Special Submissions: Predictive Agriculture
47,46, 27, 233, 90

Mid-season County-Level Corn Yield Forecast for US Corn Belt Integrating Satellite Imagery and Weather Variables

Rai Schwalbert,* Telmo Amado, Luciana Nieto, Geomar Corassa, Charles Rice, Nahuel Peralta, Bernhard Schauberger, Christoph Gornott, and Ignacio Ciampitti*

R. Schwalbert, T. Amado, G. Corassa, Federal Univ. of Santa Maria, Agricultural Engineering Dep., Santa Maria, Brazil; R. Schwalbert, L. Nieto, G. Corassa, C. Rice, I. Ciampitti, Kansas State Univ., Dep. of Agronomy, Manhattan, Kansas; T. Amado, Federal Univ. of Santa Maria, Soil Science Dep., Santa Maria, Brazil; N. Peralta, Monsanto Argentina, Dep. of Technology and Development of Corn and Sorghum, Buenos Aires, Argentina; B. Schauberger, C. Gornott, Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, 14412 Potsdam, Germany. Received 14 Mar. 2019. Accepted 2 Oct. 2019. *Corresponding author (rai.schwalbert@gmail.com, ciampitti@ksu.edu). Assigned to Associate Editor Carlos Messina.

Abbreviations: CDL, Cropland Data Layer; DOY, day of year; EVI, Enhanced Vegetation Index; GEE, Google Earth Engine; MAE, mean absolute error; MODIS, Moderate Resolution Imaging Spectroradiometer; NASS, National Agricultural Statistic Service; NDVI, Normalized Difference Vegetation Index; NSE, Nash–Sutcliffe model efficiency coefficient; VI, vegetation index; VPD, vapor pressure deficit.

ABSTRACT

Yield estimations are of great interest to support interventions from governmental policies and to increase global food security. This study presents a novel model to perform in-season corn yield predictions at the US-county level, providing robust results under different weather and yield levels. The objectives of this study were to: (i) evaluate the performance of a random forest classification to identify corn fields using Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) and weather variables (temperature, precipitation, and vapor pressure deficit, VPD); (ii) evaluate the contribution of weather variables when forecasting corn yield via remote sensing data, and perform a sensitivity analysis to explore the model performance in different dates; and (iii) develop a model pipeline for performing in-season corn yield predictions at county-scale. Main outcomes from this study were: (i) high accuracy (87% on average) for corn field classification achieved in late August, (ii) corn yield forecasts with a mean absolute error (MAE) of 0.89 Mg ha^{-1}, (iii) weather variables (VPD and temperature) highly influenced the model performance, and (iv) model performance decreased when predictions were performed early in the season (mid-July), with MAE increasing from 0.87–1.36 Mg ha^{-1} when forecast timing changed from day of year 232–192. This research portrays the benefits of integrating statistical techniques and remote sensing to field survey data in order to perform more reliable in-season corn yield forecasts.

Yield forecasts with high accuracy before harvest are extremely useful in agricultural decision-making processes, but their applicability largely depends on the spatial scale in which predictions are performed. On the one hand, within-field yield variability predictions are helpful to understand how crops respond to numerous management and environmental factors (Peralta et al., 2016; Lobell, 2013). On the other hand, yield forecast models at a larger scale (e.g., county, state, and country) are useful for questions involving global food security, government assistance in food policies, and trade of agricultural commodities. Moreover, such forecasts can enable grain traders to make informed decisions, especially in food exporting countries such as the United States (Sakamoto et al., 2014).

Remotely sensed vegetation indices (VIs) such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI) are commonly used for agricultural mapping and yield forecasting (Maselli and Rembold, 2001; Mkhabela et al., 2005; Funk and Budde, 2009). Research has been focused on a wide range of satellite imagery data to predict corn (*Zea mays*) yield worldwide (Bognár et al., 2011; Lobell, 2013; Hamada et al., 2015; Peralta et al., 2016; Schwalbert et al., 2018), demonstrating the potential of the yield forecast models based on remote sensing data, as a tool for providing quantitative and timely information on agricultural crops. Satellite images with greater spatial resolution, such as the freely available Landsat 8 (30 m) and Sentinel 2 (10 m), or commercial options RapidEye (5 m) and Skysat (2 m), are needed for field-level crop monitoring and yield forecasts. Models designed to perform predictions on county, state, or country level are based on images with an intermediate to coarse resolution, such as Advanced Very High Resolution Radiometer (AVHRR; 1km) and Moderate Resolution Imaging Spectroradiometer (MODIS; 250 m). These images have advantages in regard to their superior temporal revisit frequency and larger spatial coverage which avoid problems with cloud interference (Rembold et al., 2013).

There are two equally important characteristics that should be considered in yield forecast models: (i) accuracy of the forecasts and (ii) lead-time of the forecast relative to a decision point. Usually, those two aspects are related. Predictions issued early in the season have lower accuracy than those issued later in the season when more information is available (Hayes and Decker, 1996; Shanahan et al., 2001; Wall et al., 2008). Sakamoto et al. (2014) found high accuracy in US county- and state-level predictions using images from the beginning of the season, with error decreasing as more images were added during the progression of the crop growing season. Early yield predictions are largely dependent on weather events (e.g., heavy precipitation, drought, and heat stresses) and corresponding agronomic management decisions in the remaining growing season. Weather has a large contribution to yield variability within and between seasons, and yield forecast models usually have an improved performance when those variables are taken into account (Johnson, 2014). However, only a few studies have examined the effects of these additional input variables on the model performance (Shao et al., 2015). Precipitation, daily average, maximum and minimum air temperature (based on weather stations), and daytime and nighttime land surface temperature (derived from earth observations) are the most common variables included into crop yield forecast models (Johnson, 2014; Shao et al., 2015). A variable proven useful to explain the strong relationship between yield at regional scale and temperature is the vapor pressure deficit (VPD; Lobell et al., 2013). The VPD is the gradient between the water-vapor-saturated leaf interior and the drier bulk air (Ort and Long, 2014), and is widely used as a measure of atmospheric water demand that depends on air temperature and humidity. It has frequently been reported as one of the most decisive weather variables on historical corn yield anomalies across the US Corn Belt (Lobell et al., 2014).

One of the most critical steps to making in-season yield predictions at large scales is related to obtaining reliable information about geographical distribution of field and crop yields across large areas (Sakamoto et al., 2014; Jin et al., 2017a; Shelestov et al., 2017). For the United States, the National Agricultural Statistical Service (NASS) publishes a layer with detailed information about geographical distribution of fields since 1997 (and since 2008 for the entire country). This information is usually released 3 mo after the harvest of summer crops in the United States. While this information is valuable for training yield models, it is *not* useful for in-season yield forecasts. Recent studies using remote sensing data have focused on exploring techniques aimed at crop classification based on satellite images (Sakamoto et al., 2014; Jin et al., 2017a; Shelestov et al., 2017). This is an essential step toward the development of near real-time forecast models. Classification trees techniques, such as random forest, were proven useful as classification methods to crop mapping, presenting a high computational efficiency, and robustness against overfitting (Belgiu and Drăgut, 2016). Only two parameters, the number of variables in the random subset at each node and the number of trees in the forest, are required to run the algorithm, and the output model is usually not very sensitive to these parameters, which removes any subjectivity from the model process (Liaw and Wiener, 2002).

The objectives of this study were to: (i) evaluate the performance of a random forest classification algorithm to identify pixels where corn is grown at 250 m resolution using NDVI and EVI derived from MODIS images and weather variables; (ii) assess the effect of the weather variables: precipitation, temperature, and VPD, on corn yield predictions, and evaluate the model performance when forecasting corn yield earlier in the season; and (iii) develop a model pipeline to perform corn yield forecast at county level for the US Corn Belt. For our analyses, we select Iowa, Indiana, and Kansas to test the model at varying yield levels.

MATERIALS AND METHODS

Data Sources

Historical county-level corn yield data (2008–2017) was obtained from the USDA-NASS (https://quickstats.nass.usda.gov/, accessed 20 Feb. 2019). This database is released as point information in a county (each point is a yield record from a county in a specific year) without geographical identification such as latitude and longitude.

Vegetation indices from satellite imagery were obtained from MODIS Surface Reflectance products via the Google Earth Engine (GEE) platform (Gorelick et al., 2017). The NASA Earth Observing System Data and Information System (EOSDIS) provided 8- and 16-d imageries on a near real-time basis allowing the retrieval of satellite data with minimal cloud interference. This cloud-freeness is the main reason to choose the EOSDIS data for building the model. From those layers we retrieved NDVI and EVI. The NDVI is a widely used VI, with several applications in agriculture including crop classification in the US Corn Belt (Wardlow and Egbert, 2008), however, its ability to separate corn from soybean (*Glycine max*) has been questioned since those crops have relatively similar NDVI profiles (Shao et al., 2010; Gonzalez-Sanchez et al., 2014), and within-crop variations of season are at least as large as inter-crop differences. Moreover, when corn and soybeans reach their peak growth stage and thus high biomass, NDVI usually saturates, without further deciphering differences in biomass. For that reason, we have included the EVI, since it is more sensitive to capturing variability during high-biomass periods (Zhong et al., 2016), despite its lower image frequency (16 d) compare to the NDVI layers (8 d).

All NDVI images were generated using data from the collection MODIS/006/MOD09Q1. This collection provides images with 250-m resolution, and each MOD09Q1 pixel contains the best possible observation during an 8-d period to minimize problems with cloud interference. All EVI images were obtained from the collection MODIS/006/MOD13Q1 that provides images with 250-meter resolution, and each MOD13Q1 pixel contains the best possible observation during a 16-d period. All the images from these two collections were gathered between 1 May and 20 August from 2008–2017. The starting date was selected based on the corn planting date. The initial date was defined to capture information from the beginning of the crop growing season in the Corn Belt, which is typically from May to October (USDA-NASS), up to the date of the yield forecast. The initial date is similar to the one used by Johnson (2014) for the US Corn Belt. Shanahan et al. (2001) shows that satellite images earlier than 1 May have a weak correlation with the final yield. The final date, 20 August, was chosen to get images which cover the period of the highest reflectance of the corn canopy, and where the selected weather variables have the highest correlation with the corn yield. This period is expected to capture the two most important phenological stages, flowering and grain filling, which are usually in July and August (Johnson, 2014; Lobell et al., 2015; Peng et al., 2018). Despite the use of fixed dates for forecasting yield in a region as large as the US Corn Belt, this simple approach has produced robust results for other related studies (Schlenker and Roberts, 2009; Bolton and Friedl, 2013; Johnson, 2014; Sakamoto et al., 2014, Peng et al., 2018). Because of the high variability related to planting date, comparative relative maturity, and management, fixed dates seems to be the more robust approach.

The Cropland Data Layer (CDL) was used to retrieve information related to corn and noncorn field locations. The CDL is a raster, geo-referenced, 30-m resolution, crop-specific land-cover data layer created annually for the United States using moderate resolution satellite imagery (e.g., Landsat and MODIS) and extensive on-the-ground agricultural measurements. Its accuracy exceeds 90% for crops such as corn and soybean (Johnson and Mueller, 2010). All CDL images from 2008–2017 were obtained from the GEE platform. For this study, the CDL was reprojected to the MODIS sinusoidal projection and up-scaled to 250 m, so that all the pixels from CDL and MODIS match perfectly. When changing the scale from 30 m to 250 m, so that the values of the new pixels were equal to the average from all the smaller pixels partly or entirely overlapping with the new pixel. This process was performed to identify the pure corn pixels.

Three weather variables were selected to be potentially included on the models: daily average temperature, precipitation, and VPD. The first two refer to negative correlations of heat and positive correlations of precipitation with corn yields (Smith, 1914; Wallace, 1920; Bolton and Friedl, 2013; Johnson, 2014). Additionally, the VPD is known for having a strong influence in several processes during crop growth (Messina et al., 2015; Basso and Ritchie, 2018), and can provide important information with potential to improve the model performance in years with large yield anomalies due to weather stress events. All the weather variables were summarized (averaged for temperature and VPD, and summed for precipitation) on an 8-d period to match exactly with the NDVI derived from MODIS.

Temperature and precipitation were obtained from the Parameter-elevation Regressions on Independent Slopes Model (PRISM), and VPD was obtained from the University of Idaho gridded surface meteorological dataset (GRIDMET). Both layers are daily gridded datasets for the conterminous United States, and provide information with a resolution of ~4 km. Thus, those layers were reprojected and down-scaled to be combined with the rest of the collected information.

Data Collection and Organization

Prior to ingesting data from the aforementioned sources, a mask layer was built which contains all the pixels with a high likelihood of overlapping corn fields in any growing season (pixels entirely contained within corn fields). This mask layer was basically a mosaic of all the reprojected CDLs (process described above) from 2008–2017. The function of this mask layer was to reduce the number of nonagricultural pixels in the crop classification step. The inclusion of this layer has significantly decreased the processing time. In addition, the layer increased the model accuracy due to lower variability in the input data. For each year, all pixels that were labeled as corn at least once in a period from 2008–2017 were considered as candidates to overlap corn fields. All the collected information comprises the first step on the model development (Step 1 of Figure 1).

Crop Classification

The second step on the model development was to train a model capable to differentiate corn from noncorn field pixels (Step 2 of Figure 1). This step was necessary because the model would otherwise become largely dependent on the CDL updates commonly released 3 mo after the harvest of US summer crops (early Feb.), thus impeding any near real-time corn yield forecast. The crop classification model was based on the random forest algorithm. Random forest is an ensemble classifier that randomly selects a subset of training samples and variables to produce multiple decision trees. A larger fraction of the entire dataset (usually around two-thirds of the samples) is used to train the trees, and the remaining fraction is used in a cross-validation technique to estimate how well the resulting random forest model performs (Breiman, 2001). This technique has become common in the remote sensing community due to the accuracy of its outcomes (Belgiu and Drãgut, 2016). The algorithm was set to use 600 trees, with a minimum leaf sample size of five, to build the classification tree through the randomForest package (Liaw and Wiener, 2002) by the R program (R Core Team, 2017).

The classification model used crop types as the dependent variable (only two classes were considered: corn, 100% pure corn pixels; and noncorn). Factors such as NDVI, EVI, VPD, temperature, and precipitation were all considered as independent variables. All independent variables were only used up to the forecasting time within the season, therefore different classification models were trained for the different yield forecast dates. This implies that the crop mask can change slightly between different forecasting days. The classification model was run eight times following an assembly approach. In the first round only the multi-temporal VIs were used. In Rounds 2, 3, and 4, the weather variables were included individually into the model, then in Rounds 5, 6, and 7 two weather variables were included in pairs, and finally a full model using all the variables were tested. The efficiency of this step was obtained using a leave-1-yr-out cross-validation (removing 1 yr per round from the model and then using that year as the validation) and calculating the overall accuracy for each validated year. Overall accuracy was computed by dividing the number of correctly classified observations by the total number of observations derived from the confusion matrix. The best model was considered the model with the highest and constant accuracy over the 10 yr, and it was selected to be used on Step 3.

Empirical Relationships between Yield, Vegetation Indices, and Weather

For building the forecast model in Step 3, only the pixels tagged as corn were used, and these corn pixels were averaged to county level to be combined with the yield information from USDA-NASS. A multivariate model was fitted using corn yield as the independent

variable. The dependent variables were added following the same assembly approach used for the classification model (Step 3 of Figure 1). This process was independently repeated for all the yield forecast dates considered in this study.

Model performance was evaluated using a leave-1-yr-out cross-validation approach, and four metrics were used to assess the model accuracy: the mean absolute error (MAE), the root-mean square error (RMSE), the bias coefficient, and the Nash–Sutcliffe model efficiency coefficient (NSE). The MAE represents the average magnitude of the errors, while RMSE is a quadratic scoring rule for the average magnitude of the error, and 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 = MAE, then all the errors are of the same magnitude. Bias computes the average amount by which observed is greater than predicted, if the model is unbiased the index should be close to zero, positive values means that the model is underestimating the observed data, and negative values means that observed values are overestimated. The NSE is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance, and shows how well the prediction fits to the year-to-year yield variability, and its interpretation is analogous to the coefficient of determination (r^2).

Time Series Sensitivity Analysis

After selecting the best models for crop classification and yield forecast, a sensitivity analysis was performed to check how early in the season the forecasting yield model can be implemented and its effect on the overall model performance. For this purpose, data collected later during the growing season were subsequently removed from the model and the same validation approach aforementioned was used to compute using MAE, RMSE, bias, and NSE. Thus, we tested the model using data until day of year (DOY) 232 (August 20), DOY 224 (August 12), DOY 216 (August 4), DOY 208 (July 27), DOY 200 (July 19), and DOY 192 (July 11). 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, 5 d (Sakamoto et al., 2014), and a processing time for the proposed algorithm under parallel processing to be 1 d, totalizing a delay of 6 d.

All the data collection and organization was performed on the GEE platform. The analysis comprised in Steps 2 and 3 were performed in the R environment in the Beocat, the High Performance Computer from Kansas State University, under parallel processing.

RESULTS

Crop Classification

The importance of the weather on the crop classification was assessed using an assembly approach where the weather variables were independently added to the model. There was no significant improvement on the model performance after the inclusion of precipitation, temperature, or VPD, compared with the model using only NDVI and EVI. Accuracy of the model presented some degree of variability over the growing seasons and among the states, with Iowa and Kansas having higher accuracy relative to Indiana (Figure 2). Additionally, 2012 had the lowest accuracy related to the other years considered on this study (lowest points in Figure 2A). Overall, average for all the years on the three states, the model presented an accuracy of 87% for DOY 232 (Figure 2), 87% for DOY 224, 86.5% for DOY

216, 86% for DOY 208, 84.6% for DOY 200, and 82% for DOY 192 (data not shown) related to the CDL.

Empirical Relationships between Yield, Vegetation Indices, and Weather

There was a significant difference on the model performance driven by the inclusion of weather variables into the model. The simplest model, based only on NDVI and EVI resulted in the highest MAE (1.33 Mg ha^{-1}) and RMSE (1.04 Mg ha^{-1}), the lowest NSE (0.7), and the most negative bias (–77 kg ha^{-1}), indicating that this model presented a large dispersion of points along the 1:1 line, and tended to overestimate the observed yield in a higher proportion compared with the other models. Following a hierarchical order of importance, the weather variable with the highest contribution on enhancing the model performance was VPD, followed by temperature, and then accumulated precipitation. When the weather variables were included in pairs, the combination of precipitation + temperature did not yield better results than the model that only included VPD as the weather variable. The remaining three models had a better performance relative to the previous ones, with the model containing NDVI, EVI, temperature, and VPD having a similar performance to the full model and slightly better than the model including precipitation instead of temperature (Figure 3A). For that reason we selected the model including temperature and VPD as the weather variables, additional to the VI predictors for the last step. The inclusion of temperature and VPD resulted in a decrease of 0.15 Mg ha^{-1} in MAE, 0.18 Mg ha^{-1} in RMSE, 68 kg ha^{-1} in bias, and an increase of 0.07 in the NSE, related to the model that only included NDVI and EVI (Figure 3B). The model performance was quite stable over the years. The 2012 growing season presented the lowest model performance, mainly related to the dry conditions during this growing season (2012 had the lowest yield average among all the years considered in this study). On the other hand, 2009, 2011, and 2016 presented the better model performance according to the RMSE, MSE, NSE, and Bias metrics (Supplemental Figure 1).

Sensitivity of the Results to Forecasting Time

The accuracy of the model decreased as the county-level corn yield forecast was anticipated from the DOY 232 (August 20) to DOY 196 (July 11; Figure 4A). Mean absolute error and RMSE increased, and NSE decreased as the yield forecast was performed earlier in the season. Model performance was most affected when the predictions were performed before DOY 208 (July 27), with MAE surpassing 1 Mg ha^{-1} (Figure 4B).

Another negative effect of performing yield forecast earlier in the season was the trend to overestimate yields in a higher frequency, evidenced by the decreased (more negative values) in the bias coefficient as the predictions are performed towards the beginning of the growing season. This behavior was more evident for the lowest yields on DOY 200 (July 19) and 196 (July 11).

DISCUSSION

This study offered a novel approach using the CDL as a ground truth layer for crop classification, and remote sensing combined with weather data as predictors for estimating corn yield at the county-scale. Moreover, it provides information related to the effect on model accuracy by anticipating corn yield forecast earlier in the season relative to the projection by USDA-NASS. This study suggests that at DOY 208 (July 27) models aiming at forecasting corn yield in the US Corn Belt could be implemented with an error (MAE) lower than 1 Mg ha^{-1}.

One of the main challenges for building accurate yield forecast models is to determine the geographical distribution of the fields, herein after termed as crop mapping layer. This information is valuable since all the pixels not corresponding to corn fields should be removed, or "masked," from the image, before establishing empirical relationships between VI and yield. Different techniques focused on crop classification and crop masking have been proposed worldwide, exploring differences in emergence dates between corn and soybean (Sakamoto et al., 2014), using different machine learning algorithms (e.g., supported vector machine, decision trees, and neural networks; Shelestov et al., 2017), and exploring different resolution satellites options, such as Landsat 8, Sentinel 2, and RapidEye (Azzari et al., 2017; Jin et al., 2017a; Xiong et al., 2017). In the United States, the CDL is also an interesting option for masking crops pixels from satellite scenes due its very high accuracy exceeding 90% for corn and soybean (Johnson and Mueller, 2010). However this layer is not useful for near real-time yield forecast, since it is released with 1 yr of delay. Despite this, the CDL has great value as source of labeled data for training crop classification models. As the first outcome, this study presented a novel crop classification approach based on Random Forest and multi-temporal NDVI and EVI images from early May (DOY 128) to late August (DOY 232), and CDL from previous years as ground-truth for field geographical positions. This model achieved an accuracy of 87% (DOY 232) on average, higher than the 85% threshold, desired for most of agricultural purposes (Shelestov et al., 2017). Similar values of accuracy were reported in previous studies using MODIS images and different approaches for classifying the fields (Sakamoto et al., 2014). The crop classification model makes the yield forecast not dependent on the CDL updates and allows near real-time yield predictions.

The second outcome from this research was the development of a model to forecast corn yield at DOY 232 at county-level. The USDA-NASS usually releases the first corn yield report approximately at the 224th DOY, but on a state-level, county-level yield information is only released in the following year (usually 3 mo after harvest). Satellite imagery data are known to be a useful and a reliable information to forecast yield before harvest time (Bolton and Friedl, 2013; Sakamoto et al., 2014; Johnson, 2014; Peralta et al., 2016; Jin et al., 2017a, b), and the simplest approach to estimate crop yields by establishing empirical relationships between end-season yield observations and mid-season VIs calculated from multispectral images (Moriondo et al., 2007; Wall et al., 2008; Bognár et al., 2011; Minuzzi and Lopes, 2015; Shao et al., 2015; Hamada et al., 2015; Peralta et al., 2016; Bu et al., 2017). The model based on multi-temporal VIs was able to predict yield with a MAE of 1.04 Mg ha^{-1} (DOY 232) over ~2,500 combinations of county-years, and its model performance significantly improved with the introduction of temperature and VPD as predictors, reaching a MAE of 0.89 Mg ha^{-1} (DOY 232). Information related to inclusion of weather variables on empirical yield forecast models are still scarce in the literature. Johnson (2014) found a negative correlation between daytime surface temperature and corn yields, but a lack of improvement on the model performance was documented by including nighttime surface temperature or precipitation. Additionally, Shao et al. (2015) did not find any benefits by including precipitation and average daily maximum and minimum air temperature into the model. However, our study is one of the few studies evaluating the performance of VPD along with multi-temporal VIs as predictors for estimating corn yield at county-scale. Lobell et al. (2014) reported VPD in the third month after sowing, which is typically July for a field sown in early May, as the most influencing variable among 19 other weather variables explaining historical yield variations across the US Corn Belt. Vapor pressure deficit is a widely used measure of atmospheric water demand. It is closely related to crop evapotranspiration and consequently has major effects on crop growth and yields. It has been documented that the

photosynthetic rate declines when atmospheric VPD increases (Quick et al., 1992; Hirasawa and Hsiao, 1999; Fletcher et al., 2007). This is because plants under high VPD conditions reduce stomatal conductance which effectively saves water in the plant, at the cost of reduced carbon assimilation (Lobell et al., 2013).

Lastly, a sensitivity analysis was pursued to explore how early reliable county-level corn yield predictions can be accomplished. The importance of a yield prediction could be considered as a balance between its accuracy and the timing when the prediction are performed, considering that 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). Our results showed that corn yield can be forecasted at county-scale for the US Corn Belt at DOY 208 with a MAE <1 Mg ha^{1}, and a RMSE of 1.26 Mg ha^{-1}.

Equal RMSE was reported by Johnson (2014) when forecasting yield at DOY 305 using the CDL as the crop mask. Furthermore, the RMSEs reported in this study are within the range of the values reported by Shao et al. (2015), and below the ones reported by Sakamoto et al. (2014), ranging from approximately 1.68–1.76 Mg ha^{-1} at DOY 215, when performing prediction independently from CDL for 2002 and 2012. Therefore, these results represent a great prospect for anticipating the yield forecast in approximately 2 wk related to the first USDA-NASS yield report at state level.

Despite the model pipeline presented in this study being dependent only on remote sensing and weather data to forecast corn yield at the county-level, the layers used to train the crop classification model and to establish the yield-VI+weather empirical relationships came from extensive field surveys performed by USDA-NASS or analogous agencies. Therefore, the contribution of the research is to show the potential benefits of integrating statistical techniques and remote sensing data to standard approaches (field survey) to perform more reliable in-season yield forecasts. Moreover, it is worth acknowledging that the model developed in this study presents limitations that can be overcome in future studies. The first constraint is related to the resolution, since the MODIS pixel size is 250 m. Thus, fields below that resolution are blended with other fields and may therefore be inaccurately treated in the analysis. The second constraint is related to the model dependence on field survey data, since this study was developed for the United States, the crop classification model was trained using the CDL. For countries where this type of information is not yet available, extensive field surveys will be required to achieve high accuracy in the crop classification step. The model performance could still be enhanced by (i) adding phenology information during the crop classification process (Bolton and Friedl, 2013), (ii) exploring new sources of information combining better spatial and temporal resolutions, such as Sentinel-2, RapidEye, and Skysat, (iii) exploring new direct indicators of photosynthesis (such as solar-induced florescence) that will be available in a near future (Drusch et al., 2017), (iv) adding management information into the model scope such as selection of crop varieties, fertilizer, plant density, comparative relative maturity, or irrigation, and (v) combining remote sensing information and crop models (i.e., mechanistic- or process-based models) output to enhance predictability power and increase the spatiotemporal limits of predictability. As summarized in Sibley et al. (2014), there are at least two approaches for combining these two sources of information for forecasting crop yields. The first one is to use crop simulation models to forecast crop yields, with the remote sensing data employed to adjust inputs or parameters for the model on a pixel-by-pixel basis (Clevers, 1997; Doraiswamy et al., 2005; Doraiswamy et al., 2005; Launay and Guerif, 2005; Dente et al., 2008). The second approach is to use crop models for training empirical models under a larger variety of weather, soil, and management conditions, and access the crop yield through the empirical coefficients (Sibley et al., 2014,

Lobell et al., 2015; Azzari et al., 2017; Jin et al., 2017a, b). Both approaches result in models less dependent on third-party data such as the USDA-NASS, and more robust against weather anomalies such as the 2012 growing season. Results from this study suggest that remote sensing and weather variables (temperature and VPD) are valuable data sources to perform accurate near real-time county-level corn yield predictions even early in the season (late July), with the potential to enhance and help anticipate yield predictions from official government departments such as the USDA-NASS. Despite the fact that only three states were considered in this study (Iowa, Indiana, and Kansas), we tested the model for additional random combinations of counties (from different states) and years, and the estimated error was within the range reported in the result section.

CONCLUSIONS

Multi-temporal satellite imagery combined with weather variables can provide useful information allowing the development of models which are able to forecast and monitor corn yield at early season (after flowering) at county-scale. A decrease in accuracy is expected by anticipating the yield predictions, but this study suggests that corn yield forecast based on satellite imagery, temperature, and VPD could be implemented at 208 DOY (July 27) with an accuracy of 78%. This is ~16 d before the first corn yield report of the USDA-NASS (at state level), and approximately 122 d before the harvest. Additionally, the novel crop classification model developed in this study using the Random Forest classification technique was adequate to separate pixels from MODIS images between corn and noncorn fields with an overall accuracy higher than 85%.

The training and validation approach used in this study with data from different states and years was adequate to test the model performance in different weather and yield conditions. Despite the analysis being developed for the United States, the general approach described can potentially be applied to other regions around the globe, if a reasonable amount of survey data is available for building a solid crop mapping data layer. This could contribute to support agricultural decisions in regards to managing and transferring risks within the crop production. This can help farmers plan interventions and enable governments and traders to adjust trading schemes and thus, avoid yield failures and food shortages.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

Rai A. Schwalbert led the statistical analysis and evaluation of the crop classification and forecasting yield model, and wrote the paper. Telmo J. C. Amado, Luciana Nieto, Geomar Corassa, and Charles Rice contributed to the data discussion. Nahuel Peralta, Bernhard Schauberger, and Christoph Gornott contributed to the data analysis, discussion, and writing of the paper. Ignacio Ciampitti led the study and contributed to the data analysis, discussion, and writing of the paper.

ACKNOWLEDGMENTS

This study was supported by CAPES Foundation, Ministry of Education of Brazil, Brasilia DF, Zip Code 70.040-020, Aquarius project, and Kansas Corn Commission. This is contribution no. 20-085-J from the Kansas Agricultural Experiment Station and process 88887.130848/2016-00 from CAPES.

REFERENCES

Azzari, G., M. Jain, and D.B. Lobell. 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sens. Environ. 202:129–141. doi:10.1016/j.rse.2017.04.014

Basso, B., and J.T. Ritchie. 2018. Evapotranspiration in high-yielding maize and under increased vapor pressure deficit in the US Midwest. Arig. Environ. Lett. 3(1):170039. doi:10.2134/ael2017.11.0039.

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

Bognár, P., C. Ferencz, S. Pásztor, G. Molnár, G. Timár, D. Hamar, J. Lichtenberger, B. Székely, P. Steinbach, and O.E. Ferencz. 2011. Yield forecasting for wheat and corn in Hungary by satellite remote sensing. Int. J. Remote Sens. 32(17):4759–4767. doi:10.1080/01431161.2010.493566

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

Breiman, L. 2001. Random Forests. Mach. Learn. 45:5–32. doi:10.1023/A:1010933404324

Bu, H., L.K. Sharma, A. Denton, and D.W. Franzen. 2017. Comparison of satellite imagery and ground-based active optical sensors as yield predictors in sugar beet, spring wheat, corn, and sunflower. Agron. J. 109(1):299–308. doi:10.2134/agronj2016.03.0150

Clevers, J.G.P.W. 1997. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. Remote Sens. Environ. 61:221–228. doi:10.1016/S0034-4257(97)00004-7

Dente, L., G. Satalino, F. Mattia, and M. Rinaldi. 2008. Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. Remote Sens. Environ. 112:1395–1407. doi:10.1016/j.rse.2007.05.023

Doraiswamy, P.C., T.R. Sinclair, S. Hollinger, B. Akhmedov, A. Stern, and J. Prueger. 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sens. Environ. 97:192–202. doi:10.1016/j.rse.2005.03.015

Drusch, M., J. Moreno, U. Del Bello, R. Franco, Y. Goulas, et al. 2017. The FLuorescence EXplorer Mission Concept—ESA's Earth Explorer 8. IEEE Trans. Geosci. Remote Sens. 55(3):1273–1284. doi:10.1109/TGRS.2016.2621820

Fletcher, A.L., T.R. Sinclair, and L.H. Allen, Jr. 2007. Transpiration responses to vapor pressure deficit in well watered "slow-wilting" and commercial soybean. Environ. Exp. Bot. 61:145–151. doi:10.1016/j.envexpbot.2007.05.004

Funk, C., and M.E. Budde. 2009. Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. Remote Sens. Environ. 113(1):115–125. doi:10.1016/j.rse.2008.08.015

Gonzalez-Sanchez, A., J. Frausto-Solis, and W. Ojeda-Bustamante. 2014. Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. Sci. World J. 2014:509429. doi:10.1155/2014/509429.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202:18–27. doi:10.1016/j.rse.2017.06.031

Hamada, Y., H. Ssegane, and M.C. Negri. 2015. Mapping intra-field yield variation using high resolution satellite imagery to integrate bioenergy and environmental stewardship in an agricultural watershed. Remote Sens. 7(8):9753–9768. doi:10.3390/rs70809753

Hayes, M.J., and W.L. Decker. 1996. Using NOAA AVHRR data to estimate maize production in the United States Corn Belt. Int. J. Remote Sens. 17(16):3189–3200. doi:10.1080/01431169608949138

Hirasawa, T., and T.C. Hsiao. 1999. Some characteristics of re- duced leaf photosynthesis at midday in maize growing in the field. Field Crops Res. 62:53–62. doi:10.1016/S0378-4290(99)00005-2

Jin, Z., G. Azzari, M. Burke, S. Aston, and D. Lobell. 2017a. Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa. Remote Sens. 9(9):931. doi:10.3390/rs9090931

Jin, Z., G. Azzari, and D.B. Lobell. 2017b. Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches. Agric. For. Meteorol. 247:207–220. doi:10.1016/j.agrformet.2017.08.001

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. doi:10.1016/j.rse.2013.10.027

Johnson, D.M., and R. Mueller. 2010. The 2009 Cropland Data Layer. Photogramm. Eng. Remote Sens. 76(11):1201–1205.

Launay, M., and M. Guerif. 2005. Assimilating remote sensing data into a crop model to improve predictive performance for spatial applications. Agric. Ecosyst. Environ. 111:321–339. doi:10.1016/j.agee.2005.06.005

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2/3:18–22.

Lobell, D.B., D. Thau, C. Seifert, E. Engle, and B. Little. 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164:324–333. doi:10.1016/j.rse.2015.04.021

Lobell, D.B., G.L. Hammer, G. McLean, C. Messina, M.J. Roberts, and W. Schlenker. 2013. The critical role of extreme heat for maize production in the United States. Nat. Clim. Chang. 3:497–501. doi:10.1038/nclimate1832

Lobell, D.B. 2013. The use of satellite data for crop yield gap analysis. Field Crops Res. 143:56–64. doi:10.1016/j.fcr.2012.08.008

Lobell, D.B., M.J. Roberts, W. Schlenker, N. Braun, B.B. Little, R.M. Rejesus, and G.L. Hammer. 2014. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. Science 344(6183):516–519. doi:10.1126/science.1251423.

Maselli, F., and F. Rembold. 2001. Analysis of GAC NDVI data for cropland identification and yield forecasting in Mediterranean African countries. Photogramm. Eng. Remote Sens. 67(5):593–602.

Messina, C.D., T.R. Sinclair, G.L. Hammer, D. Curan, J. Thompson, Z. Oler, C. Gho, and M. Cooper. 2015. Limited-transpiration trait may increase maize drought tolerance in the US Corn Belt. Agron. J. 107(6):1978–1986. doi:10.2134/agronj15.0016

Minuzzi, R.B., and F.Z. Lopes. 2015. Desempenho agronômico do milho em diferentes cenários climáticos no Centro-Oeste do Brasil. Rev. Bras. Eng. Agric. Ambient. 19(8):734–740. doi:10.1590/1807-1929/agriambi.v19n8p734-740

Mkhabela, M.S., M.S. Mkhabela, and N.N. Mashinini. 2005. Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. Agric. For. Meteorol. 129(1–2):1–9. doi:10.1016/j.agrformet.2004.12.006

Moriondo, M., F. Maselli, and M. Bindi. 2007. A simple model of regional wheat yield based on NDVI data. Eur. J. Agron. 26(3):266–274. doi:10.1016/j.eja.2006.10.007

Ort, D.R., and S.P. Long. 2014. Limits on yields in the Corn Belt. Science 344(6183):484–485. doi:10.1126/science.1253884.

Peng, B., K. Guan, M. Pan, and Y. Li. 2018. Benefits of seasonal climate prediction and satellite data for forecasting U.S. maize yield. Geophys. Res. Lett. 45:9662–9671. doi:10.1029/2018GL079291

Peralta, N., Y. Assefa, J. Du, C. Barden, and I. Ciampitti. 2016. Mid-season high-resolution satellite imagery for forecasting site-specific corn yield. Remote Sens. 8(10):848. doi:10.3390/rs8100848

Quick, W., M. Chaves, R. Wendler, M. David, M.L. Rodrigues, J.A. Passaharinho, J.S. Pereira, M.D. Adcock, R.C. Leegood, and M. Stitt. 1992. The effect of wa- ter stress on photosynthetic carbon metabolism in four spe- cies grown under field conditions. Plant Cell Environ. 15:25–35. doi:10.1111/j.1365-3040.1992.tb01455.x

R Core Team. 2017. R: A Language and Environment for Statistical Computing.

Rembold, F., C. Atzberger, I. Savin, and O. Rojas. 2013. Using low resolution satellite imagery for yield prediction and yield anomaly detection. Remote Sens. 5(4):1704–1733. doi:10.3390/rs5041704

Sakamoto, T., A.A. Gitelson, and T.J. Arkebauer. 2014. Near real-time prediction of U.S. corn yields based on time-series MODIS data. Remote Sens. Environ. 147:219–231. doi:10.1016/j.rse.2014.03.008

Schlenker, W., and M.J. Roberts. 2009. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proc. Natl. Acad. Sci. USA 106(37):15594–15598. doi:10.1073/pnas.0906865106

Schwalbert, R.A., T.J.C. Amado, L. Nieto, S. Varela, G.M. Corassa, T.A.N. Horbe, C.W. Rice, N.R. Peralta, and I.A. Ciampitti. 2018. Forecasting maize yield at field scale based on high-resolution satellite imagery. Biosyst. Eng. 171:179–192. doi:10.1016/j.biosystemseng.2018.04.020

Shanahan, J.F., J.S. Schepers, D.D. Francis, G.E. Varvel, W.W. Wilhelm, J.M. Tringe, M.R. Schlemmer, and D.J. Major. 2001. Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. Agron. J. 93:583–589. doi:10.2134/agronj2001.933583x

Shao, Y., R.S. Lunetta, J. Ediriwickrema, and J. Iiames. 2010. Mapping Cropland and Major Crop Types across the Great Lakes Basin using MODIS-NDVI Data. Photogramm. Eng. Remote Sensing 75(1):73–84. doi:10.14358/PERS.76.1.73

Shao, Y., J.B. Campbell, G.N. Taff, and B. Zheng. 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. doi:10.1016/j.jag.2014.12.017

Shelestov, A., M. Lavreniuk, N. Kussul, A. Novikov, and S. Skakun. 2017. Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. Front. Earth Sci. 5:17. doi:10.3389/feart.2017.00017

Sibley, A.M., P. Grassini, N.E. Thomas, K.G. Cassman, and D.B. Lobell. 2014. Testing remote sensing approaches for assessing yield variability among maize fields. Agron. J. 106(1):24–32. doi:10.2134/agronj2013.0314

Smith, J.W. 1914. The effect of weather upon the yield of corn. Mon. Weather Rev. 42(2):78–92. doi:10.1175/1520-0493(1914)42%3C78:TEOWUT%3E2.0.CO;2

Wall, L., D. Larocque, and P.-M. Léger. 2008. The early explanatory power of NDVI in crop yield modelling. Int. J. Remote Sens. 29(8):2211–2225. doi:10.1080/01431160701395252

Wallace, H.A. 1920. Mathematical inquiry into the effect ofweather on corn yield in the eight Corn Belt states. Mon. Weather Rev. 48:439–446. doi:10.1175/1520-0493(1920)48<439:MIITEO>2.0.CO;2

Wardlow, B.D., and S.L. Egbert. 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: An assessment for the U.S. Central Great Plains. Remote Sens. Environ. 112(3):1096–1116. doi:10.1016/j.rse.2007.07.019

Xiong, J., P.S. Thenkabail, J.C. Tilton, M.K. Gumma, P. Teluguntla, A. Oliphant, R.G. Congalton, K. Yadav, and N. Gorelick. 2017. Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on google earth engine. Remote Sens. 9(10):1065. doi:10.3390/rs9101065

Zhong, L., L. Yu, X. Li, L. Hu, and P. Gong. 2016. Rapid corn and soybean mapping in US Corn Belt and neighboring areas. Sci. Rep. 6(1):36240. doi:10.1038/srep36240

FIGURE 1. Flowchart indicating all steps of the (A) model development: Step 1, data collection and organization; Step 2, variable selection and crop classification model aiming at separating corn fields from non-corn fields using random forest algorithm (crop labels is referred to the labels used in the supervised classifications, random forest); Step 3, data selection and yield forecast model based on empirical relationships; and (B) pipeline with all steps for applying the model in future conditions. NASS, National Agricultural Statistics Service; MODIS, Moderate Resolution Imaging Spectroradiometer; PRISM, parameter-evaluation regression on independent slopes model; NDVI, Normalized Difference Vegetation Index; EVI, Enhanced Vegetation Index; USDA, United States Department of Agriculture; Prec., precipitation; Temp., temperature; VPD, vapor pressure deficit; RMSE, root mean square error; MAE, mean absolute error; NSE, Nash–Sutcliffe model efficiency coefficient.

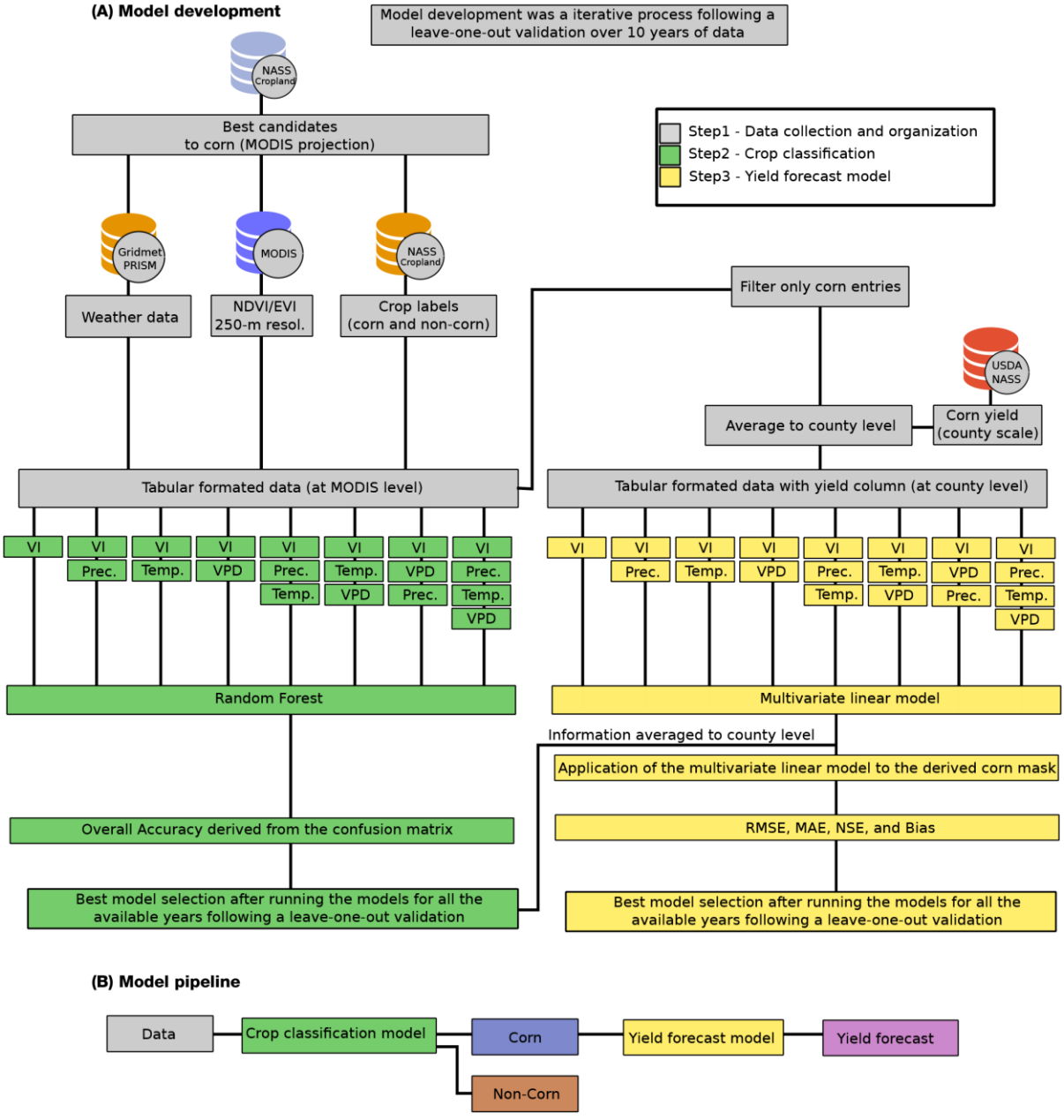

FIGURE 2. (A) Overall out-of-sample accuracy (number of correctly classified observations divided by the total number of observations derived from the confusion matrix), for all the years and states considered in this study (each point represents a year and state). The shapes around the points represent the kernel density plot for the accuracy in each condition. (B) Thematic maps representing the spatial distribution of corn pixels from Cropland Data Layer (at Moderate Resolution Imaging Spectroradiometer scale; green) and predicted by the model using Normalized Difference Vegetation Index and Enhanced Vegetation Index (blue) for 2017. VI, vegetation index; Prec., precipitation; Temp., temperature; VPD, vapor pressure deficit; USDA, United States Department of Agriculture; NASS, National Agricultural Statistics Service.

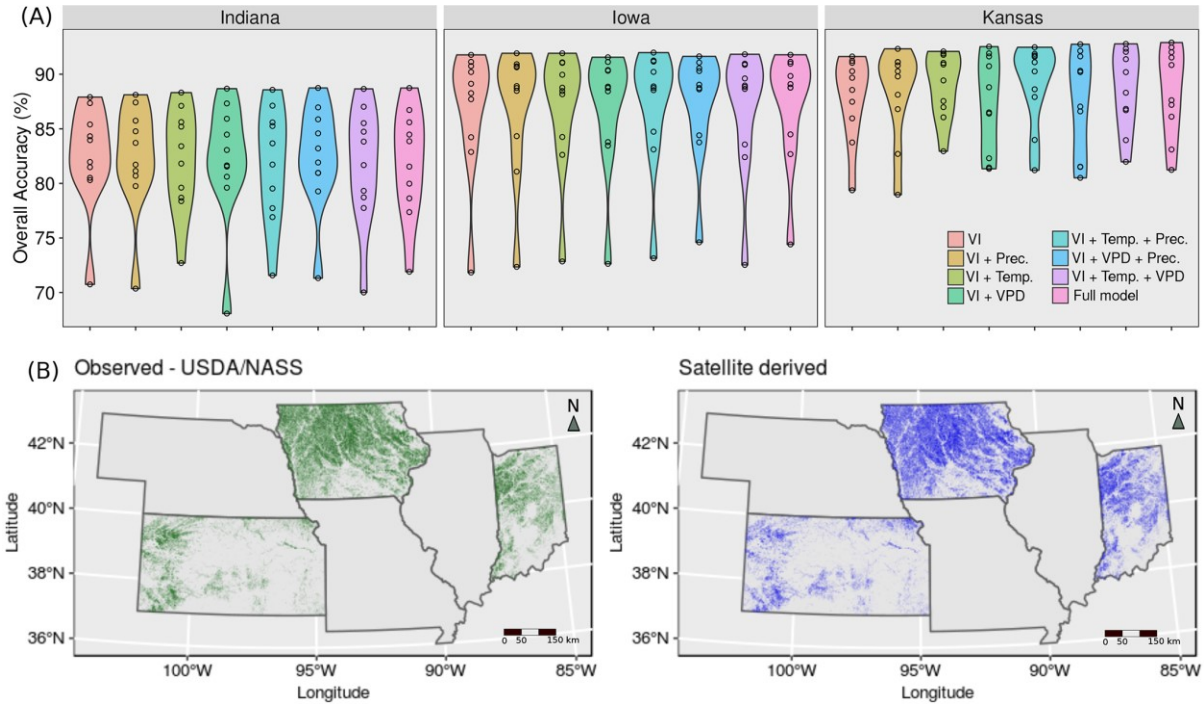

FIGURE 3. (A) Mean absolute error (MAE), Root-mean square error (RSME), Nash–Sutcliffe model efficiency coefficient (NSE), and bias coefficient for all the models tested in the assemble approach. (B) Observed versus out-of-sample forecasted corn yield from a yield forecast model with multi-temporal vegetation indices (VI); (left) and observed versus predicted corn yield from a yield forecast model with multi-temporal VI, temperature (Temp.), and vapor pressure deficit (VPD; right). Predictive yield based on aggregating data until day of year 232 (August 20) for Kansas, Indiana, and Iowa from 2008–2017. The black line is presented in panel portraying the 1:1 line for the observed-predicted relationship. The sample size is n = 2501 data points. Prec., precipitation.

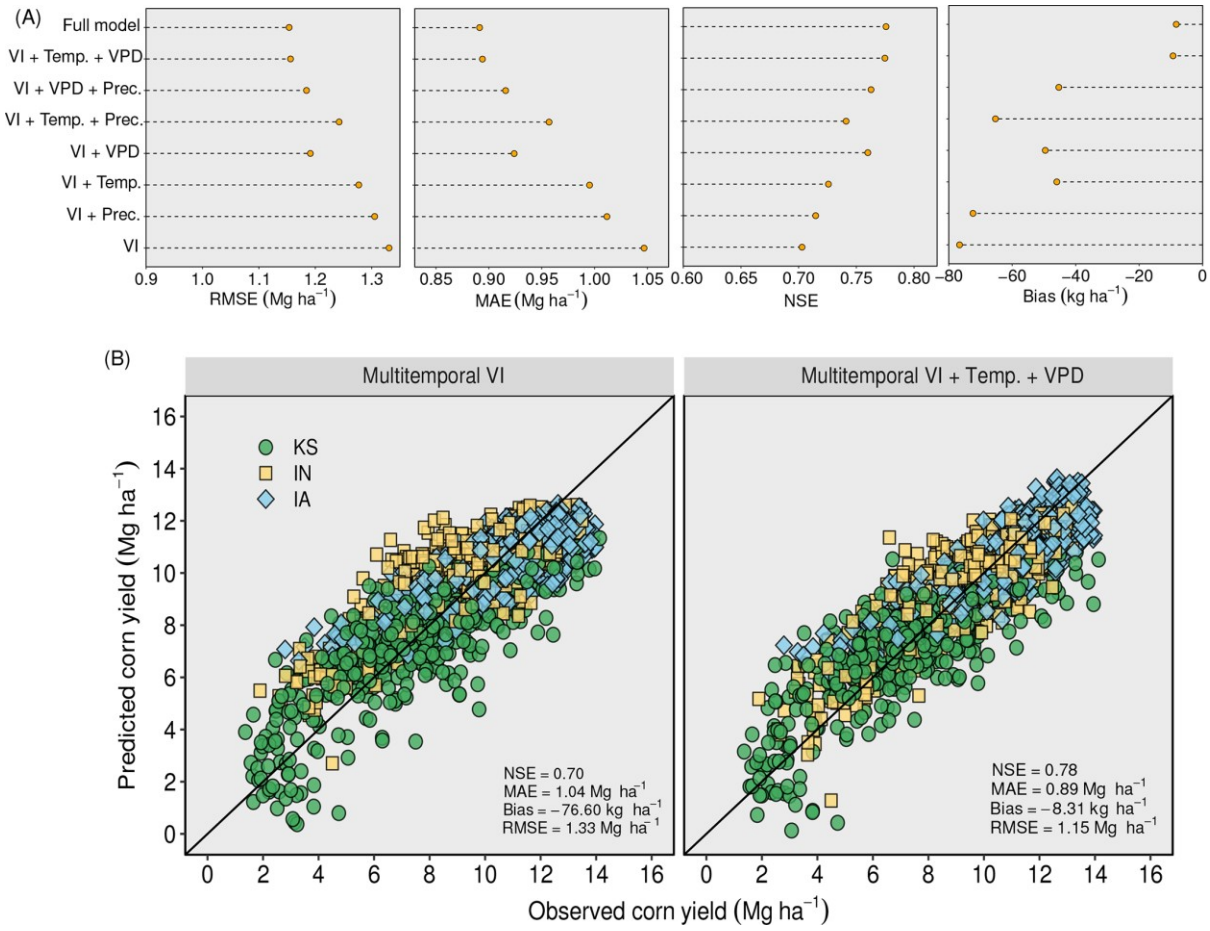

FIGURE 4. (A) Observed versus out-of-sample forecasted corn yield (forecast model with multi-temporal vegetation indices, temperature, and vapor pressure deficit) for different dates expressed in a standard month and day format and in days of year (DOY). A black dashed line is presented in panel portraying the 1:1 line for the observed-predicted relationship. The sample size is $n = 2501$ data points. (B) Variations in the mean absolute error (MAE), root-mean square error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and bias coefficient for different dates of yield prediction.