Satellite-based Soybean Yield Prediction in Argentina: a comparison between Panel Regression and Deep Learning Methods

Yuhao Wang [1], Kuishuang Feng [1,*], Laixiang Sun [1], Yiqun Xie [1] and Xiao-Peng Song [1]

[1]Department of Geographical Science, University of Maryland, College Park, Maryland 20742, United States

*Correspondence: kfeng@umd.edu

Abstract: The accurate prediction of soybean yield is vital for global food market stabilization and food security. Recent advancements in remote sensing technology have significantly amplified interest in leveraging satellite-based methods for predicting crop yield. These methods offer in-season yield estimates. By utilizing this timely information, decision-makers can formulate strategic, well-informed choices that preemptively mitigate potential food price hikes, ultimately bolstering food security. While simple regression models have been widely utilized for satellite-based yield prediction, researchers have recently begun to explore the use of deep learning algorithms. This study compares the performance of panel regression and deep learning models for in-season soybean yield prediction at the Department (county-equivalent) level in Argentina. Data sources include the latest soybean land use products and MODIS bi-weekly vegetation index products. Results indicate that deep learning models significantly outperform panel regression. Deep learning Long Short-Term Memory (LSTM) models, which incorporate attention mechanism and a series of peak NDVI images, generate more accurate and time-sensitive predictions. Among competing LSTM models, the one with attention mechanism applied to the entire growing season's NDVI data yields the highest prediction accuracy, with a Root Mean Square Error (RMSE) of 505.78 kg/ha and Normalized Root Mean Square Error (NRMSE) of 0.0726. The LSTM model with attention on the three highest NDVI images attains a satisfactory prediction accuracy (RMSE = 627.28 kg/ha, NRMSE = 0.089) six weeks prior to harvest. This study presents a robust model for predicting crop yields, promoting sustainable production of soybeans and facilitating knowledgeable choices among farmers and policymakers.

Keywords: Argentina; Deep Learning; LSTM With Attention; NDVI; Yield Prediction; Soybean

1. Introduction

As the global population grows and living standards improve, the demand for agricultural products has been increasing and this trend will continue in the future. Accurate and timely predictions of agricultural production are vital for ensuring food security worldwide. Output of crop production, the most crucial indicator of agricultural performance in growing season, has a profound impact on human society. Reliable and timely yield predictions are essential for crop mapping, market planning, and harvest management, but they remain challenging due to the complex environmental factors affecting crop growth (Pastor et al., 2019; Yu et al., 2016).

Soybean, one of the most important agricultural products as a source of protein (H. Tian et al., 2021), has gained global significance in recent years. Currently Argentina is standing as the world's third-largest producer and exporter of soybean. Sly (2017) reveals that export incomes of soybean and soybean products constitute a substantial 31.8% of the country's total export revenue in 2016. Furthermore, the FAO delineates that the global soybean production averaged 356 Mt from 2018 to 2021, with Argentina significantly contributing averaged 43 Mt during the same period (FAOSTAT, 2023). This underscores the imperative of precise soybean yield predictions for Argentina. However, compared with the top-two soybean producers – the US and Brazil – the volume of soybean yield prediction research emanating from Argentina is relatively sparse. While the USDA furnishes rich datasets to facilitate US soybean studies, accurate soybean maps become available only very recently (Song et al., 2021). This identifies a critical knowledge gap that needs to be addressed. The primary objective of this paper is to fill this gap.

Researchers have explored a variety of remote-sensing measurement to facilitate crop yield prediction, including the use of different vegetation indices, of which NDVI (Normalized difference vegetation index) has been the most used one. NDVI is a dimensionless index that captures the difference between visible red light and near-infrared regions of vegetation and is widely used to characterize the greenness of a study area (Weier & Herring, 2000). By incorporating vegetation indices into their models, researchers can consider the spectral characteristics of crops and their relationship with environmental factors, leading to improved prediction timeliness and accuracy.

Statistical-based methods have been commonly used to establish relationships between yields and selected explanatory factors, including NDVI from remote-sensing (Franch et al., 2019; Ji et al., 2021; Z. Tian et al., 2012). For example, Becker-Reshef et al. (2010) employed simple linear regression to predict winter wheat yields in Kansas and Ukraine using NDVI, achieving relatively satisfactory prediction accuracy. Cai, Yu and Oppenheimer (2014) employed a geographically weighted panel regression approach for corn yield prediction. Franch et al. (2015) improved upon the linear regression model developed by Becker-Reshef et al., applying it to the same study areas. Salehnia et al. (2020) utilized pooled panel regression for wheat yield prediction. However, these traditional regression methods exhibit limitations, as their models tend to be localized to specific regions owing to the constrained generalization abilities of linear regression models. This results in a lack of spatial generalization capability (Becker-Reshef et al., 2010; Franch et al., 2019).

With advancements of computer science, deep learning techniques have gained popularity for predicting food production. These techniques offer higher accuracy with less reliance on local survey data, making them an appealing choice in the field. Remote sensing data combined with deep learning techniques offers a better solution for yield prediction, as it provides a reliable and timely forecast (Cai et al., 2018; Khaki et al., 2020; Schwalbert et al., 2020; Sun et al., 2019; Xu et al., 2020). Long Short-Term Memory (LSTM) models, which are modifications of Recurrent Neural Network (RNN) models, are commonly used for time-series dataset classification and prediction, making it a suitable option for soybean yield prediction (Sun et al., 2019; H. Tian et al., 2021). Cai et al. (2018) introduced an in-season crop classification system using deep learning models, demonstrating higher accuracy than the USDA's Cropland Data Layer product. Xu et al. (2020) further reinforced the effectiveness of multi-temporal deep learning models in accurate crop mapping. Sun et al. (2019) leveraged a Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) framework to predict soybean yields at the county level in the United States, proving LSTM's utility in crop prediction. Continuing this trend, Tian et al. (2021) adopted an attention mechanism to forecast wheat yields in China, achieving a commendable average Root Mean Square Error (RMSE) of 502.71 kg/ha. Despite these advancements, there has been a lack of comprehensive quantitative comparison between deep learning models and traditional regression models, especially when applied to the same or similar datasets. This paper also addresses this knowledge gap. In more detail, the second objective of this paper is to evaluate a classic linear regression method, which has been widely used in the field of remote-sensing based crop harvest forecasting, against more contemporary deep learning models in the context of soybean in Argentina. Please note that this paper does not intend to make comparison across advanced machine learning techniques.

For achieving the above two objectives, we developed a set of deep learning models for in-season soybean yield predictions that utilized NDVI and other relevant and available data (Song et al., 2021). Then we compared the performance of these deep learning models with that of traditional panel regression models in terms of in-season soybean yield prediction at the Department (county-equivalent) level in Argentina, using remote sensing data captured during the growing season as inputs. This comparison is essential, as it not only evaluates the predictive accuracy of each approach but also examines their applicability in real-world agricultural practices.

The implications of our findings extend beyond academic interest. By establishing a clearer understanding of the comparative effectiveness of deep learning models versus traditional regression methods, our research contributes to the enhancement of crop yield prediction techniques. This is not only relevant for soybean production in Argentina but can also be applied to other crops and regions. By providing farmers, policymakers, and agricultural stakeholders with more accurate and reliable yield predictions, the application of the enhanced yield prediction techniques can support informed decision-making processes, leading to improved efficiency and sustainability in agricultural operations.

2. Study Area, Data and Methodology

2.1. Study Area

This study focuses on Argentina, a South American country that is renowned for its agricultural strength, particularly as one of the world's major soybean producers and exporters. The country also holds significant shares in other agricultural markets such as maize, wheat, beef, and sunflower seed. Argentina's climate is favorable for rainfed crop production, especially soybean, and it has dedicated a vast majority of its 166 million hectares of agricultural land to livestock farming and crop production. Experimental soybean plantations existed in the early 20th century, but commercial planting did not commence until the mid-20th century (Klein & Vidal Luna, 2021). Soybean cultivation has spread across the country, including in regions like the Pampas, and all provinces except Mendoza are now producing soybeans. According to Song et al. (2021), soybean plantation areas in Argentina increased from 11.4 million hectares in 2001 to 19.9 million hectares in 2015, with an average growth rate of 5.3% (0.6 million hectares) per year, before declining to 16.3 million hectares in 2019. Soybean plants in Argentina have two seasons, with the first season planted in November and harvested in April. The second season is part of the double wheat-soybean rotation, planted after wheat being harvested in late December and harvested in May. However, the planting area and total production of soybean's second season are significantly smaller than those in the first soybean season due to the relatively small extent of wheat-soybean double rotation lands. Thus, this study focuses solely on the first season of soybean. In 2020, the agricultural sector accounted for 6.1% of Argentina's GDP (World Bank, 2022).

Argentina has a rich historical legacy in soybean production, which has witnessed a considerable upsurge in cultivation across the country, driven by both agricultural innovation and growing international market demand. Notably, the record-high international soybean prices in the early 1970s played a crucial role in the expansion of soybean cultivation in Argentina (Schnepf et al., 2001). Over the period from 1970 to 2021, soybean production in Argentina has escalated substantially from 26,800 tons to 48,796,661 tons (FAOSTAT, 2023). Owing to its immense contribution as one of the largest soybean exporters worldwide, Argentina has become indispensable for the global food supply chain. Figure 1 illustrates the expansion of soybean plantations in Argentina from 2001 to 2019.

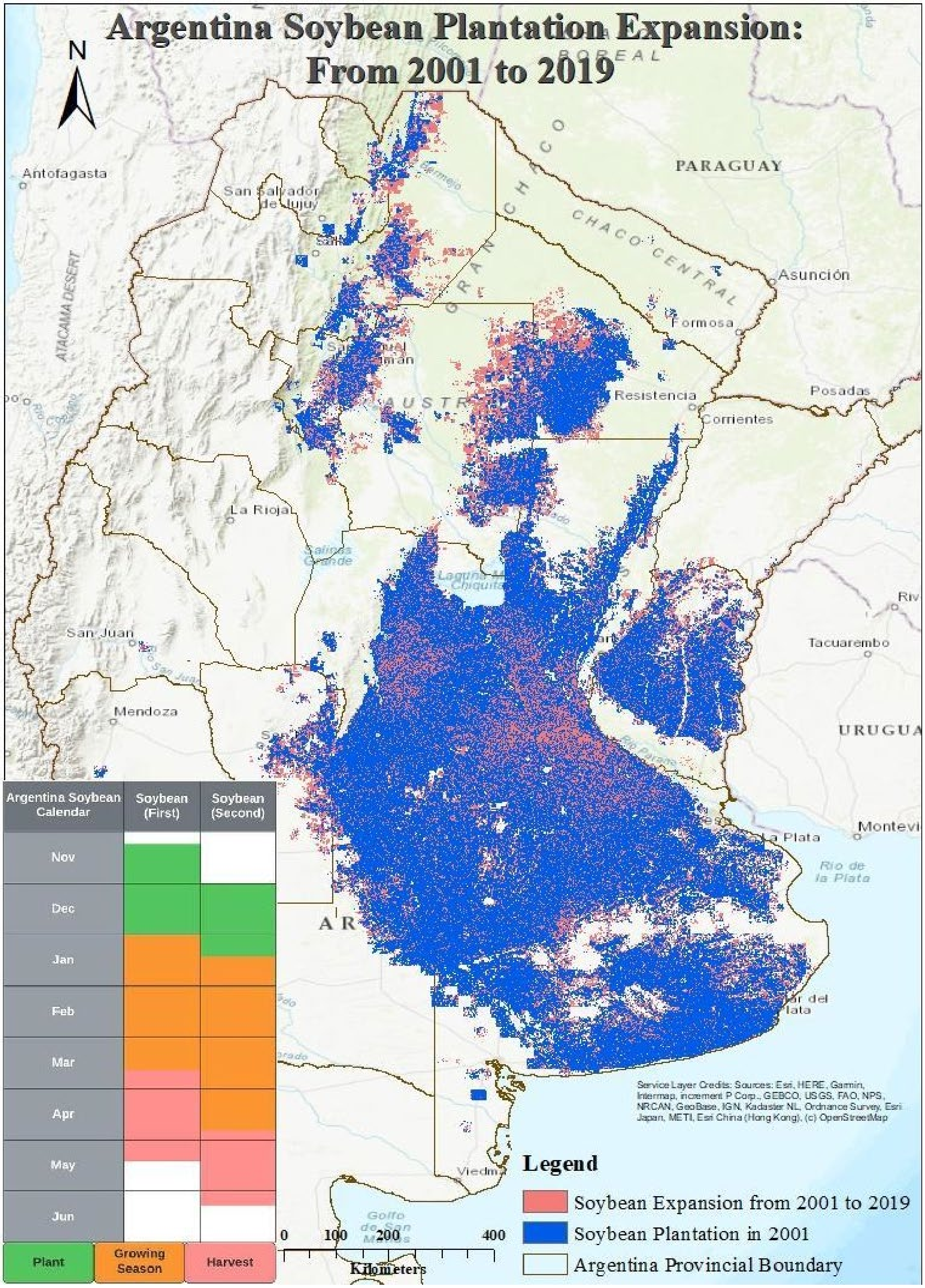


Figure 1. Argentina Soybean Plantation Expansion from 2001-2019, with a simple soybean crop calendar demonstrating two soybean seasons in Argentina.

2.2. Data collection and processing

The soybean land use dataset was produced by Song et al. (2021). Song et al. created a classification model using satellite data, machine learning and ground survey to precisely detect the presence of soybean crops throughout the South American continent. The model functions at a spatial resolution of 30 m and was utilized annually during the soybean growing season from 2000 to 2019. The map product has an overall accuracy of 96% based on a probability sample and in situ reference data. The soybean categorization map generated by the algorithm is a dependable indicator of soybean production because of the strong association between the crop regions identified in the high-resolution map and the actual soybean production.

A pixel qualifies as soybean if it undergoes a complete growth cycle within a single growing season and has a sufficient level of greenness in the spectral feature space. Therefore, the soybean pixels that have been mapped represent the cultivated fields that are farmed and have reached a stage where the crops can be harvested. Any crops that do not reach full maturity or exhibit reduced greenness as a result of abnormal weather conditions are excluded from the mapping process.

This study employed the MODIS (Moderate Resolution Imaging Spectroradiometer) product and the yearly South American soybean land-use product developed by Song et al. (2021) in conjunction with the Argentina departmental boundary data. Google Earth Engine (GEE) platform offers a variety of MODIS products, and this study utilized MOD13Q1.006, a terra vegetation indices product with a temporal resolution of 16 days and a spatial resolution of 250 meters. This product features two primary vegetation index layers, the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), as well as an additional pixel quality layer. Researchers process NDVI and EVI products based on MODIS imagery that has undergone bi-directional surface reflectance atmospheric correction and masking for water, clouds, heavy aerosols, and cloud shadow pixels (Didan et al., 2015).This study chose NDVI as the vegetation index for prediction. MOD13Q1.006 also includes four surface reflectance bands, namely Surface Reflectance Band 1 (Red), Surface Reflectance Band 2 (Near Infrared), Surface Reflectance Band 3 (Blue), and Surface Reflectance Band 7 (Mid-Infrared) (Didan et al., 2015). These bands were also employed as input parameters for the prediction model.

In this study, MOD13Q1.006 data for Argentina from 2001 to 2019 was extracted using the GEE platform based on its availability in the study area. To define the soybean land use areas in Argentina, the annual South American soybean land-use product published by Song et al. (2021)) was used as a mask. This product provides an annual classification of soybean land use for the entire South American continent between 2000 and 2019, based on Landsat imagery. This allowed the extraction of soybean land's NDVI products from the MODIS data only within the defined regions of soybean cultivation. During the aggregation process, the pixels were carefully selected based on their band summary QA, which was limited to 0, indicating data of high quality as stated in the MOD13Q1.006 user document (Didan et al., 2015).

To spatially integrate the MODIS NDVI data and soybean yield data, we employed an Argentina departmental administration boundary dataset that was published by the Food and Agriculture Organization of the United Nations. This boundary dataset was used to spatially summarize the soybean pixels' NDVI to the departmental level. We obtained the soybean yield dataset from the Argentina Ministry of Agriculture, Livestock and Fisheries, which provided departmental-level data for soybean production, including the first and second seasons of soybean harvest, total production, and yield production. The first season of soybean harvest was selected for this study since it is the primary contributor to Argentina's soybean production. We also spatially joined the selected first season soybean yields to the same departmental administration boundary dataset. The soybean yield dataset indicated that 306 departments planted soybean in the first soybean season between 2000 and 2019. For having sufficient data to train both panel regression and deep learning models, we opted to include 190 departments from nine different provinces that continuously cultivated soybean throughout the 20-year study period. Additionally, two out of the nine provinces only had a few departments with complete records, which were also dropped in order to fit the panel linear regression. As a result, we used a total of 183 departments in our study. The summarized datasets used in this study are provided in Table 1.

Table 1. Datasets Used in the Study and Their Sources

Data	Source	Year
MOD13Q1.006	NASA LP DAAC at the USGS EROS Center	2001-2019
Commodity Crop Mapping and Monitoring in South America	GLAD Landsat Analysis Ready Data and Tools	2000-2019
Soybean planting, harvesting production and yield data	Argentina Ministry of Agriculture, Livestock and Fisheries	2000-2019
Global Administrative Unit Layers 2015, Second-Level Administrative Units	FAO GUAL, UN	2015

The data processing workflow is comprised of two stages. The first stage involves filtering and aggregating spatial data from three data sources: MODIS terra vegetation indices product, soybean land-use product, and administrative boundary product. Specifically, MODIS images captured from January 1st to April 30th, which is the critical growing period of soybean, were selected. The soybean land cover products were used to mask the extracted images. Given the disparity in spatial resolution between the soybean land-use product (30 meters) and the MODIS product (250 meters), the MODIS pixels were resampled to 30 meters to facilitate the masking process. The masked images were then aggregated at the departmental level, with statistical summaries including the mean, max, min, mode, variance, quantiles and standard deviation of soybean land's NDVI and surface reflectance bands being derived from GEE. The second stage involves joining the soybean yield dataset for the period between 2000 and 2019 to Argentina departmental boundaries. However, since the MOD13Q1.006 only dates back to 2001, the yield data for the year 2000 was excluded from the analysis.

We devised several image combinations to determine the optimal model inputs, given the availability of eight MOD13Q1.006 images during each growing season. It is imperative to explore the images that contribute most to accurate crop yield estimation. Thus, we established four image combinations: In the first combination, we utilized all eight images from each growing season as input, which we will refer to as "NDVI-Eight" for clarity and brevity in the rest of this paper. In the second combination, we selected the image with the highest vegetation index among the eight images as the input, which we will refer to as "NDVI-Max". In the third combination, we chose three images which include the image at the peak, the next one before the peak, and the next one after the peak. We will refer to this combination as “NDVI-During Peak”. In the fourth combination, we selected three images that represented the growing season at and after the peak. These images included the peak NDVI and two subsequent images, which we will refer to as “NDVI-After Peak”.

We used the maximum vegetation index value to determine the peak of the soybean growing season and constructed the peak curve using data from the previous and subsequent dates. However, some regions and specific years presented issues where the peak vegetation index and the first or last two images during the growing season exceeded the allowable range. In these cases, we made necessary adjustments by shifting the three images earlier or later to resolve the issue.

To answer the question whether a simple NDVI product during the soybean's growing season can predict yield with satisfactory accuracy, we explored various independent variables settings that involve the previous year's yields. One setting solely uses the NDVI image combinations, while the other includes not only the NDVI combinations but also the previous year's yields. We also created different combinations of input explanatory variables based on the number of statistical summarization variables being input into the prediction models. In the first setting, only statistical summarization from NDVI was selected for the model, including the maximum, mean, median, and minimum. In the second setting, in addition to the previous four variables from NDVI, we added four spectral bands' (surface reflectance 1, 2,3,4) statistical summarization from MOD13Q1.006 as well, namely the maximum, mean, median, and minimum. We named the first input variable set as "Var 1" while the second input variable set as "Var 2". We combined Var 1 and Var 2 with the previous year’s yields or not as explanatory variables, resulting in a total of four groups of explanatory variables. Multiply them by four NDVI combinations, which will give us a total of 16 explanatory variable sets for prediction models. This approach provides a thorough analysis of the input variables and enables the identification of the optimal combination for predicting soybean yields accurately.

Figure 2 illustrates the soybean yields at the department level in 2019 and the most frequent NDVI peak time during 2001-2019 for the departments in Argentina. The findings demonstrate that the primary soybean production region in Argentina is concentrated in the provinces of

Buenos Aires and La Pampa, where the peak NDVI dates occur around Early to Mid-February. Notably, departments in Buenos Aires provinces are relatively smaller than those in La Pampa provinces, resulting in smaller total productions; however, these departments exhibit the highest yields. In contrast, the departments located in the Northern and Middle parts of Argentina experience later NDVI peaks and lower soybean yields.

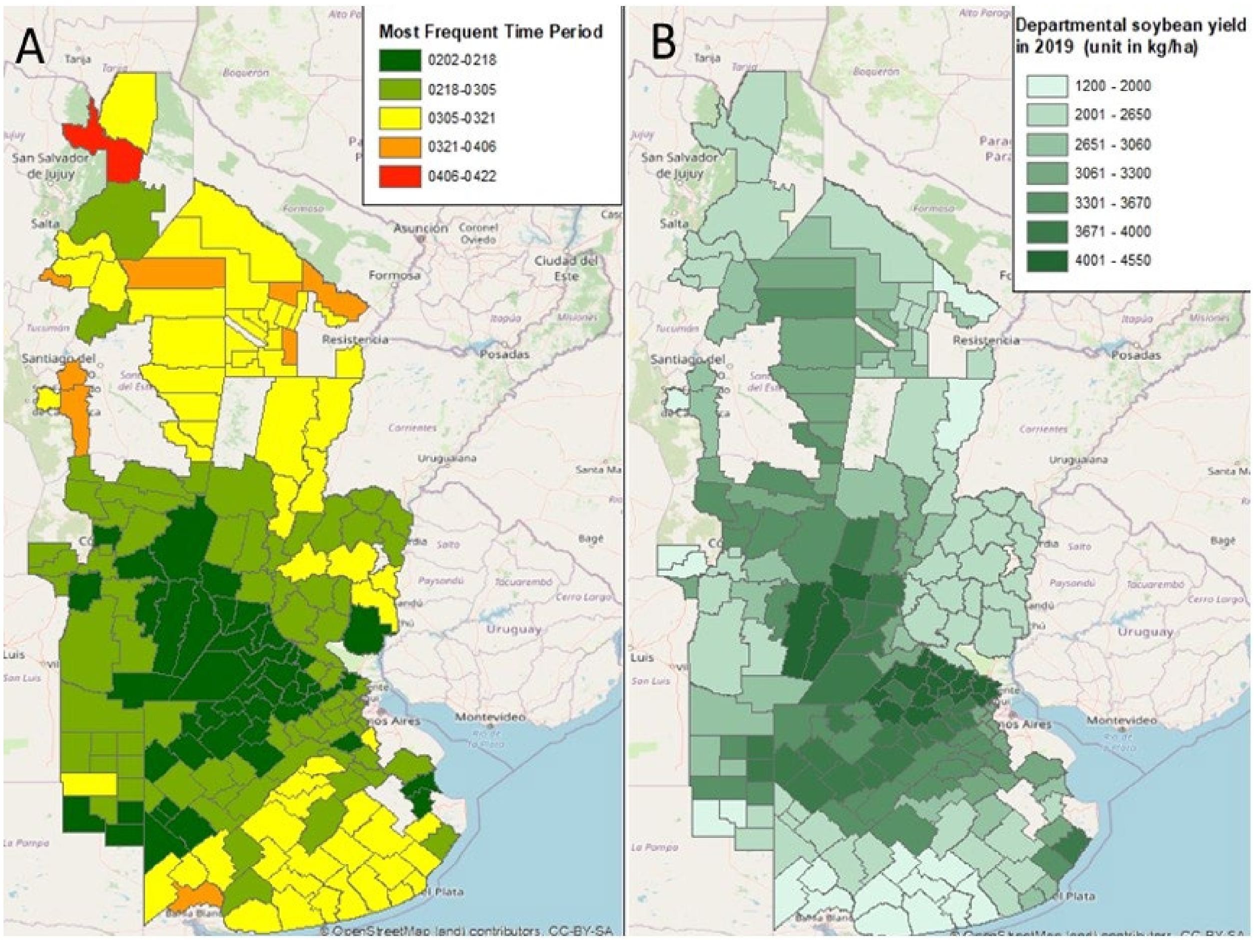


Figure 2. (**A**) Most frequent NDVI peak time period during 2001-2019; (**B**) Soybean yields at the department level (kg/ha) in 2019

2.3. Modeling Methodology

The primary objective of this study is to accurately estimate (fit) and predict soybean yields at the departmental level in Argentina. To achieve this goal, we have developed a methodology that encompasses the entire workflow of the study, starting from data retrieval and data processing to the use of processed data in different prediction models. The flow diagram below (Figure 3) depicts the methodology used in this study. First, we determine whether using the entire NDVI during the growing season or just a few key indices would achieve satisfactory accuracy in predicting soybean yield. If the latter is true, we attempt to identify which NDVI records should be utilized for accurate predictions. Then, we will evaluate the performance of the panel regression model and deep learning models for predicting yield production and identify the best model capable of predicting soybean yields with acceptable accuracy using only the first few images of the predicting season to make in-season predictions.

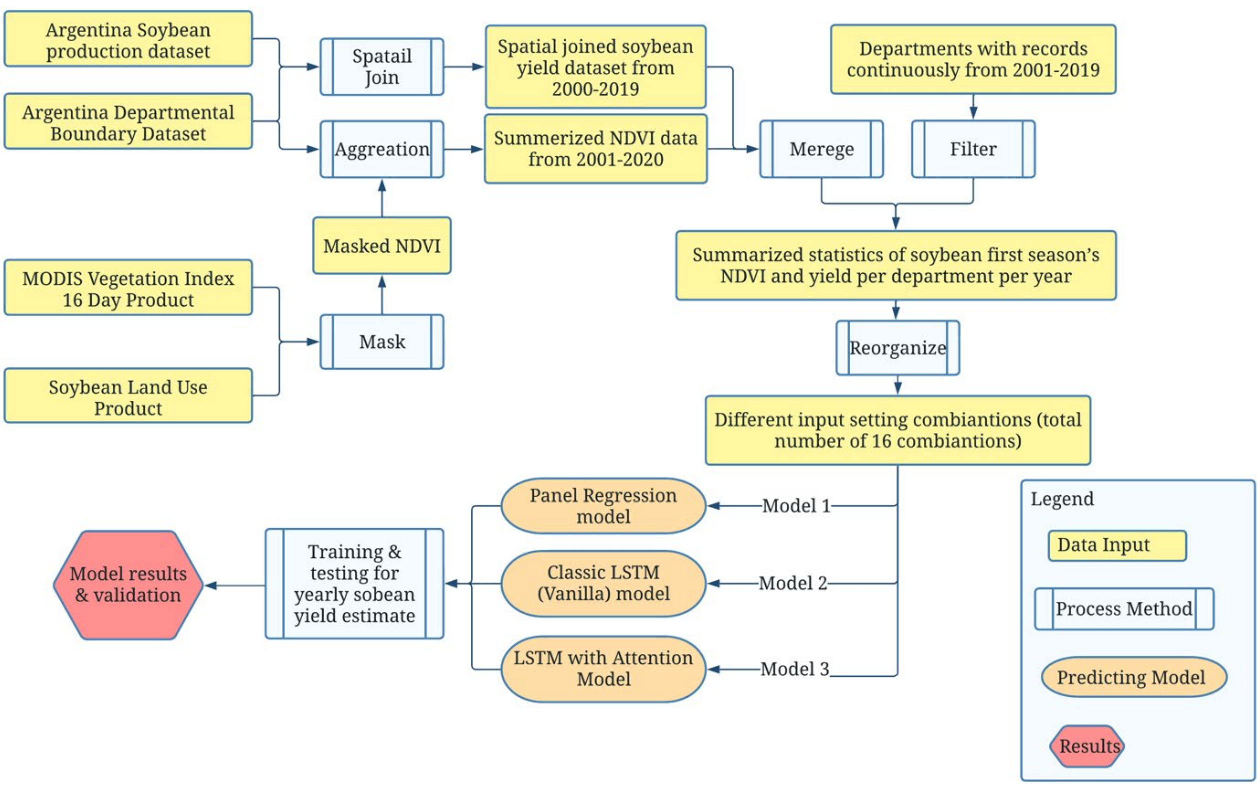


Figure 3. The flow diagram of the methodology. The legend describes what each type of shape/color represents.

Table 2 and Figure 4 demonstrate how our prediction model works using configurations 1-4 of training and testing data. Each local model has separate training and testing phases which use independent inputs. Figure 4 presents an example for predicting soybean yields for 2011 using configuration 4. The actual yield of 2011 is denoted as y_7. The entire training data has six years of yield data (denoted as y_1, y_2, y_3, y_4, y_5, y_6) paired with three years of vegetation index data (denoted as x_4, x_5, x_6), before the testing year. No testing labels have been used during the training stage. The training model is evaluated by the difference between fitted yield ($\hat{y}_4$, $\hat{y}_5$, $\hat{y}_6$) and actual yield

(y_4, y_5, y_6). For the testing year, the testing data include the testing year's vegetation index (x_7), along with yields in the previous three years (y_4, y_5, y_6) for testing evaluation between predicted yield $\hat{y}_7$ and actual yield y_7.

Table 2. Configuration of Training and Testing Data

Configuration	Training Data & Fitted Yield	Testing Data & Predicted Yield
Configuration 1	$x_{t-1} \rightarrow \hat{y}_{t-1}$	VI data $x_t \rightarrow \hat{y}_t$
Configuration 2	$x_{t-3} \rightarrow \hat{y}_{t-3}$; $x_{t-2} \rightarrow \hat{y}_{t-2}$; $x_{t-1} \rightarrow \hat{y}_{t-1}$;	$x_t \rightarrow \hat{y}_t$
Configuration 3	3-yr yield data (y_{t-4} to y_{t-2}) & 1-yr VI data (x_{t-1}) $\rightarrow \hat{y}_{t-1}$	3-yr yield data (y_{t-3} to y_{t-1}) & current year VI data (x_t) $\rightarrow \hat{y}_t$
Configuration 4	see figure 4	see figure 4

Note: 't' represents testing year, x vegetation index data, y actual yield, and $\hat{y}$ represents fitted or predicted yield.

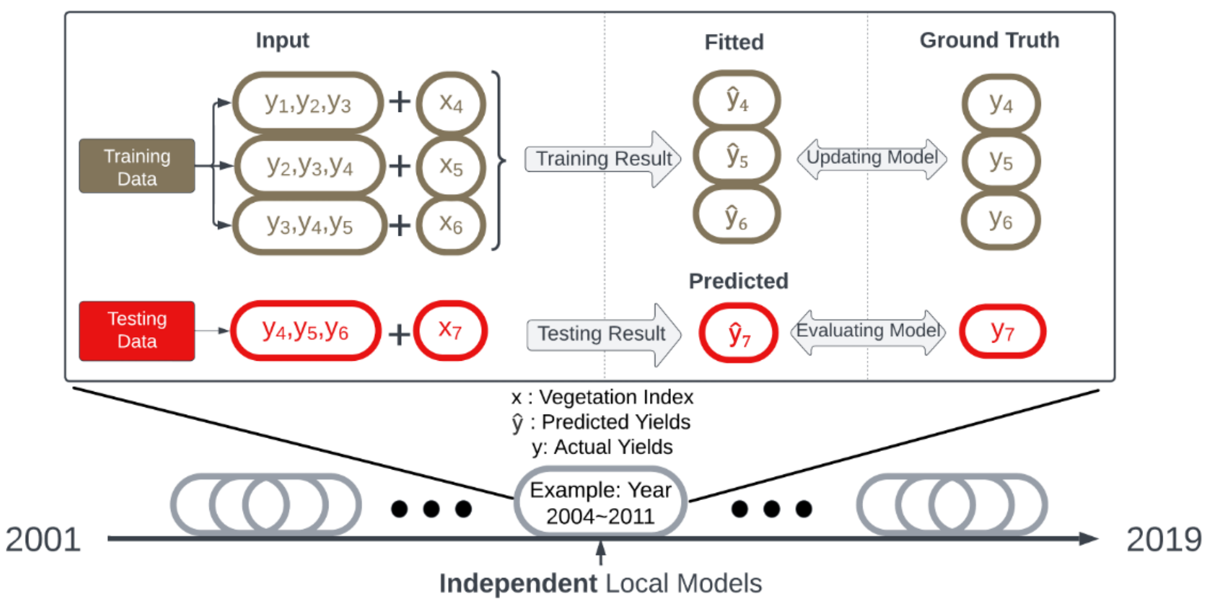


Figure 4. Example of Training, Testing and Prediction Using Configuration 4

This study uses three models as Figure 3 illustrates. The first is a panel regression model, which is arguably one of the most used models in econometrics. While the independent variables are the four combinations as the previous section explained. The dependent variable is the department's soybean yield. In each combination, the panel model is organized at the departmental level for each province, with a one-year time-step. Among alternative estimators of panel regression, we choose the fixed effects estimator since it is more suitable for prediction purposes. This fixed effect panel regression model is in the form of:

$$Y_{it} = \beta_1 x1_{it} + \beta_2 x2_{it} + \beta_3 x2_{it} + .. + \beta_k xk_{it} + a_i + \varepsilon_{it} \quad (1)$$

Where Y_{it} is soybean yield of unit i at time t; a_i is the fixed effect for unit i, which captures any time-invariant features of unit i that may affect the outcome variable; x_1, x_2, x_3, …, x_k are independent variables specified in each input combination for unit i at time t, β_1, β_2, β_3, …, β_k are the coefficients that represent the marginal effect of each independent variable on the dependent variable, ε_{it} is the error term, which captures any other factors that affect the dependent variable but are not included in the model.

The second model is a deep learning model, namely the LSTM (Long Short-Term Memory). Since this research is not trying to advance the technique itself, but rather to apply different models under the same combination of variables to determine the best suitable one for the best prediction performance in this research. A simple LSTM model and a LSTM with Attention model were chosen for this study. Both models were designed to have the same number of epochs and batch size to enable a fair comparison of their performance. Additionally, we employed an early stop mechanism to prevent overfitting and improve the generalization ability of the models.

LSTM is a type of recurrent neural network (RNN) architecture that was first introduced by Hochreiter and Schmidhuber (1997). The goal of LSTM architecture is to solve the vanishing gradient problem that arises when training traditional RNN models. Overall, the LSTM architecture is effective for modeling sequential data with long-term dependencies, such as natural language processing and time series forecasting. In this study, the classic LSTM was used as one of the deep learning models to predict soybean yields.

The third model is LSTM with Attention mechanism. The Attention mechanism in deep learning is a technique that enables the model to focus on specific parts of the input data when making predictions (Vaswani et al., 2017). The model does this by assigning different weights to different parts of the input data, which helps it prioritize significant data and downplay irrelevant data. In our study, the attention mechanism is applied to the hidden state outputs of LSTMs for more accurate predictions.

2.4. Model Comparison

RMSE (Root Mean Square Error) and its normalization (NRMSE) are commonly used metrics to evaluate the accuracy of prediction models. RMSE measures the square root of the average of the squared differences between the actual values and the predicted values. In other words, it represents the standard deviation of the residuals, or the differences between the predicted values and the actual values. It has the following form:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (2)$$

Its normalization by the range of the true results (actual yield), $\bar{o}$, is:

$$NRMSE = \frac{RMSE}{\bar{o}} \qquad (3)$$

The RMSE is a useful metric for measuring the accuracy of models in terms of their ability to predict values close to the actual values. A lower RMSE indicates a better fit between the predicted values and the actual values, and therefore a more accurate model. However, it should be noted that the RMSE may not always provide a complete picture of a model's performance, as it does not account for the direction of the errors (over- or under-predictions).

In this study, the RMSE is used to compare the performance of the panel regression, simple LSTM, and LSTM with Attention models in predicting soybean yields. By calculating the RMSE for each model, the study can identify which model has the smallest difference between predicted and actual values, indicating a more accurate prediction.

To effectively compare and determine the best prediction scenario in this study, a simple comparison of the RMSEs values is insufficient. As each prediction scenario comprises a series of RMSEs calculated over a period of 15 years (2004-2019), additional statistical tests are required to prove the effectiveness of different models and data combinations. Therefore, this study incorporates a Kruskal-Wallis H test, which is a non-parametric test, to identify any significant differences among different model and data combinations. This test is chosen over the regular t-test as the RMSEs values may not follow a normal distribution. The null hypothesis of the Kruskal-

Wallis H test is that there is no significant difference among the groups, while the alternative hypothesis is that there is a significant difference between at least one of the groups and the others.

Furthermore, to determine the preferred model combination for Argentina's soybean yield prediction, an average of RMSEs values will be calculated across all prediction model combinations. This approach will provide a more robust and comprehensive evaluation of different prediction scenarios and enable the identification of the optimal model and data combination.

Overall, the inclusion of statistical tests such as the Kruskal-Wallis H test, post-hoc analysis, and confidence intervals, as well as the use of averaged RMSE values, will provide a more thorough evaluation of different prediction scenarios and enable the identification of the best scenario for predicting Argentina's soybean yield.

3. Results

There are three distinct sub-sections in the results section. Using RMSE and H-tests, the first subsection presents and analyzes the outcomes of the predictive models aggregated at the national level. The second subsection uses two specific provinces as examples to give a more in-depth review of the performance of the models, analyzing the accuracy of the predictions at a finer spatial resolution. Lastly, the third subsection conducts a comparison analysis of the two algorithms, highlighting their respective advantages and disadvantages and providing insight into their overall performance.

3.1. Prediction results of yield at the national level

Through the utilization of panel regression and deep learning techniques, we have effectively generated predictive models for soybean yields in Argentina between 2004 and 2019. As outlined in the methodology section, our testing process encompassed 16 different explanatory variable sets, with three prediction models applied to each. This yielded a total of 64 predictions per year, covering every department in Argentina. It is important to note that, while predictions were made for each department, our training and testing procedures were conducted at the provincial level. To gauge the accuracy of our predictions, we calculated the root mean squared error (RMSE) for each prediction, which enabled us to rapidly rank the performance of different variable sets and prediction models. Table 2 presents the averaged soybean yield RMSE at the national level for all predictions conducted between 2004 and 2019.

Table 3. Mean Soybean yield RMSE (kg/ha) with NRMSE in Brackets by Different Variable Combinations over 2004-2019

Predicting combinations	Max	During Peak	Eight	After Peak
Attention with var2 & previous yields	795.48 (0.1142)	627.28 (0.0890)	505.78 (0.0726)	633.57 (0.0905)
LSTM with var2 & previous yields	878.28 (0.1247)	689.62 (0.1000)	507.32 (0.0727)	667.95 (0.0945)
Attention with var1 & previous yields	831.75 (0.1183)	721.11 (0.1492)	549.84 (0.0784)	754.35 (0.1094)
LSTM with var1 & previous yields	1422.44 (0.2012)	1049.38 (0.1181)	752.94 (0.1068)	1009.92 (0.1436)
Attention with only var2	814.88 (0.1153)	813.75 (0.1167)	824.50 (0.1196)	807.68 (0.1159)
Attention with only var1	888.02 (0.1242)	837.33 (0.2083)	828.28 (0.1173)	834.58 (0.2070)
LSTM with only var2	885.18 (0.1236)	835.80 (0.1195)	867.25 (0.1236)	833.45 (0.1186)
Panel regression with only var1	863.36 (0.1223)	911.88 (0.1306)	1030.79 (0.1465)	924.32 (0.1318)
Panel regression with var1 & previous yields	1016.19 (0.1504)	1057.81 (0.1539)	1078.54 (0.1560)	1032.25 (0.1503)
LSTM with only var1	1850.26 (0.2593)	1441.86 (0.2029)	1186.99 (0.1669)	1462.20 (0.2177)
Panel regression with var2 & previous yields	1051.95 (0.1539)	2049.53 (0.3124)	1373.80 (0.1924)	1511.10 (0.2217)
Panel regression with only var2	953.31 (0.1352)	1610.27 (0.2272)	1647.07 (0.2329)	1304.17 (0.1824)

The findings from Table 2 indicate that, at the national level, the LSTM with Attention model using an explanatory variable set of NDVI Eight on var 2 with previous year's yields is the best combination, with the lowest RMSE of 505.78 kg/ha during 2004-2019. On the other hand, the worst combination is the Panel Regression model using an explanatory variable set of NDVI During Peak on var 2 with previous year's yields, with the lowest RMSE of 2049.53 kg/ha. Overall, deep learning models perform better than Panel Regression models, with RMSE ranging from 505.78 kg/ha (NRMSE = 0. 0726) to 1850.26 kg/ha (NRMSE = 0.2593) for deep learning models and from 863.36 kg/ha (NRMSE = 0.1223) to 2049.53 kg/ha (NRMSE = 0.3124) for Panel Regression models.

It is important to note that the worst prediction made by the deep learning models is LSTM with NDVI-Max on var 1 with previous year's yields, whereas the best prediction RSME created by the Panel Regression model is made from NDVI-Max on var 1. In addition, there is a trend in deep learning models wherein the use of a larger input, be it the number of NDVI images or the number of explanatory variables, results in a lower RMSE. This trend is not observed in Panel Regression models, where the best prediction results were obtained with a smaller set of input variables, both for the number of NDVI images and explanatory variables.

Another interesting finding is that the deep learning models' prediction errors are more widely dispersed, as shown by the fact that LSTM with only var 1 or var1 with previous year's yield has worse performance than Panel Regression with only var 1 or var1 with previous year's yield. Moreover, Panel Regression shows a different trend than deep learning models, where a larger explanatory variable set would result in worse performance than a smaller input explanatory variable set. The figure below provides a clearer illustration of this trend. In order to make the representation of these models in the multiple figures below more concise, Table 3 provides the names of the combined models and (Table 4) shows the model combinations and their abbreviations used in Figures 4-10.

Table 4. Model Combinations and Their Abbreviations used in Figures 4-10.

Model Combinations	Abbreviation in figures
Attention with var2 & previous year's yields	DL1
LSTM with var2 & previous year's yields	DL5
Attention with var1 & previous year's yields	DL2
LSTM with var1 & previous year's yields	DL6
Attention with only var2	DL3
Attention with only var1	DL4
LSTM with only var2	DL7
Panel regression with only var1	LR1
Panel regression with var1 & previous year's yields	LR2
LSTM with only var1	DL8
Panel regression with var2 & previous year's yields	LR3
Panel regression with only var2	LR4

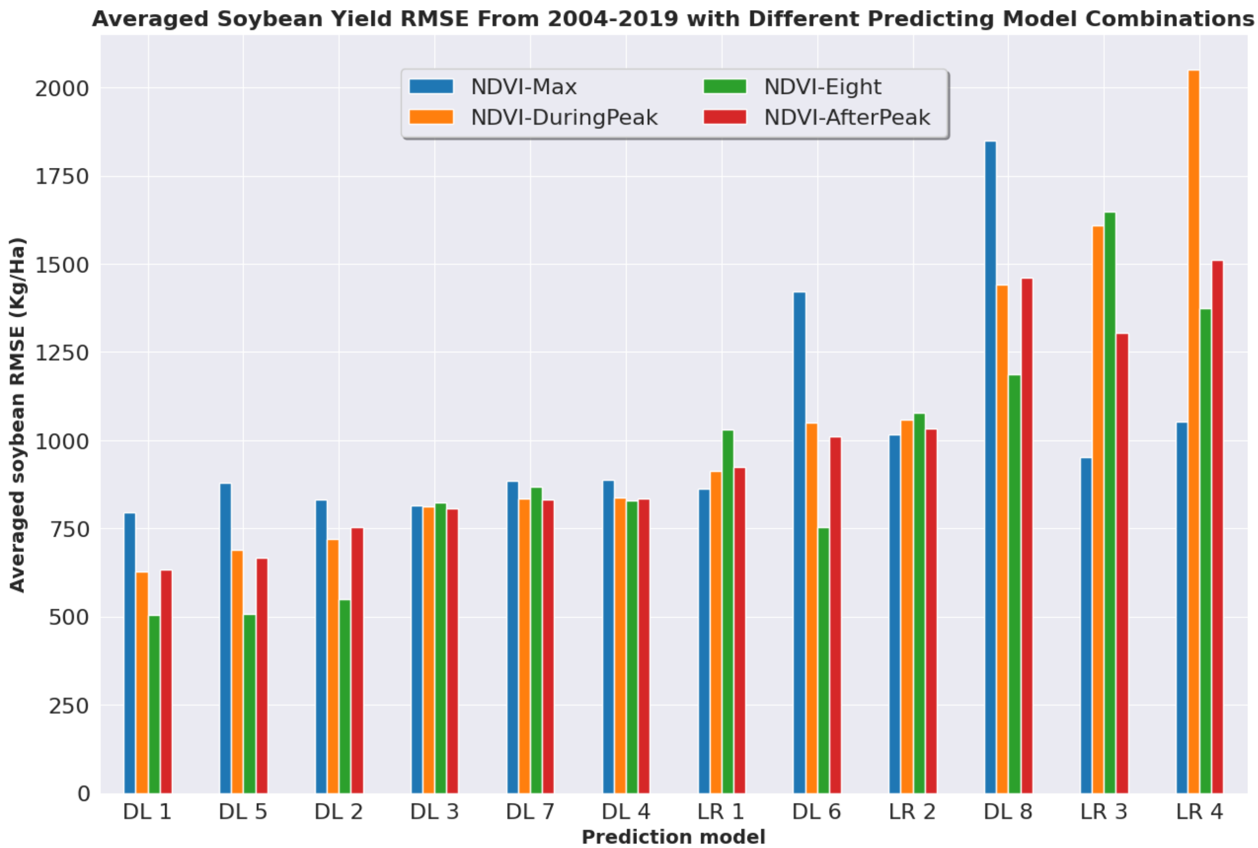


Figure 5. Averaged soybean yield RMSE over 2004-2019 by predicting model combination.

Figure 5 visualizes the averaged performance of these models as presented in Table 2. However, it is important to also examine the annual average of RMSE from the best four deep learning models' predictions and best four panel regression model's predictions during the same time period. Figure 5 demonstrates the yearly trend in prediction accuracy, with lower RMSE indicating better model performance. The figure clearly illustrates that all predictions follow a similar trend in each year's prediction accuracy, with the worst predictions occurring in 2008 and 2017. Panel Regression models also had relatively high prediction errors in 2011. Additionally, it is evident from Figure 5 that deep learning models' predictions were consistently more accurate than those of the Panel Regression models throughout the entire time period.

Interestingly, despite having similar error trends over the course of 16 years, the lowest prediction errors for deep learning models occurred in 2010 and 2019, while the lowest prediction error for the Panel Regression model occurred in 2016. The result suggests that even though each prediction model uses different input variable settings, the error trends share some similarities. It should be noted that Figure 6 only summarizes the best four deep learning models' RMSE and best four Panel Regression model's RMSE from 2004 to 2019. The worst performance of either model type is not plotted in the figure 6, as the RMSE for these worst performances were too high, rendering the predictions unreliable.

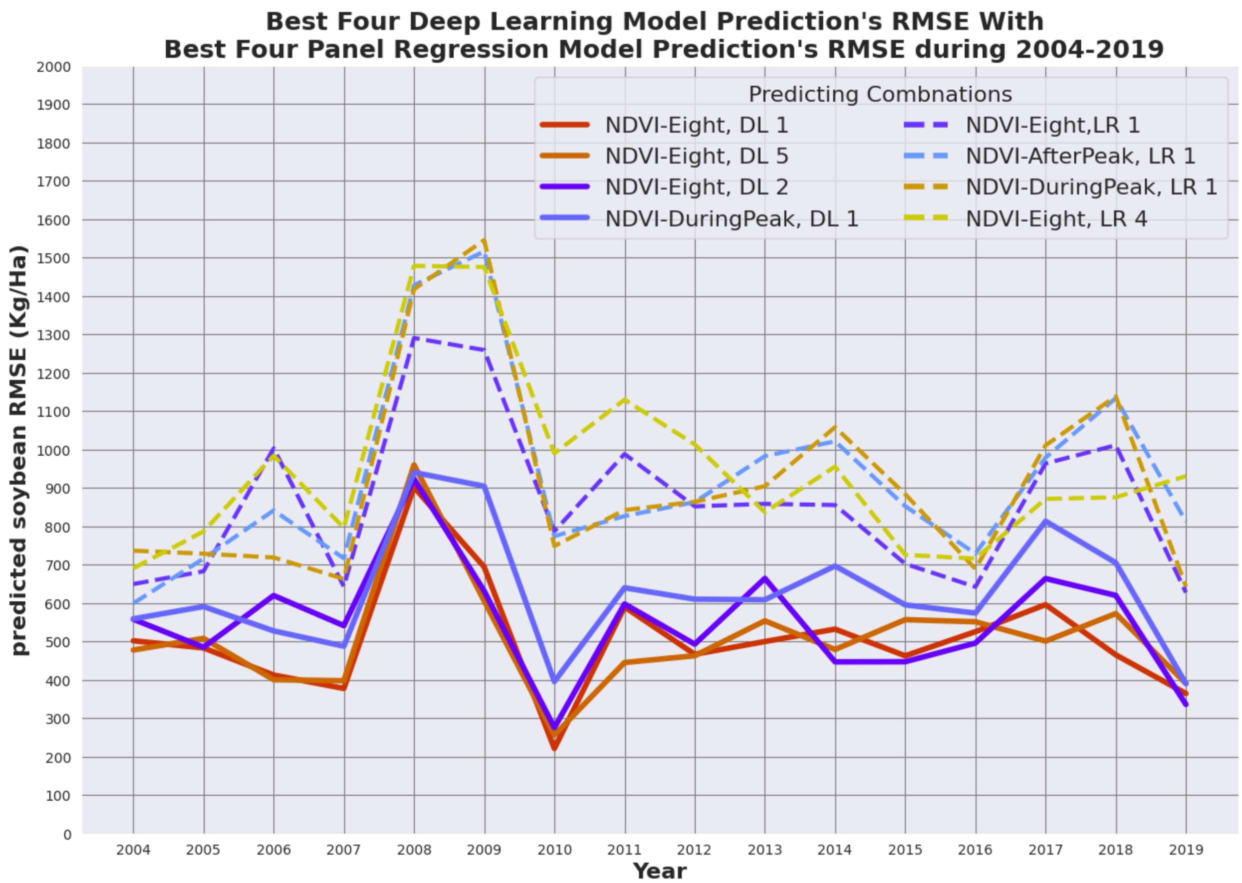


Figure 6. Comparison of the Best Four Deep Learning Model Predictions' RMSEs with the Best Four Panel Regression Model Predictions' RMSEs by year over 2004-2019

An additional key finding in this study is the importance of previous year's yield data as an explanatory variable for deep learning models in predicting soybean yield. Here, we found that in panel regression models, including previous year's yield data resulted in a worse RMSE than models without these data. As discussed earlier, deep learning models benefit from larger set of input training variables, with smaller set resulting in worse predictions. In contrast, panel regression models achieved relatively good accuracy using only key variables, but this accuracy was not as good as the deep learning models with their minimum number of input variables.

The worst deep learning prediction was made using a simple LSTM model with the input combination of NDVI-Max and only var 1, indicating that for each time step in the LSTM model, only one image with four variables was used as training data. This prediction was even worse than the panel regression model's predictions using the same input variable. Overall, the results suggest that the more training variables that are used for deep learning models, the better the model's prediction accuracy will be, while panel regression models may achieve relatively good accuracy among regression models with only key variables.

Figure 6 illustrates the national average yield changes from 2004 to 2019. The blue continuous line represents actual yields, while the dotted lines represent various prediction models. From top to bottom, different NDVI image combinations are utilized by each model. Among these, the NDVI-Eight combination demonstrates the best prediction accuracy across different prediction settings. This is because NDVI-Eight includes all images captured during the soybean growing season until harvest. However, since this study aims to establish an early in-season prediction for Argentina's soybean production, using the entire growing season's images would not meet the objective. Therefore, predictions made using NDVI-During Peak or NDVI-After Peak are preferred when employing deep learning models.

NDVI-During Peak demonstrates superior performance when compared to NDVI-After Peak. Moreover, because NDVI-During Peak is based on peak NDVI images in addition to one preceding and one subsequent image, it can generate yield predictions sooner than NDVI-After Peak, which uses the peak NDVI image and two subsequent images. In particular, NDVI-During Peak can predict yields approximately two weeks earlier than NDVI-After Peak.

Figure 6 also displays the general yield changes from 2004 to 2019, revealing three periods of soybean yield reduction in 2008, 2009-2011, and 2017. The most severe reduction occurred in 2008, followed by 2017 and 2009-2011. Comparing these reductions with the predictions, all models exhibited relatively high prediction errors for 2008 and 2017. While the prediction models successfully learned from the series of soybean reductions during 2009-2011, they failed to adjust for 2008 and 2017. This discrepancy could be attributed to the extreme weather conditions experienced in those years, when Argentina's soybean crops suffered from drought. The high prediction errors in 2008 and 2017 also led to high prediction errors in 2009 and 2018, with actual yields exceeding all predicted yields. This may be due to the three-year training data length for all models, causing the models to be misguided by the preceding drought reductions in 2009 and 2018.

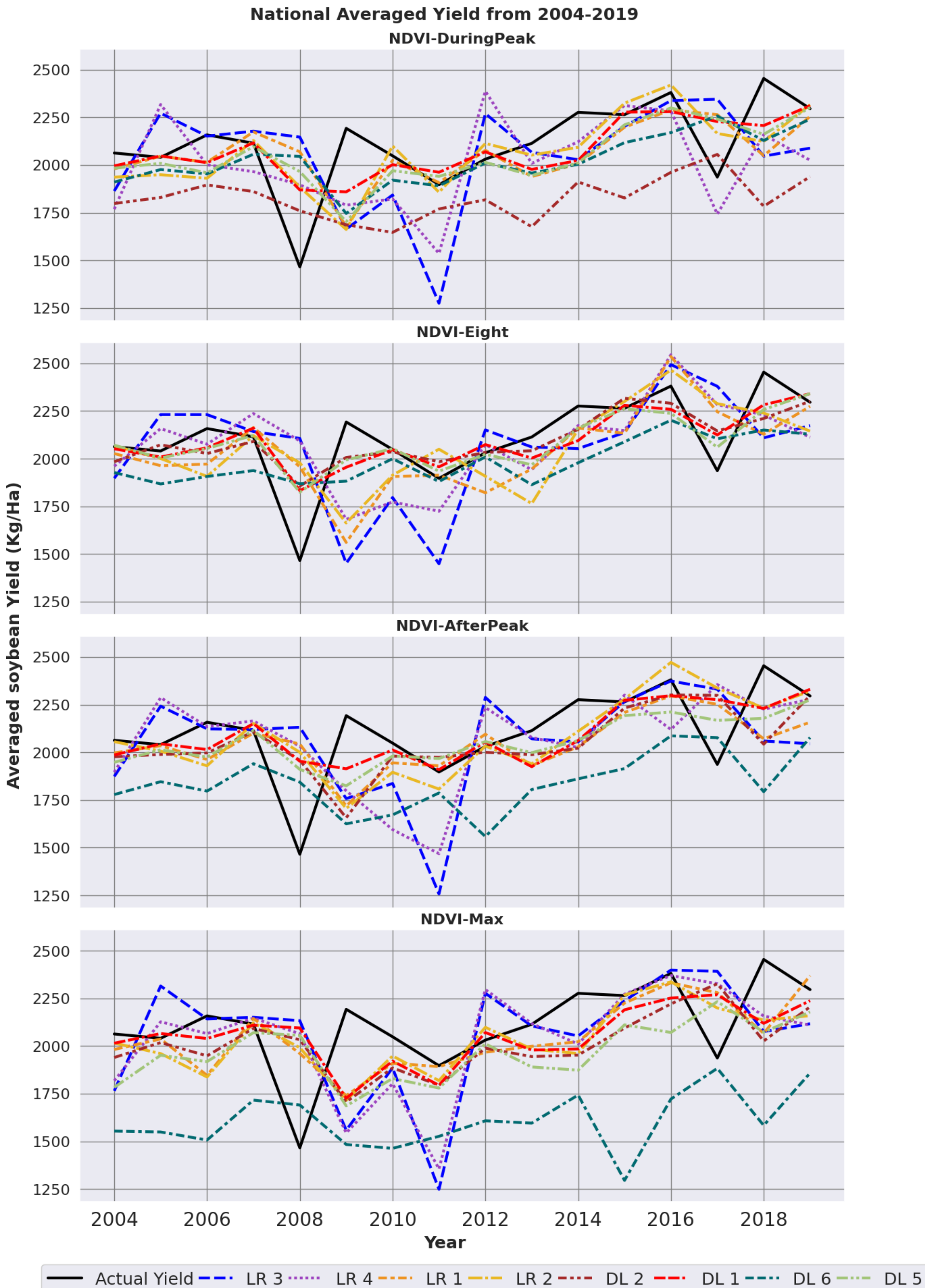
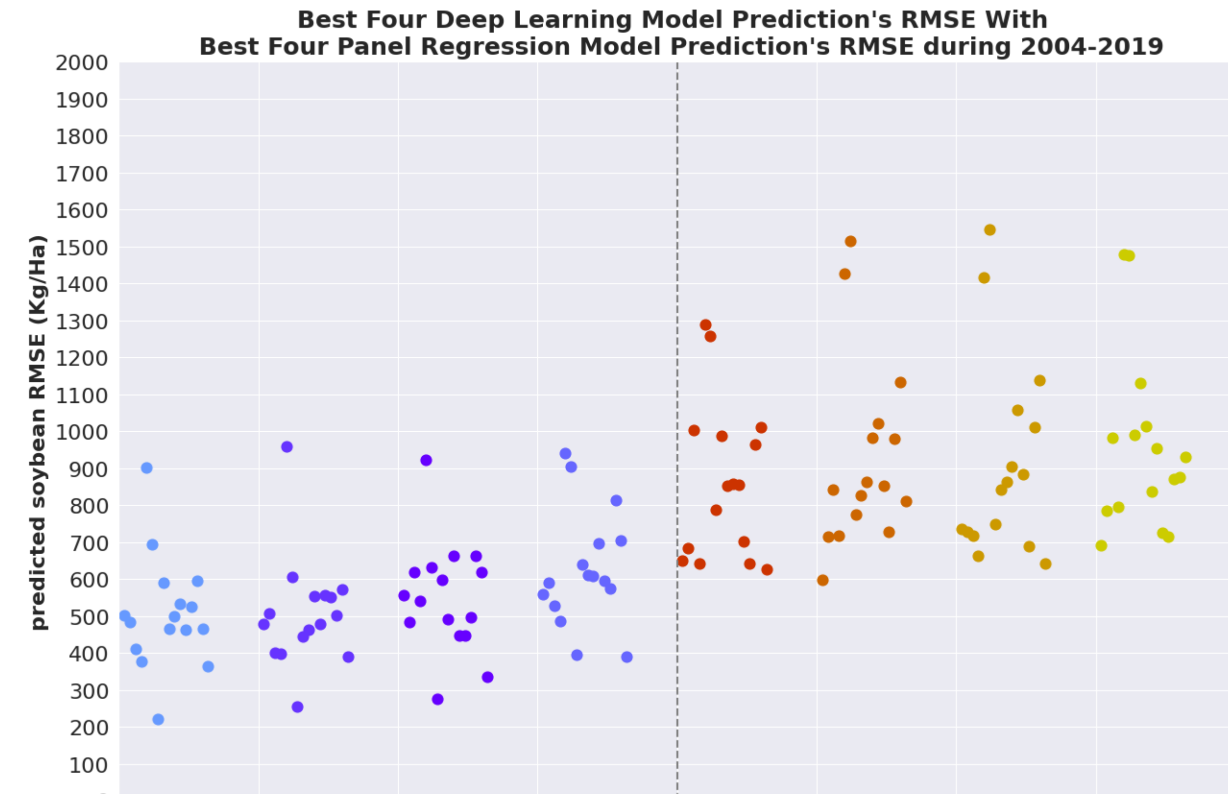


Figure 7. National Average Yield from 2004 to 2019

Large prediction errors were also observed in 2011, when Argentina's soybean production was affected by warm weather. However, the models overestimated the impact of the warm weather, leading to an underestimation of actual production. These underestimations were predominantly reflected in panel regression models, whereas deep learning models were able to generate more accurate predictions. By incorporating historical yield data, deep learning models were capable of minimizing the effects of sudden NDVI changes for in-season predictions, particularly in the context of extreme weather events. This demonstrates the potential advantages of deep learning models in capturing complex relationships and mitigating the impact of external factors on yield predictions.



	NDVI-Eight DL 1	NDVI-Eight DL 5	NDVI-DuringPeak DL 1	NDVI-AfterPeak DL 1	NDVI-Max LR 1	NDVI-AfterPeak LR 1	NDVI-DuringPeak LR 1	NDVI-Max LR 4
RMSE kg/ha 2004-2019	505.78	507.32	549.84	627.28	863.36	924.32	911.88	953.31
NRMSE 2004-2019	0.0726	0.0727	0.1183	0.0905	0.1223	0.1318	0.1306	0.1352

Prediction model combnations

Figure 8. Comparison of the Best Four Deep Learning Model Predictions' RMSEs with the Best Four Panel Regression Model Predictions' RMSEs during 2004-2019.

We further investigated the most suitable prediction combination for accurately predicting soybean yields in Argentina. Figure 8 displays the root mean square error (RMSE) of the four best deep learning model predictions and the four best panel regression model predictions from 2004 to 2019. The best prediction models are ranked from left to right based on their mean RMSE. From the plot, the LSTM with Attention model using NDVI-Eight on var2 with previous year's yields exhibits the lowest mean RMSE. However, as previously discussed, NDVI-Eight may not provide a timely pre-season prediction before soybean harvest. Therefore, the third-best prediction combination, an LSTM with Attention model using NDVI-During Peak on var2 with previous year's yields, is the most recommended soybean yield prediction model. Not only does this model have the smallest mean RMSE (548.84 kg/ha), but also the smallest RMSE variance (157.10 kg/ha).

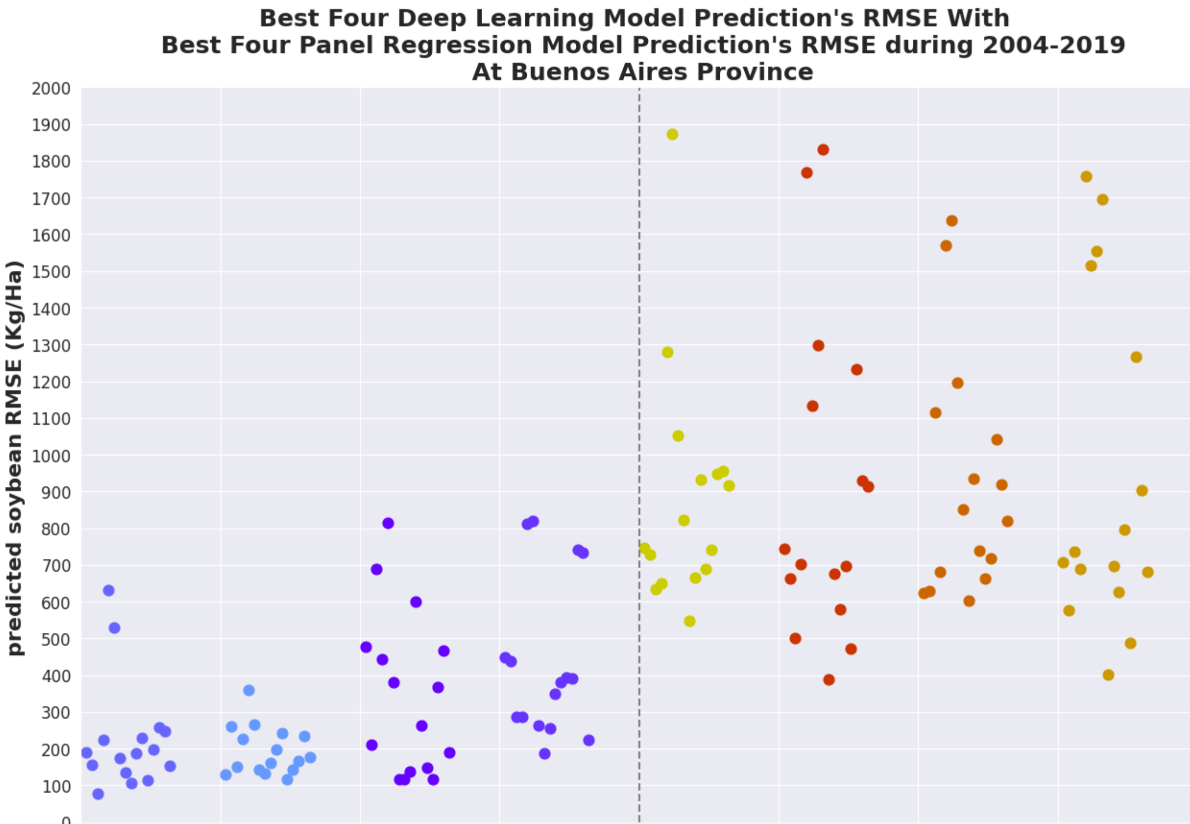
Comparing the four best deep learning models to the four best panel regression models reveals that the deep learning models not only have smaller RMSEs, but also smaller variances. To further determine whether the predictions are statistically significantly different from each other, a Kruskal-Wallis H test is conducted for these models. If the p-value of the H test is smaller than 0.05, there is a significant difference among the groups, otherwise, the null hypothesis cannot be rejected. From the H-test, the p-value is 3.504e-13 for all eight predictions; p-value is 2.05e-05 between the best deep learning model and the best panel regression model; p-value is 3.98e-05 between the LSTM with Attention model using NDVI-During Peak on var2 with previous year's yields and the best panel regression model. These results indicate a significant difference between deep learning and panel regression models in a broad comparison. With lower mean RMSE and lower RMSE variance, deep learning models outperform panel regression models.

The H test between the LSTM with Attention models using NDVI-Eight on var2 with previous year's yields and NDVI-During Peak on var2 with previous year's yields generated a p-value of 0.25, meaning that there is no substantial difference between these projections. Therefore, using NDVI-During Peak images can provide a satisfactory prediction for soybean yield in Argentina at the national level, with an average RMSE of 549.84 kg/ha from 2004 to 2019.

3.2. Prediction results of yield at provincial level

According to the methodology outlined, the model training and testing are conducted at the departmental level for each province and then aggregated to the national level for model comparison. However, it is important to note that not all provinces have the same number of departments with continuously planted soybeans from 2001-2019. For example, Buenos Aires has 88 departments, Chaco has 17 departments, Cordoba has 22 departments, Entre Rios has 15 departments, Santa Fe has 18 departments, and Santiago Del Estero has 12 departments. This uneven distribution of departments leads to varying levels of prediction accuracy across provinces.

In this section, we focus on the performance analysis of the provincial models in Buenos Aires and Cordoba. Figures 9 and 10 show the Root Mean Square Error (RMSE) of the four best deep learning models and the four best panel regression models from 2004 to 2019 in Buenos Aires and Cordoba, respectively. To test the significance of the RMSEs, the H test was performed for both provinces. In Buenos Aires, the H test result for the four best deep learning models and the four best panel regression models was 4.38e-15, while for Cordoba, the H test result was 1.39e-09.

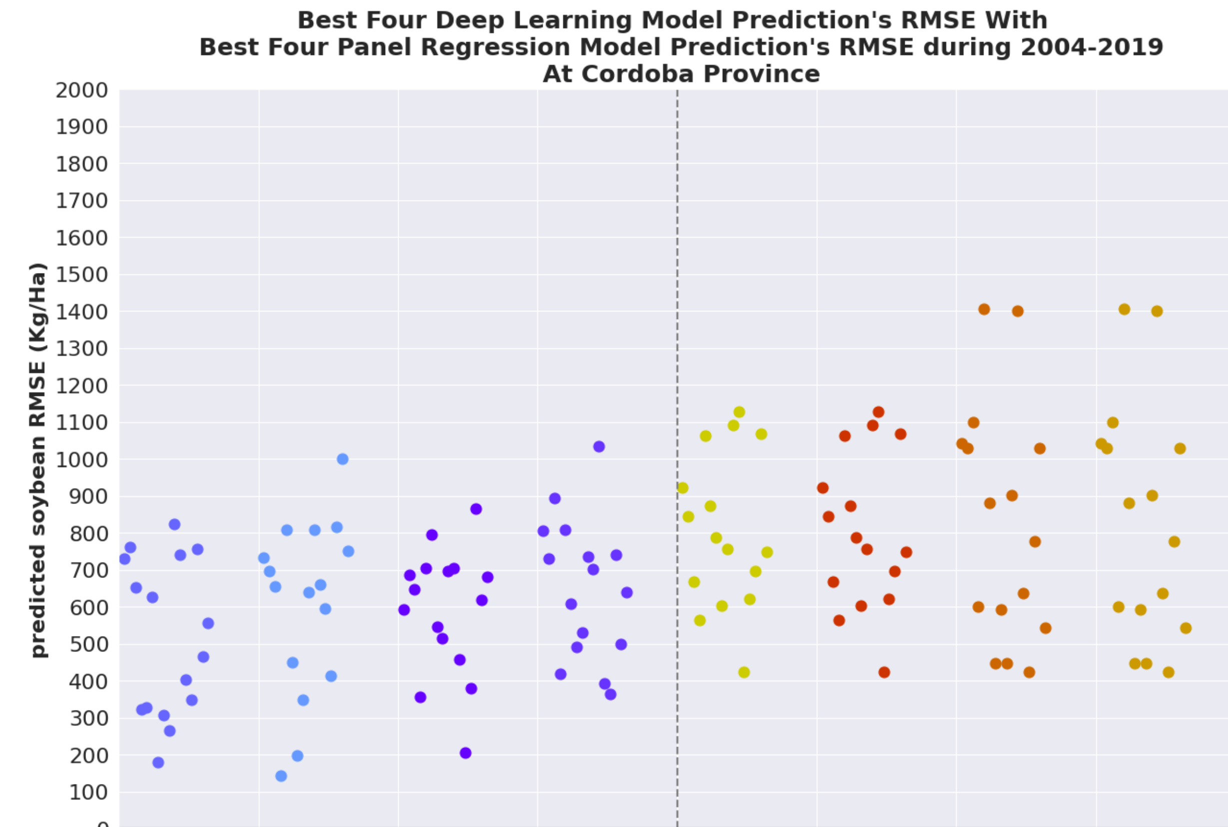


	NDVI-Eight DL 1	NDVI-Eight DL 1	NDVI-Eight DL2	NDVI-DuringPeak DL 1	NDVI-Eight LR 1	NDVI-DuringPeak LR 4	NDVI-AfterPeak LR 4	NDVI-AfterPeak LR4
RMSE kg/ha 2004-2019	225.75	194.22	346.19	438.31	886.55	908.29	921.17	943.58
NRMSE 2004-2019	0.0566	0.04322	0.0861	0.1027	0.1987	0.1827	0.2056	0.2201

Prediction model combnations

Figure 9. RMSE Comparison between the Best Four Deep Learning Models and the Best Four Panel Regression Models for Buenos Aires Province during 2004-2019

Additionally, the H test result between the NDVI-DuringPeak with Var2 and previous year's yields using an Attention model and the best panel regression model was 0.0002 in Buenos Aires and 0.045 in Cordoba. The best panel regression model in both provinces was NDVI-Eight with Var1 and previous year's yields, which requires all images to make a prediction, while the recommended deep learning model only requires the three images at, preceding and following the peak to achieve a satisfactory level of accuracy. Hence, while panel regression may provide accurate predictions close to those made during the peak NDVI, it requires the entire season's data to achieve the same level of accuracy as the Attention model, which is designed to make early in-season predictions.



	NDVI-Eight DL 1	NDVI-Eight DL 5	NDVI-DuringPeak DL 1	NDVI-AfterPeak DL 1	NDVI-Eight LR 3	NDVI-AfterPeak LR 4	NDVI-AfterPeak LR 4	NDVI-MaxThree LR 4
RMSE kg/ha 2004-2019	516.4	607.55	590.46	649.44	803.49	803.49	828.47	828.47
NRMSE 2004-2019	0.1264	0.1477	0.1396	0.1535	0.1885	0.1903	0.1884	0.2013

Prediction modelcombnations

Figure 10. RMSE Comparison between the Best Four Deep Learning Models and the Best Four Panel Regression Models for Cordoba Province during 2004-2019

4. Discussion

The comparative analysis of deep learning models and panel regression models for soybean yield prediction in Argentina reveals several key findings and implications. This discussion section will delve into the performance, advantages, and limitations of the models, as well as the potential for future research in this area.

4.1. Deep Learning Model Performance and Advantages Against Panel Regression

The results of this study demonstrate that deep learning models, particularly the LSTM with Attention module using the NDVI-DuringPeak with Var2 and previous year's yields, provide the most accurate predictions for in-season soybean yields in Argentina. The superior performance of deep learning models can be attributed to their ability to capture complex non-linear relationships between input variables and yield outcomes. This is particularly evident when multiple NDVI images are used during the prediction process, as the models can better learn and adapt to the temporal patterns in the data.

Furthermore, deep learning models exhibit a significant advantage in their ability to incorporate previous year's yield data for more accurate predictions. The Attention mechanism in the LSTM model is especially effective in capturing the contribution of previous year's yield by assigning different weights to NDVI images and attending to time steps in the data. This feature allows the model to better understand the historical context and trends in soybean yields, leading to improved prediction accuracy.

The study also highlights the importance of data availability and selection of input variables for different types of models. Deep learning models require more data to achieve good performance and benefit from a larger set of input variables. In contrast, panel regression models can perform well with only key variables, but may not necessarily benefit from a large number of input variables, which can lead to worse predictions in some cases.

4.2. Spatial Variability in Model Performance

The spatial analysis of model performance, as depicted in Figure 11, reveals interesting patterns and insights. The results demonstrate that the deep learning model outperforms the panel regression model in terms of prediction accuracy, particularly in the southeast region of the study area, which primarily consists of the Buenos Aires Province. This region exhibits a lower difference between the actual yields and predicted yields compared to the northeastern part of the study area, which includes the Santa Fe, Chaco, and Cordoba Provinces.

Within the Buenos Aires Province, only nine departments have high differences (> 60 kg/ha) between actual yield and predicted yields by the deep learning model, whereas the panel regression model has 22 departments with such high differences. This finding suggests that the deep learning model is more effective in capturing the spatial heterogeneity of soybean yields in this region, which may be attributed to its ability to learn from a larger set of input variables and its capacity to model complex, non-linear relationships.

In other provinces, the deep learning model also demonstrates better performance, with 95 departments having a difference of less than 60 kg/ha between actual and predicted yields, compared to only 36 departments in the case of the panel regression model. This further highlights the superiority of deep learning models in capturing the spatial variability of soybean yields across different regions in Argentina.

The spatial analysis also reveals that the panel regression model tends to have larger differences between actual and predicted yields, as well as a higher number of departments with differences exceeding 60 kg/ha, compared to the deep learning model. This finding underscores the limitations of traditional regression models in capturing the complex spatial patterns and relationships that influence soybean yields, and emphasizes the need for more advanced modeling techniques, such as deep learning, to improve prediction accuracy.

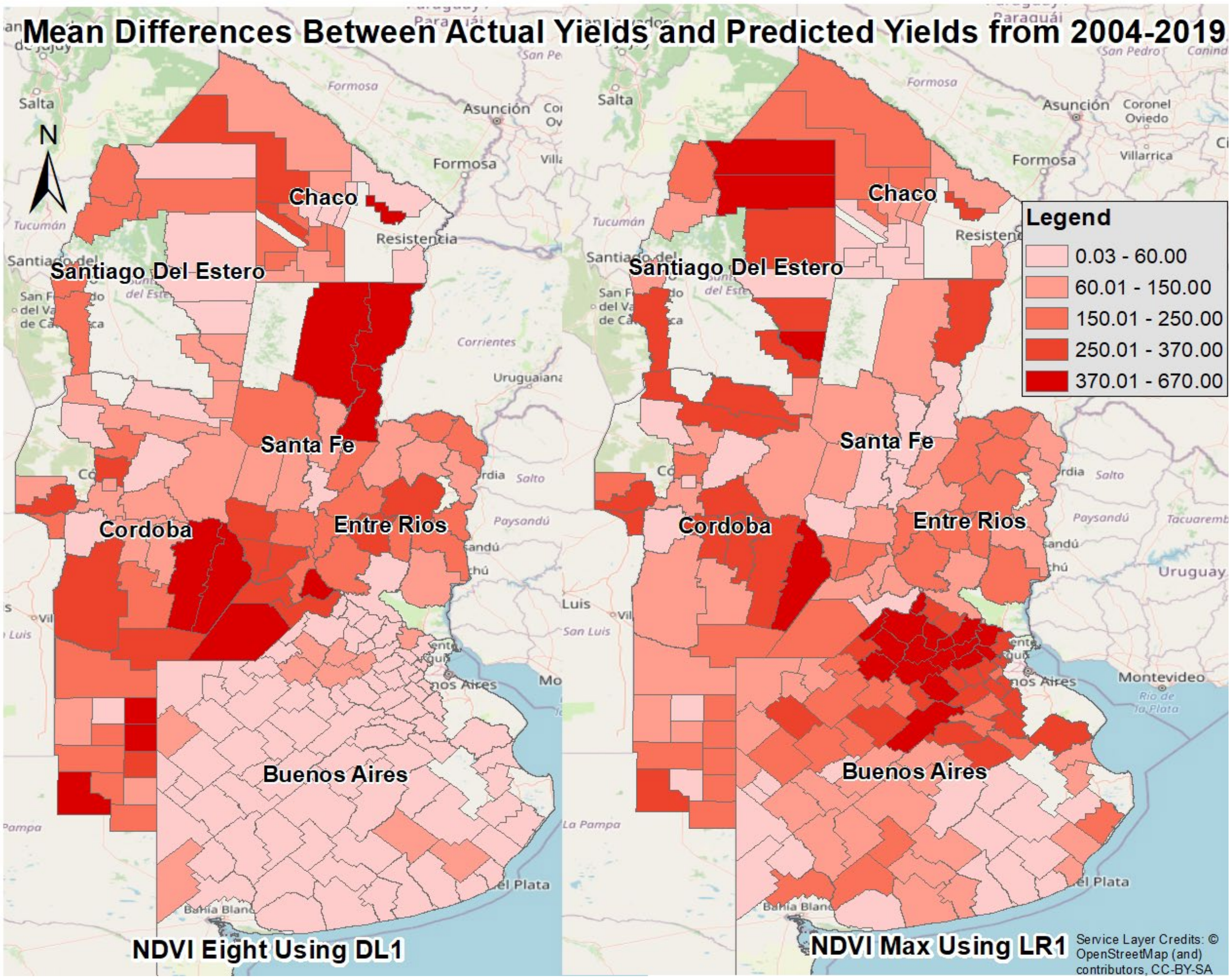


Figure 11. Mean Differences Between Actual Yields and Predicted Yields over 2004-2019

4.3. Limitations and Future Research

Despite the promising results, this study also reveals some limitations that should be addressed in future research. One major limitation is the short time span of training data, which may not fully capture the impact of agricultural technology innovation on soybean production. The three-year training period used in this study confirms that a simple soybean prediction using yields and NDVI does not require long-term data availability. However, it can also cause a delayed response effect on predictions in the event of sudden changes in yields or vegetation indices. Future research should explore the use of longer training periods and investigate methods to incorporate the effects of technological advancements in crop yield prediction models.

Another limitation of this study is the reliance on only NDVI and the previous year's yield for prediction. While these variables provide a good basis for yield prediction, incorporating additional data sources, such as precipitation, temperature, and soil moisture, could potentially improve the accuracy of the models. Future research should focus on integrating these additional variables and assessing their impact on prediction performance.

Moreover, the study highlights the challenge of predicting yields during years with extreme weather conditions, such as severe drought or flooding. Both deep learning and panel regression models exhibit relatively larger errors in these abnormal years. Developing methods to reduce prediction errors under these circumstances is crucial for providing more practical and reliable forecasting products to stakeholders in the agricultural sector. Future research should investigate techniques to better capture the effects of extreme weather events on crop yields and explore ways to improve model resilience in these situations.

The findings of this study have important implications for future research in crop yield prediction. The superior performance of deep learning models, particularly the LSTM with Attention model, underscores the potential for further exploration and refinement of these techniques. Researchers should continue to investigate novel architectures and algorithms that can better capture the complex dynamics of crop growth and yield formation. Additionally, the development of user-friendly interfaces and tools that enable stakeholders to easily access and utilize these advanced prediction models is essential for translating research findings into practical applications.

5. Conclusion

In this study, we employed two different approaches to predict soybean yields over the period of 2004-2019. We explored the explanatory power of in-season NDVI data for yield prediction in Argentina, compared the accuracy of using all growing season NDVI versus a few key NDVI, and examine the advantages and disadvantages of deep learning models compared to traditional regression models in yield prediction. The prediction results demonstrate that while using departmental NDVI data can relatively accurately predict soybean yield, the three images at, preceding and following the peak NDVI are sufficient for making a good in-season prediction as early as six weeks before the harvest. Although the LSTM model with the attention mechanism applied to the entire growing season NDVI values and three years of training data performed the best, using the entire season's NDVI for prediction may not be timely or efficient for the current season, as the results would be available after the actual harvest. Therefore, the optimal combination for accurate and more useful soybean yield prediction is the LSTM with three years of training data and the attention mechanism applied to three images at, preceding and following the maximum NDVI during the growing season.

Our comparison results demonstrated that the best-performing panel regression does not adhere to the same pattern as deep learning models, in which a larger number of training data leads to a lower RMSE. Contrariwise, larger training data sizes do not necessarily result in a lower RMSE. Possible explanations include a misspecification issue in a simple linear regression setting. Deep learning models have superior generalization abilities. The low RMSE produced by the deep learning models in this study indicates a robust capacity for generalization. Using the yield data of the previous year as a proxy for biophysical variables, the current NDVI serves as an explanatory variable to identify the potential departure of the current season from the NDVI/yield values of previous years. Thus, the incorporation of previous year’s yield information improves the accuracy of yield projections.

Here, we also highlight the limitations of this study. First, the use of only NDVI and the previous year's yield for prediction may not provide the most accurate results, and the inclusion of additional data such as precipitation and land surface temperature may improve the accuracy. Secondly, the models exhibit relatively larger errors in years with extreme weather conditions such as severe drought or flooding. As some research indicates, estimating crop yields during extreme weather conditions like drought or flooding could be very challenging (Feng et al., 2019; Prodhan, Zhang, Hasan, et al., 2022; Prodhan, Zhang, Pangali Sharma, et al., 2022). However, as accurate predictions during abnormal years would benefit local farmers and other stakeholders, additional research is required to determine how to reduce the prediction error under these circumstances in order to provide more practical forecasting products.

Funding

Xiaopeng Song is supported in part by the National Aeronautics and Space Administration’s (NASA) (Grant number: 80NSSC20K1490).).Yiqun Xie is supported in part by the National

Science Foundation (Grant number: 2105133, 2126474, and 2147195) and NASA (Grant number: 80NSSC22K1164).

CRediT authorship contribution statement

Yuhao Wang: Conceptualization, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Kuishuang Feng:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Laixiang Sun:** Conceptualization, Investigation, Supervision, Writing – review & editing. **Yiqun Xie:** Methodology, Software, Writing – review & editing. **Xiao-Peng Song:** Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Becker-Reshef, I., Vermote, E., Lindeman, M., & Justice, C. (2010). A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. *Remote Sensing of Environment*, *114*(6), 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010

Cai, R., Yu, D., & Oppenheimer, M. (2014). Estimating the Spatially Varying Responses of Corn Yields to Weather Variations using Geographically Weighted Panel Regression. *Journal of Agricultural and Resource Economics*, *39*(2), 230–252.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment*, *210*, 35–47. https://doi.org/10.1016/j.rse.2018.02.045

Didan, K., Barreto Munoz, A., Solano, R., & Huete, A. (2015). *MODIS vegetation index user's guide (MOD13 series) version 3.00*. http://vip.arizona.eduFAOSTAT. (2023). *Crops and livestock products*. License: CC BY-NC-SA 3.0 IGO. https://www.fao.org/faostat/en/#data/QCL

Feng, P., Wang, B., Liu, D. L., & Yu, Q. (2019). Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. *Agricultural Systems*, *173*, 303–316. https://doi.org/10.1016/j.agsy.2019.03.015

Franch, B., Vermote, E. F., Becker-Reshef, I., Claverie, M., Huang, J., Zhang, J., Justice, C., & Sobrino, J. A. (2015). Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. *Remote Sensing of Environment*, *161*, 131–148. https://doi.org/10.1016/j.rse.2015.02.014

Franch, B., Vermote, E. F., Skakun, S., Roger, J. C., Becker-Reshef, I., Murphy, E., & Justice, C. (2019). Remote sensing based yield monitoring: Application to winter wheat in United States and Ukraine. *International Journal of Applied Earth Observation and Geoinformation*, *76*, 112–127. https://doi.org/10.1016/j.jag.2018.11.012

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ji, Z., Pan, Y., Zhu, X., Wang, J., & Li, Q. (2021). Prediction of crop yield using phenological information extracted from remote sensing vegetation index. *Sensors (Switzerland)*, *21*(4), 1–17. https://doi.org/10.3390/s21041406

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01750

Klein, H. S., & Vidal Luna, F. (2021). The growth of the soybean frontier in South America: the case of Brazil and Argentina. *Revista de Historia Económica / Journal of Iberian and Latin American Economic History*, *39*(3), 427–468. https://doi.org/10.1017/S0212610920000269

Pastor, A. V., Palazzo, A., Havlik, P., Biemans, H., Wada, Y., Obersteiner, M., Kabat, P., & Ludwig, F. (2019). The global nexus of food–trade–water sustaining environmental flows by 2050. *Nature Sustainability*, *2*(6), 499–507. https://doi.org/10.1038/s41893-019-0287-1

Prodhan, F. A., Zhang, J., Hasan, S. S., Pangali Sharma, T. P., & Mohana, H. P. (2022). A review of machine learning methods for drought hazard monitoring and forecasting: Current research trends, challenges, and future research directions. In *Environmental Modelling and Software* (Vol. 149). Elsevier Ltd. https://doi.org/10.1016/j.envsoft.2022.105327

Prodhan, F. A., Zhang, J., Pangali Sharma, T. P., Nanzad, L., Zhang, D., Seka, A. M., Ahmed, N., Hasan, S. S., Hoque, M. Z., & Mohana, H. P. (2022). Projection of future drought and its impact on simulated crop yield over South Asia using ensemble machine learning approach. *Science of the Total Environment*, *807*. https://doi.org/10.1016/j.scitotenv.2021.151029

Salehnia, N., Salehnia, N., Saradari Torshizi, A., & Kolsoumi, S. (2020). Rainfed wheat (Triticum aestivum L.) yield prediction using economical, meteorological, and drought indicators through pooled panel data and statistical downscaling. *Ecological Indicators*, *111*. https://doi.org/10.1016/j.ecolind.2019.105991

Schnepf, R. D., Dohlman, E., & Bolling, C. (2001). *Agriculture in Brazil and Argentina: Developments and Prospects for Major Field Crops*. International Agriculture and Trade Outlook No. WRS-013, USDA, Washington DC. 85 pp. https://www.ers.usda.gov/publications/pub-details/?pubid=40353

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*. https://doi.org/10.1016/j.agrformet.2019.107886

Sly, M. J. H. (2017). The Argentine portion of the soybean commodity chain. *Palgrave Communications*, *3*(1). https://doi.org/10.1057/palcomms.2017.95

Song, X. P., Hansen, M. C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S. V., Di Bella, C. M., Conde, M. C., Copati, E. J., Fernandes, L. B., Hernandez-Serna, A., Jantz, S. M., Pickens, A. H., Turubanova, S., & Tyukavina, A. (2021). Massive soybean expansion in South America since 2000 and implications for conservation. *Nature Sustainability*, *4*(9), 784–792. https://doi.org/10.1038/s41893-021-00729-z

Sun, J., Di, L., Sun, Z., Shen, Y., & Lai, Z. (2019). County-level soybean yield prediction using deep CNN-LSTM model. *Sensors (Switzerland)*, *19*(20). https://doi.org/10.3390/s19204363

Tian, H., Wang, P., Tansey, K., Han, D., Zhang, J., Zhang, S., & Li, H. (2021). A deep learning framework under attention mechanism for wheat yield estimation using remotely sensed indices in the Guanzhong Plain, PR China. *International Journal of Applied Earth Observation and Geoinformation*, *102*. https://doi.org/10.1016/j.jag.2021.102375

Tian, Z., Zhong, H., Shi, R., Sun, L., Fischer, G., & Liang, Z. (2012). Estimating potential yield of wheat production in China based on cross-scale data-model fusion. *Frontiers of Earth Science*, *6*(4), 364–372. https://doi.org/10.1007/s11707-012-0332-0

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Weier, J., & Herring, D. (2000). *Measuring vegetation (NDVI and EVI)*. https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_1.php

World Bank. (2022). *World Development Indicators: Agriculture, forestry, and fishing, value added (% of GDP)*. https://databank.worldbank.org/reports.aspx?source=2&series=NV.AGR.TOTL.ZS&country=ARG

Xu, J., Zhu, Y., Zhong, R., Lin, Z., Xu, J., Jiang, H., Huang, J., Li, H., & Lin, T. (2020). DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sensing of Environment*, *247*. https://doi.org/10.1016/j.rse.2020.111946

Yu, Y., Feng, K., Hubacek, K., & Sun, L. (2016). Global Implications of China's Future Food Consumption. *Journal of Industrial Ecology*, *20*(3), 593–602. https://doi.org/10.1111/jiec.12392