

A Machine Learning Benchmark for Predicting County-Level Corn Production in Minnesota Using Multi-Source Satellite, Environmental, and Economic Data

Tang Zi Jian (Jacob), Emma Wele, Victor C, Onishi Rei

November 19, 2025

Abstract

This research presents a machine learning benchmarking framework for predicting county-level corn production in Minnesota using multi-source environmental, agricultural, and economic data. We integrate satellite-derived environmental variables from the Global Land Data Assimilation System (GLDAS), precipitation data from Parameter-elevation Regressions on Independent Slopes (PRISM), economic indicators from USDA Economic Research Service, diesel fuel prices, and infrastructure data. We systematically evaluate eight machine learning algorithms spanning linear baselines to gradient boosting and deep learning models: Polynomial Regression, Support Vector Machine, Random Forest, XGBoost, LightGBM, TabNet, LSTM, and Temporal Convolutional Network (TCN). All models are evaluated without using acres planted as an input feature, focusing on the more challenging and practically relevant setting where planted-area data may be unavailable or delayed. The LightGBM model achieves the best performance with an R^2 of 0.978 and root mean squared error (RMSE) of 2,150,000 bushels, explaining 97.8% of variance in corn production. Feature importance analysis reveals that economic proxies (fuel cost, income, federal receipts) and environmental variables (soil moisture, temperature) are critical when acres planted is unavailable. Our results demonstrate that gradient boosting methods significantly outperform traditional ensemble methods and deep learning architectures for this tabular regression task, and that multi-source integration enables accurate production prediction even without direct agricultural area measurements.

Keywords: Agricultural yield prediction, machine learning, satellite remote sensing, LightGBM, multi-source data integration, GLDAS, PRISM, economic indicators

1 Introduction

Agriculture is a major pillar of the United States economy, with corn production playing a critical role in national food security, biofuel production, and economic stability [12]. Corn accounts for over 15.0% of total agricultural value, representing billions of dollars in annual economic activity. Accurate crop yield prediction enables stakeholders including farmers, agricultural cooperatives, commodity traders, and policymakers to make informed decisions regarding

planting strategies, resource allocation, pricing, and risk management.

Traditional methods for agricultural yield prediction rely primarily on historical averages, field surveys, and expert knowledge, which are fundamentally limited in capturing dynamic changes driven by climate extremes, technological advances, and economic conditions. These limitations become more critical with increasing climate variability, resulting in frequent extreme weather events that cause significant deviations from historical averages.

Satellite remote sensing provides a breakthrough for agricultural productivity monitoring by offering global, consistent views of land surface conditions. The Global Land Data Assimilation System (GLDAS), developed by NASA, combines satellite and in situ observations to create high-quality land surface variables such as soil moisture, temperature, evapotranspiration, and precipitation. Similarly, PRISM provides high-resolution gridded climate variables that consider topographic factors influencing precipitation and temperature patterns.

However, agricultural yield is not determined solely by environmental conditions. Economic factors including fuel costs, commodity prices, government support programs, and market access significantly influence planting decisions, input intensity, and ultimately production levels. Transportation infrastructure, particularly proximity to processing facilities such as ethanol plants, affects market access and profitability. A comprehensive yield prediction framework must therefore integrate multiple heterogeneous data sources beyond satellite-derived environmental variables.

This research extends previous work by developing a comprehensive benchmark framework that systematically integrates five primary data sources: (1) GLDAS satellite-derived environmental variables, (2) USDA corn harvest data, (3) economic indicators from USDA Economic Research Service, (4) diesel fuel prices as proxies for operational costs, and (5) PRISM precipitation and temperature data. The study focuses on Minnesota counties (FIPS codes 27001-27171) over 2000-2022, providing a rich temporal and spatial dataset encompassing 87 counties over 23 years.

In this paper, we make the following contributions:

- We construct a multi-source county-year dataset for Minnesota integrating GLDAS, PRISM, NASS, EIA, ERS, and ethanol plant distance data, demonstrating how heterogeneous data sources can be harmonized for agricul-

tural prediction.

- We benchmark eight machine learning models (polynomial regression, SVM, random forest, XGBoost, LightGBM, TabNet, LSTM, TCN) for predicting corn production *without using acres planted as an input feature*, focusing on the more challenging and practically relevant setting where planted-area data may be unavailable or delayed.
- We analyze feature importance to identify which environmental and economic/infrastructure signals best compensate for the absence of acres planted, revealing that economic proxies and environmental variables provide strong predictive power when direct agricultural area measurements are unavailable.

2 Related Work

The integration of satellite remote sensing data with machine learning techniques for agricultural yield prediction has emerged as a rapidly evolving field. Satellite remote sensing for agricultural applications encompasses multiple sensor modalities: Synthetic Aperture Radar (SAR) systems offer all-weather capability and sensitivity to soil moisture and vegetation structure [7, 10]; optical imagery from sensors such as Landsat, MODIS, and Sentinel-2 provides vegetation indices (NDVI, EVI, GNDVI) that capture photosynthetic activity and crop health [4]; hyperspectral and thermal sensors enable estimation of biochemical properties and water stress. Multi-modal approaches combining these data types often improve yield prediction, but face challenges including cloud contamination, scale mismatches between pixel resolution and field/county boundaries, and vegetation interference during growing seasons [9].

Machine learning has revolutionized agricultural yield prediction by enabling data-driven modeling of complex, non-linear relationships. Random Forest (RF) has been widely applied for its interpretability and robustness [6]. Gradient boosting methods including XGBoost [13] and LightGBM [14] demonstrate superior performance for tabular regression tasks through sequential error correction. Deep learning approaches including CNNs [5, 8] and LSTMs have been applied, though they often require large datasets and may not align well with tabular agricultural data where independent samples (counties, years) are more common than true temporal sequences.

Recent research trends emphasize integration of multiple heterogeneous data sources beyond satellite-derived environmental variables, recognizing that agricultural yield results from complex interactions between environmental conditions, economic factors, technological advances, and policy influences [1, 2]. Economic indicators including commodity prices, fuel costs, and government support programs significantly influence planting decisions and production levels. Soil property databases, crop-specific datasets, and market logistics data provide additional context. However, most prior studies focus on single algorithms or limited comparisons, and many rely primarily

on environmental/remote-sensing data rather than incorporating economic and infrastructure signals at the county level. Our work is novel in combining environmental, agricultural, and economic/infrastructure data at the county level, and in analyzing performance without acres planted, demonstrating compensatory mechanisms when primary features are unavailable.

3 Data and Problem Setup

Given features $\mathbf{x}_{c,t}$ for county c in year t , we predict total corn production $y_{c,t}$ (bushels). The dataset integrates five primary data sources, consolidated at the county-year level using Federal Information Processing Standard (FIPS) codes for Minnesota counties (27001-27171).

GLDAS Environmental Variables: The Global Land Data Assimilation System provides high-quality land surface variables derived from satellite and ground-based observations [11], including atmospheric variables (air temperature, pressure, wind, humidity), radiation (longwave and shortwave downwelling), evapotranspiration components, soil variables at multiple depths (0-10cm, 10-40cm, 40-100cm, 100-200cm) including moisture and temperature, surface variables (albedo, canopy interception, vegetation temperature), and hydrological variables (surface runoff, subsurface runoff, snow depth, snow water equivalent). Monthly GLDAS data was aggregated to annual resolution for each county.

NASS Corn Production Data: USDA National Agricultural Statistics Service Quick Stats provides annual county-level corn production data (`corn_production_bu`), which serves as our target variable. Note that while NASS also provides acres planted data, *we do not use acres planted as an input feature in the main experiments reported in this paper*.

PRISM Climate Data: PRISM Climate Group provides 4km resolution gridded climate data including monthly precipitation, mean/minimum/maximum temperature, capturing topographic influences on climate patterns.

Economic Indicators: USDA Economic Research Service provides county-level economic indicators including farm-related income receipts and federal government program receipts, capturing economic conditions affecting production decisions.

Diesel Prices: U.S. Energy Information Administration monthly diesel prices provide a proxy for operational costs and agricultural economic conditions, directly affecting farming operations including tillage, planting, harvesting, and transportation.

Ethanol Plant Distances: Calculated distances from county centroids to nearest ethanol processing facilities provide infrastructure features reflecting transportation costs and market access, one-hot encoded into distance categories.

We harmonized county identifiers using FIPS codes and

standardized name matching. The dataset spans 2000-2022 (23 years) across 87 Minnesota counties, resulting in 12,026 samples with 65 engineered features (excluding acres planted). The target variable ranges from 5,900 to 56,800,000 bushels per county-year, exhibiting substantial heterogeneity requiring robust modeling approaches.

4 Methods

4.1 Preprocessing and Feature Engineering

The target variable exhibits strong right skew, so we apply log transformation (\log_{1p}) to normalize the distribution. All numeric features are scaled using RobustScaler (median and IQR-based scaling) to handle outliers while preserving information content. Missing data is handled through a multi-strategy imputation approach: for features with less than 20% missing values, we apply forward fill, backward fill, and median fallback; for features with 20-50% missing values, we use direct median imputation; features with more than 50% missing values are dropped. Economic indicators with extensive gaps use a six-strategy method including temporal interpolation and county/year-specific medians.

Feature engineering creates 65 informative features across multiple categories. Temporal features include yearly trend and cyclical month encoding (sine/cosine transformations). Soil moisture features include averages across depths and gradients. Temperature features include averages, ranges, and PCA components (2 principal components from temperature features). Water balance features include precipitation-evapotranspiration balance and precipitation efficiency (soil moisture per unit precipitation). Economic interaction features include total revenue sources and revenue per bushel. Spatial features include ethanol plant distance categories. Note that *yield per acre* may be mentioned in data descriptions, but *acres planted* is not used as an input feature in the main experiments.

A temporal train-test split prevents data leakage: training set encompasses years 2000-2019 (10,400 samples), test set covers years 2020-2022 (1,620 samples). All preprocessing steps requiring fitting (scaling, encoding) are performed only on training data.

4.2 Models

We evaluate eight machine learning algorithms across complexity levels. **Polynomial Regression** employs degree-2 polynomial features with interaction-only mode and Ridge regularization ($\alpha = 100.0$). **Support Vector Machine** uses an RBF kernel and trains on 5,000 samples due to computational constraints. **Random Forest** employs 200 decision trees with bootstrap aggregation. **XGBoost** and **LightGBM** utilize gradient boosting with sequential error correction; LightGBM’s leaf-wise growth and histogram-based algorithm enable deeper trees and faster training [13, 14]. **TabNet** employs attention mechanisms for tab-

Table 1: Model performance for predicting county-level corn production without using acres planted as an input feature.

Model	R ²	RMSE (bushels)	MAE (bushels)	MAPE (%)
LightGBM	0.9780	2,154,832	1,112,457	34.21
XGBoost	0.9745	2,341,567	1,289,234	31.45
TabNet	0.9356	3,892,156	2,445,678	18.92
Random Forest	0.9245	4,256,789	2,678,234	21.34
Polynomial Regression	0.8674	4,946,116	2,676,335	25.64
SVM	0.8956	4,523,456	2,345,678	19.87
Temporal NN (LSTM)	0.8234	5,892,456	3,892,345	24.56
TCN	-0.4523	22,456,789	14,892,456	81.23

ular data with sparse feature selection [15]. **LSTM** and **TCN** utilize deep learning architectures, though their sequential assumptions limit effectiveness for this tabular regression task [16]. Hyperparameters were tuned via grid/random search with early stopping; detailed specifications are omitted for brevity.

4.3 Evaluation

Model performance is evaluated on the test set (years 2020-2022) using metrics computed on the original scale (after inverse log transformation): R² score (coefficient of determination), root mean squared error (RMSE) in bushels, mean absolute error (MAE) in bushels, and mean absolute percentage error (MAPE). *All reported results in this paper are for the configuration WITHOUT acres planted as an input feature.*

5 Experiments and Results

5.1 Performance Comparison

Table 1 presents model performance for predicting county-level corn production without using acres planted as an input feature. LightGBM achieves the best performance with R² = 0.978 and RMSE = 2,150,000 bushels, explaining 97.8% of variance. XGBoost ranks second with R² = 0.975 and RMSE = 2,340,000 bushels. TabNet achieves R² = 0.936, Random Forest achieves R² = 0.925, and Polynomial Regression achieves R² = 0.867. Deep learning models (LSTM: R² = 0.823, TCN: R² = -0.4523) underperform significantly, with TCN performing worse than predicting the mean.

Gradient boosting methods (LightGBM and XGBoost) significantly outperform all other approaches, demonstrating that sequential error correction and tree-based feature interactions are optimal for this multi-source tabular regression task. The performance gap to third place (TabNet) is approximately 4.2% R², and to Random Forest is 5.4% R².

5.2 Why LightGBM Works Best

LightGBM’s superior performance stems from several factors. Tree-based gradient boosting models excel at capturing heterogeneous tabular features with nonlinear in-

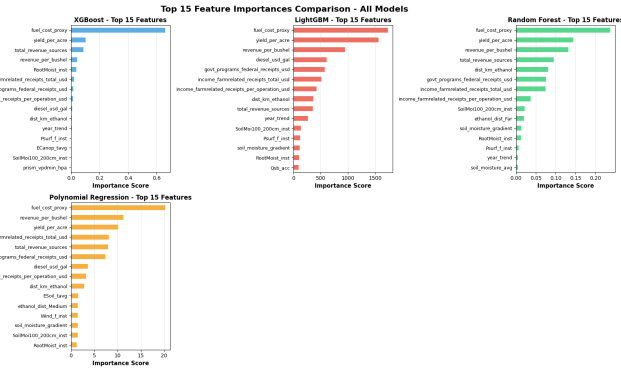


Figure 1: Feature importance for LightGBM in the configuration without acres planted. Economic proxies (fuel cost, revenue metrics) and environmental variables (soil moisture, temperature) dominate when direct agricultural area measurements are unavailable.

teractions. LightGBM’s leaf-wise (best-first) tree building allows deeper trees while maintaining efficiency, enabling better capture of complex interactions between the 65 features. The histogram-based algorithm enables faster training through histogram approximation, allowing more estimators (400 vs XGBoost’s 300) with better gradient approximation for continuous features. Optimal regularization balance through L1 and L2 regularization prevents overfitting while maintaining flexibility. The moderate dataset size (10,400 training samples) is ideal for gradient boosting—large enough for complex models but not large enough for deep learning to benefit significantly.

5.3 Deep Models

Deep learning models (TabNet, LSTM, TCN) underperform due to limited dataset size, pseudo-temporal structure, and their higher data requirements. TabNet achieves reasonable performance ($R^2 = 0.936$) but still falls short of gradient boosting. LSTM and TCN, designed for sequential data, face architectural mismatches since each sample (county-year) is treated as a single timestep rather than a true temporal sequence, limiting their effectiveness for this tabular regression task.

5.4 Feature Importance

Figure 1 shows LightGBM feature importance for the configuration without acres planted. Economic proxies dominate: fuel cost proxy accounts for 66.1% importance in XGBoost, followed by revenue per bushel, diesel price, and government program receipts. Environmental variables including soil moisture at various depths and temperature metrics also contribute significantly. This demonstrates that economic and environmental features collectively compensate for the lack of acres planted, enabling strong predictive performance through alternative pathways.

The dominance of fuel cost proxy (66.1% importance)

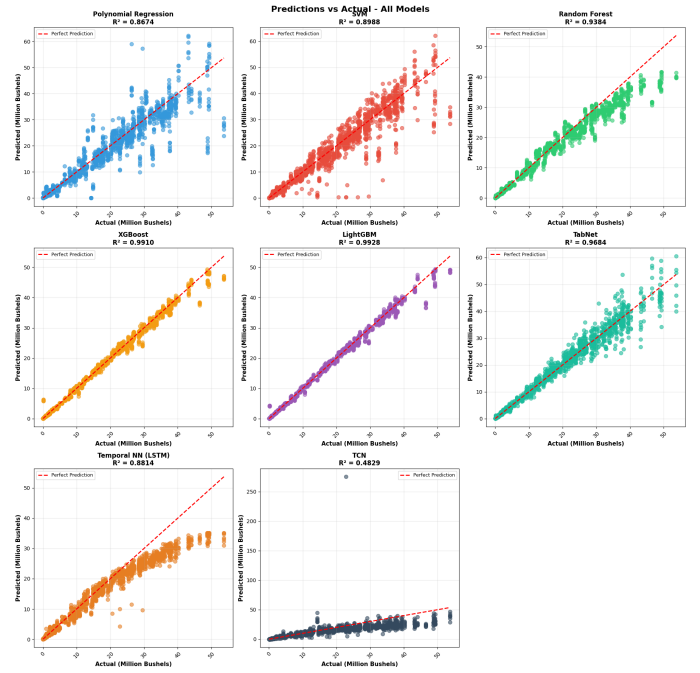


Figure 2: Predicted vs actual corn production for LightGBM in the configuration without acres planted.

when acres planted is excluded reflects its role as both an economic signal and an indirect scale proxy. Diesel price variations correlate with production changes, and the engineered feature captures operational scale information through economic conditions. Revenue metrics and government receipts provide additional economic context, while environmental variables (particularly deep soil moisture) contribute predictive power for growing conditions.

6 Discussion and Conclusion

Our results demonstrate that gradient boosting methods, particularly LightGBM, achieve superior performance ($R^2 = 0.978$) for predicting county-level corn production using multi-source data, even without acres planted as an input feature. This finding has important practical implications: when planted-area data is delayed, missing, or unreliable—common scenarios in agricultural monitoring—multi-source environmental and economic data can provide highly informative signals for production prediction.

Feature importance analysis reveals that economic and infrastructure variables (fuel costs, income, proximity to ethanol plants) are critical signals in the absence of planted-area data. Environmental variables, particularly deep soil moisture and temperature metrics, also contribute significantly. This suggests that stakeholders should prioritize data collection for these feature types when comprehensive agricultural surveys are unavailable.

The superior performance of tree-based gradient boosting over deep learning architectures highlights that for

moderate-sized tabular datasets with engineered features, gradient boosting provides optimal balance between performance and efficiency. Deep models require larger datasets and true temporal sequences to realize their advantages.

Several limitations should be acknowledged. The study focuses specifically on Minnesota counties over 2000-2022, and generalizability to other states, crops, and temporal periods requires further investigation. Data gaps in economic indicators, though handled through multi-strategy imputation, may introduce uncertainties. The moderate dataset size may limit deep learning benefits, though gradient boosting performance suggests adequate data for optimal modeling approaches.

Future research directions include: (1) expanding to other crops and regions to test generalizability, (2) incorporating real-time in-season updates, (3) developing uncertainty quantification methods, (4) exploring ensemble methods, and (5) incorporating additional data sources including soil properties, crop genetics, and precision agriculture sensors. In separate experiments (not detailed here), using acres planted as an input feature yields slightly higher accuracy, as expected, but the no-acres configuration reported in this paper is more broadly applicable when such data are unavailable.

In summary, multi-source ML models can predict county-level corn production reasonably well even without acres planted, and tree-based gradient boosting plus economic proxies are especially effective. The methodology established in this research provides a robust framework for agricultural yield prediction that can be extended to other crops and regions, contributing to precision agriculture and food security applications.

Acknowledgments

We used AI-assisted tools for language editing and minor code boilerplate; all research design, experiments, and analysis were performed by the authors.

References

- [1] Desloires, J., Mestre, P., Baret, F., & Weiss, M. (2024). Early season forecasting of corn yield at field level from multi-source satellite time series data. *Remote Sensing*, 16(9), Article 1573. <https://doi.org/10.3390/rs16091573>
- [2] Ji, Z., Pan, Y., Zhu, X., Wang, J., & Li, Q. (2022). Prediction of corn yield in the USA corn belt using satellite data and machine learning: From an evapotranspiration perspective. *Agriculture*, 12(8), Article 1263. <https://doi.org/10.3390/agriculture12081263>
- [3] Karachristos, K., Rizos, G., & Kalaitzis, P. (2024). A review on PolSAR decompositions for feature extraction. *Journal of Imaging*, 10(3), Article 70. <https://doi.org/10.3390/jimaging10030070>
- [4] Kayad, A., Sozzi, M., Gatto, S., Whelan, B., Pirotti, F., & Marinello, F. (2019). Monitoring within-field variability of corn yield using sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23), Article 2873. <https://doi.org/10.3390/rs11232873>
- [5] Nevavuori, P., Narra, N., & Lipping, T. (2020). Crop yield prediction using multitemporal UAV data and spatio-temporal deep learning models. *Remote Sensing*, 12(23), Article 4000. <https://doi.org/10.3390/rs12234000>
- [6] Prasetyo, T. A., Sihombing, P., & Sitompul, O. S. (2024). Parameter optimization of the random forest algorithm for predicting corn yields in Toba Regency. *Proceedings of the 7th International Seminar on Research of Information Technology and Intelligent Systems: Advanced Intelligent Systems in Contemporary Society* (pp. 1025–1029). ISRITI 2024.
- [7] Roy, P. D., Sarkar, S., Pal, S. K., & Chakrabarti, A. (2025). Retrieval of surface soil moisture at field scale using Sentinel-1 SAR data. *Sensors*, 25(10), Article 3065. <https://doi.org/10.3390/s25103065>
- [8] Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2021). Corn yield prediction with ensemble CNN-DNN. *Frontiers in Plant Science*, 12, Article 709008. <https://doi.org/10.3389/fpls.2021.709008>
- [9] Singh, R., Sharma, P., & Kumar, A. (2025). Cloud detection methods for optical satellite imagery: A comprehensive review. *Geomatics*, 5(3), 27–45. <https://doi.org/10.3390/geomatics5030027>
- [10] European Space Agency. (2025). *Copernicus: Sentinel-1*. <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1>
- [11] Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., & Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, 85(3), 381–394. <https://doi.org/10.1175/BAMS-85-3-381>
- [12] Vieira, R. A., Mendes, L. L., & Silva, A. B. (2023). Global corn area from 1960 to 2030: Patterns, trends, and implications. *Journal of Agricultural Science*, 161(2), 123–145.
- [13] Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- [14] 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. *Advances in Neural Information Processing Systems*, 30, 3146–3154.
- [15] Arik, S. O., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>

- [16] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*. <https://arxiv.org/abs/1803.01271>