Agricultural and Forest Meteorology 284 (2020) 107886

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil

Raí A. Schwalbert[a,b,*], Telmo Amado[c], Geomar Corassa[d], Luan Pierre Pott[a,b], P.V.Vara Prasad[b,e], Ignacio A. Ciampitti[b,**]

[a] Agricultural Engineering Department, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul, Brazil
[b] Department of Agronomy, Kansas State University, 2004 Throckmorton Plant Science Center, 1712 Claflin Road, Manhattan, KS 66506, United States
[c] Soil Science Department, Federal University of Santa Maria, Santa Maria, Rio Grande do Sul, Brazil
[d] Cooperativa Central Gaucha Ltd. – CCGL, Cruz Alta, Rio Grande do Sul, Brazil
[e] Sustainable Intensification Innovation Lab, Kansas State University, 108 Waters Hall, 1603 Old Claflin Place, Manhattan, KS 66506, United States

ARTICLE INFO

Keywords:
Yield forecast
Satellite imagery
Deep learning
Long-Short Term Memory

ABSTRACT

Soybean yield predictions in Brazil are of great interest for market behavior, to drive governmental policies and to increase global food security. In Brazil soybean yield data generally demand various revisions through the following months after harvest suggesting that there is space for improving the accuracy and the time of yield predictions. This study presents a novel model to perform in-season ("near real-time") soybean yield forecasts in southern Brazil using Long-Short Term Memory (LSTM), Neural Networks, satellite imagery and weather data. The objectives of this study were to: (i) compare the performance of three different algorithms (multivariate OLS linear regression, random forest and LSTM neural networks) for forecasting soybean yield using NDVI, EVI, land surface temperature and precipitation as independent variables, and (ii) evaluate how early (during the soybean growing season) this method is able to forecast yield with reasonable accuracy. Satellite and weather data were masked using a non-crop-specific layer with field boundaries obtained from the Rural Environment Registry that is mandatory for all farmers in Brazil. Main outcomes from this study were: (i) soybean yield forecasts at municipality-scale with a mean absolute error (MAE) of 0.24 Mg ha^{-1} at DOY 64 (march 5) (ii) a superior performance of the LSTM neural networks relative to the other algorithms for all the forecast dates except DOY 16 where multivariate OLS linear regression provided the best performance, and (iii) model performance (e.g., MAE) for yield forecast decreased when predictions were performed earlier in the season, with MAE increasing from 0.24 Mg ha^{-1} to 0.42 Mg ha^{-1} (last values from OLS regression) when forecast timing changed from DOY 64 (March 5) to DOY 16 (January 6). This research portrays the benefits of integrating statistical techniques, remote sensing, weather to field survey data in order to perform more reliable in-season soybean yield forecasts.

1. Introduction

Soybean [*Glycine max* (L.) Merrill] represents one of the world's most important sources of protein and oil, with four countries, US, Brazil, Argentina, and China, accounting for approximately 90% of the total global production (Embrapa, 2018; USDA, 2019). Brazil is currently the second largest soybean producer, only behind the US, contributing to ~34.7% of the global production. As a consequence, the soybean production from Brazil has a large impact on the global market, with seasonal fluctuations on production impacting the financial market.

In Brazil, there are two institutions responsible for providing data about the status of the crops, the National Supply Company (Conab) and the Brazilian Institute of Geography and Statistics (IBGE). Both Conab and IBGE are primarily based on field surveys and they release annually yield forecasts (before harvest) on a state-level and estimations (after harvest) on a municipality-level (the last is released only by IBGE). Alternatively, with the advent of new cloud platforms such as Google Earth Engine (GEE) (Gorelick et al., 2017) providing an easier way to access large volumes of satellite and weather data, and dramatically increasing processing power through parallel computing resources, satellite imagery became an easy alternative for providing

* Corresponding author at: Department of Agronomy, Kansas State University, 2004 Throckmorton Plant Science Center, 1712 Claflin Road, Manhattan, KS 66506, United States.
** Corresponding author.
E-mail address: rais@ksu.edu (R.A. Schwalbert).

https://doi.org/10.1016/j.agrformet.2019.107886
Received 8 July 2019; Received in revised form 1 November 2019; Accepted 19 December 2019

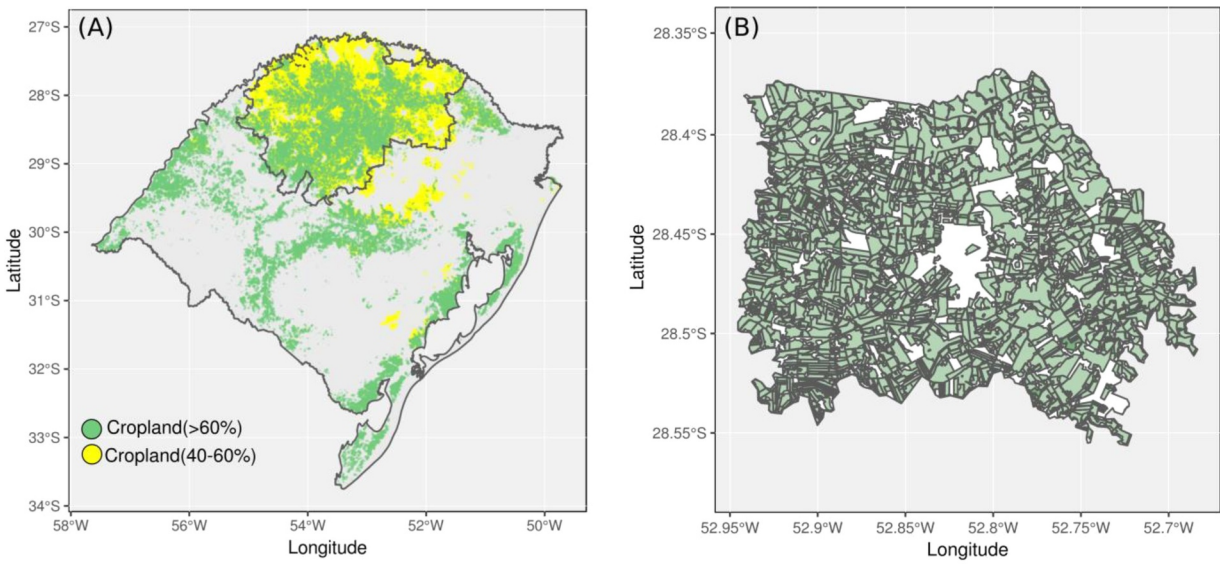

Fig. 1. (A) Annual International Geosphere-Biosphere Programme (IGBP) land cover classification generated by NASA LP DAAC (500 m - spatial resolution). Only two classes (12 and 14 from original raster file) are highlighted, with percentage of the pixel covered with cropland ranging from 40 to 100%. (B) Example of file available for downloading in CAR - Consolidated areas for the municipality of Não-Me-Toque, RS.

yield forecasts over larger domains in a near real-time basis. Research have repeatedly shown the potential of satellite imagery on providing quantitative data about yield worldwide (Ferencz et al., 2004; Hamada et al., 2015; Lobell, 2013; Peralta et al., 2016; Schwalbert et al., 2018), and improved model performance has been documented when weather data is effectively integrated on the estimations (Cai et al., 2019; Johnson, 2014; Lobell et al., 2015; Peng et al., 2018).

Along with the increase in computational processing power, more complex algorithms to data analysis also have become more popular when exploring larger and spatio-temporal datasets. Empirical relationships between soybean yield, canopy reflectance, and weather data usually present non-linearities (Johnson et al., 2016), and yield forecast models using a collection of those variables recorded over time are prone to over-fitting due to a high degree of autocorrelation. For those reasons, machine learning algorithms are able to more robustly deal with non-linearities against over-fitting. Those machine learning algorithms such as random forest and the neural networks have been successfully utilized to predict crop yield using remotely sensed vegetation indices (Alvarez, 2009; Cai et al., 2019; Johnson et al., 2016; Khaki and Wang, 2019; Li et al., 2013; Drummond et al., 2013; Shao et al., 2015). Random forest is an ensemble classifier that bootstraps training samples and variables to produce multiple decision trees performing predictions after aggregating the results from individual trees; this process is also known as bagging (Breiman, 2001). The neural networks consist of layers of highly interconnected processing units (neurons). The data moves throughout those layers across weighed connections, and each inner neuron is associated with an activation function, usually responsible for a non-linear transformation (Cai et al., 2019). A specific variation of the neural network, known as Long-Short Term Memory (LSTM) has been more recently noticed because of its large capacity to deal with sequential data (Cunha et al., 2018; You et al., 2017).

In addition to data processing, another challenge when performing yield forecast over large domains is to access to the crop geolocations. For some regions of the world such as the US, this information is easily available since it is yearly released by the National Agricultural Statistic Service (NASS) named Cropland Data Layer.. A 30-m resolution crop specific gridded layer (Johnson and Mueller, 2010) that is largely employed as a relevant layer in studies aiming at forecasting crop yield in the US (Johnson, 2014; Shao et al., 2015). In Brazil, such information is not yet available, despite the efforts of the governmental agencies. However, for most of the municipalities (similar to the county-level in US) in Brazil, it is possible to access the field boundaries of permanent agricultural fields from the Rural Environmental Registry (Cadastro Ambiental Rural - CAR) (http://www.car.gov.br). This layer despite not holding information related to crop types, provide an useful data source for removing most part of the noise from the satellite imagery, coming from areas that are not meaningful for agricultural purposes.

Thus, considering the importance of soybean in Brazil and its impact on the global economy, and the evident lack of reliable yield information in near real-time basis, the implementation of a near-real time yield forecast will provide a useful layer for agricultural purposes and policy applications. Therefore, the objectives of this research were to: i) compare the performance of three different algorithms (multivariate ordinary least square – OLS - linear regression, random forest and LSTM neural network) for forecasting soybean yield using vegetation indices such as NDVI, EVI, and weather data such as land surface temperature and precipitation as independent variables, and ii) evaluate how early (during the soybean growing season) this method is able to forecast yield with reasonable accuracy.

2. Material and methods

2.1. Region selection

The study was conducted in the northern region of the Rio Grande do Sul (RS) state, Brazil. This region was chosen due to: i) the high area and frequency of soybean crop in the soybean-corn summer crop rotation (85% of the cropland is allocated to soybean), and ii) since its represents the largest contiguous cropland area in RS state (Fig. 1A).

2.2. Data sources

Historical municipality-level soybean yield data (2003–2016) was obtained from IBGE (https://sidra.ibge.gov.br/pesquisa/pam/tabelas). This database is released as a point information in a municipality (each point is a municipality/year yield record) without geographical identification such as latitude and longitude. We used 80 municipalities once we focused only in the ones with yield data available for the entire period considered in the study.

Additionally, vegetation indices (VIs) from satellite imagery were

obtained from MODIS Surface Reflectance products. Since we are working on a large region, and we need to build a mosaic free of clouds for the entire region, the available options for satellite data are limited. The NASA Earth Observing System Data and Information System (EOSDIS) provided 8- and 16-days mosaics on a near real-time basis allowing to retrieve satellite data with minimal interference of clouds with 250-m resolution. This cloud-freeness is the main reason to choose the EOSDIS data for building our model. From those mosaics, we retrieved two VIs, the normalized difference vegetation index (NDVI) (collection MODIS/006/MOD09Q1), and enhanced vegetation index (EVI) (collection MODIS/006/MOD13Q1). Since EVI is released in a lower image frequency (every 16 days) compare to the NDVI (every 8 days) we calculated the average between each two consecutive EVI images in order to provide an EVI time series that matches with the NDVI images.All the images from these two collections were gathered between October 15 and March 5 (soybean planting and harvesting are not in the same calendar year in Brazil) from 2002 to 2016.

Two additional variables were selected to be evaluated on the models: daytime land surface temperature (LST), and precipitation. The LST is a similar, but not exactly the same, measurement as more commonly collected air temperature. The two variables (LST and air temperature) are strongly related, though, with LST having larger temperature extremes and being locally dependent on the land cover type (Mildrexler et al., 2011; Wan, 2008). The LST was produced from the 8-day composited thermal product from Aqua satellite's MODIS sensor (termed MYD11A2). Daily precipitation data was provided by the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset with resolution of ~5.5 km (Appendix A).

More details about the criteria used for selecting input variables for composing the yield forecast model are provided in Appendix B.

2.3. Data collection and organization

Since Brazil does not have a crop-specific data layer for retrieving geographical information about soybean field locations, we decided to use the data from CAR (Appendix C). This information was downloaded as individual shapefiles (one for each municipality considered in this study), and then merged via R (R Core Team, 2017) in a unique file to be uploaded on the GEE platform.

All the remote sensing and the weather data were gathered via GEE using the CAR layer as a cropland mask. All the collected information was organized in a table format and averaged to municipality level before being merged with the yield data layer, comprising the first and the second steps on the model development (Fig. 2).

2.4. Empirical relationships between yield, remote sensing and weather data

Three algorithms were tested to describe the relationship between yield, VIs, LST, and precipitation: i) multivariate OLS linear regression, ii) random forest, and iii) LSTM neural network. Multivariate OLS model was chosen as a benchmark relative to the two machine learning algorithms, since it represents the one of the simplest form to build empirical relationships between dependent and independent variables. Secondly, we chose the random forest model to explore non-linear models. Random forests are easy to train, have low sensibility to outliers, high computational efficiency and robustness against over-fitting (Belgiu and Drăgut, 2016). Lastly, we tested the model performance using the LSTM neural network. The LSTM neural network are prepared for receiving sequential data as an input and are able to extract important aspects related to the time series since it maintains a chain structure with time steps, similar to the way that crop growth modeling works. Each step takes information from previous step and outside input (from feature space – new NDVI, EVI, LST and precipitation values), and provides output for the next step. Furthermore, during the training process this algorithm is capable of retaining key information of input signals, and ignore less important parts.

For multivariate OLS and random forest, two classes of predictors were tested: i) the multi temporal EVI, NDVI, LST and precipitation, and ii) the seasonal integrated EVI, NDVI, LST and precipitation (as cumulative over the growing season). Therefore, for those two algorithms the annual municipality-level soybean yield forecasting model can be written as the following function:

$$y_{ij} = f(x_{ij}) + e_{ij} \quad (1)$$

where, y_{ij} is soybean yield for the i_{th} municipality and j_{th} year, x is the user-selected vector of predictors, f is a user-selected computer algorithm, and e_{ij} is error associated with the prediction.

The LSTM neural network received the two classes of inputs at the same time, classified as dynamic and static data. The dynamic data were related to the VIs, LST and precipitation time series, and were organized in a 3D array (samples, time steps, and features). The static data were the seasonal integrated variables. A concatenated layer was used to deal with those different input dimensions (Appendix D).

For all the algorithms, model performance was evaluated using a leave-one-year-out cross-validation approach and three metrics were used to assess the model accuracy: the mean absolute error (MAE), the mean squared error (MSE) and the root-mean squared error (RMSE) (Appendix E).

2.5. Time series sensitivity analysis

For all the models, a sensitivity analysis was performed to check how early in the crop growing season the forecasting yield model can be implemented and its impact on the overall model performance. For this purpose, data collected later during the growing season was subsequently removed from the model and the same validation approach aforementioned was used to compute the MAE, MSE, and RMSE. Thus, we tested the models using data until DOY 16 (January 16), DOY 32 (February 1), DOY 48 (February 17), and DOY 64 (March 5). We have assumed the existence of a delay in the release of the yield forecast models based on the process for uploading the MODIS product by NASA, in approximately five days (Sakamoto et al., 2014).

The model training highlighted in the step 3 of the model development framework (Fig. 2) was performed in the R environment using the RandomForest (Liaw and Wiener, 2002) and the Keras (Chollet, 2015) packages.

2.6. Relationship between model accuracy and yield/weather anomalies

Long-term yield data (1972–2017) for the entire region considered in this study (average over all the municipalities) was collected from IBGE. A regression analysis was performed using year as the independent variable and yield as the response variable. The residuals from this relationship (yield anomalies) were used in a Monte Carlo simulation in R program aiming at estimating the likelihood of any particular event to occur. We assume that the yield anomalies follow a normal distribution with mean and standard deviation estimated from the data. Residuals from the fitted model were utilized instead of using the absolute yield value to account for the genetic and technological evolution over the years.

We repeated this task using weather data instead of yield, and for doing that we extracted long-term (1982–2018) temperature and precipitation information from NASA POWER for all the municipalities considered in this study. We used NASA POWER for this analysis instead of MODIS and CHIRPS because MODIS only has information available after 2000. This information was summarized in 8-days periods (average for temperature and sum for precipitation). A Pearson correlation was performed among all the 8-days periods for precipitation and temperature, and yield in order to find a contiguous period of high correlation between these weather variables and yield. After defining this period, precipitation and temperature were summarized for

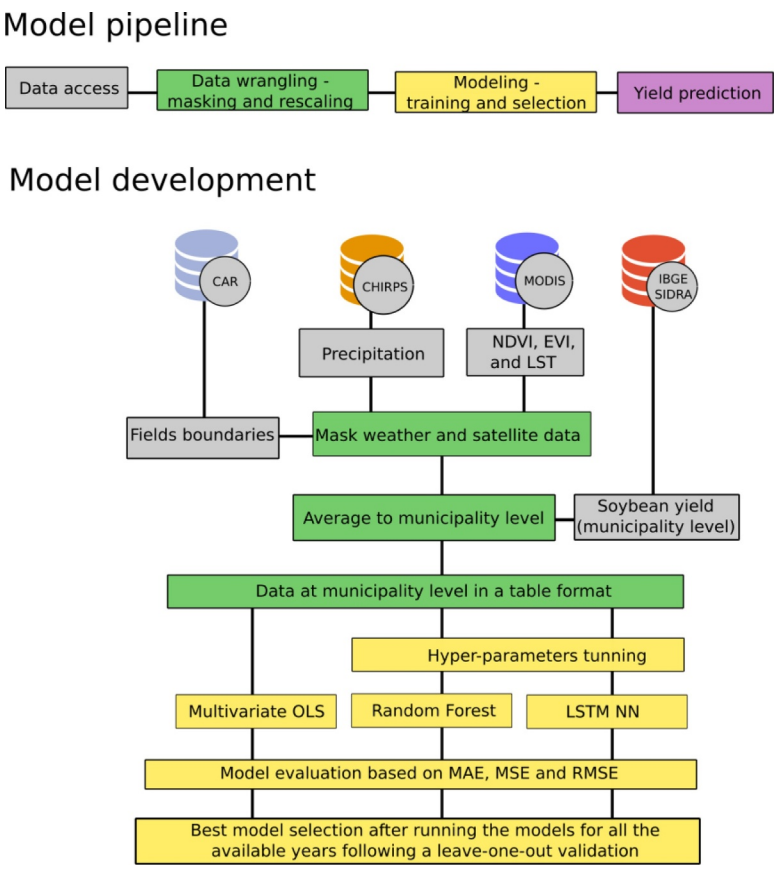

Fig. 2. Flowchart indicating all steps of the model development: 1- data access, 2- data wrangling which includes masking gridded data using CAR field boundaries and re-scaling the satellite and weather data to municipality-level before merging it with the yield data, and step 3- building the empirical relationships between soybean yield and the predictors (enhanced vegetation index - EVI, normalized difference vegetation index - NDVI, land surface temperature - LST, and precipitation) for the three considered algorithms (multivariate OLS, random forest, LSTM neural network), and selecting the best model based on metrics (MAE, MSE, and RMSE) derived from a leave-one-year-out cross validation.

the entire period and a Monte Carlo simulation was performed assuming that precipitation and temperature follow a multivariate normal distribution with μ_1, μ_2 and Σ, where: μ_1 is the precipitation mean, μ_2 is the temperature mean and Σ the variance-covariance matrix between precipitation and temperature. We decide to use a bivariate normal distribution instead a high dimensional distribution to avoid problems related to the curse of dimensionality, when the dimension is large and the sample size is moderate (Amato et al., 2013).

3. Results

3.1. Model performance at different forecast dates

Regardless of the date of the forecast, the seasonal integrated predictors outperformed the multi-temporal ones for multivariate OLS regression and random forest (data not shown). As the soybean yield forecasts were performed earlier in the growing season all the models tended to become less accurate, and as we move towards the end of the growing season all the three algorithms tended to become more assertive, represented by the decreases in all the three metrics, MAE, MSE and RMSE. Overall, the LSTM neural network presented the lowest values for MAE, MSE, and RMSE, followed by the random forest, with the OLS presenting a slightly inferior performance. The only exception was the DOY 16, where the LSTM had the least accurate performance among the three options, with the best performance for the multivariate OLS (Table 1).

The observed versus predicted soybean yield for the four dates tested in our model were explored using the best algorithm for each specific date. Based on the data presented on Table 1, we used the multivariate OLS regression model for DOY 16 and the LSTM for the remaining dates (Fig. 3A–D). The overall soybean yield data distribution for RS, Brazil from 2003 to 2016 presented a wide range of values from 0.2 to 4.2 Mg ha^{-1} with no evidence to reject the null hypothesis

Table 1
Model metrics comparison among multivariate OLS, random forest, and LSTM neural network.

Day of year	MAE (Mg ha^{-1})			RMSE (Mg ha^{-1})			MSE (kg ha^{-1})^2		
	OLS	RF	LSTM	OLS	RF	LSTM	OLS	RF	LSTM
DOY16	0.42	0.46	0.52	0.53	0.57	0.68	0.28	0.33	0.46
DOY32	0.46	0.44	0.42	0.58	0.57	0.56	0.34	0.33	0.31
DOY48	0.40	0.37	0.25	0.50	0.48	0.32	0.25	0.23	0.10
DOY64	0.32	0.32	0.24	0.40	0.39	0.32	0.16	0.15	0.10

*Values presented for OLS (multivariate OLS regression) and RF (random forest) are related to models using the seasonal integrated variables.

that the sampled yield values came from a normally distributed population (Shapiro-Wilk test p-value > 0.05). The maximum likelihood estimation for the mean and standard deviation based on the data were 2.4 and 0.8 Mg ha^{-1} respectively.

Despite residuals have been equally distributed along the 1:1 line considering all the years together for the predicted versus observed yield models, this pattern was not followed when the years were analyzed individually. Years such as 2004 and 2005 presented an error greater than the others, mainly for the early season forecasts (DOY 16 and 32) (Fig. 3). Moreover, after decomposing the MSE into its two components, the squared bias and σ^2, it can be seen that for the years presenting a greater MSE, the highest contributions came from the squared bias (lack of the capacity of the model to describe a specific phenomenon, systematic error) and not from σ^2 (non-systematic source of error) (Fig. 3).

We calculated the cumulative probability frequency for the soybean yield anomalies (residuals from the soybean yield-year relationship) for the region considered in this study (Fig. 4A–C). The analyses showed that years presenting the greatest anomalies tended to present the highest MSE values, and consequently the highest values for squared

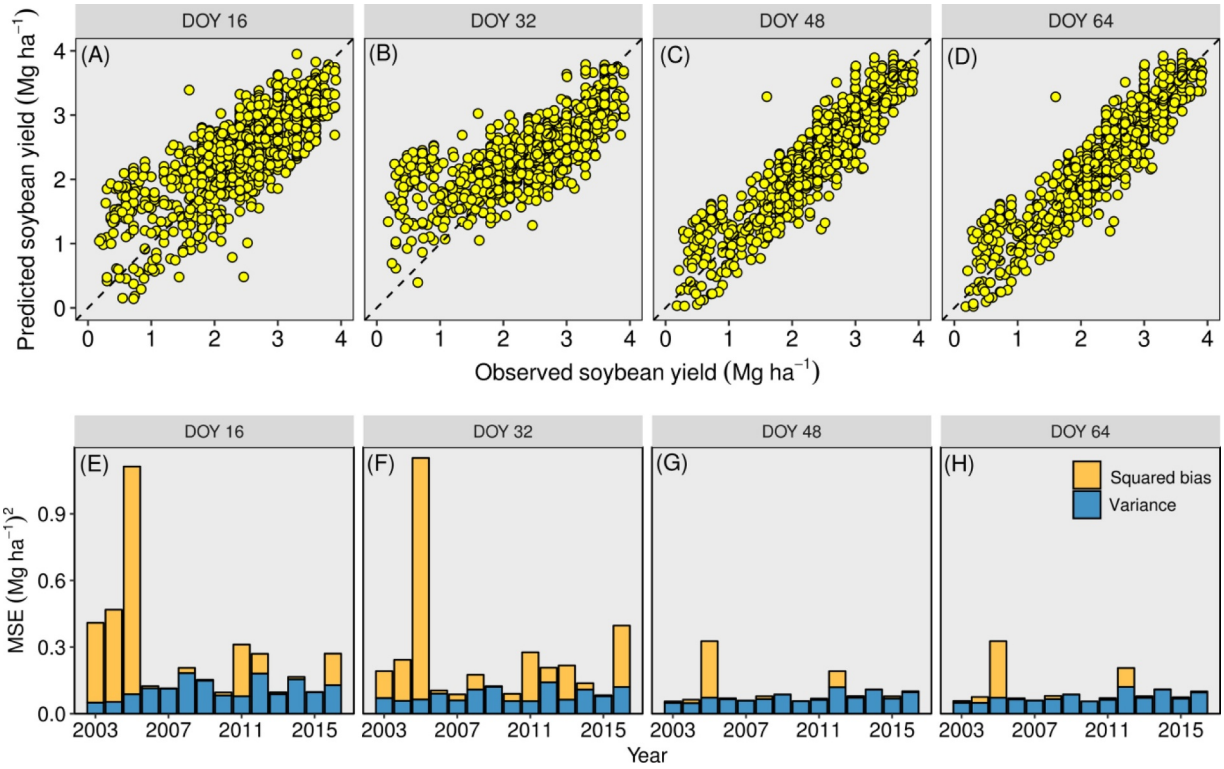

Fig. 3. Upper panels (A to D) portraying the observed versus out-of-sample forecasted corn yield (forecast model with multi-temporal vegetation indices (VIs), land surface temperature and precipitation) for different dates expressed in days of year (DOY). A black dashed line portrays the 1:1 line for the predicted-observed relationship. The Long-Short Term Memory (LSTM) Neural Network used for DOY 32, 48 and 64. Multivariate OLS regression for DOY 16. In bottom panels (E to H) variations in the mean square error (MSE) and its decomposition in squared bias and variance along the years for different dates expressed in DOY.

bias (Fig. 4D). Moreover, it was demonstrated that the frequency of occurrence of years with anomalies equal or higher than the one found in 2005 year seems to be really negligible, ~0.7% or in other words 1 in ~142 years. Following a similar approach, but using weather data instead of yield, we built a second probability density function based on temperature and precipitation. For the second approach, we focused on a specific period of the soybean growing season in Brazil - between DOY 360 and DOY 56 (usually from flowering to seed filling stages), where these variables presented the highest correlation with yield (Fig. 4E). Using this second approach the probability of occurrence for a year with an anomaly equal or higher to 2005 year was 0.3%, close to the 0.7% (but even smaller) that we estimated using the first approach.

4. Discussion

Our results clearly showed that satellite imaging combined with weather data can provide useful information to develop more accurate models to forecast yields of soybean in Brazil. Crop yield forecast based on satellite imagery have become a popular tool for providing near real-time prediction of crop status from small (field and sub-field conditions) (Azzari et al., 2016; Jin et al., 2019, 2017; Lobell et al., 2015; Peralta et al., 2016; Schwalbert et al., 2018) to medium/large domains (county/state) (Bolton and Friedl, 2013; Cai et al., 2019; Johnson, 2014; Lobell, 2013; Peng et al., 2018; Sakamoto et al., 2014; Shao et al., 2015). Furthermore, the integration of canopy reflectance (sometimes summarized as VIs), LST and precipitation, have been demonstrated as a promising approach to enhance performance of yield forecast models. In this study, the inclusion of additional variables such as LST and precipitation decreased the MAE, RMSE, and MSE by 16, 15, and 30%, respectively (averaged over all the dates for the OLS algorithm) (Fig. 5 of Appendix F). The negative correlation of heat, vapor pressure deficit, and the positive correlation of precipitation (Cai et al., 2019; Johnson, 2014; Peng et al., 2018) have been successfully explored in combination with multi-temporal VIs for providing more accurate near real-time forecasts for different crops.

Most of the algorithms used for exploring relationships between yield - multi-temporal VIs and weather variables rely on multivariate OLS (Cai et al., 2019; Lobell et al., 2015; Sakamoto et al., 2014), random forest (Cai et al., 2019; Shao et al., 2015), Rulequest Cubist, (Johnson, 2014), or supported vector machine (Cai et al., 2019). Despite those algorithms usually presents a satisfactory performance for the aforementioned task, they are not prepared for dealing with time-ordered data. Since VIs, LST, and weather variables are inherently temporal, with past state of these variables usually presenting on the future cause-effect relationship, algorithms able of learning patterns based on the sequence how the data is collected have a great potential for outperforming algorithms that treat data in a static viewpoint. In our study, the LSTM neural network outperformed the multivariate OLS regression and random forest for all the tested dates except for the earliest one. For the earliest date, there was less information from the past (related to the forecast date) to be learned by the LSTM neural network model. The use of LSTM for forecasting crop yield is still limited on literature with only a few research studies exploring this topic (Cunha et al., 2018; Wang et al., 2018; You et al., 2017).

Regardless the choice of the algorithm for modeling the yield-predictors empirical relationship, one of the main challenges on using satellite and weather data as proxies to yield at a regional level still remain on the crop field detection, mainly for countries where the crop field boundary and crop-specific layers are not available. The main outcome of this research was a soybean yield forecast model able to predict yield at the municipality level in RS state, southern Brazil. This model has proven to present a high accuracy even without using any crop specific layer, with performance comparable to the models developed in the US by Johnson (2014) using the CDL as crop mask layer and You et al. (2017) using a general world-wide land cover data derived from MODIS (DAAC, 2015), and models developed in Brazil (for

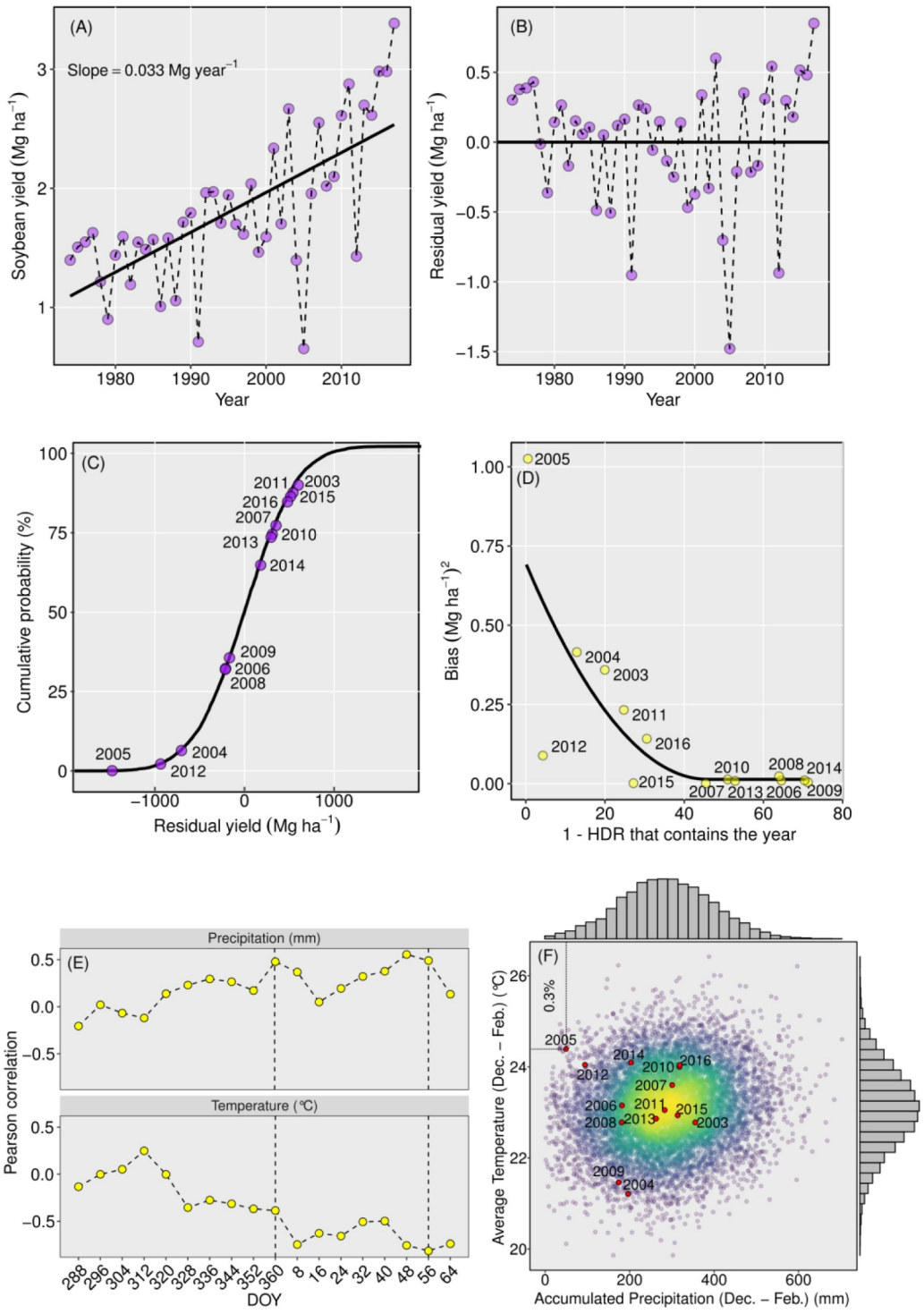

Fig. 4. (A) Relationship between soybean yield and years for the study region. (B) Relationship between residuals for panel A and growing season year. (C) Cumulative distribution function estimated through the Monte Carlo simulation for the yield residuals. (D) Relationship between the squared bias from the DOY 16 yield forecast and the 1-High Density Region (HDR) needed to overlap the considered year – 1-HDR measures how far a specific year is from the mean of the distribution towards the tails, putting equal weights for both tails. (E) Pearson's correlation between soybean yield and average air temperature and precipitation for different 8-days periods during the soybean growing season. (F) Multi-Gaussian probability density function estimated through the Monte Carlo simulation for average air temperature and precipitation (from DOY 360 to DOY 56) for the study region.

four municipalities in Paraná state), by Figueiredo et al. (2016). Similar results also have been reported for corn in the US, demonstrating that models based on multi-temporal NDVI summary statistics had similar performance either using a specific or general (e.g. summer crops, cultivated crop) crop masks (Shao et al., 2015). It is important to note that in the US Midwest and in RS state a corn-soybean rotation on an annual basis is widely adopted. More importantly, previous studies have shown that corn and soybean have relatively similar NDVI profiles (Shao et al., 2010; Wardlow and Egbert, 2008). Therefore, the inclusion of corn in the summer crop mask may still mimic the reflectance signal derived for soybean field only. In RS, the soybean/corn cultivated area is more towards to the soybean side (more frequency of this crop in the rotation), therefore most of the pixels included in this analysis came from soybean fields. The results presented in this paper represents a great prospect for providing municipality-level soybean yield data in a near real-time basis, contrasting with the frequency of the data currently released by SIDRA/IBGE, with the last yield estimation (2016/2017 growing season) announced in 2018.

Furthermore, we extended our analysis pursuing to explore the sensitivity of the time for the forecast model, considering that the importance of a yield forecast is a balance between its accuracy and the timing when the prediction is performed, and usually there is a trade-

off between the error and the date of the prediction (Bolton and Friedl, 2013; Sakamoto et al., 2014; Shao et al., 2015; You et al., 2017). Our results clearly reflected this trade-off since as the forecast is anticipated during the growing season the error of the model tended to rise. Despite of that, soybean yield still can be forecasted at municipality-level in RS, Brazil at DOY 16 with a MAE of 0.42 Mg ha^{1}, and a RMSE of 0.53 Mg ha^{-1}. The penalization in model accuracy for anticipating the yield forecast was greater for years with extreme weather (anomalies from the normal weather) but most of the error from the MSE came from squared bias instead of σ^2. The latter shows that even for years with conditions highly adverse, the model was still able to predict the most and least yielding municipalities even without accurately predicting the absolute soybean yields.

Moreover, yield anomalies such as the ones reported in the 2005 soybean growing season in southern Brazil are unlikely to happen, and the reported model performance (RMSE, MSE, and MAE) was highly penalized by the errors associated with this growing season. After dissecting MSE in σ^2 and squared bias for each one of the years, it became quite clear that years with a lower probability to occur had the highest squared bias, and the squared bias tended to decrease and get stable as the years were settled towards the middle of the yield anomalies distribution (high-density region). The relationship between the probability of a specific type of year to occur and the squared bias is in fact related to the lack of information about that event in the training dataset. Future applications of this model under conditions similar to 2005 year are expected to result in accurate soybean yield forecast, because those events (weather variation) will be already present on the training data. Despite the analysis has been developed for the state of Rio Grande do Sul (Brazil), the general approach described in this manuscript can potentially be applied to other regions around the globe if a reasonable amount of survey data is available for building a reliable crop mapping data layer. This could contribute to support agricultural decisions in regard to managing and transferring risks within the farming system. Consequently, helping farmers to plan interventions, and enable governments and traders to adjust trading schemes, ultimately, avoiding yield failures and food shortages.

5. Conclusions

Multi-temporal satellite imagery combined with weather data can provide useful information, allowing the development of more precise yield forecast models to monitor soybean yield at municipality level. A decrease in the accuracy of the yield forecast model is expected by anticipating the date for yield prediction before harvest, but this study suggests that soybean yield can be predicted by DOY 16 (January 16) with reasonable accuracy. This is approximately 70 days before harvest in RS. Better accuracy (MAE of 0.24 Mg ha^{-1}) can be obtained by DOY 48 (February 17) - 40 days before harvest in RS. The LSTM neural network has been tested to have a better performance relative to random forest or the multivariate OLS regressions, mainly for predictions towards the end of the growing season plausible due to the amount of data collected to compose the time series.

The training and validation approaches were adequate to test the model performance in different weather and yield conditions. Model performance for years with more adverse weather conditions (dramatically different from the normal years) and consequently with higher yield anomalies related to the historical yield distribution is expected to be inferior compared to the overall model accuracy for the remaining years. Under extreme weather conditions, the increase in the error was mainly associated with squared bias than σ^2. For this reason, we expect an increase in the model generalization for future extreme weather events as more data is added into the training process. Despite the analysis being developed for southern Brazil, the general approach described in this study can be potentially applied to other geographical regions around the globe with similar availability of data. This could contribute to support agricultural decisions in regard to managing and transferring risks within crop production and to improve overall crop predictions for policy makers.

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.

Acknowledgments

This study was supported by CAPES Foundation, Ministry of Education of Brazil, Brasilia - DF, Zip Code 70.040-020, process 88887.130848/2016-00. Contribution no. 20-134-J from the Kansas Agricultural Experiment Station.

Appendix A. Additional information about the data sources

Satellite imagery dates

The starting date was selected based on the soybean planting date and phenology based on the analysis of the soybean progress information and satellite images for the last 14 years. Moreover, this period was selected in order to get images covering the time series when the soybean reflectance and yield have the highest correlation (Johnson, 2014).

Climate Hazards Group Infrared Precipitation with Stations – CHIRPS

The CHIRPS provides precipitation data at ~5.5 km resolution by merging satellite and weather station information. This source of data uses satellite in three ways: first, satellite means are used to produce high-resolution rainfall climatologies; second infrared Cold Cloud Duration fields are used to estimate daily rainfall deviation from climatologies. Lastly, satellite precipitation fields are used to guide interpolation through local distance decay functions (Cunha et al., 2018). Precipitation layers were re-projected and down-scaled in order to be combined with the rest of the collected data. Precipitation was accumulated (summed) in an 8 days period to match with NDVI and LST derived from MODIS.

Appendix B. Criteria for selecting variables for composing the yield forecast model

The four criteria for variables being considering as predictors into the model were: i) availability in a spatial format with reasonable resolution, allowing us to summarize the data in a representative way as relative to the geographical inference level (e.g., municipalities), ii) availability of historical records at least until 2003, allowing us to train the models for all the years considered in the study, iii) all the data needs to be resealed in a near real-time basis, or in other words, a short lead time between the information being collected (measured) and become available for downloading, making the model able for updating the yield forecasts with a good frequency and periodicity, and iv) data availability in Google Earth Engine platform to allow future model scalability.

Appendix C. Rural Environmental Registry (Cadastro Ambiental Rural - CAR)

The CAR is an electronic national public registry, mandatory for all rural properties, with the purpose of integrating the environmental information related to the permanent preservation areas (restricted use), remnants of forests, other forms of native vegetation, and the consolidated areas, composing a database for control, monitoring, environmental and economic planning against deforestation. For the purposes of this study, we selected the consolidated rural areas, that is considered as an area of rural property with anthropogenic occupation preexisting on July 22, 2008.

Appendix D. Machine learning hyper-parameter tune

For random forest the considered hyper-parameter were the number of variables in the random subset at each node and the number of trees in the forest. For the LSTM neural network, we tuned the number of hidden layers, number of neuron on each hidden layer, dropout rate, batch size, activation function, learning rate, learning rate decay, and the gradient descent optimization algorithms. Moreover, the number of epoch was set to 60 and the training made use of the *EarlyStopping* callback function from the Keras (Chollet, 2015), with a patience parameter (the number of epochs with no improvement after which training is stopped) equal to 20 to avoid over-fitting. Four years were randomly selected from the data: 2009, 2010, 2012 and 2016 for fine-tuning the machine learning hyper-parameters (sensitivity analyses showed that the changes in the selected years did not significantly impact on the model parameterization). We performed a random search in order to find the best values for the hyper-parameters for the two considered algorithms.

Appendix E. Additional information about the metrics used for model evaluation

The mean absolute error (MAE) represents the average magnitude of the errors while root mean squared error (RMSE) is a quadratic scoring rule for the average magnitude of the error, and it is more useful when large errors are particularly undesirable. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the variance in the individual errors in the sample. If the RMSE equals to the MAE, then all the errors are of the same magnitude. The mean squared error (MSE) measures the average of the squares of the errors and can be dissected into two components, squared bias and variance (σ^2), and this decomposition is helpful to understand if the model error has a more systematic or non-systematic structure.

Appendix F. Effect on model performance metrics by adding land surface temperature (LST) and accumulated precipitation (PPT), in addition to NDVI and EVI, as independent variables

Fig. 5

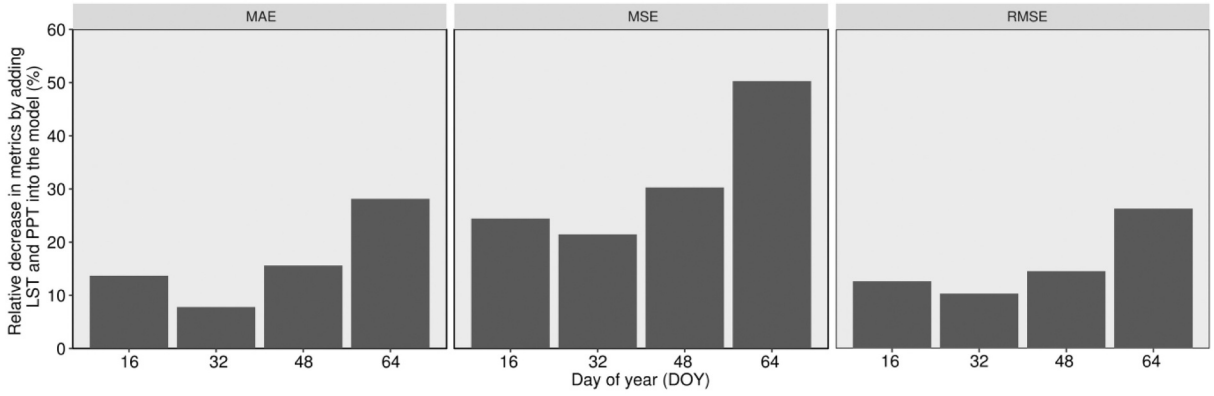

Fig. 5. Relative change in mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) by adding land surface temperature (LST) and accumulated precipitation, in addition to EVI and NDVI, as input variables, for different yield forecast dates expressed in day of year (DOY). Each bar represents the average of all the years following a leave-one-year-out validation for the OLS regression algorithm.

References

Alvarez, R., 2009. Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. Eur. J. Agron. 30, 70–77. https://doi.org/10.1016/j.eja.2008.07.005.

Amato, U., Antoniadis, A., Carfora, M.F., Colandrea, P., Cuomo, V., Franzese, M., Pignatti, S., Serio, C., 2013. Statistical classification for assessing prisma hyperspectral potential for agricultural land use. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 615–625. https://doi.org/10.1109/JSTARS.2013.2255981.

Azzari, G., Jain, M., Lobell, D.B., 2016. Towards fine resolution global maps of crop yields: testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014.

Belgiu, M., Drăgut, L., 2016. Random forest in remote sensing: a review of applications and future directions. ISPRS J. Photogramm. Remote Sens. 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007.

Breiman, L., 2001. Random forests. Mach. Learn 45, 5–32. https://doi.org/10.1007/9781441993267_5.

Chollet, F., et al., "Keras," https://github.com/fchollet/keras, 2015.

Cai, Y., Guan, K., Lobell, D.B., Potgieter, A.B., Wang, S.W., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Cunha, R.L.F., Silva, B., Netto, M.A.S., 2018. A scalable machine learning system for pre-season agriculture yield forecast. Proceedings of the IEEE Fourteenth International Conference on e-Science, (e-Science) 2018, 423–430. https://doi.org/10.1109/eScience.2018.00131.

DAAC, N.L., 2015. The MODIS land products. URLhttp://lpdaac.usgs.gov.

Embrapa, 2018. Soja em números. URLhttps://www.embrapa.br/soja/cultivos/soja1/dados-economicos.

Ferencz, C., Bognár, P., Lichtenberger, J., Hamar, D., Tarcsai, G., Timár, G., Molnár, G., Pásztor, S., Steinbach, P., Székely, B., Ferencz, O.E., Ferencz-Árkos, I., 2004. Crop

yield estimation by satellite remote sensing. Int. J. Remote Sens. 25, 4113–4149. https://doi.org/10.1080/0143116041000169870.
Figueiredo, G.K.D.A., Brunsell, N.A., Higa, B.H., Rocha, J.V., Lamparelli, R.A.C., 2016. Correlation maps to assess soybean yield from EVI data in Paraná State, Brazil. Sci. Agric. 73, 462–470. https://doi.org/10.1590/0103-9016-2015-0215.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
Hamada, Y., Ssegane, H., Negri, M.C., 2015. Mapping intra-field yield variation using high resolution satellite imagery to integrate bioenergy and environmental stewardship in an agricultural watershed. Remote Sens. 7, 9753–9768. https://doi.org/10.3390/rs70809753.
Jin, Z., Azzari, G., Lobell, D.B., 2017. Improving the accuracy of satellite-based high-resolution yield estimation: a test of multiple scalable approaches. Agric. For. Meteorol. 247, 207–220. https://doi.org/10.1016/j.agrformet.2017.08.001.
Jin, Z., Azzari, G., You, C., Di Tommaso, S., Aston, S., Burke, M., Lobell, D.B., 2019. Smallholder maize area and yield mapping at national scales with Google Earth Engine. Remote Sens. Environ. 228, 115–128. https://doi.org/10.1016/j.rse.2019.04.016.
Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128. https://doi.org/10.1016/j.rse.2013.10.027.
Johnson, D.M., Mueller, R., 2010. The 2009 cropland data layer. Photogramm. Eng. Remote Sens. 76, 1201–1205.
Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218–219, 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003.
Khaki, S., Wang, L., 2019. Crop yield prediction using deep neural networks.
Li, A., Liang, S., Wang, A., Qin, J., 2013. Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photogramm. Eng. Remote Sens. 73, 1149–1157. https://doi.org/10.14358/pers.73.10.1149.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.
Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crop. Res. 143, 56–64. https://doi.org/10.1016/j.fcr.2012.08.008.
Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021.
Mildrexler, D.J., Zhao, M., Running, S.W., 2011. A global comparison between station air temperatures and MODIS land surface temperatures reveals the cooling role of forests. J. Geophys. Res. Biogeosciences 116, 1–15. https://doi.org/10.1029/2010JG001486.
Peng, B., Guan, K., Pan, M., Li, Y., 2018. Benefits of seasonal climate prediction and satellite data for forecasting U.S. Maize Yield. Geophys. Res. Lett. 45, 9662–9671. https://doi.org/10.1029/2018GL079291.
Peralta, N., Assefa, Y., Du, J., Barden, C., Ciampitti, I., 2016. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. Remote Sens. 8, 1–16. https://doi.org/10.3390/rs8100848.
R Core Team, 2017. R: a language and environment for statistical computing.
Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2013. Statistical and neural methods for site-specific yield prediction. Trans. ASAE 46, 1–10. https://doi.org/10.13031/2013.12541.
Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2014. Near real-time prediction of U.S. corn yields based on time-series MODIS data. Remote Sens. Environ. 147, 219–231. https://doi.org/10.1016/j.rse.2014.03.008.
Schwalbert, R.A., Amado, T.J.C., Nieto, L., Varela, S., Corassa, G.M., Horbe, T.A.N., Rice, C.W., Peralta, N.R., Ciampitti, I.A., 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171, 179–192. https://doi.org/10.1016/j.biosystemseng.2018.04.020.
Shao, Y., Lunetta, R.S., Ediriwickrema, J., Iiames, J., 2010. Mapping cropland and major crop types across the great lakes basin using MODIS-NDVI data. Photogramm. Eng. Remote Sensing 75, 73–84. https://doi.org/10.14358/PERS.76.1.73.
Shao, Y., Campbell, J.B., Taff, G.N., Zheng, B., 2015. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. Int. J. Appl. Earth Obs. Geoinf. 38, 78–87. https://doi.org/10.1016/j.jag.2014.12.017.
USDA, 2019. USDA foreign agricultural service. URLhttps://www.fas.usda.gov/regions/brazil.
Wan, Z., 2008. New refinements and validation of the collection-6 MODIS land-surface temperature/emissivity product. Remote Sens. Environ. 140, 36–45. https://doi.org/10.1016/j.rse.2013.08.027.
Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series MODIS 250m NDVI data: an assessment for the U.S. Central Great Plains. Remote Sens. Environ. 112, 1096–1116. https://doi.org/10.1016/j.rse.2007.07.019.
You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 4559–4566. https://doi.org/10.1109/MWSCAS.2006.381794.
Wang, A.X., Tran, C., Desai, N., Lobell, D., Ermon, S., 2018. Deep transfer learning for crop yield prediction with remote sensing data. In Conference on Computing and Sustainable Societies (COMPASS), June 20–22, 2018, Menlo Park and San Jose, CA, USA. 1-5. https://doi.org/10.1145/3209811.3212707.