Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning

Xiao-Peng Song[1], Haijun Li[1], Peter Potapov[2], and Matthew C Hansen[2]

[1]Texas Tech University
[2]University of Maryland

November 22, 2022

Abstract

Long-term spatially explicit information on crop yield is essential for understanding food security in a changing climate. Here we present a study that combines twenty-years of Landsat and MODIS data, climate and weather records, municipality-level crop yield statistics, random forests and linear regression models for mapping crop yield in a multi-temporal, multi-scale modeling framework. The study was conducted for soybean in Brazil, the world's largest producer and exporter of this commodity crop. Using a recently developed 30 m resolution, annual (2001-2019) soybean classification map product, we aggregated multi-temporal phenological metrics derived from Landsat and MODIS data over soybean pixels to the municipality scale. We combined phenological metrics with topographic features, long-term climate data, in-season weather data and soil variables as inputs to machine learning models. We trained a multi-year random forests model using yield statistics as reference and subsequently applied linear regression to adjust the biases in the direct output of the random forests model. This model combination achieved the best performance with a root-mean-square-error (RMSE) of 344 kg/ha (12% relative to long-term mean yield) and an r2 of 0.69, on the basis of 20% withheld test data. The RMSE of the leave-one-year-out assessment ranged from 259 kg/ha to 816 kg/ha. To eliminate the artifacts caused by the coarse-resolution climate and weather data, we developed multiple models with different categories of input variables. Employing the per-pixel uncertainty estimates of different models, the final soybean yield maps were produced through per-pixel model composition. We applied the models trained on 2001-2019 data to 2020 data and produced a soybean yield map for 2020, demonstrating the predictive capability of trained machine learning models for operational yield mapping in future years. Our research showed that combining satellite, climate and weather data and machine learning could effectively map crop yield at high resolution, providing critical information to understand yield growth, anomaly and food security.

Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning

Xiao-Peng Song[1,*], Haijun Li[1] , Peter Potapov[2], Matthew C. Hansen[2]

1. Department of Geosciences, Texas Tech University, Lubbock, TX, USA
2. Department of Geographical Sciences, University of Maryland, College Park, MD, USA

*Correspondence to: xiaopeng.song@ttu.edu

Abstract

Long-term spatially explicit information on crop yield is essential for understanding food security in a changing climate. Here we present a study that combines twenty-years of Landsat and MODIS data, climate and weather records, municipality-level crop yield statistics, random forests and linear regression models for mapping crop yield in a multi-temporal, multi-scale modeling framework. The study was conducted for soybean in Brazil, the world's largest producer and exporter of this commodity crop. Using a recently developed 30 m resolution, annual (2001-2019) soybean classification map product, we aggregated multi-temporal phenological metrics derived from Landsat and MODIS data over soybean pixels to the municipality scale. We combined phenological metrics with topographic features, long-term climate data, in-season weather data and soil variables as inputs to machine learning models. We trained a multi-year random forests model using yield statistics as reference and subsequently applied linear regression to adjust the biases in the direct output of the random forests model. This model combination achieved the best performance with a root-mean-square-error (RMSE) of 344 kg/ha (12% relative to long-term mean yield) and an r^2 of 0.69, on the basis of 20% withheld test data. The RMSE of the leave-one-year-out assessment ranged from 259 kg/ha to 816 kg/ha. To eliminate the artifacts caused by the coarse-resolution climate and weather data, we developed multiple models with different categories of input variables. Employing the per-pixel uncertainty estimates of different models, the final soybean yield maps were produced through per-pixel model composition. We applied the models trained on 2001-2019 data to 2020 data and produced a soybean yield map for 2020, demonstrating the predictive capability of trained machine learning models for operational yield mapping in future years. Our research showed that combining satellite, climate and weather data and machine learning could effectively map crop yield at high resolution, providing critical information to understand yield growth, anomaly and food security.

Keywords

Crop yield map; Random forests; Landsat; MODIS; Climate; Weather

1. Introduction

Reliable and timely information on crop production can inform commodity markets, insurance companies, and policy interventions in response to natural disasters and human conflict (Benami et al. 2021; Li et al. 2022; Vroege et al. 2021). Estimating crop production over a spatial unit requires information on crop harvested area and crop yield (i.e. production per unit area). Both harvested area and yield can be derived from statistical field surveys or from satellite observations (Mulla 2013; Weiss et al. 2020) . While many methods exist in mapping crop type and estimating crop area using remote sensing (e.g. Defourny et al. 2019; Gallego 2004; Gonzáles-Alonso and Cuevas 1993; Hu et al. 2021; King et al. 2017; Massey et al. 2017; Skakun et al. 2017; Song et al. 2017; Wardlow and Egbert 2008), studies are increasingly investigating direct mapping of crop yield using remote sensing data. Crop yield maps can facilitate a number of research or practical applications, such as climate impact evaluation and yield gap analysis (Lobell 2013).

Mapping crop yield requires crop type masks as a prerequisite. When crop type masks are available, two different strategies are commonly used to produce spatially explicit information on yield: the model-data integration approach and the remote sensing-based empirical approach. The model-data integration approach seeks to integrate crop simulation models with remote-sensing-derived biophysical variables for yield forecasting (Delécolle et al. 1992; Moulin et al. 1998). Crop simulation models are developed using comprehensive measurements recorded at the plot or field level, such as crop cultivar, sowing date, soil property, water and nutrient inputs, weather, and plant physiological and morphological features (e.g. leaf area index or LAI) (de Wit et al. 2019; Holzworth et al. 2014; Jones et al. 2003; Williams et al. 1989; Yang et al. 2004). The modeled processes of crop growth can be used to predict crop productivity and to evaluate the impacts of agricultural management and environmental stressors. Various techniques have been proposed to "spatialize" crop process models using time-series of satellite-based soil, plant and environmental variables, such as soil moisture, normalized difference vegetation index (NDVI), LAI, green area index (GAI), and fraction of photosynthetically active radiation (fPAR) (Battude et al. 2016;

Claverie et al. 2012; de Wit et al. 2012; Doraiswamy et al. 2004; Duchemin et al. 2008; Huang et al. 2015; Ines et al. 2013; Kang and Özdoğan 2019; Nearing et al. 2012). Yet, a general limitation of applying crop process models over large areas is the lack of sufficient and accurate information about model inputs (Duchemin et al. 2008; Jin et al. 2018). Moreover, the model-data integration approach usually does not serve the purpose of high-resolution yield mapping. The computational cost of per-pixel crop simulation is high, but such barriers are being lifted by the recent development of cloud-computing platforms such as Google Earth Engine (Gorelick et al. 2017).

The remote sensing-based empirical approach for crop yield mapping employs regression or machine learning techniques to relate vegetation variables at key crop growth stages directly to yield. An early work by Tucker et al. (1980) showed that time-integrated NDVI had significant correlation with grain yield in a winter wheat field in Beltsville, Maryland. Becker-Reshef et al. (2010) demonstrated that seasonal peak NDVI from the Moderate Resolution Imaging Spectroradiometer (MODIS) strongly correlated with winter wheat yield in Kansas and Ukraine. Franch et al. (2015) extended the Becker-Reshef et al. (2010) approach by including Growing Degree Day (GDD) information, which enabled yield forecasting at about one month prior to peak NDVI. Funk and Budde (2009) found that time-integrated MODIS NDVI adjusted to the onset of the rainy season correlated well with maize production in Zimbabwe. Yield estimation may be improved by incorporating explicit phenology information using other vegetation indices beyond NDVI. Building on the work of Funk and Budde (2009), Bolton and Friedl (2013) suggested that MODIS-based two-band Enhanced Vegetation Index (EVI2) standardized by the greenup date correlated better than NDVI with county-level yield for maize, but indifferent for soybean, over central US. Similarly, Sakamoto et al. (2013) applied a phenology detection method to identify corn silking stage and demonstrated that MODIS-derived Wide Dynamic Range Vegetation Index (WDRVI) (Gitelson 2004) at that stage had high correlations with yield over major corn producing states of the US. Johnson (2014) proved that daytime land surface temperature (LST) negatively correlated with maize and soybean yield in the US while MODIS peak NDVI positively correlated with yield. Recently, Skakun et al. (2021) investigated the utility of Landsat-8, Sentinel-2, WorldView-3 and Planet data for corn and soybean yield mapping over a number of sample sites in Iowa, and found that surface reflectance from red-edge bands performed better than vegetation indices to reveal field-level yield variability. Lobell et al. (2015) developed an approach that used simulations from a crop model to train a regression to predict yields from satellite observations, and the approach was tested in industrial as well as smallholder systems (Jin et al. 2019).

While regression-based methods are straightforward to implement, more complex algorithms and data analytic techniques such as machine learning algorithms are being increasingly investigated. Using NDVI from the Advanced Very High Resolution Radiometer (AVHRR) and MODIS, Li et al. (2007) compared multivariate linear regression and artificial neural networks for modeling corn and soy yield over a number of sample counties in the US corn belt. Likewise, Johnson et al. (2016) compared the performance of multiple linear regression and nonlinear Bayesian neural networks and model-based recursive partitioning for forecasting barley, canola and spring wheat yields on the Canadian Prairies. Based on the finding that NDVI and LST highly correlated with crop yield, Johnson (2014) built a regression tree model using multiple years of county-level yield statistics as reference and applied the model to MODIS data to forecast corn and soybean yield at 250 m resolution in the US. Cai et al. (2019) tested the utility of the enhanced vegetation index (EVI) from MODIS and solar-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY, and regression and machine learning algorithms for wheat yield prediction in Australia, and found that the combination of MODIS EVI, climate data and support vector machines (SVM) could achieve high performance in yield prediction. Mateo-Sanchis et al. (2019) proposed a multi-sensor metric, namely the time lag between MODIS EVI and vegetation optical depth (VOD) from the Soil Moisture Active Passive (SMAP) satellite, as input to nonlinear kernel ridge regression for modeling county-scale crop yield in the US corn belt. Deep learning algorithms are also being explored in yield estimation. Schwalbert et al. (2020) developed a method for in-season soybean yield forecasting using the Long-Short Term Memory (LSTM) algorithm, MODIS-based NDVI, EVI and

LST data, and precipitation data at the municipality scale in the Brazilian state of Rio Grande do Sul. Recent research has also started to combine machine learning and crop models by incorporating output variables from crop models as input features to machine learning algorithms for yield estimation (Paudel et al. 2021; Shahhosseini et al. 2021).

These previous studies clearly show that crop yield estimation represents a continually active line of research in remote sensing. The primary goal is to improve the accuracy of yield estimation using new data and techniques, and/or to advance the date of in-season forecasting. However, most previous studies are demonstrative research with limited spatial extents and/or temporal span in their study areas. Studies exploring the long-term satellite data archives to evaluate the variability of crop yields also exist albeit over small study areas (e.g. Gao et al. 2018; Liu et al. 2020). More importantly, common to most yield mapping studies, crops in the temperate climate zone are often the target crops and target regions. Long-term, large-area crop yield mapping in the tropics does not exist. Unlike the temperate region where climate conditions are relatively homogenous and crop phenologies are largely synchronous, cropping systems in the tropics are more complex in the sense that planting and harvesting schedules could be substantially different for the same crop (e.g. soybean in Brazil) (Song et al. 2021). Statistics-based phenological metrics derived from time-series of satellite data can capture the salient features of vegetation phenology while maintaining high spatial and temporal data consistency, and thus, provide a unique advantage to large-area vegetation type mapping (DeFries et al. 1995; Hansen et al. 2013; Song et al. 2018). The main objective of this study is to explore the utility of statistical metrics derived from Landsat and MODIS data as well as machine learning algorithms for high-resolution, long-term crop yield mapping in the tropics. Producing long-term spatially explicit yield information is especially imperative in tropical countries, where agricultural production is growing rapidly, causing detrimental impacts to natural environment (Gibbs et al. 2010; Potapov et al. 2022; Song et al. 2018; Zalles et al. 2021). We focus on annual soybean yield in Brazil over 2001-2020 in this study.

2. Data and Methods

2.1. Study area

Our study area covers the southern hemisphere portion of Brazil. Brazil is the world's leading producer and exporter of soybeans, accounting for more than 35% of global production and about half of the world's total export (FAO 2020). Based on statistics from the Food and Agriculture Organization of the United Nations (FAO), soybean production in Brazil has tripled from 37.9 million tons in 2001 to 114.3 million tons in 2019 (FAO 2020). Over the same time period, soybean cultivation area in Brazil increased from 14.0 Mha to 35.9 Mha, and the national average yield increased from 2.71 to 3.18 tons/ha with the maximum yield of 3.39 tons/ha achieved in 2018 (FAO 2020). The dramatic increase in soybean cultivation in Brazil (Figure 1) has directly and indirectly caused widespread natural vegetation loss and cascading environmental impacts in the Amazon, Cerrado and other biomes (Song et al. 2021a; Zalles et al. 2019).

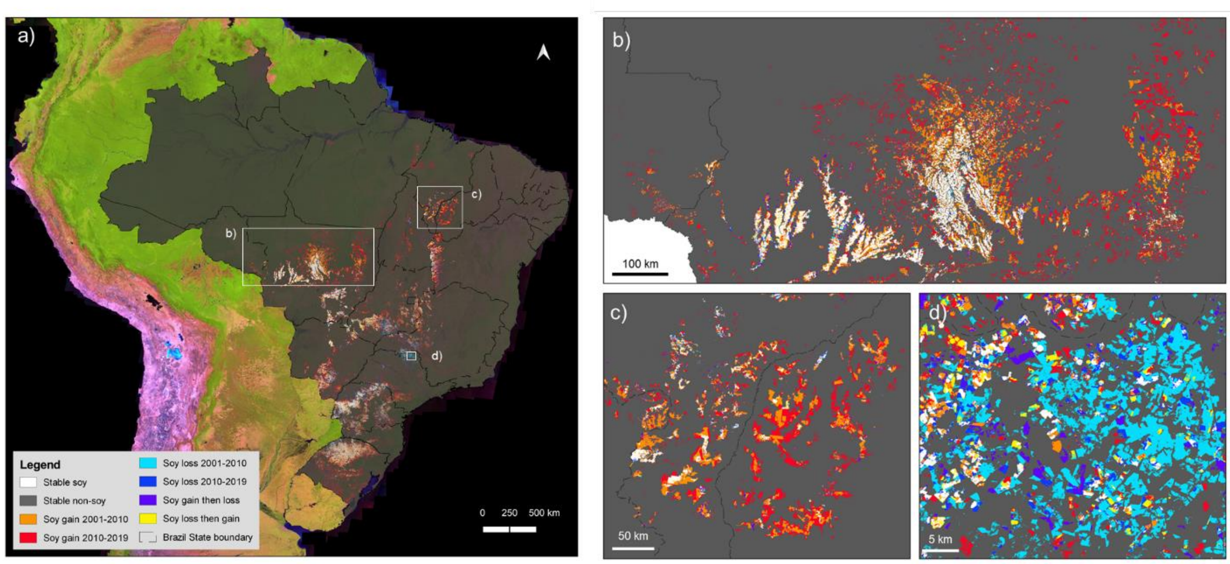

Figure 1. Soybean expansion in Brazil mapped using satellite data. (a) Soybean change during 2001-2010 and 2010-2019. For simplicity to visualize, the annual 2001-2019 classification maps are used to create bi-temporal change layers. Landsat mosaic of South America is used as the backdrop in (a), and gray shaded area represents the study area of Brazil. Regional details over two soybean expansion frontiers are

shown in (b) Mato Grosso and (c) MaToPiBa (Maranhao, Tocantins, Piaui and Bahia). Reduction in soybean cultivation was observed along the border between Sao Paulo and Minas Gerais, shown in (d).

2.2. Satellite data and products

We used Landsat and MODIS as the main satellite data to derive vegetation characteristics of soybean plants, as they represent the most consistent satellite data records over the past two decades. According to the United States Department of Agriculture (USDA) crop calendars for Brazil, soybeans in Brazil are typically planted in October to December and harvested in March to May (https://ipad.fas.usda.gov/rssiws/al/crop_calendar/br.aspx). In our study, all Landsat and MODIS observations acquired between November 1st and April 30th of the next year from 2000 to 2019 were processed. The MODIS surface reflectance (SR) data in blue (469 nm), green (555 nm), red (645 nm), near-infrared (NIR, 858 nm), shortwave infrared (SWIR, 1640 nm and 2130 nm) and thermal (11,030 nm) wavelengths were obtained as 16-day composites from the MOD44C product, same as the MOD09GA, MOD09GQ and MODTBGA v006 products (Vermote and Wolfe 2015). Landsat images acquired by the Thematic Mapper (TM), Enhanced Thematic Mapper Plus (ETM+), and Operational Land Imager (OLI), with blue, green, red, NIR, and SWIR bands, were converted from top-of-atmosphere reflectance to normalized surface reflectance (NSR) through an automated data processing system (Potapov et al. 2020). Using MODIS SR as normalization target, the system corrected atmospheric and anisotropic effects of Landsat after at-sensor radiance calculation, cloud, shadow and haze masking. The Landsat NSR, from all sensors, was then processed to 16-day composites consistent with the MODIS product. Both Landsat NSR and MODIS SR 16-day time-series were used to create seasonal phenological metrics, including NDVI, EVI, normalized difference water index (NDWI) and other band ratio indices (Table 1). A complete description of Landsat data processing and the freely available software tools to generate phenological metrics is provided in Potapov et al. (2020).

Table 1. Input features for modeling and mapping soybean yield in Brazil. Please see Supplementary Information for the complete list of variables.

Category	Input Features	N
Landsat-based	Seasonal vegetation phenological metrics derived from Blue, Green, Red, NIR, SWIR1, SWIR2 and thermal bands	50
MODIS-based	Seasonal vegetation phenological metrics derived from Blue, Green, Red, NIR, SWIR1, SWIR2 and thermal bands	24
Topographic	DEM and Slope	2
Climate	Long-term (1971-2000 average) climate data, monthly (October to May) TMP (mean 2 m temperature), DTR (diurnal 2 m temperature range), PRE (precipitation rate), VAP (vapor pressure), WET (wet days), CLD (cloud cover), TMN (minimum 2 m temperature), TMX (maximum 2 m temperature) and PET (potential evapotranspiration)	72
Weather	Annual (2000 through 2019) in-season weather data, monthly (October to May) TMP, DTR, PRE, VAP, WET, CLD, TMN, TMX and PET	72
Soil	Water storage capacity, topsoil and subsoil bulk density, cation exchange capacity of the clay fraction in the topsoil and subsoil, topsoil and subsoil clay, sand and silt fractions, topsoil and subsoil pH, and area weighted topsoil and subsoil carbon content	15

We used a recently developed 30 m resolution (0.00025° × 0.00025°), annual, 2001-2019 soybean classification map product (Song et al. 2021a) as masks to constrain the yield modeling and mapping to identified soybean pixels (Figure 1). For simplicity and consistent with the soybean classification map product, in this study we refer to a cropping year by the harvest year. For example, year 2001 indicates the 2000/01 cropping year. The soybean classification product was developed using the above Landsat and MODIS data as input in addition to 30 m resolution topographic features from the Shuttle Radar Topography Mission (SRTM) data. Continentally distributed field observations collected over three years (2017, 2018 and 2019) were used as training to calibrate a multi-year bagged decision tree model for soybean classification. The overall accuracy of the soybean classification maps for the years of 2017, 2018, and 2019, where we had probability field sample for validation, was 96%, 94% and 96%, respectively, with high and balanced producer's and user's accuracies (Song et al. 2021a).

2.3. Climate and weather data

Monthly climate and weather covariates were obtained from the Climatic Research Unit gridded Time Series (CRU TS) version 4.04 dataset (Harris et al. 2020). The variables included TMP (mean 2 m temperature), DTR (diurnal 2 m temperature range), PRE (precipitation rate), VAP (vapor pressure), WET (wet days), CLD (cloud cover), TMN (minimum 2 m temperature), TMX (maximum 2 m temperature) and PET (potential evapotranspiration) at a spatial resolution of $0.5° \times 0.5°$. We calculated monthly average values from 1971 to 2000 for the months from October to May to represent long-term climatology. For each year between 2000 to 2019, we directly used the monthly values for the months from October to May to represent in-season weather (Table 1).

2.4. Soil data

The Regridded Harmonized World Soil Database v1.2 at $0.05° \times 0.05°$ spatial resolution (FAO/IIASA/ISRIC/ISSCAS/JRC 2012; Wieder et al. 2014) were obtained and processed similar to the climate and weather data. The soil variables included available water storage capacity, topsoil (0-30 cm) and subsoil (30-100 cm) bulk density, cation exchange capacity of the clay fraction in the topsoil and subsoil, topsoil and subsoil clay, sand and silt fractions, topsoil and subsoil pH, and area weighted topsoil and subsoil carbon content (Table 1).

2.5. Municipal yield statistics

We obtained soybean yield statistics at the municipality scale for every year between 2001 and 2019 from the Brazilian Institute of Geography and Statistics (IBGE) Municipal Agricultural Production database (https://sidra.ibge.gov.br/). The size of the municipalities where soybeans are cultivated varies widely from south (small) to north (large), with a median size of approximately 48 Kha, the first quantile of 22

Kha and the third quantile of 135 Kha. These yield statistics were used as reference data for training and evaluation (Figure 2).

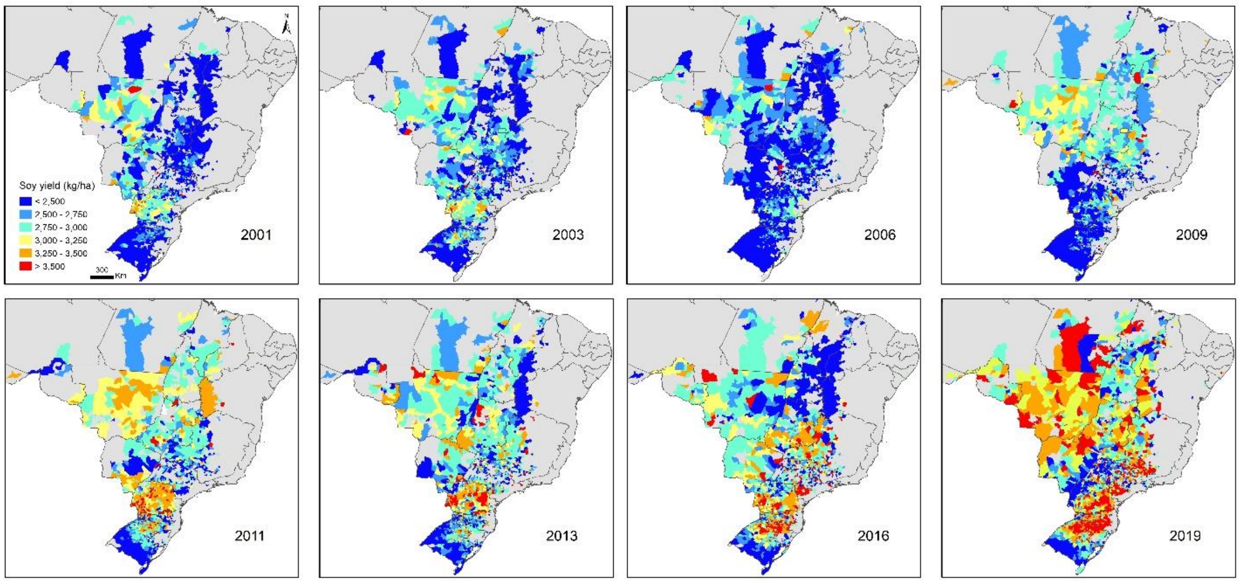

Figure 2. Municipality-level yield statistics from the Brazilian Institute of Geography and Statistics (IBGE) were used as reference for modeling and mapping soy yield.

2.6. Modeling yield

The overall workflow of modeling and mapping soybean yield is presented in Figure 3. Major steps include spatial aggregation of remote sensing (RS)-based vegetation phenological metrics, topographic (topo) features, climate, weather, and soil variables to municipal scale, categorical feature selection, random forests (RF) (Breiman 2001) model training, RF prediction, bias correction, per-pixel RF model selection and composition, and map evaluation. Details of each step are described as follows.

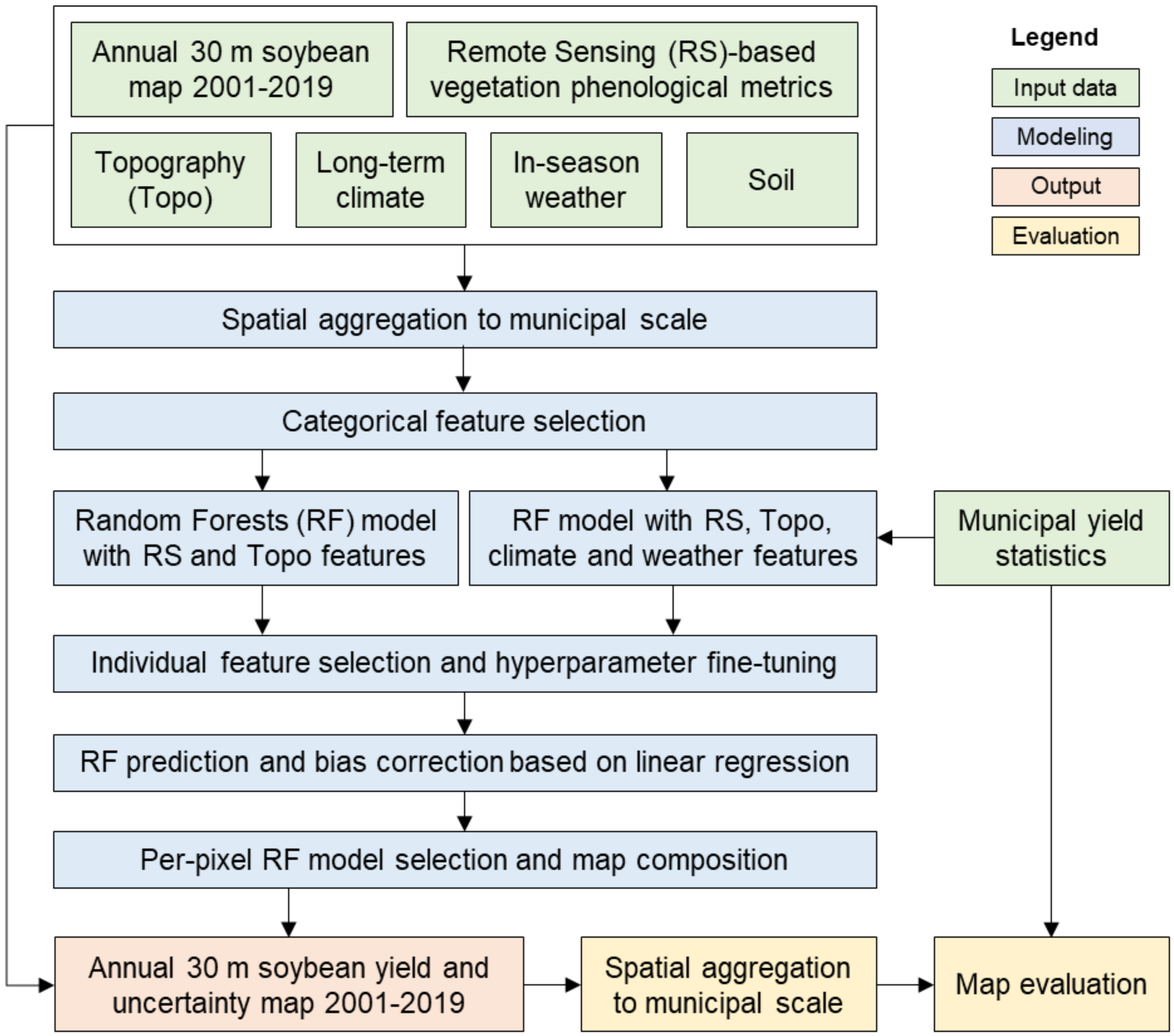

Figure 3. Overall workflow of mapping annual soybean yield 2001-2019 using satellite data, climate, weather, soil and topography data, municipality statistics, random forests and linear regression models. Two random forests models were trained and implemented with more details reported in the text.

The 0.5° × 0.5° climate and weather data, and the 0.05° × 0.05° soil data were first resampled using nearest resampling to 0.00025° × 0.00025° to match the spatial resolution of the soybean classification map, remote sensing data and topographic features. With the annual soybean classification map as a mask, we aggregated these input datasets to municipal scale by taking the average value over soybean pixels in each municipality. The spatial aggregation step was conducted for every year independently between 2001 and 2019. To remove the non-soybean and low-soybean municipalities, we selected the municipalities with annual soybean pixels ≥ 50,000, resulting in a total of 15,784 municipalities across the 19-year period. These municipalities contained 95% of all mapped soybean pixels over the study period.

To investigate the relative utilities of these multi-source, multi-resolution input datasets for yield modeling, we conducted three progressive experiments using categorical feature selection. Specifically, we built three random forests models with (1) RS and topo features as input, (2) RS, topo, climate and weather features as input, and (3) RS, topo, climate, weather and soil features as input. Performance of model #1 represents the utility of RS and topo features to model yield. Improved performance of model #2 over model #1 would represent the value of weather and climate data. Likewise, improved performance of model #3 over model #2 would represent the value of the soil variables.

Municipal yield statistics were used as reference for all three models. For each model, we randomly selected 80% municipalities as training (n = 12,649) and the remaining 20% was reserved for independent test (n = 3,135), with both training and test data covering all 19 years. We calculated root-mean-square-error (RMSE), mean bias error (MBE), mean absolute error (MAE), and r^2 using both training and test data for all three models. To further enhance the robustness of the model evaluation and to eliminate potential bias from a particular realization of sampling, we implemented a Monte Carlo method and repeated the random training/test split, model training and evaluation 100 times. The final model performance was represented using box plots of RMSE, MBE, MAE and r^2 of the 100 runs.

In addition to model evaluation with 20% withheld test data, we also conducted the leave-one-year-out model assessment. For every year between 2001 and 2019, we used 18-years of data to calibrate the random forests models and used the model to predict over the left-out year. For the left-out year, we compared the predicted yield with reference statistics and calculated error metrics.

Our model assessment revealed that climate and weather variables significantly improved model performance, but soil variables did not further improve model performance (more details are provided in the Results and Discussion sections). Therefore, the model with RS, topo, climate and weather variables as input (i.e. model #2) was selected as the primary model for yield estimation. However, due to the coarse spatial resolution (0.5° × 0.5°) of the climate and weather data, spatial grid patterns were noticed in some regions. To remove these artifacts, we implemented model #1 (RS and topo features as input) as a secondary model, and results of the two models were combined (see more details below).

To improve computational efficiency, we conducted individual feature selection for both models. For each RF model, we trained the model using all features as input, ranked each feature and selected the top features with a cumulative importance of greater than 95%. We also constructed a correlation matrix of the features and removed those less important features that had a correlation coefficient of greater than 0.95 with the more important ones. Error metrics were calculated for all as well as selected features to demonstrate the comparable performance of trained models. We implemented the random forest classifier function in the sklearn package in python. The RF parameters fine-tuned included n_estimators (number of trees), max_features (number of features to consider at every split), max_depth (maximum number of levels in a tree), min_samples_split (minimum number of samples required to split a node), min_samples_leaf (minimum number of samples required at each leaf node). We applied a randomized search on hyper-parameters followed by a grid search to determine the exact values for these parameters.

The immediate output of the two RF models include predicted soybean yield, represented as the mean value of all trees in the forest, and associated uncertainty, represented as the standard deviation of all trees in the forest. For continuous variables, random forests could generate underestimation at the high-end of the variable and overestimation at the low-end of the variable because of the effect of "regression to the mean" (Huang et al. 2016; Zhang and Lu 2012). Such is the case for our yield modeling in this study. To correct these systematic biases, we followed Zhang and Lu (2012) and Huang et al. (2016), and applied linear regression using the municipal yield statistics as the dependent variable and the RF-predicted yield as the independent variable. The derived linear equation was subsequently applied to the adjust the RF-predicted yield and uncertainty.

We implemented the two calibrated random forest models (models #1 and #2) and their associated linear regressions independently using the annual input datasets. The outputs were two sets of 30 m resolution soybean yield and uncertainty maps for every year between 2001 and 2019. We created a final soybean yield and uncertainty map for every year through per-pixel composition, where, for every pixel, the soybean yield and associated uncertainty were selected from the model with a smaller uncertainty.

2.7. Yield map evaluation

We evaluated the quality of the annual, 30 m resolution soybean yield maps at the municipal scale. Average yield was derived from the maps, and compared to municipal yield statistics as reference. We computed the difference of the two datasets and constructed a histogram. We calculated RMSE, MAE, MBE, and r^2, and created scatter plots using the 19 years of data. We also calculated these error metrics for every year to evaluate the temporal consistency of the yield map time series.

3. Results

3.1. Model selection and performance

Using remote sensing-based vegetation phenological metrics and topographic features as input to random forests (model #1) produced an r^2 of 0.74, an RMSE of 323 kg/ha, an MBE of 0 kg/ha and a MAE of 240 kg/ha for training data. Compared to the 2001-2019 national average yield of 2,869 kg/ha, this RMSE represents 11% error. Adding climate and weather variables to input (model #2) significantly improved model performance, as represented by the increase in r^2 and reduction in RMSE and MAE, for both training and test data. The improved model had an r^2 of 0.79, an RMSE of 294 kg/ha, an MBE of 0 kg/ha and a MAE of 218 kg/ha for training data, and an r^2 of 0.69, an RMSE of 356 kg/ha, an MBE of 15 kg/ha and a MAE of 264 kg/ha for test data. Adding soil variables to input (model #3) showed little to no value in further improving model performance. Therefore, we discarded model #3 and implemented model #1 and #2 in this study. Both model #1 and #2 were chosen because although climate and weather data demonstrated considerable utility in modeling soybean yield, their coarse spatial resolution (0.5° × 0.5°)

caused apparent grid patterns when the model was applied to 30 meter spatial resolution, whereas model #1 generated spatially coherent results. Moreover, individual feature selection not only improved computational efficiency but also improved model accuracy. Consistent for all model categories, there remained some differences between training and test, indicating potential overfitting of the models. This was likely due to the lack of high-quality soil data and other important agricultural management variables (e.g. fertilizer use) in the model (please see more details in the Discussion section).

Predicted yield from random forests models were highly consistent with reference yield from municipal statistics (Figure 4). However, the direct outputs of the random forests models under-estimated yield at the high end and over-estimated yield at the low end (Figure 4a and 4c). Applying a linear regression successfully corrected these systematic biases for both models (Figure 4b and 4d). Moreover, the overall model performance was also slightly improved, as demonstrated by the reduction in RMSE and MAE for both training and test results. For instance, the training accuracy in terms of RMSE was reduced from 294 to 278 kg/ha and the test accuracy was improved from 356 to 344 kg/ha for model #2 after bias adjustment (Figure 4a vs 4b).

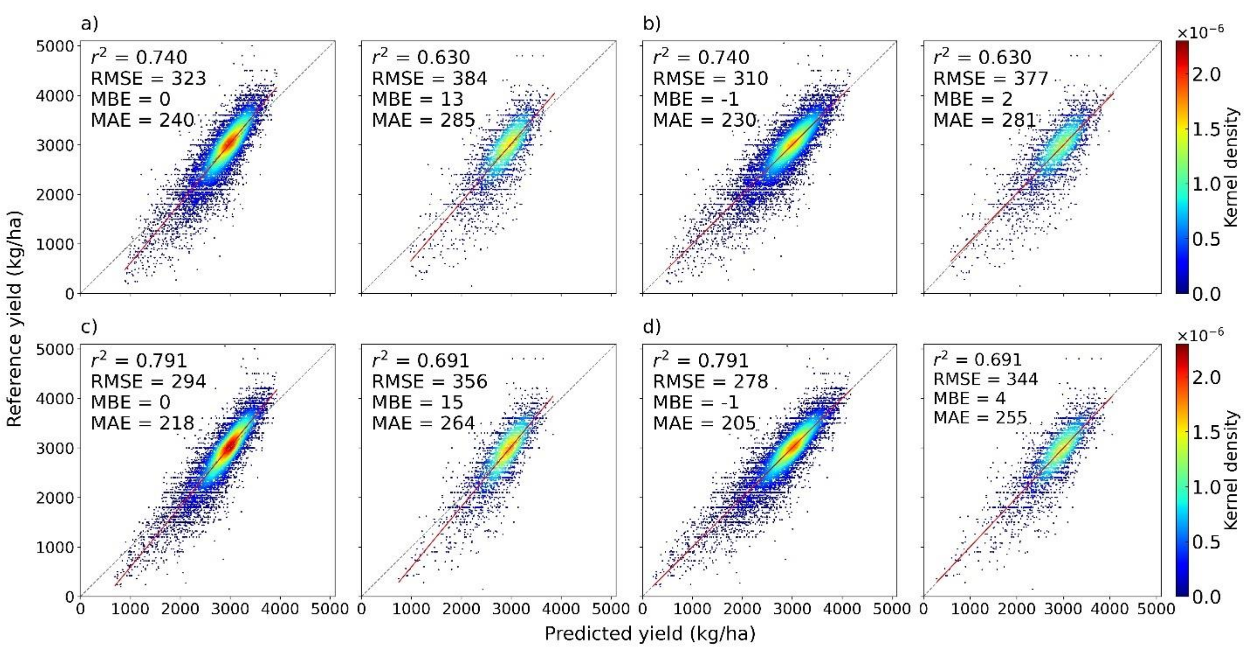

Figure 4. Performance of yield models before and after systematic bias adjustment using linear regression. a) Random forests (RF)-predicted soybean yield against reference yield from municipal statistics. Input data for RF include remote sensing, topographic features, climate and weather variables. The left panel is density scatter plots using training data and the right panel is density scatter plots of independent test data. The red lines on both panels represent the linear regression line. b) Same as a), but a linear regression was applied to adjust bias in RF outputs. c) RF-predicted soybean yield against reference yield. Input data for RF only include remote sensing and topographic features. d) Same as c), but after linear bias adjustment.

Although the model was trained using all 19-years of data as input, evaluation of model performance at the annual time scale revealed consistent model performance across all 19 years (Figure 5). Based on the withheld test data, the 19-year overall RMSE was 344 kg/ha and the r^2 was 0.69. The RMSE represents 12% error relative to long-term yield mean. The annual RMSE values ranged from 214 kg/ha in 2010 to 456 kg/ha in 2005, and the annual r^2 values ranged from 0.39 in 2003 to 0.76 in 2004. No significant systematic bias was observed for any of the years (Figure 5).

The leave-one-year-out model assessment revealed that the yield models performed well for most of the 19 years, but performed relatively poorly for 2005 and 2015 with notably higher RMSE and lower r^2, respectively (Figure 6). The RMSE of the leave-one-year-out assessment ranged from 259 kg/ha to 816 kg/ha. These results are in general comparable to regional studies of satellite-based soybean yield mapping in the Midwest of the United States (Lobell et al. 2015) and Southern Brazil (Schwalbert et al. 2020). Both 2005 and 2015 did not show notable performance deficiency when data of the two years were included in training (Figure 5). Comparison between annual accuracies of the two model assessments (Figures 5 and 6) suggests that model trained with long time series of data generally perform well for unseen years. The comparison also highlights the significance of including both good and poor harvesting years in training for enhancing the temporal generalization and predictive capability of trained models.

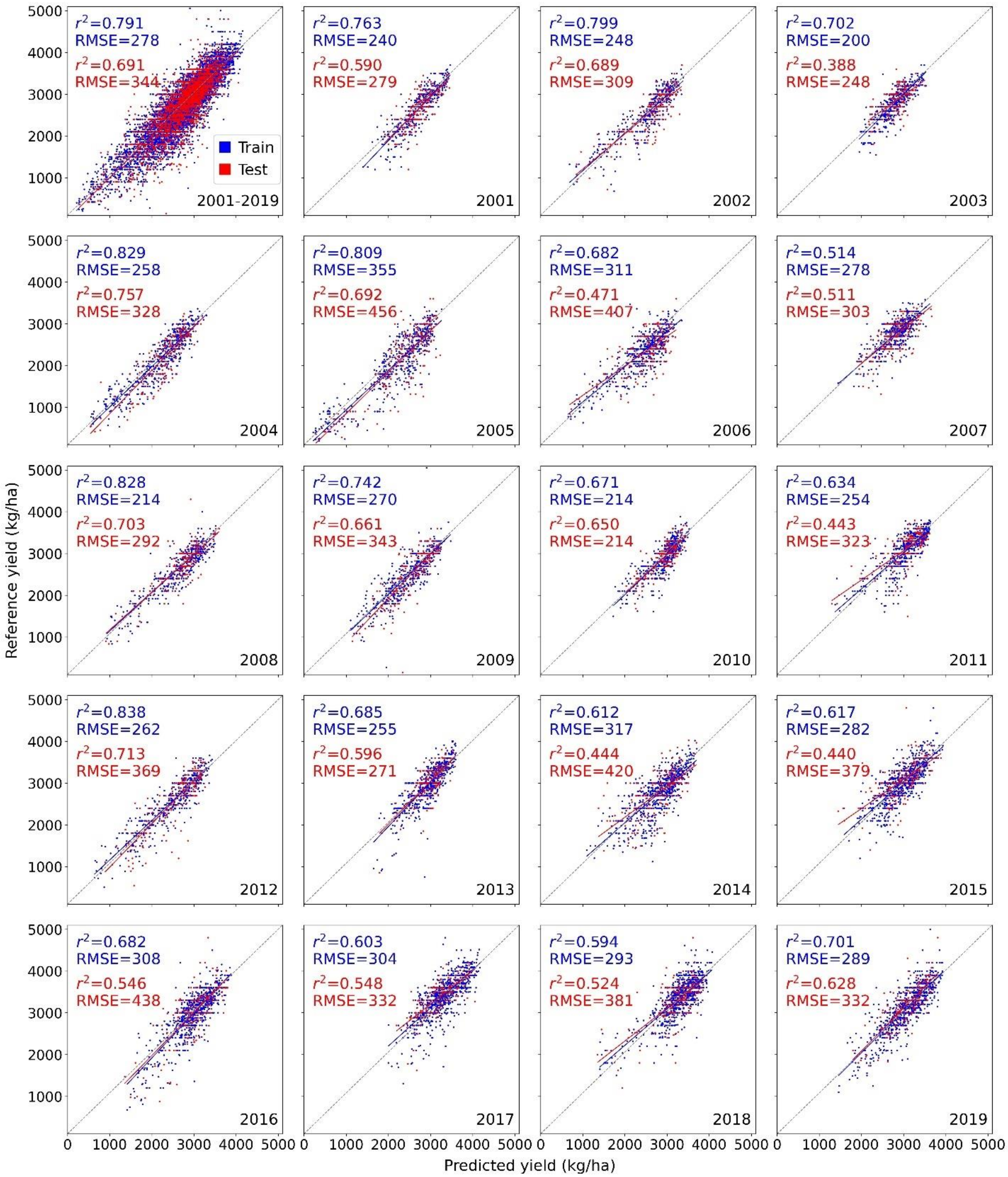

Figure 5. Performance of yield model at an annual time scale. X-axis represents model-predicted yield, and y-axis represents reference yield from municipal statistics. The top-left scatter plot is a combination of the two scatter plots in Figure 5b. Scatter plots are made using training data and withheld test data. Input data for model include remote sensing, topographic features, climate and weather variables.

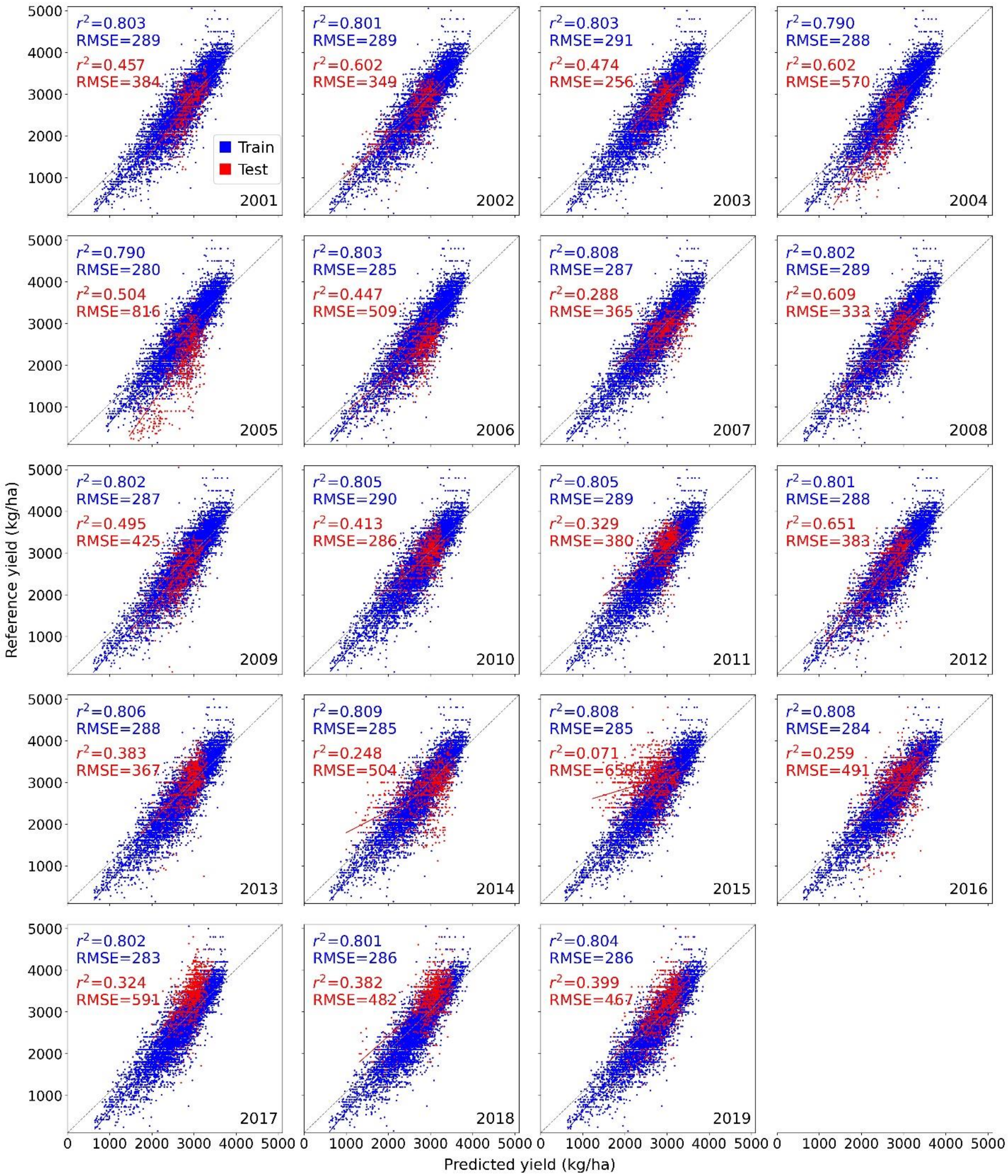

Figure 6. Leave-one-year-out model assessment. For each year between 2001 and 2019, 18-years of data were used to training the model (blue dots and text), which was used to predict over the left-out year.

Municipal statistics of the left-out year were used as reference to evaluate the model performance (red dots and text).

3.2. Annual soybean yield and uncertainty maps

Implementing the calibrated random forests and linear regression models at 30 m spatial resolution generated spatially and temporally coherent soybean yield distributions across Brazil from 2001 to 2019 (Figure 7a). Considerable spatial heterogeneity in soybean yields was observed across the country. In 2001, the highest soybean yield regions included central Mato Grosso and western Parana (also see Figure 2a), and the lowest yield regions included Rio Grande do Sul, eastern Goias, western Minas Gerais, and western Bahia. Increase in soybean yield was found in many regions, most notably in northern Rio Grande do Sul and western Bahia (also see Figure 2b). Soybeans in Mato Grosso experienced not only a substantial area expansion but also considerable yield growth. Per-pixel uncertainty of soybean yields (Figure 7b) showed that the uncertainty estimates were mostly between 300 kg/ha to 500 kg/ha. Moreover, the uncertainty distribution varied both spatially and temporally, with the south region (e.g. Rio Grande do Sul) appeared to have slightly higher uncertainties than center west (e.g. Mato Grosso).

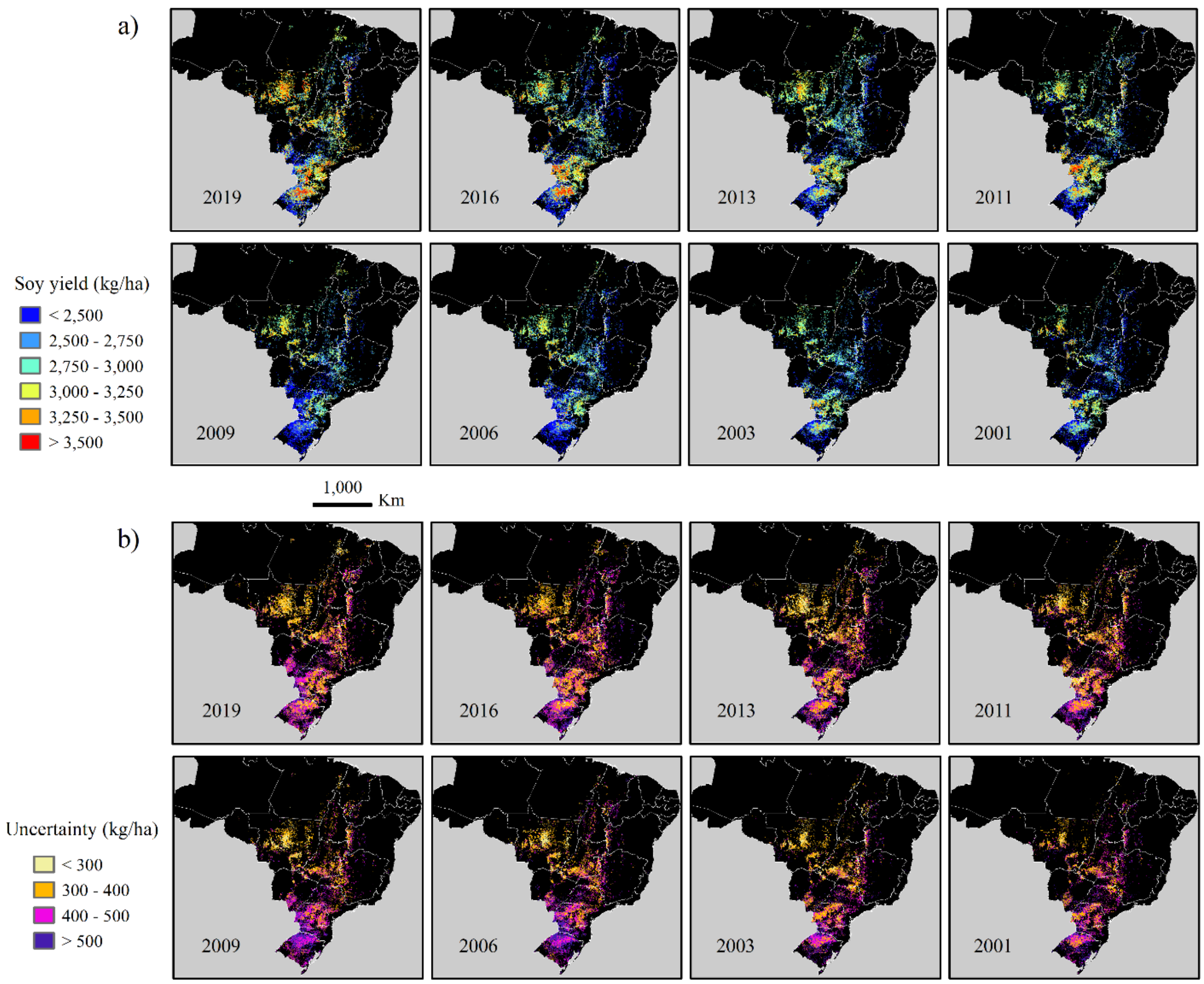

Figure 7. Annual soybean yield and uncertainty maps for selected years over Brazil. Yield and uncertainty maps were produced at 30 m spatial resolution and averaged to 1 km for the purpose of display. Regional details at 30 m resolution are shown in Figure 8.

The annual, 30 m resolution maps revealed field-level heterogeneity in soybean yields (Figure 8). Large contiguous soybean fields in central Mato Grosso have moderate-to-high yield and small variations between fields (Figure 8a), whereas smaller fragmented fields in Rio Grande do Sul show much larger variations (Figure 8b). Over the past 19 years, soybean yields in central Mato Grosso experienced an overall increase in most fields, whereas in Rio Grande do Sul, larger fields appeared to have relatively greater yield growth than smaller fields (Figure 8b).

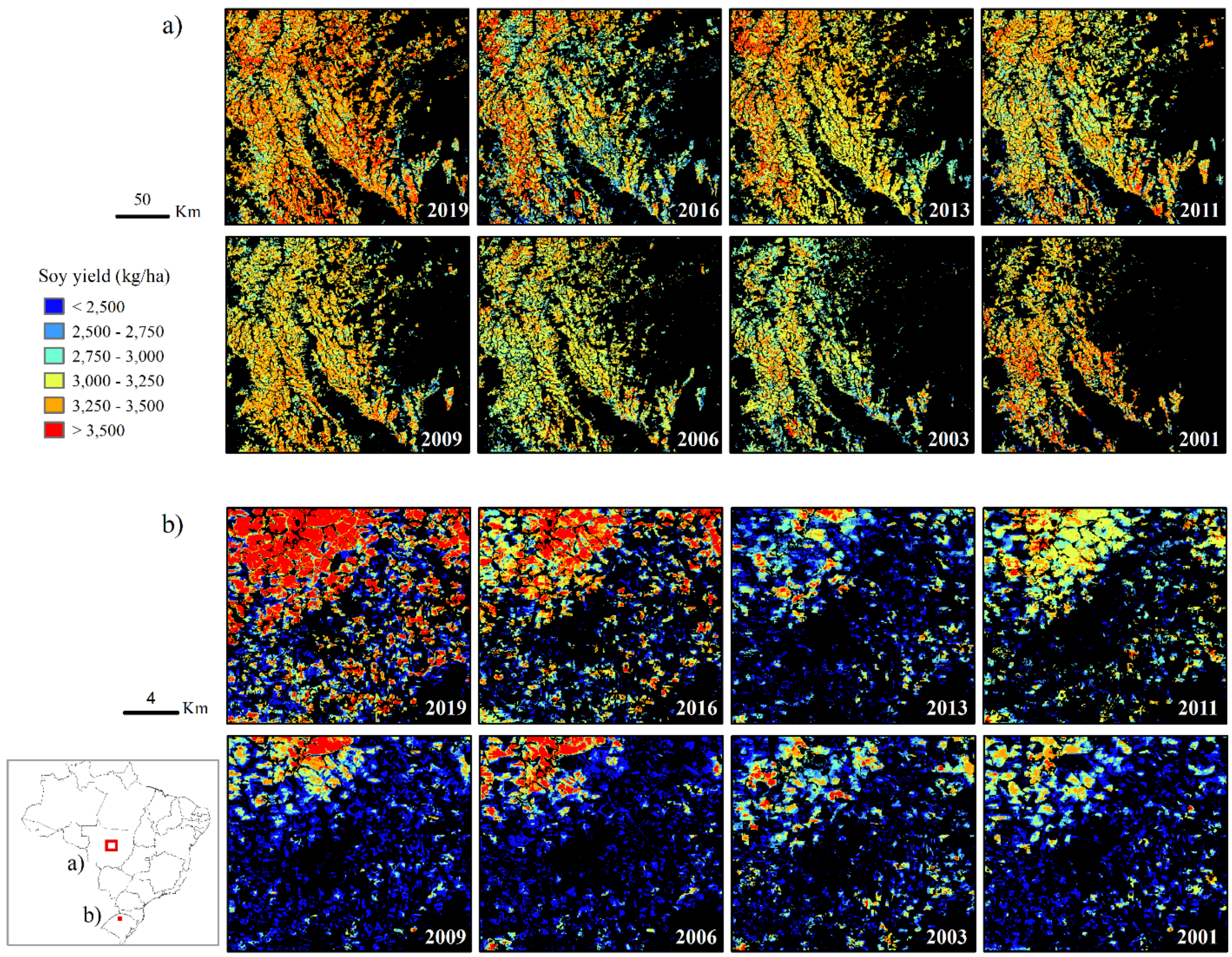

Figure 8. Spatial and temporal details of soybean yield at 30 m resolution in two selected regions: **a**) central Mato Grosso and **b**) northern Rio Grande do Sul. Field-level yield heterogeneity is revealed by the time series of high-resolution maps.

3.3. Map evaluation

The annual 30 m soybean yield maps were aggregated to municipal scale for a quantitative quality assessment. Compared to the reference data from official statistics, the yield map product had an overall RMSE of 418 kg/ha, a MAE of 311 kg/ha, an MBE of 92 kg/ha, and an r^2 of 0.60. Compared to the 2001-2019 national average yield of 2,869 kg/ha, the RMSE represents 15% error. These error metrics were all slightly worse than the model performance, with the RMSE about 20% higher (compared to 344 kg/ha; see detailed numbers of other error metrics in Figure 4). An overall slight positive bias was noted (mean bias of 92 kg/ha or 3% error compared to long-term average yield, Figure 9). Moreover, systematic underestimation was still noticed at the high end of yield and overestimation at the low end of yield (Figure 10), although a linear regression successfully corrected model bias at the training stage at the municipal level (Figure 4). At the annual time scale, the map accuracy was comparable to model performance for the majority of the 19 years (Figure 10). The comparison between model performance and map quality assessment suggested that uncertainties at the 30 m pixel scale were larger than those at the aggregated municipal scale, highlighting a general multi-scale issue in the applications of regression-based machine learning algorithms in remote sensing.

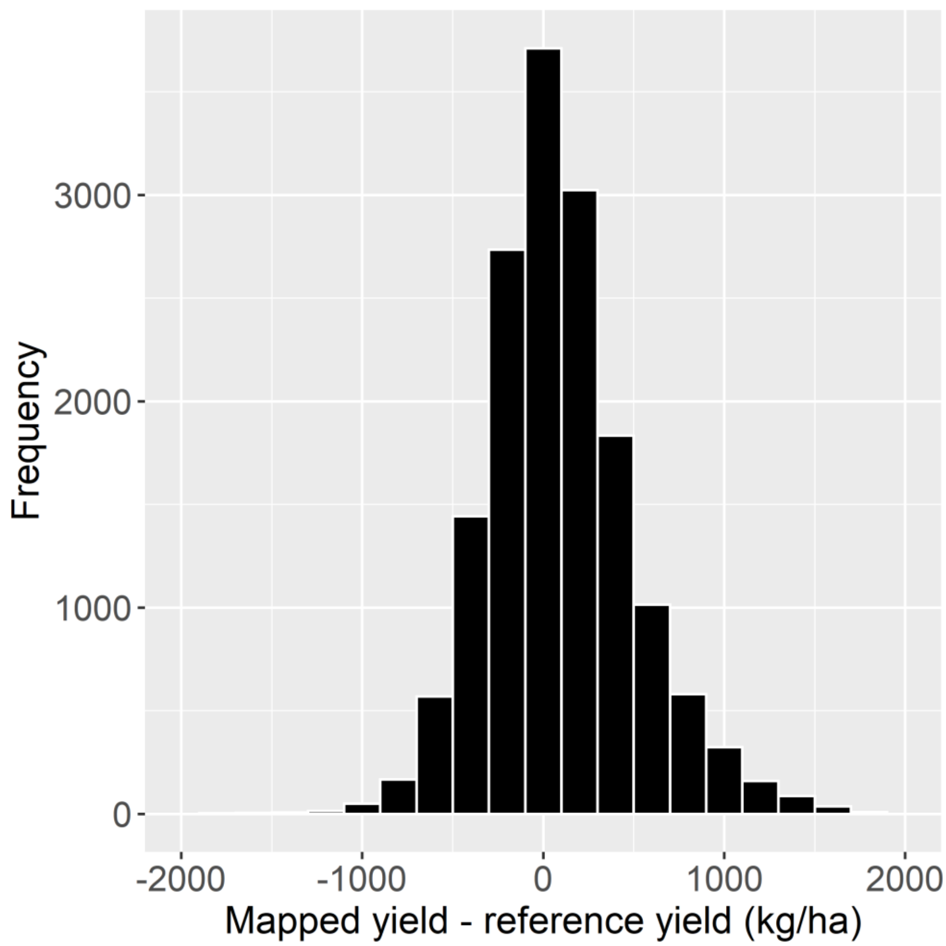

Figure 9. Histogram of the difference between predicted yield and reference yield at the municipal level between 2001 and 2019 (n=15,784) indicating a slight positive bias in the predicted yield.

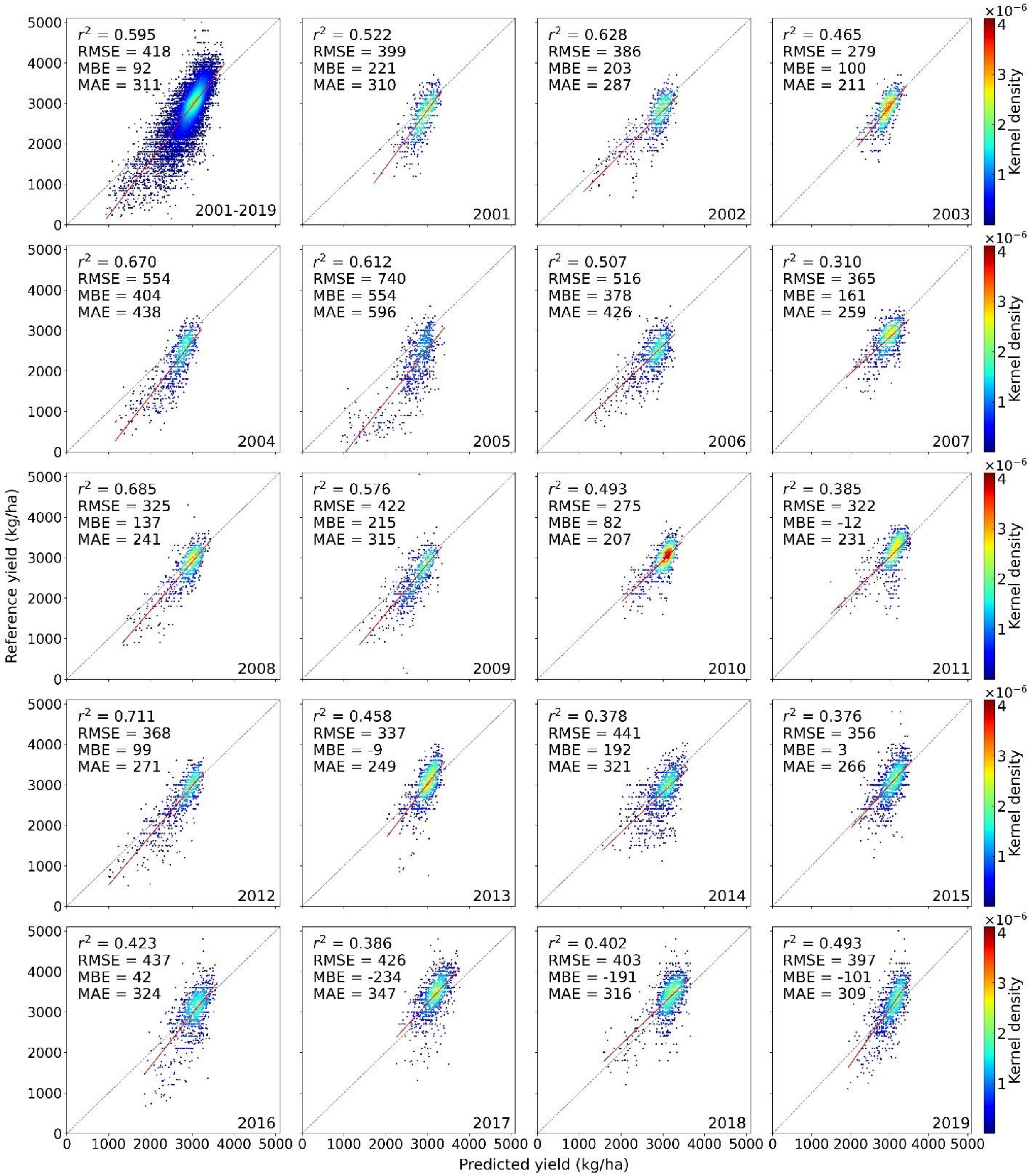

Figure 10. Quality assessment of 30 m soybean yield maps for every year between 2001 and 2019. The annual maps were averaged to the municipal scale to derive predicted yields (x-axis). Reference yields (y-axis) are official statistics.

4. Discussion

4.1. Uncertainty sources for yield modeling

Model performance and the quality of the annual yield maps are influced by a number of factors, including the temporal density of satellite observations, the coarse spatial resolution and uncertainties of climate and weather variables, lack of up-to-date soil measurements, unknown uncertainties in the official statsitics, lack of field-level reference data, missclassifications in the annual soybean masks, and the muti-scale modeling and prediction procedure. The impacts of these factors are discussed in detail as follows.

Depending on the type of cultivar, environmental conditions and agricultural management practices, soybean plants take 90 to 150 days from planting to maturity. During this short growing window, vegetation cover in the field experiences rapid transitions from bare ground to nearly closed canopy and to bare ground again. Such phenological dynamics require dense time-series data to capture the key growth stages that are critical to crop biomass accumulation and yield formation. Studies have demonstrated that the peak growing period in vegetation index is most important for modeling yield for wheat, corn and soybeans (Becker-Reshef et al. 2010; Johnson 2014). In addition, natural disasters during or after the seed-filling stage can cause severe yield reduction (Hosseini et al. 2020). In this study, we used MODIS and Landsat as the main remote sensing data source. Due to the sparse temporal interval of Landsat, cloud-free Landsat observations vary considerably in space and time (Figure 11).

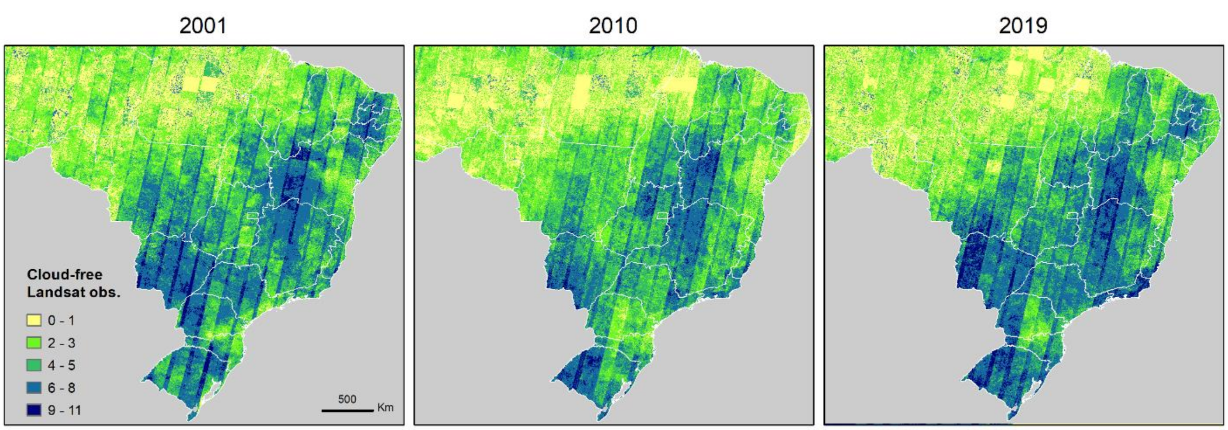

Figure 11. Cloud-free Landsat observations between November 1st and April 30th in selected years over Brazil.

On the other hand, daily MODIS acquisitions are more robust to cloud contamination. Indeed, the important features identified by random forests include many MODIS-based spectral features. The most important feature of the random forests model (model #1) was “M_NDVI_av90max”, which represented the average value of the 90th percentile and maximum NDVI (i.e. peak NDVI) derived from MODIS (Figure 12). The second and third most important features were MODIS-based peak-season NIR reflectance and middle-season NDVI, respectively. These top three features accounted for >40% of cumulative feature importance (Figure 12). Another inherent factor that enabled MODIS to be an efficient sensor for modeling soybean yield is the large field size in Brazil (Fritz et al. 2015). The feature ranking analysis suggested that improving the temporal density of high spatial resolution satellite data, such as the Harmonized Landsat and Sentinel-2 product (Claverie et al. 2018), may improve yield mapping at the field scale. Further research is also needed to investigate the utility of other freely available satellite data, particularly radar data (e.g. Sentinel 1) for yield estimation, as radar data can provide complementary infromation to optical data for crop monitoring (Song et al. 2021b; Veloso et al. 2017) in addition to their all-weather data acquisition.

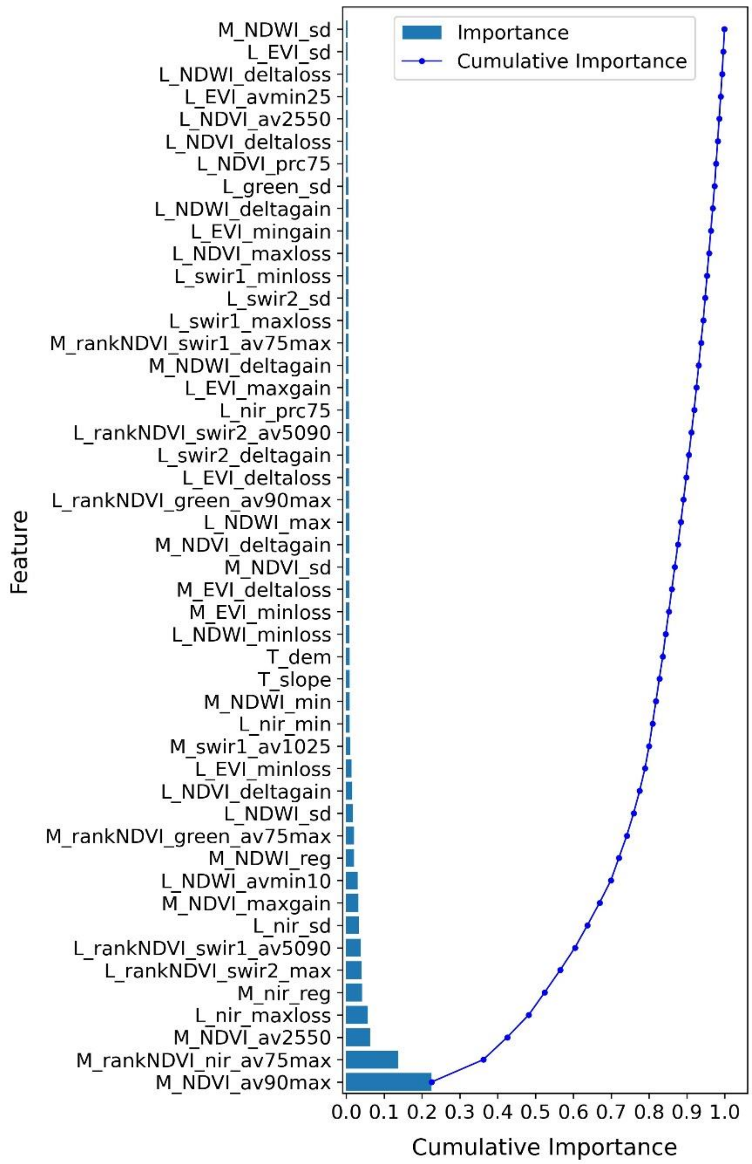

Figure 12. Cumulative feature importance for the random forests-based soybean yield modeling using MODIS and Landsat phenological metrics as input. Features with a prefix of "M*" represents MODIS-based metrics, features with a prefix of "L*" represents Landsat-based metrics, and features with a prefix of "T*" represents topographic variables. "av" stands for "average". The metrics are sorted from high to low along the vertical axis from bottom to top. Please see supplementary Table S1 for more explanation of metric names.

Our study explicitly demonstrated the value of climate and weather data for modeling crop yield. For the trained random forests model with all the features as input, climate and weather variables accounted for 36% of the total feature importance (Table 2). Compared to the models with only remote sensing data as input, adding climate and weather variables reduced RMSE by about 7 to 9%, and the improvement was statistically significant. However, adding coarse-resolution climate and weather variables could also introduce undesirable artifacts. By constructing two models and through per-pixel composition of model outputs, our strategy effectively combined the advantages of the two respective models. For any given year, the primary model (i.e. the one with climate and weather variables as input) was chosen for the majority of soybean growing regions of the country, while the secondary model (i.e. the one without climate and weather variables) was selected only for some clustered regions (Figure 13). This data-driven approach relied on the explicit uncertainty outputs associated with predictions of random forests, and the composited map had minimum uncertainties from the multi-model ensemble. Future research will evaluate the uncertainty of climate and weather variables to yield estimation, and incorporate higher-resolution weather dataset for improved yield estimation, e.g. the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) precipitation data (Funk et al. 2015).

Table 2. Importance of the five categories of input variables in random forests model for soybean yield prediction. Details of the variables are listed in Table 1. The total importance of all variables within each category was calculated and reported.

Category of variables	Importance in random forests model
Landsat-based	0.1883
MODIS-based	0.4371
Climate	0.1037
Weather	0.2539
Topographic	0.0041
Soil	0.0128

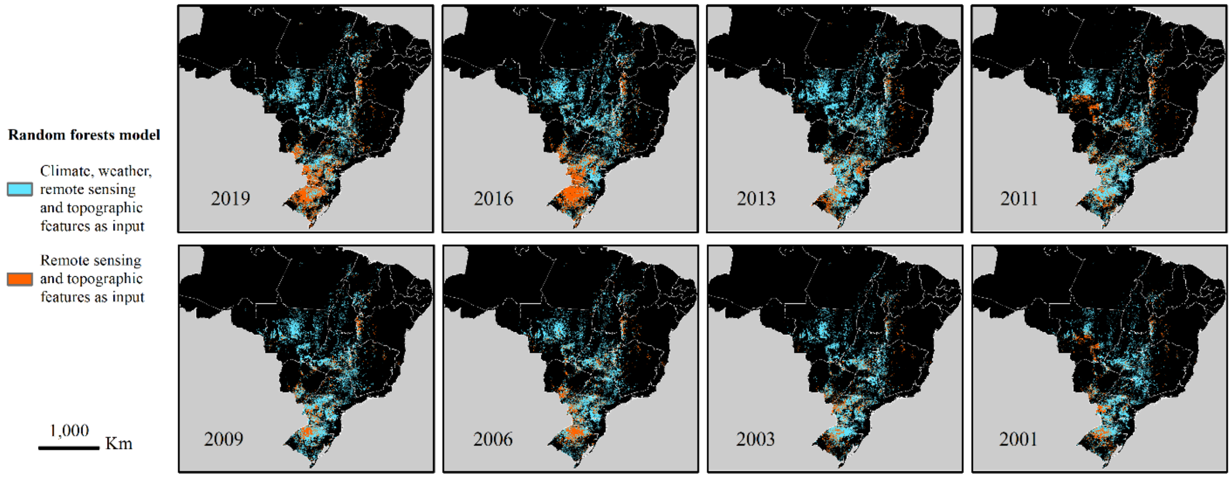

Figure 13. Maps of random forests models chosen for predicting annual soybean yield. The model with climate and weather varaibles as input was more accurate and was used in the majority of the soybean growing regions of the country in every year.

The lack of contribution by soil variables to soybean yield modeling was likely because the soil data were outdated. Soil characteristics and topography are strong determinant of cropland suitability (Ishikawa and Yamazaki 2021). We used the Harmonized World Soil Database (HWSD) in this study, which was complied from multiple data sources (FAO/IIASA/ISRIC/ISSCAS/JRC 2012). The data source for Brazil was the Soil and Terrain database for Latin America and the Caribbean, at the scale of 1:5 million and released in 1998. Therefore, HWSD represents the soil conditions in Brazil before 1998. From 2000 to 2019, soybean cultivation area in Brazil nearly tripled, and new soybean fields were mostly converted from pasture and forests (Song et al. 2021a). The conversion process involves removal of surface vegetation and extraction of the root systems. Subsequently, soil prerparation is critical for cultivating soybeans on the newly converted land. In the Cerrado, the largest soybean growing biome in Brazil, the native soil condition is poor for crop production. Most of the soils in the Cerrado are highly weathered Oxisols and Ultisols, with high acidity and serious definicieny in nutrients (Lopes 1996). Improved management practicies such as liming and fertilizerization have greatly increased soil fertility for growing soybeans (Lopes 1996). These important changes in soil property are not reflected by the HWSD soil database — likely the principal reason why the soil data did not contribute to soybean yield modelilng. Crop modeling studies suggest that soil-related yield variability outweighs the simulated year-to-year variations in yield due to weather when no fertilizer is applied (Folberth et al., 2016). Up-to-date high-quality soil data may improve modeling yiled for soybean and other crops in the tropics where agriculture is expanding (Eigenbrod et al. 2020). Future studies will investigate the utility of higher resolution soil dataset for yield mapping (Hengl et al. 2017). Generating other spatially explicit data on agricultural management that are important for crop production such as seed variety and fertilizer use, is another potential way of improving yield mapping.

Lastly, a common practice in crop yield mapping is to construct a machine learning model at an aggregated spatial scale where public yield statsitics are available, and apply the model to a finer scale at which remote sensing data are acquired (e.g. Johnson 2014). The upscaling process (e.g. spatial aggregation from pixel to municipal) can reduce uncertainties in the original data, as pixel-level errors may be averaged out. Our yield models were calibrated at the municipal scale. More problematic is the downscaling process (i.e. applying the trained model to pixels), as pixel-level errors often exist from e.g. atmospheric correction or misclassification. The discrepency between model performance (Figure 5, overall RMSE 344 kg/ha) and yield map assessment at the same municipality scale (Figure 10, overall RMSE 418 kg/ha) revealed a positive bias in the predicted yield (Figure 9), although the models were unbiased after linear adjustment (Figure 4). This bias was primarily stemmed from the downscaling process, where pixel-level errors couls corrupt the results. Such bias may be removed using field-based yield measurements. However, such datasets are traditionally held by private industry without public access especially over large areas such as the national scale (see Deines et al. (2021) for the case of the United States). Open access to field observations is rare in most parts of the world (Coutu et al. 2020). Increasing the access to historical field observations is a potentially effective way of advancing crop yield research.

4.2. Towards operational yield mapping

Achieving operational yield prediction using satellite data alone is a cost-effective approach of generating timely information on crop production. To demonstrate the predictive capability of our yield models, we applied the models, trained on 2001-2019 data, to 2020 data and produced a 30 m resolution soybean yield map for 2020 (Figure 14). We also collected municipal yield statistics for 2020 and compared with our 2020 yield map. Our random forests models, trained on 2001-2019 data, were able to predict 2020 yield with comparable accuracy as the withheld 2001-2019 test data. The RMSE, MBE and r^2 of the direct output of random forests predictions for 2020 was 555 kg/ha, -145 kg/ha and 0.66, respectively. Consistent with the model performance on 2001-2019 test data, an overall bias was noted. To eliminate this bias, we applied the linear regression approach as reported above. We randomly selected 3% of municipalities (n=34) from the 1,136 municipalities, and constructed a linear regression model using the random forests-predicted yield as the independent variable and the 2020 municipal yield statistics as the dependent variable. After bias correction, the MBE was reduced to -37 kg/ha, and RMSE was reduced to 462 kg/ha (Figure 14b). The RMSE represents 13% error relative to the national average of 3480 kg/ha in 2020. This result suggests that our pre-trained models can be used to generate high-resolution soybean yield maps for future years with the caveat that a small amount of reference data are still needed for the final bias correction. Given the continued operational satellite data acquisitions, including Landsat 8, Landsat 9, MODIS and Visible Infrared Imaging Radiometer Suite (VIIRS), the demonstrated predictive capability of our pre-trained yield models may be used for future yield mapping in a semi-operational mode.

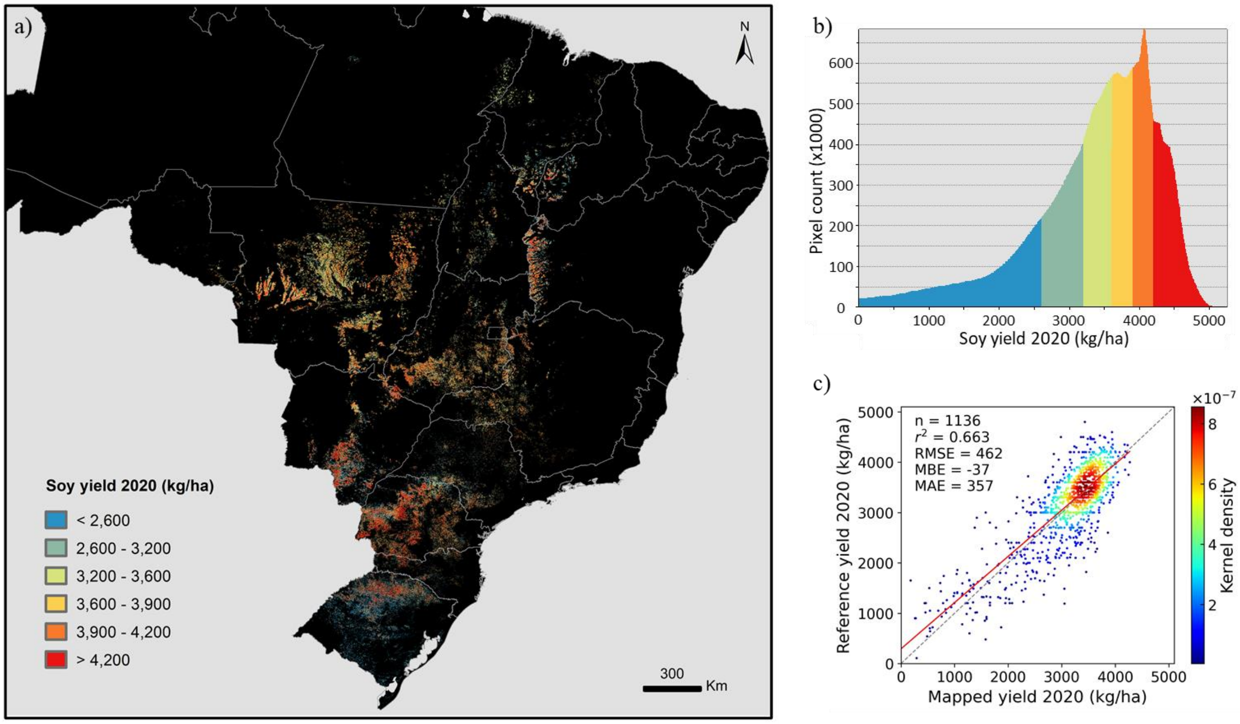

Figure 14. Soybean yield in year 2020 predicted using models trained on 2001-2019 data. a) 30-m map of soy yield 2020. b) Density distribution of the soy yield map. The colors match those shown on the map, and each color corresponds to approximately 1/6 of the total soy pixels. c) Comparison between predicted yield and municipal yield statistics as reference.

The rapidly developing technology of satellite remote sensing is transforming global agriculture. Earth observation data are increasingly used in research and operational settings for mapping crop types, monitoring crop growth, improving agricultural management and forecasting food production. Increasing the comprehensiveness within a single data product, including area, yield, cropping intensity and calendar, at high spatial and temporal resolution has been identified as one of the future research areas in developing global gridded cropping system data product (Kim et al. 2021). We showed in a previous study that satellite data could be used retrospectively mapping soybean over South America since 2001 (Song et al. 2021a). Our 30 m South America soybean map product is being updated at an annual frequency in an operatioanl mode as new satellite data are acquired. This study extends our research from crop type mapping to yield mapping, and we demonstrated that pre-trained machine learning models could be applied for yield mapping in future years. Our current approach for yield mapping and updating uses satellite data of the entire growing season as input. This post-season mapping can generate highly relabile data products, but lacks sufficient timeliness to capture production shocks resulted from e.g. extreme weather events within the growing season. Recent research has demonstrated that early- and in-season crop type mapping and crop yield forecasting could be achieved using advanced machine learning algorithms (e.g. Lin et al. 2022), seasonal climate forecast (Iizumi et al. 2021), and in-season weather observations (Schauberger et al. 2017). Implementing robust in-season forecasting methods in monitoring systems is needed to mitigate the adverse impacts of climate change (Fritz et al. 2019; Kim et al. 2021; Li et al. 2019; Lobell and Burke 2010; Nakalembe et al. 2021).

5. Conclusions

We developed a machine learning-based approach to map annual soybean yield in Brazil over the past two decades. Consistent satellite observations from the open Landsat and MODIS data archives were used to calibrate unbiased yield models using random forests followed by linear regression. Soybean yield maps were generated at 30-meter spatial resolution for every year from 2001 to 2020. NDVI at the peak of the growing season was found to be the most important variable for modeling soybean yield. Our study explicitly demonstrated the utility of climate and weather variables for crop yield estimation. Our multi-scale approach was effective in integrating official yield statistics at political unit level with remote sensing data. Our study demonstrated that models trained on long-term historical data could be employed to predict yield for future years. Our research also highlights that improving the temporal density of high-resolution satellite observations, and enhancing the accessibility to field-level yield measurements are viable ways to improve crop yield mapping over large areas.

Acknowledgements

This research was supported by the start-up funding to X.P.S. from Texas Tech University. We thank Marcos Adami for collecting the yield statistics data used in this study. Satellite data used to create the yield maps were processed using the GLAD ARD tools https://glad.umd.edu/ard/.

References

Battude, M., Al Bitar, A., Morin, D., Cros, J., Huc, M., Marais Sicre, C., Le Dantec, V., & Demarez, V. (2016). Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. *Remote Sensing of Environment, 184*, 668-681

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*, 1312-1323

Benami, E., Jin, Z., Carter, M.R., Ghosh, A., Hijmans, R.J., Hobbs, A., Kenduiywo, B., & Lobell, D.B. (2021). Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth & Environment, 2*, 140-159

Bolton, D.K., & Friedl, M.A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology, 173*, 74-84

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology, 274*, 144-159

Claverie, M., Demarez, V., Duchemin, B., Hagolle, O., Ducrot, D., Marais-Sicre, C., Dejoux, J.-F., Huc, M., Keravec, P., Béziat, P., Fieuzal, R., Ceschia, E., & Dedieu, G. (2012). Maize and sunflower biomass estimation in southwest France using high spatial and temporal resolution remote sensing data. *Remote Sensing of Environment, 124*, 844-857

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.-C., Skakun, S.V., & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment, 219*, 145-161

Coutu, S., Becker-Reshef, I., Whitcraft, A.K., & Justice, C. (2020). Food security: underpin with public and private data sharing. *Nature, 578*, 515

de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., & van Diepen, K. (2019). 25 years of the WOFOST cropping systems model. *Agricultural Systems, 168*, 154-167

de Wit, A., Duveiller, G., & Defourny, P. (2012). Estimating regional winter wheat yield with WOFOST through the assimilation of green area index retrieved from MODIS observations. *Agricultural and Forest Meteorology, 164*, 39-52

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S., & Koetz, B. (2019). Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sensing of Environment, 221*, 551-568

DeFries, R.S., Field, C.B., Fung, I., Justice, C.O., Los, S., Matson, P.A., Matthews, E., Mooney, H.A., Potter, C.S., Prentice, K., Sellers, P.J., Townshend, J.R.G., Tucker, C.J., Ustin, S.L., & Vitousek, P.M. (1995). Mapping the land surface for global atmosphere‐biosphere models: Toward continuous distributions of vegetation's functional properties. *Journal of Geophysical Research: Atmospheres, 100*, 20867-20882.

Deines, J.M., Patel, R., Liang, S.-Z., Dado, W., & Lobell, D.B. (2021). A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. *Remote Sensing of Environment, 253*, 112174

Delécolle, R., Maas, S., Guerif, M., & Baret, F. (1992). Remote sensing and crop production models: present trends. *ISPRS Journal of Photogrammetry and Remote Sensing, 47*, 145-161

Doraiswamy, P.C., Hatfieldb, J.L., Jacksona, T.J., Akhmedova, B., Pruegerb, J., & Sterna, A. (2004). Crop condition and yield simulations using Landsat and MODIS. *Remote Sensing of Environment, 92*, 548-559

Duchemin, B., Maisongrande, P., Boulet, G., & Benhadj, I. (2008). A simple algorithm for yield estimates: Evaluation for semi-arid irrigated winter wheat monitored with green leaf area index. *Environmental Modelling & Software, 23*, 876-892

Eigenbrod, F., Beckmann, M., Dunnett, S., Graham, L., Holland, R.A., Meyfroidt, P., Seppelt, R., Song, X.P., Spake, R., Vaclavik, T., & Verburg, P.H. (2020). Identifying Agricultural Frontiers for Modeling Global Cropland Expansion. *One Earth, 3*, 504-514

FAO (2020). FAOSTAT database. In. Rome: FAO

FAO/IIASA/ISRIC/ISSCAS/JRC (2012). Harmonized World Soil Database (version 1.2). In. FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Folberth, C., Skalsky, R., Moltchanova, E., Balkovic, J., Azevedo, L.B., Obersteiner, M., & van der Velde, M. (2016). Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nat Commun, 7*, 11872

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

Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I., & McCallum, I. (2019). A comparison of global agricultural monitoring systems and current gaps. *Agricultural Systems*, *168*, 258-272

Fritz, S., See, L., McCallum, I., You, L., Bun, A., Moltchanova, E., Duerauer, M., Albrecht, F., Schill, C., Perger, C., Havlik, P., Mosnier, A., Thornton, P., Wood-Sichra, U., Herrero, M., Becker-Reshef, I., Justice, C., Hansen, M., Gong, P., Abdel Aziz, S., Cipriani, A., Cumani, R., Cecchi, G., Conchedda, G., Ferreira, S., Gomez, A., Haffani, M., Kayitakire, F., Malanding, J., Mueller, R., Newby, T., Nonguierma, A., Olusegun, A., Ortner, S., Rajak, D.R., Rocha, J., Schepaschenko, D., Schepaschenko, M., Terekhov, A., Tiangwa, A., Vancutsem, C., Vintrou, E., Wenbin, W., van der Velde, M., Dunwoody, A., Kraxner, F., & Obersteiner, M. (2015). Mapping global cropland and field size. *Glob Chang Biol*, *21*, 1980-1992

Funk, C., & Budde, M.E. (2009). Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. *Remote Sensing of Environment*, *113*, 115-125

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations--a new environmental record for monitoring extremes. *Sci Data*, *2*, 150066

Gallego, F.J. (2004). Remote sensing and land cover area estimation. *International Journal of Remote Sensing*, *25*, 3019-3047

Gao, F., Anderson, M., Daughtry, C., & Johnson, D. (2018). Assessing the Variability of Corn and Soybean Yields in Central Iowa Using High Spatiotemporal Resolution Multi-Satellite Imagery. *Remote Sensing*, *10*, 1489.

Gibbs, H.K., Ruesch, A.S., Achard, F., Clayton, M.K., Holmgren, P., Ramankutty, N., & Foley, J.A. (2010). Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proceedings of the National Academy of Sciences, USA*, *107*, 16732-16737

Gitelson, A.A. (2004). Wide Dynamic Range Vegetation Index for remote quantification of biophysical characteristics of vegetation. *Journal of Plant Physiology*, *161*, 165-173

Gonzáles-Alonso, F., & Cuevas, J.M. (1993). Remote sensing and agricultural statistics: crop area estimation through regression estimators and confusion matrices. *International Journal of Remote Sensing, 14*, 1215-1219

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18-27

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., & Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. *Science, 342,* 850-853.

Harris, I., Osborn, T.J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci Data, 7*, 109

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One, 12*, e0169748

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., Moore, A.D., Brown, H., Whish, J.P.M., Verrall, S., Fainges, J., Bell, L.W., Peake, A.S., Poulton, P.L., Hochman, Z., Thorburn, P.J., Gaydon, D.S., Dalgliesh, N.P., Rodriguez, D., Cox, H., Chapman, S., Doherty, A., Teixeira, E., Sharp, J., Cichota, R., Vogeler, I., Li, F.Y., Wang, E., Hammer, G.L., Robertson, M.J., Dimes, J.P., Whitbread, A.M., Hunt, J., van Rees, H., McClelland, T., Carberry, P.S., Hargreaves, J.N.G., MacLeod, N., McDonald, C., Harsdorf, J., Wedgwood, S., & Keating, B.A. (2014). APSIM – Evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software, 62*, 327-350

Hosseini, M., Kerner, H.R., Sahajpal, R., Puricelli, E., Lu, Y.-H., Lawal, A.F., Humber, M.L., Mitkish, M., Meyer, S., & Becker-Reshef, I. (2020). Evaluating the Impact of the 2020 Iowa Derecho on Corn and Soybean Fields Using Synthetic Aperture Radar. *Remote Sensing, 12*, 3878

Hu, Q., Yin, H., Friedl, M.A., You, L., Li, Z., Tang, H., & Wu, W. (2021). Integrating coarse-resolution images and agricultural statistics to generate sub-pixel crop type maps and reconciled area estimates. *Remote Sensing of Environment, 258*, 112365

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., & Wu, W. (2015). Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. *Agricultural and Forest Meteorology, 204*, 106-121

Huang, X., Schneider, A., & Friedl, M.A. (2016). Mapping sub-pixel urban expansion in China using MODIS and DMSP/OLS nighttime lights. *Remote Sensing of Environment, 175*, 92-108

Iizumi, T., Shin, Y., Choi, J., van der Velde, M., Nisini, L., Kim, W., & Kim, K.-H. (2021). Evaluating the 2019 NARO-APCC Joint Crop Forecasting Service Yield Forecasts for Northern Hemisphere Countries. *Weather and Forecasting, 36*, 879-891

Ines, A.V.M., Das, N.N., Hansen, J.W., & Njoku, E.G. (2013). Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. *Remote Sensing of Environment, 138*, 149-164

Ishikawa, Y., & Yamazaki, D. (2021). Global high-resolution estimation of cropland suitability and its comparative analysis to actual cropland distribution. *Hydrological Research Letters, 15*, 9-15

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., & Wang, J. (2018). A review of data assimilation of remote sensing and crop models. *European Journal of Agronomy, 92*, 141-152

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 Sensing of Environment, 228*, 115-128

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 Sensing of Environment*, *141*, 116-128

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. *Agricultural and Forest Meteorology, 218-219*, 74-84

Jones, J.W., Hoogenboom, G., Porter, C.H., Boote, K.J., Batchelor, W.D., Hunt, L.A., Wilkens, P.W., Singh, U., Gijsman, A.J., & Ritchie, J.T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*, *18*, 235–265

Kang, Y., & Özdoğan, M. (2019). Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sensing of Environment, 228*, 144-163

Kim, K.-H., Doi, Y., Ramankutty, N., & Iizumi, T. (2021). A review of global gridded cropping system data products. *Environmental Research Letters, 16*

King, L., Adusei, B., Stehman, S., Potapov, P.V., Song, X.-P., Krylov, A., Bella, C.D., Loveland, T.R., Johnson, D.M., & Hansen, M.C. (2017). A multi-resolution approach to national-scale cultivated area estimation of soybean. *Remote Sensing of Environment, 195*, 13-29

Li, A., Liang, S., Wang, A., & Qin, J. (2007). Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing*, *73*, 1149-1157

Li, X.-Y., Li, X., Fan, Z., Mi, L., Kandakji, T., Song, Z., Li, D., & Song, X.-P. (2022). Civil war hinders crop production and threatens food security in Syria. *Nature Food*, *3*, 38-46

Li, Y., Guan, K., Schnitkey, G.D., DeLucia, E., & Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. *Glob Chang Biol*, *25*, 2325-2337

Liu, J., Huffman, T., Qian, B., Shang, J., Li, Q., Dong, T., Davidson, A., & Jing, Q. (2020). Crop yield estimation in the Canadian Prairies using Terra/MODIS-derived crop metrics. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 2685-2697

Lobell, D.B. (2013). The use of satellite data for crop yield gap analysis. *Field Crops Research, 143*, 56-64

Lobell, D.B., & Burke, M.B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology, 150*, 1443-1452

Lobell, D.B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment, 164*, 324-333

Lopes, A.S. (1996). Soils under Cerrado: a success story in soil management. *Better crops international, 10*, 10

Massey, R., Sankey, T.T., Congalton, R.G., Yadav, K., Thenkabail, P.S., Ozdogan, M., & Sánchez Meador, A.J. (2017). MODIS phenology-derived, multi-year distribution of conterminous U.S. crop types. *Remote Sensing of Environment, 198*, 490-503

Mateo-Sanchis, A., Piles, M., Munoz-Mari, J., Adsuara, J.E., Perez-Suay, A., & Camps-Valls, G. (2019). Synergistic integration of optical and microwave satellite data for crop yield estimation. *Remote Sens Environ, 234*, 111460

Moulin, S., Bondeau, A., & Delecolle, R. (1998). Combining agricultural crop models and satellite observations: From field to regional scales. *International Journal of Remote Sensing, 19*, 1021-1036

Mulla, D.J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering, 114*, 358-371

Nakalembe, C., Becker-Reshef, I., Bonifacio, R., Hu, G., Humber, M.L., Justice, C.J., Keniston, J., Mwangi, K., Rembold, F., Shukla, S., Urbano, F., Whitcraft, A.K., Li, Y., Zappacosta, M., Jarvis, I., & Sanchez, A. (2021). A review of satellite-based global agricultural monitoring systems available for Africa. *Global Food Security, 29*

Nearing, G.S., Crow, W.T., Thorp, K.R., Moran, M.S., Reichle, R.H., & Gupta, H.V. (2012). Assimilating remote sensing observations of leaf area index and soil moisture for wheat yield estimates: An observing system simulation experiment. *Water Resources Research, 48*, W05525

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I.N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, *187*, 103016

Potapov, P., Hansen, M.C., Kommareddy, I., Kommareddy, A., Turubanova, S., Pickens, A., Adusei, B., Tyukavina, A., & Ying, Q. (2020). Landsat Analysis Ready Data for Global Land Cover and Land Cover Change Mapping. *Remote Sensing, 12*, 426

Potapov, P., Turubanova, S., Hansen, M.C., Tyukavina, A., Zalles, V., Khan, A., Song, X.-P., Pickens, A., Shen, Q., & Cortez, J. (2022). Global maps of cropland extent and change show accelerated cropland expansion in the twenty-first century. *Nature Food, 3*, 19-28

Sakamoto, T., Gitelson, A.A., & Arkebauer, T.J. (2013). MODIS-based corn grain yield estimation model incorporating crop phenology information. *Remote Sensing of Environment, 131*, 215-231

Schauberger, B., Gornott, C., & Wechsung, F. (2017). Global evaluation of a semiempirical model for yield anomalies and application to within-season yield forecasting. *Glob Chang Biol, 23*, 4750-4764

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*, 107886

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S.V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep, 11*, 1606

Skakun, S., Franch, B., Vermote, E., Roger, J.-C., Becker-Reshef, I., Justice, C., & Kussul, N. (2017). Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a Gaussian mixture model. *Remote Sensing of Environment, 195*, 244-258

Skakun, S., Kalecinski, N.I., Brown, M.G.L., Johnson, D.M., Vermote, E.F., Roger, J.-C., & Franch, B. (2021). Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. *Remote Sensing, 13*, 872

Song, X.-P., Hansen, M.C., Potapov, P.V., 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. (2021a). Massive soybean expansion in South America since 2000 and implications for conservation. *Nature Sustainability*, *4*, 784-792

Song, X.-P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F., & Townshend, J.R. (2018). Global land change from 1982 to 2016. *Nature, 560*, 639-643

Song, X.-P., Huang, W., Hansen, M.C., & Potapov, P. (2021b). An evaluation of Landsat, Sentinel-2, Sentinel-1 and MODIS data for crop type mapping. *Science of Remote Sensing*, *3*, 100018

Song, X.-P., Potapov, P.V., Krylov, A., King, L., Di Bella, C.M., Hudson, A., Khan, A., Adusei, B., Stehman, S.V., & Hansen, M.C. (2017). National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. *Remote Sensing of Environment*, *190*, 383-395

Tucker, C.J., Holben, B., Elgin Jr, J., & McMurtrey III, J. (1980). Relationship of spectral data to grain yield variation. *Photogrammetric Engineering and Remote Sensing*, *46*, 657-666

Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., & Ceschia, E. (2017). Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. *Remote Sensing of Environment*, *199*, 415-426

Vermote, E., & Wolfe, R. (2015). MOD09GQ MODIS/Terra Surface Reflectance Daily L2G Global 250m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2021-11-27 from https://doi.org/10.5067/MODIS/MOD09GQ.006

Vroege, W., Vrieling, A., & Finger, R. (2021). Satellite support to insure farmers against extreme droughts. *Nature Food*, *2*, 215-217

Wardlow, B.D., & Egbert, S.L. (2008). Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. *Remote Sensing of Environment*, *112*, 1096-1116

Weiss, M., Jacob, F., & Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. *Remote Sensing of Environment*, *236*

Wieder, W.R., Boehnert, J., Bonan, G.B., & Langseth, M. (2014). Regridded Harmonized World Soil Database v1.2. In: Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA.

Williams, J., Jones, C., Kiniry, J., & Spanel, D.A. (1989). The EPIC crop growth model. *Transactions of the ASAE, 32*, 497-0511

Yang, H.S., Dobermann, A., Lindquist, J.L., Walters, D.T., Arkebauer, T.J., & Cassman, K.G. (2004). Hybrid-maize—a maize simulation model that combines two crop modeling approaches. *Field Crops Research*, *87*, 131-154

Zhang, G., & Lu, Y., (2012). Bias-corrected random forests in regression. *Journal of Applied Statistics, 39*, 151-160

Zalles, V., Hansen, M.C., Potapov, P.V., Parker, D., Stehman, S.V., Pickens, A.H., Parente, L.L., Ferreira, L.G., Song, X.-P., Hernandez-Serna, A., & Kommareddy, I. (2021). Rapid expansion of human impact on natural land in South America since 1985. *Science Advances, 7*, eabg1620

Zalles, V., Hansen, M.C., Potapov, P.V., Stehman, S.V., Tyukavina, A., Pickens, A., Song, X.-P., Adusei, B., Okpa, C., Aguilar, R., John, N., & Chavez, S. (2019). Near doubling of Brazil's intensive row crop area since 2000. *Proceedings of the National Academy of Sciences, USA, 116*, 428-435