Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest

Yanghui Kang, Mutlu Ozdogan, Xiaojin Zhu, Zhiwei Ye, Christopher Hain and Martha Anderson

1 Department of Geography, University of Wisconsin-Madison, 550 N. Park St., Madison, WI 53706 United States of America
2 Nelson Institute Center for Sustainability and the Global Environment, University of Wisconsin-Madison, 1710 University Avenue, Madison, WI 53726 United States of America
3 Department of Forest and Wildlife Ecology, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI 53706 United States of America
4 Department of Computer Sciences, University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, WI 53706 United States of America
5 NASA Marshall Space Flight Center, Earth Science Branch, Huntsville, AL 35805 United States of America
6 Hydrology and Remote Sensing Laboratory, USDA ARS, Beltsville, MD 20705 United States of America
7 Author to whom any correspondence should be addressed.
E-mail: kanyanghui@gmail.com

Keywords: crop yields, climate impact, machine learning, deep learning, data-driven

Abstract
Crop yield estimates over large areas are conventionally made using weather observations, but a comprehensive understanding of the effects of various environmental indicators, observation frequency, and the choice of prediction algorithm remains elusive. Here we present a thorough assessment of county-level maize yield prediction in U.S. Midwest using six statistical/machine learning algorithms (Lasso, Support Vector Regressor, Random Forest, XGBoost, Long-short term memory (LSTM), and Convolutional Neural Network (CNN)) and an extensive set of environmental variables derived from satellite observations, weather data, land surface model results, soil maps, and crop progress reports. Results show that seasonal crop yield forecasting benefits from both more advanced algorithms and a large composite of information associated with crop canopy, environmental stress, phenology, and soil properties (i.e. hundreds of features). The XGBoost algorithm outperforms other algorithms both in accuracy and stability, while deep neural networks such as LSTM and CNN are not advantageous. The compositing interval (8-day, 16-day or monthly) of time series variable does not have significant effects on the prediction. Combining the best algorithm and inputs improves the prediction accuracy by 5% when compared to a baseline statistical model (Lasso) using only basic climatic and satellite observations. Reasonable county-level yield forecasting is achievable from early June, almost four months prior to harvest. At the national level, early-season (June and July) prediction from the best model outperforms that of the United States Department of Agriculture (USDA) World Agricultural Supply and Demand Estimates (WASDE). This study provides insights into practical crop yield forecasting and the understanding of yield response to climatic and environmental conditions.

1. Introduction
Timely and reliable crop yield forecasts at local to regional scales are becoming increasingly important as the Earth warms, and the population continues to climb (Foley et al 2011, Mueller et al 2012, Rosenzweig et al 2014). Early and robust estimates of crop yield inform decision making about commodity trading and insurance assessment. Spatially explicit information on historical crop yield also facilitates the understanding of yield gaps, crop response to environmental stress, and the adaptation of cropping systems under climate change scenarios (Lobell et al 2008, Sacks and Kucharik 2011, Sayago et al 2017). The availability of long-term historical statistics of crop production at sub-national units allows the development of data-driven models for crop yield prediction (Johnson 2014, Franch...
et al 2015, López-Lozano et al 2015, Pagani et al 2017). The idea is to associate reported crop yields with various climate and satellite observations (Bolton and Friedl 2013, Johnson et al 2016, Mladenova et al 2017). Climatic variables such as air temperature and precipitation are shown to explain one third of the variation in global county-level crop production (Ray et al 2015). Vapor pressure deficit (VPD) closely relates to crop water and heat stress and thus potentially assists in crop yield forecasting (Lobel et al 2014, Guan et al 2017, Li et al 2019b). Besides weather variables, satellite-derived vegetation indices (VIs) significantly improve the skill of yield predictions in many studies (Becker-Reshef et al 2010, Johnson 2014, Guan et al 2017, Peng et al 2018). In addition, a few studies show that satellite-based evapotranspiration (ET) and soil moisture observations possess unique signatures in quantifying environmental stress and might provide additional information than conventional optical-based VIs (Anderson et al 2016a, 2016b; Guan et al 2017, Mladenova et al 2017, Yang et al 2018).

Relationships between crop yields and the environmental variables are typically described by statistical or machine learning algorithms. Traditional linear or non-linear statistical models provide high interpretability and simplicity to describe the response of crop productivity to predictor variables. As a result, these models are widely used in crop yield forecasts with the help of monthly weather observations and VIs (Becker-Reshef et al 2010, Bolton and Friedl 2013, Sakamoto et al 2014, Peng et al 2018, Li et al 2019b). In recent years, machine learning algorithms, especially deep neural networks, have received increased attention given their ability to describe complex relationships (Yang et al 2019, Zhong et al 2019). Compared to statistical models, machine learning algorithms require no prior assumption about the relationships between response and predictor variables and allow for higher-order interactions, resulting in improved predictive power (Hengl et al 2017, Zhang et al 2019). Several studies have examined the performance of machine learning algorithms for county or state/province-level crop yield estimation mainly using climate variables, VIs, and Land Surface Temperature (LST) derived from satellite with mixed results (Johnson 2014, Kuwata and Shibasaki 2015, Johnson et al 2016, You et al 2017, Wang et al 2018, Kaneko et al 2019). For an instance, Johnson et al (2016) observed comparable forecast skills between multiple linear regression, Bayesian neural networks, and model-based recursive partitioning, while You et al (2017) found neural networks (e.g. Convolutional Neural Network (CNN) and Long-short Term Memory (LSTM)) to outperform decision tree and ridge regression.

While data-driven methods (i.e. statistical and machine learning methods) have become the dominant technique for crop yield forecasting at municipal levels, less attention has been paid to the comparative performance of different models and the behavior of different environmental variables including those that are not commonly used for yield predictions, such as crop progress statistics, soil properties, drought indices and variables from land data assimilation models. In this letter, we present a comprehensive analysis that compares a set of state-of-the-art statistical and machine learning models in combination with an extensive set of climatic and environmental features for their skill in crop yield forecasting. We evaluated the models and inputs using county-level maize yield statistics from 2001 to 2016 for the U.S. Midwest obtained from the United States Department of Agriculture (USDA).

We designed experiments to address the following questions: (1) How does yield respond to various environmental variables? (2) Does an extensive set of variables with higher temporal observation frequency benefit seasonal crop yield predictions? And (3) Is there a significant difference between the performance of different statistical and machine learning algorithms? It is worth noting that county-level yield forecasts are not publicly available during the growing season in the Midwest. The USDA World Agricultural Supply and Demand Estimates (WASDE) publishes monthly forecasts of US national yield primarily based on surveys. We used the WASDE reports as a baseline to benchmark our seasonal yield forecasts at national level.

2. Data and methods

2.1. Study area and yield data

The study area includes 12 states in U.S. Midwest: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. We obtained county-level maize yield statistics from the National Agricultural Statistics Service (NASS) of USDA for 2001 to 2016. For counties with both rainfed and irrigated maize yield, we used the reported overall yield value. Counties that had less than one percent cultivated maize areas were discarded.

2.2. Climatic and environmental datasets

We compiled an extensive set of variables extracted from gridded weather data, optical, thermal, and microwave remotely sensed observations, land surface data assimilation systems, drought indices, gridded soil maps, as well as crop progress reports (table 1). These include time-series of land surface temperature (LST), surface reflectance, and VIs including EVI, Green Chlorophyll Index (GCI) (Gitelson 2005), and Normalized Difference Water Index (NDWI) (Gao 1996), derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite images (Schaaf and Wang 2015, Wan et al 2015); air temperature, precipitation, and VPD from
the Parameter-elevation Relationships on Independent Slopes Model (PRISM) climate dataset (Daly et al 2008, 2015); surface soil moisture from the European Space Agency (ESA) Climate Change Initiative (CCI) dataset (Gruber et al 2013); rootzone soil moisture, potential ET, and ET from the Global Land Data Assimilation System (GLDAS-v2.1) dataset (Rodell et al 2004); and ET from the Atmosphere-Land Exchange Inverse (ALEXI) remote sensing model (Anderson et al 2011b). We derived a water stress (WS) indicator defined as the ratio between ET and PET from the GLDAS dataset. We incorporated three drought indices including the Palmer Drought Severity Index (PDSI) (Abatzoglou et al 2014), Standardized Precipitation Index (SPI) derived from PRISM data (Guttman 1999), and the Soil Moisture Deficient Index (SMDI) derived from the CCI soil moisture time series (Narasimhan and Srinivasan 2005). A heat index, Extreme Degree Day (EDD), was computed as the cumulative temperature above 30 °C during summer months (June–August) using the PRISM temperature data (Lobell et al 2013, Jin et al 2016). Soil properties such as soil available water content (AWC), organic matter content (SOM) and nitrogen content were derived from gridded soil datasets SoilGrids250 m (Hengl et al 2017) and SoilGrid + (Ramcharan et al 2018). We included county-level statistics such as maize harvested area, irrigated fraction, crop progress, and previous 5 year yield mean extracted from the USDA NASS. The crop progress data at state-level were converted to the dates when a certain percentage of acreage reaches a specific growth stage. Detailed description of each variable and its relation to yield is provided in table S1 of supporting information

| Dataset | Spatial resolution | Temporal resolution | Variables |
|---------|-------------------|---------------------|-----------|
| PRISM   | 1 km              | Daily               | Air Temperature, Precipitation, Vapor Pressure Deficit, Standardized Precipitation Index (SPI), Extreme Degree Day (EDD) |
| MODIS Nadir BRDF\(^a\)-Adjusted Reflectance (MCD43A4) | 500 m | Daily | Surface Reflectance, EVI, Green Chlorophyll Index (GCI), Normalized Difference Water Index (NDWI) |
| MODIS Terra (MOD11A2\(^b\)) | 1 km | 8-day | Daytime and Nighttime Land Surface Temperature (LST) |
| ESA CCI\(^c\) Microwave Soil Moisture v04.2 | 0.25 arc degree | Daily | Surface Soil Moisture (top 5 cm) |
| ALEXI\(^d\) model | 4 km | Daily | Evapotranspiration (ET) |
| GLDAS-v2.1 | 0.25 arc degree | 3 hourly | ET, Potential ET, Root Zone Soil Moisture, Water Stress\(^e\) |
| University of Idaho gridMET dataset | 4 km | 10 d | Palmer Drought Severity Index (PDSI) |
| SoilGrids250 m | 250 m | — | Soil Available Water Content, Organic Matter |
| Soil Properties and Class 100 m Grids United States | 100 m | — | Soil Nitrogen Content |
| NASS database | County | — | Crop progress, Harvest Area, Percentage of Harvest Area out of County Area, Percentage of Irrigated Area out of County Area, Last 5 year Yield Mean |

\(^a\)Bidirectional Reflectance Distribution Function (BRDF).
\(^b\)MODIS Aqua product (MYD11A2) was used when the Terra product was not available from 2001 to early 2002.
\(^c\)The ESA CCI soil moisture product (http://www.esa-soilmoisture-cci.org/node/139) merges surface soil moisture measures from various active and passive microwave observations, including ERS-1, ERS-2, SSM/I, TRMM TMI, and Aqua AMSR-E.
\(^d\)The ALEXI is an ET estimation model primarily driven by satellite remote sensing products. ALEXI mainly uses surface radiometric temperature sensed by geostationary satellites and Leaf Area Index (LAI) from MODIS.
\(^e\)Water stress was derived at the ratio of ET and Potential ET (PET).

2.3. Machine learning algorithms

We examined six machine learning algorithms including least absolute shrinkage and selection operator (LASSO) regression, Support Vector Regression (SVR), Random Forest (RF), eXtreme Gradient Boost (XGBoost), as well as two deep neural networks—LSTM and CNN. LASSO is a linear regression model that performs L1 regularization and variable selection in favor of simple and sparse models. SVR is a variant of the Support Vector Machine (SVM) that optimizes hyperplanes to separate training samples and uses kernels to incorporate non-linear relationships. Random forest is an ensemble machine learning algorithm that creates regression trees using subsets
of features and bootstrap samples. XGBoost is an algorithm that implements an optimized gradient boost technique over decision/regression trees (Chen and Guestrin 2016). Random forest and XGBoost are the most widely used machine learning algorithms in Geoscience and other areas (Anderson and Lucas 2018, Fernández-Delgado et al 2019, Hengl et al 2017; Shangguan et al 2017). LSTM network is a recurrent neural network (RNN) where a sequence of inputs is processed by nodes with directed connections (Lecun et al 2015). CNN is a specialized class of neural networks mostly used in image recognition. Figure S1 shows the LSTM and CNN architectures used in our analyses. Detailed descriptions of the models and parameterization can be found in table S2 and text S1.

2.4. Experiment settings and model evaluation
In exploratory analysis, we assessed the relationship between individual environmental variables and yield. The temporal dynamics of both Pearson’s ($r$) and Spearman’s ($r_s$) correlation coefficient was analyzed for sequential variables at an 8-day time interval as well as for non-sequential variables. The general trends of yield response to environmental variables were established using Multivariate Adaptive Regression (MARS) models with the year as a secondary predictor. The full dataset (2001–2016) was used for the exploratory analysis, while a separate assessment on the subset (2001–2006) involved only in the training sets showed similar results.

Based on the correlation analysis, we constructed three sets of environmental variables to build the feature space. A ‘Minimum (Min) set’ contains variables that are commonly used in previous yield prediction studies, including EVI, NDWI, daily average air temperature, average daily VPD, precipitation, daytime and nighttime LST, as well as the year and geographic coordinates to model spatiotemporal relationships. A ‘Good set’ extends the minimum set with variables that show a substantial correlation with yield (Section 3.1). These include ET, water stress, SMDI drought index, EDD, AWC, previous 5 year average yield, harvested and dented progress, etc. A third level—‘Full set’—includes all the variables. In addition, we tested the effect of temporal aggregations over sequential variables at 8-day, 16-day, or monthly time steps. These resulted in nine experimental levels based on three levels of variable selection and three time steps. These resulted in nine experimental levels over sequential variables at 8-day, 16-day, or monthly addition, we tested the effect of temporal aggregations with variables that show a substantial correlation with yield in summer consistent with previous findings (Lobell et al 2017). Heat stress variables, especially daytime LST, maximum VPD, and maximum daily air temperature in the last week of July are between yield and daytime LST, maximum VPD, and maximum daily air temperature, have strong negative correlations with yield during the whole growing season (p-value below 0.001) (figure S2). Based on their similarities in the temporal relationships with yield, variables are grouped into three categories (figure S3). The first group includes surface reflectance and VIs representing the amount of green foliage in the crop canopy. These variables possess strong correlations with yield during the whole growing season, especially in summer. The second group indicates crop heat stress, including air temperature, LST, and VPD. The last group is comprised of other variables that are associated with crop water stress, including ET, soil moisture, water stress index, and drought index SMDI. Besides, there are also high inter-correlations between variables within these groups (figure S4).

Overall, VIs (GCI, NDWI, EVI) and the NIR band from MODIS show the strongest signal (figure 1(a)). The correlation between yield and GCI/NDWI/NIR is between 0.72–0.74 from mid-July to mid-August, attributed to their high sensitivity to canopy green leaf volume, which directly contributes to grain growth. The negative correlations between NIR/VIs and yield (positive for the visible and SWIR bands) in April and May are consistent with previous studies (Johnson 2016) and related to the increasing trend of yield from low (earlier planting dates) to high latitudes (later planting dates). Heat stress variables, especially daytime LST, maximum and mean VPD, and maximum daily air temperature, have strong negative correlations with yield in summer consistent with previous findings (Lobell et al 2013). The correlation between yield and daytime LST, maximum VPD, and maximum air temperature in the last week of July are $-0.53, -0.47$, and $-0.41$, respectively. The superior performance of LST over air temperature agrees with previous studies, as canopy temperatures incorporate
the effect of soil water moisture and evaporative cooling, and are thus better indicators for heat stress and drought impact (Siebert et al 2014, Pede et al 2019). Among the variables related to water stress, ALEXI ET exhibits the highest correlation between mid-July and late-August ($r \sim 0.38-0.42$). Among the drought indices, SMDI based on CCI microwave surface soil moisture explain more yield variation than SPI and PDSI (maximum $r = 0.30$ vs. 0.28). But SMDI based on GLDAS root zone soil moisture shows only weak correlations with yield (maximum $r = 0.22$). Other variables, including nighttime LST, minimum VPD,
Figure 2. Comparison of county-level maize yield estimation (end-of-season) between different models and feature levels. (a) Average RMSE of yield estimation from 2007 to 2016 by different models and feature levels. The error bars refer to (mean ± standard deviation). (b) Average MAPE of yield estimation from 2007 to 2016 by different models and feature levels. Heatmaps for each model show MAPE of each feature levels. The number of features increase from the bottom left to the top right of the heatmaps. Refer to table S3 for the exact number of features.

potential ET, rainfall, and root zone soil moisture, show only weak correlations. Note that the correlation between rainfall and yield (maximum $r = 0.26$) is higher when rainfall is aggregated over a longer period (i.e. month) rather than 8-day (figure S5) since the effect of rainfall on yield is often cumulative.

Most non-sequential variables are highly correlated with yield (figure 1(b)). The previous five-year yield average has the highest correlation ($r = 0.67$) with yield, followed by the percentage of county area planted with maize and total harvested area. EDD negatively impacts yield with $r = -0.46$, stronger than the correlation with LST. The correlation between AWC and yield ($r = 0.23$) is also significant. Most crop progress variables show statistically significant correlations with yield. The dates of harvested and dented stage both exhibit high correlations with yield at $r \sim 0.24$–0.28, indicating that longer growing season favors crop yield (figure 1). The positive correlation between planting/emergence/silking and yield is a manifest of the relationship between yield and latitude (Long et al 2017) since we did not find any significant correlation when fixing the effect of latitude (figure S6).

A separate assessment using Spearman’s correlation coefficient, which considers non-linear relationships, shows similar results for both sequential and non-sequential variables (figure S7).

Figure 1(c) shows the yield response curves of selected variables by fixing the year in 2008 (full result in figure S8). Most variables exhibit expected behaviors. The response curve of rainfall suggests that both drought conditions and excessive rainfall are harmful to yield (Lobell et al 2014, Li et al 2019a). High maximum temperature of July harms yield, especially when it exceeds 33.5 °C. A linear reduction in yield is found with increasing extreme heat events (i.e. EDD). In addition, we also found increasing trends in the correlation between heat and water stress variables and yield since 2001 (figure S9), suggesting increasing sensitivity of yield to extreme drought and heat conditions (Hawkins et al 2013, Lobell et al 2014).

3.2. Comparison of machine learning models and features

All machine learning models show reasonable accuracy in predicting end-of-season county-level maize yield (figures 2, S10). Overall, Lasso has the lowest accuracy for all feature levels (figure S11; $p < 0.01$), with average RMSE (2007–2016) between 16.7 and 24 bu ac$^{-1}$ and a large year-to-year fluctuation. SVR performs better than Lasso but not as good as the other models. XGBoost has the best overall performance (figure S11; $p < 0.01$) and the smallest year-to-year variation. It achieves the highest accuracy with the Good set and 16-day aggregation: RMSE = 14.8 bu ac$^{-1}$, MAE = 11.4 bu ac$^{-1}$, and MAPE = 9%. The superiority of XGBoost over RF is attributed to the boosting technique that iteratively reduces bias and variance. In addition, XGBoost also runs significantly faster than SVR and RF given the same computing resources (table S4). Interestingly, we did not find any significant difference between the performance of deep neural networks and RF and SVM (figure S11). LSTM only marginally
Regression models were trained with XGBoost algorithm using three different feature sets: Min set, Good set, and Full set. Forecasts were made twice in a month with an early-month (E) and a late-month (L) estimates. (a) Yield estimation error (mean RMSE across 2007–2016) at different forecast time with different feature levels. The inset shows the difference in RMSE between Good/Full set and the Min set. (b) RMSE of maize yield estimation by county and different forecast time in U.S. Midwest using the Good set.

Figure 3. Within-season maize yield forecast error across U.S. Midwest. Regression models were trained with XGBoost algorithm using three different feature sets: Min set, Good set, and Full set. Forecasts were made twice in a month with an early-month (E) and a late-month (L) estimates. (a) Yield estimation error (mean RMSE across 2007–2016) at different forecast time with different feature levels. The inset shows the difference in RMSE between Good/Full set and the Min set. (b) RMSE of maize yield estimation by county and different forecast time in U.S. Midwest using the Good set.

outperforms RF and CNN (not statistically significant, figure S11), but it is not as good as XGBoost especially when a small number of features (monthly or Min set) was involved (figure 2). CNN and RNN are designed to capture specific patterns in structured data: CNN specializes in recognizing spatial locality in images and RNN in detecting sequential dependencies. However, the tabular style dataset used in this analysis has no spatial-related patterns for the convolutional kernels of CNN, although the sequential features might have favored LSTM, explaining its slightly better performance over CNN. This observation is consistent with previous studies (Wang et al 2017, Fernández-Delgado et al 2019, Zhang et al 2019).

The yield prediction accuracy using the Good and Full sets is significantly higher than that of the Min set across all models, while the difference between the Good and Full sets is minimal (figures 2, S11). Additional information contained in the Good/Full set, such as previous yield records, ET, soil moisture, EDD, soil properties, and crop progress reduces the MAPE by more than 1% for RF, XGBoost, CNN, LSTM and 2%–5% for Lasso and SVR (figure 2(b)). The performance of non-neural-network algorithms does not always improve as sequential features are aggregated at smaller time steps (figures 2(b), S11). For example, RF achieves the lowest accuracy using 8-day aggregated observations. While frequent observations contain more information during rapid growth periods, aggregation over a longer time period reduces random errors and the frequency of missing data. On the contrary, the accuracy of CNN and LSTM increases with smaller temporal aggregation steps and more variables (figure 2(b)), indicating the ability of these deep neural networks to work on large feature spaces. Nevertheless, there is not a significant overall effect of temporal aggregation frequency (figure S11).
Feature importance was evaluated using RF and XGBoost with the Full set and monthly aggregation. NDVI, GCI, and NIR reflectance in July and August were found to be the most important features for yield prediction (figure S12). Other important features include previous 5 year yield average, harvest area, percentage of irrigated area, potential ET, drought conditions (PDSI, SMDJ), June and July average VPD, as well as soil properties (organic matter content, available water content). XGBoost also leverages the information from crop progress data such as the time of emergence and silking. The feature importance rankings are different each year, as the controlling environmental factors of yield vary (figures S13, S14).

3.3. Within-season yield forecast
Within-season yield forecasting was evaluated using XGBoost, 16-day temporal aggregation of features, and three variable selection levels (figure 3). Using only information from March to May, the early June prediction achieves a reasonable estimation accuracy with 23.9 bu ac$^{-1}$ RMSE (16% MAPE) based on the Good or Full sets, attributed to the strong correlations between yield and MODIS reflectance and ALEXI ET from March and May (figure 3). As more information becomes available in the summer (June–August), the prediction accuracy gradually improves (figure 3(a)). Again, there is not a significant difference in the performance between the Good and Full sets. In addition, we tested a remote sensing Evaporative Stress Index (ESI), defined as the climatological anomalies in the ratio between actual ET and potential ET (Anderson et al 2007, 2011b). Compared to other ET related variables, such as ALEXI ET and GLDAS water stress, ESI shows a higher correlation with yield early in the season (i.e. before June) (figure S15). However, adding ESI to the Good sets does not significantly improve the within-season yield forecasting (figure S15).

The evolution of the forecasting accuracy during the growing season has significant spatiotemporal patterns (figures 3(b), S16,S17). The benefit of summer information is the most pronounced in southern states, such as Kansas, Missouri, southern Illinois, and Indiana (figures 3(b), S16). The prediction accuracy is higher in northern states, including North Dakota, Minnesota, Wisconsin, Iowa, and Michigan (figure 3(b)), where more than half of the counties could achieve RMSE below 10 bu ac$^{-1}$ by early September. While almost all the years see a decrease of RMSE from June to August, the decline is especially dramatic in 2012, when a summer drought significantly depressed yields (Grigg 2014) (figure S17). As a result, early-season forecasting in 2012 is inefficient with a significant overestimation of yield (figure S18). Although the yield forecast skill increases when late June and July data are provided, the late-season prediction accuracy of 2012 is still much lower compared to other years, especially for counties with significant irrigation presence (figure S18). This indicates that the data-driven models might not extrapolate well to unseen climatic conditions, and thus could lead to undermined performance.

The within-season yield forecasts produced by the Good set was further compared to that of the USDA WASDE estimates at national level. XGBoost significantly outperforms that of the WASDE in the early season (June and July) by 2 to 2.5 bu ac$^{-1}$ in RMSE, especially in extreme years such as 2012 (figure S19). XGBoost results are comparative to WASDE in August, and only slightly worse in September. Note that starting from August, the WASDE was primarily driven by surveys consisting of growers’ assessment and field measurements, and thus became much more accurate.

4. Conclusion
This study analyzed the performance of data-driven methods for seasonal crop yield estimation using an extensive set of environmental variables and six machine learning models. Almost all variables possess significant correlations with yield from March to October, with the strongest being surface reflectance, VIS, daytime LST, VPD, and ALEXI ET. The profound signals in surface reflectance and ALEXI ET during March to May contribute to the reasonably high accuracy of county-level yield forecasting in June—4 months before harvest. The national-level forecasts outperform that of the USDA WASDE reports in the early season (June and July).

The XGBoost model outperforms other algorithms both in accuracy and stability, while deep neural networks such as CNN and LSTM are not advantageous, especially for datasets involving small feature space. The XGBoost also computationally scales better than RF and SVM given the same workload. Using Lasso and commonly used monthly observations (climatic, LST, and VIS) (min set) as the baseline, we found that including variables about environmental stresses and phenology reduces the percentage error by 2.8% (Full set); using stronger estimators sees an error reduction of 1.6% (SVR) to 3.8% (XGBoost). Combining ‘more estimators’, ‘better variables’, and ‘appropriate temporal frequency’ results in an improvement of accuracy up to 5% (XGBoost, 16-day).

These results contribute to a comprehensive understanding of the response of maize yield to various environmental variables and the comparative performance of seasonal crop yield forecasts across
different types of machine learning algorithms, different predictor variables, and different temporal aggregation steps. All the data inputs are freely accessible to the public. Further improvement of the yield prediction modeling could benefit from more explicit information about crop type and irrigation practices, seasonal climate prediction data and ensemble stacking techniques.

Acknowledgments

Y. K. was supported by the NASA Space and Earth Science Fellowship (NNX12AN59H). The CNN models were implemented using the computing resources and assistance of the UW-Madison Center for High Throughput Computing (CHTC) in the Department of Computer Sciences. CHTC is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy’s Office of Science.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID ID

Yanghui Kang https://orcid.org/0000-0001-8563-1503

References

Abatzoglou J T, Barbero R, Wolf J W and Holden Z A 2014 Tracking interannual streamflow variability with drought indices in the U.S. Pacific Northwest J. Hydrometeorol. 15 1900–12

Anderson G J and Lucas D D 2018 Machine learning predictions of a multiresolution climate model ensemble Geophys. Res. Lett. 45 4273–80

Anderson M C, Zolin C A, Sentelhas P C, Hain C R, Semmens K, Tugrul Yilmaz M, Gao E, Otkin J A and Tetraart R 2016b The evaporative stress index as an indicator of agricultural drought in Brazil: an assessment based on crop yield impacts Remote Sens. Environ. 174 82–99

Becker-Reshef I, Vermote E, Lindeman M and Justice C 2010 A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data Remote Sens. Environ. 114 1312–21

Bolton D K and Friedl M A 2013 Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics Agric. For. Meteorol. 173 74–84

Chen T and Guerzoni C 2016 XGBoost: a scalable tree boosting system Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining—KDD ’16 pp 785–94

Daly C, Halbleib M, Smith J J, Gibson W P, Doggett M K, Taylor G H, Curtis J and Pasteris P P 2008 Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States Int. J. Climatol. 28 2031–2064

Daly C, Smith J I and Olson K V 2015 Mapping atmospheric moisture climatologies across the conterminous United States PLoS One 10 e0141140

Fernández-Delgado M, Sirsat M S, Cernadas E, Alawadi S, Barro S and Febrero-Bande M 2019 An extensive experimental survey of regression methods Neural Networks 111 11–34

Foley J A et al 2011 Solutions for a cultivated planet Nature 478 337–42

Franch B, Vermote E F, Becker-Reshef I, Claverie M, Huang J, Zhang J, Justice C and 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 Sens. Environ. 161 131–48

Friedl M and Sulla-Menashe D 2018 MCD12Q1 MODIS/terra+aqua land cover type yearly L3 global 500m SIN grid V006 [data set]. NASA EOSDIS Land Processes DAAC (https://doi.org/10.5067/MODIS/MCD12Q1.006)

Gao B C 1996 NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space Remote Sens. Environ. 58 257–66

Gitelson A A 2005 Remote estimation of canopy chlorophyll content in crops Geophys. Res. Lett. 32 L08403

Grigg N S 2014 The 2011–2012 drought in the United States: new lessons from a record event Int. J. Water Resour. Dev. 30 183–99

Gruber A, Scanlon T, Van Der Schalie R, Wagner W and Dorigo W 2019 Evaluation of the ESA CCCI soil moisture climate data records and their underlying merging methodology Earth Syst. Sci. Data 11 717–39

Guan K, Wu J, Anderson M C, Kimball J, Frolking S, Li B and Lobell D 2017 The shared and unique value of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields Remote Sens. Environ. 199 333–49

Guttman N B 1999 Accepting the standardized precipitation index: a calculation algorithm J. Am. Water Res. Assoc. 35 311–22

Hawkins E, Fricker T E, Challinor A J, Ferro C A T, Ho C K and Osborne T M 2013 Increasing influence of heat stress on French maize yields from the 1960s to the 2030s Glob. Change Biol. 19 937–47

Hengl T et al 2017 SoilGrids250m: global gridded soil information based on machine learning PLoS One 12 e0169748

Jin Z, Zhuang Q, Tan Z, Dukes J S, Zheng B and Melillo J M 2016 Do maize models capture the impacts of heat and drought stresses on yield? Using algorithm ensembles to identify successful approaches Glob. Change Biol. 22 3112–26

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. 141 116–28

Johnson D M 2016 A comprehensive assessment of the correlations between field crop yields and commonly used
MODIS products Int. J. Appl. Earth Obs. Geoinf. S2 65–81
Johnson M D, Hsieh W W, Cannon A J, Davidson A and Bécard F 2016 Crop yield forecasting on the canadian prairies by remotely sensed vegetation indices and machine learning methods Agric. For. Meteorol. 218–219 74–84
Kaneko A, Kennedy T, Mei L, Sintek C, Burke M, Ermon S and Lobell D 2019 Deep Learning for Crop Yield Prediction in Africa Int. Conf. on Machine Learning AI for Social Good Workshop (Long Beach, United States) pp 1–5
Kuwata K and Shibasaki R 2015 Estimating crop yields with deep learning and remotely sensed data IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS) pp 858–61
Lecun Y, Bengio Y and Hinton G 2015 Deep learning Nature 521 436–44
Li Y, Guan K, Schnitkey G D, Delucia E and Peng B 2019a Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States Glob. Chang. Biol. 25 2325–37
Li Y, Guan K, Yu A, Peng B, Zhao L, Li B and Peng J 2019b Toward building a transparent statistical model for improving crop yield prediction: modeling rainfall in corn in the U.S. (under review) F. Crop. Res. 234 55–65
Lobell D B, Burke M B, Tebaldi C, Mastrandrea M D, Falcon W P and Naylor R L 2008 Prioritizing climate change adaptation needs for food security in 2030 Science 319 607–10
Lobell D B, Hammer G L, Mclean G, Messina C, Roberts M J and Schlenker W 2013 The critical role of extreme heat for maize production in the United States Nat. Clim. Change 3 497–501
Lobell D B, Roberts M J, Schlenker W, Braun N, Little B B, Rejeus R M and Hammer G J 2014 Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest Science 344 516–19
Long N V, Assefa Y, Schwulbert R and Ciampitti I A 2017 Maize yield and planting date relationship: a synthesis-analysis for US high-yielding contest-winner and field research data Front. Plant Sci. 8 1–9
López-Lozoano R, Duveiller G, Seguin L, Meroni M, García-Condado S, Hooker J, Leo O and Baruth B 2015 Towards regional grain yield forecasting with 1km-resolution EO biophysical products: strengths and limitations at pan-European level Agric. For. Meteorol. 206 12–32
Mladenova I E, Bolten J D, Crow W T, Anderson M C, Hain C R, Johnson D M and Mueller R 2017 Inter-comparison of soil moisture, water use and vegetation indices for estimating corn and soybean yields over the U.S. IEEE J. 20 10 1–16
Mueller N D, Gerber J S, Johnston M, Ray D K, Ramankutty N and Foley J A 2012 Closing yield gaps through nutrient and water management Nature 490 254–7
Narasimhan B and Srinivasan R 2005 Development and evaluation of soil moisture deficit index (SMDI) and evapotranspiration deficit index (ETDI) for agricultural drought monitoring Agric. For. Meteorol. 133 69–88
Pagani V et al 2017 Improving cereal yield forecasts in Europe—the impact of weather extremes Eur. J. Agron. 89 57–106
Pede T, Mountirasik G and Shaw S B 2019 Improving corn yield prediction across the US corn belt by replacing air temperature with daily MODIS land surface temperature Agric. For. Meteorol. 276–277 107615
Peng B, Guan K, Pan M and Li Y 2018 Benefits of seasonal climate prediction and satellite data for forecasting US maize yield. Geophys. Res. Lett. 1–10
Ramcharan A, Hengl T, Nauman T, Brungard C, Waltman S, Wills S and Thompson J 2018 Soil property and class maps of the conterminous US at 100 meter spatial resolution based on a compilation of national soil point observations and machine learning Soil Sci. Soc. Am. J. 186–201
Ray D K, Gerber J S, MacDonald G K and West P C 2015 Climate variation explains a third of global crop yield variability Nat. Commun. 6 1–9
Rodell M et al 2004 The global land data assimilation system Bull. Am. Meteorol. Soc. 85 381–94
Rosenzweig C et al 2014 Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison Proc. Natl. Acad. Sci. U.S.A. 111 3268–73
Sacks W J and Kucharik C J 2011 Crop management and phenology trends in the U.S. corn belt: impacts on yields, evapotranspiration and energy balance Agric. For. Meteorol. 151 882–94
Sakamoto T, Gitelos A A and Arkebauer T J 2014 Near real-time prediction of U.S. corn yields based on time-series MODIS data Remote Sens. Environ. 147 219–31
Sayago S, Ovando G and Bocco M 2017 Landsat images and crop model for evaluating water stress of rainfed soybean Remote Sens. Environ. 198 30–39
Schaff H and Wang Z. 2015, MCD43A4 MODISterra +aqua BRDF/albedo nadir BRDF adjusted ref daily L3 global—500m V006 [data set] distributed by NASA EOSDIS Land Processes DAAC (https://doi.org/10.5067/MODIS/ MCD43A4.006) Shangguan W, Heng T, de Jesus J M, Yuan H and Dai Y 2017 Mapping the global depth to bedrock for land surface modeling J. Adv. Model. Earth Syst. 9 65–88
Siebert S, Ewert F, Eyshi Rezaei F, Kage H and Graß R 2014 Impact of heat stress on crop yield—on the importance of considering canopy temperature Environ. Res. Lett. 9 044012
Wan Z, Hook S and Hulley G 2015 MOD11A2 MODIS/terra land surface temperature/emissivity 8-day L3 global 1km SIN grid V006 [data set] Wang A X, Tran C, Desai N, Lobell D and Ermon S 2018 Deep transfer learning for crop yield prediction with remote sensing data COMPASS ’18: Proc. of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies 50 pp 1–5
Wang S, Aggarwal C and Liu H 2017 Using a random forest to map the global depth to bedrock for land surface modeling Proc. of the Thirty-First AAAI Conf. on Artificial Intelligence pp 1–9
Yang Q, Shi L, Han J, Zha Y and Zhu P 2019 Deep convolutional neural networks for rice grain yield estimation at the ripening stage using UAV-based remotely sensed images F. Crop. Res. 235 142–53
Yang Y, Anderson M C, Gao F, Wardlow H, Bain C R, Otkin J A, Alfieri J, Yang Y, Sun L and Dulaney W 2018 Field-scale mapping of evaporative stress indicators of crop yield: an application over mead, NE, USA Remote Sens. Environ. 210 387–402
You J, Li X, Low M, Lobell D and Ermon S, 2017. Deep gaussian process for crop yield prediction based on remote sensing data Proc. of the Thirty-First AAAI Conf. on Artificial Intelligence pp 4559–65
Zhang Z, Jin Y, Chen B and Brown P 2019 California almond yield prediction at the orchard level with a machine learning approach Front. Plant Sci. 10 1–18
Zhong L, Hu L and Zhou H 2019 Deep learning based multi-temporal crop classification Remote Sens. Environ. 221 430–45