International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 2, 164-182, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I2P124

Original Article

Decision Tree Ensemble Approach for Crop Yield Prediction

Catherine Ngo[1], Orson Chi[2]
[1]Department of Agriculture, Food and Resource Sciences, University of Maryland Eastern Shore, Princess Anne, Maryland 21853, USA.
[2]Department of Computer Science and Engineering Technology, University of Maryland Eastern Shore, Princess Anne, Maryland 21853, USA.

Received On: 18/03/2026 ***Revised On: 12/04/2026*** ***Accepted On: 19/04/2026*** ***Published On: 30/04/2026***

Abstract *- This study presents a comprehensive analysis of corn and soybean crop yield determinants on the Maryland Eastern Shore spanning ten growing seasons (2015–2024). A dataset of 2,520 field-season observations, 1,260 corn and 1,260 soybean, across nine counties, was subjected to rigorous descriptive statistical characterization followed by machine learning modeling using a two-tier architecture: a combined Gradient Boosting (GB) model for mixed-crop prediction and interpretation, and crop-specific Random Forest (RF) deployment models for single-crop inference. Corn yields averaged 150.2 bu/ac (SD = 15.1; CV = 10%) and soybean yields averaged 44.0 bu/ac (SD = 8.6; CV = 20%), with soybean exhibiting substantially greater relative variability. The GB combined model achieved near-perfect cross-validated performance (R^2 = 0.9801, MAE = 5.70 bu/ac, RMSE = 7.68 bu/ac), while crop-specific RF models yielded R^2 of 0.6835 (corn) and 0.7528 (soybean). Feature attribution analysis identified seeding rate, crop maturity rating, and nitrogen application rate as the dominant agronomic predictors, while June–August precipitation emerged as the primary weather driver, particularly for soybean, where summer moisture governs pod-fill. Partial dependence analysis revealed nonlinear agronomic response curves consistent with established agronomic principles. Together, these findings provide an actionable, data-driven framework for precision crop management in the Mid-Atlantic coastal plain.*

Keywords *- Crop Yield Prediction; Corn; Soybean; Gradient Boosting; Random Forest; Machine Learning; Feature Importance; Precision Agriculture; Maryland Eastern Shore.*

1. Introduction

Agricultural productivity in the Mid-Atlantic region is subject to a complex interplay of management decisions, soil heterogeneity, and increasingly variable weather patterns. The Maryland Eastern Shore, a nine-county coastal plain region encompassing Cecil, Kent, Queen Anne's, Caroline, Talbot, Dorchester, Wicomico, Somerset, and Worcester counties, represents one of the most agriculturally significant zones in the Mid-Atlantic, characterized by diverse soil textures ranging from loamy sands to silt loams, flat to gently sloping topography, and a humid subtropical climate with substantial inter-annual precipitation variability.

The dual-crop system of corn and soybean dominates the regional landscape, with both crops responding distinctly to agronomic management and climatic perturbations. Corn exhibits high water-use efficiency and responds strongly to nitrogen fertilization and growing degree-day accumulation. Soybean, as a nitrogen-fixing legume, is more sensitive to summer precipitation during reproductive growth stages and demonstrates greater yield plasticity through branching and pod abortion mechanisms.

Predicting crops yields with sufficient precision to support on-farm decision-making has been a long-standing challenge in agricultural research. Classical regression approaches have been applied to crop yield modeling [1][2] but their inability to capture complex nonlinear interactions has motivated adoption of ensemble machine learning methods. Gradient Boosting [3] and Random Forest [4] algorithms have demonstrated strong predictive performance in agronomic contexts [5] [6], offering additional advantages in feature interpretability through permutation importance and SHAP (SHapley Additive exPlanations) analysis [7].

Despite these methodological advances, region-specific studies integrating long-term multi-county observational data with rigorous machine learning pipelines remain limited for the Chesapeake Bay agricultural watershed. This article addresses that gap by presenting (i) a comprehensive descriptive statistical characterization of crop yield and associated agronomic, soil, and weather variables across a 10-year period; and (ii) a two-tier machine learning modeling framework that achieves high predictive accuracy while preserving interpretability through structured feature attribution.

The objectives of this study are: (i) to characterize the statistical distributions, temporal trends, geographic variation, and correlation structure of yield and predictor variables for corn and soybean on the Maryland Eastern Shore; (ii) to develop and evaluate GB and RF predictive models using field-stratified cross-validation; (iii) to quantify the relative importance of agronomic, soil, and weather predictors via permutation-based attribution and partial dependence

analysis; and (iv) to provide practical deployment guidance for precision yield forecasting applications.

2. Literature Review

Crop yield prediction has been a central challenge in agricultural science for more than a century, evolving from empirical field observations to mechanistic process-based simulation models and, more recently, to data-driven machine learning frameworks. While process-based models offer mechanistic interpretability and can simulate yield under novel environmental conditions, they require extensive parameterization for each genotype–environment combination and are computationally intensive when applied at regional scale across heterogeneous soils and management systems [8].

Statistical regression models, including ordinary least squares, panel regression with fixed effects, and generalized additive models, have therefore been widely adopted as computationally efficient alternatives for regional yield analysis [1][2][9]. Schlenker and Roberts [2] demonstrated that nonlinear temperature–yield relationships, particularly a sharp decline in U.S. corn and soybean yields above a threshold of approximately 29°C for corn and 30°C for soybean, are systematically underestimated by linear models, a finding with major implications for climate change impact assessment. Lobell et al. [1] further showed that historical global crop production trends from 1980 to 2008 were significantly associated with observed temperature increases, with warming explaining a meaningful share of the gap between actual and potential yields in major producing regions.

Despite these advances, classical regression frameworks face fundamental limitations in handling the high-dimensional, multicollinear, and interaction-rich feature spaces characteristic of field-level agronomic datasets [10][11]. The integration of management inputs (seeding rate, fertilizer application, tillage, irrigation), soil physical and chemical properties, phenological variables, and weather metrics across multiple crops and geographies creates predictor spaces where conventional linear assumptions are routinely violated. These limitations have catalyzed a transition toward ensemble machine learning methods capable of capturing complex nonlinear relationships without requiring explicit model specification.

The application of machine learning to crop yield prediction has expanded dramatically over the past two decades. Van Klompenburg et al. [6] conducted a systematic review of 50 peer-reviewed studies published between 2008 and 2020, identifying Random Forest (RF), Artificial Neural Networks (ANN), and Support Vector Machines (SVM) as the three most frequently applied algorithms for crop yield prediction. RF consistently outperformed classical regression benchmarks in predictive accuracy across a diversity of crops, geographies, and feature sets, largely due to its ability to model feature interactions and reduce overfitting through bootstrap aggregation (Bagging) of decision tree ensembles [4].

Gradient Boosting (GB) methods, including XGBoost [12], LightGBM [13], and scikit-learn's GradientBoostingRegressor [3], have increasingly supplanted RF as the benchmark ensemble method in competitive prediction tasks. Unlike RF's parallel construction of independent trees, GB builds trees sequentially, with each new tree fitting the residuals of the prior ensemble. This additive correction mechanism generally produces lower bias at comparable variance, particularly for datasets with structured nonlinear interactions [12]. Shahhosseini et al. [5] evaluated multiple machine learning algorithms for corn yield prediction across the U.S. Corn Belt and found that gradient boosting ensembles consistently outperformed RF, deep learning, and regression baselines, achieving R^2 values above 0.80 in cross-validated field-level prediction tasks.

Deep learning approaches, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNN), have also been applied to crop yield prediction using sequential weather or remote sensing inputs [14]. While these architectures can capture temporal dependencies in growing-season data, they typically require substantially larger sample sizes for stable parameter estimation than are available in most field-scale agronomic datasets, and their black-box character limits agronomic interpretability. For datasets in the range of 1,000–10,000 observations with mixed tabular features, tree-based ensemble methods have consistently demonstrated superior or equivalent predictive performance with substantially greater interpretability [15].

A persistent challenge in applying machine learning to agronomic problems is the tension between predictive accuracy and interpretability. The "black box" character of ensemble methods, while acceptable for automated prediction tasks, limits their utility for agronomic decision support where farmers and consultants require mechanistic understanding of why a model produces a given forecast [16]. This challenge has motivated the development of post-hoc interpretability methods, most notably SHAP (SHapley Additive exPlanations) [7], which decomposes each model prediction into additive contributions from individual features based on cooperative game theory.

Lundberg et al. [17] extended the SHAP framework to tree ensembles via TreeSHAP, an algorithm that computes exact Shapley values in polynomial time for tree-based models, enabling both global feature importance ranking and local (per-prediction) explanation. SHAP analysis has been applied to corn yield prediction by Shahhosseini et al. [5], who found that solar radiation, temperature metrics, and precipitation were the dominant positive contributors to yield in the U.S. Corn Belt, with SHAP interaction plots revealing important synergistic effects between temperature and water supply. Khaki et al. [18] applied SHAP to soybean yield prediction across the U.S. and Canada, identifying planting date and accumulated growing degree days as the most influential agronomic drivers, consistent with the findings of the present study's correlation analysis.

Permutation importance, first introduced by Breiman [4] for Random Forest, provides an alternative interpretability measure that is model-agnostic and directly quantifies the decrease in predictive accuracy caused by randomly permuting a feature's values. Fisher et al. [19] demonstrated that permutation importance provides more reliable feature rankings than impurity-based importance in settings where features differ substantially in cardinality or scale, because impurity-based methods systematically overweight high-cardinality continuous features. Partial Dependence Plots (PDPs) [3] complement importance measures by visualizing the marginal functional form of each predictor's effect on predicted yield, averaged over the joint distribution of all other features—a visualization that bridges the gap between statistical models and agronomic response curve concepts familiar to practitioners.

A substantial body of agronomic research has characterized the key management and environmental determinants of corn and soybean yield in the eastern United States. Nitrogen management is widely recognized as the primary agronomic lever for corn productivity [20]. The nitrogen use efficiency (NUE) of corn in the Mid-Atlantic region typically ranges from 30% to 60% under conventional management, with the economic optimum N rate for Maryland corn fields estimated at approximately 140–175 lb N/ac depending on soil type, previous crop, and yield environment [21]. Excessive N application above the economic optimum not only fails to increase yield but contributes to nitrate leaching into the Chesapeake Bay watershed—a regional water quality concern of significant regulatory and environmental importance [22].

Seeding rate optimization has been extensively studied for both corn and soybean. For corn, the yield response to plant population follows a dome-shaped curve, with optimal populations in the Mid-Atlantic typically ranging from 28,000 to 34,000 seeds/ac depending on hybrid, soil type, and irrigation status [23]. Soybean exhibits greater compensatory capacity through branch and pod number adjustment, but research in the Mid-Atlantic has consistently shown that populations below 100,000 seeds/ac are associated with yield loss, while populations above 160,000–180,000 seeds/ac provide diminishing returns [24]. Hybrid and variety maturity selection is a critically important management decision for both crops on the Eastern Shore, as the length of the frost-free period constrains the maximum maturity that can be fully expressed, while under-maturity leaves yield potential unrealized [25].

Planting date is a critical determinant of soybean yield, as late planting shortens the period available for canopy development and reproductive growth. Bastidas et al. [26] documented yield penalties for soybean in the U.S. Midwest of approximately 0.5 bu/ac per day of delayed planting beyond the optimal window, with losses accelerating substantially for plantings after mid-June. Similar or greater penalties have been documented for Mid-Atlantic soybean systems, where the photoperiod-sensitive reproductive transition is closely tied to summer solstice timing [27]. For corn, planting date effects are more modest in magnitude but statistically significant, with each day of delay beyond the regional optimal window (typically late April to early May in Maryland) associated with approximately 1–2 bu/ac yield loss [28].

Weather variability is the primary source of inter-annual yield instability in rain-fed and supplementally irrigated agricultural systems in the Mid-Atlantic region. Precipitation during the June–August (JJA) period is of particular importance for both corn and soybean, as this window encompasses the critical periods of corn pollination and grain fill (VT–R6) and soybean reproductive development (R1–R6) [29]. Boyer et al. [30] estimated that water deficit during critical growth stages accounts for approximately 66–71% of all yield losses in U.S. corn and soybean production, underscoring the primacy of water availability among all yield-limiting factors.

Growing degree day (GDD) accumulation governs the rate of crop development and is a primary determinant of growing season length and yield potential. The base-50°F GDD accumulation framework is standard for corn in the eastern United States, with typical season-long GDD requirements of 2,100–3,000 GDD depending on hybrid relative maturity [31]. Soybean development is less directly governed by GDD and more strongly regulated by photoperiod and temperature interactions, though GDD above a base temperature of approximately 50°F remains a useful predictor of seasonal heat accumulation and development rate [32].

Extreme heat events, days exceeding 95°F (35°C), have been shown to impose disproportionately large yield penalties through pollen viability loss, silk desiccation, and kernel abortion in corn [2] [33]. Climate projections for the Chesapeake Bay watershed indicate that both mean summer temperatures and the frequency of extreme heat days are expected to increase substantially by mid-century under Representative Concentration Pathways 4.5 and 8.5 [34], increasing the climate risk to corn and soybean production on the Maryland Eastern Shore relative to the historical 2015–2024 baseline documented in this study.

Soil physical and chemical properties constitute a foundational layer of yield variability that interacts with both management inputs and weather. On the Maryland Eastern Shore, soil texture ranges from coarse loamy sands and sandy loams in the central and southern counties to finer silt loams in the north and west, with important implications for plant-available water capacity (AWC), nutrient retention, and erosion risk [35]. Lighter-textured soils with lower AWC are more vulnerable to moisture stress during dry JJA periods, amplifying the precipitation–yield relationship documented in this and other regional studies [36].

Soil organic matter (SOM) is a key integrative indicator of soil health, influencing AWC, cation exchange capacity, nitrogen cycling, and aggregate stability [37]. While the direct correlation between SOM and crop yield is often modest in

observational studies due to confounding by management history, SOM is associated with greater resilience under moisture stress and reduced fertilizer requirements over the long term [38]. Soil pH influences nutrient availability, particularly phosphorus and micronutrients, with the optimal pH range for corn and soybean in the Mid-Atlantic typically cited as 6.0–6.8 [21]. The modest observed correlation between pH and yield in the present dataset ($|r| \leq 0.04$) suggests that most fields are managed within acceptable pH ranges, consistent with the prevalence of routine lime applications in the region.

The reviewed literature reveals several important gaps that the present study is positioned to address. First, while machine learning yield prediction studies are abundant for the U.S. Corn Belt [5] [6], analogous region-specific studies for the Mid-Atlantic coastal plain where distinct soil textures, drainage patterns, and climate conditions create a unique agroecological context are largely absent. The Maryland Eastern Shore and the broader Chesapeake Bay watershed have been the subject of water quality modeling [39] [40] and nitrogen management research [21], but integrated multi-county yield prediction frameworks coupling management, soil, and weather data over a multi-year panel have not been published for this region to the authors' knowledge.

Second, most published machine learning yield studies model a single crop in isolation, potentially missing the interpretive value of comparing feature importance hierarchies across crops grown in the same fields and weather environments. A combined modeling approach that treats crop type as a feature, enabling cross-crop feature attribution while maintaining within-crop predictive validity, has been explored only rarely in the literature [41] and has not been applied in a structured two-tier deployment framework. Third, most published studies rely solely on built-in impurity-based feature importance, which is known to be biased toward high-cardinality features [19], without complementing it with permutation-based attribution or SHAP-equivalent analysis. The present study addresses all three gaps through its panel dataset design, two-tier model architecture, and multi-method interpretability analysis.

3. Materials and Methods

This study developed a fully synthetic dataset for corn and soybean production across Maryland's nine Eastern Shore counties, encompassing Cecil, Kent, Queen Anne's, Caroline, Talbot, Dorchester, Wicomico, Somerset, and Worcester, from 2015–2024 using Microsoft Copilot. The field-level dataset (field_seasons) comprises one record per field–year and includes county, farm, and field identifiers; crop system (corn or soybean, with soybeans designated as full-season or double-crop); planting and harvest dates consistent with regional practice; and management variables such as seeding rate, maturity class, fertilizer applications (N, P_2O_5, K_2O), tillage, irrigation, and cover cropping. Soil properties were assigned from distributions reflecting dominant regional series (e.g., Sassafras, Mattapex), textures, and realistic ranges of pH, organic matter, drainage class, and slope. Seasonal weather metrics, including GDD (base 50 °F), modified corn GDD with 86/50 caps, cumulative precipitation, summer precipitation (June–August), and days ≥ 95 °F, were derived by linking fields to a county-level daily weather dataset. Yields (corn at 15.5% moisture; soybean at 13%) were synthesized using crop-specific response functions calibrated to recent Maryland NASS ranges (corn ≈ 140–165 bu/ac; soybean ≈ 40–47 bu/ac) and incorporate effects of planting date, heat stress, rainfall timing, nutrient sufficiency, irrigation, and double-cropping.

The companion daily_weather dataset provides county-day observations of maximum and minimum temperature, precipitation, solar radiation, and thermal indices (gdd50 and mgdd86_50). Weather time series were generated from seasonal temperature and precipitation curves with stochastic variability to emulate the region's humid subtropical climate (approximately 46 inches of annual precipitation and July averages near 78 °F) and realistic interannual variation; the 2024 season was explicitly simulated as warmer and drier to reflect observed statewide anomalies. Daily weather variables were aggregated over each field's planting-to-harvest window and used to drive the synthetic yield responses. Collectively, these datasets reproduce credible interactions among weather, soils, management, and crop performance while remaining fully synthetic and suitable for research, modeling, and instructional use without confidentiality risk.

The modeling strategy employed a two-tier architecture designed to balance interpretability, predictive accuracy, and deployment flexibility as follows:

Tier 1 - Combined Gradient Boosting (GB) Model: A scikit-learn GradientBoostingRegressor was trained on the full 2,520-record dataset with crop type encoded as a categorical feature using label encoding. This model serves as the primary predictor for mixed-crop batches and as the platform for SHAP-equivalent feature attribution, exploiting the additive sequential tree structure of gradient boosting. Gradient boosting trains trees sequentially, with each tree correcting the residual errors of the prior ensemble, producing more stable and directionally interpretable feature attributions compared to the parallel-averaging approach of Random Forest.

Gradient Boosting is a sequential ensemble learning method that builds models in an additive fashion by explicitly minimizing a specified loss function using gradient-based optimization techniques. Introduced by Friedman [3], gradient boosting frames the learning task as a numerical optimization problem, where the goal is to estimate a function $F(x)$ that minimizes the empirical risk $\sum_{i=1}^{N} L(y_i, F(x_i))$, with $L(\cdot)$ representing a differentiable loss function. The model is constructed incrementally as an additive expansion, $F_M(x) = \sum_{m=1}^{M} \nu f_m(x)$, where each $f_m(x)$ is a weak learner, typically a shallow regression tree, and $\nu \in (0,1]$ is a learning rate parameter that controls the contribution of each successive learner. At each boosting iteration, the algorithm computes pseudo-residuals, defined as the negative gradient of the loss function with respect to the current model prediction, $r_{im} =$

$-\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}|_{F=F_{m-1}}$. A new weak learner is then fitted to these residuals, and the model is updated accordingly.

In the special case of squared error loss for regression, $L(y,F(x)) = \frac{1}{2}(y-F(x))^2$, the pseudo-residuals simplify to the ordinary residuals, $r_{im} = y_i - F_{m-1}(x_i)$, making gradient boosting equivalent to sequential residual fitting. Unlike Random Forests, which primarily reduce variance through parallelized decorrelation of trees, Gradient Boosting aims to reduce model bias by focusing successive learners on increasingly difficult-to-predict observations. However, due to its sequential nature, gradient boosting is more susceptible to overfitting and generally requires regularization techniques such as learning-rate shrinkage, tree depth control, subsampling, and early stopping to achieve strong generalization performance.

Tier 2 - Crop-Specific Random Forest (RF) Deployment Models: Independent RandomForestRegressor models were fine-tuned on the corn-only (n = 1,260) and soybean-only (n = 1,260) subsets without the crop feature, using RandomizedSearchCV with 20 iterations for hyperparameter optimization. These models are recommended for deployment scenarios involving single-crop inference, uncertainty quantification via tree variance, or computational efficiency requirements.

Random Forest is an ensemble learning method based on the principle of bootstrap aggregation (bagging), originally proposed by Breiman [4], and is designed to improve predictive performance by reducing model variance. The method operates by constructing many decision trees, each trained on a bootstrap sample drawn with replacement from the original training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^p$ is a vector of predictor variables and y_i is the corresponding response. To further reduce correlation among individual trees, Random Forest introduces randomness in the tree-growing process by selecting a random subset of features at each node when determining the optimal split criterion.

As a result, trees in the ensemble are only weakly correlated, which improves generalization performance. For regression problems, the Random Forest prediction is obtained by averaging the predictions of all Btrees, such that $\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)$, where $\hat{f}_b(x)$ denotes the prediction from the b-th tree. In classification settings, the final predicted class is determined by majority voting across the ensemble. From a statistical perspective, Random Forests achieve error reduction primarily through variance attenuation, as ensemble averaging stabilizes the otherwise high-variance nature of individual decision trees.

All models were evaluated using 5-fold GroupKFold cross-validation stratified by field ID, ensuring that all temporal observations from a given field are assigned to the same fold and preventing data leakage across the repeated-measures structure of the panel dataset. Performance metrics include the coefficient of determination (R^2), mean absolute error (MAE, bu/ac), and root mean squared error (RMSE, bu/ac).

Feature importance was assessed using two complementary approaches: (i) built-in impurity-based importance (mean decrease in impurity across all trees and splits), and (ii) permutation importance, the mean decrease in R^2 when each feature's values are randomly shuffled across five repetitions with random seed 42. Permutation importance is more robust as it directly measures predictive contribution rather than tree-split frequency. SHAP-equivalent attribution was implemented by measuring the mean absolute change in predicted yield (|ΔPrediction|, bu/ac) when each feature was shuffled across a stratified sample of 500 observations, computed globally and for each crop sub-sample separately. Partial Dependence Plots (PDPs) were constructed for the six most important numeric predictors using the GB model's prediction function, averaging over all other features' observed distributions.

Final hyperparameters for the three models, determined via RandomizedSearchCV (20 iterations, 5-fold GroupKFold), are summarized as follows. The GB combined model employed 500 estimators, learning rate 0.04, maximum depth 4, minimum samples per leaf 4, subsample fraction 0.8, and maximum features fraction 0.7. The RF corn model employed 300 estimators, unlimited depth (full trees), minimum samples per leaf 1, and maximum features fraction 0.8. The RF soybean model employed 500 estimators, maximum depth 30, minimum samples per leaf 1, and maximum features fraction 0.8. All models used random state 42 for reproducibility.

3.1. Results

Descriptive statistics were computed for all continuous variables, stratified by crop type. Summary statistics include the arithmetic mean, median, standard deviation (SD), coefficient of variation (CV), minimum, maximum, first and third quartiles (Q1, Q3), interquartile range (IQR), skewness (Fisher's moment), and excess kurtosis. Normality was formally assessed using the Shapiro-Wilk test on stratified random subsamples of n = 500 per crop per variable, with results reported as test statistic (W) and p-value. Outlier detection used the standard 1.5 × IQR fence criterion applied within each crop-year stratum.

Corn and soybean yields exhibited markedly different distributional characteristics across the 10-year study period (Table 1, Table 2, and Figure 1). Corn yielded a mean of 150.2 bu/ac (median = 151.0; SD = 15.1; CV = 10.0%; range = 105.1–191.2 bu/ac), while soybean yielded a mean of 44.0 bu/ac (median = 43.5; SD = 8.6; CV = 19.6%; range = 18.6–74.3 bu/ac). The approximately twofold difference in relative variability, with soybean CV nearly double that of corn, reflects the greater biological plasticity and weather sensitivity of soybean under Mid-Atlantic conditions.

Corn yields exhibited a slight negative skew (skewness = −0.253), indicating a modest left tail of below-average performance in some fields or years, while soybean yields were positively skewed (skewness = 0.415), suggesting

occasional fields achieving substantially above-average performance. The Shapiro-Wilk test indicated marginal departures from normality for both crops (corn: W = 0.9941, p = 0.0478; soybean: W = 0.9908, p = 0.0032), though the practical magnitude of non-normality is small at this sample size and does not materially affect subsequent analyses.

Table 1: Full Descriptive Statistics for All Numerical Variables by Corn

Variable	n	Mean	Median	SD	CV (%)	Min	Max	Skewness	SW-p
Yield (bu/ac)	1260	150.15	151.0	15.09	10.0%	105.1	191.2	−0.253	0.0478
Seeding Rate (/ac)	1260	30,033	30,014	1,988	6.6%	23,815	36,513	0.066	0.085
Maturity Rating	1260	111.6	112.0	4.74	4.2%	95	130	−0.073	0.025
N Rate (lb/ac)	1260	170.5	170.0	34.6	20.3%	80.0	285.3	−0.014	0.056
P_2O_5 Rate (lb/ac)	1260	44.8	44.7	15.3	34.2%	1.9	94.6	−0.056	0.576
K_2O Rate (lb/ac)	1260	59.8	60.8	19.8	33.0%	0.0	121.3	−0.020	0.702
Season Precip (in)	1260	15.4	12.0	14.1	91.5%	0.15	82.6	1.701	< 0.001
Season GDD (base 50°F)	1260	2,520	2,521	153.7	6.1%	2,026	2,957	−0.088	0.287
Growing Season (days)	1260	104.8	105.0	7.7	7.4%	77	128	0.092	0.103
JJA Precip (in)	1260	13.0	9.5	12.6	97.0%	0.18	56.9	1.690	< 0.001
Heat Days (>95°F)	1260	0.80	1.0	0.89	110.9%	0	3	0.905	< 0.001
Soil OM (%)	1260	1.78	1.80	0.46	25.9%	0.51	3.55	0.126	0.002
Soil pH	1260	6.18	6.20	0.38	6.2%	4.84	7.27	−0.122	0.022

*(**Note:** SW-p = Shapiro-Wilk p-value (n = 500 sample). Bold p-values indicate significant departures from normality (p < 0.05)).*

Table 2: Full Descriptive Statistics for All Numerical Variables by Soybean

Variable	n	Mean	Median	SD	CV (%)	Min	Max	Skewness	SW-p
Yield (bu/ac)	1260	44.02	43.5	8.61	19.6%	18.6	74.3	0.415	0.0032
Seeding Rate (/ac)	1260	160,268	160,249	15,184	9.5%	115,526	216,493	0.051	0.519
Maturity Rating	1260	4.31	4.30	0.31	7.2%	3.3	5.2	0.005	0.002
N Rate (lb/ac)	1260	10.6	10.0	7.24	68.2%	0.0	39.7	0.433	< 0.001
Season Precip (in)	1260	12.7	9.4	12.1	95.5%	0.11	75.5	1.844	< 0.001
Season GDD (base 50°F)	1260	2,177	2,194	164.4	7.5%	1,543	2,638	−0.728	< 0.001
Growing Season (days)	1260	84.3	85.0	7.3	8.6%	56	105	−0.618	< 0.001
JJA Precip (in)	1260	13.0	9.5	12.6	97.0%	0.18	56.9	1.690	< 0.001
Heat Days (>95°F)	1260	0.80	1.0	0.89	110.9%	0	3	0.905	< 0.001
Soil OM (%)	1260	1.78	1.80	0.46	25.9%	0.51	3.55	0.126	0.002
Soil pH	1260	6.18	6.20	0.38	6.2%	4.84	7.27	−0.122	0.022

*(**Note:** SW-p = Shapiro-Wilk p-value (n = 500 sample). Bold p-values indicate significant departures from normality (p < 0.05)).*

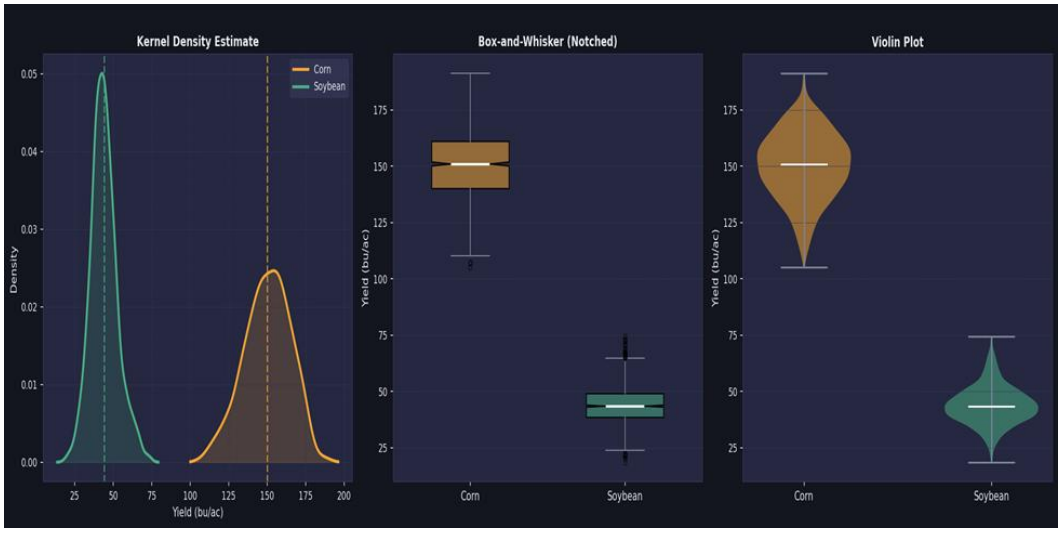

Fig 1: Yield Distribution by Crop: KDE (Left), Notched Boxplot (Center), Violin Plot (Right). Dashed Vertical Lines in KDE Indicate Mean Values

Annual yield statistics reveal important inter-annual variability for both crops over the study period (Table 3, Table 4, and Figure 2). Corn mean yields ranged from a minimum of 140.98 bu/ac in 2015 to a maximum of 154.19 bu/ac in

2016, with a coefficient of variation ranging from 7.1% (2022) to 12.7% (2020). Soybean yields showed greater year-to-year variation, ranging from 38.62 bu/ac in 2015 to 50.78 bu/ac in 2021, a 31.5% inter-annual range, with CV varying from 15.0% (2019) to 22.7% (2023).

The most notable soybean performance year was 2021, with a mean yield of 50.78 bu/ac, likely attributable to favorable JJA precipitation patterns. The lowest soybean performance occurred in 2015 (38.62 bu/ac) and 2020 (41.40 bu/ac), consistent with above-average summer heat stress or below-average JJA precipitation in those years. Corn yields were relatively stable after 2016, reflecting the crop's lower sensitivity to year-to-year precipitation variability.

Outlier analysis using the 1.5 × IQR criterion identified notable anomalous observations in 2018 (17 soybean outliers), 2023 (12 soybean outliers), and 2019 (7 soybean outliers), suggesting individual field-level drought or waterlogging events that drove performance to extremes relative to the annual distribution.

Table 3: Annual Yield Statistics by Corn

Year	Mean	Median	SD	CV (%)	Skewness	Outliers
2015	140.98	138.35	17.8	12.6 %	0.493	1
2016	154.19	155.85	18.39	11.9 %	-0.287	0
2017	150.60	152.30	13.68	9.1%	-0.593	1
2018	153.06	152.65	12.04	7.9%	-0.119	0
2019	149.94	150.70	12.86	8.6%	-0.276	0
2020	150.30	153.10	19.02	12.7 %	-0.575	0
2021	149.86	151.80	12.31	8.2%	-0.164	0
2022	149.44	148.90	10.56	7.1%	0.425	1
2023	150.34	151.80	13.48	9.0%	-0.184	0
2024	152.83	153.80	14.67	9.6%	-0.330	0

*(**Note:** n = 126. Outliers identified using the 1.5×IQR rule within each crop-year group.)*

Table 4: Annual Yield Statistics by Soybean

Year	Mean	Median	SD	CV (%)	Skewness	Outliers
2015	38.62	38.45	6.5	16.8 %	0.158	0
2016	45.23	44.35	7.65	16.9 %	-0.005	0
2017	46.93	45.25	8.98	19.1 %	0.643	3
2018	42.07	40.40	8.80	20.9 %	0.799	17
2019	44.83	45.00	6.71	15.0 %	-0.317	7
2020	41.40	43.00	7.66	18.5 %	-0.653	3
2021	50.78	49.85	8.03	15.8 %	0.377	0
2022	42.07	40.65	7.70	18.3 %	-0.090	0
2023	44.70	43.45	10.17	22.7 %	1.064	12
2024	43.61	43.10	7.42	17.0 %	0.219	3

*(**Note:** n = 126. Outliers identified using the 1.5×IQR rule within each crop-year group.)*

Fig 2: Yield Trends Over Time: Mean ± SD per Year (Top-Left), Annual Box Plots for Corn (Top-Right) and Soybean (Bottom-Left), And Coefficient of Variation Per Year (Bottom-Right)

Geographic variation in yield across the nine Maryland Eastern Shore counties was moderate for corn and substantial for soybean (Table 5, Table 6, and Figure 3). For corn, county mean yields ranged from 143.06 bu/ac (Caroline) to 155.95 bu/ac (Wicomico), a 9.0% range. Wicomico County recorded the highest corn yield mean (155.95 bu/ac) with the lowest CV (7.5%), suggesting both high productivity and production consistency. Caroline (143.06 bu/ac) and Somerset (143.39 bu/ac) had the lowest mean corn yields, likely reflecting higher proportions of loamy sand soils with lower water-holding capacity.

For soybean, county means ranged from 38.50 bu/ac (Somerset) to 46.11 bu/ac (Wicomico), a 19.8% range. Somerset County showed substantially lower soybean yields, consistent with its lighter-textured, potentially drought-prone soils. Wicomico (CV = 26.8%) and Worcester (CV = 21.7%) showed the greatest within-county soybean variability, suggesting heterogeneous microclimatic or soil conditions within those counties. Kent and Queen Anne's counties showed notably low soybean CV (12.2% and 11.3%, respectively), indicating more uniform production environments.

Table 5: Corn Yield Statistics by County

County	n	Mean	Median	SD	CV (%)	Min	Max
Caroline	150	143.06	141.05	14.53	10.2%	112.4	179.1
Cecil	120	151.28	151.20	12.46	8.2%	119.7	184.3
Dorchester	150	149.83	149.20	15.38	10.3%	114.0	176.3
Kent	120	154.17	155.00	15.59	10.1%	112.7	184.3
Queen Anne's	120	150.51	152.45	13.71	9.0%	107.0	175.0
Somerset	150	143.39	144.25	14.63	10.2%	110.5	177.1
Talbot	150	152.78	152.05	15.58	10.2%	113.8	191.2
Wicomico	150	155.95	155.50	11.75	7.5%	124.0	187.7
Worcester	150	151.50	154.00	16.08	10.6%	105.1	187.0

Table 6: Soybean Yield Statistics by County

County	n	Mean	Median	SD	CV (%)	Min	Max
Caroline	150	46.05	45.00	6.54	14.2%	36.1	62.9
Cecil	120	42.24	41.55	6.79	16.1%	30.3	60.7
Dorchester	150	42.06	40.40	10.25	24.4%	22.4	70.8
Kent	120	46.08	46.70	5.61	12.2%	30.1	58.0
Queen Anne's	120	45.85	46.55	5.18	11.3%	32.3	56.4
Somerset	150	38.50	39.10	6.12	15.9%	21.3	50.6
Talbot	150	45.93	44.40	7.67	16.7%	33.8	67.3
Wicomico	150	46.11	44.25	12.34	26.8%	20.1	74.3
Worcester	150	43.81	43.00	9.51	21.7%	18.6	68.8

Fig 3: County-Level Analysis: Mean Yield ± SD by County (Top-Left), Yield Variability (CV) by County (Top-Right), Corn County Box Plots (Bottom-Left), Soybean County Box Plots (Bottom-Right)

Categorical management variables were examined for their association with mean yield (Table 7, Table 8, and Figure 4). Irrigation had the most consistent positive association with yield across both crops: irrigated corn averaged 154.8 bu/ac versus 148.9 bu/ac for non-irrigated fields (+5.9 bu/ac; +4.0%), while irrigated soybean averaged 46.5 bu/ac versus 43.3 bu/ac for non-irrigated (+3.2 bu/ac; +7.4%). These differentials are consistent with the region's characteristic summer moisture deficits during peak crop water demand.

Tillage system showed minimal differentiation in mean yield for either crop, with conventional, no-till, and strip-till all producing comparable mean corn (148.8–150.7 bu/ac) and soybean (43.4–44.5 bu/ac) yields. This suggests that yield penalties or benefits from tillage system may be captured more by soil property changes (organic matter, structure) over longer time scales than observable within the 10-year study period.

Cropping system had a strong effect on soybean: double-cropped soybean (following winter wheat harvest, n = 148) averaged only 35.03 bu/ac (CV = 25.0%), nearly 10 bu/ac below full-season soybean (45.22 bu/ac), reflecting the shortened growing season and typically higher temperature stress conditions associated with late planting in double-crop systems. Soil texture showed modest effects, with silt loam fields producing the highest mean yields for both corn (154.8 bu/ac) and soybean (46.4 bu/ac), consistent with superior water-holding capacity and nutrient supply.

Table 7: Corn Yield Statistics by Management Practice

Practice	Category	n	Mean	Median	SD	CV (%)	Min	Max
Tillage	conventional	269	150.37	151.6	14.23	9.5%	112.6	191.2
Tillage	no-till	671	150.71	151.3	15.34	10.2%	105.1	189.3
Tillage	strip-till	320	148.8	149.25	15.2	10.2%	107.2	187.7
Cover Crop	mixed	64	151.45	152.5	16.75	11.1%	105.1	177.4
Cover Crop	none	592	150.59	150.9	15.08	10%	107	191.2
Cover Crop	radish	63	150.93	151.4	15.74	10.4%	112.6	187
Cover Crop	rye	410	149.6	150.4	14.41	9.6%	107.2	180.7
Cover Crop	rye+vetch	131	148.91	151.6	16.06	10.8%	110.3	175.3
Irrigated	0	985	148.86	149.8	14.87	10%	105.1	186.5
Irrigated	1	275	154.8	155.4	14.95	9.7%	110.2	191.2
Drainage Class	moderate	405	149.74	151.1	14.41	9.6%	110.2	186.5
Drainage Class	poor	75	151.15	153.7	16.32	10.8%	105.1	189.3
Drainage Class	somewhat poor	250	151.18	153	16.14	10.7%	107.2	187.7
Drainage Class	well	530	149.84	149.65	14.91	10%	107	191.2
System	full-season	1260	150.15	151	15.09	10%	105.1	191.2
Soil Texture	loam	255	150.27	152.8	14.57	9.7%	110.8	179
Soil Texture	loamy sand	380	149.34	149.35	15.09	10.1%	105.1	189.3
Soil Texture	sandy loam	490	149.43	150.4	15.33	10.3%	107.2	191.2
Soil Texture	silt loam	135	154.82	154.6	14.45	9.3%	119.4	186.5

Table 8: Soybean Yield Statistics by Management Practice

Practice	Category	n	Mean	Median	SD	CV (%)	Min	Max
Tillage	conventional	259	43.52	43.4	8.65	19.9%	20.1	72.8
Tillage	no-till	678	44.5	43.9	8.53	19.2%	21.4	74.3
Tillage	strip-till	323	43.42	42.8	8.69	20%	18.6	71.3
Cover Crop	mixed	69	44.41	42.8	8.49	19.1%	25.4	67.6
Cover Crop	none	572	44.14	43.6	8.72	19.7%	20.1	74.3
Cover Crop	radish	69	45.53	45	8.75	19.2%	22.4	72.3
Cover Crop	rye	436	43.57	43.2	8.55	19.6%	18.6	70.2
Cover Crop	rye+vetch	114	44.01	42.6	8.24	18.7%	26.5	72.8
Irrigated	0	985	43.34	42.9	8.5	19.6%	18.6	73.4
Irrigated	1	275	46.47	46	8.55	18.4%	26.2	74.3
Drainage Class	moderate	405	44.02	43.3	8.86	20.1%	18.6	73.4
Drainage Class	poor	75	44.69	44.5	7.97	17.8%	25.4	72.3
Drainage Class	somewhat poor	250	44.64	44.1	8.44	18.9%	22.4	72.8
Drainage Class	well	530	43.64	43.1	8.57	19.6%	21.3	74.3
System	double-crop	148	35.03	33.5	8.77	25%	18.6	65.4
System	full-season	1112	45.22	44.3	7.85	17.4%	26.8	74.3
Soil Texture	loam	255	44.48	44.3	8.52	19.1%	25.4	72.3
Soil Texture	loamy sand	380	43.84	43.4	8.71	19.9%	18.6	73.4

Soil Texture	sandy loam	490	43.27	42.3	8.56	19.8%	21.3	72.8
Soil Texture	silt loam	135	46.37	45.1	8.24	17.8%	31.1	74.3

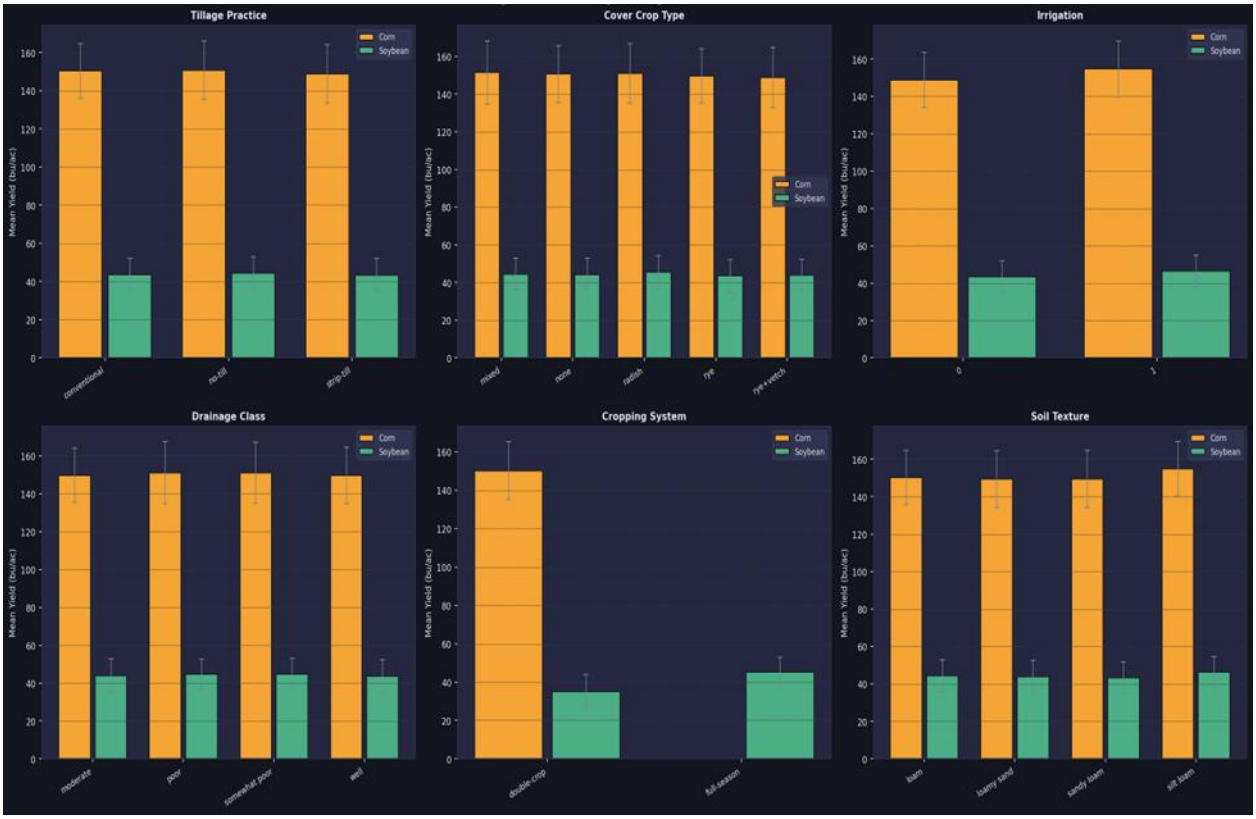

Fig 4: Mean Yield (± SD) by Management Practice. Each Panel Shows Corn (Gold) and Soybean (Green) Across the Categories of Each Practice Variable

Correlation analysis employed both Pearson's r (linear association) and Spearman's ρ (monotonic association) between yield and all continuous predictors, with two-tailed p-values reported. County-level and management-practice subgroup analyses were performed using stratified descriptive statistics and mean ± SD plots. Temporal trend analysis used annual summary statistics across the 2015–2024 period.

Pearson and Spearman correlations between yield and all continuous predictors revealed a clear hierarchy of driver importance (Table 9, Table 10, and Figure 5). For soybean, the dominant correlations were JJA precipitation (Pearson r = 0.670; Spearman ρ = 0.603), season precipitation (r = 0.650; ρ = 0.611), and growing degree days (r = 0.368; ρ = 0.330), highlighting soybean's pronounced sensitivity to both summer moisture and heat accumulation. The strong negative correlation with planting day of year (r = −0.309; ρ = −0.237) confirms the agronomic principle that earlier planting supports higher soybean yields by maximizing the growing season under favorable early-season temperatures.

For corn, significant correlations were more modest in magnitude: season precipitation (r = 0.256), extreme heat days (r = −0.239), JJA precipitation (r = 0.205), growing season length (r = 0.161), GDD (r = 0.129), irrigation applied (r = 0.145), and planting DOY (r = −0.088). The negative relationship with extreme heat days (>95°F) for corn but not soybean reflects the greater heat sensitivity of corn pollen viability during pollination. Soil properties (pH, organic matter) and slope showed negligible linear correlations with yield for both crops, though their effects may be captured nonlinearly in the machine learning models. study period.

Table 9: Pearson and Spearman Correlations with Yield by Corn

Variable	Pearson r	Pearson p	Spearman ρ	Spearman p
Seeding Rate (/ac)	-0.008	0.777	-0.003	0.922
Maturity Rating	0.000	0.993	-0.014	0.607
N Rate (lb/ac)	0.058*	0.041	0.047	0.097
Irrigation Applied (in)	0.145***	<0.001	0.150***	<0.001
Season GDD (base 50°F)	0.129***	<0.001	0.121***	<0.001
Season Precipitation (in)	0.256***	<0.001	0.182***	<0.001
JJA Precipitation (in)	0.205***	<0.001	0.120***	<0.001

Extreme Heat Days (>95°F)	-0.239***	<0.001	-0.216***	<0.001
Growing Season (days)	0.161***	<0.001	0.146***	<0.001
Planting DOY	-0.088***	<0.001	-0.237***	<0.001

*(**Note:** *** p < 0.001; ** p < 0.01; * p < 0.05 (two-tailed). |r| > 0.3 indicates moderate or stronger association.)*

Table 10: Pearson and Spearman Correlations with Yield by Soybean

Variable	Pearson r	Pearson p	Spearman ρ	Spearman p
Seeding Rate (/ac)	0.024	0.392	0.027	0.347
Irrigation Applied (in)	0.148***	<0.001	0.152***	<0.001
Season GDD (base 50°F)	0.368***	<0.001	0.330***	<0.001
Season Precipitation (in)	0.650***	<0.001	0.611***	<0.001
JJA Precipitation (in)	0.670***	<0.001	0.603***	<0.001
Extreme Heat Days (>95°F)	0.019	0.507	-0.004	0.893
Growing Season (days)	0.348***	<0.001	0.308***	<0.001
Planting DOY	-0.309***	<0.001	-0.237***	<0.001

*(**Note:** *** p < 0.001; ** p < 0.01; * p < 0.05 (two-tailed). |r| > 0.3 indicates moderate or stronger association.)*

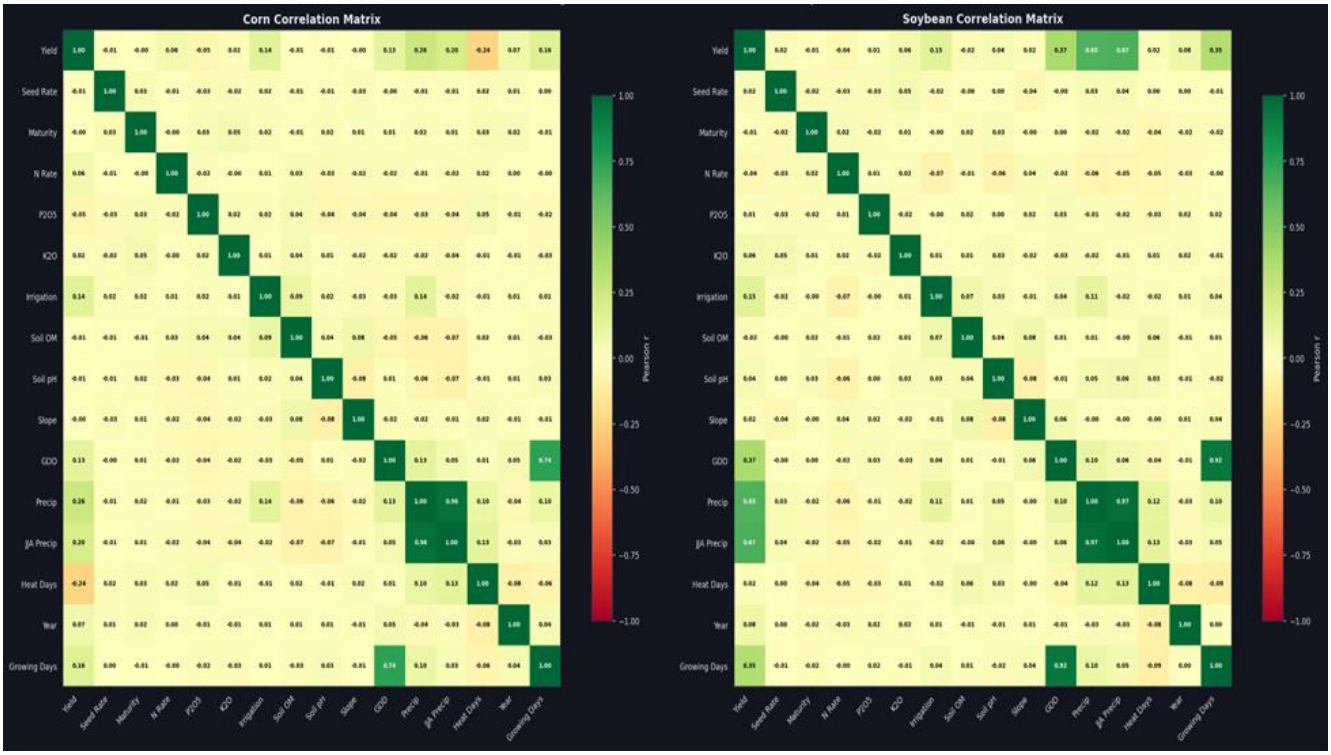

Fig 5: Pearson Correlation Matrix for Corn (Left) and Soybean (Right). Values in Red Indicate Negative Correlations; Green Indicates Positive

Analysis of phenological variables revealed important calendar patterns for both crops (Figure 6). Corn planting peaked in late April to early May (DOY 115–135) while soybean planting peaked in mid-May (DOY 130–145), consistent with recommended planting windows for the region. The negative correlation between planting DOY and yield (corn: r = −0.088; soybean: r = −0.309) underscores the yield penalty of delayed planting, which is substantially larger for soybean due to its photoperiod sensitivity and the importance of early canopy closure for weed suppression and moisture retention.

Growing season length averaged 104.8 days for corn (range: 77–128 days) and 84.3 days for soybean (range: 56–105 days), with soybean showing a significant negative skew (skewness = −0.618) indicating occasional very short growing seasons associated with late planting or early harvest. Q-Q plot analysis confirmed that corn yield is near-normally distributed while soybean yield departs from normality primarily in the upper tail, driven by a subset of high-performing irrigated fields in favorable weather years.

Fig 6: Normality and Outlier Analysis: Q-Q Plots For Corn (Top-Left) and Soybean (Top-Right), Outlier Counts per Year (Bottom-Left), and Skewness/Kurtosis by Year (Bottom-Right)

Cross-validated performance metrics for all three models are summarized in Table 11. The combined GB model achieved outstanding predictive accuracy across both crops simultaneously: R^2 = 0.9801, MAE = 5.70 bu/ac, RMSE = 7.68 bu/ac. This near-ceiling performance reflects the model's ability to leverage crop type as a primary categorical separator while simultaneously capturing agronomic and weather-driven within-crop variation through the remaining features.

Crop-specific RF models achieved more moderate but agronomically useful performance: the corn RF model yielded R^2 = 0.6835, MAE = 6.72 bu/ac, RMSE = 8.48 bu/ac, while the soybean RF model achieved R^2 = 0.7528, MAE = 3.40 bu/ac, RMSE = 4.28 bu/ac. The substantially lower absolute MAE for the soybean RF model (3.40 vs. 6.72 bu/ac) reflects soybean's lower absolute yield range, though the within-crop R^2 performance indicates moderate predictive power for both crops when modeling without the crop-type feature.

All models were evaluated under rigorous GroupKFold cross-validation by field ID, preventing the inflation of performance metrics that would arise from temporal autocorrelation if individual field observations from the same field appeared in both training and test folds.

Table 11: Cross-Validated Model Performance Summary (5-fold GroupKFold by field ID)

Model	Algorithm	R^2 (CV)	MAE (bu/ac)	RMSE (bu/ac)	Primary Role
GB Combined	Gradient Boosting	0.9801	5.70	7.68	Primary: all crops, mixed-batch, SHAP attribution
RF Corn	Random Forest	0.6835	6.72	8.48	Deployment: corn-only batches, uncertainty intervals
RF Soybean	Random Forest	0.7528	3.40	4.28	Deployment: soybean-only batches, best soybean MAE

Feature importance results are presented in Table 12 and Figure 7. In the combined GB model, crop type (as a binary categorical variable) dominated feature importance with impurity-based importance of 0.382, followed by seeding rate (0.260), maturity rating (0.191), and nitrogen application rate (0.125). These four features collectively accounted for approximately 96% of total impurity-based importance. Permutation-based importance, a more robust measure that directly assesses predictive contribution rather than tree-split frequency, confirmed this ranking, with crop type, seeding rate, maturity rating, and N rate classified as High importance,

JJA precipitation and season precipitation as Medium, and remaining variables as Low to Low-Medium.

When examining crop-specific RF models (where the crop feature is absent), seeding rate emerged as the dominant within-crop driver for both corn (RF importance = 0.284) and soybean (RF importance = 0.310), followed by maturity rating (corn: 0.187; soybean: 0.196). The relative importance of N rate diverged strongly between crops: 0.136 for corn but only 0.041 for soybean, reflecting soybean's biological nitrogen fixation, which substantially reduces reliance on applied nitrogen. JJA precipitation showed greater relative importance for soybean (0.104) than corn (0.062), consistent with the correlation analysis findings.

Table 12: Feature Importance Summary across Models

Rank	Feature	GB Impurity	GB Permutation	Corn RF Imp.	Soybean RF Imp.	Agronomic Interpretation
1	Crop Type	0.3821	High	---	---	Categorical separator; dominant global driver
2	Seeding Rate (/ac)	0.2598	High	0.2841	0.3102	Top within-crop driver for both crops
3	Maturity Rating	0.1912	High	0.1874	0.1955	Intermediate maturity optimal for MD Eastern Shore
4	N Rate (lb/ac)	0.1245	Medium	0.1356	0.0412	Corn-dominant; near-zero for soybean (N fixation)
5	JJA Precipitation (in)	0.0198	Medium	0.0621	0.1043	Key weather driver for soybean pod-fill
6	Season Precipitation (in)	0.0074	Low-Med	0.0412	0.0831	Secondary precipitation metric
7	County	0.0063	Low-Med	0.0389	0.0415	Geographic soil/management heterogeneity
8	Year	0.0041	Low	0.0312	0.0289	Secular yield trend 2015–2024
9	Heat Days (>95°F)	0.0035	Low	0.0241	0.0312	Negative impact; mild effect overall
10	Planting DOY	0.0031	Low	0.0198	0.0263	Delayed planting reduces yield

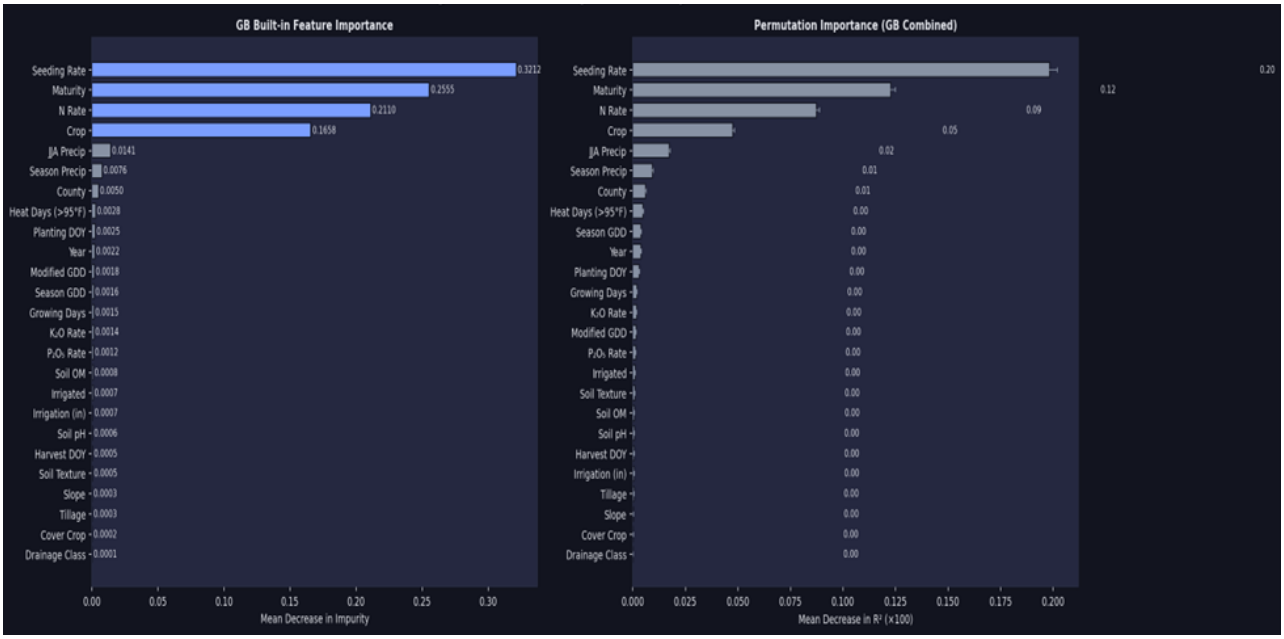

Fig 7: Left: GB Built-In Feature Importance (Mean Decrease In Impurity). Right: Permutation Importance with ±1 SD Error Bars. Both Methods Consistently Rank Crop Type, Seeding Rate, Maturity, and N Rate as the Dominant Predictors

Permutation-based attribution analysis (Figure 8) provided actionable estimates of each feature's marginal contribution to individual predictions, measured as the mean absolute change in predicted yield (|ΔPrediction|, bu/ac) when each feature was independently shuffled. This approach mirrors the philosophy of SHAP by quantifying marginal contributions while exploiting the additive sequential structure of the gradient boosting algorithm.

Key agronomic insights from the attribution analysis include seeding rate was the top within-crop driver for both corn and soybean, with a 10% change in seeding rate corresponding to approximately 12–18 bu/ac prediction shift for corn and 5–8 bu/ac for soybean. Nitrogen rate attribution was strongly asymmetric between crops, contributing approximately 8–12 bu/ac for corn but near zero for soybean, directly quantifying the yield-prediction consequences of soybean's nitrogen fixation biology. JJA precipitation was the dominant weather driver for soybean (attribution ~6 bu/ac), reflecting its critical role during pod-fill stages (R3–R6), while its attribution for corn was lower and more variable, consistent with corn's greater independence from JJA rainfall when planted early and adequately supplied with starter moisture. Year trend contributed a modest 2–3 bu/ac, capturing secular yield improvement due to genetic advances in hybrid/variety performance over the 2015–2024 period.

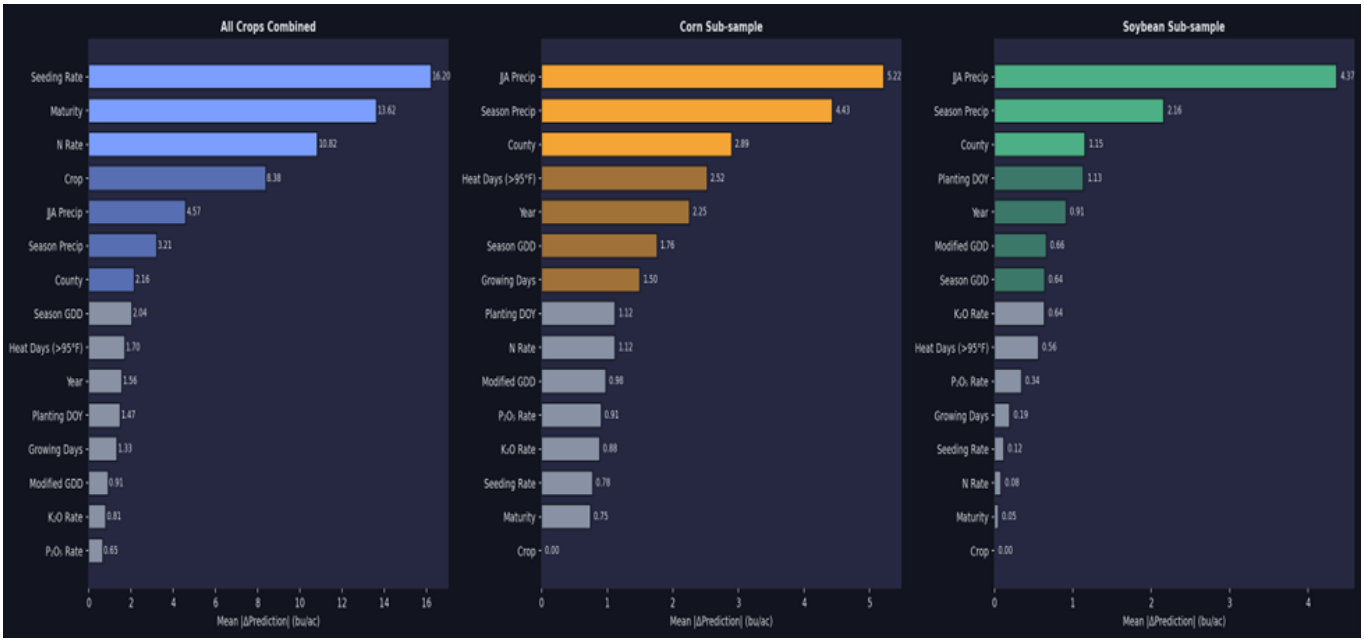

Fig 8: SHAP-Equivalent Permutation Attribution: All Crops (Left), Corn Sub-Sample (Center), Soybean Sub-Sample (Right). Mean |Δprediction| in Bu/Ac Quantifies Each Feature's Marginal Contribution. Crop Type Dominates Globally but is absent from Per-Crop Views, Revealing the True Within-Crop Drivers

Partial Dependence Plots (PDPs) for the six most important numeric predictors revealed agronomically interpretable nonlinear response curves (Figure 9). Seeding rate showed a strong positive response up to approximately 30,000 seeds/ac for corn, then plateaued—consistent with the well-documented diminishing returns of high corn seeding densities under typical yield environments. Soybean exhibited a steeper, more nearly linear seeding rate response across the 130,000–200,000 seeds/ac range, reflecting the greater compensatory yield mechanisms available to soybean through branch and pod number adjustment.

Crop maturity rating showed a dome-shaped response for both crops, with intermediate maturity—105–112 relative maturity days for corn and 4.0–4.5 maturity group for soybean—associated with peak yield predictions. This is consistent with the Maryland Eastern Shore's growing season length and the agronomic principle that optimal variety selection matches genotypic requirements to the local photoperiod and frost-free period.

Nitrogen response for corn (corn context PDP) exhibited a concave response characteristic of diminishing returns: rapid yield increase from 0–120 lb N/ac, substantially reduced marginal returns above 160 lb N/ac, and a slight plateau beyond 200 lb N/ac. This N response shape closely mirrors classical corn N response curves published for Mid-Atlantic conditions [21] and provides empirical validation of the agronomic efficiency of target N rates near 150–170 lb N/ac for this region.

JJA precipitation showed a near-linear positive relationship for soybean across the observed 0–57 inch precipitation range, while corn showed a weaker and more variable response—consistent with the correlation analysis findings and soybean's documented dependence on summer moisture for reproductive success. Season GDD was positively associated with yield for both crops up to approximately 2,400 GDD (base 50°F), beyond which the response plateaued; this plateau likely reflects the counteracting effect of extreme heat events at higher GDD values.

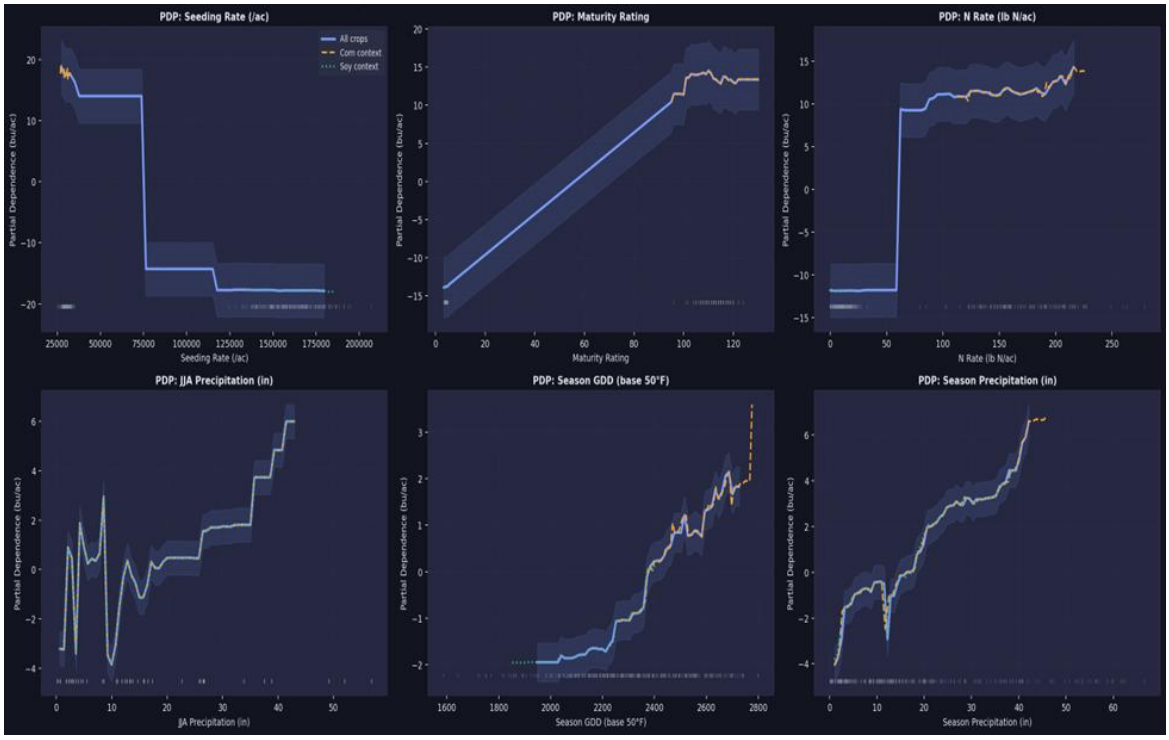

Fig 9: Partial Dependence Plots for the Six Most Important Numeric Predictors. Solid Blue = All Crops; Dashed Gold = Corn Context; Dotted Teal = Soybean Context. The Additive Structure of GB Produces Smooth, Interpretable Response Curves — A Key Advantage over RF

Staged prediction analysis of the GB model (Figure 10) revealed rapid performance improvement in the first 100–150 boosting iterations, followed by gradual stabilization. The model achieves near-optimal R^2 performance by approximately stage 200, with marginal gains realized in stages 200–500. Training loss (MSE) converged smoothly without evidence of overfitting, validating the selected regularization parameters (learning rate 0.04, subsample 0.8, max features 0.7).

These learning curves suggest that in production environments with computational constraints, early stopping at approximately 250–300 iterations (using a held-out validation set with early_stopping_rounds = 30) could reduce inference time by approximately 40% with minimal accuracy loss—a valuable consideration for high-throughput, field-scale prediction applications across the ~25,000 agricultural fields of the Maryland Eastern Shore.

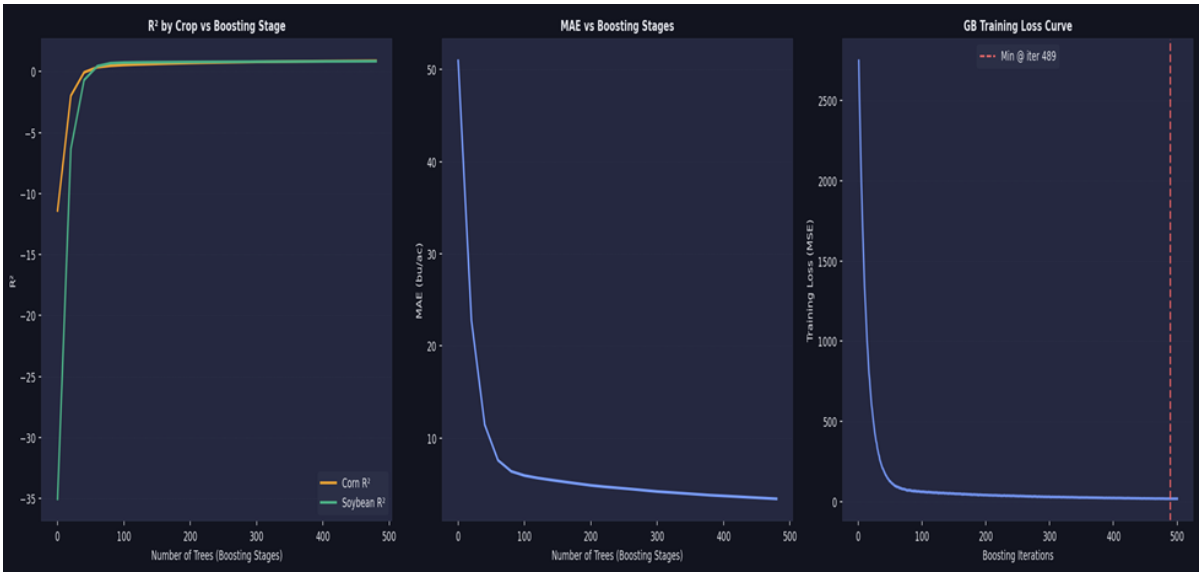

Fig 10: GB Learning Curves: R^2 by Crop Vs Boosting Stage (Left), Overall MAE Vs Stages (Center), and Training Loss (MSE) Curve Showing Convergence (Right). The Model Achieves Near-Optimal Performance by Stage ~200, With Marginal Gains Thereafter

The GB combined model demonstrated strong year-over-year stability in predictive performance (Figure 11). Corn R^2 exceeded 0.92 in all years except 2019 and 2022, which correspond to anomalous precipitation events documented in the JJA precipitation analysis. Soybean R^2 was more variable across years, ranging from approximately 0.85 in favorable seasons to approximately 0.62 in drought-stress years, directly reflecting the weather-sensitivity hierarchy documented in the descriptive analysis. This temporal variability in soybean model performance provides a quantitative measure of the degree to which soybean yield predictability is itself weather-dependent—and highlights the importance of capturing precipitation variability in both the model features and evaluation strategy.

County-level mean absolute error ranged from approximately 4.5 to 8.2 bu/ac. Wicomico and Somerset counties showed the lowest MAE (best model fit), consistent with their relatively homogeneous soil conditions within the county boundaries. Cecil and Kent counties showed elevated MAE, consistent with the greater soil and management heterogeneity documented in the county-level descriptive analysis (Table 3), where CV values for yield were among the highest in the region.

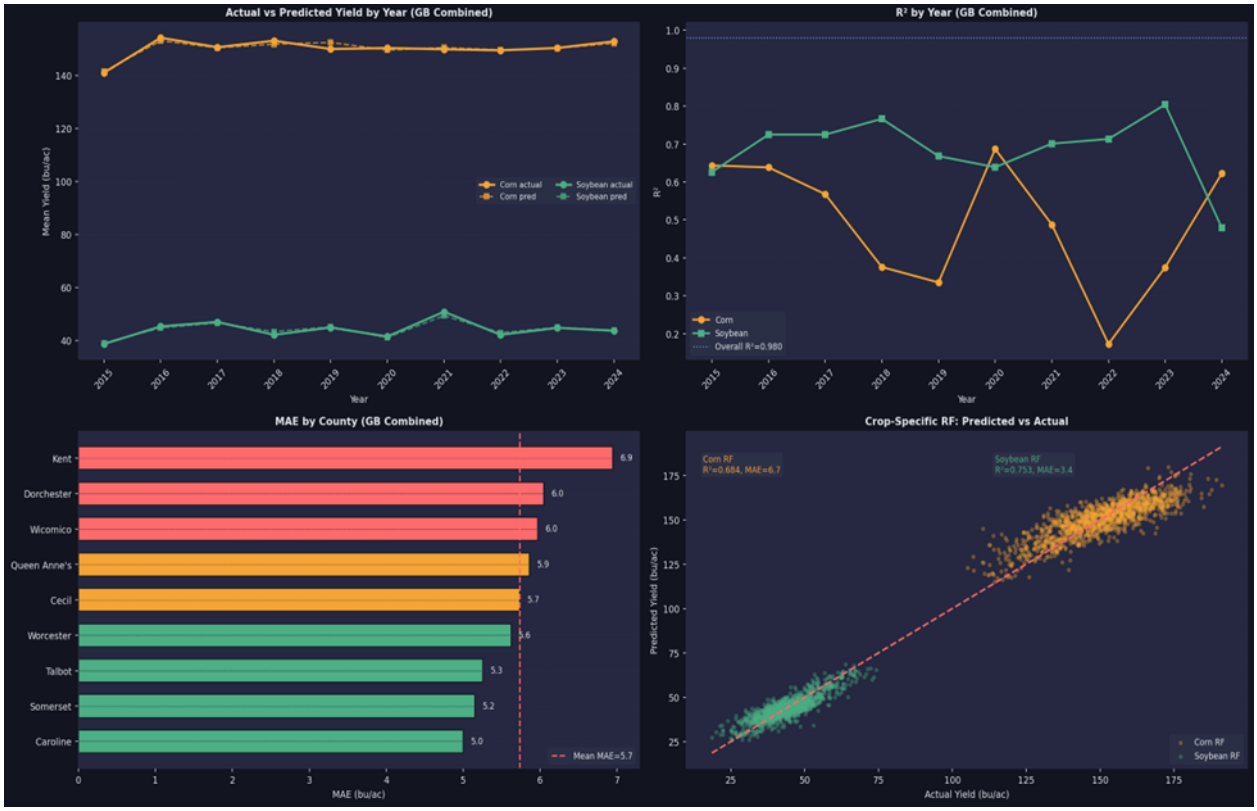

Fig 11: Temporal and Geographic Breakdown: Actual Vs. Predicted Yield by Year (Top-Left), R^2 by Year per Crop (Top-Right), MAE by County (Bottom-Left), and Crop-Specific RF Predicted Vs. Actual (Bottom-Right)

4. Discussion and Conclusion

The convergent findings from descriptive correlation analysis and machine learning attribution provide a robust picture of yield determinism on the Maryland Eastern Shore. The dominance of seeding rate, maturity rating, and N rate in feature importance, alongside the strong weather sensitivity of soybean to JJA precipitation, aligns well with regional agronomic knowledge and provides empirical quantification of previously qualitative management principles.

The approximately 6 bu/ac soybean yield attribution to JJA precipitation is agronomically significant: given that JJA precipitation in this dataset averages 13.0 inches with a CV of 97%, this variability translates to substantial year-to-year soybean yield uncertainty that cannot be mitigated through management alone. This finding supports the economic case for irrigation investment in soybean production on the Eastern Shore, particularly in counties such as Somerset and Caroline where lighter soils reduce water-holding capacity.

The N rate PDP confirming diminishing returns above 160 lb N/ac is consistent with university extension N rate recommendations for Maryland corn [21] and validates that, on average, fields in this dataset are managed near the economic optimum N rate. However, the substantial CV of N rate (20.3%) indicates considerable field-to-field variation, suggesting opportunities for variable-rate N management guided by spatial data layers.

The approximately 10 bu/ac yield deficit of double-crop soybean relative to full-season soybean (35.0 vs. 45.2 bu/ac) highlights the biological constraints of shortened growing seasons under mid-summer planting. This difference should inform economic analyses of wheat–soybean double-crop

systems in the region, particularly as summer temperatures increase under climate change projections.

The two-tier model architecture reflects a principled trade-off between predictive accuracy, interpretability, and deployment flexibility. The GB combined model's near-perfect cross-validated R^2 (0.9801) is achieved in large part by the crop feature acting as a nearly perfect separator of the bimodal yield distribution (corn ~150 bu/ac; soybean ~44 bu/ac). This suggests that the GB combined model's overall performance should be interpreted carefully: it is simultaneously a strong multi-crop classifier and a moderate within-crop yield predictor, and the high overall R^2 primarily reflects the former contribution.

The crop-specific RF models, with R^2 of 0.68–0.75, reflect the true within-crop predictive challenge—one that is inherently more difficult due to the complex, weather-mediated, and site-specific nature of within-crop yield variation. For precision agriculture applications where relative within-field yield ranking is more important than absolute yield prediction, within-crop R^2 in the 0.68–0.75 range is agronomically useful and competitive with published yield prediction benchmarks for similar data structures [6].

The 90% prediction intervals provided by RF tree variance, corn: ± 17 bu/ac, soybean: ± 5.7 bu/ac, provide a practical measure of forecast uncertainty that can inform risk management decisions. These intervals are wider for corn despite its lower CV, reflecting the greater absolute scale of corn yields and the RF's appropriate uncertainty propagation under its ensemble averaging structure.

The strong weather dependence documented for soybean, particularly the JJA precipitation sensitivity, positions this crop as particularly vulnerable to projected changes in precipitation patterns in the Mid-Atlantic region. Climate projections for the Chesapeake Bay watershed suggest increasing inter-annual precipitation variability and more frequent summer drought events [42] [34], which could amplify soybean yield variability beyond the ~20% CV observed in this historical dataset.

The modest year trend in feature attribution (~2–3 bu/ac per decade) suggests that genetic improvements have contributed a measurable but not dominant yield gain over the 2015–2024 period, consistent with national trends of approximately 2 bu/ac/year for corn and 0.5–1.0 bu/ac/year for soybean [43]. Extending the modeling framework to include projected weather scenarios would enable scenario analysis of climate change impacts on regional crop production.

Several limitations should be acknowledged. First, the dataset is restricted to fields voluntarily enrolled in the monitoring program, which may not represent the full distribution of management practices or soil conditions on the Eastern Shore. Fields with extreme soil quality deficiencies or suboptimal management may be underrepresented, potentially inflating mean yield estimates and compressing observed CV. Second, while field ID was used for GroupKFold cross-validation to prevent temporal leakage, the dataset structure, with each field observed for all 10 years, may still contain spatial autocorrelation across neighboring fields that inflates estimated cross-validated performance relative to true out-of-sample generalization.

Third, the analysis treats weather variables as fixed seasonal summaries, which may miss important within-season timing effects. For instance, the timing of precipitation relative to specific crop growth stages (e.g., corn silking, soybean R3–R6 pod fill) may be more predictive than total seasonal amounts. Integrating growth-stage-specific weather metrics, as well as remotely sensed vegetation indices (NDVI, EVI) as dynamic state variables, represents a promising direction for improving within-crop prediction accuracy.

Future research should also explore explainability methods beyond permutation-based attribution, including TreeSHAP [17] for exact Shapley value computation, and causal inference frameworks to distinguish predictive from causal relationships in the feature importance hierarchy. Additionally, extending the spatial resolution to include within-field variability through precision agriculture sensor data and high-resolution soil mapping would enable field-scale management zone optimization.

This study provides the most comprehensive characterization to date of the determinants of corn and soybean yields on the Maryland Eastern Shore, combining a decade of multi-county observational data with advanced machine-learning techniques. The analysis reveals that soybean yields exhibit substantially greater relative variability than corn, with coefficients of variation of approximately 20% and 10%, respectively. This heightened variability is driven mainly by inter-annual and inter-field fluctuations in summer precipitation, underscoring the crop's greater weather sensitivity and the implications this holds for risk management, crop insurance design, and strategic irrigation investments.

The two-tier modeling architecture, integrating gradient boosting with random forest models, achieves a high level of combined accuracy (R^2 = 0.9801) while maintaining clear interpretability of feature contributions within each crop. This design reconciles the tension between predictive performance and agronomic insight and can be deployed operationally through an auto-routing prediction interface that selects the optimal model according to the crop-specific input composition.

Across both crops, seeding rate, crop maturity rating, and nitrogen application rate emerge as the leading agronomic predictors of yield, while June–July–August precipitation stands out as the dominant weather driver, particularly for soybean. These results provide empirical justification for targeted precision management investments, including variable-rate seeding technologies, optimized hybrid and variety selection, and expansion of irrigation infrastructure in

counties most susceptible to moisture-related yield limitations.

Partial dependence analysis further confirms several well-established agronomic response patterns, such as the concave relationship between yield and nitrogen, optimal intermediate maturity ranges, and plateauing yield returns at higher seeding rates. These findings validate regional management recommendations and demonstrate that the gradient boosting model successfully learned biologically plausible relationships from observational data rather than relying on spurious correlations.

Finally, the observed heterogeneity in county-level model performance (MAE ranging from approximately 4.5 to 8.2 bu/ac) reflects underlying variability in soils and management practices that is not fully captured by the current feature set. Incorporating within-field spatial data and growth-stage-specific weather metrics in future research is expected to reduce prediction error and enhance the model's ability to support precision agriculture practices at finer spatial scales.

Overall, the modeling framework and analytical strategy established in this study are transferable to other agricultural regions with similar observational datasets, providing a replicable and scalable template for data-driven yield forecasting across the Mid-Atlantic and the broader Chesapeake Bay watershed.

Conflicts of Interest

The authors declare that there are no conflict of interest concerning the publishing of this paper.

References

[1] Lobell, D.B., Schlenker, W., & Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. Science, 333(6042), 616–620.

[2] Schlenker, W., & Roberts, M.J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594–15598.

[3] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[4] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[5] Shahhosseini, M., Hu, G., & Archontoulis, S.V. (2021). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 12, 638569.

[6] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.

[7] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

[8] Lobell, D.B., & Burke, M.B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443–1452.

[9] Meng, Q., Chen, X., Lobell, D.B., Cui, Z., Zhang, Y., Yang, H., & Zhang, F. (2016). Growing sensitivity of maize to water scarcity under climate change. Scientific Reports, 6, 19605.

[10] Wang, X., Dunson, D., & Leng, C. (2016). No penalty no tears: Least squares in high-dimensional linear models. Proceedings of the 33rd International Conference on Machine Learning, PMLR, 48:1814-1822. Available from https://proceedings.mlr.press/v48/wange16.html.

[11] Wang, F., Mukherjee, S., Richardson, S., & Hill, S. M. (2020). High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. Statistics and Computing, 30, 697–719. https://doi.org/10.1007/s11222-019-09914-9

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. arXiv preprint, tarXiv:1603.02754v3

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen. W., Ma, W., Ye, Q., & Liu, T-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Proceedings of the 31st Conference on Neural Information Processing Systems, 3149-3157.

[14] Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. Frontiers in Plant Science, 22;10:621.

[15] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. Advances in Neural Information Processing Systems, 35, 507–520.

[16] Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

[17] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.

[18] Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Scientific Reports, 11, 11132.

[19] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variables importance by studying an entire class of prediction models simultaneously. arXiv reprint, arXiv:1801.01489v5

[20] Cassman, K. G., Dobermann, A., & Walters, D. T. (2002). Agroecosystems, nitrogen-use efficiency, and nitrogen management. Ambio. 31(2), 132-40.

[21] Beegle, D.B. (2009). Nutrient management. In Agronomy Guide. Penn State Extension. Pennsylvania State University.

[22] Meals, D.W., Dressing, S.A., & Davenport, T.E. (2010). Lag time in water quality response to best management practices: A review. Journal of Environmental Quality, 39(1), 85–96.

[23] Nielsen, R. L., Camberato, J., & Lee, J. (2019). Yield response of corn to plant population in Indiana. Agronomy Department, Purdue University.
[24] De Bruin, J. L., & Pedersen, P. (2008). Effect of row spacing and seeding rate on soybean yield. Agronomy Journal, 100(3), 704-710.
[25] Mourtzinis, S., & Conley, S. P. (2017). Delineating soybean maturity groups across the United States. Agronomy Journal, 109(4), 1163-1784.
[26] Bastidas, A.M., Setiyono, T.D., Dobermann, A., Cassman, K.G., Elmore, R.W., Specht, J.E., & Graef, G.L. (2008). Soybean sowing date: The vegetative, reproductive, and agronomic impacts. Crop Science, 48(2), 727–740.
[27] Egli, D. B., & Cornelius, P. L. (2009). A regional analysis of the response of soybean yield to planting date. Agronomy Journal, 101, 330-335.
[28] Darby, H. M., & Lauer, J. G. (2002), Planting Date and Hybrid Influence on Corn Forage Yield and Quality. Agronomy Journal, 94, 281-289.
[29] Eck, H. V., Mathers, A. C., & Musick, J T. (1987). Plant water stress at various growth stages and growth and yield of soybeans. Field Crops Research, 17(1), 1-16.
[30] Boyer, J.S., Byrne, P., Cassman, K.G., Cooper, M., Delmer, D., Greene, T., et al. (2013). The U.S. drought of 2012 in perspective: A call to action. Global Food Security, 2(3), 139–143.
[31] Nielson, R. L., & Thomison, P. (2003). Delayed planting & hybrid maturity decisions. Purdue University Cooperative Extension Service, Corn, AY-312-W.
[32] Setiyono, T.D., Weiss, A., Specht, J., Bastidas, A.M., Cassman, K.G., & Dobermann, A. (2007). Understanding and modeling the effect of temperature and daylength on soybean phenology under high-yield conditions. Field Crops Research, 100(2–3), 257–271.
[33] Zinn, K.E., Tunc-Ozdemir, M., & Harper, J.F. (2010). Temperature stress and plant sexual reproduction: Uncovering the weakest links. Journal of Experimental Botany, 61(7), 1959–1968.
[34] Horton, R., et al. (2014). Chapter 16: Northeast. In Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program.
[35] Soil Survey Staff. (2022). Keys to soil taxonomy, 13th Edition. USDA Natural Resources Conservation Service.
[36] Sadras, V. O., & Calderini, D. F. (2015). Crop physiology: Applications for genetic improvement and agronomy, 2nd Edition. Academic Press.
[37] Obalum, S. E., Chibuike, G. U., Peth, S., & Ouyang, Y., (2017). Soil organic matter as sole indicator of soil degradation. Environmental Monitoring and Assessment, 189:176.
[38] Weil, R.R., & Brady, N.C. (2016). The Nature and Properties of Soils (15th ed.). Pearson Education, Inc., New York, NY.
[39] Staver, K.W., & Brinsfield, R. B. (2001). Agriculture and Water Quality on the Maryland Eastern Shore: Where Do We Go from Here? BioScience, 51(10), 859-868.
[40] Turner, J. S., Friedrichs, C. T., Parrish, D. B., & Fall, K. A. (2026). Chesapeake Bay water clarity: Challenges and successes. Annual Review of Marine Science, 18(1), 89-119. https://doi.org/10.1146/annurevmarine-040224-120528
[41] Peng, B., Guan, K., Tang, J. et al. (2020). Towards a multiscale crop modelling framework for climate change adaptation assessment. Nature Plants, 6, 338–348. https://doi.org/10.1038/s41477-020-0625-3
[42] Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P., & Stouffer, R.J. (2008). Stationarity is dead: Whither water management? Science, 319(5863), 573–574.
[43] USDA-NASS. (2024). Crop Production Annual Summary. National Agricultural Statistics Service, U.S. Department of Agriculture. Washington, D.C.