Machine Learning for Agricultural Supplies

Chelsea Wang*
Virginia Tech
Arlington, VA 22203
chelseawang@vt.edu

Yan Dong*
Virginia Tech
Arlington, VA 22203
yandong55@vt.edu

Abstract

In the past decade, American agriculture has been buffeted by many events, such as natural disasters, trade wars, and a pandemic. Such unprecedented events have created uncertainties throughout the food supply chain, starting at the farm for agricultural producers and culminating at the consuming household or international ports. The main drivers behind the shocks in agricultural productivity include weather-related events and other transitory factors such as transportation, shipping, and policy shifts. To quantify and protect against adverse events, we deployed machine learning models to measure and manage uncertainties proactively and evaluate multiple causal scenarios. We found that the DeepAR model had a good performance in predicting annual yields of corn, wheat, soybeans, and cotton in the United States. The overall MAPE (Mean Absolute Percentage Error) is 0.077, 0.097, 0.147, and 0.133 for the above four crops. Moreover, we analyzed the impact of extreme weather on crop yield. We found that the soybean is the most sensitive crop that responds to weather outliers mixed with temperature and precipitation anomalies. Given the advantage of the DeepAR algorithm for learning multiple time series, we predicted crop yields for individual states. Our results show that crop yield responses to climatic factors vary by geographical region and crop species. Such information can provide a guidance for future work that would forecast crop yields in different regions of the US.

1 Statement of the Problem

The United States is one of the largest agricultural producers and largest importers and exporters of agricultural commodities. In recent decades, agricultural productivity estimates in the US fluctuated considerably from year to year. Changes in agricultural production could affect food supply, food prices, and food consumption and introduce uncertainties in supply chains. Weather events and other transitory factors, such as transportation and crop prices, could be the main drivers of productivity changes. Ongoing climate change is expected to increase the frequency and severity of extreme weather events, particularly extreme heat and excessive rainfall [3]. Therefore, stresses from climate change are likely to suppress crop yields increasingly.

Machine learning (ML) has emerged as an important tool to study the impact of climate change on crop yields [5] [4] [8] [2]. Unlike conventional linear regression models, ML does not assume a specific shape between the predictive and responsive variables. ML models can show better predictions than linear regression models because they can capture nonlinear and complex relationships between input and output. We assume that these features of ML can help better predict the spatio-temporal variability of crop yield under the context of climate change. In this study, we aim to solve the two problems:

- Evaluate the accuracy of machine learning method on predicting crop yield.

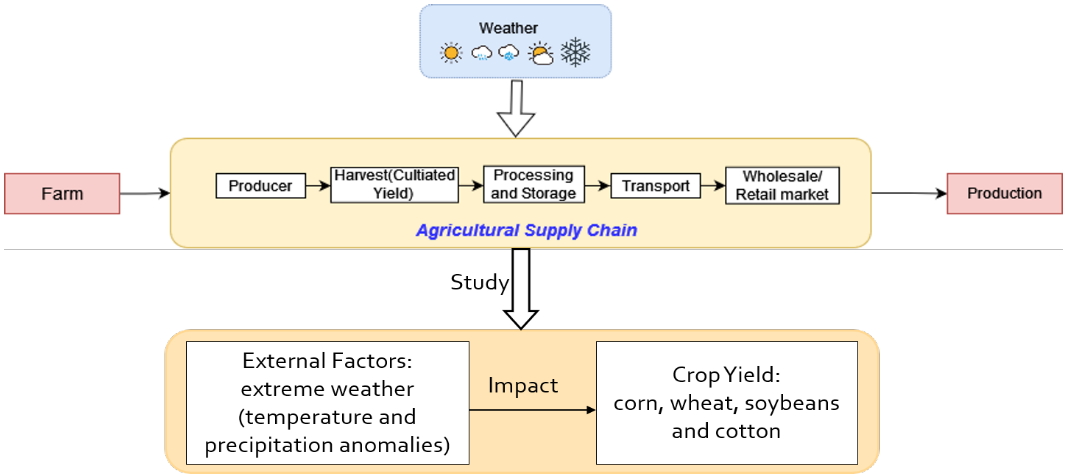

Figure 1: Agricultural Supply Chain

- Compare the model performance of predictions for different crops.

2 Data and Method

2.1 Data Collection

We used agricultural data and weather data between 1991 and 2020 in the United States. The agricultural data was downloaded from National Agricultural Statistics Service (NASS, USDA). We gathered state-level yield data measured in bu/acre for corn, wheat, soybean, and cotton. In addition, we collected acres planted and harvested which present conditions of the farmland stage. The weather data were obtained from U.S. Climate Divisional Database (NClimDiv, NOAA) and included monthly maximum temperature, minimum temperature, and precipitation over states [6]. We summarizes all data information in Table 1. Both agricultural and weather data are time series over the 30-year study period. For example, the time-series data we used in IOWA are shown in Figure 2.

Table 1: Data Dictionary

Columns	Description	Source	Supply Chain	Notes
state_name	name of states			
year	year			date value
month	month			date value
acres_planted	acres planted	NASS	producer	annual values
acres_harvested	acres harvested	NASS	harvested	annual values
buacre_yield	yield, measured in bu/acre	NASS	**target**	annual values
pcp_val	accumulated precipitation (inch)	NOAA	weather	monthly values
tmax_val	maximum temperature (fahrenheit)	NOAA	weather	monthly values
tmin_val	minimum temperature (fahrenheit)	NOAA	weather	monthly values

2.2 Feature Engineering

We assumed that the entire agricultural supply chain consists of farm, producer, harvest (cultivated yield), processing and storage, transportation, wholesale/retail market, and production, as shown in the figure 1. However, we encountered data missing issues when collecting data for all supply chain stages. After initial data quality checks, the following decisions were made to finalize the data set structure. In order to maximize the number of records after combining multiple data with various frequencies, the data panel was formatted to a monthly frequency for the best granularity. Crops planted, harvested, yield, and production were at the annual level; thus, the data was stretched to have repeated each month for a specific year to achieve a monthly dataset. Crop conditions and price data were originally weekly data and were subsequently aggregated to an average at the monthly level. However, not all states report crop conditions consistently. Transportation data recovered from

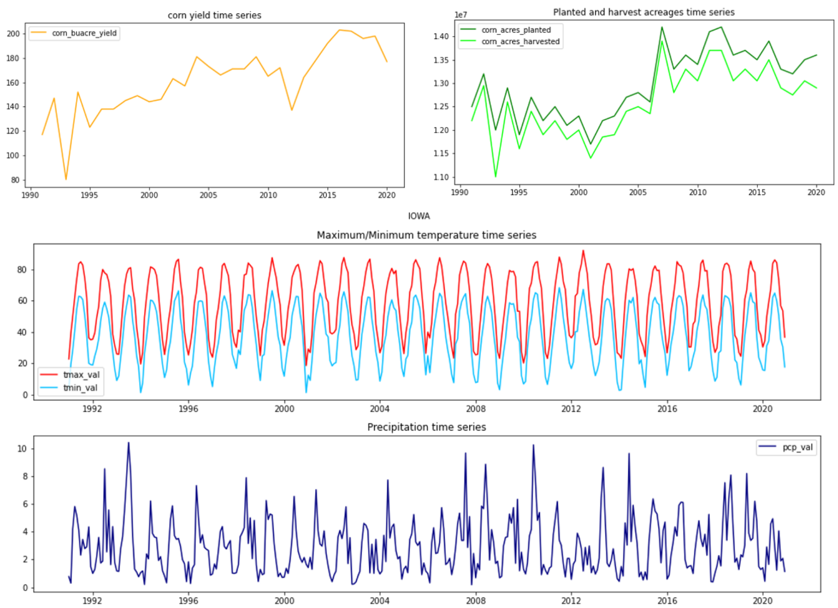

Figure 2: Time-series from 1991 to 2020 in IOWA

the USDA did not have enough coverage for 50 states or 30 years and thus was excluded from our datasets. From the data perspective, crop conditions and transportation are weak in the supply chain.

We performed feature engineering to generate extreme weather event indicators for monthly accumulated precipitation, maximum temperature, and minimum temperature across different states. Following Li et al [3], we identified extreme weather based on the precipitation/temperature z-score that measure how far a data point is away from the mean as a signed multiple of the standard deviation. Large absolute values of the Z-score suggest climate anomalies. Precipitation and temperature anomalies are defined as z-score $<$ -1.5 and z-score $>$ 2, respectively.

2.3 ML Algorithm

We used the DeepAR algorithm, a supervised machine learning method developed by Amazon Research to predict corp yields [7]. The DeepAR algorithm employs recurrent neural networks for probabilistic forecasting. The model implements long short-term memory cells [1] in an architecture that allows for simultaneous training of multiple related time series. Moreover, it implements an encoder-decoder setup commonly used in sequence-to-sequence models. The DeepAR model also takes a set of input covariates, which are assumed to be known for all points of the target time series. The model can predict the conditional distribution of the future of each time series given a certain point in time. Moreover, it learns a global model from time-series historical data. Since we have several similar crop yield time series across different states, we can benefit from training a single model jointly over all of the time series with DeepAR. The training data set was from 1991 to 2016, and the test data set was from 2017 to 2020. The model trained weather outlier and baseline datasets separately. The baseline model was also tested on the GAN data. The model hyperparameters were trained at 200 epochs, with an early stopping point and a gradient clip value of 0.1 to reduce the influence of exploding gradients. The best learning rate was 0.001 with 32 hidden layers.

2.4 Evaluation Metric

Mean absolute error (MAE) and mean absolute percentage error (MAPE) are used to determine forecast accuracy. Both MAE and MAPE are the most commonly used metrics. However, MAE is scale-dependent while MAPE is an average of percentage errors over a number of data points. Therefore, MAPE is preferred for model comparison.

$$
\begin{aligned}
MAE &= \frac{1}{n}\sum_{t=1}^{n} \mid Actual_t - Forecast_t \mid \\
MAPE &= \frac{1}{n}\sum_{t=1}^{n} \mid \frac{Actual_t - Forecast_t}{Actual_t} \mid
\end{aligned} \tag{1}
$$

3 Results and Analysis

3.1 Overall analysis of yield predictions

Our primary goal is to investigate the impacts of weather (minimum temperature, maximum temperature, and precipitation) and farmland (planted area and harvested area) on crop yields using the DeepAR algorithm. After training the model, we used MAPE to evaluate yield predictions of corn, wheat, soybeans, and cotton in 2020. In this paragraph, we report prediction results calculated from original data (weather outliers included) to compare the model performance across different crops and states. Table 2 shows the overall MAPE values of all states for four kinds of crops. We found that corn has the lowest MAPE (0.077), followed by wheat and cotton. The highest MAPE of 0.147 was found for soybean. Overall, the model performed best for forecasting corn yields than any other studied crops at the state level. These results indicate that the DeepAR method is an applicable choice available to predict state-level crop yields.

When using weather outliers smoothed datasets to predict crop yields based on the DeepAR model, we saw some differences in MAPEs of our studied crops compared with the evaluation results obtained from original datasets. The most apparent difference occurred in soybeans, which changed from 0.147 to 0.115. MAPE changes were very comparable for the other three crops. This comparison implies that soybean yield responds the most sensitively to weather outliers (mixed with temperature and precipitation anomalies). Using the DeepAR model, we found that external factors, such as extreme weather events, impose a perturbation in predicting yields of different kinds of crops. However, how a particular weather index impacts crop yield is still unclear based on our machine learning model.

Table 2: Mean Absolute Percentage Error of Yield Predictions in Test Dataset

Crop	Original with outliers	Weather outliers smoothed
Corn (41 states)	0.077	0.082
Wheat (35 states)	0.097	0.104
Soybean (28 states)	0.147	0.115
Cotton (17 states)	0.133	0.128

3.2 DeepAR Predictions across States

Given the DeepAR model could provide predictions for individual states, we made a further look at how this model performs geographically using our original datasets. Figure 3 and Figure 4 display actual and predicted yields of corn, wheat, soybeans and cotton, respectively, at the state level. In these two figures, crop yields are divided into five quantile classifications. We see that spatial patterns of observed and predicted yields are generally similar across states. For example, in yield observation and yield prediction maps (Figure 3 left), the states with high corn yields are gathered in the western U.S, and the state with low corn yields are mainly located in the central U.S. The result reveals that the DeepAR method could accurately forecast crop yields for individual states. Moreover, it shows the capability of the DeepAR algorithm to learn multiple time series.

When predicting corn yields in 41 states, the best prediction was in Kansas with an MAE of 0.969, and the worst prediction was in South Carolina with an MAE of 38.907 (Figure 3 left). For wheat predictions in 35 states, North Dakota had the lowest MAE of 0.714, and Idaho had the highest MAE of 17.776 (Figure 3 right). In addition, we found that the states with the lowest and highest MAEs were Pennsylvania and Kentucky for soybean (28 states, Figure 4 left), and Mississippi and Florida for cotton (17 states, Figure 4 right), respectively. Our results show that the predicting performance of our DeepAR model varies geographically, although this model has the capability of predicting multiple states' crop yields separately.

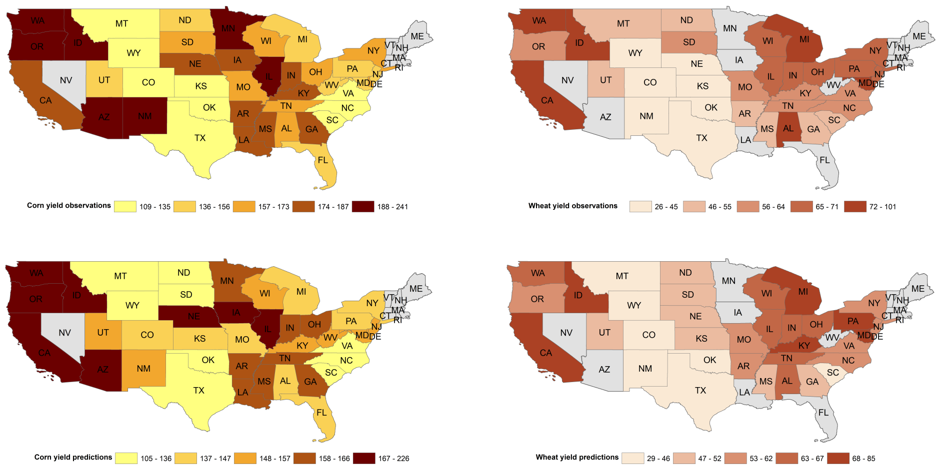

Figure 3: Corn (left) and wheat (right) yield observations and predictions in 2020 across states in the United States.

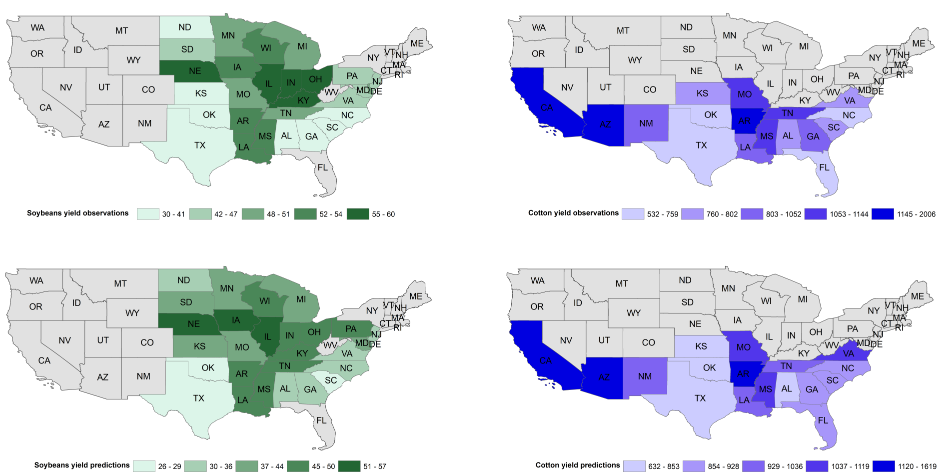

Figure 4: Soybeans (left) and cotton (right) yield observations and predictions in 2020 across states in the United States.

4 Conclusion

Given the above analysis, we can make three conclusions. Firstly, the DeepAR method had a good performance in predicting time-series yields of corn, wheat, and soybeans at the state level in the U.S. Secondly, temperature and precipitation outliers can impact the accuracy of our model for predicting crop yields, which indicates the effects of extreme weather on agricultural production and supplies. Lastly, our model shows that crop yield responses to climatic factors vary by geographical region and crop species.

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[2] Saeed Khaki and Lizhi Wang. "Crop yield prediction using deep neural networks". In: *Frontiers in plant science* 10 (2019), p. 621.

[3] Yan Li et al. "Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States". In: *Global change biology* 25.7 (2019), pp. 2325–2337.

[4] Maitiniyazi Maimaitijiang et al. "Soybean yield prediction from UAV using multimodal data fusion and deep learning". In: *Remote sensing of environment* 237 (2020), p. 111599.

[5] Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. "Crop yield prediction with deep convolutional neural networks". In: *Computers and electronics in agriculture* 163 (2019), p. 104859.

[6] NOAA. *NOAA Monthly U.S. Climate Divisional Database (NClimDiv)*. 2022. URL: `https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00005`.

[7] David Salinas et al. "DeepAR: Probabilistic forecasting with autoregressive recurrent networks". In: *International Journal of Forecasting* 36.3 (2020), pp. 1181–1191.

[8] Anna X Wang et al. "Deep transfer learning for crop yield prediction with remote sensing data". In: *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. 2018, pp. 1–5.