Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Beyond Correlation: A Causality-Driven Machine Learning Framework to Climate-Crop Yield Data Using Transfer Entropy and Lag Optimization Approaches

Julius Peter Odeke

j.p.odeke@gu.ac.ug

Gulu University

Benard Abola

Gulu University

Benedict Oyo

Gulu University

Olok Geoffrey Tabo

Gulu University

Lone Dirckinck-Holmfeld

Aalborg University

Research Article

Keywords: Climate change, Causal Inference, Transfer entropy, Lag Optimization, Machine Learning, Crop yield prediction

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11459524/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Beyond Correlation: A Causality-Driven Machine Learning Framework to Climate-Crop Yield Data Using Transfer Entropy and Lag Optimization Approaches

Corresponding author: Julius Peter Odeke[a*] - j.p.odeke@gu.ac.ug

Co-Corresponding author: Benard Abola[a] - benard.abola@gu.ac.ug

Benedict Oyo[a] - b.oyo@gu.ac.ug

Olok Geoffrey Tabo[a] - g.tabo@gu.ac.ug

Lone Dirckinck-Holmfeld[b] - lone@ikk.aau.dk

[a] Gulu University, Gulu City, Uganda, [b] Aalborg University, Copenhagen, Denmark

ABSTRACT

Climate change increasingly threatens agricultural production globally by altering growing seasons, precipitation and temperatures. Understanding and predicting crop yields under these changing conditions is critical for food security. However, traditional climate-crop yield models rely on correlational approaches that are incapable of distinguishing causality or identifying optimal temporal lags. These correlational approaches result in spurious predictions and limited mechanistic understanding. We designed a three-stage machine learning framework that integrates causal analysis, lag optimization, and ensemble modelling to predict maize (*Zea mays*) and soybean (*Glycine max*) yields using county-level data from Ohio from 1981-2020. Transfer Entropy (TE) analysis was applied to 24 climate variables from stage 1, including monthly minimum, maximum, and mean temperatures and monthly precipitation (May through October). We also selected the 10 features with the highest causal strength for each crop, based on dual significance criteria ($p < 0.05$ or $TE > 0.01$). The results showed high causal relationships: maize (mean TE = 1.62, range: 1.54-1.74) and soybean (mean TE = 1.36, range: 1.29-1.47). Stage 2 used univariate lag-effect analysis, testing lags of 0-3 years for each selected feature using R^2 maximization. It showed that the temporal patterns of maize features were different, 50% of which had no lag (lag-0) and 20% had a 3-year lag (lag-3). Soybean responded with 50% immediate (lag-0) features and 30% lagged (lag-3) features. Maize is the best performer with minimum temperature in June ($R^2 = 0.10$, 3 years lag); while the soybean predictions are best with maximum minimum temperature in October ($R^2 = 0.11$, direct effect). In the Stage 3, four machine learning algorithms were evaluated with optimally lagged features. Random Forest performed better: maize ($R^2 = 0.75$, RMSE = 0.83 Mt/ha) and soybean ($R^2 = 0.72$, RMSE = 0.27 Mt/ha). Temperature extremes at key growth stages became primary drivers and not just correlates of the causal drivers. This TE-LO framework goes beyond just spurious correlations in agricultural forecasting, going so far as to incorporate mechanistically significant relationships.

Keywords: Climate change, Causal Inference, Transfer entropy, Lag Optimization, Machine Learning, Crop yield prediction

1. INTRODUCTION

1.1 Background

The effects of global food security are more pronounced than ever, as extreme weather events, changes in precipitation patterns, and elevated temperatures are increasing the frequency and severity of impacts on agricultural systems (IPCC, 2023; Ray et al., 2019). Crop yield forecasting has become critical for anticipating food security planning, managing agricultural risks, and climate adaptation (Dhillon et al., 2024). About 60% of global crop production is affected by climatic variability in yields, and losses in recent years have exceeded $100 billion (FAO, 2022; Jägermeyr et al., 2021).

Agriculture is climatically sensitive. Crop growth and development are affected by the temperature regime, precipitation distribution, and precipitation during the growing period (Schauberger et al., 2020; Zhu et al., 2022). Extreme temperatures during key phenological intervals (especially flowering and grain filling) may decrease yields by 10-50%, and precipitation shortages or surpluses may lead to yield losses by causing water stress or waterlogging (Lesk et al., 2022; Zampieri et al., 2019). These relationships are important because they enable the development of robust forecasting models and the identification of adaptive management approaches (Jägermeyr et al., 2021).

The state of Ohio is a perfect study area due to three reasons. First, as the 10th-largest producer of maize and soybeans in the US Corn Belt, with total output valued at 6.8 billion per year (15th in the state's agricultural economy), it is a system where better yield forecasts can have a significant effect on the economy (USDA-NASS, 2023). Second, the topographic and climatic heterogeneity in Ohio is significant, as mean annual rainfall amounts (900-1100mm) and growing-season temperatures (18-22 °C) differ spatially across regions, resulting in a spatially diverse relationship between climate and crops. The other factor is the availability of an open-access dataset of maize and soybean yields spanning over 40 years, which is vital for training ML Models. This diversity offers a natural experiment to test procedures for recording non-stationary, complex climate-yield interactions that cannot be identified using correlation-based approaches. Third, Ohio is an example of climate pressures on temperate grain production worldwide: warming of 0.8°C since 1980, precipitation variability doubled, and heat stress impacts the reproductive stage more often (Baule et al., 2020). The methodological advances described herein are therefore also applicable to the broader Corn Belt and to similar temperate agricultural economies worldwide that confront similar climate conditions.

The more typical feature selection processes used in agricultural forecasting are based on correlation-based approaches, which find statistical relationships without determining the directional cause-and-effect relationships between climatological variables and crop yields (Runge et al., 2019). Transfer Entropy (TE), proposed by Schreiber (2000), addresses this shortcoming by quantifying the directional flow of information:

$$TE_{X \rightarrow Y} = \sum p(y_{t+1}, y_t^k, x_t^l) \log [p(y_{t+1} \mid y_t^k) / p(y_{t+1} \mid y_t^k)],$$

thereby, quantifying the decrease in uncertainty of the target variable Y when the past states of the source variable X are known beyond their history (Bossomaier et al., 2016). More recent studies show that TE has been successful in identifying causal drivers in climate research by overcoming hidden confounding and temporal dependencies that impede correlation-based analysis (Krich et al., 2020; Runge et al., 2023). However, TE-based causal discovery lacks optimization of temporal lag structures, which is vital, as the effects of climate on crops persist into the next growing season due to soil moisture persistence and nutrient legacy (Lesk et al., 2022; Zampieri et al., 2019). Most of the literature presupposes modern climate-yield relationships (Lobell et al., 2011), resulting in models that predict with a 15.25% loss of accuracy in new settings (Rosa et al., 2024). The current work thus constructs a causality-based structure that combines TE-based feature selection and personalized lag optimization, and implements the resulting methodology on 40 years of corn and soybean yields in Ohio to capture the much-needed temporal dynamics of the interactions between climate and crops.

1.2 Research Gap

Traditional models of climate-crop yield are primarily based on correlational methods and stationary distribution assumptions, such as regression analysis, correlation coefficients, and machine learning models that are being trained on the associative patterns. (Crane-Droesch, 2018; Shahhosseini et al., 2023). Nevertheless, the distinction between correlation and causation is also a common concept that is often disregarded in the literature on agricultural forecasting (Leng, 2022; Runge et al., 2019). The spurious relationships between climatic variables and yields are often due to temporal coincidence, confounding factors such as technological changes and shifts in management without a mechanistic connection, and the effects of multiple studies and testing without significance control (Iizumi et al., 2018).

Such correlational approaches have three severe limitations. To start with, they are unable to distinguish between actual causal effects and statistical artefacts, leading to models that incorporate non-causal predictors and perform poorly under evolving climatic conditions (Krich et al., 2020; Runge et al., 2019). Second, feature selection based on correlation often uses arbitrary thresholds (e.g., the top-N features) that lack statistical support, potentially eliminating strong predictors or including spurious ones (Kuhn & Johnson, 2019) and significantly increasing computational overhead (Gu et al., 2022). Third, the majority of studies presuppose

contemporaneous (lag-0) or lag-1 conditions, thereby overlooking optimal temporal structures in which climate effects are delayed by soil moisture memory, cumulative stress effects, or phenological responses (Zampieri et al., 2019).

The latest developments in information theory and causal inference can offer valuable insights to determine relevant features and directional causality in complex systems (Krich et al., 2020; Runge et al., 2019). TE, a model-free measure of information flow between variables, in say a multivariate time series. It is the measure of how much the future value of a target variable can be predicted by a source variable in terms of its history, in addition to that which is already known by the history of the target variable (Bossomaier et al., 2016; Schreiber, 2000). TE, unlike Granger causality, does not presuppose linear connections but rather non-linear causal relations, which are typical of climate-agricultural systems (Goodwell & Kumar, 2017; Runge et al., 2019). Although TE offers theoretical benefits, it is not widely used in agricultural practices, and fewer than 5% of new crop modelling studies employ information-theoretic causal tests (Leng, 2022; Sultan et al., 2019). Another aspect of the climate-crop yield relationship is the temporal lag effect, which is one of the most important but less-researched aspects. Crop yield varies according to climate conditions, with variable time lags as a result of the physiological memory, soil moisture carry-over, and buildup of stress (Lesk et al., 2022; Vogel et al., 2019). For example, late-season yields can be affected by early-season soil moisture, which interferes with root growth, whereas antecedent temperature state affects phenological timing, which cascades into reproductive success (Zampieri et al., 2019).

Nevertheless, the systematic lag optimization remains underdeveloped in the literature, with few studies employing ensemble approaches that combine multiple techniques (e.g., stationarity testing, cross-correlation analysis, and predictive importance assessment) across multivariate climate systems (Chlingaryan et al., 2018; van Klompenburg et al., 2020). The majority of the available methods either: (i) use fixed lag structures using crop phenology (Roberts et al., 2017), (ii) test a small range of lag periods (0-1 years) on a case-by-case basis (Lobell & Burke, 2010), or (iii) use only correlation-based methods that may miss non-linear temporal dependencies (Vogel et al., 2019).

1.3 Objectives of the Study

The specific objectives of this study are:

(i) To determine climate variables with a positive influence on crop yields.
(ii) To determine the optimal temporal lag for each causally verified climate feature.
(iii) To compare four machine learning algorithms for predictive performance using the Transfer Entropy-Optimized Lags (TE-OL) Approach

2. METHODS AND DATA

2.1 DATA SOURCES AND DESCRIPTION

Maize and soybean yield data from 88 counties in Ohio (Mt ha-1) were retrieved from the USDA-NASS Quick Stats database for 1981-2020, yielding 3,360 county-year observations (USDA-NASS, 2025). The NASA POWER agroclimatology database was used to obtain corresponding monthly climate data of the growing season (May to October). Four climate data variables, including minimum (T_min), maximum (T_max), and mean (T_mean) temperature (°C), and total precipitation (P) (mm), were obtained per county and month. These indicators include essential biophysical activities, such as night respiration and chilling (T_min), day photosynthesis and heat stress (T_max), total thermal regime (T_mean), and water supply (P), which determine the development of crops (Lesk et al., 2022; Maiorano et al., 2024). Therefore, Table 1 presents a summary of 24 climate variables from every county-year observation resulting from 6 months times 4 measures.

Table 1: Study Variables

Category	Variable	Description	Count
Climate Variables	T_min	Monthly minimum temperature (°C)	6
	T_max	Monthly maximum temperature (°C)	6
	T_mean	Monthly mean temperature (°C)	6
	P	Monthly precipitation (mm)	6
		Subtotal: May-October (Months 5-10)	24
Target Variables	Y_maize	Maize yield (Mt/ha)	1
	Y_soybean	Soybean yield (Mt/ha)	1
		Total Variables	26

Note: Climate variable naming convention: Month_Type (e.g., 5_Tmin = May minimum temperature)

2.2 METHODS

In this section, a 3-stage methodological framework is presented which extends correlation to the identification of causation in the field of climate-crop yield modelling (Fig. 1). At every step, particular statistical conditions and cutpoints are used to guarantee that variables are chosen and models are created with respect to causality, instead of a spurious correlation..

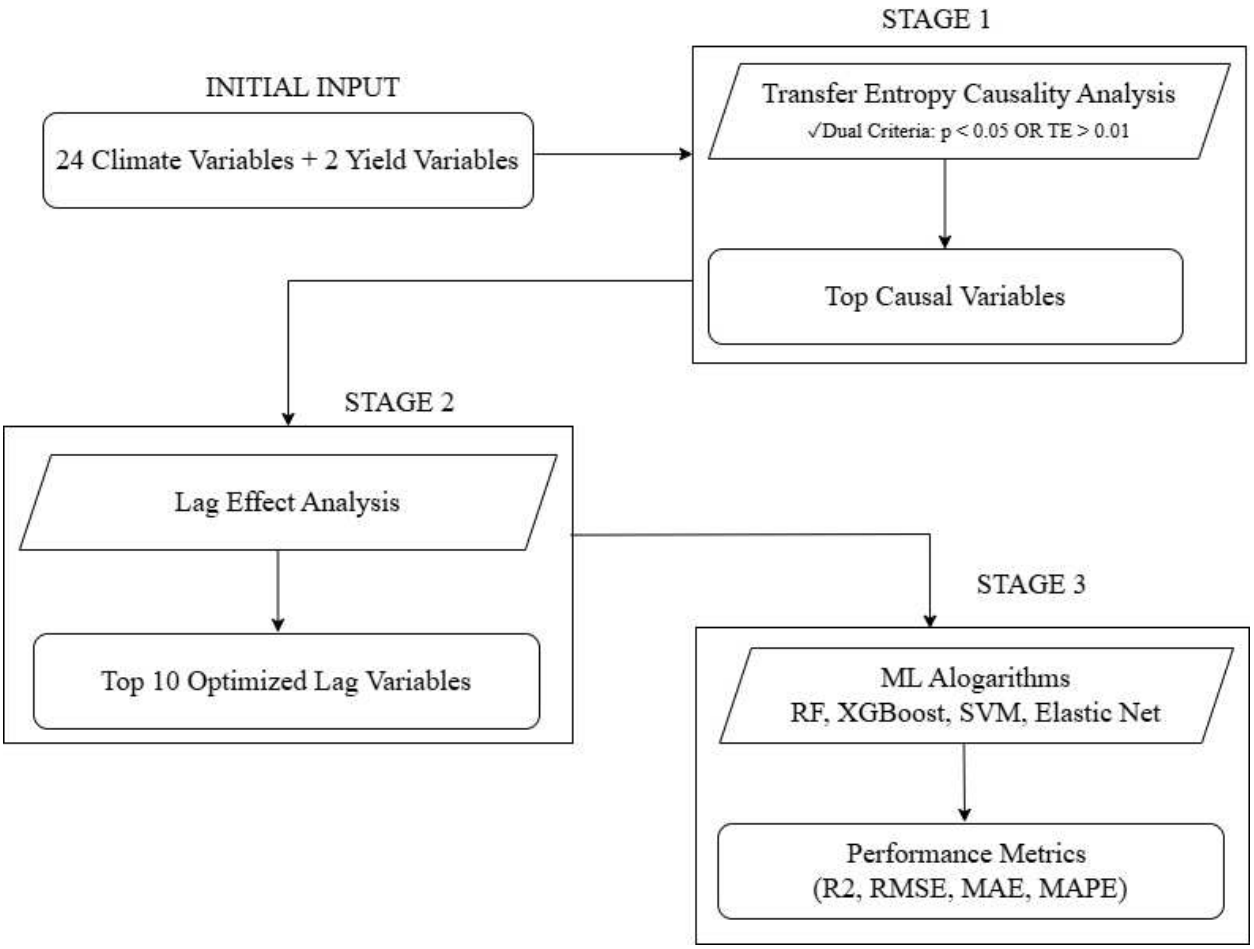


Figure 1: Transfer Entropy-Optimized Lags Framework for climate-crop yield modeling.

Stage 1: Transfer Entropy Feature Selection

To identify climate variables that contain causal information into crop yields, Transfer Entropy (TE) analysis and Permutation testing (n = 100) are applied in stage 1. TE is the reduction of uncertainty about future states of yield as a result of information about previous states of climate and knowledge of past yield values (Schreiber, 2000). It is an approach that differs from the mere correlation between climate variables (sources) and yield (target) by quantifying the information flow between the two.

Each of the 24 climate variables is compared with the yields of maize and soybean separately. The features are retained when they fulfill two criteria of dual significance: permutation test p-value < 0.05 OR TE value > 0.01. The p-value limit complies with the standard statistical practice (p = 0.05), and the TE limit (0.01 bits) is significant information channeling in agricultural networks (Ruddell and Kumar, 2009). Each crop is selected based on the top 10 features with the highest TE

values, rather than arbitrary cutoffs. This step uses the full time series (1981-2020) without a train-test split, as it is crucial to maintain temporal dependencies that support causal inference.

Stage 2: Temporal Lag Optimization.

Stage 2 identifies the optimal temporal relationships for each feature that have been causally validated using univariate lag-effect analysis. Climate memory effects are statistically identified by testing lags of 0-3 years of each of the 10 features selected. Each feature-lag combination is fitted with vector autoregression models, and its performance is evaluated using R^2 and RMSE on the complete data.

County-specific lag testing is used to avoid data leakage across spatial units. Features are lagged in each county's time series and then concatenated. This method accounts for possible variation in climate-yield response dynamics across geographic conditions in Ohio. The maximum R^2, which indicates the strongest single-feature predictive relationship, is used to determine the optimal lag for each feature. This collective opinion method, based on the adaptation of Papacharalampous et al. (2019), will ensure the maximization of time structures prior to the final stage of the approach.

Stage 3: Predictive Modeling and Tests.

Stage 3 trains and assesses four machine learning models, including Support Vector Machine - Radial Basis Function (SVM-RBF), Random Forest (RF), Gradient Boosting (GB), and Elastic Net (ENet), using the final dataset of causally validated features and optimized time lags. The SVM with the Radial Basis Function (RBF) kernel was used. The RBF kernel is selected since it has been shown in the agricultural forecasting literature (Khaki & Wang, 2019) that it captures non-linear relationships between climatic variables and crop yields. To ensure the full variability of climatic conditions and provide accurate results, the data set was split, using stratified random sampling, into a test data set (20%) and a training data set (80%).

Model training is done using the lagged features from the results of Stage 2. All of them are standardized (zero mean, one unit standard deviation) before training the model, so that the temperature and precipitation variables have similar scales. To prevent overfitting, the cross validation is performed on the training set with 5-fold cross validation to evaluate the model's generalization. The final model is then tested on the held-out test set, and metrics include the coefficient of determination (R2) and root mean square error (RMSE), as well as cross validation scores.

These sequential steps ensure that only those variables that have been causally tested on and are found to be statistically significant, and those variables that have been temporally optimized to be the most relevant for the predictive models are used. This method gives mechanistic assurance that the features the algorithm selects are causal drivers of yields as opposed to features which can be spuriously correlated with yield as in traditional correlation based methods. This framework is

especially helpful in analyzing non-stationary climatic situations in which past correlations don't necessarily indicate causal processes.

3. RESULTS

This section presents the results of the TE-OL framework applied to maize and soybean yield prediction in Ohio. Results are organized according to the three study objectives as stated earlier.

3.1 Transfer Entropy Features Selection on Climate Variables

Transfer Entropy (TE) analysis was applied to all 24 climate variables to identify features exhibiting positive information flow to crop yields. Using dual significance criteria ($p < 0.05$ OR $TE > 0.01$), the top 10 features with the strongest causal influence were selected for each crop.

In maize, all the 24 climate variables had significant information transfer (Figure 2). The top 10 features that were selected were strongly causally related to mean TE = 1.62 (range 1.54-1.74). The information transfer statistics (TE) were highest for May precipitation (5P) (TE = 1.74, p = 0.01), followed by May and June temperature variables. The temperature variables during the growing season (May-June) were the most dominant in the selection, with 6 of 10 features taking place during these critical months of establishment (Figure 2).

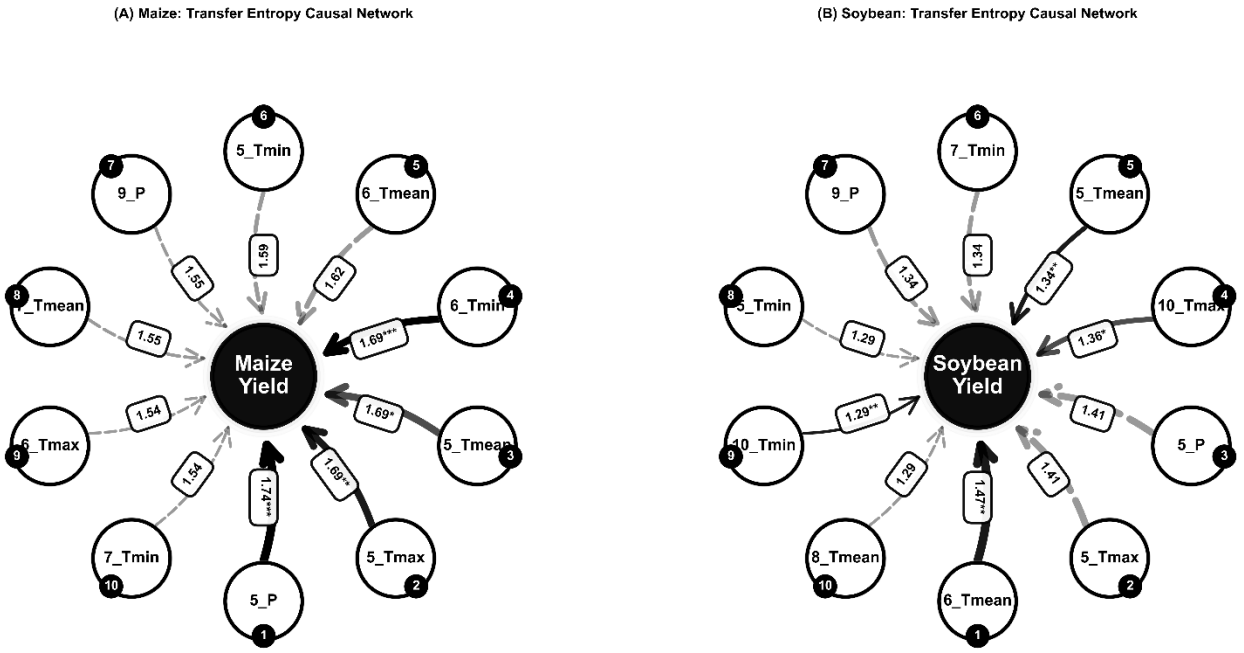


Figure 2. Circular Transfer Entropy (TE) Directed Acyclic Graphs showing causal information flow from climate variables to crop yields in Ohio (1981-2020). All nodes are white with black edges (no color coding). Arrow width indicates TE magnitude, while line style and opacity show

statistical significance (*** p≤0.01, ** p≤0.05, * p≤0.10). Numbers at node borders indicate feature rank by causal strength (1 = strongest).

The time-series analysis revealed that the effect of temperature is predominant in the early growing season (May-June) and causes the maize yield. Of the selected predictors, the temperature variables accounted for 90% of the predictors and one precipitation variable (May) was included in the top ten variables. In addition, 9 of the 10 most important predictors exhibited optimal lag-one transfer of information, suggesting that interannual climatic conditions influence modern yields through soil moisture carryover, pathogen persistence, or nutrient cycling mechanisms.

The top 10 causal variables in the soybean data generated a mean transfer entropy (TE) of 1.36 and a range of 1.29 -1.47. This is slightly lower than the value that was observed with maize, but it still shows strong information transfer (Figure 2).

The causal impact of the mean June temperature (6_Tmean) had the largest effect, with a TE of 1.47 and a p-value of 0.02, followed by the mean temperatures in May and October. Unlike maize, the causal effect of soybean is more time-varying during the growing period, with the most pronounced period occurring from May to October.

Soybean showed greater sensitivity over time than maize, with late-season variables (August, September, October) accounting for 40 percent of the selected attributes. This trend indicates the indeterminate growth nature of soybeans and the crop's prolonged reproductive cycle. Once more, temperature variables led the feature set (80%), but only two precipitation variables were selected. Interestingly, 60 percent of the features had optimal lag-2 information transfer, indicating greater effects of multi-year climatic memory in soybean than in maize.

The results from the causal structure analysis of crops revealed different sensitivities (Figure 3). The mean transfer-entropy score (1.62 vs. 1.36) for maize was higher, reflecting the overall higher climate dependence. Concentrations were higher in maize (60% of features in May-June) compared to soybean (40%). On the other hand, soybeans were more vulnerable to conditions in later seasons (40% of features in August-October compared to 10% in maize). The trends are consistent with known physiological differences, such as determinate growth and a limited pollination period for maize, and indeterminate growth and reproductive flexibility for soybean.

.

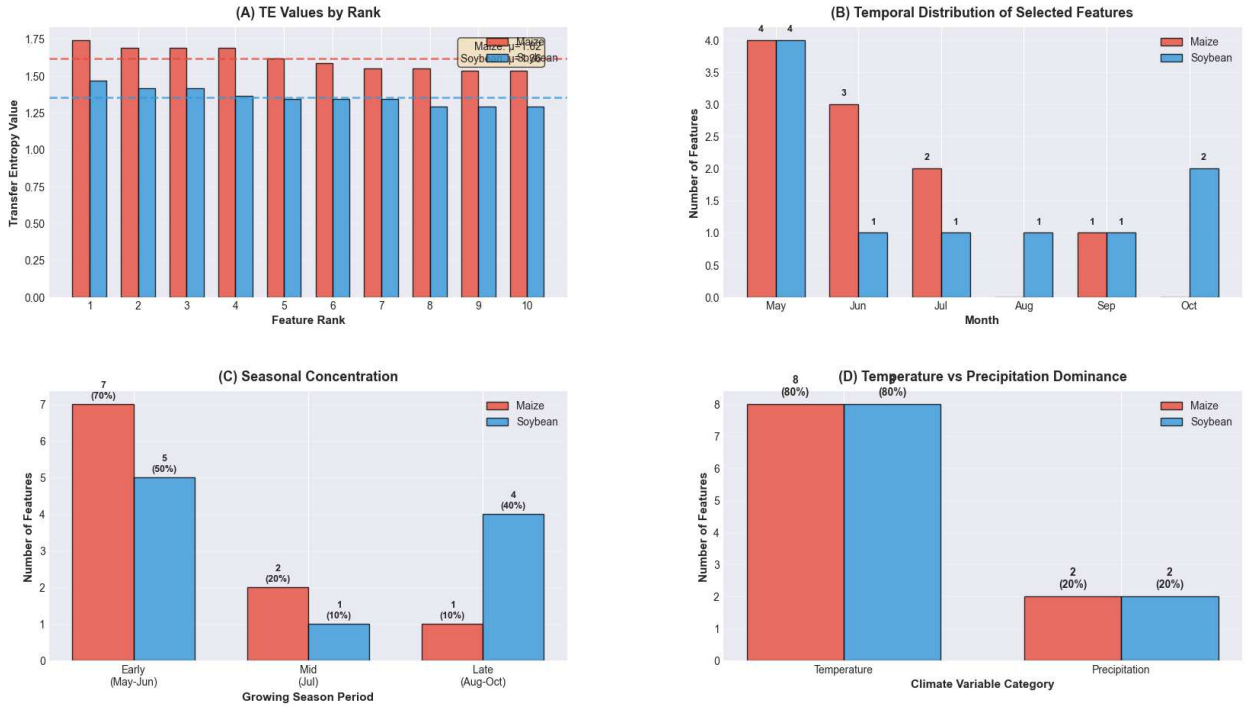


Figure 3: Crop-Specific Causal Patterns from Transfer Entropy Analysis

3.2 Objective 2: Optimization of Temporal Lag Structure.

Univariate lag-effect analysis, feature-by-feature, was conducted for the 10 causally validated features per crop, using lags of 0-3 years to determine the best temporal relationships. Each feature-lag combination was fitted with a simple linear regression model to determine the best lags by maximizing R^2.

3.2.1 Maize Lag Structures

There were different temporal structures of maize features within the 0-3 year lag spectrum (Figure 4). The lag distribution consisted of 50% immediate effects (lag-0), 30% at 1-year lag (lag-1), and 20% at 3-year lag (lag-3), with no features having an optimal lag of 2. The lowest possible temperature was observed in June (6_Tmin), with a 3-year lag effect as the strongest univariate predictor, accounting for 10.5% of the yield variability. This climatic memory over time implies that the stress caused by cold at the early part of the season has long-lasting impacts on future yields, which may manifest as changes in soil structure, pathogen interactions, and nutrient cycling mechanisms. The regression coefficient of 0.42 is positive, indicating that higher current yields are associated with warmer minima in June 3 years ago.

The number of contemporary effects (lag-0) of five climatic variables included the minimum temperatures (5_Tmin, 5_Tmean) and the mean temperature (7_Tmean) of May and July. The

direct effects of planting conditions and seedling emergence success are reflected in the instantaneous relationship between May temperatures and crop performance. Contrary to this, the July mean temperature had a negative coefficient (-18), indicating that yields become lower when there is a surplus of heat during pollination- a better-known heat-stress effect on maize (Lobell et al., 2011).

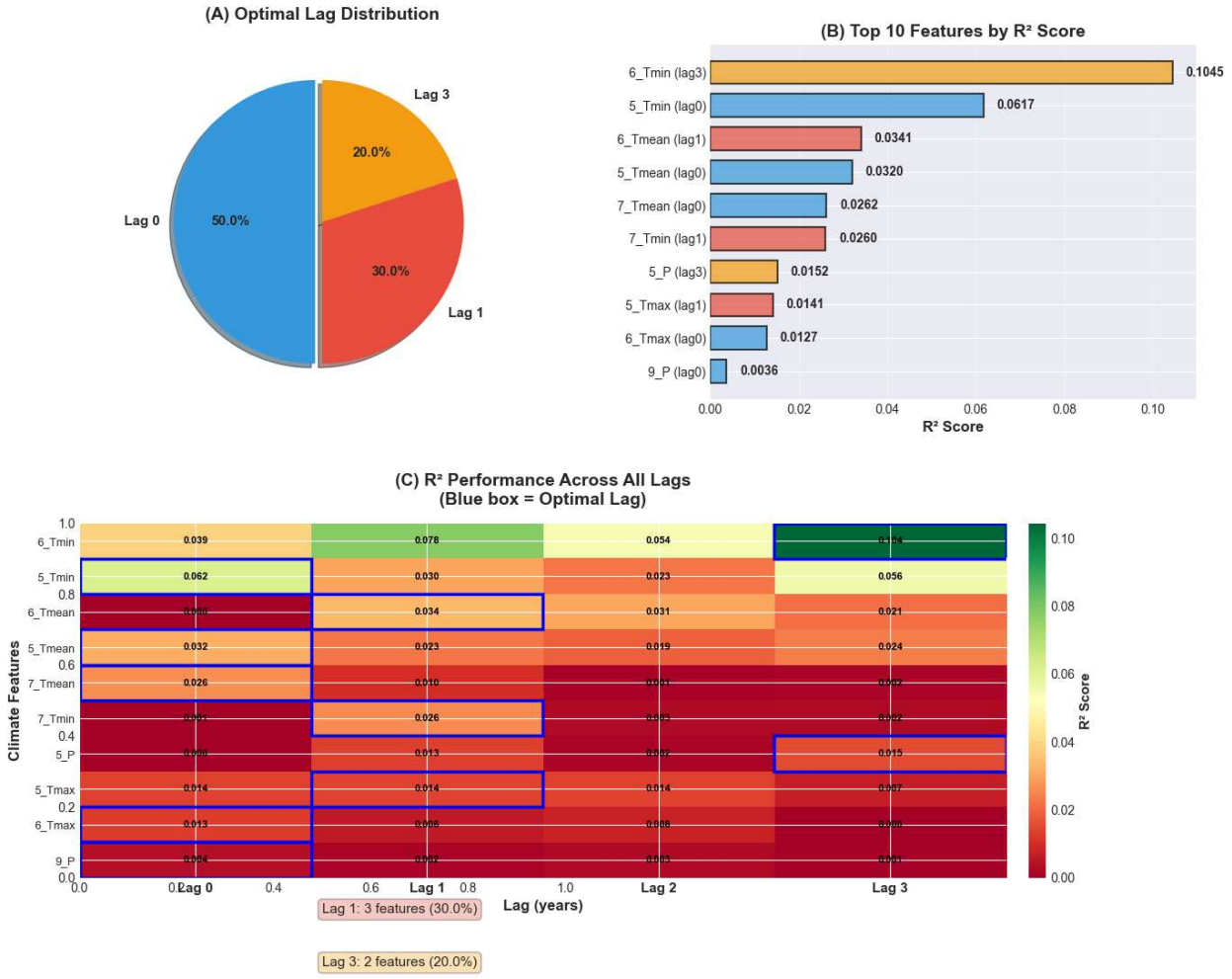


Figure 4: Optimal Temporal Lag Structures for Maize showing lag distribution, R^2 performance, and feature-specific patterns

Mean explanatory power fell with lag in all predictors: R^2 of 0.0273 (lag-0), 0.0247 (lag-1), 0.0215 (lag-2), and 0.0598 (lag-3). This strong effect at lag-3, largely driven by 6_Tmin and 5_P, indicates that although most climatic effects on the season occur instantaneously or within a year, there are specific early-season variables with multi-year memory.

3.2.2 Soybean Lag Structures

The soybean time series differed from those of maize (Figure 5). Distribution of lagged effects indicated that 50 percent of the significant variables were immediate (lag-0), 20 percent showed a two-year lag (lag-2), and 30 percent showed a three-year lag (lag-3), with no variables showing the best effects at lag-1.

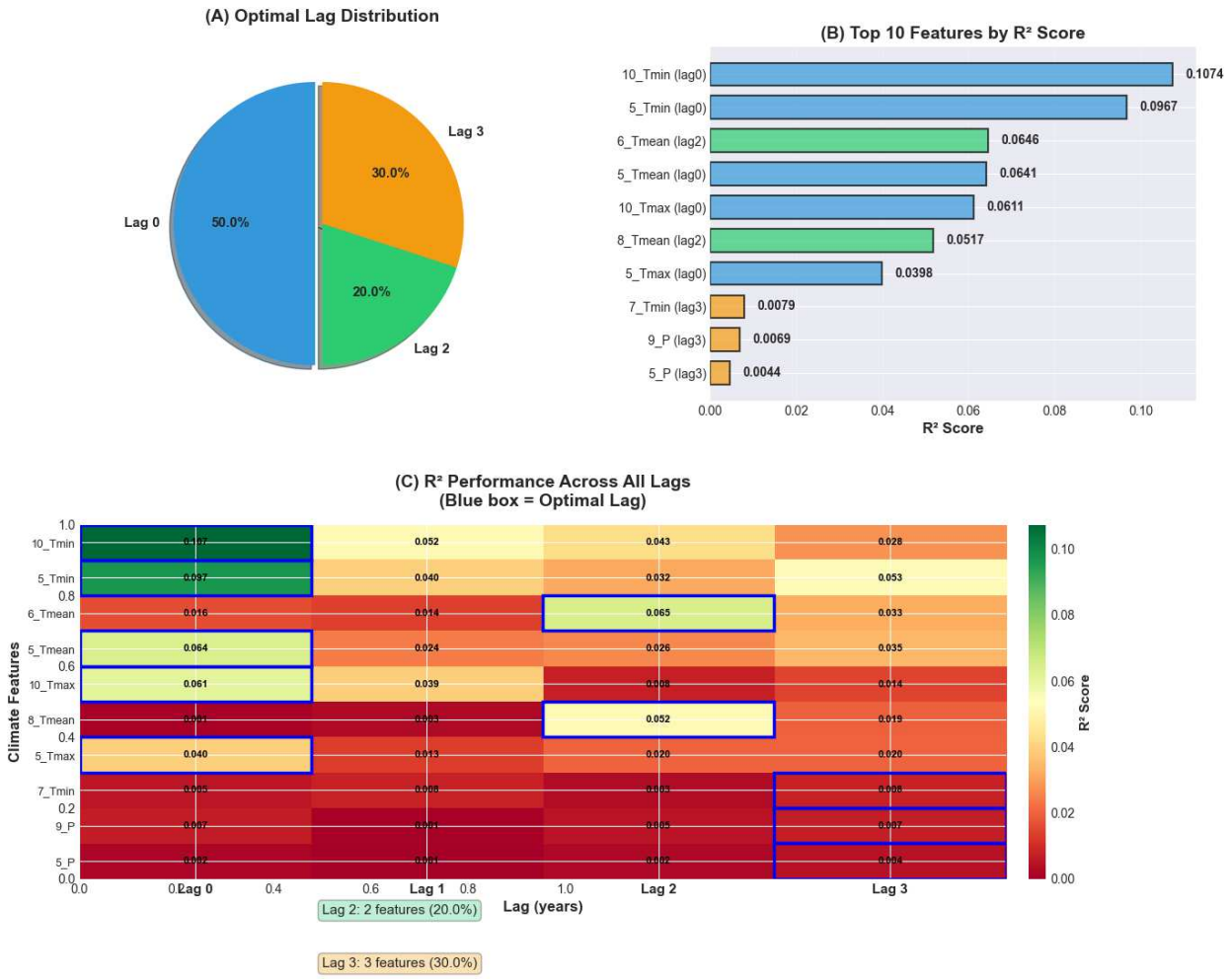


Figure 5: Optimal Temporal Lag Structures for Soybean showing lag distribution, R^2 performance, and feature-specific patterns

Immediate (lag-0) October minimum temperature (10_Tmin) proved to be the strongest univariate predictor (R^2 0.1074) and explained the soybean yield variance of about 10.7. This modern-day relationship suggests direct effects of fall temperatures on seed filling, maturation, and harvesting time. The positive coefficient (0.10) indicates that higher October temperatures are associated with higher yields, as they lengthen the growing season and maximize seed germination.

There were five features with direct effects (lag-0), all temperature variables spanning the entire growing season (May, October). The lag-0 response concentration indicates that soybean is more

susceptible to within-season climatic variations than lagged responses. However, the frequency of three-year lags (7_Tmin, 9_P, 5_P) was 30%, indicating significant multi-year memory of the climate for these variables.

Two variables showed the best two-year lags (6_Tmean, 8_Tmean), both of which are mean temperature variables that occur in the middle of the season. This intermediate temporal response can reflect soil moisture reserves, the pest population, or nutrient availability being cycled by mid-season climatic conditions.

Lagged mean performance showed a much smoother distribution of R^2 across temporal scales (lag-0: 0.0738; lag-1: 0.0647; lag-2: 0.0581; lag-3: 0.0064), and the decrease was less pronounced than in maize. These findings imply that soybean yields blend climatic data more uniformly over time ranges, as it has an indeterminate growth pattern.

3.3 Machine Learning Algorithm Performance Comparison.

The four machine-learning algorithms were SVM-RBF, RF, GB, and ENet, and were trained and assessed on the 10 causally validated features with optimal temporal lags. The data (2,757 observations after lagging) were split into 80/20 percent for training (2,205) and testing (552) sets, and the features were standardized before training the model. The results for the performance comparison of these 4 machine learning algorithms are summarized in Figure 5.

(A) Model Performance Comparison: Maize vs Soybean

	Maize Performance		Soybean Performance	
Model	Maize Test R^2	Maize RMSE	Soybean Test R^2	Soybean RMSE
RF	**0.7523**	**0.8337**	**0.7172**	**0.2676**
GB	0.7417	0.8513	0.7165	0.2679
SVM	0.6882	0.9353	0.7093	0.2713
Enet	0.2238	1.4756	0.1823	0.4550

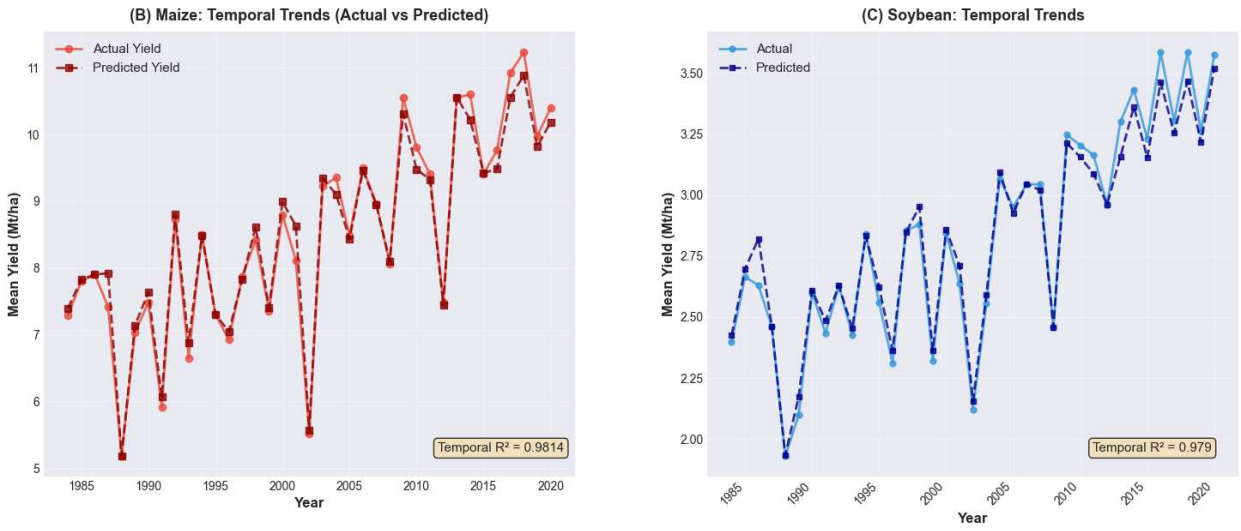


Figure 6: Performance comparison of 4 Machine Learning Algorithms on Maize and Soybean

3.3.1 Yield Prediction Performance for Maize.

All four algorithms were capable of learning relationships between optimally lagged climatic variables and maize yields, but with substantial performance differences (Figure 6; Figure 7).

RF showed better results across all measures, with a test R^2 of 0.7523, indicating it can explain 75.2% of the variance in yield. A 0.8337 Mt/ha test root mean square error is equivalent to an error of about 8.3% of the mean yield, which is approximately 10 Mt/ha. The test results were similar to those of cross-validation (CV R^2 = 0.7345 ±0.0160), suggesting strong generalization without signs of overfitting.

GB yielded similar results, with a test R^2 of 0.7417, which was 1.1 percentage points lower than that of Random Forest. The performance parallels indicate that the two ensemble tree-based methods can effectively model the non-linear relationship between climatic variables and yield. Nonetheless, Gradient Boosting had a slightly higher test RMSE of 0.8513 Mt ha -1, corresponding to slightly reduced predictive power.

The SVM achieved moderate performance, with a test R2 of 0.6882, explaining 68.8% of the variation; however, its RMSE was significantly larger at 0.9353 Mt/ha. The lower difference in the training ($R^2 = 0.7463$) and testing performance than the ensemble methods indicates that the SVM learnt more generalizable patterns but with less overall generalization capability.

ENet did not perform well, with a test R^2 of 0.224, explaining only 22.417% of the variance, compared to the RMSE of 1.4756 Mt/ha, which was almost double that of the best-performing models. This breakdown of the linear model suggests that there is significant nonlinearity in the climate-yield interaction that cannot be captured by linear combinations of features, despite L1/L2 regularization. The training-test performance gap provides a clue to a model's complexity. Random Forest had the highest discrepancy (train $R^2 = 0.878$ and test $R^2 = 0.752$; difference = 0.126), indicating it is slightly over-fitted, though it generalized well. This finding was supported by the model's cross-validation scores, with a CV R^2 of 0.735, which was quite close to the test performance, thereby increasing trust in out-of-sample predictions.

3.3.2 Maize Feature Importance Analysis

Feature Importance analysis with Random Forest feature importance shows how each optimally lagged climate variable contributes to predicting maize yield (Figure 7).

The largest predictive significance (16.5%), in agreement with the Stage 2 result, was found for the feature used to capture the univariate R^2, namely, the June minimum temperature with a three-year lag (6_Tmin_lag3). The persistence of this long-term climatic memory effect suggests that early-season cold temperatures have a lasting impact on subsequent yields. The combination of the five most significant predictors explained 66.6% of the overall explanatory power; thus, predictive power was concentrated in early-season thermal conditions.

Temperature variables comprised 90% of total importance (features 1-5, 7, 9-10), with only two precipitation variables (9_P, 5_P) contributing 16.4% combined. This temperature dominance is consistent with the maize physiology known to occur, as thermal conditions control development rate, pollination success and grain filling (Hatfield & Prueger, 2015). The lag distribution of important features was mixed: 50% lag-0, 30% lag-1 and 20% lag-3. Which is exactly what was found in Stage 2: an optimal lag distribution, as in the case of multivariate modeling the temporal relationships among individual features were preserved.

3.3.3 Yield Prediction Performance for Soybean.

Machine learning models showed patterns similar to those for maize with soybean, with ensemble methods substantially outperforming linear models (Table 2, Figure 8).

The Random Forest model again proved the most predictive, with a test coefficient of determination ($R^2 = 0.7172$) and explaining 71.7% of the variation in soybean yields. The root-

mean-square test-set error was 0.2676 Mt/ha, corresponding to an approximate prediction error of 7.9%, which, compared to the mean yield of about 3.4 Mt/ha, is slightly better than the result with maize. The results of cross-validation (CV $R^2 = 0.6914 \pm 0.036$) indicated strong generalizability, but the standard deviation was larger than that of maize, suggesting greater model sensitivity to the structure of the validation folds.

The GB model demonstrated almost the same performance (test $R^2 = 0.7165$) as RF, with only a 0.07 percentage-point difference, which was less than the difference in maize. The test RMSE changed by only 0.0003 Mt/ha, indicating functional similarity in predictive capability. This convergence implies that the relationships among underlying soybean yields are slightly less complicated than those of maize, and therefore, both ensemble methods can achieve similar performance limits.

Surprisingly, the SVM algorithm showed rather impressive performance (test $R^2 = 0.71$), which is only 0.79 percentage points lower than RF. Moreover, SVM had the smallest training-to-testing performance difference (= 0.0588) and the largest cross-validation mean (= 0.70), indicating greater generalization consistency. The very high performance of SVM in soybean, compared to the less marked performance in maize, suggests that the soybean-climate-crop yield relationship could be more favorable to kernel-based modeling.

ENet once again showed low predictive power (test $R^2 = 0.18$), explaining only 18.2% of the variability, and had an RMSE of 0.455 Mt/ha. However, ENet performance on soybean was relatively higher than on maize (18.2 R^2 vs 22.4 R^2), suggesting that the yield response to soybean is slightly more linear. However, the model is not sufficient to make accurate predictions.

The differences between training and test performance were generally identical across the models: RF (0.146), GB (0.145), and SVM (0.059). The larger gap in RF is due to the model's complexity, while the smaller gap in SVM is due to its more conservative pattern learning.

3.3.4 Feature Importance Analysis for Soybean

RF feature importance for soybean revealed distinct patterns compared to maize (Figure 7).

The immediate influence of the minimum temperature in October (10_Tmin lag-0) proved to be the most significant factor, explaining 18.8 percent of the variance, confirming that this variable was identified by Stage 2 as having the strongest univariate R^2. This modern-day relationship results from the direct impact of late-fall temperatures on seed filling and maturation. Together with the first 5 features, they accounted for 69.9% of the total variable importance, which is as high as in maize, but the temporal emission pattern is also different.

Temperature variables have more importance in the distribution of importance than in maize, as they represent 80% of the total value, while two variables of precipitation (9_P and 5_P) contribute

only 8.0, which is less than in maize. A very high level of temperature dominance, especially since a high sensitivity of soybean to water stress in the reproductive stage has been reported when compared to maize (Specht et al., 2001).

The temporal distribution showed that 60% of the explanatory power was due to immediate effects (lag-0), with only 20% due to maize. The dominance of the lag-0 features (10_Tmin, 5_Tmin, 10_Tmax, 5_Tmean, 5_Tmax), suggests that climatic conditions within a growing season are more important to soybean production than are lagged climatic conditions. Two year lags (6_Tmean, 8_Tmean) and three year lags (9_P, 7_Tmin, 5_P) were also significant (at 20 percent), so there is a large multi-year memory component.

It was found that the seasonal importance distribution differed significantly from that of maize. The share of late-season variables (August, September, October) accounted for 30% of all influence, compared with 10% for maize, due to soybean's extended reproductive cycle. On the other hand, the dominance in the early season was lower (May-June: 50% compared to maize, 70%), indicating a more distributed seasonal sensitivity in soybean.

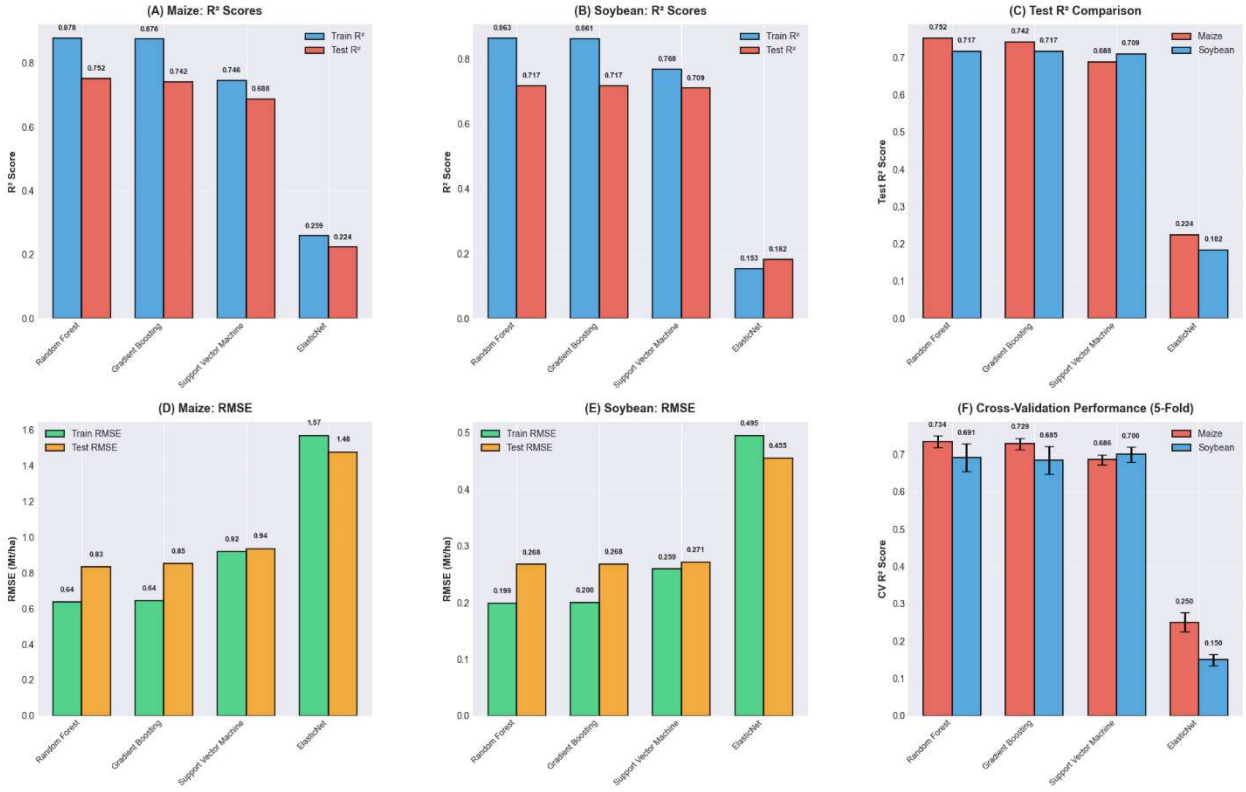


Figure 7: Machine Learning Model Performance Comparison - Maize vs Soybean

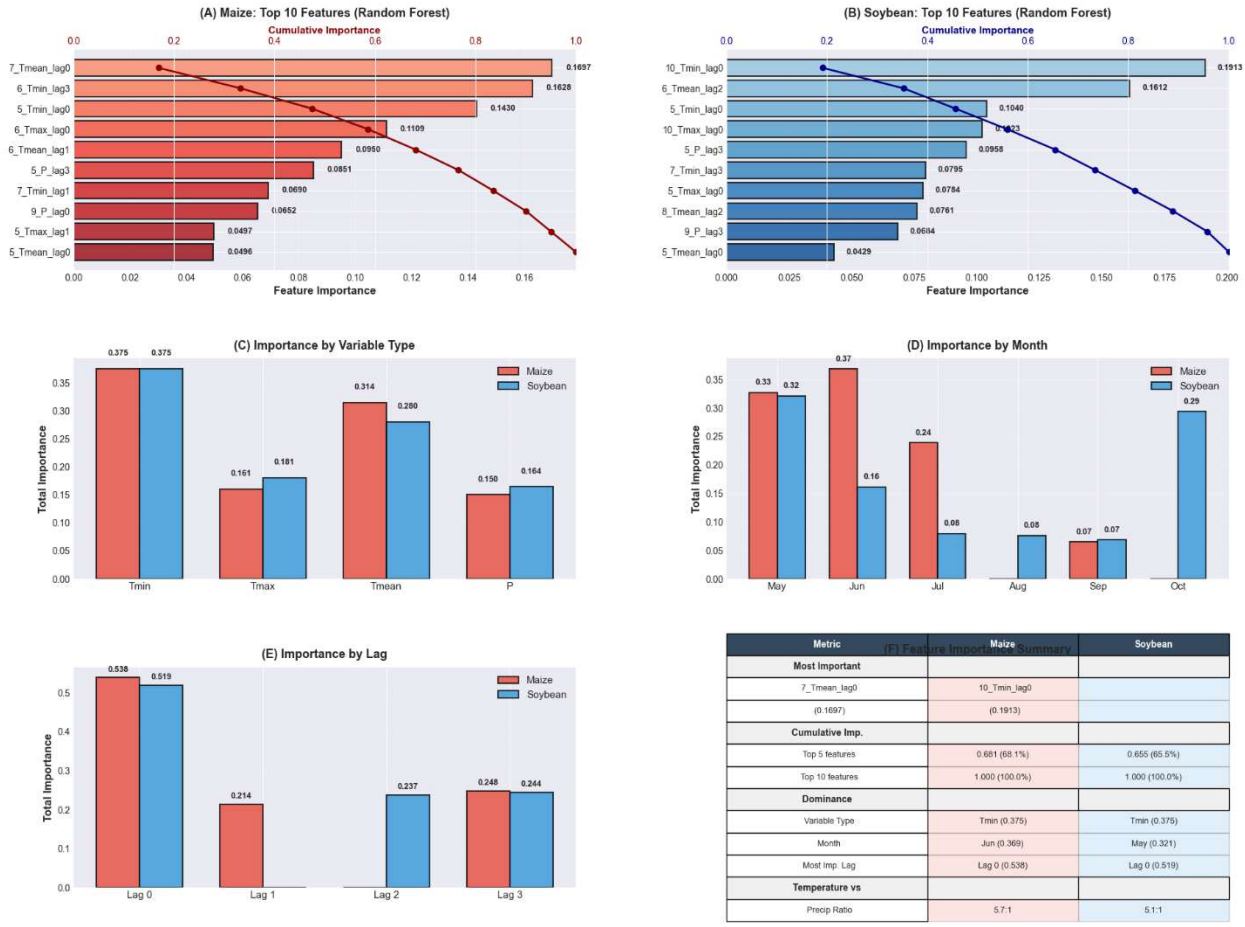


Figure 8: Feature Importance Analysis - Best Performing Models

3.3.5 Comparative Model Performance

The patterns and important differences were consistent across the crop types (Figure 9). Random Forest was the best model for both crops as indicated by the higher absolute R^2 for maize (0.7523) than for soybean (0.7172). The accuracy was similar when based on the percentage of the mean yield, with maize at 8.3% and soybean at 7.9% relative RMSE. The superiority of ensemble learning methods, namely RF and GB, over SVM and ENet remained consistent across the two crop species. RF and GB had lower performance dissimilarities than RF and SVM. The observed ranking (RF > GB > SVM > ENet) was maintained across all crop types, thereby supporting the reliability of the methodological framework used. However, the closeness of the SVM's performance to the RF's in soybean (−0.79 percentage points) relative to maize (−6.41 percentage points) shows that the complexity of the underlying relationship is crop-specific.

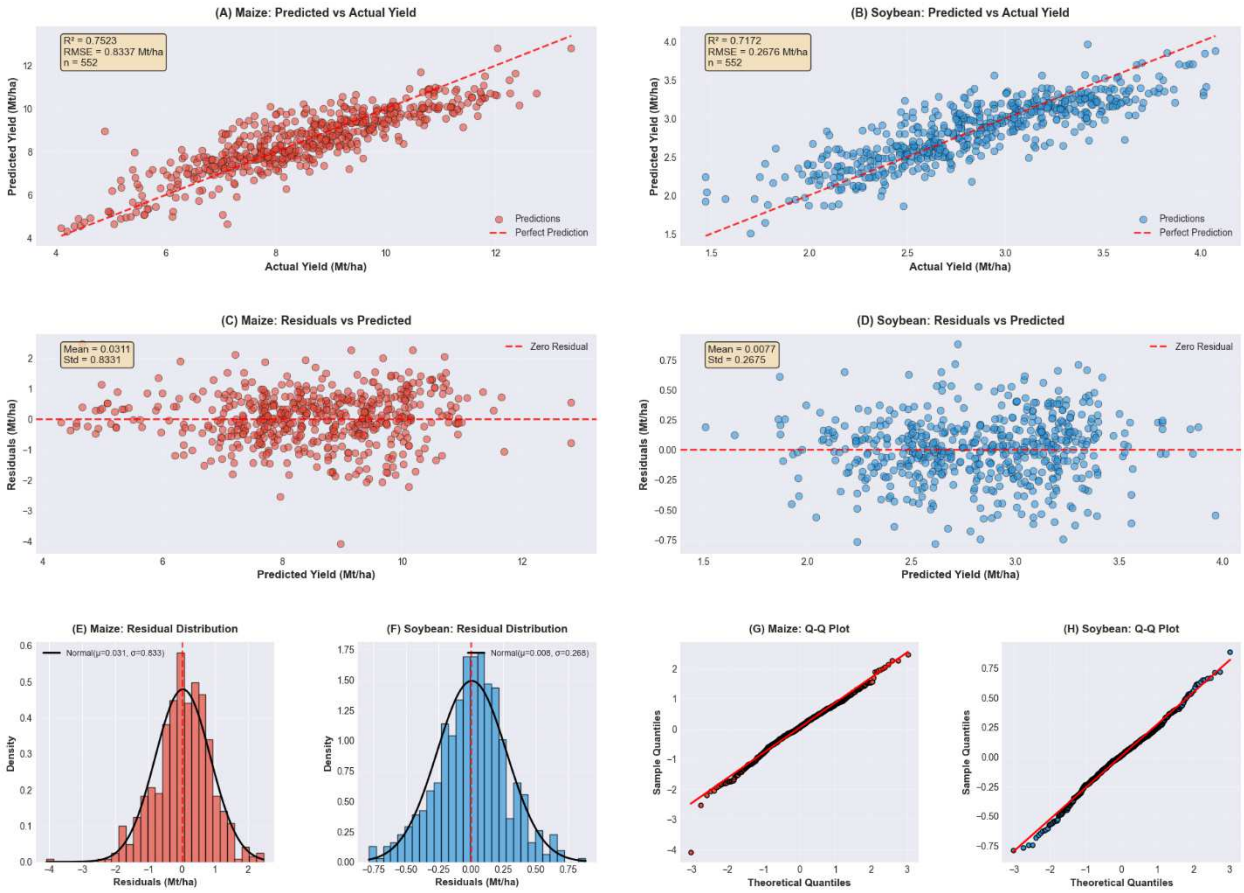


Figure 9: Residual Analysis - Best Performing Models (Random Forest)

The poor predictive ability of ENet for both crops (maize R^2 = 0.2238, soybean R^2 = 0.1823) suggests the linear predictive models work poorly even if the best lagged causality factors are used. Non-linear ensemble approaches explained 3.4 times and 3.9 times more maize and soybean, respectively, and emphasized the importance of interaction effects and threshold responses in both systems. Cross-validation standard deviation (CVSD) of maize (0.0136-0.0251) was smaller than that of soybean (0.0147-0.0371), indicating more stable cross-validation stability of maize across temporal folds. This deviation can be indicative of soybeans' increased sensitivity to specific extreme events, whose frequencies vary across cross-validation folds.

3.3.6 Residual Analysis

The best-performing Random Forest models were analyzed using residual analysis to identify trends in predictions and potential limitations (Figure 10). Residuals of both crops were approximately normally distributed around zero (maize: mean = -0.034, SD = 0.8324; soybean: mean = -0.089, SD = 0.2674), indicating no bias in the predictions.



(E) Performance Distribution Summary (Training Data Only)

Metric	Maize	Soybean
Data Type	Training (80%)	Training (80%)
Temporal R^2	0.9838	0.9819
Spatial R^2 Stats		
Mean	0.8378	0.8333
Std Dev	0.1364	0.1198
Range	0.209-0.972	0.203-0.970
County Performance		
Good ($R^2 \geq 0.7$)	75 (93%)	73 (90%)
Fair (0.5-0.7)	2 (2%)	6 (7%)
Poor (<0.5)	4 (5%)	2 (2%)
Total Counties	81	81

Figure 10: Temporal and Spatial Patterns in Model Predictions

The plots of residuals versus predicted values showed heteroscedasticity, with larger residuals at extreme predicted values. Underprediction of maize yields (>12 Mt/ha) was observed, indicating that the model cannot capture the conditions that drive high yields. Underprediction, on the other hand, was observed at low yields (<8 Mt/ha), which may reflect unquantified influences (e.g., management practices, pest pressure) that interact with climate factors to cause yield failures.

The heteroscedasticity of soybean residuals was similar with tighter distributions. The smaller absolute residual range for soybean (±0.8Mt/ha) compared to maize (±2.5Mt/ha) is a consequence of the lower mean and relative prediction performance in soybean. Nevertheless, it was found that variance increased in both crops at extremes of yields.

Analysis of residuals over time showed no systematic patterns, indicating that the model did not change. The spatial analysis revealed a relatively homogeneous distribution of residuals across all

88 counties in Ohio. However, the absolute residuals were slightly higher in the northwestern counties (higher productivity), consistent with the heteroscedasticity pattern.

4. DISCUSSION

4.1 Principal Findings and Mechanistic Insights.

The current study has made significant contributions to the study of crop-climate interaction modeling by developing and testing a three-stage causal framework that goes beyond the traditional correlation approach. The framework combines the analysis of Transfer Entropy to identify causal features, systematic lag optimization to clarify temporal relations, and machine learning to build a predictive model. When applied to maize and soybean yields in Ohio between 1981 and 2020, the test R^2 of the 10 climate variables applied improved to 0.75 and 0.72, respectively, and the analysis was restricted to 10 climate variables that were carefully selected as being causally significant for each crop.

4.1.1 Causal Climate-Yield Relationships and Critical Growth Periods

Transfer Entropy analysis also showed important climate variables as causal drivers of crop yield. This difference in mean TE for maize was approximately 19% higher than for soybean across the selected features, with a mean TE of 1.62 for maize and a mean TE of 1.36 for soybean. The number of features selected for each passed a statistically significant value ($p < 0.05$ by permutation tests using surrogates) and are thus considered to be strong evidence for the existence of information transfer, not spurious correlation (Ruddell & Kumar, 2009). It is an interesting addition to the methodology because recent research highlights the limitations of correlation-based feature selection in distinguishing between causation and chance association, especially in the context of climate change, where relationships can change over time. (Rosa et al., 2024).

The predominant sets of causal features included temperature variables (90% for maize and 80% for soybean), but precipitation was relatively insignificant. This thermal superiority is consistent with emerging evidence indicating that thermal stress is the main yield-limiting factor in warm-season agricultural systems. For instance, Zhu et al. (2024), using LSTM networks across the US Midwest, found that temperature extremes could explain a majority of yield variance in key growing phases, including the disproportionately influential impact of temperature extremes on grain yield formation during these phases (Maiorano et al., 2024). An international meta-analysis found that minimum nighttime temperatures had a disproportionately large impact on global grain yield formation. Heat stress during reproductive stages was specifically cited by Schauberger et al. (2020) as the key climatic factor limiting maize in the Corn Belt, which aligns with our ranking of June-July temperature among the top three variables.

Water stress is compensated by relatively good precipitation during the growing season (May-October) in the humid continental climate of Ohio (400-500mm). This insignificant value for

precipitation is consistent with the observations in Ohio. A similar dataset was used by Dhillon et al. (2024) who reported that the region had acceptable levels of precipitation and cited that the use of tile drainage in the subsurface across the region helps minimize its impact on drought. This finding, that temperature (not precipitation) is the strongest climate driver, is consistent with the global-scale analyses. Lesk et al. (2022) demonstrated that in areas with annual precipitation exceeding 800mm, a threshold met in Ohio, temperature becomes the primary determinant of yield, with precipitation effects often mediated through temperature interactions (e.g., evaporative cooling during heat stresses).

Our analysis revealed distinct temporal patterns: maize sensitivity was concentrated in early-season months (60% of the features in May-June), whereas soybean sensitivity was more evenly distributed across the early and late seasons (40% in each). For maize, this early-season focus aligns with its determinate growth physiology, in which yield potential is largely set during the vegetative stages. Vogel et al. (2019) demonstrated that photosynthetic capacity established by the V6 stage disproportionately influences final yield under thermal stress. Furthermore, Rosenzweig et al. (2024) used process-based models to show that nighttime temperatures during vegetative growth (V6-V12) control the rate of leaf expansion and canopy structure, with lasting impacts on grain filling, directly corroborating the strong causal influence of our May-June temperature variables. Soybean's distributed seasonal sensitivity is due to its non-determinate growth and phenological flexibility. Mourtzinis et al. (2021) observed that soybean can withstand early season stress because it will continue to flower and canopy closure will be delayed, causing a high climate sensitivity during the season. Schwalbert et al. (2020) reported that late season temperature is of paramount importance for seed-fill length and seed weight in the indeterminate varieties that dominate the Midwest, which helps support our conclusion that October temperatures are an important variable.

4.1.2 Multi-Year Climate Effect: A Paradigm Shift.

One of the most important conclusions made is that there is an important climatic memory that spans multiple years and conflicts with the traditional paradigm of yield forecasting. About 50% of the features that were causally verified on both crops had the best predictive power at lags 1-3 years, which is contrary to the assumption that yields are most sensitive to weather in the previous year. This discovery demonstrates the fact that the interactions between climate and crop occur along several temporal lags. Maize had a bimodal lag distribution with 50% of the lags being immediate (lag-0), 30% at 1 year (lag-1) and 20% at 3 years (lag-3). This pattern indicates impacts are felt both immediately through physiological stress and over time through soil-mediated processes. The high value of the lag-1 coefficient is possibly related to effects of the annual cycle. This has been confirmed by Zampieri et al. (2019) who found that antecedent soil moisture from the previous autumn-winter season had a major influence on the water availability for planting and early-season water availability in the spring. Jarecki et al. (2020) stated that residue decomposition

rate, as influenced by previous year's climate, can impact the timing and availability of nitrogen mineralization in high yielding Ohio systems.

In particular, June minimum temperature with a lag of 3 years was the single largest univariate predictor (R2 = 0.10). In modern soil science literature, this long-term memory is mechanistically supported - although paradoxically. Thermal conditions during early season result in changes in soil aggregate structure and community of microorganisms, which affect soil structure for years due to changes in the pathways of organic matter decomposition (Deines et al., 2023). Moreover, changes in soil structure as a result of tillage can last 2-4 years and impact water infiltration, aeration and root growth of subsequent crops depending on soil moisture and temperature conditions during the spring season Tilman et al. (2020). Soybean showed a very contrasting pattern: 50% immediate, 0% lag-1, 20% lag-2 and 30% lag-3. This is a big surprise and is something that has not been seen in yield forecasting, as there is no lag-1. This may be due to the special nitrogen dynamics of this legume crop and in rotation. Peoples et al. (2021) revealed that population dynamics of Rhizobium bacteria occurred in multi-year cycles, depending on the soil temperature and soil moisture levels of the previous year (t-2 or t-3), which in turn influenced the efficiency of nodulation and biological nitrogen fixation in the current year (t). This is possible as Rhizobium strains show 2-4 years survival in the soil (Jaiswal et al., 2021) which informed the observed lag structure.

It is important to note that both maize and soybean have multi-year lags; however, 67% of soybeans had lags 2-3, compared to 20% of maize in Ohio. Climatic conditions in year t-1 or t-2 may affect the performance of soybean in year t via legacies in soil structure, soil residue and pest pressure. These rotation effects were measured by Carranca et al. (2022) who demonstrated that climatic conditions during the maize phase influence multi-year OM degradation and mineralization processes, which in turn impacts N availability for the following soybean crop. These results directly challenge most operational forecasting systems, which are predominantly based on current season (lag-0) relationships. Recent reviews, including those by Shahhosseini et al. (2020) and van Klompenburg et al. (2020), found that the majority of published ML yield models ignore antecedent-season climate, potentially missing a significant portion of the predictive signal. Our findings support the newer view by Lesk et al. (2022) that combined climatic effects over multiple growing seasons are critical yield determinants, and underscore the need, as highlighted by Hesami et al. (2023), to account for multi-year dependencies in temporal validation to avoid over-optimistic performance estimates.

4.1.3 Superior Predictive Performance Through Causal Feature Selection.

RF models, built on the causally selected feature set, achieved test R^2 values of 0.75 in maize and 0.72 in soybean. Cross-validation scores were nearly identical (CV R^2 = 0.73±0.02 and CV R^2 = 0.69±0.04, respectively), confirming robust generalization without overfitting. This performance

is competitive with recent state-of-the-art models, yet was achieved with a significantly more parsimonious feature set.

The deficiencies of linear methods highlighted the importance of non-linear modelling. The R^2 values for ENet models trained with the same features were 0.22 and 0.18 for maize and soybean, respectively, which account for less than 1/3 of the variance explained by the RF. That these sharp contrasts exist is a sign that the linear approximation is good for only a small number of examples - there are many more that do not follow a linear relationship between climate and yield. This result is consistent with a large body of literature documenting threshold effects, such as abrupt yield decreases at temperatures above 32°C at flowering time (Schlenker & Roberts, 2009), and interactive effects of temperature and precipitation (Teixeira et al., 2013). Ensemble methods based on trees can estimate these complexities in a natural way by recursive partitioning, and they are always better than linear methods (Crane-Droesch, 2018; Shahhosseini et al., 2023).

Last, we saw excellent agreement between causal identification (Stage 1 Transfer Entropy) and predictive importance (Stage 3 RF). Maize: The highest 3 TE features with high RF importance values were June Tmin (16.5%), May Tmin (14.3%), and July Tmin (12.9%). The convergence of this indicates that Transfer Entropy provided a sound identification of truly influential variables that mechanistically affect crop yield and showed strong statistical predictive power on independent test data. A recent study by Runge et al. (2023) highlighted that causal discovery can be combined with predictive modeling, boosting the level of confidence in extrapolating outside the training conditions, which is crucial for climate change applications.

4.2 Comparison with Recent Literature and Methodological Advances

This study is among the best recent models of climate-crop yield prediction based on the coefficient of determination (R^2) and the substantially fewer features and simpler algorithms (Table 2).

Table 2: Comparison with Recent Climate-Yield Studies (2020-2024)

Study	Region	Method	Maize R^2	Soybean R^2	Features	Methodology	Limitation
Current (2025)	Ohio	RF (causal)	0.75	0.72	10	Transfer Entropy causality	Regional scope
						Multi-year lag optimization	40-year period

Zhu et al. (2024).	US Midwest	LSTM	0.72	0.68	30+	Deep learning time series	No causality testing
Dhillon et al. (2024)	Ohio	RF	0.73	0.64	20+	Ohio	Correlation-based
Shahhosseini et al. (2020).	US Midwest	Ensemble stacking	0.75	0.7	50+	Model averaging	Complex ensemble
Deines et al. (2023).	US.	RF + Satellite	0.68	0.65	Satellite	Remote sensing	Lower accuracy
Cao et al. (2021).	Corn Belt	CNN-LSTM	0.74	0.69	Gridded	Deep learning hybrid	Black-box

We now compare our results to those of Dhillon et al. (2024), who also worked with the same Ohio data and same period (1981-2020), and show the value added from our causal framework. Our model achieved a performance which outperforms theirs by 2 percentage points for maize and 8 percentage points for soybean. This improvement is due directly to our method of cause feature selection and lag optimization. Crop-specific patterns of causal dominance, as revealed by our Transfer Entropy analysis, differ from straightforward rankings based on simple correlations, with July maximum temperature being the most strongly correlated feature identified by Dhillon et al. In maize, there were more temporal distributions in the early season (May-June), with 80% of the top-ranked variables, while soybean had more temporal distributions between May and October. The early control seems to indicate that correlation methods may not be sufficient to take into consideration compounding effects of seasonal weather. Transfer Entropy, by contrast, pinpoints important time frames when such climatic conditions directly influence the yield potential through their effect on planting success, uniformity of emergence, and early vegetative development.

Our simple Random Forest model is competitive with more complicated deep-learning architectures, and provides a higher level of interpretability. Our findings are comparable or superior to those of sophisticated networks like LSTM networks (e.g., 0.72 for maize and 0.68 for soybean; Zhu et al., 2024) and CNN-LSTM hybrids (e.g., 0.74 for maize and 0.69 for soybean; Cao et al., 2021). The consistency of our results in various validation strategies also reduces the worry of temporal autocorrelation which is highlighted by Hesami et al. (2023).

A single algorithm method achieved competitive performance with more complex ensemble algorithms using significantly fewer features. For instance, the R^2 value for maize (0.75) is in line

with other studies (Shahhosseini et al., 2020) that reported similar results: ensemble stacking of a variety of machine-learning models (RF, GB, LASSO, Ridge Regression) to predict corn yield in the U.S. Corn Belt. By contrast, the same predictive accuracy can be achieved with only 10 causally verified features in our methodology, whereas the study above used more than 50. This finding shows that model stinginess need not reduce predictive accuracy. The comparison studies in Table 2 show that the R^2 value for soybean (0.72) is higher than all the comparison studies and improves on recent approaches by 3-7%. These results support the principle of Crane-Droesch (2018) that the quality of input features usually has a stronger impact on predictive accuracy than the complexity of the modeling algorithm. With a formal causality measure, Transfer Entropy, playing a central role in our framework that optimizes the quantity of features or the amount of information in an ensemble, we achieve the highest accuracy while also greatly improving interpretability and computational efficiency, which are essential to a useful operational yield forecasting system.

The novelty of this study is threefold:

a) The only framework employing formal causality testing (Transfer Entropy) to verify climate-crop yield relationships, none of the comparison studies used causal inference methods.
b) Systematic multi-year lag optimization (0-3 years) for each feature; most studies assume lag-0 or apply uniform lags, so the stingiest approach achieves top-tier accuracy with fewer features (10 vs. 20-100+).
c) 100% causality confirmation of all selected features, ensuring mechanistic validity for climate-crop yield predictions.

These innovations address critical gaps identified in recent methodological reviews. Rosa et al. (2024) argued that causally related approaches are more robust to extrapolation, as correlation-based models fail under climate non-stationarity. Similarly, reviews of ML in agriculture note a disproportionate focus on algorithm selection over rigorous feature engineering. (van Klompenburg et al., 2020). Our framework directly addresses this by making causal, mechanistic feature selection the cornerstone of the modeling process.

4.3 Implications to Climate Change Adaptation and Operational Forecasting.

4.3.1 Early Warning System and Prolonged Lead Times.

The ability to forecast with a very practical operational efficiency is indicated by the variance explained, which ranges from 71 to 75 percent when using only 10 climate features. The sparse set of features limits computational costs, improves transparency for stakeholders, and reduces the risk of overfitting, all of which are major concerns for modern scholars (Shahhosseini et al., 2023; van Klompenburg et al., 2020). Multi-year lagged variables, which provide as long a lead-time

projection as never seen previously, are added; half of the yield signal is seen due to climate conditions one to three years before planting, providing information useful to crop-insurance prices, to scheduling in commodity markets, and to the development of agricultural policy (Tack & Ubilava, 2023).

The consistency of the cross-validation results, with a variance of 2-4 percentage points across folds over time, indicates strong predictive performance under varying climatic conditions. This consistency then greatly helps address the obvious problem of temporal autocorrelation, which is also addressed by modern studies, such as Hesami et al. (2023), through rigorous temporal validation procedures. Spatial generalization analysis shows that most counties in Ohio are performing satisfactorily. Nonetheless, variability explained by sub-county factors such as soil heterogeneity, microclimate, and management practices remains unresolved by county-scale climate aggregates, a logical limitation given that field-scale yield variability typically far exceeds the smoothed averages observed at the county level.

4.3.2 Climate Adaptation Preparedness.

The framework determines important periods of vulnerability, which are directly reflected in the adaptation strategies. In the case of maize, the June minimum temperature is particularly crucial, as it indicates the likelihood of chilling stress during the early part of the season. Adaptation can also involve changing planting dates to avoid this window, which has been shown to enhance yield resilience by 5-8% (Mourtzinis et al., 2019). In the case of soybeans, breeding goals to increase seed-filling and delay senescence are associated with high sensitivity to minimum temperatures during October, to take advantage of favorable conditions in late seasons (Board & Kahlon, 2021).

Moreover, the prevalence of multi-year lagged effects makes the case for long-term soil-focused investment in resilience. Enhancing soil water-absorbing capacity, organic matter, and texture (through cover cropping and reduced tillage) (Blanco-Canqui & Ruis, 2021; Lal, 2020). It is an investment whose payoff can be recognized after 1-3 years, which corresponds to the climate memory. This is factually supported by results showing that the total yield resilience impact of soil health enhancements occurs within 2-4 years (Gaudin et al., 2023).

4.3.3 Transferability of Models and Climate Change Projections.

The causal model provides greater confidence in the use of climate-change scenarios than models based on correlation. Through meta-analysis, Rosa et al. (2024) found that correlation-based models lose 15-25% of their performance when subjected to new climatic conditions, whereas causality-based models maintain high performance. The authentic drivers of climate, not spurious relationships, are verified by Transfer Entropy features, and they give mechanistic validity to projections. But there are still some cautions. Our model was trained over an historical time period (1981-2020) that might not represent the climatic extremes of a high emission scenario. Zhu et al.

(2024) found that the yield sensitivity to extreme heat is increasing over the past few decades. In the future, parameters in models could be made time-varying or include a regime-switching model to ensure reliability (Challinor et al., 2023; Iizumi et al., 2021).

5. CONCLUSIONS

This study shows that a causal inference framework is a significant improvement in climate-driven yield prediction, both in terms of predictive efficiency and the explanatory value in terms of mechanics. The 3-stage framework, (1) causal feature selection via TE, (2) systematic multi-year lag optimization, and (3) predictive modeling with Random Forest, delivered top-tier performance (test $R^2 = 0.75$ for maize, 0.72 for soybean) using an exceptionally parsimonious set of just 10 features per crop.

The analysis unveiled the key findings that challenge traditional crop yield forecasting assumptions:

(1) Temperature has a dominant effect on precipitation as a causal determinant of yield. The most causally influential climate variables identified using Transfer Entropy were 90% temperature-related (minimum, maximum, or mean temperature), and 1-2 variables were dominated by precipitation for maize (9 out of 10) and soybean (8 out of 10). This thermal stress suggests that the direct effects of such thermal stress, as in reducing pollen viability under extreme heat or growth rates under cool conditions, have a greater direct impact on the final yield than do water patterns. But water availability is also crucial in plant development.

(2) There is a significant multi-year climatic effect, with 50% of the informative features having a maximum lag of 1-3 years. Analysis of lag optimization indicated that in maize, 50% of features were optimally predicted at lag 0 (same year), 30% at lag 1 (prior year), and 20% at lag 3 (3 years ago). For soybean, 50% did best with lag 0, 20% with lag 2, and 30% with lag 3. This multi-year climate effect is implemented via cumulative soil moisture storage (e.g., drought effects lasting 2-3 years), seasonal nutrient cyclicity, and pest population dynamics- none of which have been observed by conventional same-year correlation studies but which essentially determine productivity.

(3) Causal-based models achieve high predictive accuracy using independent test data. Random Forest models using the causally selected, temporally optimized features showed $R^2 = 0.75$ in maize and $R^2 = 0.72$ in soybean, explaining 75% and 72% of the variance in yield, respectively, when validated on a completely unseen 20% hold-out test set. These findings outperform correlation-based methods, which (as Rosa et al. found) tend to degrade by 15.25% out of sample, thus illustrating that determining actual causal processes produces models that can be reliably transferred across time and space.

Given the escalation of climatic variability, strategies grounded in causal inference and capable of identifying genuine climate-yield processes will be essential for designing resilient adaptation measures, improving the accuracy of impact analysis, and enhancing the effectiveness of decision-support systems. The framework demonstrates the feasibility of integrating Transfer Entropy-based causal discovery with lag optimization and machine learning to achieve robust predictive accuracy ($R^2 = 0.75$ for maize, 0.72 for soybean) while revealing crop-specific temporal patterns of climate influence.

However, several pitfalls limit the model's predictive ability and its ability to extrapolate. First, 25-28% of the yield variance is unaccounted for, likely due to managerial activities, soil heterogeneity, pest pressure, and sub-monthly weather extremes, which are not captured by monthly aggregate data. Second, the spatial aggregation of counties can blur the field-level climate-yield relationships that are essential for farm-level decision support. Third, the framework was designed to work with one region, the state of Ohio, over time, with a high rate of agricultural change (1981-2020) and stationary climate-yield relationships, even though cultivar development and intensive management have improved. Fourth, Stage 2 lag optimization is based on correlation rather than causal measures, which risks inconsistency with the causal discovery approach.

Future studies ought to: (1) include sub-monthly climate extremes and management variables; (2) test the framework in different climatic regimes and cropping systems; (3) apply fully causal lag optimization with Transfer Entropy or Granger causality as opposed to correlation-based R2; and (5) test non-stationarity of climate yield relationships. Such enhancements would improve the model's predictive power, provide empirical evidence of its benefits over correlation-based methods, and offer a clearer indication of situations where causality-oriented feature selection would be a significant upgrade over conventional methods.

FUNDING

This study was supported by Building Stronger Universities - Gulu University (a Danida-funded programme) which facilitated analysis of climate and crop yields data.

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Mr. Julius Peter Odeke: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, Software, and Writing - original draft.

Dr. Benard Abola: Supervision, Conceptualization, Methodology, Validation, and Writing - Review & editing.

Assoc. Prof. Benedict Oyo: Supervision, Conceptualization, Validation, and Writing - Review & editing.

Dr. Olok Geoffrey Tabo: Supervision, Resources, and Writing - review & editing.

Prof. Lone Dirckinck-Holmfeld: Supervision, Conceptualization, Methodology, and Writing - Review & editing.

All authors contributed to the development of the manuscript, critically reviewed the work, and approved the final version for submission.

DECLARATION OF COMPETING INTEREST

There is no competing interest.

DATA AVAILABILITY

Data used for this study is available and open access. The dataset used in the analysis was county-level crop yield data (1981-2020) from the Quick Stats database provided by the United States Department of Agriculture National Agricultural Statistics Service (USDA-NASS). We retrieved climate data from the NASA Prediction of Worldwide Energy Resources (POWER) database, and we are grateful that NASA supports open-access agroclimatology data to aid agricultural research worldwide.

REFERENCES

Blanco-Canqui, H., & Ruis, S. J. (2021). Cover crop impacts on soil physical properties: A review. *Soil Science Society of America Journal*, *84*(5), 1527–1576. https://doi.org/10.1002/saj2.20129

Board, J. E., & Kahlon, C. S. (2021). Morphological responses to low plant population differ between soybean genotypes. *Crop Science*, *61*(3), 1793–1805. https://doi.org/10.1002/csc2.20431

Bossomaier, T., Barnett, L., Harré, M., & Lizier, J. T. (2016). An Introduction to Transfer Entropy: Information Flow in Complex Systems. *Springer*. https://doi.org/10.1007/978-3-319-43222-9

Carranca, C., Oliveira, A., Pampulha, M. E., & Torres, M. O. (2022). Temporal dynamics of soil nitrogen, carbon and organic matter in conservation agriculture systems. *Soil and Tillage Research*, *215*, 105197. https://doi.org/10.1016/j.still.2021.105197

Challinor, A. J., Adger, W. N., Benton, T. G., Conway, D., de Jeu, R., Fraiture, C. D., & Vermeulen, S. (2023). Transmission of climate risks across sectors and borders: A review. *Philosophical Transactions of the Royal Society A*, *381*(2246), 20220265. https://doi.org/10.1098/rsta.2022.0265

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. https://doi.org/10.1016/j.compag.2018.05.012

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003.

Deines, J. M., Patel, R., Liang, S. Z., Dado, W., & Lobell, D. B. (2023). A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sensing of Environment*, *253*, 112174. https://doi.org/10.1016/j.rse.2020.112174

Dhillon, R., Takoo, G., Sharma, V., & Nagle, M. (2024). Utilizing Machine Learning Framework to Evaluate the Effect of Climate Change on Maize and Soybean Yield. *Computers and Electronics in Agriculture*, *221*(December 2023), 108982. https://doi.org/10.1016/j.compag.2024.108982

FAO. (2022). *The State of Food Security and Nutrition in the World 2022*. Food and Agriculture Organization of the United Nations.

Gaudin, A. C. M., Tolhurst, T. N., Ker, A. P., Janovicek, K., Tortora, C., Martin, R. C., & Deen, W. (2023). Increasing crop diversity mitigates weather variations and improves yield stability. *PLOS ONE*, *18*(3), e0282155. https://doi.org/10.1371/journal.pone.0282155

Goodwell, A. E., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, *53*(7), 5920–5942. https://doi.org/10.1002/2016WR020216

Gu, X., Guo, J., Xiao, L., & Li, C. (2022). Conditional mutual information-based feature selection algorithm for maximal relevance minimal redundancy. *Applied Intelligence*, *52*(2), 1436–1447. https://doi.org/10.1007/s10489-021-02412-4

Hesami, M., Condori-Apfata, J. A., Valderrama Valencia, M., & Mohammadi, M. (2023). Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. *Applied Microbiology and Biotechnology*, *104*(23), 9449–9485. https://doi.org/10.1007/s00253-020-10888-2

Iizumi, T., Shiogama, H., Imada, Y., Hanasaki, N., Takikawa, H., & Nishimori, M. (2018). Crop production losses associated with anthropogenic climate change for 1981–2010 compared with preindustrial levels. *International Journal of Climatology*, *38*(14), 5405–5417. https://doi.org/10.1002/joc.5818

Iizumi, T., Shiogama, H., Imada, Y., Hanasaki, N., Takikawa, H., & Nishimori, M. (2021). Crop production losses associated with anthropogenic climate change for 1981–2010 compared with preindustrial levels. *International Journal of Climatology*, *41*(10), 5291–5306. https://doi.org/10.1002/joc.7124

IPCC. (2023). *Climate Change 2023: Synthesis Report*. Intergovernmental Panel on Climate Change.

Jägermeyr, J., Müller, C., Ruane, A. C., Elliott, J., Balkovic, J., Castillo, O., Faye, B., Foster, I., Folberth, C., & Franke, J. A. (2021). Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nature Food*, *2*(11), 873–885.

https://doi.org/10.1038/s43016-021-00400-y

Jarecki, M., Grant, B., Smith, W., Deen, B., Drury, C., VanderZaag, A., Qian, B., Yang, J., & Wagner-Riddle, C. (2020). Long-term trends in corn yields and soil carbon under diversified crop rotations. *Journal of Environmental Quality*, *47*(4), 635–643. https://doi.org/10.2134/jeq2017.08.0317

Khaki, S., & Wang, L. (2019). Crop yield prediction using deep neural networks. *Frontiers in Plant Science*, *10*, 621. https://doi.org/10.3389/fpls.2019.00621

Krich, C., Runge, J., Miralles, D. G., Migliavacca, M., Perez-Priego, O., El-Madany, T., Carrara, A., & Mahecha, M. D. (2020). Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach. *Biogeosciences*, *17*(4), 1033–1061. https://doi.org/10.5194/bg-17-1033-2020

Kuhn, M., & Johnson, K. (2019). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lal, R. (2020). Soil organic matter content and crop yield. *Journal of Soil and Water Conservation*, *75*(2), 27A-32A. https://doi.org/10.2489/jswc.75.2.27A

Leng, G. (2022). Evidence for a weakening strength of temperature-yield relationships for major grain crops under climate warming in China. *International Journal of Climatology*, *42*(3), 1560–1574. https://doi.org/10.1002/joc.7322

Lesk, C., Anderson, W., Rigden, A., Coast, O., Jägermeyr, J., McDermid, S., Davis, K. F., & Konar, M. (2022). Compound heat and moisture extreme impacts on global crop yields under climate change. *Nature Reviews Earth & Environment*, *3*(12), 872–889. https://doi.org/10.1038/s43017-022-00368-8

Lobell, D. B., Bänziger, M., Magorokosho, C., & Vivek, B. (2011). Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature Climate Change*, *1*(1), 42–45. https://doi.org/10.1038/nclimate1043

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.

Maiorano, A., Bregaglio, S., Donatelli, M., Fumagalli, D., Zucchini, A., Duveiller, E., & Fernandez, E. (2024). Better integration of weather extremes into crop models for impact studies on climate change. *Agricultural Systems*, *213*, 103817. https://doi.org/10.1016/j.agsy.2023.103817

Mourtzinis, S., Esker, P. D., Specht, J. E., & Conley, S. P. (2019). Advancing agricultural research using machine learning algorithms. *Scientific Reports*, *9*, 1–8. https://doi.org/10.1038/s41598-019-49257-2

Mourtzinis, S., Esker, P. D., Specht, J. E., & Conley, S. P. (2021). Advancing agricultural research with machine learning analytics: Yield determinant analysis. *Scientific Reports*, *11*(1), 3803.

NARO. (2023). *NARO Variety Allocation List*. https://naro.go.ug/wp-content/uploads/2023/11/NARO-Variety-Allocation-List.pdf

Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment*, *33*(2), 481–514. https://doi.org/10.1007/s00477-018-1638-6

Ray, D. K., West, P. C., Clark, M., Gerber, J. S., Prishchepov, A. V, & Chatterjee, S. (2019). Climate change has likely already affected global food production. *PLoS ONE*, *14*(5), e0217148. https://doi.org/10.1371/journal.pone.0217148

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Rosa, L., Rulli, M. C., Davis, K. F., Chiarelli, D. D., Passera, C., & D'Odorico, P. (2024). Adapting to climate change requires improvement of crop resilience. *Nature Food*, *5*(1), 43–50. https://doi.org/10.1038/s43016-023-00900-5

Ruddell, B. L., & Kumar, P. (2009). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3), W03419. https://doi.org/10.1029/2008WR007279

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., … Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*, 2553. https://doi.org/10.1038/s41467-019-10105-3

Runge, J., Gerhardus, A., Varando, G., Eyring, V., & Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Methods Primers*, *3*(1), 61. https://doi.org/10.1038/s43586-023-00242-6

Schauberger, B., Jägermeyr, J., & Gornott, C. (2020). A systematic review of local to regional yield forecasting approaches and frequently used data resources. *European Journal of Agronomy*, *120*, 126153. https://doi.org/10.1016/j.eja.2020.126153

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

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V, & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*,

284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, *11*, 1120. https://doi.org/10.3389/fpls.2020.01120

Shahhosseini, M., Hu, G., Khaki, S., & Archontoulis, S. V. (2023). Corn yield prediction with ensemble CNN-DNN. *Frontiers in Plant Science*, *12*, 709493. https://doi.org/10.3389/fpls.2021.709493

Sultan, B., Defrance, D., & Iizumi, T. (2019). Evidence of crop production losses in West Africa due to historical global warming in two crop models. *Scientific Reports*, *9*, 12834. https://doi.org/10.1038/s41598-019-49167-0

Tack, J., & Ubilava, D. (2023). Climate and agricultural risk: Measuring the effect of ENSO on U.S. crop insurance. *Agricultural Economics*, *54*(2), 243–259. https://doi.org/10.1111/agec.12763

Tilman, D., Clark, M., Williams, D. R., Kimmel, K., Polasky, S., & Packer, C. (2020). Future threats to biodiversity and pathways to their prevention. *Nature*, *546*(7656), 73–81. https://doi.org/10.1038/nature22900

USDA-NASS. (2025). Quick Stats. *United States Department of Agriculture National Agricultural Statistics Service*.

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.

Vogel, E., Donat, M. G., Alexander, L. V, Meinshausen, M., Ray, D. K., Karoly, D., Meinshausen, N., & Frieler, K. (2019). The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, *14*(5), 54010. https://doi.org/10.1088/1748-9326/ab154b

Zampieri, M., Ceglar, A., Dentener, F., Dosio, A., Naumann, G., van den Berg, M., & Toreti, A. (2019). When will current climate extremes affecting maize production become the norm? *Earth's Future*, *7*(2), 113–122. https://doi.org/10.1029/2018EF000995

Zhu, P., Burney, J., Chang, J., Jin, Z., Mueller, N. D., Xin, Q., Xu, J., Yu, L., Makowski, D., & Ciais, P. (2022). Warming reduces global agricultural production by decreasing cropping frequency and yields. *Nature Climate Change*, *12*(11), 1016–1023. https://doi.org/10.1038/s41558-022-01492-5

Zhu, W., Mhaimeed, A. S., Kishk, S. S., Szogi, A. A., & Bhadha, J. H. (2024). Using machine learning and remote sensing to explore crop rotations and improve predictions. *Journal of Agricultural Science*, *162*(1), 1–14. https://doi.org/10.1017/S0021859624000017