International Journal of Biometeorology (2021) 65:489–502
https://doi.org/10.1007/s00484-020-02039-z

ORIGINAL PAPER

In-season weather data provide reliable yield estimates of maize and soybean in the US central Corn Belt

Vijaya R. Joshi[1,2,3] · Maciej J. Kazula[1] · Jeffrey A. Coulter[1] · Seth L. Naeve[1] · Axel Garcia y Garcia[1,2]

Received: 8 June 2020 / Revised: 13 October 2020 / Accepted: 27 October 2020 / Published online: 21 November 2020

Abstract

Weather conditions regulate the growth and yield of crops, especially in rain-fed agricultural systems. This study evaluated the use and relative importance of readily available weather data to develop yield estimation models for maize and soybean in the US central Corn Belt. Total rainfall (Rain), average air temperature (Tavg), and the difference between maximum and minimum air temperature (Tdiff) at weekly, biweekly, and monthly timescales from May to August were used to estimate county-level maize and soybean grain yields for Iowa, Illinois, Indiana, and Minnesota. Step-wise multiple linear regression (MLR), general additive (GAM), and support vector machine (SVM) models were trained with Rain, Tavg, and with/without Tdiff. For the total study area and at individual state level, SVM outperformed other models at all temporal levels for both maize and soybean. For maize, Tavg and Tdiff during July and August, and Rain during June and July, were relatively more important whereas for soybean, Tavg in June and Tdiff and Rain during August were more important. The SVM model with weekly Rain and Tavg estimated the overall maize yield with a root mean square error (RMSE) of 591 kg ha^{-1} (4.9% *n*RMSE) and soybean yield with a RMSE of 205 kg ha^{-1} (5.5% *n*RMSE). Inclusion of Tdiff in the model considerably improved yield estimation for both crops; however, the magnitude of improvement varied with the model and temporal level of weather data. This study shows the relative importance of weather variables and reliable yield estimation of maize and soybean from readily available weather data to develop a decision support tool in the US central Corn Belt.

Keywords Yield forecasting · Statistical modeling · Weather index · Crop modeling

Introduction

Maize (*Zea mays* L.) and soybean [*Glycine max* (L.) Merr.] production in the US central Corn Belt is predominantly rain fed and highly dependent on in-season weather conditions (Franzluebbers et al. 2011; Green et al. 2018). In several areas of this region, weather variability has been reported to account for more than 75% of maize yield variation (Ray et al. 2015). Across the USA, 40 and 37% yield variability in maize and soybean was explained by in-season weather conditions, respectively, with an even greater percentage in the Corn Belt (Leng et al. 2016). Precipitation and air temperature during the growing season, particularly at late vegetative and early reproductive stages of development, are reported to cause significant yield deviations from average in both maize and soybean (Teasdale and Cavigelli 2017). Given the strong dependency of rain-fed maize and soybean yield on weather conditions, readily available meteorological data, such as average air temperature (Tavg) and rainfall (Rain), can be used to develop yield estimation models as decision support tools. Yield estimation is one of the most important factors for decision-making on crop insurance, crop management, storage needs, and marketing (Peng et al. 2018). Thus, the development of a yield estimation model based on easily accessible weather data can aid farmers in making informed decisions (Chen et al. 2019) and policy-makers in drafting policies on trade and food security (Basso et al. 2013; van der Velde et al. 2019).

Several attempts have been made at developing weather-based crop yield estimation models. For example, Kaul et al. (2005) used field-specific weekly rainfall and United States

✉ Axel Garcia y Garcia
axel@umn.edu

[1] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, MN 55108, USA

[2] Southwest Research and Outreach Center, University of Minnesota, 23669 130th Street, Lamberton, MN 56152-1326, USA

[3] Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL 32611, USA

Department of Agriculture (USDA) soil rating values to estimate maize and soybean yield at multiple locations in Maryland, USA. Using Tavg and Rain of June, July, and August with mid-May planting progress, Westcott and Jewison (2013) developed yield models for maize and soybean in the USA. The effects of delayed planting from unfavorable weather conditions at the beginning of the growing season were accounted in their model. However, mid-May planting progress, as measured by the percent of maize plantings by mid-May, can vary within a county and its effect can be difficult to estimate. In addition, the June Rain in their model was used as the deviation amount from normal and only in the years when the June Rain was lower than 10% tail of its distribution. Data on such weather variable are not always easily available and requires calculation before it can be used in the yield estimation model. Mathieu and Aires (2016, 2018) compared more than 50 agro-climatic indices and different weather-impact models for maize yield estimation in the USA. They found that the Standardized Precipitation Evapotranspiration Index (SPEI) and Tavg in July were the best predictors. Derivation of agro-climatic indices such as SPEI, however, requires data of several other parameters, e.g., potential evapotranspiration, which may not be readily available. Apart from yield estimation, weather-based yield models have also been used to assess crop yield sensitivity to weather variation (Cai et al. 2014; Mourtzinis et al. 2015) and to predict yield in future or altered climatic conditions (D'Agostino and Schlenker 2016; Verón et al. 2015).

Weather-based modeling is one of several methods for crop yield estimation. Broadly, this method can be considered a statistical approach for yield estimation, in which empirical regression equations are developed between yield and other predictors such as several weather parameters collected across multiple site-years (Mathieu and Aires 2016; Thompson 1969). In weather-based models, weather data such as Tavg, Rain, or agroclimatic indices (e.g., Palmer drought severity index, SPEI, and growing degree days) are used as predictors in regression equations. In recent decades, with the advancement of remote sensing tools, several spectral indices such as normalized difference vegetation index (Kriegler et al. 1969; Rouse et al. 1974) and enhanced vegetation index (Huete et al. 2002) are also being used to develop crop yield estimation models (Franch et al. 2019; Prasad et al. 2006). Other approaches for crop yield estimation include field observation/sampling and crop simulation modeling. The field observation/sampling is based on careful observation or destructive measurements of representative crop samples. For example, the USDA estimates in-season crop yield based on grower-reported surveys and field-measurement surveys. These surveys are designed to collect data on cultivated area, crop density, row spacing, and yield components (USDA-NASS 2012). In crop simulation modeling, computer-based models are run with soil, weather, cultivar, and crop management information in order to estimate crop yield (Lobell and Asseng 2017; Morell et al. 2016). Crop models ranging in complexity and data requirements have been used in scientific research and as decision support tools (Jame and Cutforth 1996; Jones 1993; Setiyono et al. 2011).

Weather-based yield estimation models have several advantages over field observation/sampling and crop simulation modeling approaches. Field observation/sampling often entails trained personnel to collect representative samples from the field for meticulous measurements (USDA-NASS 2012). Therefore, the field observation/sampling approach can be time consuming and expensive. While crop modeling can be advantageous for estimating crop yield from the sub-field to national level (Fraisse et al. 2001; Morell et al. 2016; van Wart et al. 2013), it requires several soil and crop parameters for calibration and evaluation (Mathieu and Aires 2016; Verón et al. 2015). In contrast, weather-based models rely only on weather parameters or comparatively fewer parameters. Therefore, these models can be advantageous in terms of cost and scalability in both temporal and spatial domains. In previous studies, weather-based models have been used in yield estimation ranging from short term, such as the current year, to long term, such as at the mid or end of the century under future climate change scenarios (Chen et al. 2019; Schlenker and Roberts 2009). Similarly, regarding spatial scale, these models have been applied at the field, regional, and global levels (Lobell et al. 2008; Powell and Reinhard 2016; Ray et al. 2015). Despite these comparative advantages, weather-based models have some limitations. One of these arises from the nature of space-time data, which are often spatially and temporally correlated (Mathieu and Aires 2016; Verón et al. 2015). In addition, collinearity and interaction among the weather variables can make such data difficult to model (Shi et al. 2013). Moreover, the spatial and temporal levels of data used to develop a model can strongly influence the accuracy of yield estimation (Lobell and Burke 2010; Mathieu and Aires 2016). Even with these limitations, weather-based models or statistical models in general, are still widely used by the scientific community (Lobell and Burke 2010; Mathieu and Aires 2016; Shi et al. 2013).

For model development, multiple linear regression has been widely used, but is gradually being replaced by other more advanced approaches such as mixed-effects models (Mathieu and Aires 2016) and general additive models (Chen et al. 2019). Recent studies have also explored the potential of machine learning approaches, such as random forest model (Jeong et al. 2016), support vector machine regression (Oguntunde et al. 2018), and neural network (Crane-Droesch 2018). As the availability of weather data has increased with improvement in temporal frequency and spatial resolution of data, more advanced statistical models provide the opportunity to develop more accurate crop yield estimation models. Most studies have used Rain, Tavg, solar radiation, and vapor pressure on monthly or seasonal levels as predictors (Lobell et al. 2014; Mathieu and Aires 2016; Schlenker and Roberts 2009; Tack et al. 2015; Verón et al. 2015). A limited number of studies have used weather data at daily, weekly, and biweekly timescales and with more advanced statistical models for crop yield estimation (Kaul et al. 2005; Tack et al. 2015). While the difference between daily maximum and minimum air temperature (Tdiff) has been shown to significantly affect crop yield, its use in yield estimation models is less frequent (Hu et al. 2003; Lobell 2007; Tack et al. 2015; Verón et al. 2015). With relatively greater increment in daily minimum air temperature as compared to maximum air temperature, studies have shown a decrease in Tdiff on a global scale in the last 50 years (Lewis et al. 2013; Vose et al. 2005). However, the role of Tdiff in maize and soybean yield estimation model development is poorly understood. Understanding the implication of Tdiff on maize and soybean production is crucial for designing resilient cropping systems in the US Corn Belt and global food security, in general. The overarching goal of this study is to assess yield estimation models for maize and soybean grown in the US central Corn Belt based on the hypothesis that in rain-fed agricultural systems of the region, in-season weather data can be used to estimate maize and soybean yield. The specific objectives are to (1) compare the accuracy of multiple linear regression (MLR), general additive model (GAM), and support vector machine (SVM) regression model to estimate maize and soybean yields and (2) evaluate if inclusion of Tdiff in the model improves model performance.

Materials and methods

Study period and study area

Historical weather data and grain yield of maize and soybean for a period of 28 years (1990–2017) were included for this study. The study year was selected until 2017 since the daily weather data at county level were available until 2017 at the time of data analysis. The study was limited up to 1990 only as it allowed considering uniform effects from the genetic, agronomic, and technological advancements in crop yield over the study period. The study area focused on the central Corn Belt region of the USA and included the major rain-fed maize and soybean agricultural districts of Iowa (IA), Illinois (IL), Indiana (IN), and Minnesota (MN). These four states accounted for almost 50 and 44% of total US maize and soybean grain production in 2018 (USDA-NASS 2019). The western Corn Belt region has a more irrigated production system and, therefore, was not considered for this study. From IA, IL, IN, and MN, only those agricultural districts with non-irrigated maize and soybean production were included. The number of counties included from IA, IL, IN, and MN were 68, 39, 33, and 29, respectively (Fig. 1). Altogether, the study involved a 28-year period and 169 counties from 16 agricultural districts in four states comprising a total data set of 4732 site-years.

Data collection

Daily data on Tavg and Rain were obtained at the county level from the fine-scaled weather data set for contiguous USA developed by Schlenker and Roberts (2009) and Schlenker (2018) based on the PRISM Climate Group (2018) weather data. From the difference between daily maximum and minimum air temperatures, Tdiff was calculated. The data were prepared at weekly, biweekly, and monthly timescales.

County-level annual grain yield data of maize and soybean for the study period were obtained from USDA-NASS (2019) using “nass_data” function from “nassR” package (Dinterman and Eyer 2018) in R (R Core Team 2018).

Model development

Yield estimation models were developed for each crop at the total study area level and at individual state level. For model development, in-season (May–August) weather data of Tavg and Rain with/without Tdiff were used as weather variables. Three different statistical models were assessed, which included MLR, GAM, and SVM. Each model was developed with weather data at weekly, biweekly, and monthly timescales.

First, time-series yield data were linearly detrended for each of the 169 counties to remove non-weather effects such as advancement in technology, improved agronomic practices, and cultivar changes over years. In this study, we used US maize and soybean production data from 1990 to 2017, a relatively short period during which yield increase was largely linear, so the detrending approach used was linear. Linear detrending is a common approach for trend removal in crop yield-climate relationship studies, especially for the removal of yield increment in a shorter time period. Other frequently used detrending approaches are first differencing and non-linear regressions. After trend removal, MLR, GAM, and SVM yield estimation models were developed.

In MLR, the mean response variable, y, was modeled as a linear function of n explanatory variables x_1, x_2, x_3, x_n, and can be written as

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n + \epsilon \tag{1}$$

where b refers to a regression coefficient and ϵ is the residual. The regression coefficient in MLR is estimated based on the ordinary least square regression that minimizes the sum of squared errors. For this study, MLR was modeled as

$$y_{i,j} = b_0 + \sum(\beta.X)_{i,j} + \epsilon_{i,j} \tag{2}$$

where $y_{i,j}$ denotes the estimated yield of the ith county in the jth year, b_0 refers to the intercept or mean yield, and β denotes a vector of regression coefficients for X vector of weather variables, namely Tavg, Rain, and Tdiff from May to August.

In GAM, the linear relationship as described in MLR was modeled as a smooth function in order to capture any nonlinearities in the model (Hastie and Tibshirani 1990), and can be written as

$$y = b_0 + f_1 x_1 + f_2 x_2 + f_3 x_3 + \cdots + f_n x_n + \epsilon \tag{3}$$

where f refers to smooth functions on explanatory variable x, which is estimated using a scatterplot smoother (Hastie and Tibshirani 1990). For this study, GAM was modeled as

$$y_{i,j} = b_0 + \sum \ (F.X)_{i,j} + \epsilon_{i,j} \tag{4}$$

where F denotes a vector of smoother functions for X vector of weather variables as mentioned above.

In SVM, a regression function is set up from a training dataset with a margin of tolerance defined by ϵ as the ϵ-insensitive zone. The SVM ignores any error from data points within this margin of tolerance. Non-negative slack variables outside this margin are used to measure the deviation of training samples. The goal is to find an optimal hyperplane by choosing the function that minimizes the deviation from the insensitivity parameter (Cortes and Vapnik 1995; Vapnik 1999). In a given dataset of (x_i, y_i) pairs, where i refers to 1,.., n observations, a linear function in SVM can be expressed as

$$y = f(x) = \omega.x + b \tag{5}$$

where ω. x denotes the product between weighing vector ω and input vector x, and b refers to the bias term. The objective function in regression analysis using SVM is to minimize ω. Errors less than ϵ (tolerance margin) are not considered for the objective function, and only the non-negative slack variables across the tolerance margin or ϵ-insensitive zone are considered (Vapnik 1999). In nonlinear regression, the data are first linearized through incorporation of kernel functions to apply linear functions. Some of the commonly used kernels include polynomial, radial basis function, and sigmoid. In this study, radial basis function kernel was used for SVM.

All data analyses were performed in R v. 3.5.1 (R Core Team 2018) using the caret package (Kuhn 2008). The datasets were randomly allocated for training (75% of total data) and testing (25% of total data) for model development and evaluation, respectively. For MLR and GAM, data were transformed using the BoxCox transformation (Box and Cox 1964), whereas for SVM, the data were centered and scaled before processing. Each model was trained using fivefold cross validation with five repetitions. For the MLR model, the "lmStepAIC" method was used for stepwise selection of variables and selection of the model with the lowest AIC value. For the GAM model, the "gam" method from the "mgcv" package (Wood 2018) was used. For the SVM model, the "svmRadial" method from the "kernlab" package (Karatzoglou et al. 2018) was used. The MLR did not have tuning parameters. For GAM, the smoothing parameter estimation method was set to "REML" and the selection type to "TRUE," which penalizes for new terms in the model and allows removal of terms (Wood 2017, 2018). For SVM, the sigma and cost parameters were tuned using the "tuneLength" command, which was set to test 15 default values for each parameter.

Variable importance

The relative importance of weather variables in estimating maize and soybean yields was determined using the "varImp" function in the caret package (Kuhn 2008) in R (R Core Team 2018). The relative importance was determined by using a locally weighted least square regression (loess) smoother between crop yield and each weather variable. The R^2 value from each model was compared with the intercept-only null model, which was used to determine the relative importance of weather variables. Only the weather variables grouped at monthly scale were used for this purpose.

Model evaluation

An independent dataset (not used for model development) was used for model evaluation. Model fitness was evaluated using the root mean square error (RMSE), normalized RMSE (nRMSE), correlation coefficient (r), and Willmott's index of agreement (d) (Willmott 1981).

The RMSE between the estimated and actual yield was calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(E_i - A_i\right)^2}{n}} \tag{6}$$

where n refers to the number of observations, and E and A denote estimated and actual yield. Then, nRMSE was calculated as $RMSE$/mean(actual yield). Lower RMSE and nRMSE values indicate better model performance.

The r ranges from − 1 to + 1, with − 1 indicating perfect negative linear correlation and + 1 indicating perfect positive linear correlation between estimated and actual yields. It was calculated using the following equation:

$$r = \frac{\sum_{i=1}^{n}\left(A_i - \overline{A}\right)\left(E_i - \overline{E}\right)}{\sqrt{\sum_{i=1}^{n}\left(A_i - \overline{A}\right)^2}\sqrt{\sum_{i=1}^{n}\left(E_i - \overline{E}\right)^2}} \tag{7}$$

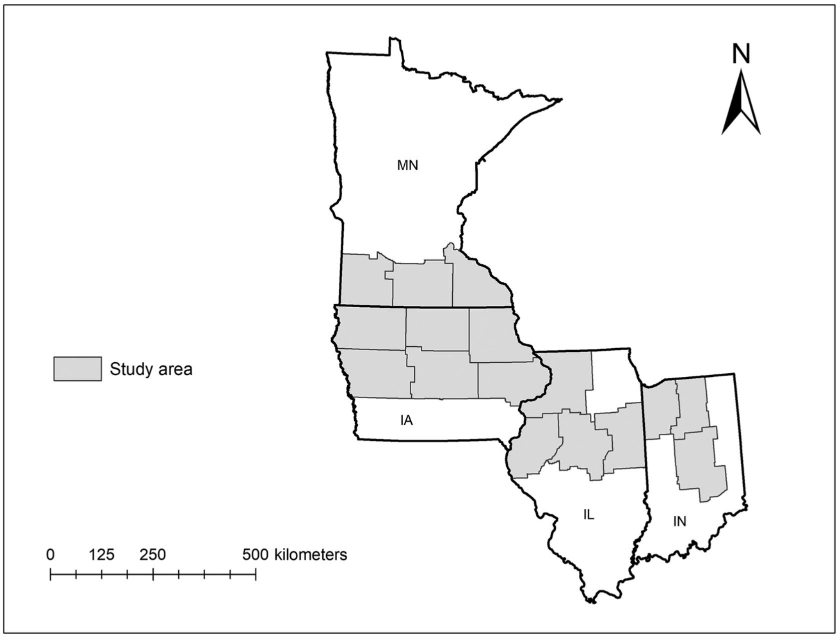


Fig. 1 Map of the US states and the agricultural districts included in this study. In the map, MN, IA, IL, and IN refer to Minnesota, Iowa, Illinois, and Indiana, respectively

The d varies between 0 and 1, with 1 indicating a perfect match and 0 indicating no match between estimated and actual yields and was calculated using the following equation:

$$d = 1 - \frac{\sum_{i=1}^{n} \left(A_i - E_i\right)^2}{\sum_{i=1}^{n} \left(\left| E_i - \overline{A} \right| + \left| A_i - \overline{A} \right|\right)^2} \quad (8)$$

Results

Average maize and soybean yield and yield deviation

Over the total study area, the average yield for maize and soybean was 11,950 and 3680 kg ha^{-1}, respectively (Table 1). The coefficient of variation (CV) was 12% for maize and 11.3% for soybean. For maize, the highest average yield was from MN, followed by IA, IL, and IN. The CV ranged from 9.1 to 13.3% across states. For soybean, the highest average yield was in IL, followed by IA. In IN and MN, the soybean yield was similar at around 3580 kg ha^{-1}. The CV for soybean yield was 10.4–11.5% in all states.

Considerable maize and soybean yield loss occurred in 1993, to a greater extent in IA and MN, due to flooding (Phillips 1994). Lower-than-average yield in soybean was observed in 2003, and can be attributed to dry conditions during August and September (Brumm and Hurburgh 2003). In all states, except MN, the highest negative maize and soybean yield deviation from average occurred in 2012, a historic drought year that caused a significant reduction in US crop yields (USDA-NASS 2013) (Figs. 2 and 3). In 1994, 2016, and 2017, comparatively, yield in all states was above average ranging from 293 to 1635 kg ha^{-1} in maize and from 49 to 508 kg ha^{-1} in soybean (Figs. 2 and 3).

Model development

The SVM model predicted the maize and soybean yield with the lowest RMSE and therefore, performed the best followed by GAM and MLR at all temporal levels of weather data (Fig. 4). Inclusion of Tdiff in the model improved yield estimation for both crops; however, the magnitude of improvement varied with model and temporal level of weather data. With the SVM model for maize, inclusion of Tdiff at weekly level did not improve the yield estimation. Similarly, adding Tdiff in the MLR model did not improve yield estimation at monthly level. However, estimation improved significantly ($P < 0.05$) in all other cases. For soybean, inclusion of Tdiff in the SVM model did not improve the estimation until monthly level. Adding Tdiff to the GAM and MLR models significantly ($P < 0.05$) improved yield estimation at all levels.

Within SVM models for maize, no significant difference in yield estimation occurred between weekly and biweekly models; yet, the RMSE from both was significantly lower compared to the monthly model (Fig. 3). Similar results were obtained for soybean yield estimations. This reveals the greater efficiency and superior performance of the SVM model compared to the GAM and MLR models. At the state level, the SVM model produced the lowest RMSEs as shown in Figs. S1–S8 (Electronic Supplementary Material).

Variable importance

In estimating maize yield across the study area, rainfall during June and July and temperature during July and August were more important than other monthly weather variables (Fig. 5). The Tavg and Tdiff during July and August were both important. Rainfall during May and August, which coincides with early vegetative and late reproductive stages in maize, respectively, were comparatively less important in estimating maize yield.

In estimating soybean yield, Tavg during June, when plants are in early reproductive stages, was the most important weather variable; rainfall during August was the second-most important variable (Fig. 5). As in maize, Tdiff during July and August was important for soybean yield estimation.

Model evaluation

The best-performing models based on the lowest RMSE (Fig. 4) were used to estimate maize and soybean yield using an independent test dataset that was not used previously for model development (Fig. 6). As in model development, the SVM model outperformed the other models. The SVM model with weekly weather data of only Tavg and Rain as predictors estimated maize yield with the lowest RMSE of 591 kg ha^{-1} (4.9% *n*RMSE). The *r* and *d*-index were also the highest for the SVM model. After the SVM model, the lowest RMSE was obtained from the GAM model (792 kg ha^{-1}), followed by the MLR with RMSE of 1065 kg ha^{-1} (Fig. 6). The GAM and MLR models used weekly weather data of Tavg, Rain, and Tdiff.

As in maize, the SVM model outperformed the other models in soybean yield estimation. The SVM model with weekly Tavg and Rain estimated soybean yield with the lowest RMSE of 205 kg ha^{-1} (5.5% *n*RMSE). Like in maize, the *r* and *d*-index were also highest for the SVM model. Similarly, the next lowest RMSE was obtained from the GAM model (274 kg ha^{-1}) and the MLR model produced the highest RMSE of 320 kg ha^{-1} (Fig. 7).

Yield estimation at state level

Models were also evaluated for both crops at individual state level. Since the SVM models gave the lowest RMSEs during model development (Figs. S1–S8, Electronic Supplementary Material), only those models were used for evaluation at the state level.

Overall, maize yield was estimated with a RMSE lower than 800 kg ha^{-1} in all states (Fig. 8). The lowest RMSE of 472 kg ha^{-1} (3.8% *n*RMSE) was obtained for IA, followed by MN with 505 kg ha^{-1} (4.2% *n*RMSE), IL with 660 kg ha^{-1} (5.4% *n*RMSE), and IN with 733 kg ha^{-1} (6.8% *n*RMSE). The *d*-index values were above 0.9 in all states. Except for IN, the values for *r* were equal to or greater than 0.9 in other states.

In all states, soybean yield was estimated with RMSE lower than 250 kg ha^{-1} (Fig. 9). The lowest RMSE for a soybean yield of 178 kg ha^{-1} (4.8% *n*RMSE) was obtained for IA, followed by IL with 197 kg ha^{-1} (5.1% *n*RMSE), MN with 199 kg ha^{-1} (5.7% *n*RMSE), and IN with 219 kg ha^{-1} (6.1% *n*RMSE). As for maize, the *d*-index values for soybean were also above 0.9 in all states. The values of *r* were 0.9, 0.86, 0.84, and 0.89 for IA, IL, IN, and MN respectively.

Discussion and conclusions

Crop yield estimation is an important factor in the decision-making process for farm management. In this study, weather-based models for maize and soybean yield estimation were developed using easily available weather data of air temperature and rainfall. Weather variables show varied degree of correlation. For example, we found that average temperatures in June and July are positively correlated, whereas average air temperature and rainfall in June are negatively correlated (Fig. S9). Still, collinearity primarily influences the coefficients of a parameter and associated level of significance, but does not inhibit the predictive potential of the model (Neter et al. 1996). Additionally, covariance among weather variables in our training and test data is assumed to be similar as it is randomly allocated and large enough.

The SVM model clearly outperformed the GAM and MLR models. Using only Tavg and Rain data at the weekly level, the SVM model estimated county average maize yield with less than 5% *n*RMSE and soybean yield with less than 6% *n*RMSE (Figs. 6 and 7). Comparatively better yield prediction from the SVM than MLR has been reported in other studies as well. Chen et al. (2016) found that the SVM model was more accurate at predicting rice yield from weather variables than the MLR model. Similarly for maize, better model predictions with the SVM model compared to the MLR were reported by Karimi et al. (2008). Superior performance of the SVM model can be attributed to its ability to model nonlinear functions and high dimensional data (Vapnik 1999). Crop yield data in relation to weather parameters are nonlinear in nature (Lobell et al. 2011; Schlenker and Roberts 2009; Tack et al. 2015). Therefore, unsurprisingly, the MLR model which only modeled the linear relationship between weather variables and yield performed the worst in this study (Figs. 6 and 7). The GAM model, despite being able to model nonlinear relationships, did not give better estimations compared to the SVM model, but were better than those with the MLR model.

This study demonstrated the importance of Tdiff in maize and soybean yield estimation. Although the lowest RMSEs were obtained from models with Tdiff, the differences from models with and without Tdiff were not always significant

Table 1 Total study area and state average detrended yield and coefficient of variation (CV) of maize and soybean during the study period (1990–2017)

State	Maize		Soybean	
	Average yield (kg ha^{-1})	CV (%)	Average yield (kg ha^{-1})	CV (%)
Iowa	12,260	9.8	3700	11.1
Illinois	12,160	12.2	3830	10.4
Indiana	10,740	13.3	3580	10.9
Minnesota	12,300	9.1	3580	11.5
Total area	11,950	12.0	3680	11.3

(Fig. 4). Improvement in yield estimates with Tdiff were dependent on the temporal level of weather data, statistical model, and crop. Regardless of the temporal level of weather data, inclusion of Tdiff always significantly improved yield estimation with the GAM model (Fig. 4). Inclusion of Tdiff in the SVM model at the weekly level for maize and the weekly and biweekly levels for soybean did not improve yield estimates. However, including Tdiff in the model at the

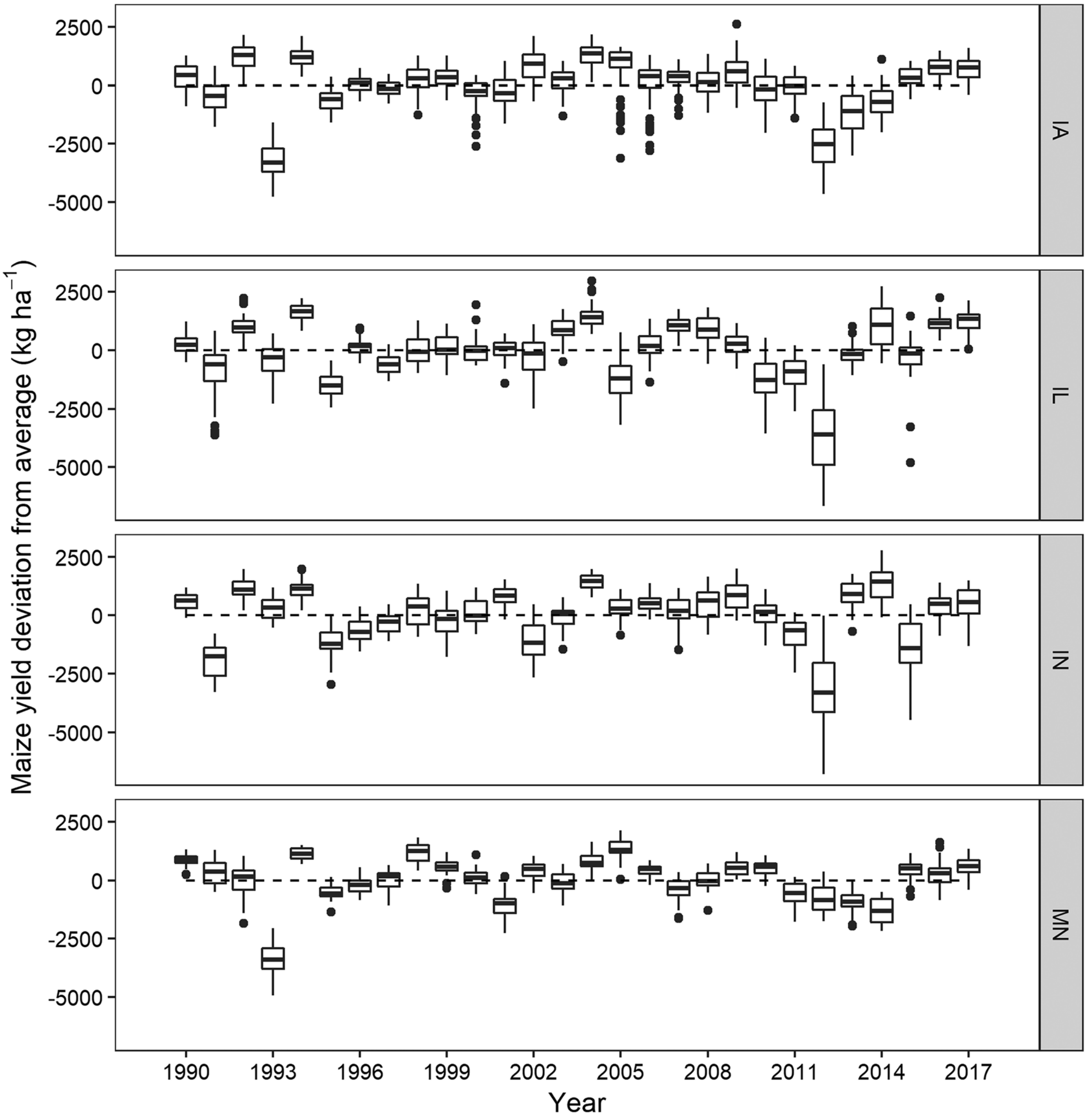


Fig. 2 Maize yield deviation from average in Iowa (IA), Illinois (IL), Indiana (IN), and Minnesota (MN) from 1990 to 2017. The dotted black line represents the zero-reference line

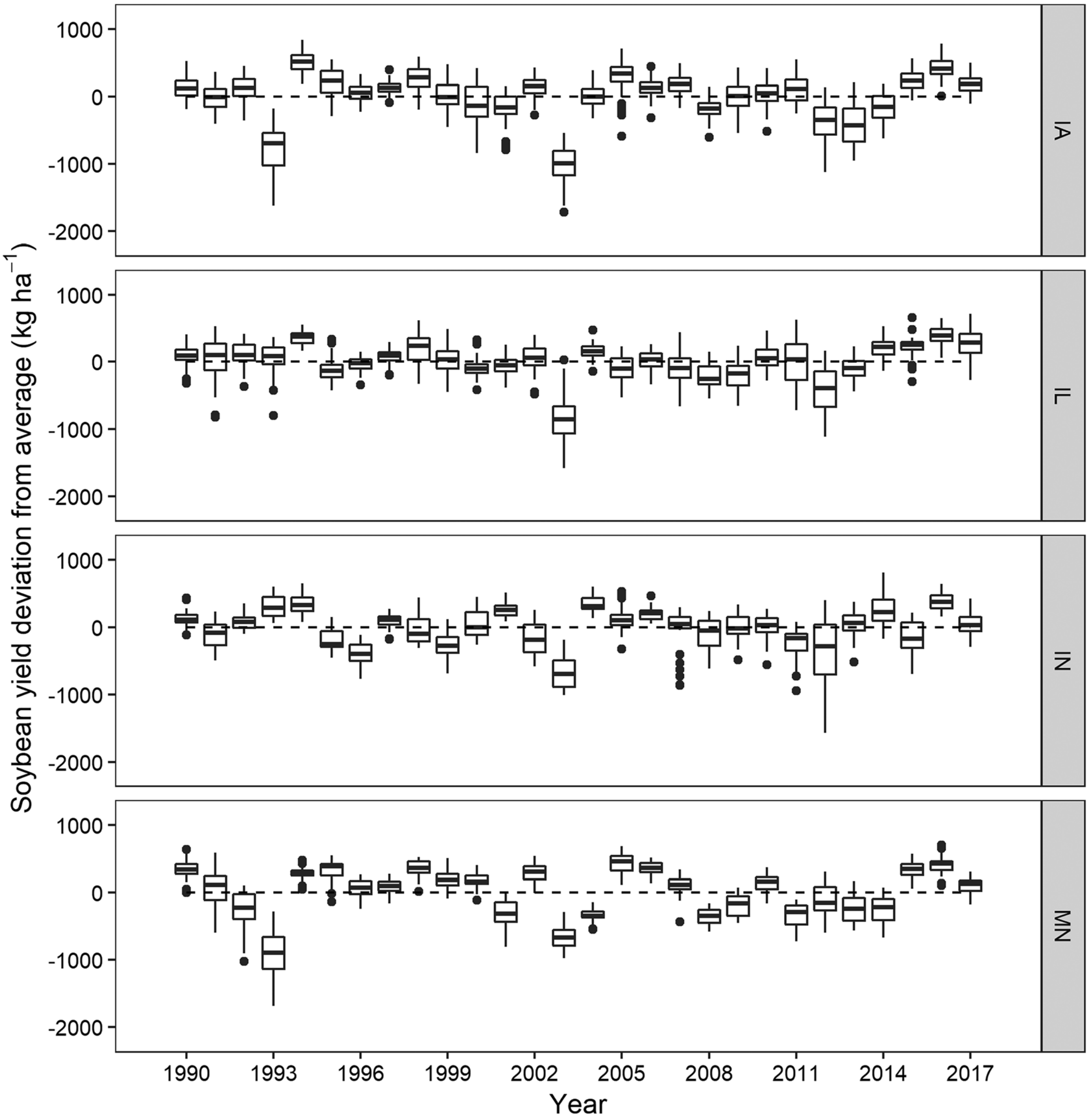


Fig. 3 Soybean yield deviation from average in Iowa (IA), Illinois (IL), Indiana (IN), and Minnesota (MN) from 1990 to 2017. The dotted black line represents the zero-reference line

monthly level produced significantly lower RMSE for both crops. Yield estimates for maize and soybean obtained from monthly weather data with the SVM model were more accurate (lower RMSEs) than those from weekly and biweekly weather data from the GAM and MLR models.

The variable importance analysis showed that Tdiff played a key role in model development (Fig. 5). For maize, Tavg and Tdiff during July and August were important weather variables. Similarly, Tdiff during August and July were important for soybean. This highlights the important implication of Tdiff on maize and soybean yield. Lesser Tdiff owing to higher night temperature has been reported to reduce maize yield (Cantarero et al. 1999; Chang 1981; Peters et al. 1971) and soybean yield (Peters et al. 1971). Higher night temperature, especially during the reproductive phase, can reduce crop yield by increasing respiration and the duration of the reproductive phase (Cantarero et al. 1999; Chang 1981). Increased respiration reduces stored photo-assimilates in crops, thereby reducing their translocation to grains. Higher values of Tdiff have also been shown to reduce crop yields. Lobell (2007) found a nonlinear response of Tdiff to cereal grain yield and a negative response of maize yield to increased Tdiff. This finding can have an important implication on crop production and climate change studies. Most studies on climate change have focused primarily on the effects of increased air temperature, but the difference between maximum and minimum air temperatures has received less attention. Results from this study as well as similar results from past studies (Cantarero

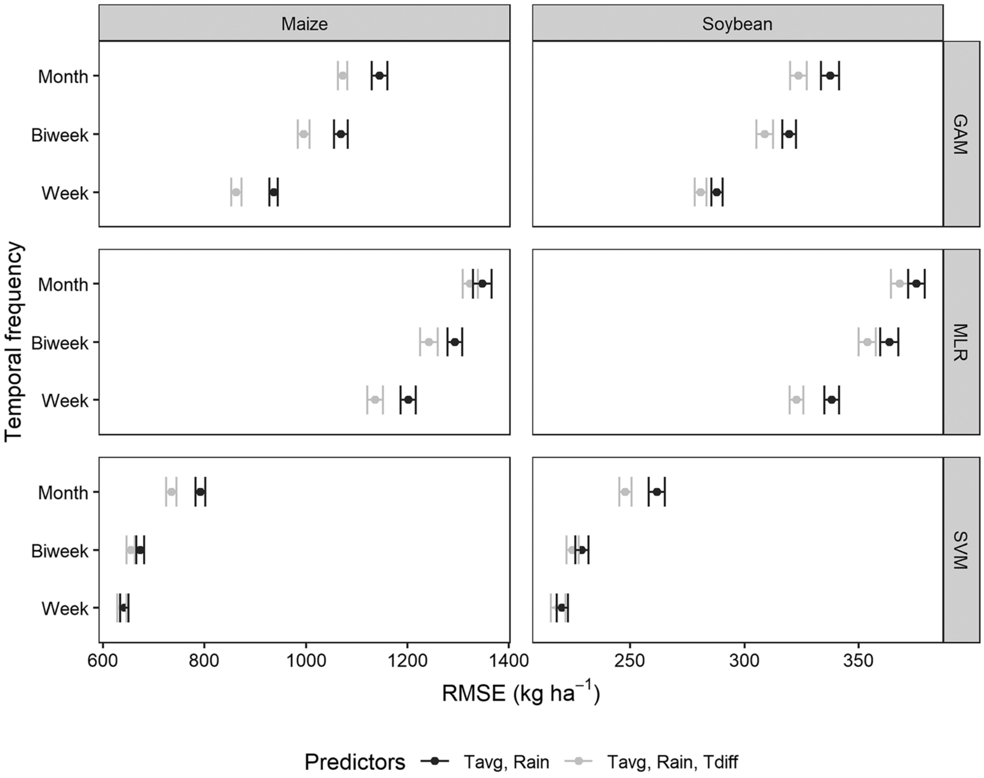


Fig. 4 Root mean square error (RMSE) of maize and soybean yield estimation during model development for the generalized additive model (GAM), multiple linear regression (MLR), and support vector machine (SVM) models trained with weather predictors at weekly, biweekly, and monthly timescales

et al. 1999; Chang 1981; Lobell 2007) suggest that Tdiff is important to consider while assessing the impact of climate change on maize and soybean production. As the Tdiff has been decreasing on a global scale (Vose et al. 2005), it is critical to assess its impacts on future maize and soybean production in the US Corn Belt. A decrease in Tdiff due to warmer nights can be advantageous in some regions with too low air temperature as in northern MN but can be damaging to crops in regions with high daily air temperature near optimal.

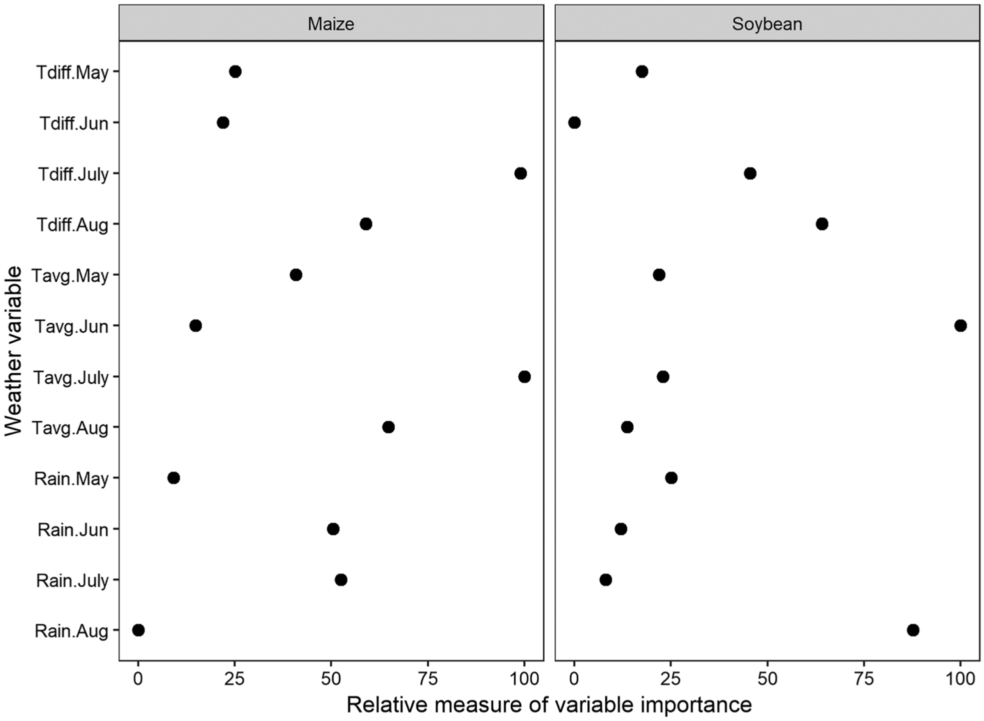


Fig. 5 Relative measure of monthly weather variable importance scores in estimating maize and soybean yield across Iowa, Illinois, Indiana, and Minnesota. (Importance scores were normalized and scaled between 0 and 100)

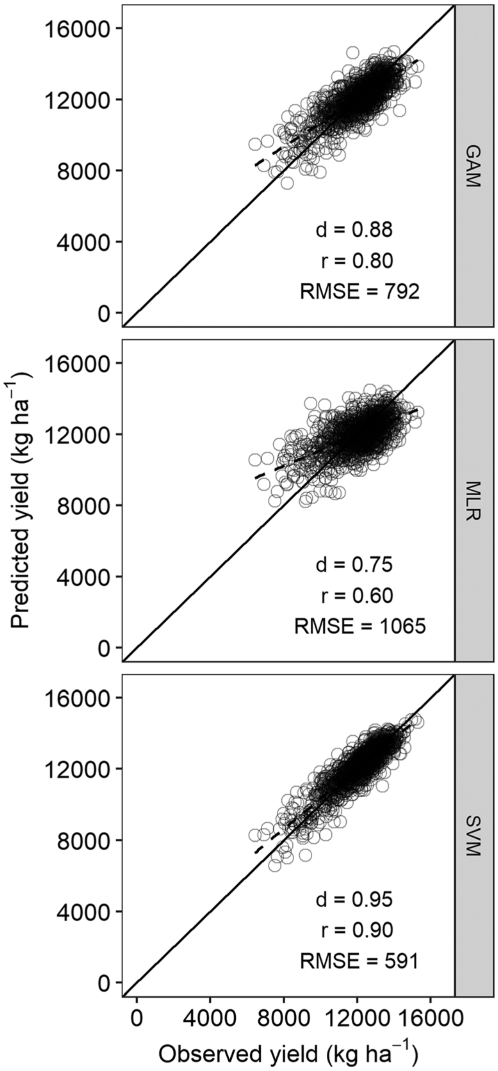


Fig. 6 Scatterplots of observed versus predicted maize yield from the generalized additive model (GAM), multiple linear regression (MLR), and support vector machine (SVM) models. The diagonal black line is the 1:1 line. The dashed black line represents the linear regression between observed and predicted yields

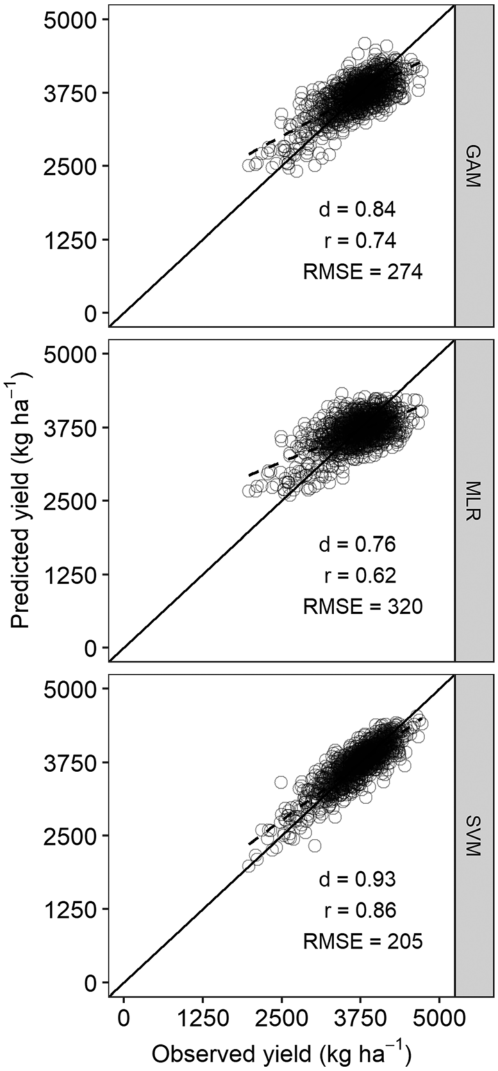


Fig. 7 Scatterplots of observed versus predicted soybean yield from the generalized additive model (GAM), multiple linear regression (MLR), and support vector machine (SVM) models. The diagonal black line is the 1:1 line. The dashed black line represents the linear regression between observed and predicted yields

Understanding such implications of weather conditions on crop production will help to design cropping systems that are resilient to future weather situations.

This study demonstrates the usefulness of weather data for reliable estimation of maize and soybean yield in the US central Corn Belt. This study also highlights the comparative significance of weather variables in estimating maize and soybean yield. Improvements in weather-based crop yield estimation are highlighted with respect to different statistical models and weather variables. The comparative analyses of model performances indicated that the support vector machine model with radial kernel is better as compared to multiple linear regression and general additive model in estimating maize and soybean yields. The inclusion of the difference in daily maximum and minimum air temperature as an explanatory variable improved the performance of all three models assessed. However, such improvement was shown to be dependent on the temporal level of weather data, statistical model, and crop type. Due to inherent micro-environments in the agricultural landscape, future research should investigate the performance of such weather-based models in explaining spatial crop yield variability occurring within the field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00484-020-02039-z.

Acknowledgments The authors would like to thank the two anonymous reviewers for their insightful suggestions and comments on the manuscript. We appreciate the financial support from the University of Minnesota, Saint Paul, MN.

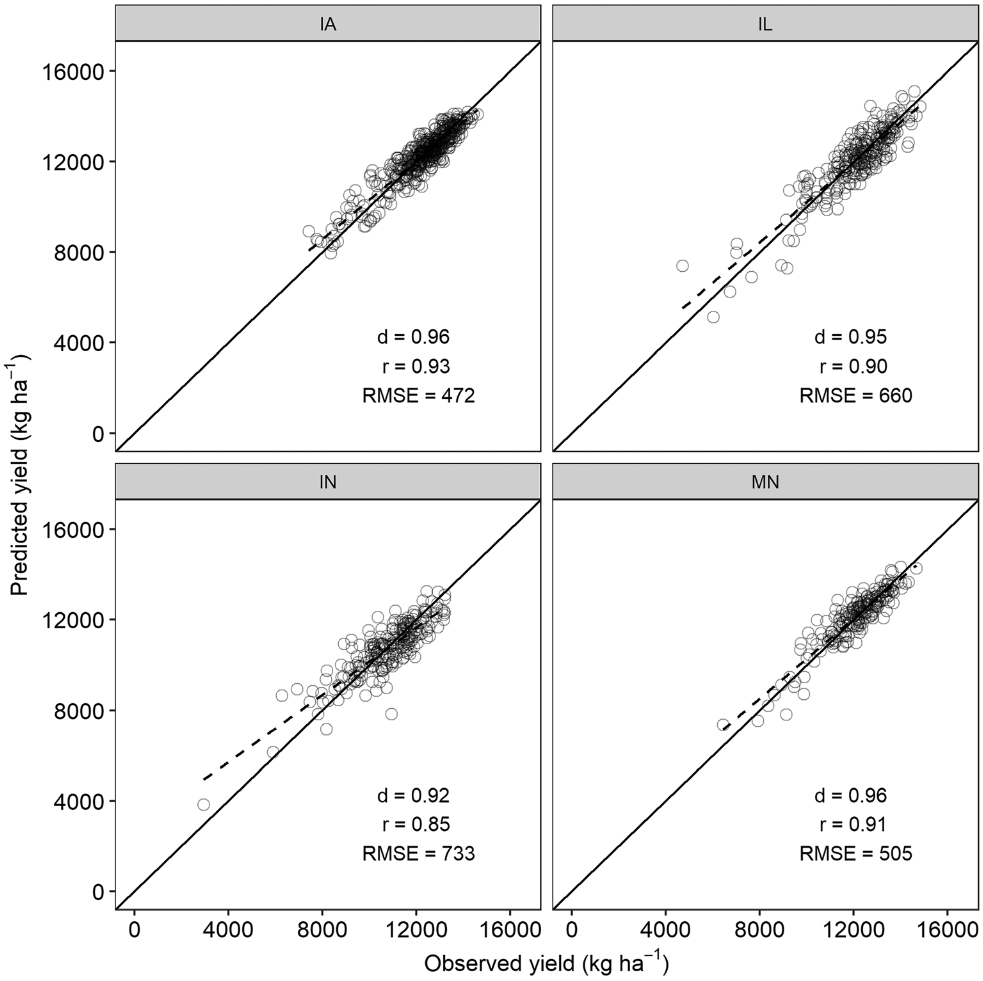


Fig. 8 Scatterplots of observed versus predicted yield of maize using support vector machine model in Iowa (IA), Illinois (IL), Indiana (IN), and Minnesota (MN). The diagonal black line shows the 1:1 line. The dashed black line represents the linear regression between observed and predicted yields

Funding This study was supported with funds provided by the University of Minnesota, Saint Paul, MN.

Data availability All data analyzed during the study are included in this published article and its supplementary information files.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Code availability All data analyses were carried out in R software for statistical computing. All R packages used for data analyses are mentioned in the “Materials and methods” section of this published article. R codes are available from the corresponding author on reasonable request.

References

Basso B, Cammarano D, Carfagna E, (2013) Review of crop yield forecasting methods and early warning systems. In: Report presented to first meeting of the scientific advisory committee of the global strategy to improve agricultural and rural statistics. FAO, U.N. Rome, Italy, 18 - 19 July

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

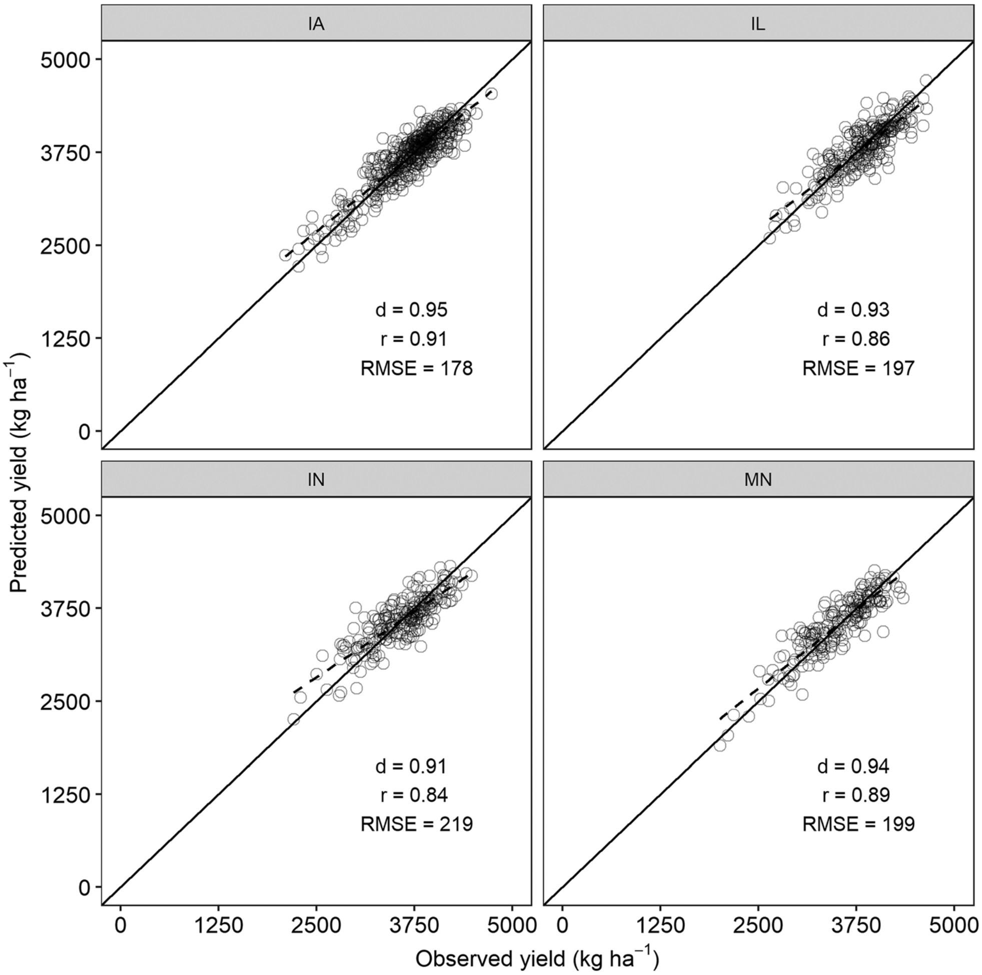
Brumm TJ, Hurburgh CRJ (2003) Quality of the 2003 soybean crop in the United States. Agricultural and biosystems engineering technical reports and white papers-19, Iowa State University

Cai R, Yu D, Oppenheimer M (2014) Estimating the spatially varying responses of corn yields to weather variations using geographically weighted panel regression. J Agric Resour Econ 39:230–252

Cantarero MG, Cirilo AG, Andrade FH (1999) Night temperature at silking affects set in maize. Crop Sci 39:703–710. https://doi.org/10.2135/cropsci1999.0011183X003900020017x

Chang J (1981) Corn yield in relation to photoperiod, night temperature, and solar radiation. Agric Meteorol 24:253–262. https://doi.org/10.1016/0002-1571(81)90049-2

Fig. 9 Scatterplots of observed versus predicted yield of soybean using support vector machine model in Iowa (IA), Illinois (IL), Indiana (IN), and Minnesota (MN). The diagonal black line shows the 1:1 line. The dashed black line represents the linear regression between observed and predicted yields



Chen H, Wu W, Liu H-B (2016) Assessing the relative importance of climate variables to rice yield variation using support vector machines. Theor Appl Climatol 126:105–111. https://doi.org/10.1007/s00704-015-1559-y

Chen K, O'Leary RA, Evans FH (2019) A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool. Agric Syst 173:140–150. https://doi.org/10.1016/J.AGSY.2019.02.009

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20: 273–297. https://doi.org/10.1007/BF00994018

Crane-Droesch A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ Res Lett 13:114003. https://doi.org/10.1088/1748-9326/aae159

D'Agostino AL, Schlenker W (2016) Recent weather fluctuations and agricultural yields: implications for climate change. Agric Econ 47:159–171. https://doi.org/10.1111/agec.12315

Dinterman R, Eyer J (2018) nassR: an alternative for downloading various USDA data from Quick Stats through R

Fraisse CW, Sudduth KA, Kitchen NR (2001) Calibration of the CERES-Maize model for simulating site-specific crop development and yield on claypan soils. Appl Eng Agric 17:547–556. https://doi.org/10.13031/2013.6453

Franch B, Vermote EF, Skakun S, Roger JC, Becker-Reshef I, Murphy E, Justice C (2019) Remote sensing based yield monitoring: application to winter wheat in United States and Ukraine. Int J Appl Earth Obs Geoinf 76:112–127. https://doi.org/10.1016/J.JAG.2018.11.012

Franzluebbers A, Steiner J, Karlen D, Griffin T, Singer J, Tanaka D (2011) Rainfed farming systems in the USA. In: Rainfed farming systems. Springer Netherlands, Dordrecht, pp 511–560. https://doi.org/10.1007/978-1-4020-9132-2_20

Green TR, Kipka H, David O, McMaster GS (2018) Where is the USA Corn Belt, and how is it changing? Sci Total Environ 618:1613–1618. https://doi.org/10.1016/j.scitotenv.2017.09.325

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. In: Monographs on statistics and applied probability, 1st edn. Chapman and Hall/CRC, Boca Raton

Hu Q, Buyanovsky G, Hu Q, Buyanovsky G (2003) Climate effects on corn yield in Missouri. J Appl Meteorol 42:1626–1635. https://doi.org/10.1175/1520-0450(2003)042<1626:CEOCYI>2.0.CO;2

Huete A, Didan K, Miura T, Rodriguez E, Gao X, Ferreira L (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 83:195–213. https://doi.org/10.1016/S0034-4257(02)00096-2

Jame YW, Cutforth HW (1996) Crop growth models for decision support systems. Can J Plant Sci 76:9–19. https://doi.org/10.4141/cjps96-003

Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ, Shim K-M, Gerber JS, Reddy VR, Kim S-H (2016) Random forests for global and regional crop yield predictions. PLoS One 11: e0156571. https://doi.org/10.1371/journal.pone.0156571

Jones JW (1993) Decision support systems for agricultural development. Springer, Dordrecht, pp 459–471. https://doi.org/10.1007/978-94-011-2840-7_28

Karatzoglou A, Smola A, Hornik K (2018) kernlab (v 0.9-27)-An S4 package for kernel methods in R

Karimi Y, Prasher SO, Madani A, Kim S, Madani SO, Kim A (2008) Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. Can Biosyst Eng 50:1–8

Kaul M, Hill RL, Walthall C (2005) Artificial neural networks for corn and soybean yield prediction. Agric Syst 85:1–18. https://doi.org/10.1016/j.agsy.2004.07.009

Kriegler FJ, Malia WA, Nalepka RF, Richardson W (1969) Preprocessing transformations and their effects on multispectral recognition. In: Proceedings of the Sixth International Symposium on Remote Sensing of Environment, Ann Arbor, MI, USA, 13–16 October. pp 97–131

Kuhn M (2008) Building predictive models in R using the caret package. J Stat Softw 28:1–26

Leng G, Zhang X, Huang M, Asrar GR, Leung LR (2016) The role of climate covariability on crop yields in the conterminous United States. Sci Rep 6:1–11. https://doi.org/10.1038/srep33160

Lewis SC, Karoly DJ, Lewis SC, Karoly DJ (2013) Evaluation of historical diurnal temperature range trends in CMIP5 models. J Clim 26: 9077–9089. https://doi.org/10.1175/JCLI-D-13-00032.1

Lobell DB (2007) Changes in diurnal temperature range and national cereal yields. Agric For Meteorol 145:229–238. https://doi.org/10.1016/j.agrformet.2007.05.002

Lobell DB, Asseng S (2017) Comparing estimates of climate change impacts from process-based and statistical crop models. Environ Res Lett 12:015001. https://doi.org/10.1088/1748-9326/aa518a

Lobell DB, Burke MB (2010) On the use of statistical models to predict crop yield responses to climate change. Agric For Meteorol 150: 1443–1452. https://doi.org/10.1016/J.AGRFORMET.2010.07.008

Lobell DB, Cahill KN, Field CB (2008) Weather-based yield forecasts developed for 12 California crops. Calif Agric 60:211–215. https://doi.org/10.3733/ca.v060n04p211

Lobell DB, Bänziger M, Magorokosho C, Vivek B (2011) Nonlinear heat effects on African maize as evidenced by historical yield trials. Nat Clim Chang 1:42–45. https://doi.org/10.1038/nclimate1043

Lobell DB, Roberts MJ, Schlenker W, Braun N, Little BB, Rejesus RM, Hammer GL (2014) Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. Science 344:516–519. https://doi.org/10.1126/science.1251423

Mathieu JA, Aires F (2016) Statistical weather-impact models: an application of neural networks and mixed effects for corn production over the United States. J Appl Meteorol Climatol 55:2509–2527. https://doi.org/10.1175/JAMC-D-16-0055.1

Mathieu JA, Aires F (2018) Assessment of the agro-climatic indices to improve crop yield forecasting. Agric For Meteorol 253–254:15–30. https://doi.org/10.1016/J.AGRFORMET.2018.01.031

Morell FJ, Yang HS, Cassman KG, Van Wart J, Elmore RW, Licht M, Coulter JA, Ciampitti IA, Pittelkow CM, Brouder SM, Thomison P, Lauer J, Graham C, Massey R, Grassini P (2016) Can crop simulation models be used to predict local to regional maize yields and total production in the U.S. Corn Belt? Field Crop Res 192. https://doi.org/10.1016/j.fcr.2016.04.004

Mourtzinis S, Specht JE, Lindsey LE, Wiebold WJ, Ross J, Nafziger ED, Kandel HJ, Mueller N, Devillez PL, Arriaga FJ, Conley SP (2015) Climate-induced reduction in US-wide soybean yields underpinned by region- and in-season-specific responses. Nat. Plants 1:1–4. https://doi.org/10.1038/nplants.2014.26

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models, 4th edn. Irwin, Chicago

Oguntunde PG, Lischeid G, Dietrich O (2018) Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis. Int J Biometeorol 62:459–469. https://doi.org/10.1007/s00484-017-1454-6

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

Peters DB, Pendleton JW, Hageman RH, Brown CM (1971) Effect of night air temperature on grain yield of corn, wheat, and soybeans. Agron J 63:809–809. https://doi.org/10.2134/agronj1971.00021962006300050046x

Phillips S (1994) The Soil Conservation Service responds to the 1993 Midwest floods. Historical notes number 4, Enonomic and Social Sciences Division, United States Department of Agriculture

Powell JP, Reinhard S (2016) Measuring the effects of extreme weather events on yields. Weather. Clim. Extrem 12:69–79. https://doi.org/10.1016/J.WACE.2016.02.003

Prasad AK, Chai L, Singh RP, Kafatos M (2006) Crop yield estimation model for Iowa using remote sensing and surface parameters. Int J Appl Earth Obs Geoinf 8:26–33. https://doi.org/10.1016/J.JAG.2005.06.002

PRISM Climate Group (2018) PRISM gridded climate data, Parameter-elevation Regression on Independent Slopes Model (PRISM) Climate Group, Oregon State University [WWW Document]. Prism.

Ray DK, Gerber JS, MacDonald GK, and West PC (2015) Climate variation explains a third of global crop yield variability. Nat Commun 6(5989):1–9. https://doi.org/10.1038/ncomms6989

R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rouse JWJ, Hass RH, Schell JA, Deering DW (1974) Monitoring vegetation systems in the Great Plains with ERTS. In: Third Earth Resources Technology Satellite-1 Symposium- Volume I: Technical Presentations. NASA SP-351. Washington D.C. pp 309–317

Schlenker W (2018) Data sources and links [WWW Document]. Fine-scaled Weather data set

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

Setiyono TD, Yang H, Walters DT, Dobermann A, Ferguson RB, Roberts DF, Lyon DJ, Clay DE, Cassman KG (2011) Maize-N: a decision tool for nitrogen management in maize. Agron J 103:1276–1283. https://doi.org/10.2134/AGRONJ2011.0053

Shi W, Tao F, Zhang Z (2013) A review on statistical models for identifying climate contributions to crop yields. J Geogr Sci 23:567–576. https://doi.org/10.1007/s11442-013-1029-3

Tack J, Barkley A, Nalley LL (2015) Effect of warming temperatures on US wheat yields. Proc Natl Acad Sci U S A 112:6931–6936. https://doi.org/10.1073/pnas.1415181112

Teasdale JR, Cavigelli MA (2017) Meteorological fluctuations define long-term crop yield patterns in conventional and organic production systems. Sci Rep 7:1–10. https://doi.org/10.1038/s41598-017-00775-8

Thompson LM (1969) Weather and technology in the production of corn in the U. S. Corn Belt. Agron J 61:453–456. https://doi.org/10.2134/agronj1969.00021962006100030037x

USDA-NASS (2012) The yield forecasting program of NASS. SMB staff report number SMB 12-01, Statistical Methods Branch, National Agricultural Statistics Service, United States Department of Agriculture

USDA-NASS (2013) Crop production 2012 annual summary. National Agriculture and Statistics Service, United States Department of Agriculture

USDA-NASS (2019) Statistics by state. National Agricultural Statistics Service, United States Department of Agriculture van der Velde M, Biavetti I, El-Aydam M, Niemeyer S, Santini F, van den Berg M (2019) Use and relevance of European Union crop monitoring and yield forecasts. Agric Syst 168:224–230. https://doi.org/10.1016/J.AGSY.2018.05.001

van Wart J, Kersebaum KC, Peng S, Milner M (2013) Estimating crop yield potential at regional to national scales. Field Crop Res 143:34–43. https://doi.org/10.1016/J.FCR.2012.11.018

Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999

Verón SR, de Abelleyra D, Lobell DB (2015) Impacts of precipitation and temperature on crop yields in the Pampas. Clim Chang 130:235–245. https://doi.org/10.1007/s10584-015-1350-1

Vose RS, Easterling DR, Gleason B (2005) Maximum and minimum temperature trends for the globe: an update through 2004. Geophys Res Lett 32:L23822. https://doi.org/10.1029/2005GL024379

Westcott PC, Jewison M (2013) Weather effects on expected corn and soybean yields. FDS-13g-01. A report from the Economic Research Service, United States Department of Agriculture

Willmott CJ (1981) On the validation of models. Phys. Geogr. 2(2):184–194. https://doi.org/10.1080/02723646.1981.10642213

Wood SN (2017) Generalized additive models : an introduction with R, 2nd edn. Chapman and Hall/CRC, New York

Wood SN (2018) Mixed GAM computation vehicle with automatic smoothness estimation. Package “mgcv” v. 1.8-26

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.