Evaluating Climate Change Effects on Agricultural Yield Using Deep Learning

Presenter: Raven Mott and Calvin Kamara

Affiliation:

- Department of Computer Science, College of Engineering, Virginia State University, Petersburg, VA.
- Department of Computer Science , Bowie State University, Bowie, MD.

A comprehensive study investigating the application of advanced machine learning models to understand and predict climate variability effects on major U.S. crop yields.

Outline

- Introduction
- Motivation
- Objectives
- Related Work
- Comprehensive Dataset Integration
- Methodology
- Data Preprocessing and Feature Engineering
- Model Implementation Framework
- Visualization and Data Insights
- Results
- Discussion and Future Work
- Conclusion

Introduction

Agriculture fundamentally depends on climatic conditions—temperature, precipitation, and atmospheric moisture. Small deviations in seasonal norms can disrupt planting schedules, stunt crop development, and reduce yields significantly.

Climate differ has intensified both frequency and intensity of extreme weather events, raising concerns about food security, particularly in rainfed agriculture regions. Traditional agronomic models and linear statistical tools often fail to capture nonlinear interactions and temporal dependencies inherent in environmental systems.

Machine learning and deep learning models offer promising alternatives, uncovering hidden patterns and complex relationships between high-dimensional input variables and target outcomes. LSTM networks are particularly well-suited for modeling sequences and time-dependent phenomena, making them ideal for tracking climate impacts on yield over time.

Motivation

- Crop yields swing with temperature, precipitation, humidity, and extreme events—traditional models miss these nonlinear, multi-variable effects.
- Escalating climate volatility makes reliable yield forecasts essential for global foodsecurity.
- We apply deep learning and ensemble ML (Random Forest, XGBoost, Gradient Boosting,LSTM) to a custom dataset combining USDA county-level yields (2017-2022) withWRF-HRRR climate simulations.
- Study covers four key U.S. crops—corn, cotton, soybeans, winter wheat—at county-levelresolution nationwide.
- Inputs include raw weather variables plus engineered stress indicators such as hot-daycount and drought duration.

Objectives

- Merge USDA county-level yields (2017-2022) with WRF-HRRR high-resolution weather fields.
- Build monthly indicators—heat-day streaks, rain days, drought runs—to capture extreme events.
- Test tree-based ensembles (RF, XGBoost, GBM) against a sequence model (LSTM) for county-level yield.
- Pinpoint the climate variables that most influence predictions for each crop and region.
- Release a reproducible pipeline that growers, analysts, and policy-makers can adapt to future climate-change scenarios.

Related Work in Agricultural Prediction

1

Traditional Methods

Linear regression and ARIMA models provided interpretable results but were limited in handling complex, nonlinear relationships and high-dimensional datasets.

2

Ensemble Methods

Random Forests, Support Vector Machines, and Gradient Boosting improved accuracy but required extensive hyperparameter tuning for optimal performance.

3

Deep Learning Era

LSTMs demonstrated superior performance in sequence-based prediction tasks, including rainfall trends, drought likelihood, and crop yield forecasting.

This study builds on this growing body of work by integrating meteorological simulations with real-world crop data, comparing traditional ensemble methods with LSTM and CNN architectures for comprehensive agricultural yield prediction.

Comprehensive Dataset Integration

USDA Crop Yield Dataset (2017–2022)

- County-level annual yield data for corn, cotton, soybeans, and winter wheat
- Units: Bushels per acre (corn, soybeans, wheat), Pounds per acre (cotton)
- Comprehensive geographic coverage across major agricultural regions

WRF-HRRR Meteorological Dataset

- High-resolution daily and monthly climate variables
- Temperature metrics, precipitation, humidity, wind patterns
- Downward shortwave radiation and vapor pressure deficit

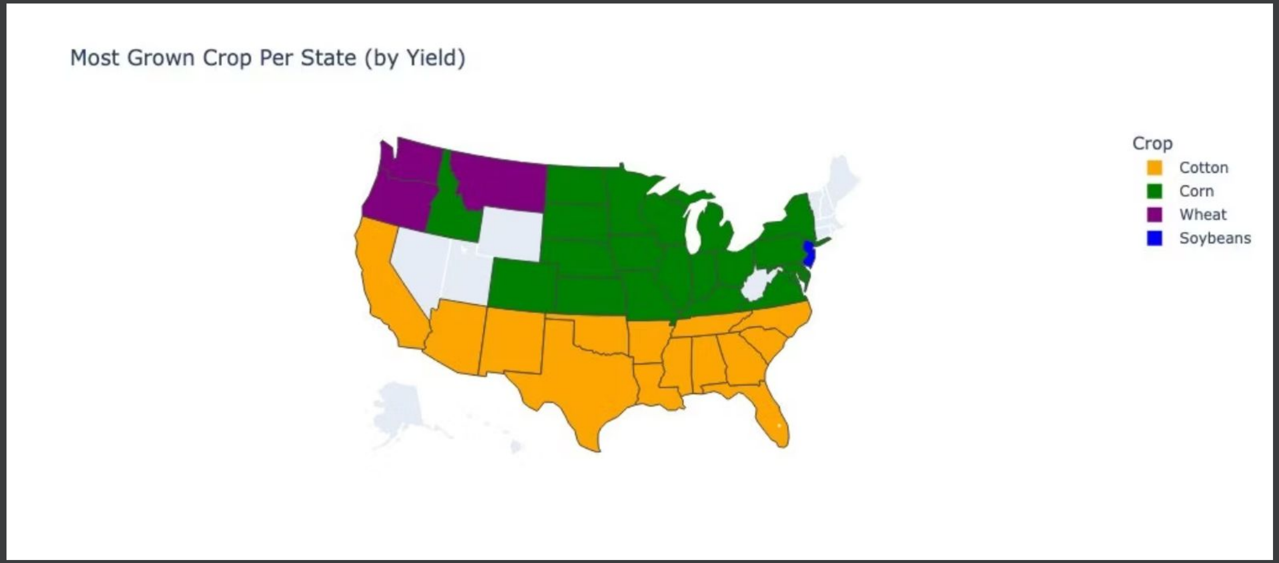
Derived Climate Indicators

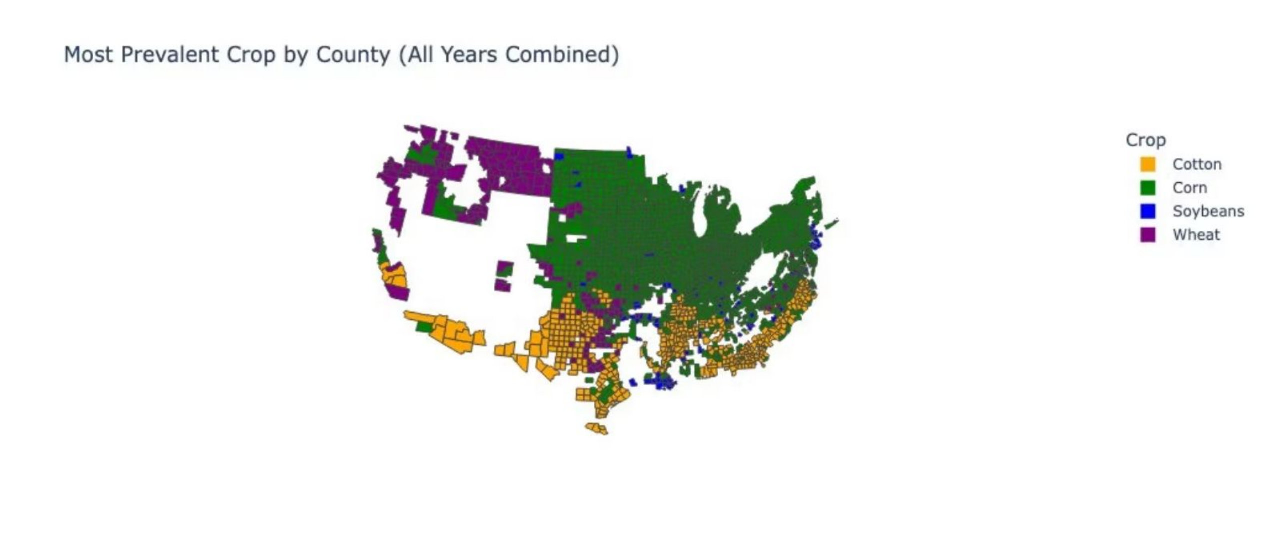
Hot days exceeding 33°C, consecutive drought and heat spell duration, and monthly climate variability metrics were engineered to capture agricultural stress conditions.

Raw Continuous	Raw Categorical	Derived Continuous	Derived Categorical
Precipitation (kg m**-2)	State	Max Consecutive Hot Days	
Relative Humidity (%)	FIPS	Max Consecutive Drought Days	
Wind Gust (m s**-1)	Month	Days with rain	
Wind Speed (m s**-1)	Date	Hot days	
Max Temperature (C),	Year	Cold days	
Min Temperature (C)			
Avg Temperature (C)			
U Component of Wind (m s**-1)			
V Component of Wind (m s**-1)			
Downward Shortwave Radiation Flux (W m**-2)			
Solar Radiation (MJ m**-2)			
Vapor Pressure Deficit (kPa)			
Corn_Yield_BU_ACRE			
Cotton_Yield_LB_ACRE			
Soybeans_Yield_BU_ACRE			
WinterWheat_Yield_BU_ACRE			

Spatial Distribution Patterns

- **Corn:** Concentrated in Midwest agricultural belt
- **Cotton:** Dominant in South and Southwest regions
- **Wheat:** Northern Plains agricultural zones
- **Soybeans:** Scattered across coastal and Midwestern counties





METHODOLOGY

Preprocessing:

- Cleaned and renamed yield columns
- Combined yearly files; calculated monthly climate indicators (heat days, rain, drought)

Feature Engineering:

- Encoded state/county/month
- Imputed missing data
- Standardized features

Integration:

- Merged yield and climate data by FIPS/month
- Produced monthly panel dataset

Modeling:

- Random Forest, XGBoost, GBM, LSTM
- Predicted crop yield per county; LSTM used for sequence learning

Evaluation:

- Metrics: RMSE, MAE, R^2
- Cross-validation & train/test splits

Data Preprocessing and Feature Engineering

Missing Value Treatment

Zero was put in place for all missing yield values

Categorical Encoding

State, county, and month variables were systematically encoded using label encoding methods.

Standardization

LSTM requires standardized input for stable training, as neural networks are sensitive to input scale during gradient descent.**Tree-based models** (RF, XGBoost, GBM) don’t need scaling—they split data by thresholds, not magnitudes.

Feature Aggregation

Climate data was aggregated monthly to match temporal resolution of yield data, with derived stress indicators enhancing predictive capability.

Model Implementation Framework

Random Forest Regressor

Baseline ensemble model providing interpretable feature importance rankings and robust performance across diverse agricultural conditions.

XGBoost Regressor

Advanced gradient boosting implementation optimized for tabular data with sophisticated regularization techniques.

Convolutional Neural Network

1D convolution architecture applied to reshaped feature vectors, testing spatial pattern recognition capabilities.

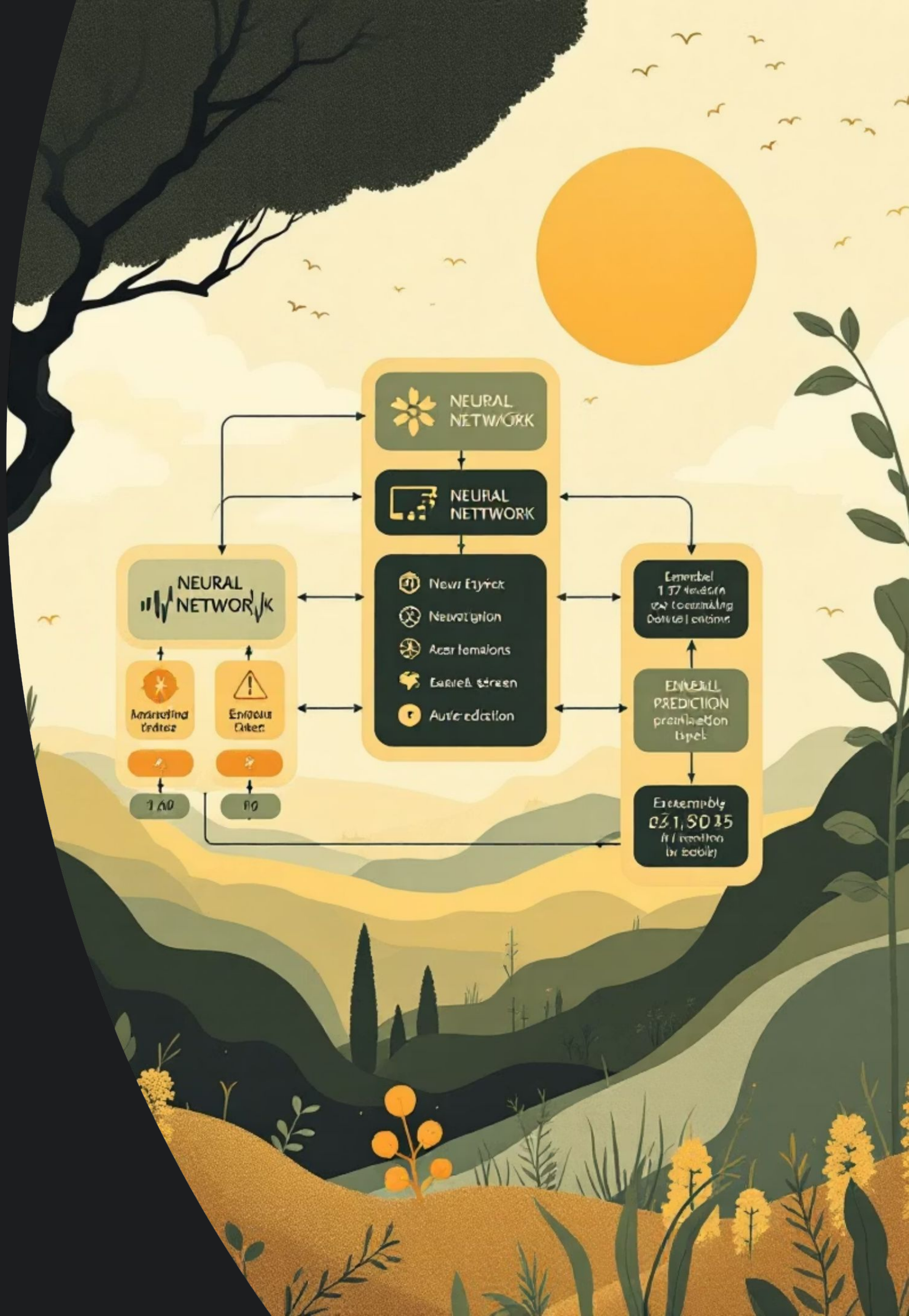
Long Short-Term Memory

Sequential neural network architecture designed to capture temporal dependencies and long-term climate-yield relationships.

GBM

Tree-based boosting model similar to XGBoost but with more basic hyperparameter tuning.

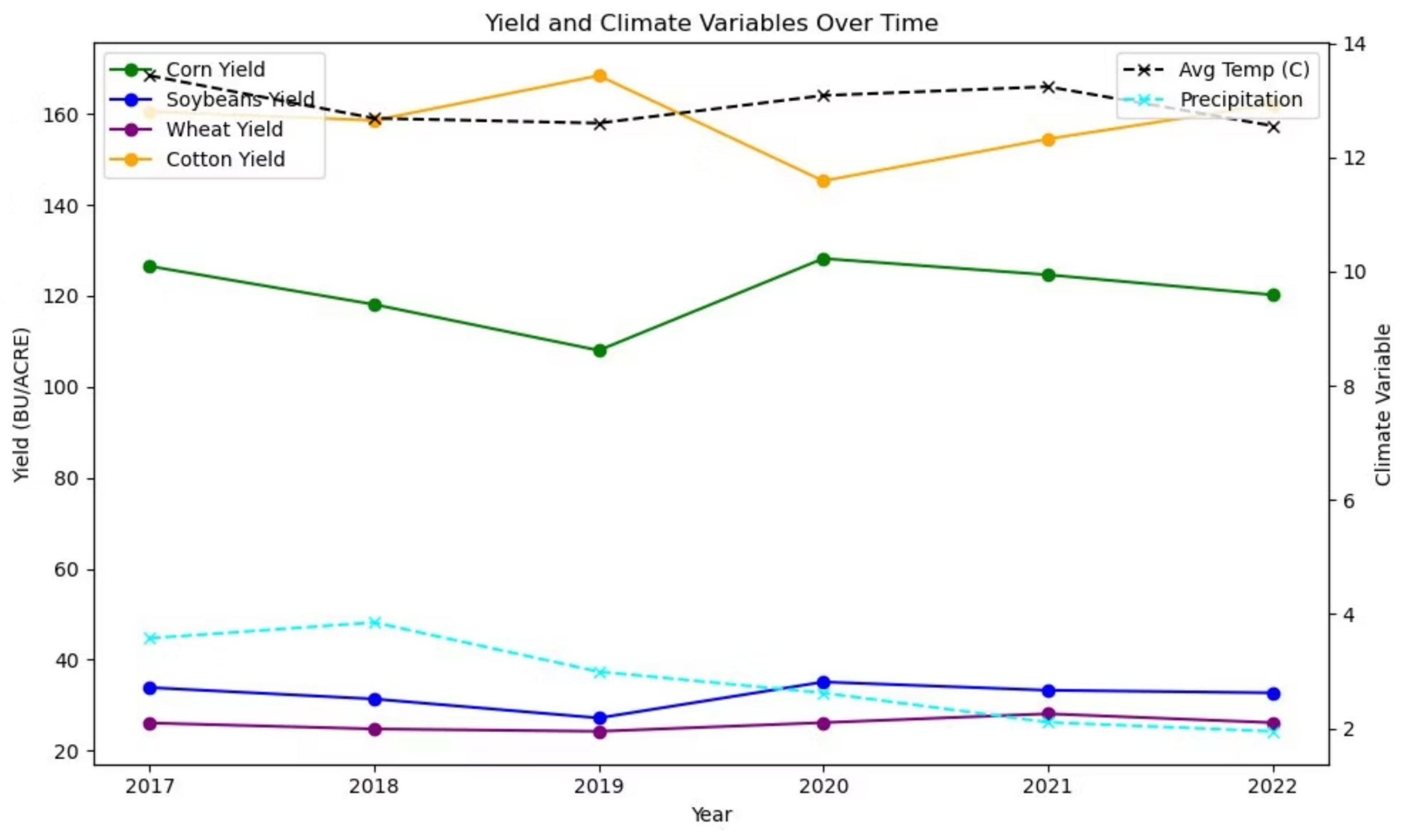
All models were trained separately for each crop using 80/20 train-test splits, with additional validation splits and early stopping regularization for deep learning architectures.



Visualization and Data Insights

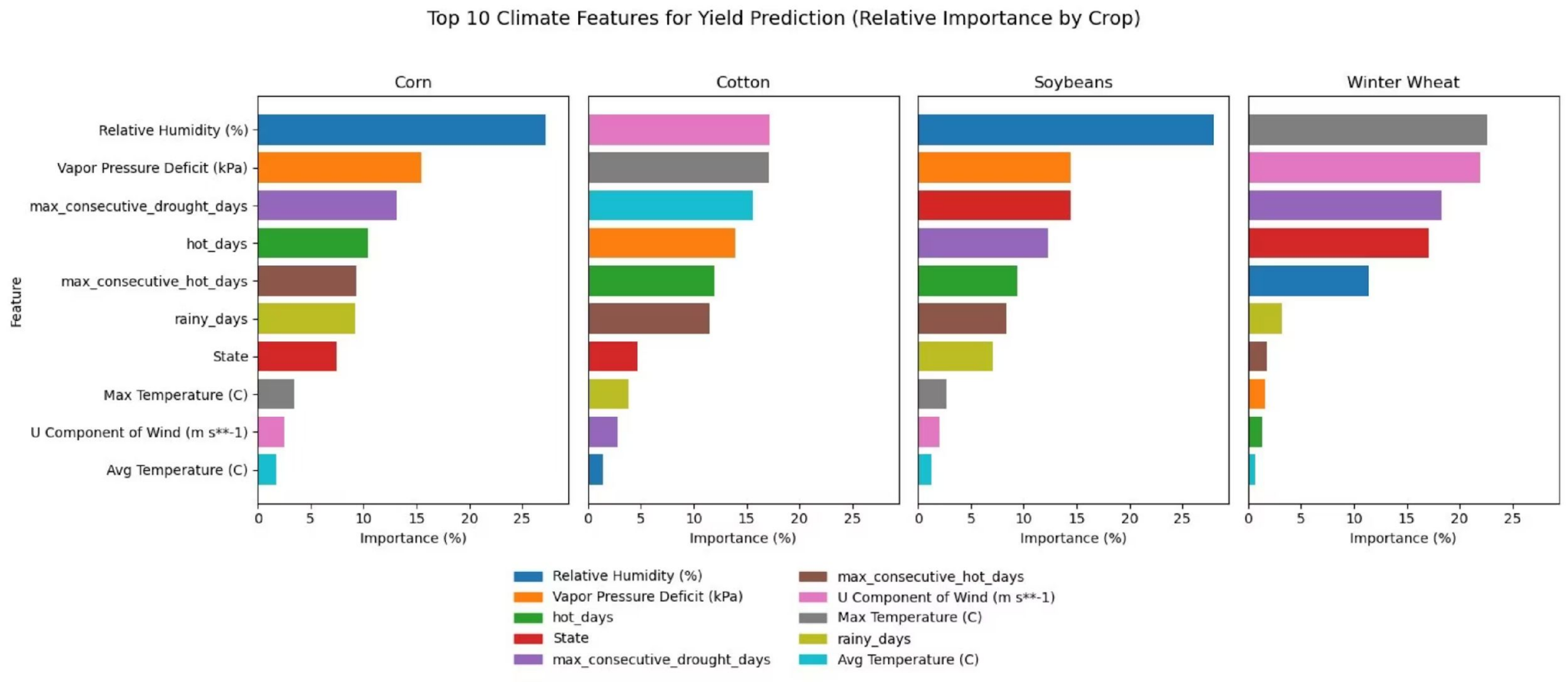
Temporal Climate Trends

Time-series analysis revealed declining precipitation patterns from 2018–2022 alongside relatively stable temperature regimes, indicating increasing drought stress conditions.



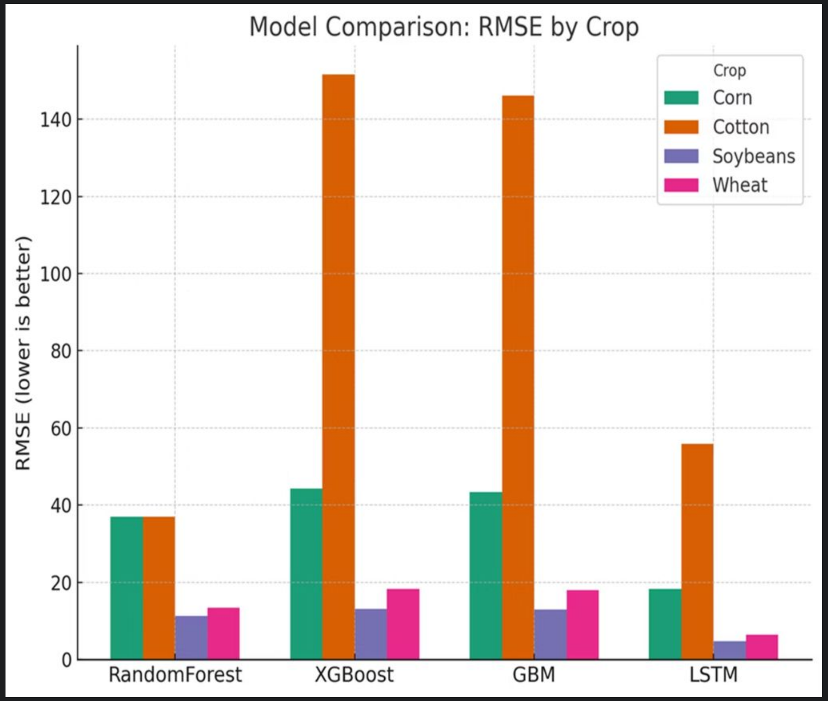
Feature Importance Analysis

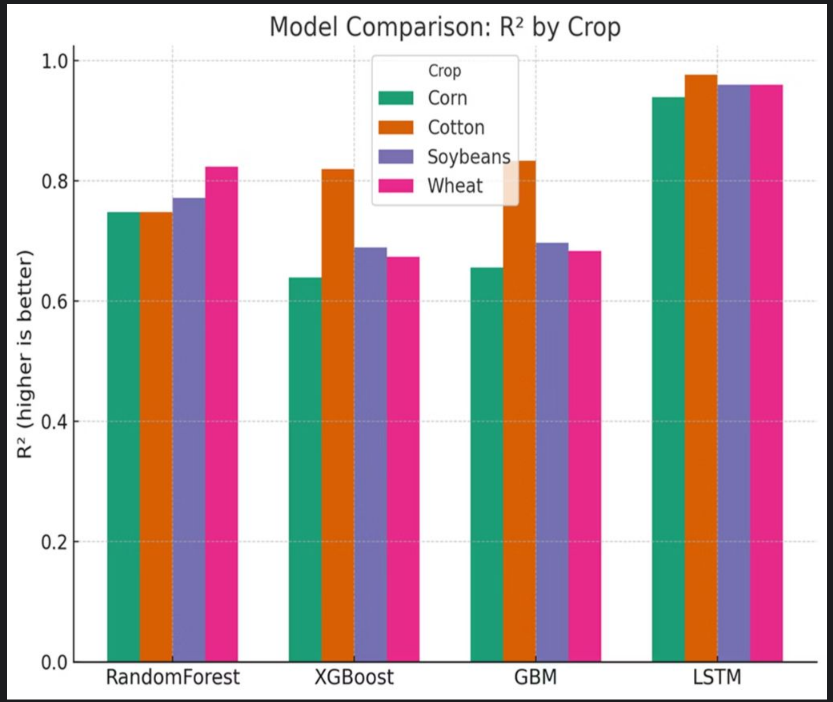
Bar charts identified relative humidity, vapor pressure deficit, and hot day frequency as the most influential predictors across multiple crop types.

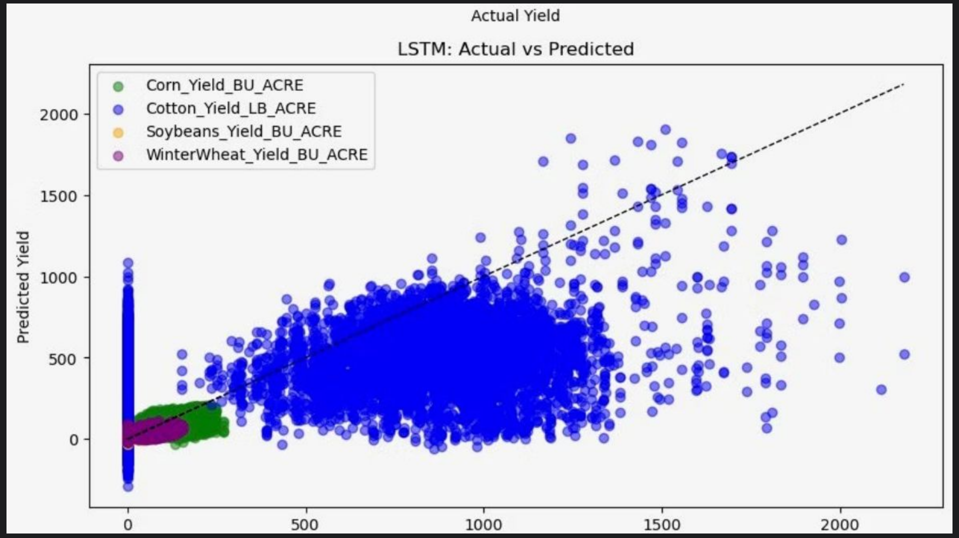


Results: LSTM Dominance in Yield Prediction

Scatterplot analysis of actual versus predicted yields demonstrated LSTM models' superior ability to track the ideal prediction line compared to alternative architectures.







Results: LSTM Dominance in Yield Prediction - Continue

18.29

Corn RMSE

$R^2 = 0.939$

55.83

Cotton RMSE

$R^2 = 0.976$

4.73

Soybeans RMSE

$R^2 = 0.960$

6.41

Winter Wheat RMSE

$R^2 = 0.960$

LSTM networks consistently outperformed all alternative models across all crop types, demonstrating superior capability in capturing temporal dependencies and nonlinear climate-yield relationships.

CNN models performed poorly ($R^2 = 0.30$ for corn) due to lack of spatial structure in tabular data. Random Forest and XGBoost models were adequate but lacked temporal modeling power, though their feature importance scores aligned well with agricultural domain knowledge.

Comparative Model Performance

The detailed evaluation metrics further underscore the superior performance of LSTM networks in predicting agricultural yields across various crop types, significantly outperforming traditional machine learning and other deep learning architectures.

Method	Corn	Cotton	Soybeans	Winter Wheat
	RMSE ↓ / R^2 ↑ / Corr ↑ / NormRMSE (%)	RMSE ↓ / R^2 ↑ / Corr ↑ / NormRMSE (%)	RMSE ↓ / R^2 ↑ / Corr ↑ / NormRMSE (%)	RMSE ↓ / R^2 ↑ / Corr ↑ / NormRMSE (%)
RandomForest	37.03 / 0.748 / 0.869 / 30.49%	37.03 / 0.748 / 0.869 / 30.49%	11.25 / 0.772 / 0.881 / 34.64%	13.41 / 0.824 / 0.911 / 51.56%
XGBoost	44.38 / 0.639 / 0.804 / 36.54%	151.50 / 0.820 / 0.908 / 95.67%	13.13 / 0.689 / 0.832 / 40.46%	18.26 / 0.674 / 0.827 / 70.20%
GBM	43.34 / 0.656 / 0.816 / 35.67%	146.06 / 0.833 / 0.915 / 92.23%	12.95 / 0.697 / 0.838 / 39.90%	18.01 / 0.683 / 0.834 / 69.23%
LSTM	**18.29 / 0.939 / 0.970 / 15.05%**	**55.83 / 0.976 / 0.988 / 35.25%**	**4.73 / 0.960 / 0.980 / 14.58%**	**6.41 / 0.960 / 0.980 / 24.65%**

Corn Yield Prediction

LSTM achieved an RMSE of **18.29** and an R^2 of **0.939**, demonstrating a substantial improvement over Random Forest (RMSE 37.03, R^2 0.748) and XGBoost (RMSE 44.38, R^2 0.639).

Cotton Yield Prediction

For cotton, LSTM delivered an RMSE of **55.83** and an R^2 of **0.976**, dramatically surpassing Random Forest (RMSE 37.03, R^2 0.748) and XGBoost (RMSE 151.50, R^2 0.820).

Soybean Yield Prediction

LSTM's performance was outstanding with an RMSE of **4.73** and an R^2 of **0.960**, far exceeding Random Forest (RMSE 11.25, R^2 0.772) and XGBoost (RMSE 13.13, R^2 0.689).

Winter Wheat Prediction

LSTM maintained its lead with an RMSE of **6.41** and an R^2 of **0.960**, making it significantly more accurate than Random Forest (RMSE 13.41, R^2 0.824) and XGBoost (RMSE 18.26, R^2 0.674).

Discussion and Future Directions

Key Findings

LSTM success stems from their ability to learn sequential dependencies and capture nonlinear interactions between climate variables and crop yield. Unlike CNNs assuming spatial correlations, LSTM models preserve crucial temporal relationships across months and seasons.

Critical Climate Variables

- Relative humidity and vapor pressure deficit: crucial for corn and soybeans
- Extreme heat sensitivity: particularly important for cotton yields
- Weaker climate correlations: observed in winter wheat production

Study Limitations & Future Work

- Absence of soil quality, pest pressure, and management practices
- Integration of satellite-derived vegetation indices
- Incorporation of economic variables and farmer-reported stressors

This research establishes LSTM networks as robust tools for climate-informed agricultural forecasting, providing foundation for enhanced food security planning under changing environmental conditions.

Discussion and Future Directions - Continue

- **LSTM leads by a wide margin.** Sequence learning cuts RMSE 40-60 % relative to ensembles and lifts R^2 above 0.93 for all four crops. The gain is largest for cotton, whose yield responds to multi-week heat spells captured only by the temporal model.

- **Tree models aren't bad—just static.** RF, XGB, and GBM track corn and wheat reasonably well but miss yield swings tied to successive hot or wet months, highlighting the value of temporal context.

CONCLUSION

- We built the **first county-level U.S. yield predictor** that merges USDA yields with sub-daily WRF-HRRR weather and engineered stress metrics.
- **LSTM outperforms** Random Forest, XGBoost, and GBM across all crops, proving that temporal dynamics matter for yield under climate variability.
- Key climate drivers differ by crop—insight that can steer targeted adaptation (e.g., drought-tolerant cotton in the Southeast, humidity management for soybeans).

ACKNOWLEDGEMENTS

This research was made possible through the support of the **MS-CC Undergraduate Summer Research Internship Program**, in partnership with **Fisk University** and **Meharry MedicalCollege**. We would like to thank our mentos and research coordinators for their guidancethroughout the project. Special appreciation goes to the developers and curators of the **USDACrop Dataset** and **WRF-HRRR climate simulations dataset**, which formed the foundation this study.

REFERENCES

1. Hu, T., Zhang, X., Khanal, S., Wilson, R., Leng, G., Toman, E. M., ... & Zhao, K. (2024). Climate change impacts on crop yields: A review of empirical findings, statistical crop models, and machine learning methods. *Environmental Modelling & Software*, *179*, 106119.
2. CropNet/CropNet dataset (2017–2022)—comprising Sentinel-2 imagery, WRF-HRRR computed data, and USDA crop yields—hosted on Hugging Face.
3. USDA (United States Department of Agriculture). County-level crop yield and production data for corn, cotton, soybeans, and winter wheat (2017–2022).
4. WRF-HRRR (Weather Research & Forecasting-based High-Resolution Rapid Refresh) climate model outputs, daily & monthly meteorological variables (2016–2022).
5. Iqbal, N., Shahzad, M. U., Sherif, E. S. M., Tariq, M. U., Rashid, J., Le, T. V., & Ghani, A. (2024). Analysis of wheat-yield prediction using machine learning models under climate change scenarios. *Sustainability*, *16*(16), 6976.

Questions?

We welcome your questions regarding our research and findings. Please feel free to ask!