Predicting Future Crop Yield Under Climate Change with ML Methods

Ctrl+Alt+Deheat

May 8, 2025

Abstract

Climate change threatens global food security by altering crop suitability patterns. To address this, we developed Random Forest, XGBoost, and LSTM models for predicting future crop viability using historical climate and soil data. Our analysis reveals that under certain constraints, all models are capable of obtaining accurate predictions of crop yields, with soil nitrogen and short-wave radiation being key predictors. This work provides a scalable framework for adapting agriculture to climate uncertainty, bridging traditional and deep learning approaches.

Background

Climate change poses a significant risk to global agriculture. This risk stems from a dual challenge: the rising global demand for food supply under population surge, and the climate threats that jeopardize food security. This multifaceted challenge necessitates data-driven approaches to predict future crop suitability. In response to *The Future Crop Challenge* hosted on Kaggle, we developed and evaluated multiple machine learning models - Random Forest, XGBoost, and LSTM networks - to forecast crop viability under varying environmental shifts. We leverage RF and XGBoost for interpretability on tabular data, and LSTMs to model temporal trends. Our methodology integrates historical climate data, soil properties, and crop yield records to train robust predictive models that yield accurate predictions (low RMSE) with high generalization ability (high R^2).

Similar work has been conducted in the past to understand the impact of climate on agricultural outcomes. Ren et al. (2023) analyzed long-term panel data from over 150 countries to assess the effects of temperature, precipitation, and farm size on nitrogen use efficiency (NUE), fertilization, and nitrogen surplus. They estimated that global warming led to an average annual decline of 6.5% in crop yield and found that the impacts of temperature and precipitation on nitrogen use and losses varied significantly across regions. By combining historical data with counterfactual climate scenarios, they highlighted the role of farm size in improving NUE and reducing nitrogen surplus (Ren et al., 2023). Building on this foundation, our work applies machine learning models, including Random Forests, XGBoost, and LSTMs, on more local climate and soil data, to capture detailed interactions and improve crop yield predictions across varying environmental conditions.

Method

We utilized the datasets that were posted by the Agricultural Model Intercomparison and Improvement Project (AgMIP) Machine Learning team (AgML). The source is widely available on Kaggle.

The data consists of static and dynamic features for both maize and wheat. More specifically
We implemented several models, including both tree-based models and LSTM-based ANNs:

Tree-Based Models
In our study, we processed climate and soil datasets to create a structured feature matrix suitable for Random Forest and XGBoost. The climate data included daily values for maximum temperature (tasmax), minimum temperature (tasmin), precipitation (pr), and surface solar radiation (rsds) for each agricultural grid cell. Since tree-based models do not natively handle sequential inputs, we aggregated these daily time series into summary statistics per numeric variable: mean, maximum, minimum, and standard deviation. These summary features effectively capture seasonal trends, extremes, and variability that are known to influence crop development and yield.

We then concatenated these climate-derived summaries with soil texture class, which do not vary over time but significantly impact yield potential. All numeric features were standardized using z-score normalization to ensure consistency. Lastly, encoded year values were mapped to actual calendar years to support time-aware model validation and projection. This preprocessing pipeline resulted in a compact and interpretable feature set optimized for tree-based yield prediction models.

Using this structured feature set, we implemented RF and XGB models to estimate maize and wheat yields. To evaluate model performance under realistic conditions, we held out the year 2020 as a validation set and trained the models on all prior years. This allowed us to assess generalization to the unseen while preserving the temporal integrity of the dataset. For both models, we used 100 estimators to prevent overfitting. For XGBoost, we additionally set the learning rate to 0.1, maximum tree depth to 6, and subsample ratio to 0.8. These hyperparameters are good starting points that balance learning speed, model complexity, and generalization. Lastly, the models were used to generate yield projections under future climate scenarios from 2021 to 2100.

LSTMs

For our first attempt, we implemented an LSTM using feature engineering methods to feed into both our Unidirectional and Bidirectional LSTMs. We created the following features with our raw climate-related data:

- **GPP (Gross Primary Productivity)**: The total amount of carbon compounds produced by the photosynthesis of plants in an ecosystem in a given period of time
- **Heat Stress Days**: Counts of daily maximum temperatures exceeding 30°C, which are critical in maize and wheat due to their role in pollen sterility and grain filling reduction.
- **Frost Days**: Number of sub-zero days during sensitive growth phases, which can damage early leaf and root development or halt reproductive processes.

To improve spatial generalization, we also partitioned the dataset by global region (USA, South America, Europe, and others), allowing the model to focus on geographically coherent patterns.

The unsatisfying first attempt inspired our second attempt at implementing LSTM models, and led to a significant performance improvement over our initial trials. In contrast to the earlier implementation, which incorporated hand-engineered features like GPP, heat stress days, and frost days but struggled to generalize beyond the training set, this version relied on a streamlined feature pipeline and architectural simplification. We directly leveraged the full temporal resolution of daily climate variables (e.g., tas, rsds, pr) without extensive biological transformations, and used a lighter-weight LSTM architecture with a feedforward fusion of static soil properties.

Results

Random Forest Results

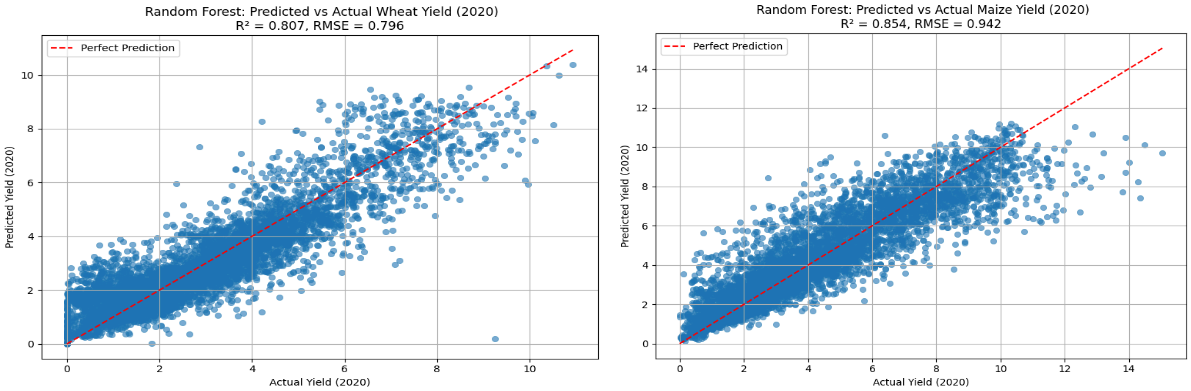


Figure 1: Random Forest Performance Results on Held Out Validation Year (2020)

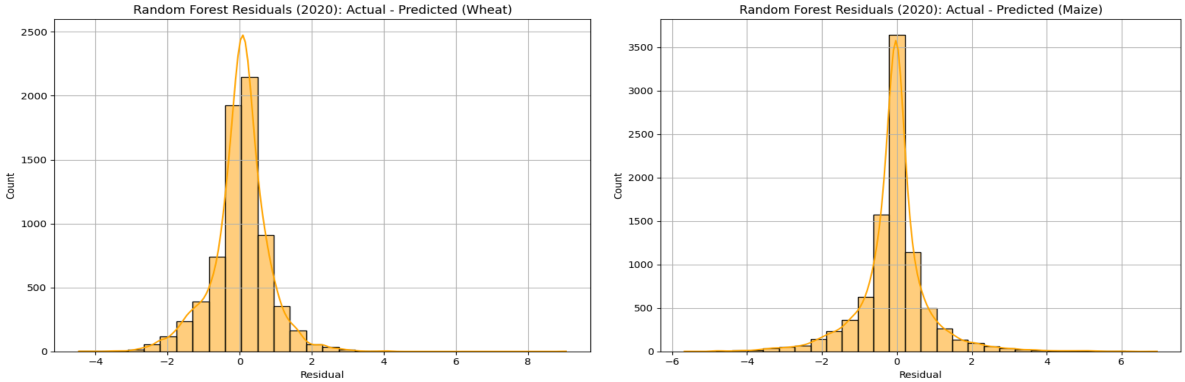
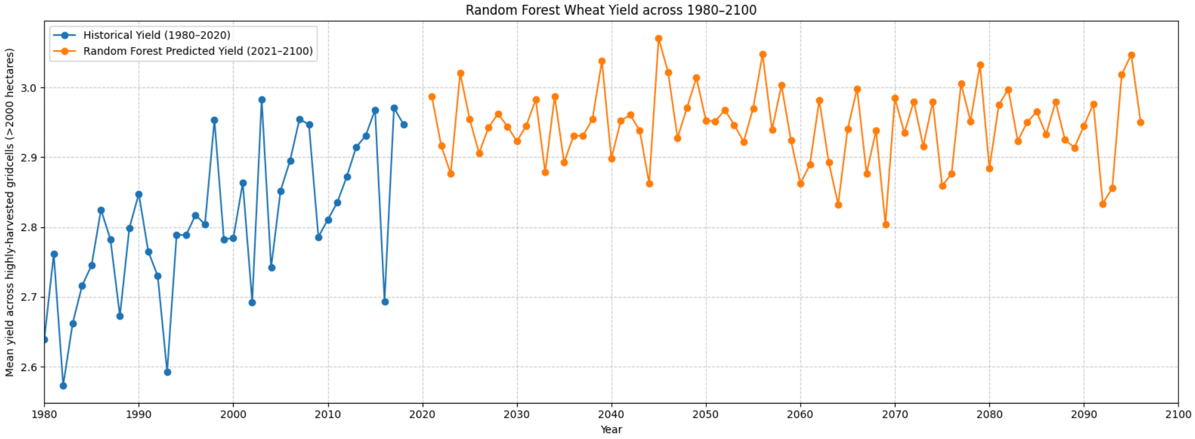


Figure 2: Random Forest Residuals on Held Out Validation Year (2020)

On the 2020 validation set, Random Forest achieved R² scores of 0.81 for wheat and 0.85 for maize, indicating strong predictive performance. Residual distributions for both crops were centered near zero, indicating unbiased predictions.



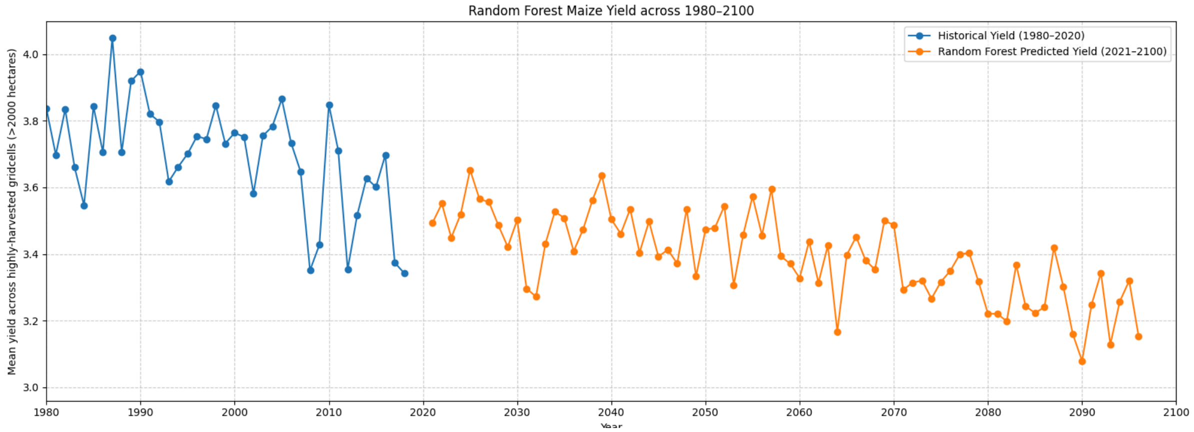


Figure 3: Random Forest Predictions of Future Yields

The future yield projections (2021–2100) suggest wheat yield remains relatively stable, but maize yield shows a gradual decline.

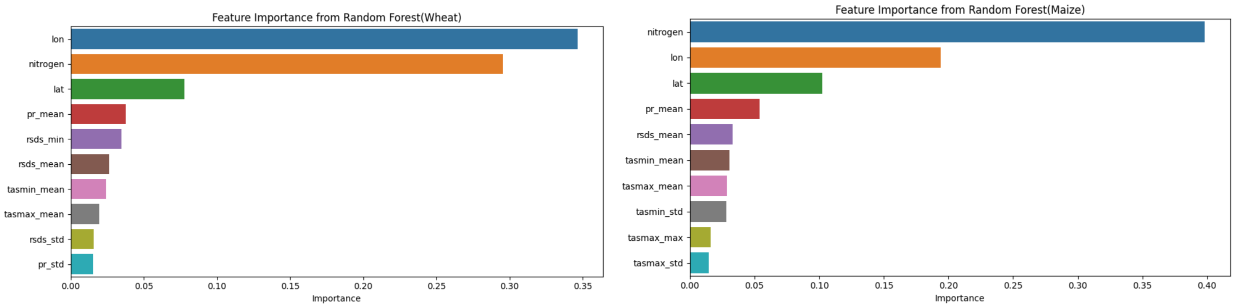


Figure 4: Random Forest Feature Importance

The bar charts show which features were most important for predicting wheat and maize yield in the Random Forest model. For wheat, Longitude ranks at the top, indicating strong spatial dependence, likely due to regional differences in soil or weather. Nitrogen application is also highly influential. Climate variables like precipitation mean and shortwave radiation contribute as well. For maize, nitrogen is the most important feature, which highlights maize's strong response to fertilizer level. Longitude and latitude follow, suggesting regional patterns are still relevant. Weather features like precipitation mean, shortwave radiation, and temperature stats also matter, confirming mthat aize is sensitive to climate.

XGBoost Results

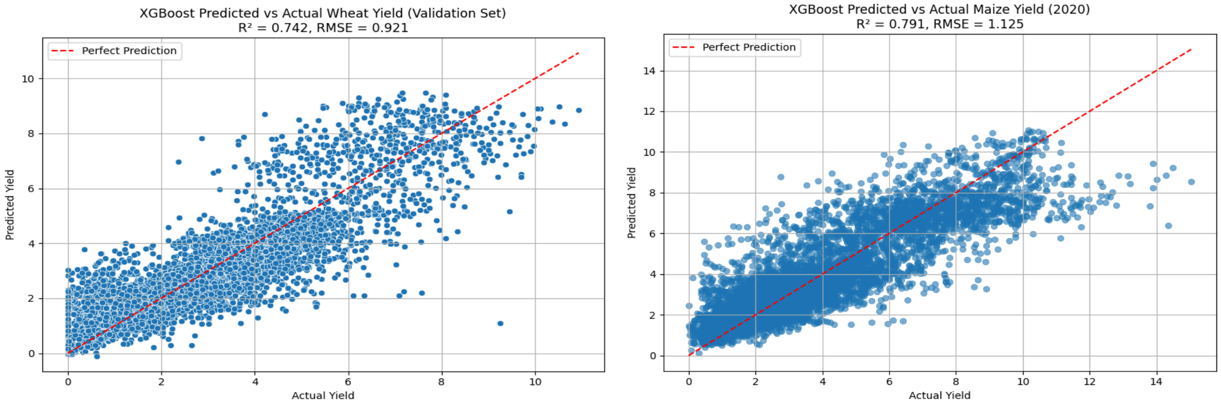


Figure 5: XGBoost Performance Results on Held Out Validation Year (2020)

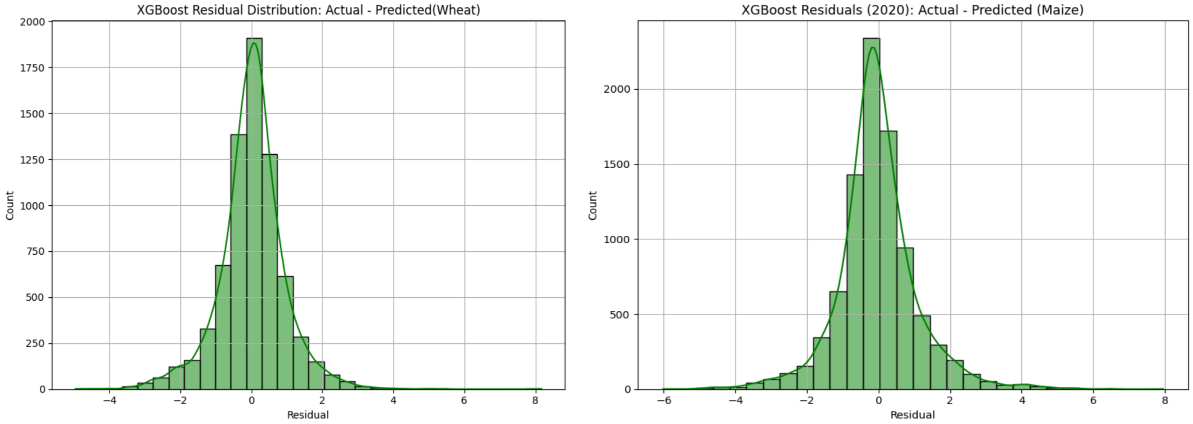
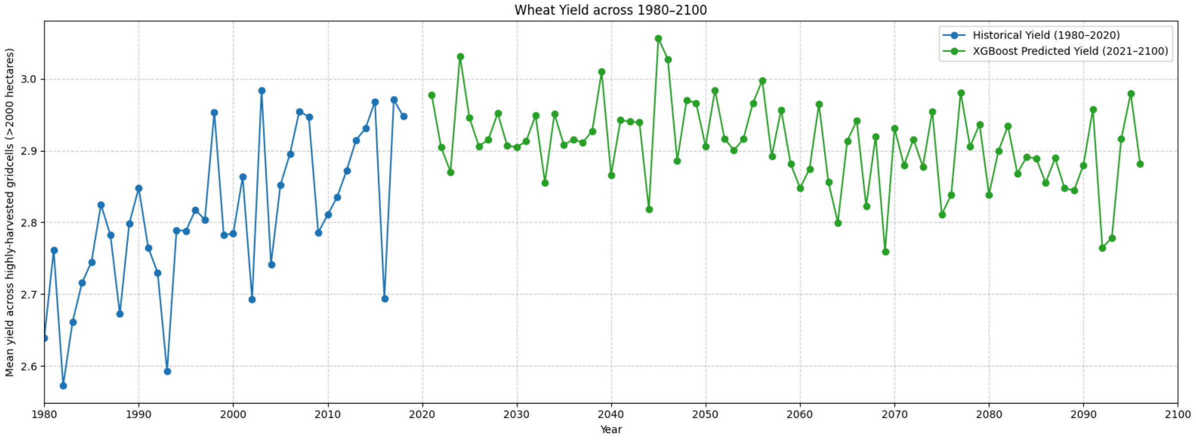


Figure 6: XGBoost Residuals on Held Out Validation Year (2020)

On the 2020 validation set, XGBoost achieved R² of 0.74 for wheat and 0.79 for maize. Residual distributions for both crops were centered near zero, indicating unbiased predictions.



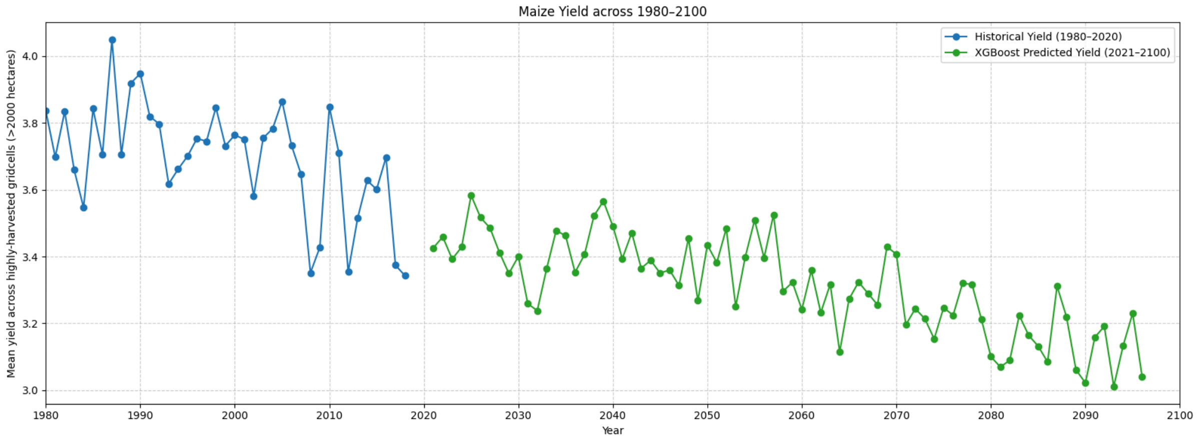


Figure 7: XGBoost Predictions of Future Wheat and Maize Yields

For the prediction of XGBoost, we can find a similar trend to the trend of Random Forest. The wheat yield remains stable, and the maize yield shows a clear decline.

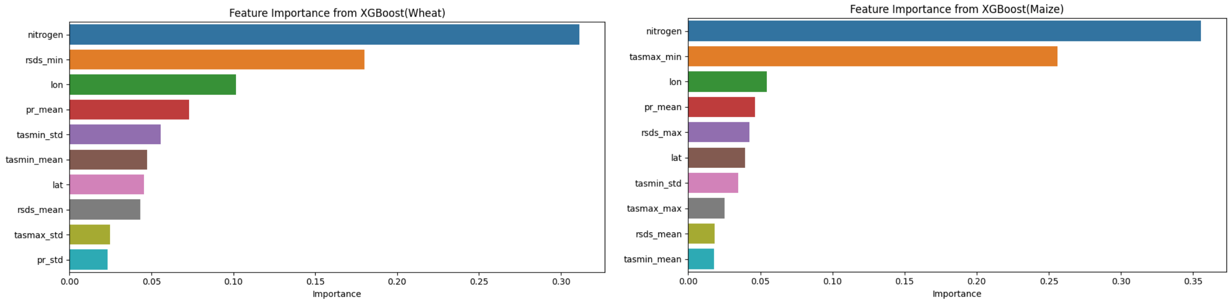


Figure 8: XGBoost Feature Importance

For both crops, nitrogen application is the most important feature, showing how critical nutrient input is for productivity. For wheat, shortwave and longitude also ranked high. For maize, minimum temperature emerged as a major factor, highlighting that it is sensitive to heat stress.

LSTM Results

First Attempt: Feature-engineered Unidirectional and Bidirectional LSTM

Initial unidirectional LSTM runs showed modest improvements in RMSE over baseline predictors. However, R^2 scores on the 2020 validation set were limited: 0.31 for maize and just 0.14 for wheat. Subsequent bidirectional LSTM trials yielded similar results, with significant gaps between training and validation metrics. This discrepancy suggests mild overfitting in maize and severe underperformance in wheat, likely due to limited year-to-year signal capture.

While LSTMs did reduce prediction error, they struggled to generalize climate variability over time, highlighting the difficulty of modeling seasonality and climate extremes in sparse agronomic datasets.

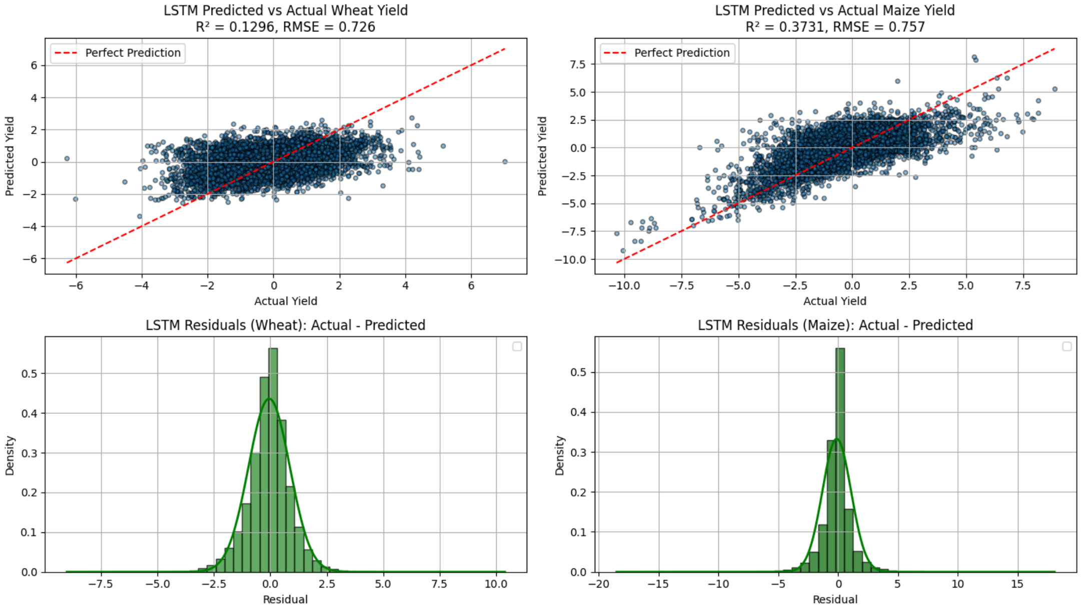


Figure 9: Unidirectional LSTM with Feature Engineering Performance on Held Out Year (2020)

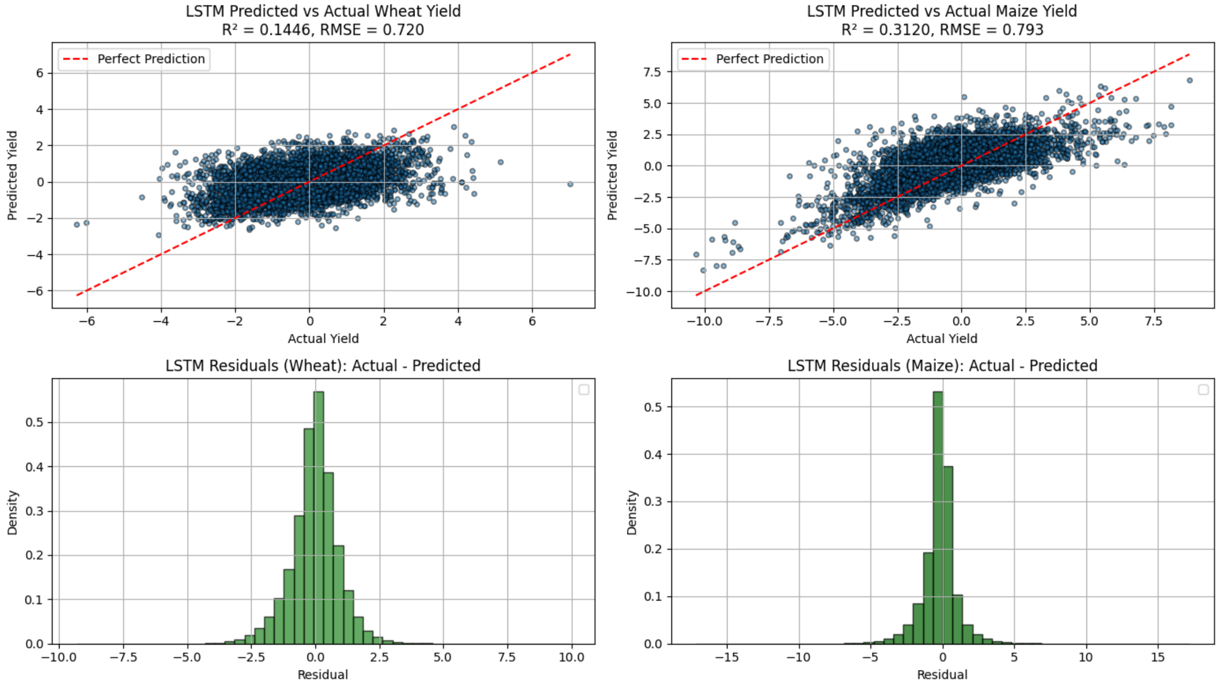


Figure 10: Bidirectional LSTM with Feature Engineering Performance on Held Out Year (2020)

Second Attempt: Model Refinement for Unidirectional LSTM

As shown in the validation scatter plot and residual histogram, this revised model achieved an **R^2 of 0.7236 and RMSE of 0.953 on the 2020 wheat validation set, R^2 of 0.7172 and RMSE of 0.929 on the maize validation set**, a substantial gain over the previous R^2 of 0.14 and 0.31. Predictions aligned more closely with actual values across a wide yield range, and the residuals were more tightly distributed around zero, suggesting reduced bias and improved generalization. These improvements are likely attributed to cleaner input scaling, a better balance between static and dynamic features, and regularized training with early stopping. While we didn't include hand-crafted stress indicators this time, the model appeared to learn these patterns directly from the raw time series, validating the potential of deep learning to discover latent agronomic drivers in climate sequences.

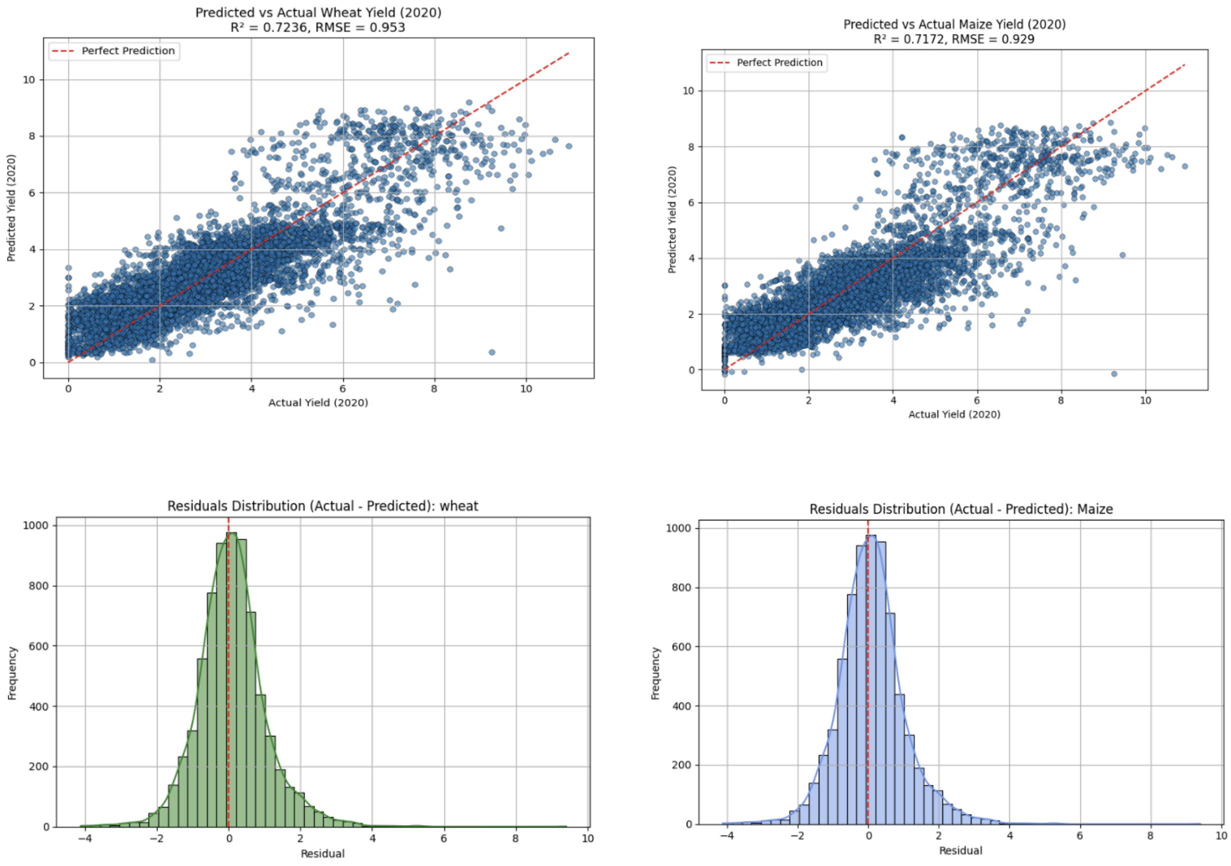


Figure 11: Refined LSTM Performance on Held Out Year (2020)

To facilitate the understanding of our RMSE, we compared them against a trivial “always‑predict‑the‑mean” baseline, our LSTM with raw data input achieved dramatic error reductions: for maize, RMSE fell from 2.754 t/ha to 1.116 t/ha ($R^2 = 0.7755$); for wheat, RMSE fell from 1.9056 t/ha to 0.9250 t/ha ($R^2 = 0.7396$). These 3-fold decreases in RMSE results place our models in the competitive range of published crop-yield emulators, confirming that deep sequence models can capture key weather–soil interactions driving inter-annual yield variability.

Discussion

Random Forest & XGBoost

For both Random Forest and XGBoost models, the residual distributions for both crops are centered near zero, indicating generally unbiased predictions. However, Random Forest achieved higher R^2 and lower RMSE than XGBoost, demonstrating stronger overall performance. Future yield projections suggest that wheat yields remain relatively stable under the high-emissions scenario, while maize yields show a gradual decline. This pattern aligns with empirical findings by Ren et al. (2023), who found that climate change exerts nonlinear and crop-specific effects on NUE and yield. In addition, wheat-growing regions often benefit from moderate warming, while maize-dominated regions are more vulnerable to heat and nutrient stress.

Feature importance analysis from both models supports this interpretation, with Nitrogen emerging as the most critical factor for both crops. This reinforces Ren et al.'s conclusion that nitrogen remains the dominant limiting input under climate change and that rising temperatures may further reduce crop NUE while increasing nitrogen losses.

Additionally, our XGBoost model highlighted minimum temperature as a key feature for maize. Similarly, Ren et al. (2023) reported that heat stress during critical stages like silking can significantly reduce yield and nitrogen use efficiency in maize-producing regions. Our results reinforce the need for targeted adaptation strategies, for example, developing heat-tolerant maize cultivars and improving nitrogen use practices in vulnerable zones.

LSTM

While bidirectional LSTMs are theoretically advantageous due to their ability to capture both forward and backward temporal dependencies, our experiments showed that they did not yield substantial performance gains over the unidirectional variant. In fact, validation R² scores remained low (0.13 for wheat, 0.37 for maize) and did not improve meaningfully across training configurations. One reason for this could be the nature of the climate-yield relationship itself: future climate values have no causal influence on current crop development, making backward temporal signals less useful or even misleading. Additionally, the bidirectional architecture may have increased model complexity without sufficient additional signal in the data, leading to overfitting and poor generalization, especially in wheat, where the data distribution was more variable and sparse.

In contrast, our second attempt using raw daily climate sequences directly as input to a simpler LSTM architecture yielded significantly better results (R² = 0.7172 on wheat validation). This outcome suggests that removing overly engineered features such as GPP, heat stress days, and frost days, while intuitive, may have inadvertently limited the model's capacity to learn more nuanced temporal patterns. By providing the full-resolution climate input, the model had more flexibility to internally learn feature representations relevant to yield prediction. This architecture also benefited from cleaner data scaling, better regularization (e.g., early stopping), and a streamlined fusion of static and dynamic features. The success of this approach reinforces the value of allowing sequence models to extract temporal dependencies directly from data, rather than relying too heavily on manual abstractions.

Author Contributions

Tree-Based Models:

- Random Forest: Simon and Xinkun
- XGBoost: Simon and Xinkun

LSTMs:

- Unidirectional LSTM with feature engineering: Ruoda and Richard
- Bidirectional LSTM with feature engineering: Richard
- Refined LSTM: Ruoda (Plotting done by Richard)

Project Report & Presentation:

Simon, Xinkun, Ruoda, Richard

Work Cited

Ren, C., Zhang, X., Reis, S., Wang, S., Jin, J., Xu, J., & Gu, B. (2023). Climate change unequally affects nitrogen use and losses in global croplands. *Nature Food, 4*(4), 294–304. https://doi.org/10.1038/s43016-023-00730-z