Precision Agriculture (2024) 25:51–64
https://doi.org/10.1007/s11119-023-10057-1

Sweet corn yield prediction using machine learning models and field-level data

Daljeet S. Dhaliwal[1] · **Martin M. Williams II**[2]

Accepted: 21 July 2023 / Published online: 29 July 2023

Abstract

The advent of modern technologies, acquisition of large amounts of crop management and weather data, and advances in computing are reshaping modern agriculture. These advancements have unlocked the power of data by providing valuable insights and more accurate yield predictions. This study utilizes a historic US sweet corn dataset to: (a) evaluate machine learning model performances on sweet corn yield prediction and (b) identify the most influential variables for crop yield predictions. The sweet corn data comprised field-level data for over a quarter-century period (1992–2018) from two primary commercial sweet corn production regions for processing, namely the Upper Midwest and the Pacific Northwest. Several machine learning models were trained to predict field-level sweet corn yield from 67 variables of crop genetics, management, weather, and soil factors. The random forest model outperformed all trained models with the lowest RMSE (3.29 Mt/ha) and the highest Pearson's correlation coefficient (0.77) between predicted and observed yields. Variable importance plots revealed the top three most influential predictor variables as year (time), location (space), and seed source (genetics). Season long total precipitation and average minimum temperature during anthesis were the two most important weather variables in yield prediction. This is the first report of using fine-scale (time and space) crop data and advanced data analytics to leverage insights into commercial sweet corn production.

Keywords Machine learning · Random forest · Weather · Yield prediction

✉ Martin M. Williams II
martin.williams@usda.gov

[1] Department of Crop Sciences, University of Illinois at Urbana Champaign, Urbana, IL, USA

[2] Global Change and Photosynthesis Research Unit, USDA-ARS, Urbana, IL, USA

Introduction

Recent advances in machine learning techniques coupled with copious amounts of data have fostered a new niche of big data analytics in agriculture. Crop yield prediction plays a crucial role in optimizing production, resource allocation, and monitoring crop performance in real-time. With the increasing availability of data in agriculture, including remote sensing data, sensor data, and historical crop yield data, there is a growing need to leverage machine learning techniques to translate complex data into actionable insights for improving agricultural practices and enhancing crop yields.

Previous studies have attempted to understand yield variability in sweet corn; accounting for different planting dates (Williams, 2008), in-row spacing and genotype (Rangarajan et al., 2002; Williams, 2015; Dhaliwal & Williams, 2019), and planting date–weed control interactions (Williams and Lindquist, 2007). However, these factors reflect narrow aspects of the production of sweet corn for processing. More recent studies use long-term observational datasets to examine variability in historical yield trends of cereal crops (Jeong et al., 2016; Iwanska et al., 2018). To our knowledge, this is the first study to use fine-scale yield data (field-level observations) from commercial sweet corn production fields, harvested mechanically, to expand on our understanding of yield variability.

Crop yield is a product of complex interactions among genotype, environmental, and management practices, typically predicted by process-based or statistical modeling. Process-based models utilize detailed amounts of field-level measurements to simulate crop growth and development to environmental and management practices (Muchow et al., 1990). On the contrary, statistical models rely on historic climate & soil data and observational yield data to predict crop yields while considering the underlying eco-physiological conditions (Schlenker & Roberts, 2009). Process-based models provide more accurate outcomes compared to statistical models; however, statistical models are powerful regarding the volume, velocity, and variety of data (Roberts et al., 2017).

Statistical models implemented using machine learning models provide a better alternative to traditional regression models and are accompanied by tools to gain deeper insights into data. Different machine learning models have been used for crop yield forecasting, ranging from linear regression and decision trees (Jeong et al., 2016; Osman et al., 2017; Ranjan & Parida, 2019; Shahhosseini et al., 2020) to deep learning algorithms (Wang et al., 2018; Rao & Manasa, 2019; Khaki et al., 2021). For instance, Jeong et al. (2016) used random forest (RF) models to predict wheat, field corn, and potato yield at regional and global scales. Deep learning algorithms, such as convolutional neural networks (Rao & Manasa, 2019; Khaki et al., 2021), have shown substantial breakthroughs in improving crop yield predictions due to advancements in computational power and accessibility to larger volumes of data. However, the interpretability of higher-order deep learning algorithms remains a challenge compared to other machine learning models.

Machine Learning approaches have proven successful in identifying genetic variations (Yoosefzadeh-Naifabadi et al., 2021; Xu et al., 2022), understanding weather impacts, and determining effective management practices (Crane-Droesch, 2018; Shook et al., 2021) in agriculture. These models learn from historical information, considering factors such as environmental conditions, genetics, and management practices to make accurate predictions. However, machine learning models can suffer from overfitting, where they become overly specialized to the training data and struggle to generalize to new data. To address

this challenge, ensemble techniques have emerged as a solution by combining multiple machine learning models to improve prediction accuracy and mitigating overfitting (Dietterich, 2000).

Ensemble techniques are valuable approaches to combat overfitting in machine learning models. They involve combining multiple models to enhance prediction accuracy and reduce overfitting. Bagging trains independent models on different subsets of the training data and combines their predictions, effectively smoothing out biases and reducing variance. Boosting builds models sequentially, focusing on correcting errors made by previous models, thus improving the overall performance of the ensembles. Stacking trains multiple models on the same dataset and uses their predictions as input features for a meta-model, capturing complex patterns and reducing overfitting. Regularization techniques, such as adding penalty terms to the model's objective function, can be applied to each individual model within the ensemble to control complexity and prevent overfitting. By utilizing these ensemble techniques, the combined models can provide more robust and accurate predictions on new data, improving generalization and mitigating the risk of overfitting.

By leveraging advanced machine learning techniques to analyze vast amounts of data from commercial sweet corn fields, combined with historic weather and genotype data, it is possible to gain a deeper understanding of the complex interactions and variables that significantly impact sweet corn yield. This enables farmers, breeders, and industry professionals to make data-driven decisions that can optimize crop performance and ultimately improve yields.

The objectives of this study were to: (a) evaluate machine learning model performances on sweet corn yield prediction and (b) identify the most influential variables for crop yield predictions.

Materials and methods

Data description

Field-level historic sweet corn yield data were obtained from multiple US vegetable processors from 1992 to 2018. The dataset (hereafter referred to as 'US sweet corn data') contains sweet corn yields from two primary regions of commercial sweet corn production for processing in the US, i.e., the Upper Midwest (the states of IL, MN, WI) and the Pacific Northwest (the state of WA). Furthermore, these regions were classified into five production areas: IL-Irrigated, IL-Rainfed, MN-Rainfed, WA-Irrigated, and WI-Irrigated (Fig S1). Sweet corn yields, i.e., green ear mass (Mt/ha) were recorded from contract growers' fields by the processor. Contract growers' fields reflect typical standards of commercial sweet corn production regarding plant density, nutrient management, pest and weed control, etc. The Materials Transfer Agreement governing the use of this dataset dictates strict confidentiality, including names of processors, contract growers, and hybrids.

The US sweet corn dataset accompanied observed information on hybrid grown, cultural practices, and important agronomic planting dates, tasseling, and harvest (Table 1). Later, hybrid information was matched to the seed source/company, and this new variable (seed source) would become a proxy for a hybrid. This was done to reduce possible confounding bias arising from the likelihood of similar genetic material contributed by individual

Table 1 Description of genetics and crop management variables in the US sweet corn data. Seed source denotes the parent seed company that makes the hybrid grown in the contract growers' fields

Genetics & crop management	Description
Production area	IL-Rainfed, IL-Irrigated, MN-Rainfed, WA-Irrigated, WI-Irrigated
Seed Source	Abbot & Cobb, Crookham, DelMonte, Harris Moran, IFSI, Seminis, Seneca, Snowy River, Syngenta
Cultural practice	Grower rating: Poor, Good, Excellent; based on how closely they followed contract recommendations for crop production
Important dates	Planting date, tassel date, harvest date

Table 2 Weather and soil characteristics included in the US sweet corn data. Using field location coordinates, weather and soil data were extracted from Daymet and SSURGO databases, respectively

Variable	abbreviation	units
[1]**Weather data (Daymet dataset)**		
Minimum air temperature	Tmin	°C
Maximum air temperature	Tmax	°C
Average air temperature	Tavg	°C
Precipitation	prec	mm
Growing degree days	GDD	
Shortwave solar radiation	SWRAD	W/m^2
Average vapor pressure deficit	VPD	Pa
Average potential evapotranspiration	PET	mm
[2]**Soil data (SSURGO database)**		
Clay		g/kg
Sand		g/kg
Silt		g/kg
Cation exchange capacity	cec	mmol/kg
Soil organic carbon	soc	dg/kg

[1]Weather variables were calculated for the length of crop growing season (season length), first (planting + 10 days), second (tassel date – 30 days), third (tassel date + 10 days), and fourth (harvest – 21 days) crop growth intervals.

[2]Soil variables were aggregated for four soil depths: dep1 (0–30 cm), dep2 (30–60 cm), dep3 (60–100 cm) and dep4 (100–200 cm).

seed companies, thereby shrinking around 100 hybrids into nine unique seed sources as described in Table 1.

Important information on growing season weather and soil characteristics was added to the US sweet corn data using field location coordinates. Weather data were obtained from Daymet daily surface weather database on a 1-km grid for the North Americas (Thornton et al., 2021). The following eight variables were included to represent growing season weather conditions: daily minimum, maximum, and average air temperatures; precipitation; shortwave solar radiation; growing degree days; average vapor pressure deficit; and average potential evapotranspiration (see Table 2). Daily average potential evapotranspiration was calculated using the Priestly-Taylor method, which captures both diurnal and seasonal variations in seasonal evaporative demand (Priestley & Taylor, 1972). All weather variables were calculated for four different intervals corresponding to different crop growth and development stages. The intervals were estimated using observed planting, tassel, and harvest date recorded for each of the contract growers' fields. First, second, third, and fourth intervals represented the periods of 0–10 days after planting, 0–30 days before tasseling,

0–10 days after tasseling, and 0–20 days before harvest, respectively. The critical stages of sweet corn growth and development captured by the four intervals include seed germination and emergence, exponential growth, pollen shed and silking (anthesis), and kernel development (R2 or blister stage), respectively. Weather variables were also calculated for the growing season duration, representing season long estimates of all eight weather variables.

Soil characteristics used to describe the environmental conditions included soil texture (clay, sand, silt content), cation exchange capacity, and soil organic carbon (see Table 2) obtained from the SSURGO database (Web Soil Survey, 2020). All soil characteristics were obtained at different depths along the soil profile (0–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm, see Table 2), hence, resulting in total twenty soil features in the US sweet corn dataset.

After data augmentation, the US sweet corn dataset consisted of 16,040 unique green ear yield observations (i.e., fields) with 67 different explanatory features detailed above. The explanatory variables comprised a time component (year), spatial component (production area), genetics (seed source), crop management, and environmental conditions (weather and soil). This comprehensive set of explanatory variables captures sufficient information to explain spatio-temporal variabilities in sweet corn yield and will be used in model building to predict sweet corn yields.

Data pre-processing

Prior to model building, the US sweet corn dataset was split into two sets — a training set to build the model and a test set to provide an unbiased evaluation of the training model. A random subset consisting of 70% of yield observations was assigned to the training set, and the remaining was used as the test set. A k-fold cross-validation with $k = 10$ was used on the training dataset. This involved dividing the training data into 10 subsets, where 9 subsets are used for training and the remaining subset is used for validation. This process was repeated 10 times, with each subset used as the validation set once. This allowed us to obtain an estimate of the model's performance and its variability by averaging the performance over k validation sets.

Machine learning model building

Principal components regression

Principal components regression is an unsupervised technique that performs dimensionality reduction by featuring principal component analysis (PCA) in the first step and then building a linear regression model on the transformed principal components (Jolliffe, 1986). We utilized the 'prcomp' function from the 'stats' package in R for PCA and the 'lm' function for linear regression (R Core Team, 2021).

Partial least squares regression

Partial least squares regression (PLS) is a supervised dimensionality reduction technique that reduces the predictors to a smaller set of uncorrelated components and performs least squares regression on these components. While principal component analysis chooses linear

components that summarize the maximum variation of the predictors, PLS finds components that summarize the maximum variability in predictors while maintaining maximum correlation between components and the response (Kuhn & Johnson, 2013). We utilized the 'plsr' function from the 'pls' package in R (Mevik & Wehrens, 2007).

Multiple linear regression

Multiple linear regression accommodates multiple predictors to predict a measurable response variable **(*Y*)**. The multiple linear regression model can be expressed as:

$$\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{X}_1 + \boldsymbol{\beta}_2 \boldsymbol{X}_2 + \cdots + \boldsymbol{\beta}_{\boldsymbol{P}} \boldsymbol{X}_{\boldsymbol{P}} + \varepsilon,$$

Where $\boldsymbol{X_j}$ represents the ***j*th** predictor and $\boldsymbol{\beta_j}$ quantifies the association between that variable and the response. The model assumes a linear relationship between the predictors and response variable, normality, no multicollinearity, and homoscedasticity (Hastie et al., 2009). The 'lm' function from the 'stats' package in R was used to build the regression model (R Core Team, 2021).

Regularized regression

Regularized regression constrains the slope coefficient estimates and shrinks them towards zero to build a more parsimonious model. We used the 'glmnet' package in R to implement regularized regression (Friedman et al., 2010).

Multivariate adaptive regression splines

The multivariate adaptive regression splines (MARS) model creates a piecewise linear model which captures nonlinear relationships in data by automatically determining the number and location of breakpoints (or knots) (Friedman, 1991). The 'earth' package in R was used to implement the MARS model (Milborrow, 2019).

Random forest

Random forest (Breiman, 2001) is built on the concept of bootstrap aggregation, which is a tree-based ensemble model. Bootstrap aggregation (or Bagging) attempts to reduce the variance for notoriously noisy decision trees by utilizing bootstrap procedure, i.e., averaging predictions from many random sub-samples with replacement. Random forest algorithm uses a random number of features to construct each tree and repeats this procedure many times and eventually averaging all the predictions made by all trees. Thus, the RF model addresses both bias and variance components of the error and is proved to be powerful. Random forest model was implemented using 'randomForest' package in R.

For the purpose of reproducibility, the random forest model was built with carefully chosen hyperparameters. The forest consisted of 300 trees; a number determined through a validation process to ensure optimal performance. The 'mtry' parameter, which determines the random number of features considered at each split, was set to the square root of the total number of variables in the dataset. This selection aimed to strike a balance between model

complexity and feature diversity, enabling the random forest model to effectively capture the relationships within the data.

Model performance comparison metrics

To assess model performance, the following two metrics were calculated on the test dataset.

Root Mean Square Error (RMSE) is calculated using the equation:

$$RMSE = \sqrt{\sum_{i=1}^{N} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}}_i)^2 / \boldsymbol{N}}$$

Where, $\boldsymbol{y_i}$ represents the observations in the test data sets, $\widehat{\boldsymbol{y}}_i$ are the predictions in the test dataset. $\boldsymbol{N}$ is the total number of observations.

Pearson's Correlation Coefficient (r) was calculated between yield observations and yield predictions obtained from the test dataset.

Results

Model performance comparisons

Model performances for accurate sweet corn yield predictions were compared using root mean square error (RMSE) and Pearson's correlation coefficient (r) metrics. Using the same test dataset across all models, the performance comparison metrics were obtained between predicted and observed yield values. The random forest model outperformed all trained models with the lowest RMSE of 3.29 Mt/ha and the highest Pearson's correlation coefficient of 0.77 (Fig. 1). Multivariate adaptive regression splines (MARS), regularized regression, and partial least squares models did not show significant improvements from the benchmark multiple linear regression model for RMSE or Pearson's correlation coefficient (Fig. 1). Principal components regression reported the worst accuracy of parameters for yield predictions among the six trained models.

Variable importance plot

Variable importance plots provide a better understanding of data by quantifying the importance of explanatory variables in predictive modeling. Since the RF model was the best performing model, this section will discuss the implications of important variables extracted from the RF model. The metric used to quantify the importance of variables is a loss function in RMSE as described in Fisher et al. (2019). If eliminated from the predictive model, the more influential variables result in a greater loss in model performance metric. To eliminate the effect of a variable, values of the variable were permuted, and model performance was evaluated using the RMSE metric. Fifty permutations were carried out, and the average loss in RMSE is reported (Fig. 2). The top fifteen variables in predicting sweet corn yields, ranked by descending order of importance, are shown in Fig. 2.

Year, production area, and seed source were the top three influential variables in predicting sweet corn yield. Season long precipitation and average minimum temperature during

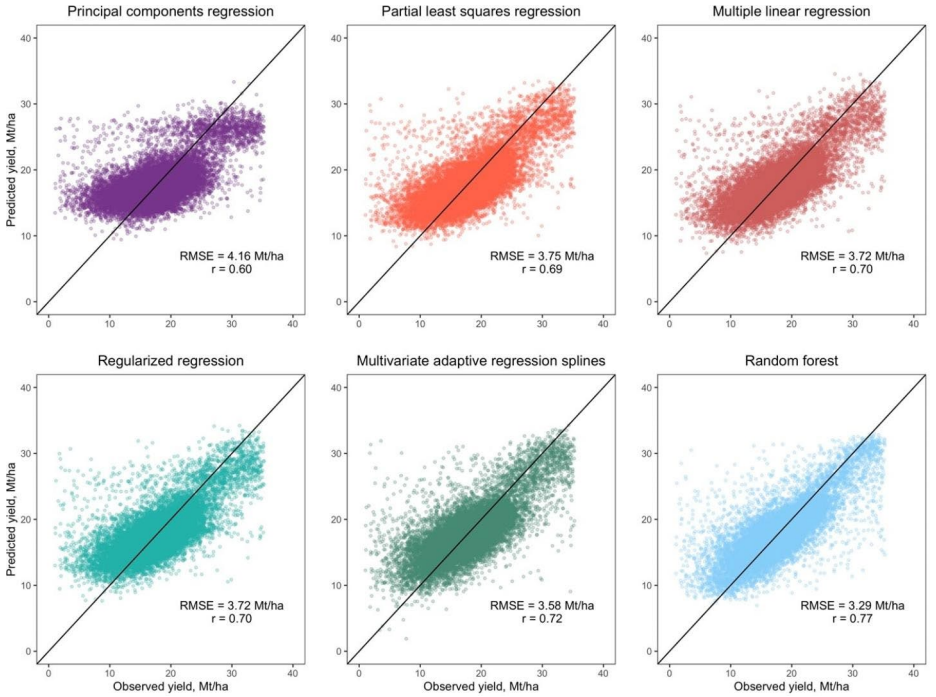

Fig. 1 Comparison of machine learning model performances on test dataset (sample size = 4,800). Scatterplots for observed vs. predicted yields are shown. The solid line represents 1:1 relation between observed and predicted yields. Model comparison metrics shown are RMSE and Pearson's correlation coefficient (r)

the third crop growth interval were the top two important weather variables in predicting sweet corn yield. The only soil characteristic that showed up in the top fifteen important predictor variables was average soil organic carbon at soil depth 0–30 cm (Fig. 2).

Partial dependence plots

Partial dependence plots show the marginal effect of continuous predictor variables on sweet corn yield (Fig. 3). On average, sweet corn yield increased from nearly 16.5 Mt/ha to 20.0 Mt/ha over one decade (2000–10). Since 2010, crop yields have not reported any improvements. Season long precipitation stabilizes crop yields until the total precipitation exceeds ~ 500 mm.

During the third crop growth interval, the average minimum temperature shows detrimental effect on crop yields beyond 16 °C. The dependence of crop yield on remaining important weather variables is shown in Fig. 3.

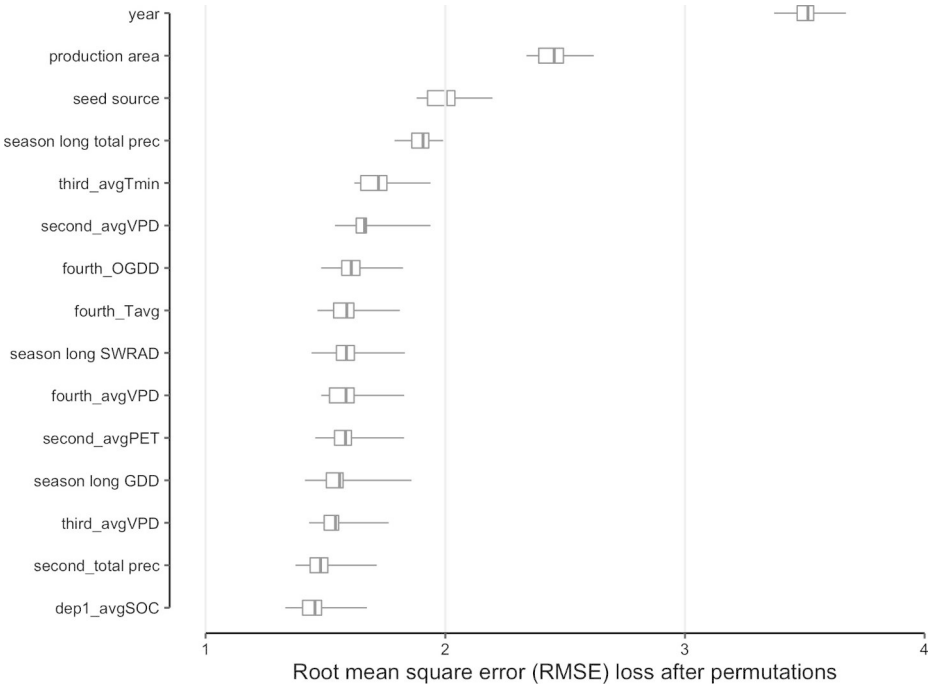

Fig. 2 Variable importance plot calculated by using fifty permutations and the RMSE loss function for a random forest model. Top fifteen selected variables are shown, with year being the most important and soil organic carbon content (0–30 cm) being the least important

Discussion

This study utilizes machine learning models to understand the temporal and spatial heterogeneities in sweet corn yield as a function of genetics, management, weather, and soil factors. Machine learning models were compared using common model performance metrics: RMSE and Pearson's correlation coefficient. Our results illustrate that the RF model is highly effective for sweet corn yield prediction. The RF model outperformed all machine learning models we tested. The accuracy obtained in this study ($r=0.77$) is similar to that obtained in field corn yield prediction using different machine learning models (Ahalawat and Minsker, 2016; Pantazi et al., 2016). However, this is the first report of data-driven analysis to predict sweet corn yields utilizing yield data at a finer scale (i.e., field-level) to build machine learning models.

The high performance of the RF model is likely more evident when the response is a result of complex interactions among multiple predictors, such as crop management, environmental, and edaphic factors, that can complicate classical linear regression modeling. Crop production variables from weather, management, and soil factors are usually highly correlated and violate the assumption of independent variables in traditional linear regression models. In such instances, RF models are highly advantageous. Additionally, RF models allow flexibility in using both categorical and continuous data in model-building frameworks.

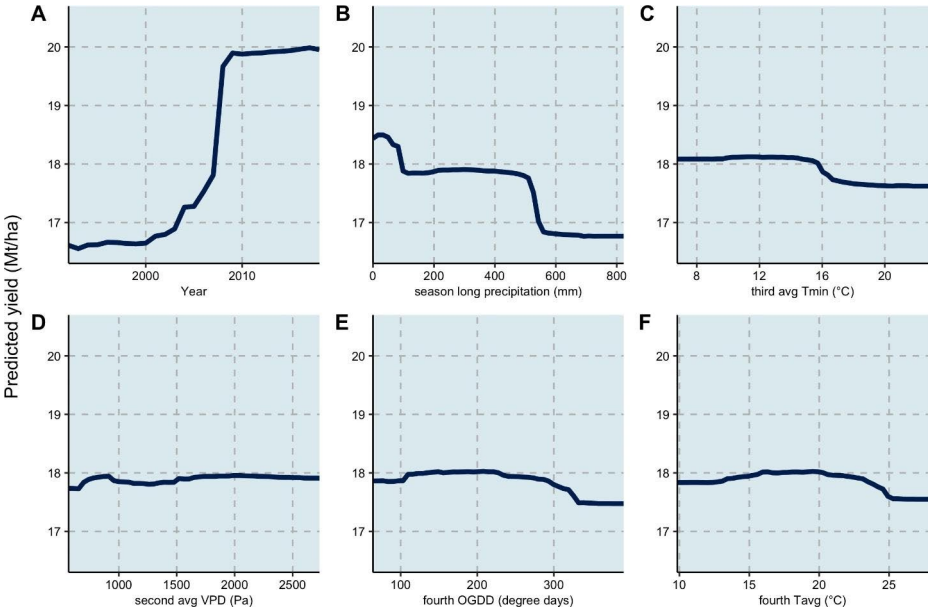

Fig. 3 Partial dependence plots, based on results from the random forest model, showing the mean marginal influence of top six explanatory variables on sweet corn yield prediction, including (A) year, (B) season long precipitation, (C) third interval avg. minimum temperature, (D) second interval avg. vapor pressure deficit, (E) fourth interval growing degree days, (F) fourth interval avg. temperature. Each plot represents the effect of a single variable while holding the other variables constant

The second objective of this study was to understand which variables contribute the most to sweet corn yield prediction. Variable importance and partial dependence plots are key utilities of RF models that allow comparisons of variable importance and dependence. Our results indicate that year and production area, representing broad temporal and spatial heterogeneities, were the top two influential variables predicting sweet corn yield. The year variable captured technological advancements and improved agronomic recommendations over time. Sweet corn yields increased significantly in the first decade of the 21st century and have stagnated since 2010. The lack of yield improvement in processing sweet corn in recent years is also supported by USDA–NASS data (USDA–NASS, 2021). However, the partial dependence plot (Fig. 3a) doesn't capture the yield trends at the regional level, i.e., production area.

Production area represents the spatial component in US sweet corn data. Production areas in Pacific Northwest were higher yielding than those in the Upper Midwest (Fig. S2). Historically, production areas in the inland Pacific Northwest have led in fruit and vegetable yields, mainly by abundant seasonal daylight, moderate nighttime temperatures, and an ample irrigation water. Furthermore, irrigated production areas in the Upper Midwest yield higher than rainfed production areas in the same region (Fig. S2). Previous studies have shown the cooling effect of irrigation to negate extreme heat stress on field corn growth and development (Lobell et al., 2008; Siebert et al., 2017; Li et al., 2020).

One of the unique attributes of this observational dataset is the availability of hybrid information represented by seed sources. The Materials Transfer Agreement dictates the

strict confidentiality of hybrid names. Nonetheless, seed source (genetics) is a significant driver of sweet corn yield heterogeneities across the production areas in the Upper Midwest and Pacific Northwest (Fig. S3). This can be attributed to differences in breeding programs (i.e., eating quality, yield improvement, host-disease resistance) and other unknown organizational dynamics within seed companies.

Our results show season long precipitation is the most influential weather variable on sweet corn yields. The partial dependence plot for the effect of season long precipitation on crop yield integrates information from all production areas in the Upper Midwest and Pacific Northwest. The initial decline in crop yield with increasing precipitation represents the shift in data from high-yielding WA-Irrigated to production areas in the Upper Midwest (Fig. S43). The yield decline observed near 500 mm precipitation underscores the importance of extreme weather conditions like floods on sweet corn yield (Fig. 3b). These data provide an excellent opportunity to study regional effects of precipitation anomalies on sweet corn yield.

Average minimum (i.e., nighttime) temperature during the third crop growth interval was the top temperature-related variable in sweet corn yield prediction. The third crop growth interval (anthesis) represents the phenological stage most vulnerable to heat stress. Higher nighttime temperatures during the third crop growth interval are associated with reduced sweet corn yields (Fig. 3c). In C4 crop species, including sweet corn, yield losses from high nighttime temperatures are a result of complex eco-physiological processes involving increased plant respiration rates and, potentially, changes in crop evaporative demands (Atkin & Tjoelker, 2003; Sadok & Jagadish, 2020). Vyn (2010) stated that grain fill is a 24-h process, so both day and nighttime temperatures are important. High nighttime temperatures (>21 °C) induce nighttime respiration, resulting in a lower amount of dry matter accumulation in plants. With high nighttime temperatures, more photoassimilates produced during the day are lost; less is available to fill developing kernels, thereby lowering grain yield (Thomison, 2005). The lower threshold for yield loss in sweet corn, 16 °C compared to >21 °C reported in field corn, suggests sweet corn has a higher sensitivity to yield losses from excessive temperatures than field corn.

Conclusions

This study used a rich spatiotemporal sweet corn dataset with field-level yield observations to evaluate various machine learning model prediction capabilities. Our results demonstrate that the RF model provides the most accurate yield predictions among the models tested. Variable importance plots quantify the importance of predictor variables, with year, production area, and seed source being the top three influential variables on sweet corn yield. Season long precipitation and average minimum (i.e., nighttime) temperature during anthesis were the top two essential weather variables in sweet corn yield prediction. Improving model prediction accuracy may be possible with the inclusion of additional crop management information, such as fertilizer application rate, plant density, and amount of irrigation, where applicable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11119-023-10057-1.

Acknowledgements We would like to extend our sincere gratitude to anonymous vegetable processors for graciously providing us with the historic sweet corn data used in this study. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the US Department of Agriculture (USDA) and does not imply its approval to the exclusion of other products or vendors that also may be suitable. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the USDA. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal opportunity provider and employer.

Declarations

Competing interests The authors declare that they have no conflict of interest.

References

Ahalawat, J. (2016). Data driven modeling of corn yield: A machine learning approach [Master's Thesis]. University of Illinois. http://hdl.handle.net/2142/90600.

Atkin, O. K., & Tjoelker, M. G. (2003). Thermal acclimation and the dynamic response of plant respiration to temperature. *Trends in Plant Science*, *8*(7), 343–351. https://doi.org/10.1016/S1360-1385(03)00136-5.

Breiman, L. (2001). Random Forests. *Machine Learning 2001*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003.

Dhaliwal, D. S., & Williams, M. M. II. (2019). Optimum plant density for crowding stress tolerant processing sweet corn. *Plos One*, *14*(9), https://doi.org/10.1371/journal.pone.0223107.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*. https://arxiv.org/abs/1801.01489v5.

Friedman, J. H. (1991). Multivariate adaptive regression splines. (1), 1–67. https://doi.org/10.1214/AOS/1176347963.

Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization Paths for generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Linear methods for regression. In T. Hastie, R. Tibshirani, J. Friedman (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, Prediction* (pp. 261–294). Springer Series in Statistics. New York, NY. pp. 261–294. doi: https://doi.org/10.1007/978-0-387-84858-7_8.

Iwańska, M., Oleksy, A., Dacko, M., Skowera, B., Oleksiak, T., & Wójcik-Gront, E. (2018). Use of classification and regression trees (CART) for analyzing determinants of winter wheat yield variation among fields in Poland. *Biometrical Letters*, *55*(2), 197–214. https://doi.org/10.2478/BILE-2018-0013.

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K. M., Gerber, J. S., Reddy, V. R., & Kim, S. H. (2016). Random forests for global and regional crop yield predictions. *Plos One*, *11*(6), 1–15. https://doi.org/10.1371/journal.pone.0156571.

Jolliffe, I. T. (1986). Principal components in regression analysis. *Principal component analysis* (pp. 129–155). New York, NY: Springer.

Khaki, S., Pham, H., & Wang, L. (2021). Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Scientific Reports*, *11*(1), 1–14. https://doi.org/10.1038/s41598-021-89779-z.

Kuhn, M., & Johnson, K. (2013). Linear regression and its cousins. *Applied Predictive modeling*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-6849-3_6.

Li, Y., Guan, K., Peng, B., Franz, T. E., Wardlow, B., & Pan, M. (2020). Quantifying irrigation cooling benefits to maize yield in the US Midwest. *Global Change Biology*, *26*(5), 3065–3078. https://doi.org/10.1111/GCB.15002.

Lobell, D. B., Bonfils, C. J., Kueppers, L. M., & Snyder, M. A. (2008). Irrigation cooling effect on temperature and heat index extremes. *Geophysical Research Letters*, *35*(9), 9705. https://doi.org/10.1029/2008GL034145.

Mevik, B. H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, *18*(2), 1–23.

Milborrow, M. S. (2019). *Package "earth*. R Software package.

Muchow, R. C., Sinclair, T. R., & Bennett, J. M. (1990). Temperature and solar radiation effects on potential maize yield across locations. *Agronomy Journal*, *82*(2), 338–343. https://doi.org/10.2134/AGRONJ1990.00021962008200020033X.

Osman, T., Psyche, S. S., Kamal, M. R., Tamanna, F., Haque, F., & Rahman, R. M. (2017). Predicting early crop production by analysing prior environment factors. In M. Akagi, T. T, Nguyen, D. T, Vu, T. N. Phung, V. N. Huynh (Eds.), Advances in information and communication technology. International Conference on Advances in Information and Communication Technology 2016. Advances in intelligent systems and computing, vol 538. Springer, Cham. https://doi.org/10.1007/978-3-319-49073-1_51.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65. https://doi.org/10.1016/J.COMPAG.2015.11.018.

Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100>0081:OTAOSH<2.3.CO;2.

Rangarajan, A., Ingall, B., Orfanedes, M., & Wolfe, D. (2002). In-row spacing and cultivar affects ear yield and quality of early-planted sweet corn. *HortTechnology*, *12*(3), 410–415.

Ranjan, A. K., & Parida, B. R. (2019). Paddy acreage mapping and yield prediction using sentinel-based optical and SAR data in Sahibganj district, Jharkhand (India). *Spatial Information Research*, *27*(4), 399–410. https://doi.org/10.1007/S41324-019-00246-4.

Rao, D. T. V. N., & Manasa, S. (2019). Artificial neural networks for soil quality and crop yield prediction using machine learning. *International Journal on Future Revolution in Computer Science & Communication Engineering*, *5*(1), 57–60. http://www.ijfrcsce.org/index.php/ijfrcsce/article/view/1835.

Roberts, M. J., Braun, N. O., Sinclair, T. R., Lobell, D. B., & Schlenker, W. (2017). Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environmental Research Letters*, *12*(9), 095010. https://doi.org/10.1088/1748-9326/AA7F33.

Sadok, W., & Jagadish, S. V. K. (2020). The hidden costs of nighttime warming on yields. *Trends in Plant Science*, *25*(7), 644–651. https://doi.org/10.1016/J.TPLANTS.2020.02.003.

Schlenker, W., & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, *106*(37), 15594–15598. https://doi.org/10.1073/PNAS.0906865106.

Shahhosseini, M., Hu, G., & Archontoulis, S. (2020). Forecasting corn yield with machine learning ensembles. *Frontiers in Plant Science*, *11*(July), 1–16. https://doi.org/10.3389/fpls.2020.01120.

Shook, J., Gangopadhyay, T., Wu, L., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2021). Crop yield prediction integrating genotype and weather variables using deep learning. *Plos One*, *16*(6), e0252402. https://doi.org/10.1371/journal.pone.0252402.

Siebert, S., Webber, H., Zhao, G., & Ewert, F. (2017). Heat stress is overestimated in climate impact studies for irrigated agriculture. *Environmental Research Letters*, *12*(5), 054023. https://doi.org/10.1088/1748-9326/AA702F.

Thomison, P. (2005). Impact of warm night temperatures on corn grain yields. CORN newsletter 25. 12 Oct. 2021. http://corn.osu.edu/newsletters/2005/article?issueid=97&articleid=574

Thornton, P. E., Shrestha, R., Thornton, M., Kao, S. C., Wei, Y., & Wilson, B. E. (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data*, *8*(1), https://doi.org/10.1038/S41597-021-00973-0.

USDA National Agricultural Statistics Service (2021). NASS - Quick Stats. USDA National Agricultural Statistics Service. https://data.nal.usda.gov/dataset/nass-quick-stats. Accessed 2021-09-18.

Vyn, T. J. (2010). Excessive heat and humidity not ideal for corn. *Pest & Crop Newsletter*. Issue 19. 12 Oct. 2021. http://extension.entm.purdue.edu/pestcrop/2010/issue19/index.html.

Wang, A. X., Tran, C., Desai, N., Lobell, D., & Ermon, S. (2018). Deep transfer learning for crop yield prediction with remote sensing data. ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS), pp. 1–5. https://doi.org/10.1145/3209811.3212707.

Web Soil Survey (2020). Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available at https://websoilsurvey.sc.egov.usda.gov/App/HomePage.htm.

Williams, M. M. II. (2008). Sweet corn growth and yield responses to planting dates of the North Central United States. *Hortscience*, *43*(6), 1775–1779. https://doi.org/10.21273/HORTSCI.43.6.1775.

Williams, M. M. II. (2015). Identifying crowding stress-tolerant hybrids in processing sweet corn. *Agronomy Journal*, *107*(5), 1782–1788. https://doi.org/10.2134/agronj15.0011.

Williams, M. M., II, & Lindquist, J. L. (2007). Influence of planting date and weed interference on sweet corn growth and development. *Agronomy Journal*, *99*(4), 1066–1072. https://doi.org/10.2134/AGRONJ2007.0009.

Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., Varshney, R. K., Prasanna, B. M., & Qian, Q. (2022). Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction. *Molecular Plant*, *15*, 1664–1695.

Yoosefzadeh-Najafabadi, M., Tulpan, D., & Eskandari, M. (2021). Application of machine learning and genetic optimization algorithms for modeling and optimizing soybean yield using its component traits. *Plos One*, *16*(4), e0250665. https://doi.org/10.1371/journal.pone.0250665.