 812–818 (2010).
5. Tilman, D., Balzer, C., Hill, J. & Befort, B. L. Global food demand and the sustainable intensification of agriculture. *PNAS* **108**, 20260–20264 (2011).

6. Persson, U. M. The impact of biofuel demand on agricultural commodity prices: a systematic review. in *Advances in Bioenergy* 465–482 (John Wiley & Sons, Ltd, 2016).

7. Lauer, J. G. *et al.* The scientific grand challenges of the 21st century for the Crop Science Society of America. *Crop Sci.* **52**, 1003–1010 (2012).

8. Reddy, B. V. S., Sanjana Reddy, P., Bidinger, F. & Blümmel, M. Crop management factors influencing yield and quality of crop residues. *Field Crops Res.* **84**, 57–77 (2003).

9. Peng, B. *et al.* Towards a multiscale crop modelling framework for climate change adaptation assessment. *Nat. Plants* **6**, 338–348 (2020).

10. Andorf, C. *et al.* Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* **132**, 817–849 (2019).

11. Boehm Jr., J. D. *et al.* Genetic improvement of US soybean in maturity groups V, VI, and VII. *Crop Sci.* **59**, 1838–1852 (2019).

12. Prasanna, B. M. Diversity in global maize germplasm: characterization and utilization. *J. Biosci.* **37**, 843–855 (2012).

13. Xavier, A., Thapa, R., Muir, W. M. & Rainey, K. M. Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genet. Resour.* **16**, 513–523 (2018).

14. Heinemann, J. A., Massaro ,Melanie, Coray ,Dorien S., Agapito-Tenfen ,Sarah Zanon & and Wen, J. D. Sustainability and innovation in staple crop production in the US Midwest. *Int. J. Agric. Sustainability* **12**, 71–88 (2014).

15. Egli, D. B. Comparison of corn and soybean yields in the United States: historical trends and future prospects. *Agron. J.* **100**, S-79-S-88 (2008).

16. Yost, M. A. *et al.* A long-term precision agriculture system sustains grain profitability. *Precis. Agric.* **20**, 1177–1198 (2019).

17. Gage, J. L. *et al.* The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun.* **8**, 1348 (2017).

18. Kang, Y. & Özdoğan, M. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sens. Environ.* **228**, 144–163 (2019).

19. Lobell, D. B. & Burke, M. B. On the use of statistical models to predict crop yield responses to climate change. *Agric. For. Meteorol.* **150**, 1443–1452 (2010).

20. Khaki, S. & Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **10**, (2019).

21. Shahhosseini, M., Hu, G. & Archontoulis, S. V. Forecasting corn yield with machine learning ensembles. *Front. Plant Sci.* **11**, (2020).

22. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. (2025).

23. Kaspar, T. C. *et al.* Relationship between six years of corn yields and terrain attributes. *Precis. Agric.* **4**, 87–101 (2003).

24. Jagadamma, S., Lal, R., Hoeft, R. G., Nafziger, E. D. & Adee, E. A. Nitrogen fertilization and cropping system impacts on soil properties and their

relationship to crop yield in the central Corn Belt, USA. *Soil Tillage Res.* **98**, 120–129 (2008).

25. Erickson, N. *et al.* AutoGluon-Tabular: robust and accurate AutoML for structured data. Preprint at https://doi.org/10.48550/arXiv.2003.06505 (2020).

26. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).

27. Ke, G. *et al.* LightGBM: a highly efficient gradient boosting decision tree. in *Advances in Neural Information Processing Systems* vol. 30 (2017).

28. Dhillon, R., Takoo, G., Sharma, V. & Nagle, M. Utilizing machine learning framework to evaluate the effect of climate change on maize and soybean yield. *Comput. Electron. Agric.* **221**, 108982 (2024).

29. Chang, Y., Latham, J., Licht, M. & Wang, L. A data-driven crop model for maize yield prediction. *Commun. Biol.* **6**, 1–9 (2023).

30. Chaney, N. W. *et al.* POLARIS soil properties: 30-m probabilistic maps of soil properties over the contiguous United States. *Water Resour. Res.* **55**, 2916–2938 (2019).

31. Lv, X. *et al.* Heat stress and sexual reproduction in maize: unveiling the most pivotal factors and the greatest opportunities. *J. Exp. Bot.* **75**, 4219–4243 (2024).

32. L Hoffman, A., R Kemanian, A. & E Forest, C. The response of maize, sorghum, and soybean yield to growing-phase climate revealed with machine learning. *Environ. Res. Lett.* **15**, 094013 (2020).

33. Ray, D. K., Gerber, J. S., MacDonald, G. K. & West, P. C. Climate variation explains a third of global crop yield variability. *Nat. Commun.* **6**, 5989 (2015).

34. Ashworth, A. J., Allen, F. L. & Saxton, A. M. Using partial least squares and regression to interpret temperature and precipitation effects on maize and soybean genetic variance expression. *Agronomy* **13**, 2752 (2023).

35. Bhattarai, B., Leasor, Z. & Reis, A. F. D. B. Incorporating soil moisture data into a machine learning framework improved the predictive accuracy of corn yields in the U.S. *Agric. Water Manage.* **319**, 109762 (2025).

36. Kravchenko, A. N. & Bullock, D. G. Correlation of corn and soybean grain yield with topography and soil properties. *Agron. J.* **92**, 75–83 (2000).

37. Cairns, J. E. *et al.* Identification of drought, heat, and combined drought and heat tolerant donors in maize. *Crop Sci.* **53**, 1335–1346 (2013).

38. Valliyodan, B. *et al.* Genetic diversity and genomic strategies for improving drought and waterlogging tolerance in soybeans. *J. Exp. Bot.* **68**, 1835–1849 (2017).

39. Safi, A. R., Karimi, P., Mul, M., Chukalla, A. & de Fraiture, C. Translating open-source remote sensing data to crop water productivity improvement actions. *Agric. Water Manage.* **261**, 107373 (2022).

40. Jin, Z. *et al.* Smallholder maize area and yield mapping at national scales with Google Earth Engine. *Remote Sens. Environ.* **228**, 115–128 (2019).

41. Celis, J., Xiao, X., Wagle, P., Adler, P. R. & White, P. A review of yield forecasting techniques and their impact on sustainable agriculture. in

Transformation towards circular food systems: sustainable, smart and resilient citrus supply chains in mediterranean areas 139–168 (Springer Nature, 2024).

42. Bullock, D. S. *et al.* The data-intensive farm management project: changing agronomic research through on-farm precision experimentation. *Agron. J.* **111**, 2736–2746 (2019).

43. Vega, A., Córdoba, M., Castro-Franco, M. & Balzarini, M. Protocol for automating error removal from yield maps. *Precis. Agric.* **20**, 1030–1044 (2019).

44. Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E. & Kao, S.-C. Daymet: daily surface weather data on a 1-km grid for North America, version 4 R1. ORNL Distributed Active Archive Center https://doi.org/10.3334/ORNLDAAC/2129 (2022).

45. U.S. Geological Survey. 3D elevation program 1-meter resolution digital elevation model. https://www.usgs.gov/the-national-map-data-delivery (2019).

46. Safanelli, J. L. *et al.* Terrain analysis in Google Earth Engine: a method adapted for high-performance global-scale analysis. *ISPRS Int. J. Geo-Inf.* **9**, 400 (2020).

47. Lehner, B., Verdin, K. & Jarvis, A. New global hydrography derived from spaceborne elevation data. *Eos, Transactions American Geophysical Union* **89**, 93–94 (2008).

48. Gorelick, N. *et al.* Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).

49. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, (2011).

50. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* vol. 30 (2017).

51. Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D. & Groh, G. SHAP-based explanation methods: a review for NLP interpretability. *Proceedings of the 29th International Conference on Computational Linguistics* 4593–4603 (2022).

52. Ying, X. An overview of overfitting and its solutions. *J. Phys. Conf. Ser.* **1168**, 022022 (2019).

Acknowledgements

The authors would like to acknowledge the data contributions of farmers from the Data Intensive Farm Management project, without which this project would not have been possible. The U.S. Department of Agriculture (USDA) prohibits discrimination in all its programs and activities on the basis of race, color, national origin, age, disability, and where applicable, sex, marital status, familial status, parental status, religion, political beliefs, reprisal, or because all or part of an individual's income is derived from any public assistance. (Not all prohibited bases apply to all programs.) Persons with disabilities who require alternative means for communication of

program information (Braille, large print, audiotape, etc.) should contact USDA's TARGET Center at 202-720-2600 (voice and TDD).

Funding

This research was supported by the intramural research program of the U.S. Department of Agriculture, National Institute of Food and Agriculture, Agriculture and Food Research Initiative Predoctoral Fellowship Program, Award Number 2023-67011-40361. This research was also funded in part by USDA-NRCS On-Farm Trials Conservation Innovation Grant, “Improving the Economic and Ecological Sustainability of US Crop Production through On-Farm Precision Experimentation,” award number NR213A7500013G021, and by USDA-NIFA Hatch Project 470-362.

Author Contributions

HWS wrote the main manuscript text. HWS and CJH wrote software used in the work and conducted analyses. Data acquisition was conducted by DSB, AJA and PRO. HWS, AJA, LLN, PRO, DSB, and JT contributed to project conception. All authors contributed to data interpretation. All authors reviewed and approved the final manuscript.

Data Availability

Yield monitor data has been kept private at the request of farmer participants. All other data used in this study are available on Google Earth Engine (https://developers.google.com/earth-engine/datasets) or through the Google Earth

Engine Community Catalogue (https://gee-community-catalog.org/). Please contact Harrison Smith at hws001@uark.edu to request data from this study.

Code availability

Code, documentation, and metadata are available from the corresponding author's GitHub repository: https://github.com/harrisonwsmith/harvesting_insights, or contact Harrison Smith at hws001@uark.edu to request the code used in this study.