frontiers in Plant Science

ORIGINAL RESEARCH
published: 21 September 2021
doi: 10.3389/fpls.2021.701192

Data Driven Explanation of Temporal and Spatial Variability of Maize Yield in the United States

*Lizhi Wang**

Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA, United States

Maize yield has demonstrated significant variability both temporally and spatially. Numerous models have been presented to explain such variability in crop yield using data from multiple sources with varying temporal and spatial resolutions. Some of these models are data driven, which focus on approximating the complex relationship between explanatory variables and crop yield from massive data sets. Others are knowledge driven, which focus on integrating scientific understanding of crop growth mechanism in the modeling structure. We propose a new model that leverages the computational efficiency and prediction accuracy of data driven models and incorporates agronomic insights from knowledge driven models. Referred to as the GEM model, this model estimates three independent components of (G)enetics, (E)nvironment, and (M)anagement, the product of which is used as the predicted crop yield. The aim of this study is to produce not only accurate crop yield predictions but also insightful explanations of temporal and spatial variability with respect to weather, soil, and management variables. Computational experiments were conducted on a data set that includes maize yield, weather, soil, and management data covering 2,649 counties in the U.S. from 1980 to 2019. Results suggested that the GEM model is able to achieve a comparable prediction performance with state-of-the-art machine learning models and produce meaningful insights such as the estimated growth potential, effectiveness of management practices, and genetic progress.

Keywords: crop yield prediction, machine learning, crop models, temporal and spatial variability, heuristic algorithm

OPEN ACCESS

Edited by:
Zhanwu Dai,
Institute of Botany, Chinese Academy of Sciences, China

Reviewed by:
Yingjie Xiao,
Huazhong Agricultural University, China
Alpha Kamara,
International Institute of Tropical Agriculture (IITA), Nigeria

***Correspondence:**
Lizhi Wang
lzwang@iastate.edu

Specialty section:
This article was submitted to Crop and Product Physiology, a section of the journal Frontiers in Plant Science

Received: *27 April 2021*
Accepted: *09 August 2021*
Published: *21 September 2021*

Citation:
Wang L (2021) Data Driven Explanation of Temporal and Spatial Variability of Maize Yield in the United States. Front. Plant Sci. 12:701192. doi: 10.3389/fpls.2021.701192

1. INTRODUCTION

Crop yield prediction plays an important role in agriculture. On the economic front, agriculture stakeholders such as farmers, insurance companies, and breeders rely on yield predictions to make informed operational decisions. On the societal front, yield predictions help governments and organizations make effective policies to strength global security, support famine-prevention efforts, and protect environmental sustainability, especially in an era of global climate change and pandemics (Messina et al., 2010; Marko et al., 2016). On the scientific front, underlying crop yield prediction is a fundamental research question of understanding how phenotype is determined by genotype, environment, and their interactions. In particular, the relationship between genetics, weather, soil, and management variables and crop yield has been the topic of extensive studies. The pursuit of more accurate crop yield prediction techniques has and will continue to motivate innovations at the intersection of plant science, engineering, and data analytics.

Most crop yield prediction models can be categorized as either data driven or knowledge driven. Machine learning models, epitomized by neural networks, consist of large numbers of simple computational units that grow into complex model structures for data driven analysis. With little pre-programmed knowledge or biases, a machine learning algorithm treats crop yield as an unknown function of genotype and environment and attempts to approximate the underlying function by learning its own lesson from large data sets. Khaki and Wang (2019) used a deep neural network model for the 2018 Syngenta crop challenge (Syngenta, 2020), in which participants were challenged to predict the 2017 crop yield of 2,247 fields using historical data of genotype, weather, soil, and yield. Their approach outperformed other popular machine learning methods such as LASSO, shallow neural networks, and regression tree. Shahhosseini et al. (2019) evaluated four machine learning algorithms and their ensembles in predicting maize yield and nitrate losses. Using experimental data from seven locations in four Midwestern states in the U.S. over 5–7 years and a large scenario analysis data set generated by the agricultural production systems simulator (APSIM) (Holzworth et al., 2014), they achieved the following RMSEs in bu/ac for yield prediction: 22.19 for LASSO regression, 22.42 for ridge regression, 20.82 for random forests, 20.04 for extreme gradient boosting, and 18.44 for their optimal ensemble. Kang et al. (2020) compared six machine learning algorithms in predicting county level maize yield in 12 states in the U.S. using data from 2001 to 2016. The RMSE ranged from 14.8 to 24 bu/ac. Crane-Droesch (Crane-Droesch, 2018) proposed a semiparametric variant of a deep neural network and compared it with multiple other machine learning algorithms using data from 9 states in the U.S. Corn Belt from 1979 to 2016; RMSEs for unseen years ranged from 15.9 to 19.1 bu/ac. More detailed reviews of machine learning models for crop yield prediction can be found in Chlingaryan et al. (2018) and van Klompenburg et al. (2020).

Knowledge driven models, epitomized by crop models such as APSIM (Holzworth et al., 2014) and CERES-Maize (Hodges et al., 1987), build upon physiological understanding of plant growth processes and develop biologically meaningful non-linear equations to predict crop yield (among other outputs) as a complex function of plant traits (e.g., leaf appearance rate, total leaf number, grain fill duration, grain number, and root front velocity) and environmental parameters (Batchelor et al., 2002; Heslot et al., 2014; Schauberger et al., 2017). Crop models offer biological insights into causes of phenotypic variability by providing explicit explanations of the interactions between traits and environmental conditions in different phases of the crop growth cycle. As such, knowledge driven models are more commonly evaluated based on their qualitative reflection of crop responses to agrometeorological effects (Lalić et al., 2014) than quantitative prediction accuracy (Kiniry et al., 1997). Blanc (2017) built an emulators of crop yields based on an ensemble of five crop models and evaluated its performance in replicating spatial patterns of yields crop levels and changes overtime. Schauberger et al. (2017) used an ensemble of nine crop models to enhance the ability of individual process-based crop models to represent effects of high temperature on crop yield. Durand et al. (2018) assessed the ability of 21 crop models to capture the impact of elevated carbon dioxide concentration on maize yield and found evidence that more mechanistic modeling approaches led to better performances.

Data driven and knowledge driven models have complementary strengths and limitations. On the one hand, the ability to approximate complex functions to fit data (Hornik et al., 1990) enables machine learning models to achieve relatively high prediction accuracy, but a major limitation is the difficulty to explain the results. For example, many studies use a separate model for each geographic region with different parameters; each feature is used in hundreds or thousands of equations to produce the final prediction, and the importance of different features changes by year and by location (Kang et al., 2020). As a result, when the predictions are accurate, they offer little insights that are explainable, much less transferable temporally or spatially; when the predictions are inaccurate, it is hard to identify the cause of the errors. On the other hand, knowledge driven models have the potential to propose scientifically and biologically meaningful hypotheses that can form the basis of experimental validation. The pre-programmed human input in knowledge driven models can greatly simplify the learning process while achieving a reasonable performance, but it also restricts what can be learned from data and consequently limits the prediction accuracy. Parameter calibration is also challenging due to the complex structures of these models.

Attempts have been made to integrate more human knowledge in data driven models. Khaki et al. (2020) designed a novel machine learning model that uses convolutional neural networks to extract interactions between weather and soil variables and recurrent neural networks to capture time dependencies of genetic improvement of seeds. Using data from 13 states in the U.S. Corn Belt, with 1980 to 2015 being training years, the model achieved RMSEs of 16.48, 15.74, and 17.64 bu/ac for the respective test years of 2016, 2017, and 2018. Coupled with the backpropagation method, the model could reveal the extent to which weather conditions, accuracy of weather predictions, soil conditions, and management practices were able to explain the variability in the crop yields. Several other studies have used regression models with manually extracted features to estimate effects of temperature (Zhao et al., 2017; Butler et al., 2018; Tigchelaar et al., 2018), solar brightening (Tollenaar et al., 2017), plant density (Lacasa et al., 2020), and flowering time (Parent et al., 2018) on crop yield.

We propose a new model for not only predicting crop yield but also attributing spatial and temporal yield variability to contributions of genetics, environment, and management variables. This is also a largely data driven model, but the model structure was specifically designed to incorporate basic agronomic knowledge to ensure that parameters from the trained model can be used to explain temporal and spatial yield variability with respect to genetic, environment, and management components. In contrast, parameters from most machine learning models are typically much more numerous and harder to provide meaningful interpretations beyond predicted yield. A large set of weather, soil, and management data were collected from public sources, which covered 41 states in the U.S.

from 1980 to 2019. The model was designed for this data set in order to strike a balance among four competing objectives. First, modeling resolution takes full advantage of available data to accurately quantify the effects of daily weather changes and variability in soil conditions and management practices on crop yield. Second, modeling structure is based on scientific facts or reasonable simplifying assumptions. Third, modeling results can be used to explain causes of temporal and spatial yield variability, allowing users to either gain meaningful insights or pinpoint flaws in specific model components for further improvement. Fourth, modeling parameters can be optimally calibrated using state-of-the-art machine learning and optimization techniques to extract useful information from data that is transferable temporally and spatially. We focused on maize yield prediction because maize is one of the most important food, feed, and fuel crops in the U.S., and its production has demonstrated temporal and spatial variability and strong sensitivity to both environmental and management conditions (Meng et al., 2016).

2. DATA

This section provides details on the data we collected for this study. More detailed weather, soil, or management data that were only available at smaller temporal and spatial scales were not included in this study. Moreover, we were not able to locate publicly accessible data sources for seed genetics. Instead, we used modeling techniques to estimate the genetics component in crop yield based on available data sets. This was achieved by using a polynomial function to explain part of the historical yield that could not be explained by spatial or temporal variability in soil and weather conditions. More details about the modeling approach can be found in section 3.

2.1. Yield and Geography Data

County level corn yield in the U.S. from 1980 to 2019 were collected from NASS (2020). After removing a few data points with incomplete information, the entire data set contained 2,649 counties in 295 crop reporting districts of 41 states (all 50 states in the U.S. excluding AK, CT, HI, MA, ME, NV, NH, RI, and VT; the list of the 41 states can be found in the horizontal axis of **Figure 19**) with a total of 78,169 corn yield records for county-year combinations. **Figure 1** shows the temporal trend of national average corn yield (as well as areas planted) from 1980 to 2019. **Figure 2** visualizes the spatial variability of county level average corn yield in 6 representative years: 1980, 1990, 2000, 2010, and 2019 in approximate 10-year time lapse plus 2012, in which year severe drought resulted in historic reductions in corn yield.

Shape files of U.S. counties were collected from National Weather Service (2020). This information was used to determine the membership of counties in crop reporting districts and states and also to locate weather stations and soil map units for calculating average weather and soil variables within each county.

2.2. Weather Data

Daily surface weather data on a 1-km grid from 1980 to 2019 were collected from Daymet (Thornton et al., 2020). The data set included 7 variables, the names and descriptions of which from Thornton et al. (2020) are summarized as follows.

- dayl: duration of the daylight period in seconds per day.
- prcp: daily total precipitation in millimeters per day, sum of all forms converted to water-equivalent.
- srad: incident shortwave radiation flux density in watts per square meter, taken as an average over the daylight period of the day.
- swe: snow water equivalent in kilograms per square meter.
- tmax: daily maximum 2-meter air temperature in degrees Celsius.
- tmin: daily minimum 2-meter air temperature in degrees Celsius.
- vp: water vapor pressure in pascals.

Figure 3 shows the temporal trends of national averages of the 7 weather variables from 1980 to 2019. **Figure 4** shows the spatial variability of these weather variables between planting and harvesting weeks in 2019.

2.3. Soil Data

Soil data were collected from the latest version of Gridded Soil Survey Geographic (gSSURGO) Database released in July 2020 (USDA, 2020). We used 10 soil variables measured in the Value Added Look Up Table (Valu1) in the database, the names and descriptions of which from USDA (2020) are summarized as follows.

- aws: available water storage, expressed in mm, the volume of plant available water that the soil can store in this layer based on all map unit components. This variable was measured at 11 standard depth layers: standard zone 1 (0–5 cm depth), layer 2 (5–20 cm depth), layer 3 (20–50 cm depth), layer 4 (50–100 cm depth), layer 5 (100–150 cm depth), layer 6 (150 cm to the reported depth of the soil profile), zone 2 (0–20 cm depth), zone 3 (0–30 cm depth), zone 4 (0–100 cm depth), zone 5 (0–150 cm depth), and total soil profile (0 cm to the reported depth of the soil profile).
- tka: thickness of soil components, expressed in cm for the available water storage calculation. This variable was measured at 11 standard depth layers.
- soc: soil organic carbon stock estimate, expressed in grams C per square meter. This variable was measured at 11 standard depth layers.
- tks: thickness of soil components, expressed in cm for the soil organic carbon calculation. This variable was measured at 11 standard depth layers.
- nccpi3corn: National Commodity Crop Productivity Index for Corn (weighted average). Values range from 0.01 (low productivity) to 0.99 (high productivity).
- pctearthmc: National Commodity Crop Productivity Index for major earthy components, which are those soil series or higher level taxa components that can support crop growth.
- rootznemc: Root zone depth, expressed in mm, is the depth within the soil profile that commodity crop roots can effectively extract water and nutrients for growth.

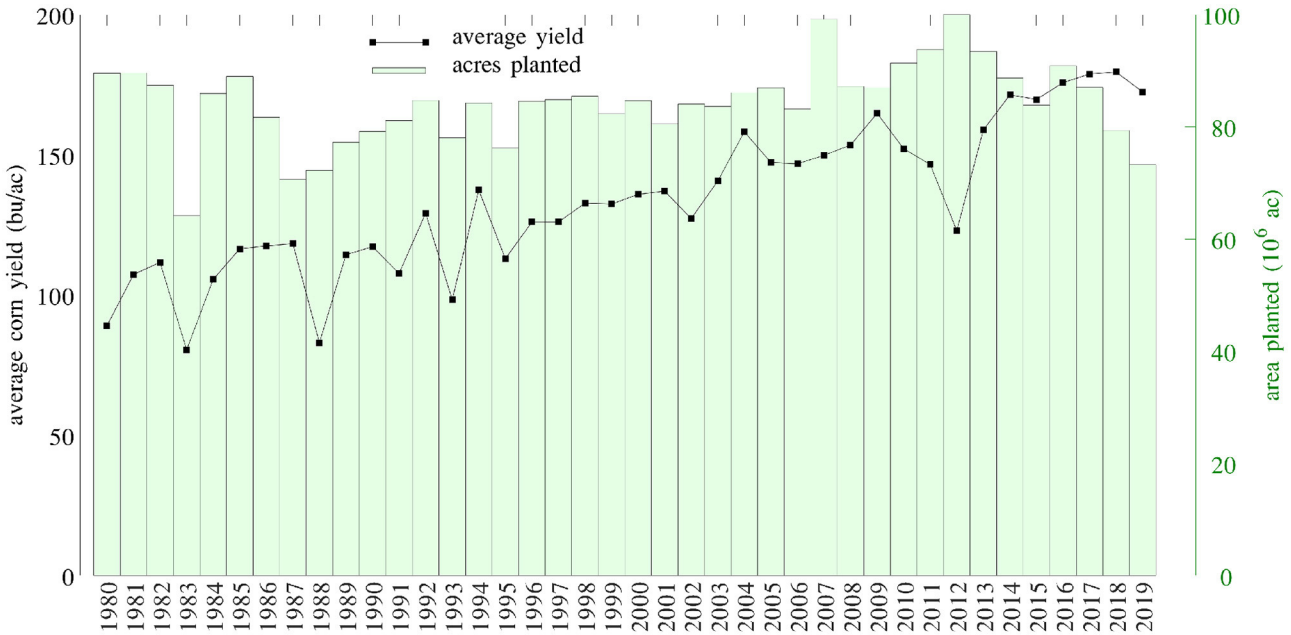

FIGURE 1 | National average corn yield and areas planted from 1980 to 2019.

- rootznaws: Root zone available water storage estimate, expressed in mm, is the volume of plant available water that the soil can store within the root zone based on all map unit earthy major components.
- droughty: Drought vulnerable landscapes comprise those map units that available water storage within the root zone for commodity crops is less than or equal to 6 inches (152 mm), expressed as "1" for a drought vulnerable soil landscape map unit or "0" for a non-droughty soil landscape map unit.
- pwsl1pomu: Potential Wetland Soil Landscapes is expressed as the percentage of the map unit that meets the PWSL criteria.

Figure 5 shows the spatial variability of these 10 soil variables.

2.4. Management Data

Data for areas planted in the 2,649 counties in the U.S. from 1980 to 2019 were collected from NASS (2020). The temporal trend of this information was shown in **Figure 1** together with the yield trend. **Figure 6** visualizes the spatial variability of areas planted in 6 representative years. Due to limited data availability, several important management variables were not included in the model, such as seed genotype, irrigation, fertilization, tillage, and disease/weed control.

Data for corn plant population density (number of plants per acre) in the U.S. from 1980 to 2019 were collected from NASS (2020). Data were available at the state level with more than 60% missing values, and we used mean of non-missing data (other years for the same state, if available) for data imputation. **Figure 7** visualizes the spatial variability of plant density in 6 representative years. Imputed data were not shown in the figure.

Planting and harvesting time data from 1980 to 2019 were collected from NASS (2020), which were at the state level given as percentages of planted areas having finished planting or harvesting in each week. Although 30% data were missing, the format of the planting and harvesting time data allowed imputation to be done in a relatively straightforward manner using the mean of non-missing data. Fortunately, the planted areas of corn belt states with more complete data far exceeded other states with more missing data, as shown in **Figure 19**, which limited the negative impact of the missing data on this case study. **Figures 8**, **9** visualize the spatial variability of planting and harvesting times in 6 representative years (1980 was replaced with 1981 because no harvesting data were available for 1980).

3. MODEL

The model was designed for explaining temporal and spatial variability of corn yield in the U.S. using available data summarized in section 2. Compared with existing crop yield prediction models, this model has three salient features. First, it integrates domain knowledge of plant science in the design of the model. Second, it deploys advanced machine learning and optimization algorithms as solution techniques. Third, it was designed to make data-driven discoveries on the interactions among genetics, environment, and management variables, which must be validated in 2,649 counties in 41 states over the past 40 years. We refer to this model as the GEM model, due to its capability to dissect and quantify (G)enetics, (E)nvironment, and (M)anagement components of crop yield.

3.1. Assumptions

Assumption 1: Crop yield is jointly determined by three mutually independent components: genetics,

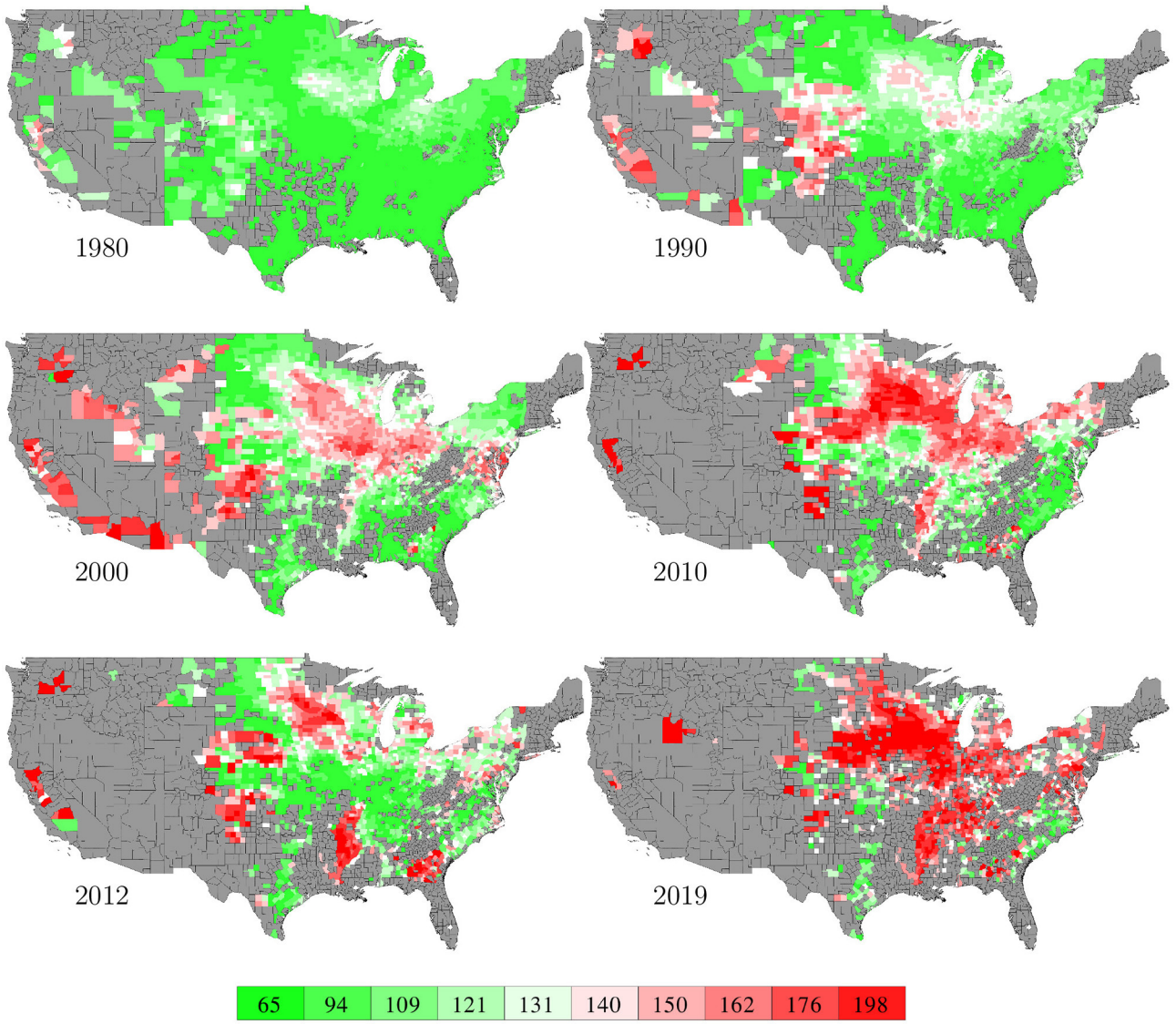

FIGURE 2 | County level average corn yield in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations from 1980 to 2019. Counties without yield data are shown in gray.

environment, and management. The environment component, determined by weather and soil variables and their interactions, sets a growth potential. The genetics and management components reflect the proportion of growth potential that is actually captured and converted to crop yield.

Assumption 2: The genetics component within a crop reporting district is the same for each year. This assumption enables the model to capture the effects of seed selections in different geographic regions; it also gives each crop reporting district sufficient data to learn a separate model for its own genetics trend. The change of genetic performance over the previous year is within a subjectively estimated range of $[-2.5\%, 5\%]$.

Assumption 3: Environmental and management effects on crop growths during different time periods are additive. As such, the model calculates the amount of actual growth in each week based on G, E, and M variables of a given county, and the sum of the growths over 52 weeks gives the total yield for that county. Similar assumptions are commonly made when analyzing the effect of environmental variables. For example, growing degree days and killing degree days are used to measure the cumulative beneficial and damaging effects, respectively, of thermal time (Butler et al., 2018).

Assumption 4: The amount of crop growth potential achieved by management practices depends on the growth stages of the crop. For simplicity, we consider only two stages: the vegetative and reproductive stages, the division for which is approximated

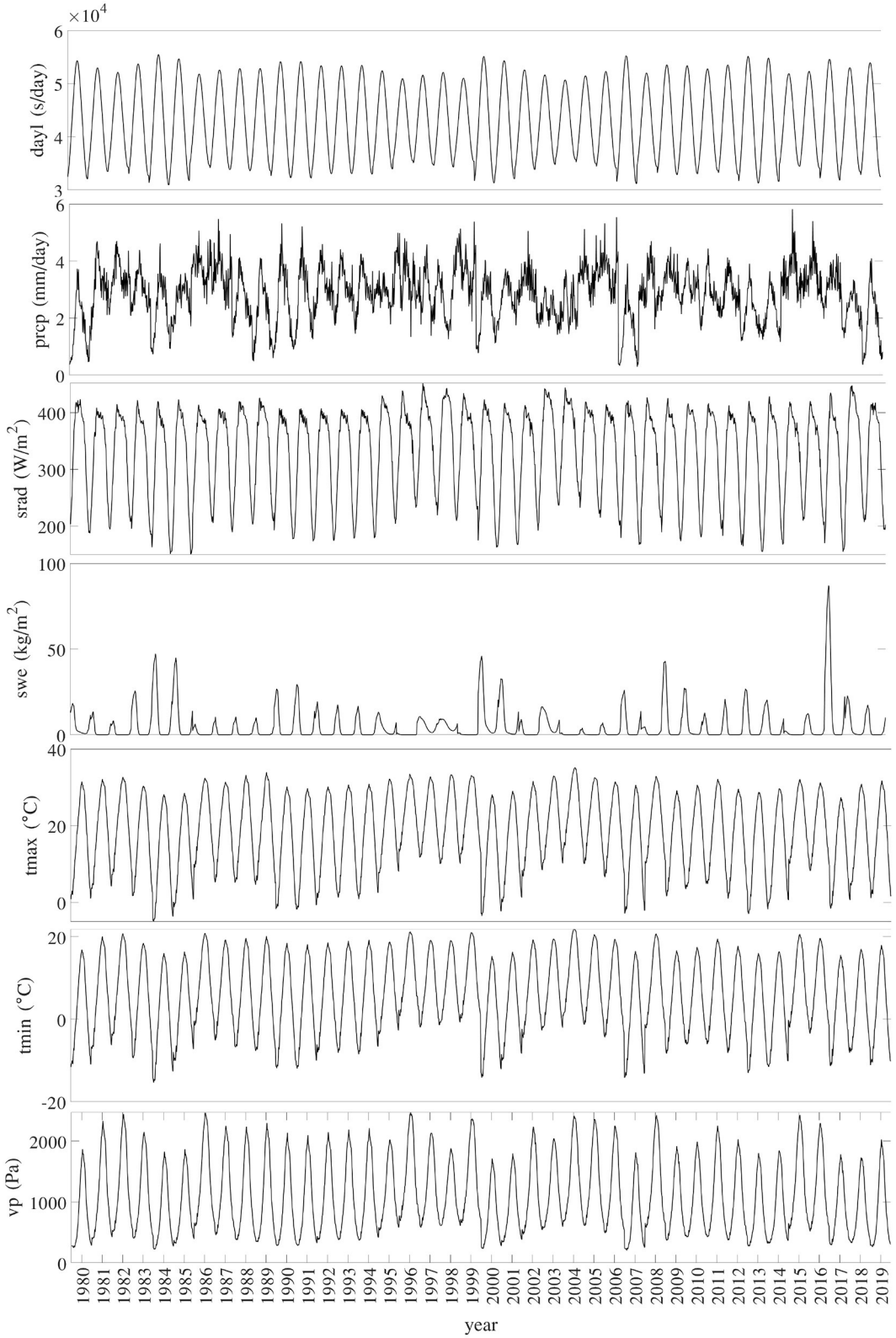

FIGURE 3 | Trend of national averages of 7 weather variables from 1980 to 2019.

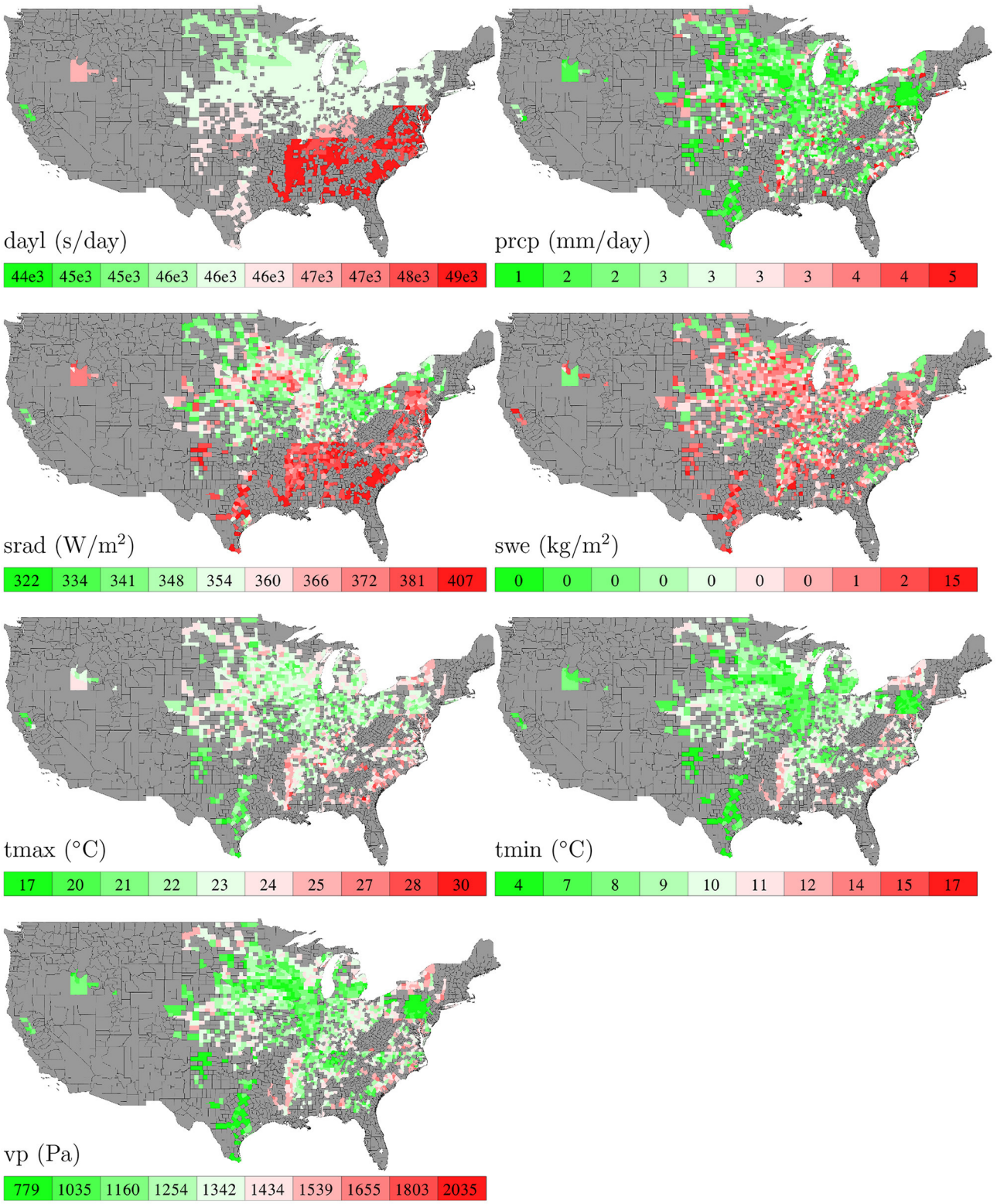

FIGURE 4 | County averages of 7 weather variables between planting and harvesting weeks in 2019. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations from 1980 to 2019. Although weather data are available for all counties in all years, those without yield data are shown in gray.

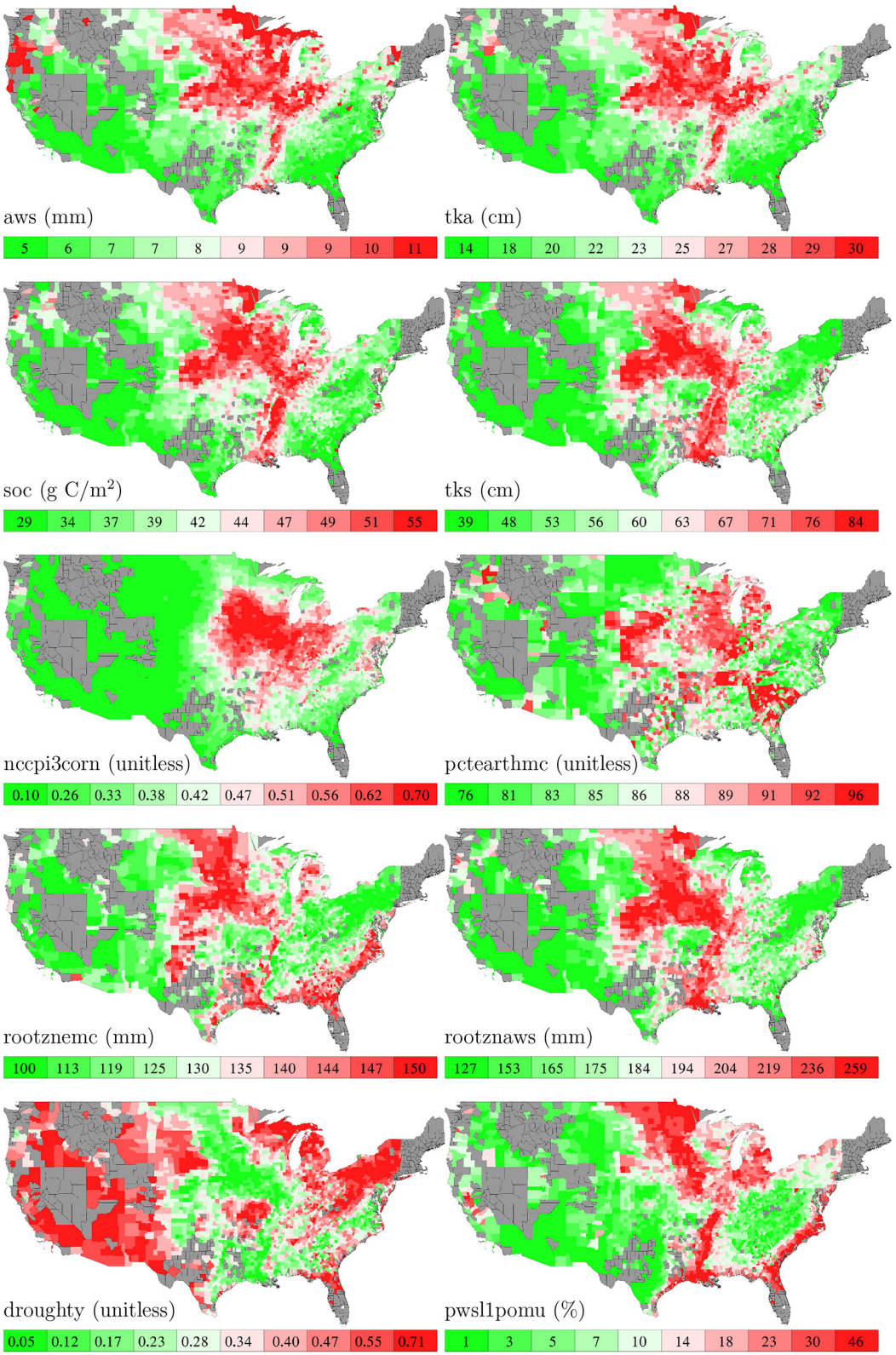

FIGURE 5 | County averages of 10 soil variables. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations. Although soil data are available for all counties, those without yield data are shown in gray.

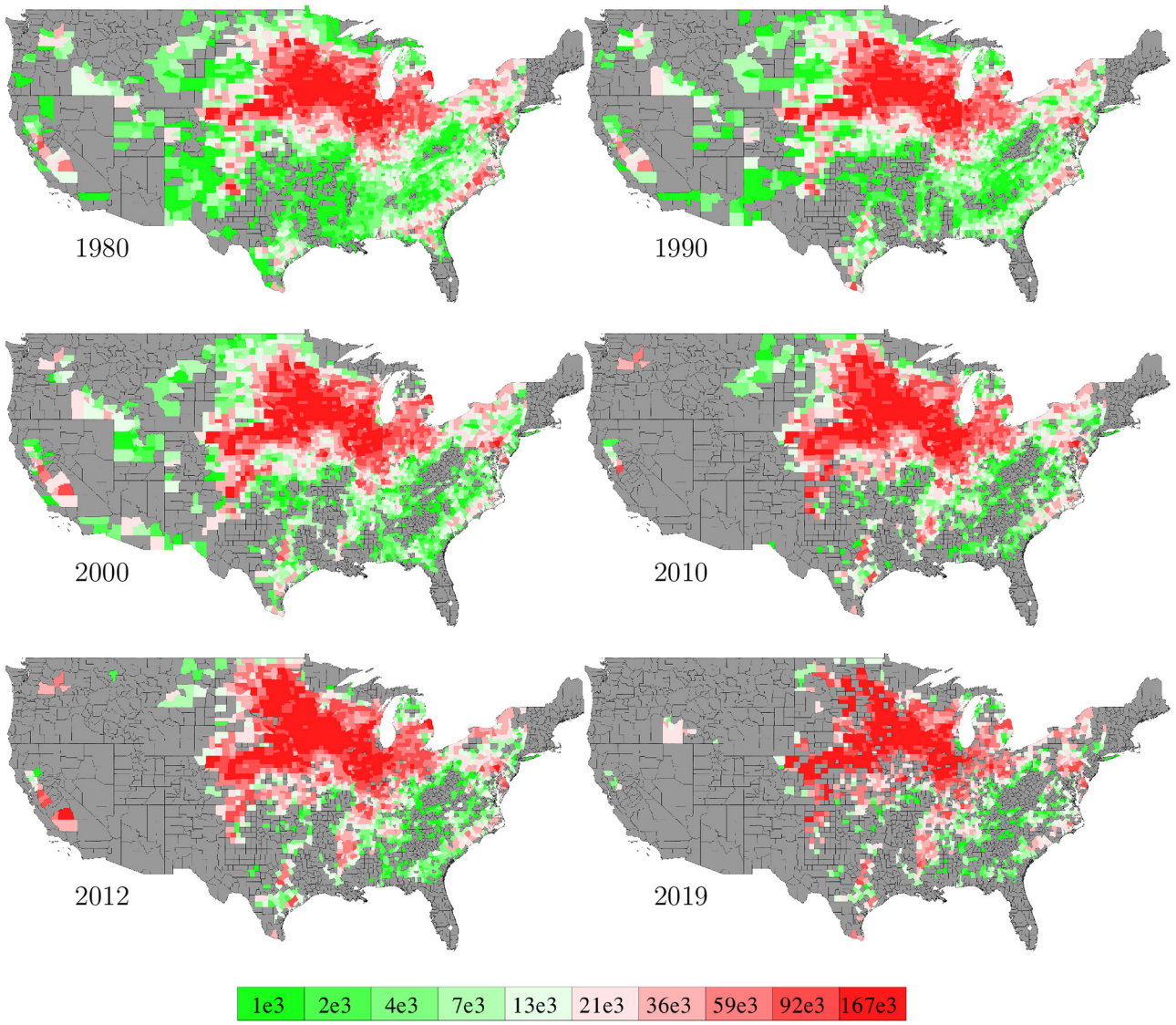

FIGURE 6 | County averages of areas planted in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations. Counties without such data are shown in gray.

as the first week when 100% planting has been finished in a given county.

Assumption 5: Up to quadratic additive effects and bilinear interactions between weather and soil variables are considered. This assumption makes the model formulation relatively simple for computational efficiency yet sufficiently flexible for data-driven discovery.

3.2. Nomenclature

Known parameters:

- $\mathcal{I}$: set of indices for 7 weather variables in the weather data set.
- $\mathcal{J}$: set of indices for 10 soil variables in the soil data set.
- $\mathcal{C}$: set of indices for 2,649 counties in the corn yield data set.
- $\mathcal{R}(c)$: set of indices for all counties that belong to the same crop reporting district as county c.
- $\mathcal{T}$: set of years 1980 to 2019.
- $\mathcal{CT}$: set of 78,169 county-year combinations for which historical corn yield data are available in the collected data set.
- $\mathcal{W}$: set of 52 weeks in a year.
- $\mathcal{W}^{\mathrm{V}}(c,t)$: subset of 52 weeks in year $t \in \mathcal{T}$ that crop in county $c \in \mathcal{C}$ is in or before the vegetative stage.
- $\mathcal{W}^{\mathrm{R}}(c,t)$: subset of 52 weeks in year $t \in \mathcal{T}$ that crop in county $c \in \mathcal{C}$ is in or after the reproductive stage.
- K: the highest polynomial order of genetic progress as a function of time. $K = 10$ was used in this study.
- $A_{c,t}$: area planted in county $c \in \mathcal{C}$ and year $t \in \mathcal{T}$.

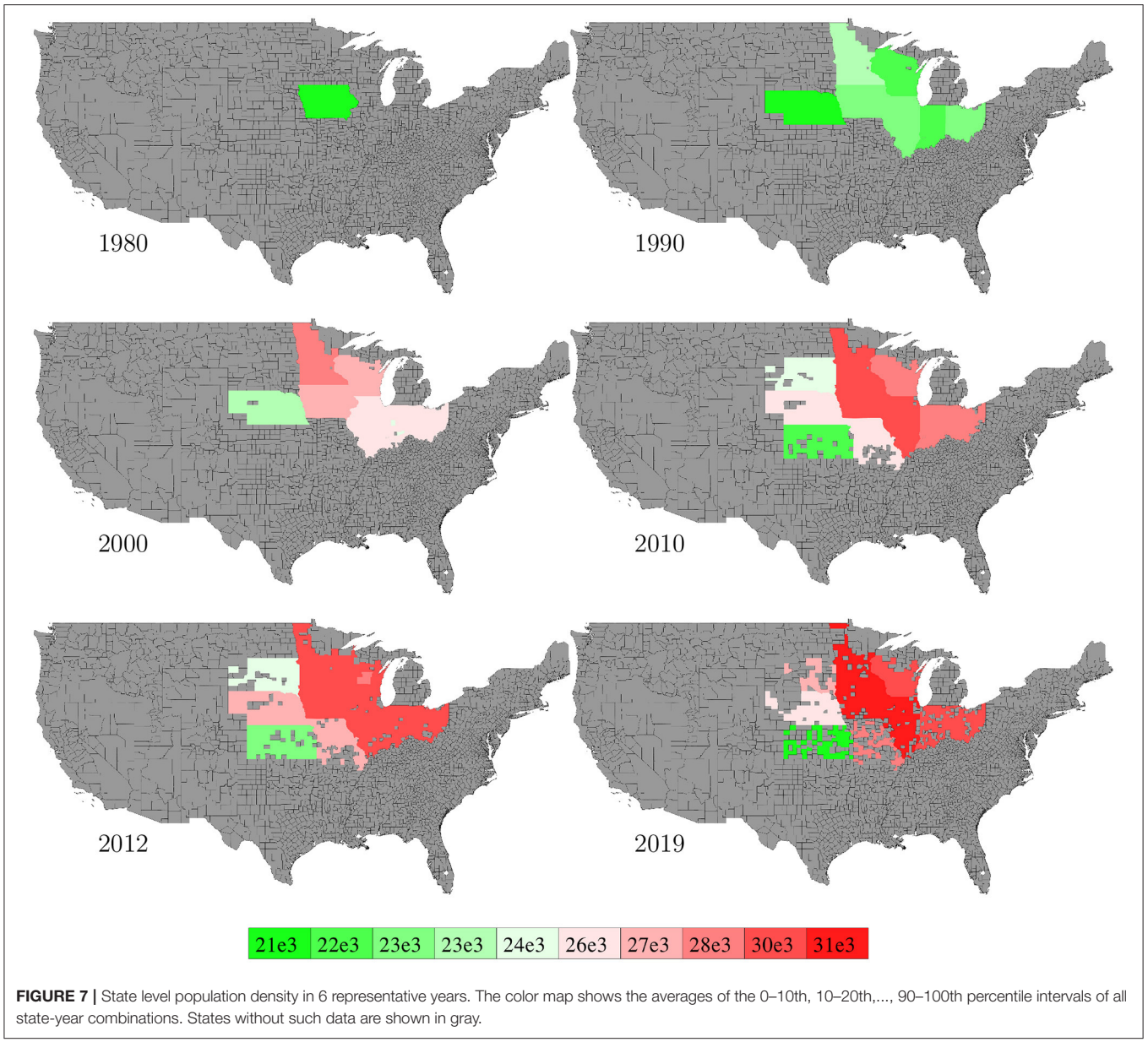

FIGURE 7 | State level population density in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all state-year combinations. States without such data are shown in gray.

Explanatory variables:

- $t \in \mathcal{T}$: year variable.
- $W_{c,i,w}$: weather variable $i \in \mathcal{I}$ in county $c \in \mathcal{C}$ and week $w \in \mathcal{W}$. This variable is averaged over 7 days and across all weather stations within county c. In case no weather stations were located inside a small county, the nearest one was used.
- $S_{c,j}$: soil variable $j \in \mathcal{J}$ in county $c \in \mathcal{C}$. This variable is averaged across all soil map units within county c. In case no soil map units were located inside a small county, the nearest one was used.
- $D_{c,t}$: population density (number of plants per acre) in county $c \in \mathcal{C}$ and year $t \in \mathcal{T}$.
- $P_{c,w}$: percentage of planting finished in county $c \in \mathcal{C}$ by week $w \in \mathcal{W}$, which monotonically increases from 0 to 100% during the planting season and stays at 100% to the end of the year.
- $H_{c,w}$: percentage of harvesting finished in county $c \in \mathcal{C}$ by week $w \in \mathcal{W}$, which monotonically increases from 0 to 100% during the harvesting season and stays at 100% to the end of the year.

Response variables:

- $\hat{y}_{c,t}$: predicted corn yield in county $c \in \mathcal{C}$ and year $t \in \mathcal{T}$.
- $y_{c,t}$: observed corn yield in county $c \in \mathcal{C}$ and year $t \in \mathcal{T}$.

Unknown parameters:

- $\alpha_{c,k}$: genetic progress parameter for t^k in county $c \in \mathcal{C}$.

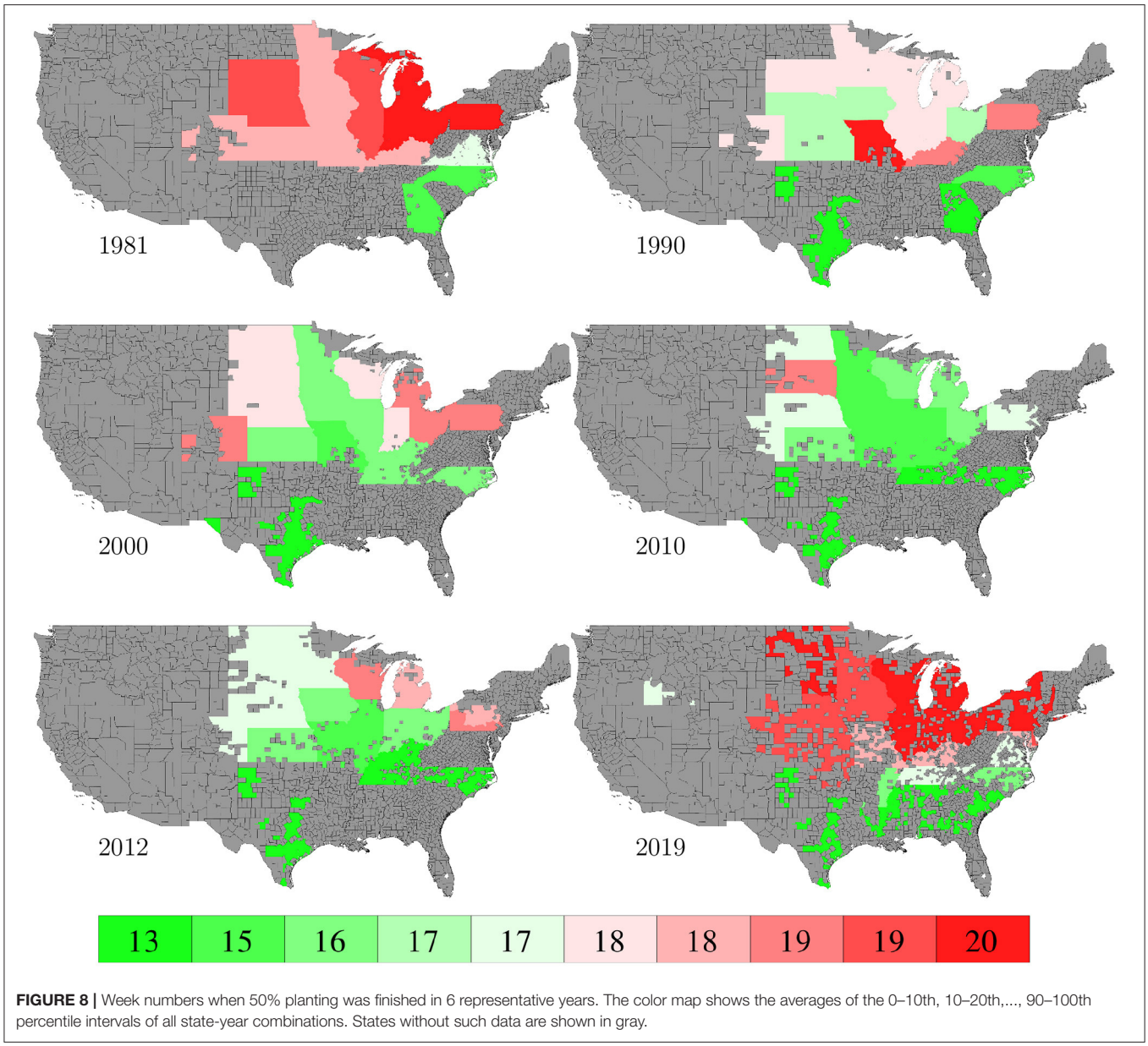

FIGURE 8 | Week numbers when 50% planting was finished in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all state-year combinations. States without such data are shown in gray.

- $\gamma^{V}_{i,j}$: parameter of interaction between (weather and/or soil) variables $i \in \mathcal{I}$ or $\mathcal{J}$ and $j \in \mathcal{I}$ or $\mathcal{J}$ on growth potential during the vegetative stage. In particular, $\gamma^{V}_{0,0}$ is a constant term, $\gamma^{V}_{i,0}$ is the coefficient for linear effect of variable i, $\gamma^{V}_{i,i}$ is the coefficient for quadratic effect of variable i, and $\gamma^{V}_{i,j}$ is the coefficient for bilinear interaction between variables i and j.
- $\gamma^{R}_{i,j}$: parameter of interaction between variables i and j on growth potential during the reproductive stage.

3.3. Crop Yield Model

The GEM model predicts the corn yield in county c and year t as follows:

$$\hat{y}_{c,t} = \left(\sum_{k=0}^{K} \alpha_{c,k} t^k\right) \cdot \left[\sum_{w \in \mathcal{W}(c,t)} D_{c,t}(P_{c,w} - H_{c,w}) G_{c,w}\right]. \quad (1)$$

The term $\sum_{k=0}^{K} \alpha_{c,k} t^k$ estimates the relative genetic performance of seeds in county c and year t using a polynomial function of the year number (normalized to $[0, 1]$). The term $D_{c,t}(P_{c,w} - H_{c,w})$ reflects management practices of plant population density and planting/harvesting progress, which directly affect the amount of growth potential that can be captured and converted to grain yield. In particular, $(P_{c,w} - H_{c,w})$ calculates the percentage of crop in county c and week w that has been planted and not yet harvested, as the crop continues to accumulate growth. The composite variable $G_{c,w}$ calculates the growth potential in county c and week w, defined as $G_{c,w} = \gamma^{V}_{0,0} + \sum_{i \in \mathcal{I}} \gamma^{V}_{0,i} W_{c,i,w} + \sum_{i_1 \le i_2 \in \mathcal{I}} \gamma^{V}_{i_1,i_2} W_{c,i_1,w} W_{c,i_2,w} + \sum_{j \in \mathcal{J}} \gamma^{V}_{0,j} S_{c,j} + \sum_{j_1 \le j_2 \in \mathcal{J}} \gamma^{V}_{j_1,j_2} S_{c,j_1} S_{c,j_2} + \sum_{i \in \mathcal{I}, j \in \mathcal{J}} \gamma^{V}_{i,j} W_{c,i} S_{c,j}$ for all $w \in \mathcal{W}^{V}(c,t)$

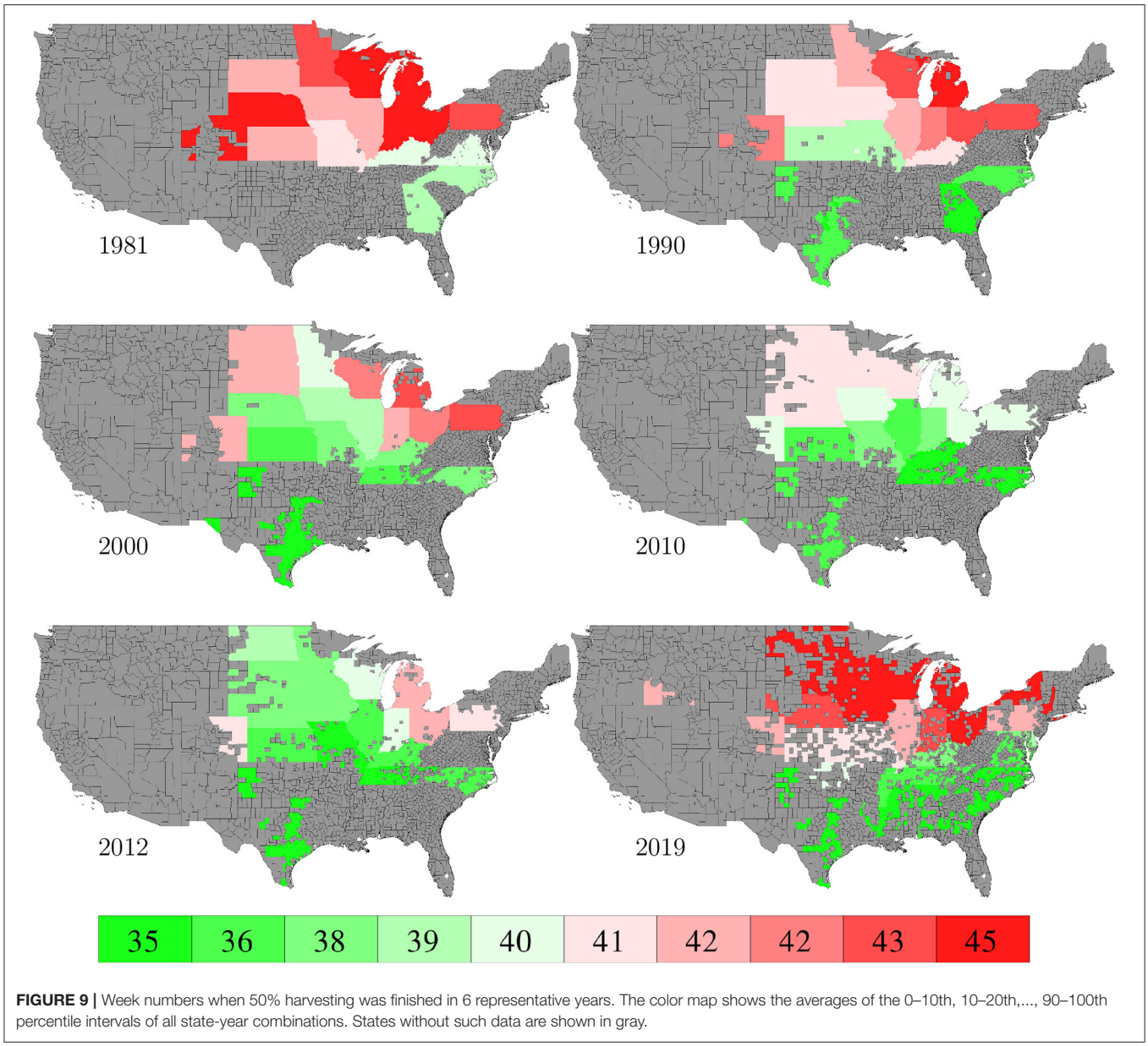

FIGURE 9 | Week numbers when 50% harvesting was finished in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all state-year combinations. States without such data are shown in gray.

and $G_{c,w} = \gamma^{\mathrm{R}}_{0,0} + \sum_{i\in\mathcal{I}} \gamma^{\mathrm{R}}_{0,i} W_{c,i,w} + \sum_{i_1\le i_2\in\mathcal{I}} \gamma^{\mathrm{R}}_{i_1,i_2} W_{c,i_1,w} W_{c,i_2,w} + \sum_{j\in\mathcal{J}} \gamma^{\mathrm{R}}_{0,j} S_{c,j} + \sum_{j_1\le j_2\in\mathcal{J}} \gamma^{\mathrm{R}}_{j_1,j_2} S_{c,j_1} S_{c,j_2} + \sum_{i\in\mathcal{I},j\in\mathcal{J}} \gamma^{\mathrm{R}}_{i,j} W_{c,i} S_{c,j}$ for all $w \in \mathcal{W}^{\mathrm{R}}(c,t)$.

For a given county c and year t, we dissect the crop yield into components G, E, and M as follows.

- Component E is defined as $\sum_{w\in\mathcal{W}(c,t)} \max\{G_{c,w}, 0\}$, which is the maximally achievable growth potential determined by weather and soil variables and their interactions. The max function is used to narrow the range of summation of non-negative weekly growth potential terms during the favorable growing season.
- Component G is defined as $\sum_{k=0}^{K} \alpha_{c,k} t^k$. Year number t and parameter α are normalized so that component G is within $[0, 1]$, indicating the proportion of component E that is achieved by genetics. This component captures what is not explained by environment and management variables, which is mostly due to genetic improvement over time. This component is estimated with a separate polynomial function for each crop reporting district.
- Component M is defined as $\dfrac{\sum_{w\in\mathcal{W}(c,t)} D_{c,t}(P_{c,w} - H_{c,w}) G_{c,w}}{\sum_{w\in\mathcal{W}(c,t)} \max\{G_{c,w}, 0\}}$, where parameters D, P, and H are all normalized to $[0, 1]$. As

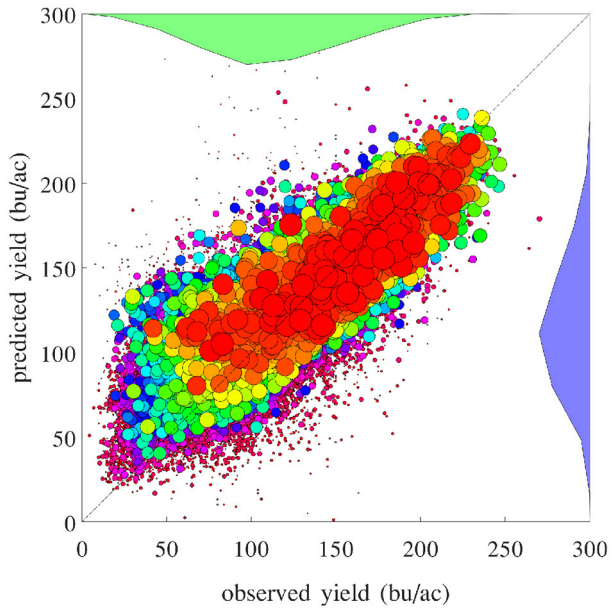

FIGURE 10 | Predicted vs. observed corn yield for all 78,169 county-year combinations are plotted against the 45 degree line, with the sizes of the dots approximately proportional to the areas planted. Histograms of the two sets of yield are also shown.

such, component M reflects the proportion of component E that is captured by management.

By definition, predicted yield is equal to the product of components G, E, and M. Notice that component E represents the maximally achievable growth potential from 1980 to 2019 when components G and M are normalized to $[0, 1]$. For future years when both seed genetics and management practices continue to improve, such growth potential may be exceeded.

3.4. Model Performance Evaluation

We used the following definition of the root mean square error (RMSE) to measure the prediction accuracy of the model:

$$r(\alpha, \gamma, \mathcal{CT}) = \sqrt{\frac{\sum_{(c,t)\in\mathcal{CT}} A_{c,t}^2 (y_{c,t} - \hat{y}_{c,t})^2}{\sum_{(c,t)\in\mathcal{CT}} A_{c,t}^2}}. \quad (2)$$

Here, parameters α and γ are used in equation (1) to calculate $\hat{y}$. Parameter $A_{c,t}$, the area planted in county c and year t, is used as the weight.

The proposed model was evaluated based on both descriptive and predictive performances in the case study. The descriptive performance measures how well the model can fit the training data and explain the spatial and temporal variability of crop yield. Results from models (3)-(7) and (8)-(12) can provide insights on weather and soil interactions, trends of components of G, E, and M, and growth potential. The predictive performance evaluates the ability of the model to predict crop yield for unseen years or counties, for which training data have not been used to train the model. For this purpose, we trained the GEM model 40 times in a leave-one-year-out manner to validate its temporal prediction performance and 2,649 times in a leave-one-county-out manner to validate its spatial prediction performance. Moreover, the model was also used to provide in season prediction with daily updates of weather conditions.

Although numerous crop yield prediction models can be found in the literature, most were designed for different sets of explanatory variables and more focused geographic regions or time periods. To provide a meaningful benchmark comparison, we used the nearest-neighbor approach (Cover and Hart, 1967), which is popular for machine learning studies and intuitive in the crop yield prediction context. The nearest-neighbor approach for crop yield prediction was implemented as follows. To predict the yield of county c in an unseen year t, we identified historical yield data for county c in the nearest-year (before or after year t) and used that yield as the prediction. Similarly, to predict the yield of an unseen county c in year t, we identified historical yield data for the geographically nearest-county in the same year t and used that yield as the prediction. As such, the nearest-neighbor approach can be referred to more specifically as nearest-year and nearest-county for temporal and spatial predictions, respectively.

3.5. Algorithm

The GEM model (1) is a complex nonlinear optimization problem that is not readily solvable by standard machine learning algorithms. Herein, we present a heuristic algorithm that can efficiently obtain a high quality solution (without optimality guarantee) for unknown parameters α and γ. The strategy is not to simultaneously optimize α and γ but to iteratively update one of them at a time while keeping the other fixed. As such, solving model (1) reduces to solving two smaller quadratic optimization models multiple times, which are readily solvable by state-of-the-art optimization solvers such as Gurobi (Gurobi Optimization, LLC, 2021) and Cplex (IBM ILOG Cplex, 2009). Detailed steps of the algorithm are explained as follows.

Step 0: Initialization. Pre-process data by normalizing them to $[0, 1]$. Initialize the incumbent α^* as $\alpha^*_{c,0} = 1, \forall c \in \mathcal{C}$, $\alpha^*_{c,k} = 0, \forall c \in \mathcal{C}, k \in \{1, ..., K\}$, and $\gamma^*_{i,j} = 0, \forall i, j$. Go to step 1.

Step 1: Update γ^*. Randomly select a subset $\mathcal{CT}^1 \subset \mathcal{CT}$ with approximately 80% samples. Solve the following quadratic optimization model using $\mathcal{CT}^1$ while keeping the incumbent α^* as a constant.

$$\max_{\gamma, G, \hat{y}} \quad \sum_{(c,t)\in\mathcal{CT}^1} A_{c,t}^2 (y_{c,t} - \hat{y}_{c,t})^2 \quad (3)$$

$$s.t. \quad \hat{y}_{c,t} = \left(\sum_{k=0}^{K} \alpha^*_{c,k} t^k\right) \cdot \left[\sum_{w\in\mathcal{W}(c,t)} D_{c,t}(P_{c,w} - H_{c,w}) G_{c,w}\right] \quad \forall (c,t) \in \mathcal{CT}^1 \quad (4)$$

$$G_{c,w} = \gamma^{\mathrm{V}}_{0,0} + \sum_{i\in\mathcal{I}} \gamma^{\mathrm{V}}_{0,i} W_{c,i,w} + \sum_{i_1 \le i_2 \in \mathcal{I}} \gamma^{\mathrm{V}}_{i_1,i_2} W_{c,i_1,w} W_{c,i_2,w} + \sum_{j\in\mathcal{J}} \gamma^{\mathrm{V}}_{0,j} S_{c,j} + \sum_{j_1 \le j_2 \in \mathcal{J}} \gamma^{\mathrm{V}}_{j_1,j_2} S_{c,j_1} S_{c,j_2}$$

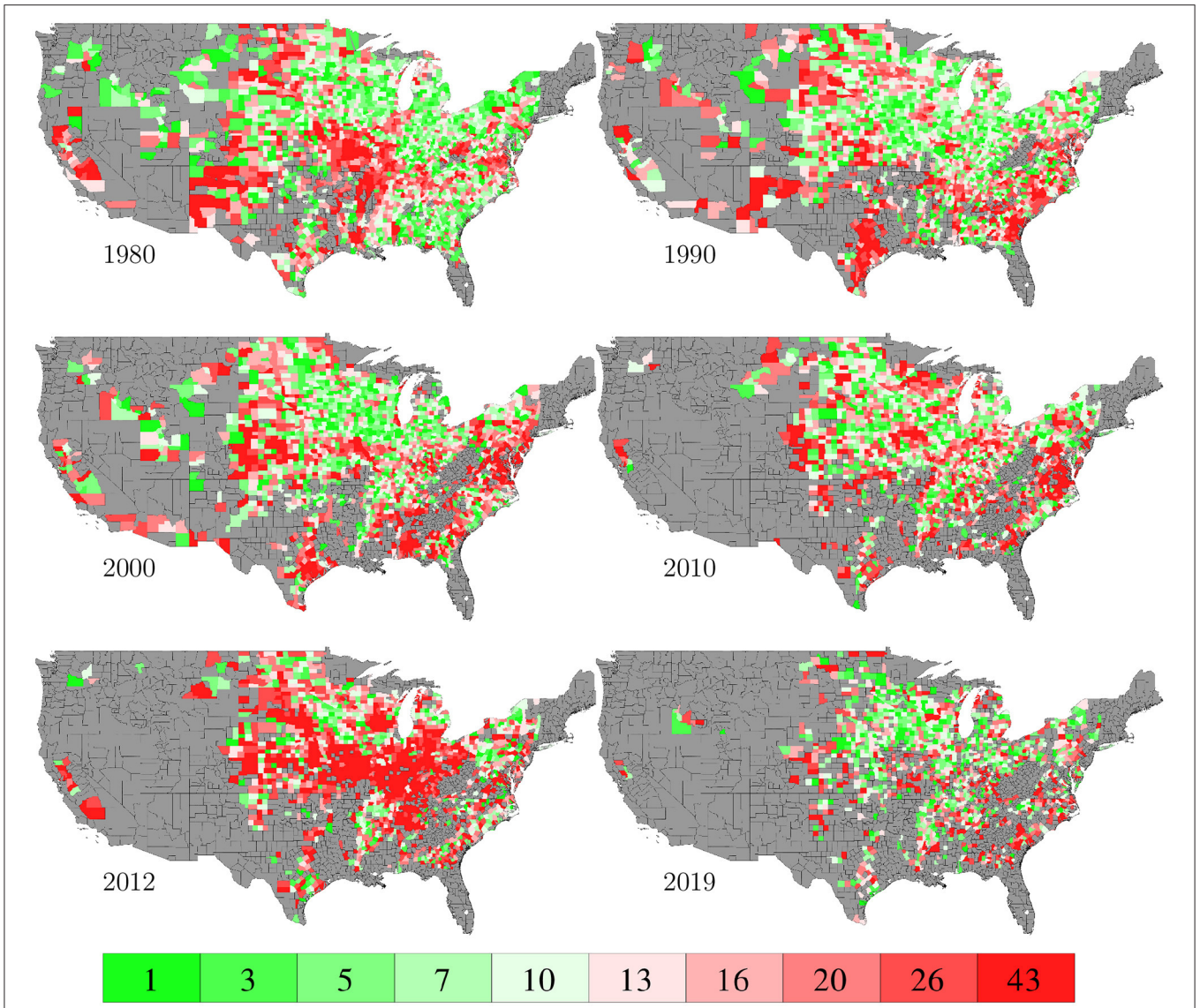

FIGURE 11 | Training RMSEs. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations, weighted by the areas planted. Counties without yield data are shown in gray.

$$+ \sum_{i\in\mathcal{I},j\in\mathcal{J}} \gamma^{\mathrm{V}}_{i,j} W_{c,i} S_{c,j} \quad \forall w \in \mathcal{W}^{\mathrm{V}}(c,t) \tag{5}$$

$$G_{c,w} = \gamma^{\mathrm{R}}_{0,0} + \sum_{i\in\mathcal{I}} \gamma^{\mathrm{R}}_{0,i} W_{c,i,w} + \sum_{i_1\le i_2\in\mathcal{I}} \gamma^{\mathrm{R}}_{i_1,i_2} W_{c,i_1,w} W_{c,i_2,w} + \sum_{j\in\mathcal{J}} \gamma^{\mathrm{R}}_{0,j} S_{c,j} + \sum_{j_1\le j_2\in\mathcal{J}} \gamma^{\mathrm{R}}_{j_1,j_2} S_{c,j_1} S_{c,j_2} + \sum_{i\in\mathcal{I},j\in\mathcal{J}} \gamma^{\mathrm{R}}_{i,j} W_{c,i} S_{c,j} \quad \forall w \in \mathcal{W}^{\mathrm{R}}(c,t) \tag{6}$$

$$0 \le \hat{y}_{c,t} \le 300 \quad \forall (c,t) \in \mathcal{CT}^1 \tag{7}$$

Let $\tilde{\gamma}$ denote an optimal solution and use it to define a new incumbent candidate as $\tilde{\gamma}^* = 0.2\gamma^* + 0.8\tilde{\gamma}$. If $r(\alpha^*, \tilde{\gamma}^*, \mathcal{CT}) < r(\alpha^*, \gamma^*, \mathcal{CT})$, then update $\gamma^* \leftarrow \tilde{\gamma}^*$. Go to step 2.

Step 2: Update α^*. Randomly select a subset $\mathcal{CT}^2 \subset \mathcal{CT}$ with approximately 80% samples. Solve the following quadratic optimization model using $\mathcal{CT}^2$ while keeping G^* determined by the incumbent γ^* as a constant.

$$\max_{\alpha,\hat{y}} \quad \sum_{(c,t)\in\mathcal{CT}^2} A^2_{c,t}(y_{c,t} - \hat{y}_{c,t})^2 \tag{8}$$

$$s.t.\ \hat{y}_{c,t} = \left(\sum_{k=0}^{K} \alpha_{c,k} t^k\right) \cdot \left[\sum_{w\in\mathcal{W}(c,t)} D_{c,t}(P_{c,w} - H_{c,w}) G^*_{c,w}\right] \quad \forall (c,t) \in \mathcal{CT}^2 \tag{9}$$

$$\alpha_{c,k} = \alpha_{d,k} \quad \forall d \in \mathcal{R}(c), k \in \{0,...,K\} \tag{10}$$

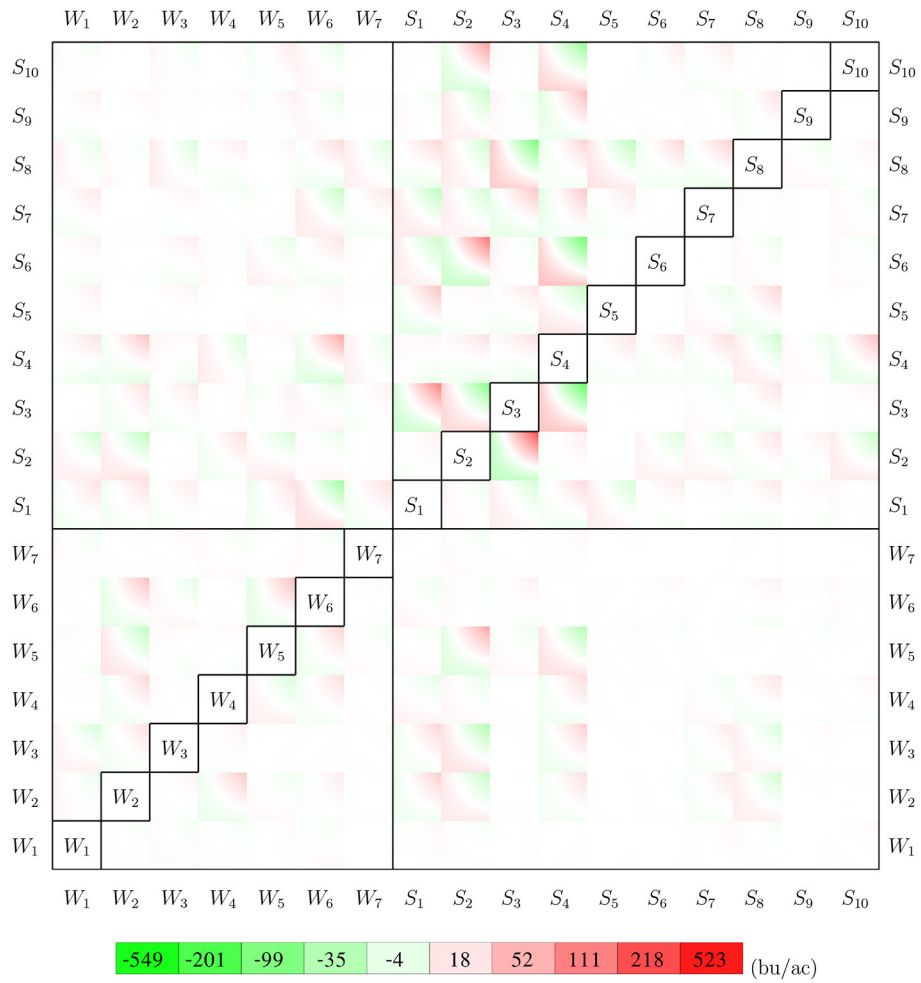

FIGURE 12 | Interactions between 7 weather variables and 10 soil variables. The top left side of the diagonal shows the interactive effects for the vegetative stage, and the bottom right is for the reproductive stage. All weather and soil variables are normalized to [0, 1].

$$0\% \leq \sum_{k=0}^{K} \alpha_{c,k} t^k \leq 100\%$$
$$\forall (c,t) \in \mathcal{CT}^2 \quad (11)$$
$$-2.5\% \leq \sum_{k=0}^{K} \alpha_{c,k} [t^k - (t-1)^k] \leq 5\%$$
$$\forall (c,t) \in \mathcal{CT}^2 \quad (12)$$

Let $\tilde{\alpha}$ denote an optimal solution and use it to define a new incumbent candidate as $\tilde{\alpha}^* = 0.2\alpha^* + 0.8\tilde{\alpha}$. If $r(\tilde{\alpha}^*, \gamma^*, \mathcal{CT}) < r(\alpha^*, \gamma^*, \mathcal{CT})$, then update $\alpha^* \leftarrow \tilde{\alpha}^*$.

Terminate the algorithm if the incumbent solution (α^*, γ^*) has not been updated for two consecutive iterations; otherwise go back to Step 1 for a new iteration.

Remark for model (3)-(7): Minimizing the objective function (3) is equivalent to minimizing the RMSE since the square root function is a monotonically increasing one. With parameter α^* being a constant, all constraints (4)-(7) are linear. Constraint (7) avoids the predicted yield to be negative or unrealistically high.

Remark for model (8)-(12): Constraint (10) requires that all $\alpha_{c,k}$ values in a same crop reporting district be the same. Constraint (11) normalizes the genetic progress within [0%, 100%]. Constraint (12) restricts the change in genetic performance over the previous year to be between −2.5% and 5%. These lower and upper bounds were subjectively estimated to reflect changes in genetic perform, and the optimal γ was found to be insensitive to these parameters.

Remark for incumbent updates: The incumbent solution (α^*, γ^*) is not automatically updated with optimal solution $(\tilde{\alpha}, \tilde{\gamma})$ from the two quadratic optimization models. Rather, it is only partially updated by $(\tilde{\alpha}, \tilde{\gamma})$ if the new incumbent candidate passes a cross validation. In particular, the optimal solutions $\tilde{\alpha}$ and $\tilde{\gamma}$ are obtained using random subsets of $\mathcal{CT}$, then new incumbent candidate solutions are defined as $\tilde{\gamma}^* = 0.2\gamma^* + 0.8\tilde{\gamma}$ and $\tilde{\alpha}^* = 0.2\alpha^* + 0.8\tilde{\alpha}$, which will not be accepted unless they improve the RMSE on the entire data set $\mathcal{CT}$. As such, this technique reduces overfitting by cross validating the generalizability of any updates made to the incumbent solution.

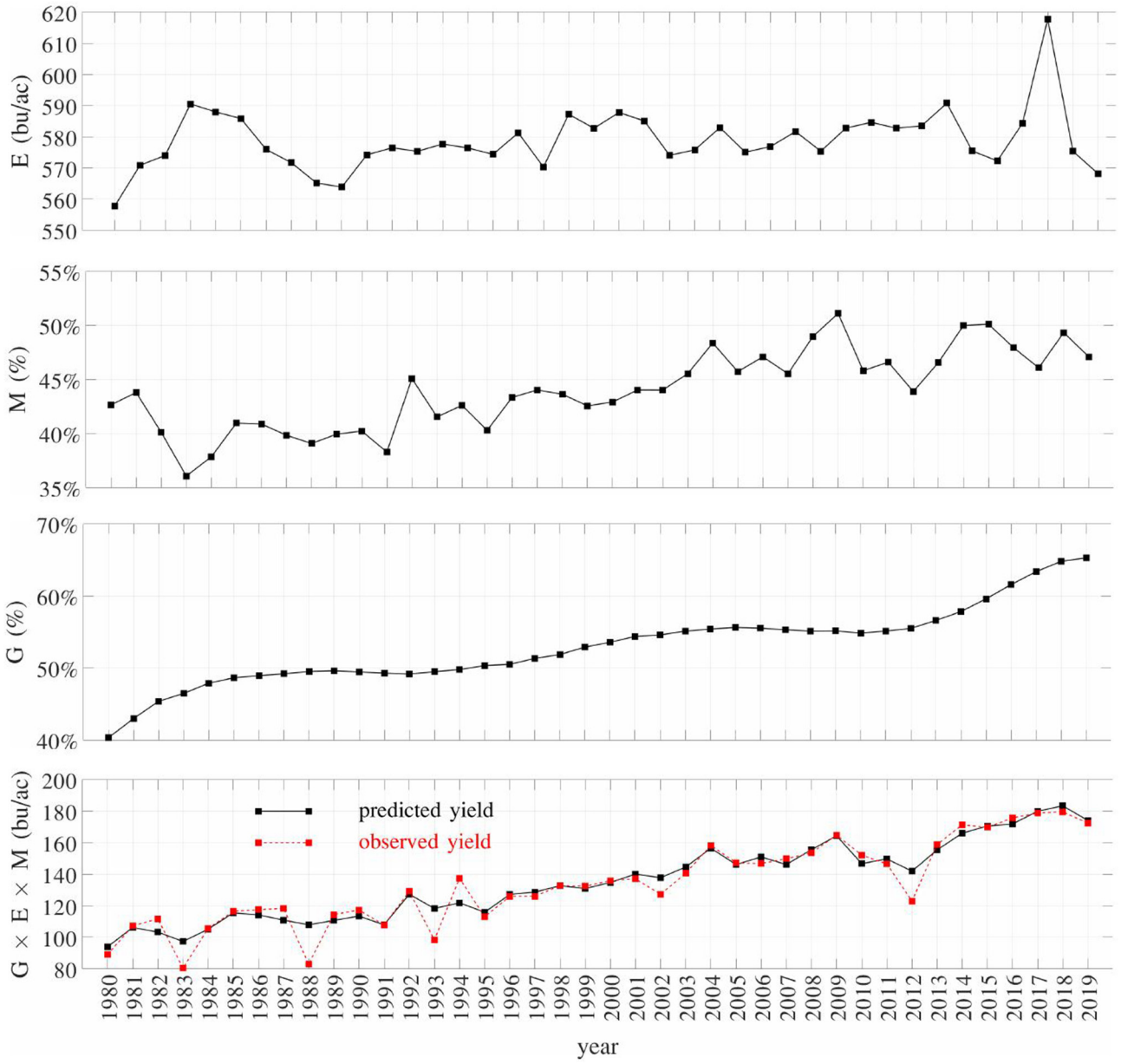

FIGURE 13 | Trend of national averages of G, E, M from 1980 to 2019. The predicted yield is the product of G, E, and M.

Performance of the algorithm was found to be insensitive to the parameters 0.2 and 0.8.

4. RESULTS AND DISCUSSIONS

Computational experiments were carried out using Matlab as the main platform and Gurobi 9.1 as the quadratic programming solver. The proposed heuristic algorithm took approximately 10 min to find a high quality solution on an average laptop. Data and Matlab code used in this study were shared at https://github.com/lzwang2017/maizeyield. Section 4.1 gives the results of fitting the entire data set and section 4.2 presents the results of predicting for unseen years or counties.

4.1. Descriptive Performance

4.1.1. Training Error

The RMSE of fitting the entire data set with the GEM model is 17.84 bu/ac, which is 10.34% of the average yield in 2019. **Figure 10** plots the predicted and observed yield for all 78,169 county-year combinations against the 45 degree line; it also compares the histograms of the two sets of yields. **Figure 11** visualizes the spatial variability of training RMSEs in 6 representative years. In order to offset the potentially misleading

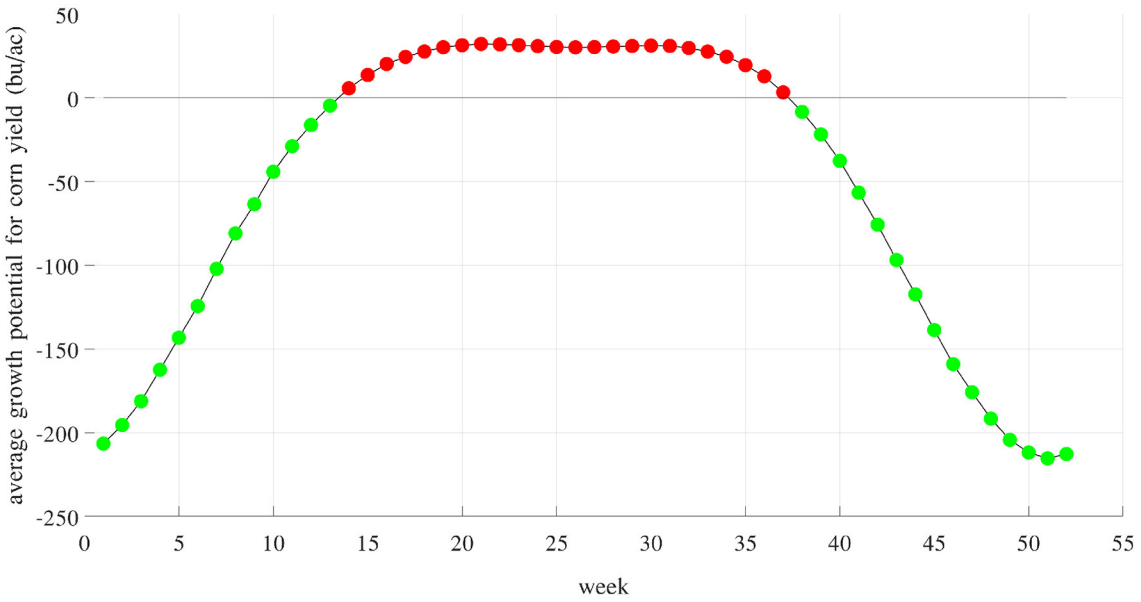

FIGURE 14 | Weekly growth potential determined by weather and soil variables and their interactions. Red and green colors indicate positive and negative effects, respectively. These potentials are positive from week 14 (late March) to week 37 (early September).

discrepancy between the geographic area of a county shown on the map and the area planted in the county, we designed the color map in such a way that each color represents 10 percent of total areas planted. Similarly designed color maps are also used in other figures. **Figure 11** suggests that more than 70% of historical yield data were explained by the GEM model within a 10% relative error; counties with larger planted areas had lower errors than those with smaller planted areas.

4.1.2. Weather and Soil Interactions

Parameter γ from the GEM model reveals how weather and soil variables jointly determine the growth potential in the vegetative and reproductive stages. **Figure 12** visualizes such pair-wise interactive effects with a color map. The color square for variables i and j shows the combined effects of $\gamma_{i,j} + \frac{1}{16}\left(\gamma_{i,0} + \gamma_{i,i} + \gamma_{0,j} + \gamma_{j,j}\right)$. For a given set of observations for 7 weather variables and 10 soil variables of a given week, the growth potential for that week can be calculated using information from **Figure 12** as follows. First, locate the 272 intersections with the 17 variables in vertical and horizontal directions less the diagonal. Then determine if the crop is in vegetative or reproductive stage in the given week. Finally, the growth potential for the vegetative or reproductive stage can be calculated as the summation of the 136 values at the intersections inside the 136 squares on the top left, or bottom right, side of the diagonal, respectively. The asymmetry in the figure reveals how maize may respond differently to the same environmental conditions during the vegetative and reproductive stages.

Results from **Figure 12** can be used to determine the sensitivity of crop yield to combinations of specific weather and/or soil variables. For example, the (S_4, W_6) square suggests that the combination of higher tks and higher tmin in the vegetative stage had an unfavorable effect on crop yield, but the effect of the same combination was negligible in the reproductive stage, as suggested by the (W_6, S_4) square.

4.1.3. Trends of Components G, E, and M

Figure 13 shows the trends of the average components G, E, and M defined in section 3.3, as well as predicted and observed yield from 1980 to 2019. Component E shows an average growth potential of 576 bu/ac, which fluctuates from year to year with a slight increasing trend in the long term and a sharp decrease in recent years. This result is consistent with recent observations by meteorologist Takle and atmospheric scientist Gutowski (Takle and Gutowski, 2020). Both components M and G demonstrate clear increasing trends; the former reflects the improvement in population density and planting/harvesting timing, whereas the latter explains the increasing trend of yield unaccounted for by components E and M. The product of components G, E, and M accurately fits the observed yield except for several years with extreme weathers (e.g., 1983, 1988, 1993, and 2012). This may be caused by the lack of sufficient data with extreme weathers for the model to learn how maize responds to stressful environmental conditions. Improving prediction accuracy under extreme weather conditions has been widely recognized as a challenging yet important topic for future research (van der Velde et al., 2012; Eitzinger et al., 2013; Blanc, 2017; Schauberger et al., 2017), especially in the face of global climate changes.

4.1.4. Growth Potential

Figure 14 shows the average component E for 52 weeks of the year. Averaged across 2,649 counties and 40 years, these

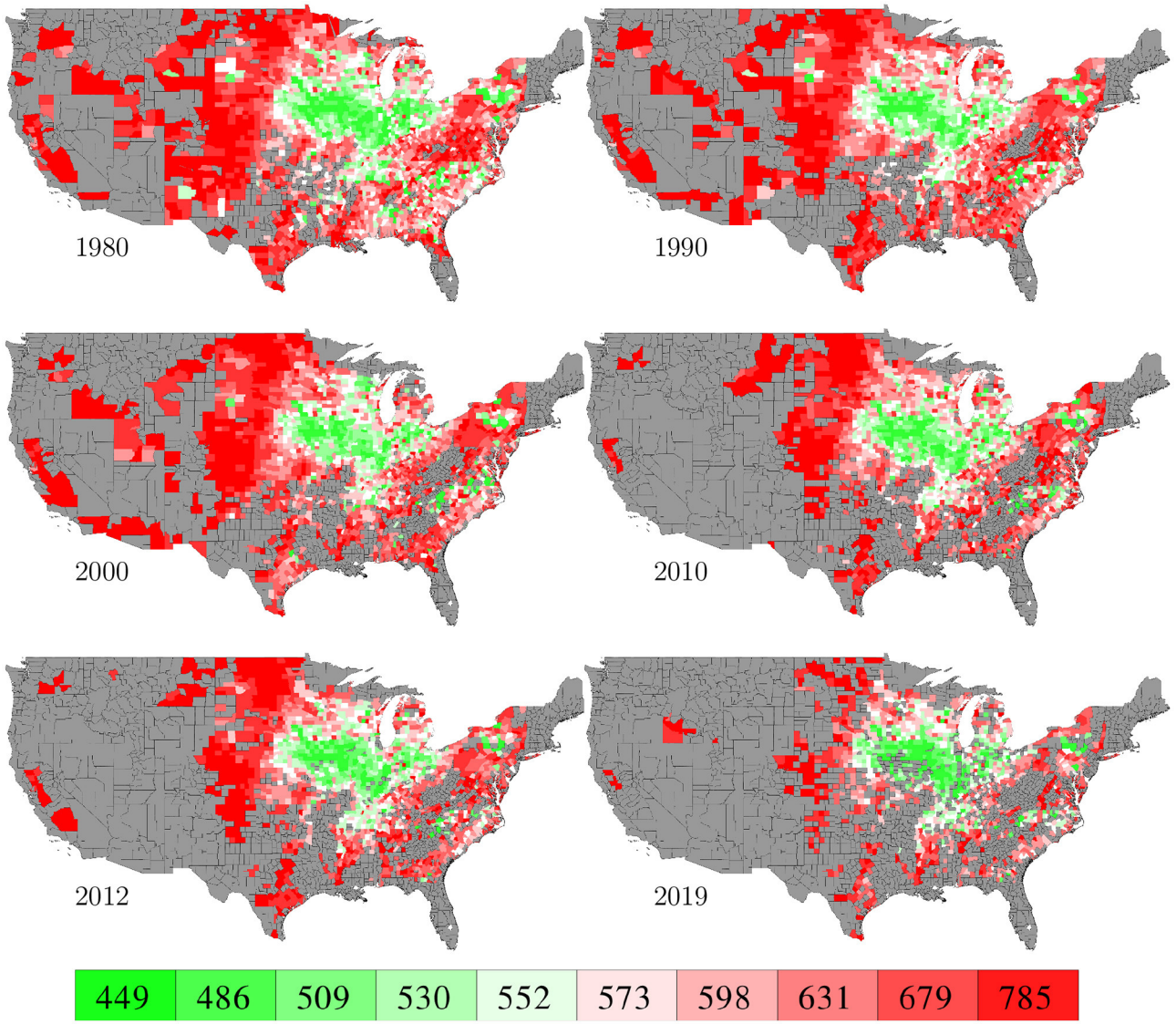

FIGURE 15 | Growth potential determined by weather and soil in 6 representative years. The color map shows the averages of the 0–10th, 10–20th,..., 90–100th percentile intervals of all county-year combinations, weighted by areas planted. Counties without yield data are shown in gray.

growth potentials are positive from week 14 (late March) to week 37 (early September), indicating a favorable time window for maize growth. These results are consistent with prior work that documented the yield benefits of earlier planting and longer season varieties (Kucharik, 2006; Zhu et al., 2018). **Figure 15** visualizes the spatial variability of growth potential in 6 representative years, ranging from 456 to 714 bu/ac. As a reality check, according to the National Corn Growers Association (NCGA, 2020), the winners of the National Corn Yield Contest in 2019 and 2020 achieved, respectively, 616 (new record) and 476 bushels per acre.

4.2. Predictive Performance

4.2.1. Temporal Prediction

Forty experiments were conducted, for 1 year at a time, to test how accurately the GEM model can be used to predict the yield of an unseen year for which historical yield were held out of the training data. Complete weather data for the target year were provided to the prediction model. Section 4.2.3 presents results for in season prediction with daily updated weather information. These forty experiments took approximately seven CPU hours.

Figure 16 compares the performances of temporal prediction of the GEM model and nearest-year approach, in which predicted and observed yields for all 78,169 county-year combinations are

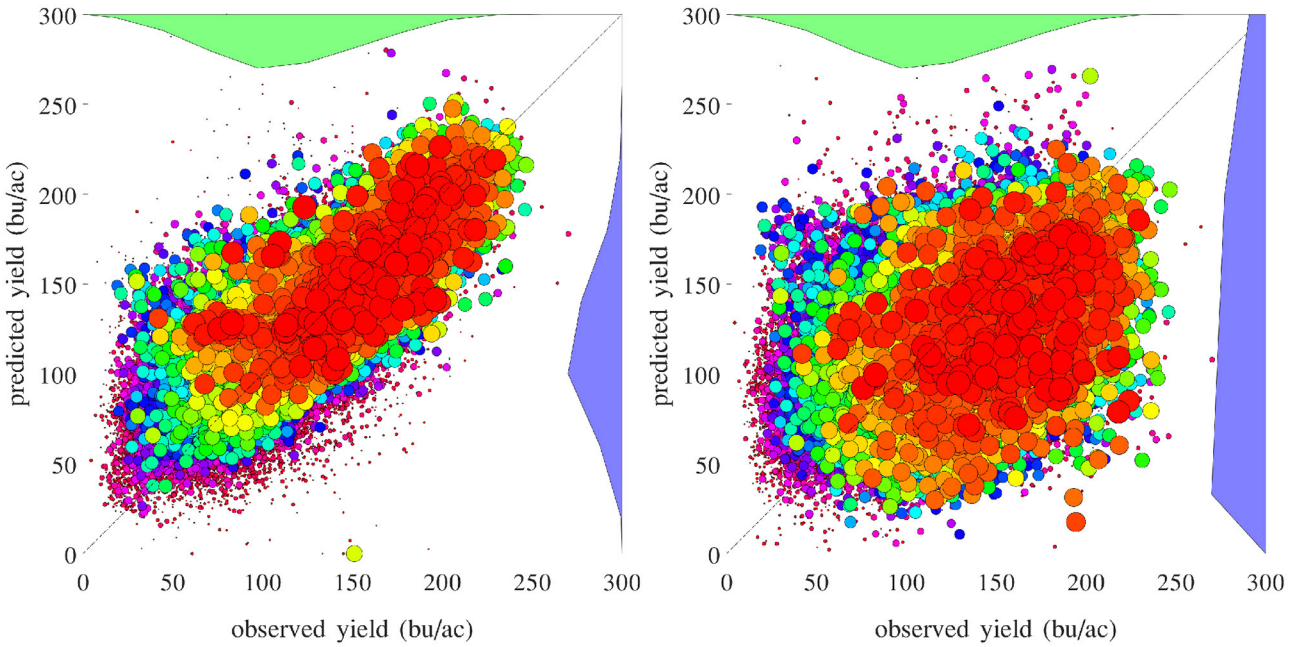

FIGURE 16 | Performances of temporal prediction of the GEM model (left) and nearest-year approach (right). Predicted vs. observed corn yield for all 78,169 county-year combinations are plotted against the 45 degree line, with the sizes of the dots approximately proportional to the areas planted. Histograms of the two sets of yield are also shown.

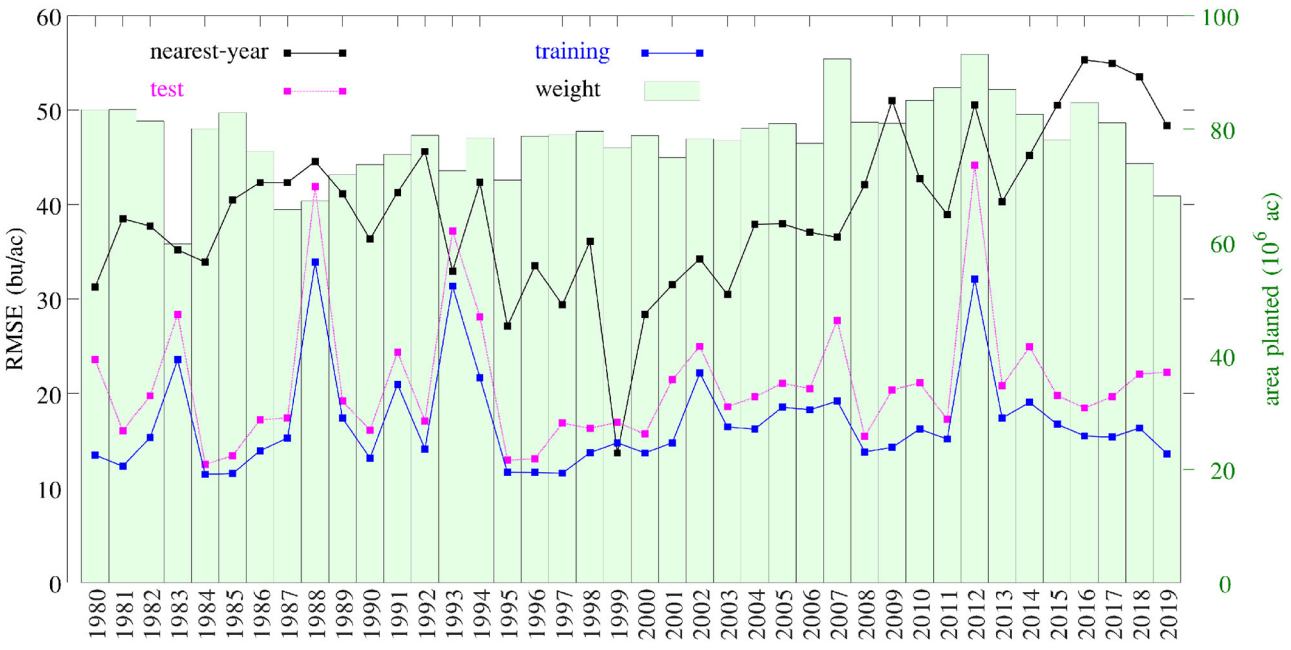

FIGURE 17 | RMSEs of training, test, and nearest-year models. Total planted areas were shown to illustrate the weight for each year.

plotted against the 45 degree line. The RMSEs for these two models were 22.25 and 40.66 bu/ac, respectively. The weighted R^2 values for these two models were 0.79 and 0.30 bu/ac, respectively. **Figure 17** compares the RMSEs for training (using full data), test (leaving 1 year out at a time) using the GEM model, and test using the nearest-year approach from 1980 to 2019. This figure shows that the GEM model clearly outperformed the nearest-year approach. Moreover, the small gap between

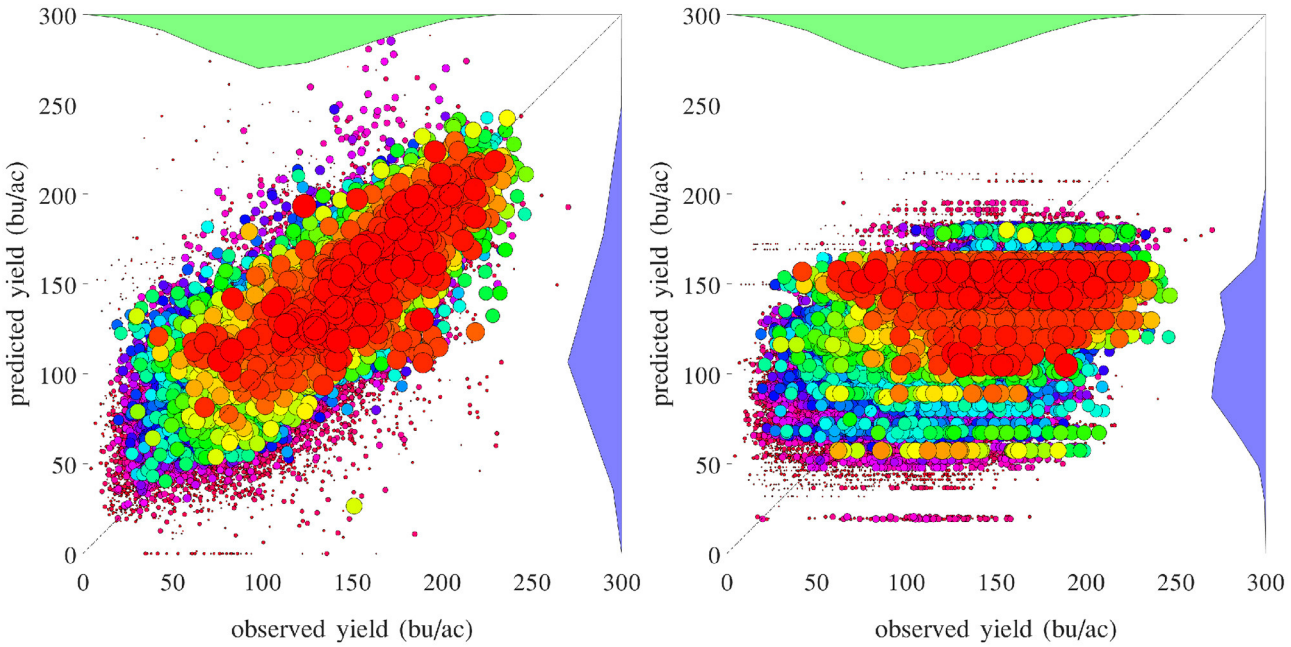

FIGURE 18 | Performances of spatial prediction of the GEM model (left) and nearest-county approach (right). Predicted vs. observed corn yield for all 78,169 county-year combinations are plotted against the 45 degree line, with the sizes of the dots approximately proportional to the areas planted. Histograms of the two sets of yield are also shown.

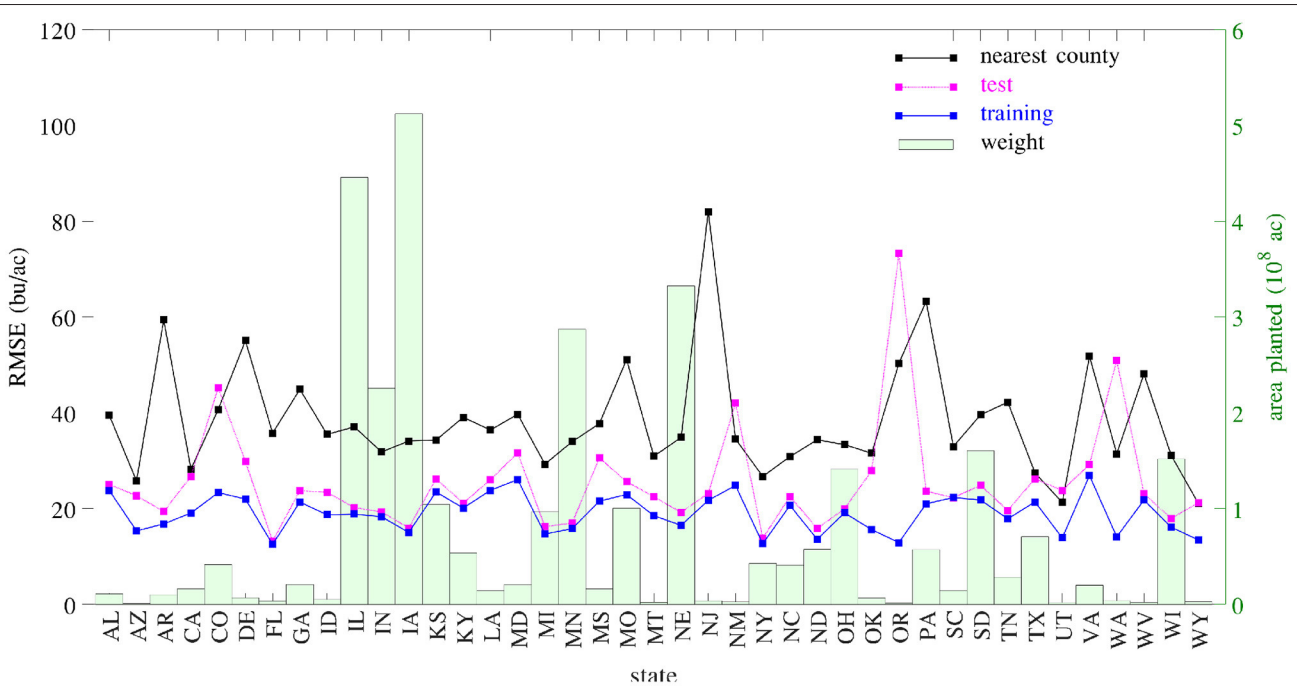

FIGURE 19 | RMSEs of training, test, and nearest-county models. Total planted areas were shown to illustrate the weight for each state.

training and testing errors also indicates very little over-fitting in the model, thanks to the integration of domain knowledge in the design of the model as well as the large data set that allows the model to extract temporally and spatially transferable information. These results suggested that the proposed GEM model can be used to produce crop yield predictions of a known

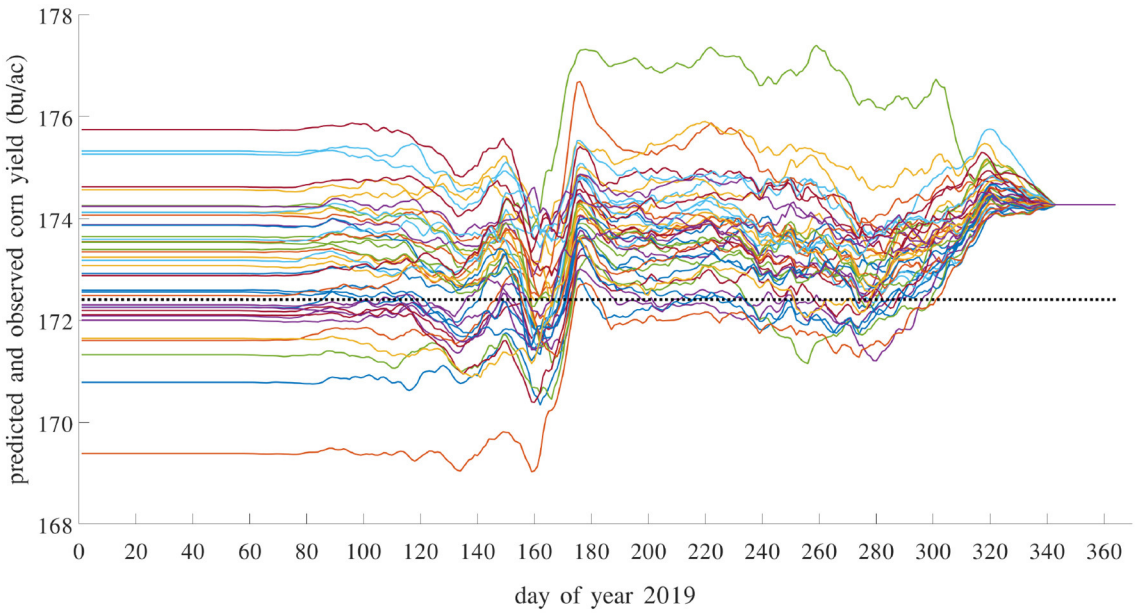

FIGURE 20 | Predicted and observed average yield in 2019 with daily updates. Prediction model was trained using data from 1980 to 2018. There are 39 solid curves, representing 39 predictions using historical weather scenarios from 1980 to 2018, each giving a daily updated prediction using weather data from a historical year to impute the unobserved weather variables in 2019. The dashed flat line is the observed average yield over 1,256 counties with yield data for 2019.

county (with training data) for a new year based on its soil, weather, and management variables.

4.2.2. Spatial Prediction

A set of 2,649 experiments were conducted, for one county at a time, to test how accurately the GEM model can be used to predict the yield of an unseen county for which historical yield were held out of the training data. Complete weather data for all years were provided to the prediction model. These experiments took approximately 18 CPU days. **Figure 18** compares the performances of spatial prediction of the GEM model and nearest-neighbor approach, in which predicted and observed yields for all 78,169 county-year combinations are plotted against the 45 degree line. The RMSEs for these two models were 19.97 and 35.67 bu/ac, respectively. The weighted R^2 values for these two models were 0.83 and 0.46 bu/ac, respectively. **Figure 19** compares the RMSEs for training (using full data), test (leaving one county out at a time) using the GEM model, and test using the nearest-county approach for 41 states. The GEM model demonstrated superior performance over the nearest-county approach and very little overfitting. Therefore, the GEM model can be used to produce crop yield predictions for a new county based on its soil, weather, and management variables.

4.2.3. In Season Prediction With Daily Updates

Most existing methods for in season yield prediction use remote sensing data (Teal et al., 2006; Jagmandeep et al., 2020). The proposed GEM model can be used to provide daily updated yield predictions using daily updated weather data. We demonstrated this approach for the test year 2019 using a GEM model trained with data from 1980 to 2018. For each day in 2019, we made a yield prediction by combining the observed weather data (from January 1st to that day) in 2019 with unobserved weather data (from the next day to December 31st) from a historical year. As such, 39 different predictions were made using 39 historical weather scenarios from 1980 to 2018. These predictions differ widely on day 1 and then gradually converge as more actual weather information in 2019 has been observed. These results are shown in **Figure 20**. Similar daily updated crop yield predictions can be produced using the proposed GEM model for any known county with specific soil and management conditions as long as daily weather variables are available.

5. CONCLUSIONS

Crop yield is a complex trait jointly determined by numerous genotype, environment, and management variables and their interactions. Being able to accurately predict crop yield under changing environmental conditions in a wide range of geographic locations is increasingly important to agriculture stakeholders. It also poses a formidable academic challenge, which can only be overcome by integrating insights from crop science with data analytical methodology.

In our attempt to explain the temporal and spatial variability of maize yield in the U.S., we collected a large data set and designed the GEM model to analyze the data. The data covers 40 years of yield, weather, soil, and management information from 41 states. The GEM model was specifically designed for this data set to extract insights that are explainable and transferable both spatially and temporally.

Computational results suggest that the GEM model is a reasonable attempt to combine the strengths of data driven models in prediction accuracy and the advantage of knowledge driven models in explainability. Compared with data driven models in the literature, the GEM model achieved a prediction

accuracy on par with state-of-the-art machine learning models, thus the advantage of the GEM model is the explainable insights. For example, predicted yield is dissected into genetics, environment, and management components. Compared with knowledge driven models in the literature, the GEM model has a more flexible modeling structure that allows unknown parameters to be efficiently and optimally calibrated using advanced computational techniques, extracting more data-driven information and less human biases.

The GEM model has several limitations and caveats. First, the model was specifically designed for the current data set and not directly applicable to other data sets without necessary modifications, although similar modeling and solution techniques will still be applicable. Second, several assumptions listed in section 3.1 are the backbone of the model, which allow the model to be constructed and solved efficiently but may also limit the effectiveness of the model to a certain extent. Third, several important variables (such as seed genotype, irrigation, fertilization, tillage, and disease/weed control) were not included in the model due to lack of public data sources at the desired scale.

Several followup research directions deserve further investigation. Similar modeling and solution techniques can be applied to other crops, other regions, other time periods, and other data sets. More crop growth stages can be considered to incorporate more crop physiological insights and to give the GEM model additional features. Further effort should also be made to train the model to learn from low frequency but high impact extreme weather scenarios.

DATA AVAILABILITY STATEMENT

The original contributions generated for the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was partially supported by NSF under LEAP HI and GOALI programs (grant no. 1830478) and EAGER program (grant no. 1842097) and by the Plant Sciences Institute at Iowa State University.

ACKNOWLEDGMENTS

The author is grateful to the editor and reviewers for their insightful feedback.

REFERENCES

Batchelor, W. D., Basso, B., and Paz, J. O. (2002). Examples of strategies to analyze spatial and temporal yield variability using crop models. *Eur. J. Agronomy* 18, 141–158. doi: 10.1016/S1161-0301(02)00101-6

Blanc, É. (2017). Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models. *Agric. Forest Meteorol.* 236, 145–161. doi: 10.1016/j.agrformet.2016.12.022

Butler, E. E., Mueller, N. D., and Huybers, P. (2018). Peculiarly pleasant weather for US maize. *Proc. Natl. Acad. Sci. U.S.A.* 115, 11935–11940. doi: 10.1073/pnas.1808035115

Chlingaryan, A., Sukkarieh, S., and Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput. Electron. Agric.* 151, 61–69. doi: 10.1016/j.compag.2018.05.012

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

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

Durand, J.-L., Delusca, K., Boote, K., Lizaso, J., Manderscheid, R., Weigel, H. J., et al. (2018). How accurately do maize crop models simulate the interactions of atmospheric CO2 concentration levels with limited water supply on water use and yield? *Eur. J. Agronomy* 100, 67–75. doi: 10.1016/j.eja.2017.01.002

Eitzinger, J., Thaler, S., Schmid, E., Strauss, F., Ferrise, R., Moriondo, M., et al. (2013). Sensitivities of crop models to extreme weather conditions during flowering period demonstrated for maize and winter wheat in austria. *J. Agric. Sci.* 151, 813. doi: 10.1017/S0021859612000779

Gurobi Optimization, LLC (2021). Gurobi optimizer reference manual. Available online at: https://www.gurobi.com/

Heslot, N., Akdemir, D., Sorrells, M., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* 127, 463–480. doi: 10.1007/s00122-013-2231-5

Hodges, T., Botner, D., Sakamoto, C., and Haug, J. H. (1987). Using the CERES-Maize model to estimate production for the US Cornbelt. *Agric. Forest Meteorol.* 40, 293–303. doi: 10.1016/0168-1923(87)90043-8

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., et al. (2014). APSIM-evolution towards a new generation of agricultural systems simulation. *Environ. Model. Softw.* 62, 327–350. doi: 10.1016/j.envsoft.2014.07.009

Hornik, K., Stinchcombe, M., and White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* 3, 551–560. doi: 10.1016/0893-6080(90)90005-6

IBM ILOG Cplex (2009). V12. 1: user's manual for CPLEX. *Int. Bus. Mach. Corporation* 46, 157. Available online at: http://citebay.com/how-to-cite/cplex/

Jagmandeep, D., Lawrence, A., Eickhoff, E., and Raun, W. (2020). Predicting in-season maize (zea mays l.) yield potential using crop sensors and climatological data. *Sci Rep.* 10.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., and Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environ. Res. Lett.* 15, 064005. doi: 10.1088/1748-9326/ab7df9

Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. *Front. Plant Sci.* 10:621. doi: 10.3389/fpls.2019.00621

Khaki, S., Wang, L., and Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. *Front. Plant Sci.* 10:1750. doi: 10.3389/fpls.2019.01750

Kiniry, J. R., Williams, J. R., Vanderlip, R. L., Atwood, J. D., Reicosky, D. C., Mulliken, J., et al. (1997). Evaluation of two maize models for nine US locations. *Agron. J.* 89, 421–426. doi: 10.2134/agronj1997.00021962008900030009x

Kucharik, C. J. (2006). A multidecadal trend of earlier corn planting in the central USA. *Agron. J.* 98, 1544–1550. doi: 10.2134/agronj2006.0156

Lacasa, J., Gaspar, A., Hinds, M., Don, S. J., Berning, D., and Ciampitti, I. A. (2020). Bayesian approach for maize yield response to plant density from both

agronomic and economic viewpoints in North America. *Sci. Rep*. 10, 1–9. doi: 10.1038/s41598-020-72693-1

Lalić, B., Eitzinger, J., Thaler, S., Vučetić, V., Nejedlik, P., Eckersten, H., et al. (2014). Can agrometeorological indices of adverse weather conditions help to improve yield prediction by crop models? *Atmosphere* 5, 1020–1041. doi: 10.3390/atmos5041020

Marko, O., Brdar, S., Panic, M., Lugonja, P., and Crnojevic, V. (2016). Soybean varieties portfolio optimisation based on yield prediction. *Comput. Electron. Agric*. 127, 467–474. doi: 10.1016/j.compag.2016.07.009

Meng, Q., Chen, X., Lobell, D. B., Cui, Z., Zhang, Y., Yang, H., et al. (2016). Growing sensitivity of maize to water scarcity under climate change. *Sci. Rep*. 6:19605. doi: 10.1038/srep19605

Messina, C., Podlich, D., Dong, Z., Samples, M., and Cooper, M. (2010). Yield-trait performance landscapes: from theory to application in breeding maize for drought tolerance. *J. Exp Bot*. 62, 855–868. doi: 10.1093/jxb/erq.329

NASS, U. (2020). *National Agricultural Statistical Service*. Available online at: https://quickstats.nass.usda.gov

National Weather Service (2020). U.S. Counties.

NCGA (2020). *National Corn Growers Association*. Available online at: https://ncga.com

Parent, B., Leclere, M., Lacube, S., Semenov, M. A., Welcker, C., Martre, P., et al. (2018). Maize yields over Europe may increase in spite of climate change, with an appropriate use of the genetic variability of flowering time. *Proc. Natl. Acad. Sci. U.S.A*. 115, 10642–10647. doi: 10.1073/pnas.1720716115

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., et al. (2017). Consistent negative response of US crops to high temperatures in observations and crop models. *Nat. Commun*. 8:13931. doi: 10.1038/ncomms13931

Shahhosseini, M., Martinez-Feria, R. A., Hu, G., and Archontoulis, S. V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. *Environ. Res. Lett*. 14, 124026. doi: 10.1088/1748-9326/ab5268

Syngenta (2020). *Syngenta Crop Challenge in Analytics*. Available online at: https://www.ideaconnection.com/syngenta-crop-challenge

Takle, E. S., and Gutowski, W. J. Jr. (2020). Iowa's agriculture is losing its goldilocks climate. *PhT* 73, 26–33. doi: 10.1063/PT.3.4407

Teal, R., Tubana, B., Girma, K., Freeman, K., Arnall, D., Walsh, O., et al. (2006). In-season prediction of corn grain yield potential using normalized difference vegetation index. *Agron. J*. 98, 1488–1494. doi: 10.2134/agronj2006.0103

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wei, Y., Devarakonda, R., Vose, R. S., et al. (2020). *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3*. Oak Ridge, TN: ORNL DAAC.

Tigchelaar, M., Battisti, D. S., Naylor, R. L., and Ray, D. K. (2018). Future warming increases probability of globally synchronized maize production shocks. *Proc. Natl. Acad. Sci. U.S.A*. 115, 6644–6649. doi: 10.1073/pnas.1718031115

Tollenaar, M., Fridgen, J., Tyagi, P., Stackhouse Jr, P. W., and Kumudini, S. (2017). The contribution of solar brightening to the US maize yield trend. *Nat. Clim. Chang* 7, 275–278. doi: 10.1038/nclimate3234

USDA (2020). *The Gridded Soil Survey Geographic*. Available online at: https://www.nrcs.usda.gov/wps/portal/nrcs/site/soils/home

van der Velde, M., Tubiello, F. N., Vrieling, A., and Bouraoui, F. (2012). Impacts of extreme weather on wheat and maize in France: evaluating regional crop simulations against observed data. *Clim. Change* 113, 751–765. doi: 10.1007/s10584-011-0368-2

van Klompenburg, T., Kassahun, A., and Catal, C. (2020). Crop yield prediction using machine learning: a systematic literature review. *Comput. Electron. Agric*. 177:105709. doi: 10.1016/j.compag.2020.105709

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., et al. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci. U.S.A*. 114, 9326–9331. doi: 10.1073/pnas.1701762114

Zhu, P., Jin, Z., Zhuang, Q., Ciais, P., Bernacchi, C., Wang, X., et al. (2018). The important but weakening maize yield benefit of grain filling prolongation in the US Midwest. *Glob. Chang Biol*. 24, 4718–4730. doi: 10.1111/gcb.14356

Author Disclaimer: Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author and do not necessarily reflect the view of the funding agencies.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.