communications
biology

ARTICLE

https://doi.org/10.1038/s42003-023-04833-y

OPEN

A data-driven crop model for maize yield prediction

Yanbin Chang[1], Jeremy Latham[1], Mark Licht[2] & Lizhi Wang[1✉]

Accurate estimation of crop yield predictions is of great importance for food security under the impact of climate change. We propose a data-driven crop model that combines the knowledge advantage of process-based modeling and the computational advantage of data-driven modeling. The proposed model tracks the daily biomass accumulation process during the maize growing season and uses daily produced biomass to estimate the final grain yield. Computational studies using crop yield, field location, genotype and corresponding environmental data were conducted in the US Corn Belt region from 1981 to 2020. The results suggest that the proposed model can achieve an accurate prediction performance with a 7.16% relative root-mean-square-error of average yield in 2020 and provide scientifically explainable results. The model also demonstrates its ability to detect and separate interactions between genotypic parameters and environmental variables. Additionally, this study demonstrates the potential value of the proposed model in helping farmers achieve higher yields by optimizing seed selection.

[1] Department of Industrial and Manufacturing Systems Engineering, Iowa State University, 2529 Union Drive, Ames 50011 IA, USA. [2] Department of Agronomy, Iowa State University, 716 Farm House Lane, Ames 50011 IA, USA. ✉email: lzwang@iastate.edu

Predicting crop yield is central to addressing emerging challenges in food security, particularly in an era of global climate change[1,2]. Accurate yield predictions help farmers make informed economic and management decisions and can support famine-prevention efforts and the global food security. Early crop model pioneers have developed research[3–5] to categorize many relevant factors that are needed by crop models, such as temperature, humidity and leaf area index (LAI). To date, underlying yield prediction is one of the greatest challenges of biology: understanding how phenotype is determined by genotype, environment, and their interactions. Specifically, the relationship between genetics, weather, soil and management variables and crop yield has been the subject of extensive studies[6–16]. Pursuing more accurate crop yield prediction techniques has and will continue to motivate innovation at the intersection of plant science and data analytics.

Majority of the literature on crop yield prediction falls into two categories: processed-based crop models and data-driven machine learning models, both of which have their salient strengths and weaknesses. Process-based crop models, such as APSIM[17–19] and DSSAT[20–22], describe the crop growth process and development as a complex function of weather, soil, and management. As such, process-based crop models reflect human knowledge of plant biology and are easily explainable in terms of physiological mechanisms. For example, Yield Prophet[23], an APSIM-based online crop simulation service, was set up to help farmers avoid over- or under-investing in their crops by forecasting potential yields with detailed inputs such as nitrogen application types and altered sowing dates. Since crop models can be experimentally validated, their results provide not only crop yield predictions but also scientific explanation of such predictions. However, these models face several serious challenges. Calibration of the numerous parameters of the a crop model typically requires time-consuming and resource intensive field experiments, yet these parameters are hardly generalizable across different varieties and environmental conditions. Oftentimes, the large variability of environmental conditions, coupled with choices of model structure and parameters, limits the predictive performance of these models beyond the spatiotemporal variability of observed yields in a large area[24,25]. Application of process-based crop models is also limited by the paucity of spatially detailed input data[24]. In the absence of spatial data on the distribution of key model inputs such as information on crop cultivars and management (e.g. irrigation, planting, fertilization, tillage, weed control), modelers often make broad assumptions across large geographic regions that may or may not reflect on-the-ground decision-making of individual producers in the context of economic opportunities and policy incentives[26].

In contrast to the process-based methodology, machine learning models take a data-driven approach to approximate the complex relationship between input (genotype and environment) and output (crop yield) without relying on human knowledge on crop science, which is incomplete and sometimes incorrect. Several machine learning models have been successfully deployed to produce remarkable prediction accuracy, including multiple linear regression[27], partial least squares regression[28], random forest regression[29,30], convolutional neural networks[31], deep neural network[10,32,33], among others. The sophisticated and powerful model structures of these data-driven models, when trained with high quality large datasets, are able to implicitly account for both additive effects and interactions among genotype, environment, and crop management practices, allowing them to outperform most crop models in terms of prediction accuracy. Some satellite-based indicators have also been utilized in the data-driven crop model to study the crop yield in a large area, such as Gross Primary Productivity (GPP)[34], Normalized Difference Vegetation Index (NDVI)[35–37], and Enhanced Vegetation Index (EVI)[38,39]. Some recent research has incorporated the remotely sensed data derived indicators into the machine learning crop model[40–45]. However, these models also inevitably suffer from the common limitations of machine learning models. They are sensitivity to data quantity and quality[46], limiting their applicability to crops with sufficient datasets. Machine learning models often include a huge number of parameters in a blackbox structure, but it is hard to discern how the parameters are used to incorporate input data into the model to predict a particular outcome such as crop yield; as such, it is difficult to extract scientific insights from the results or transfer them spatially, temporally, or genetically[47–49].

An emerging and promising research direction is to integrate process-based models and data-driven ones. Huang et al.[50] used Bayesian averaging method to construct a process-based ensemble model to provide a reliable maize yield forecast in Liaoning Province, China. Feng et al.[51] combined the APSIM and statistical regression-based model to improve the accuracy of wheat yield prediction by dynamically tracking climate and remote sensing indices during the growing season. Shahhosseini et al.[52] integrated the APSIM model and machine learning models and achieved improved yield prediction accuracy. Saha et al.[53] used regression-based machine learning models integrated with the crop growth model to improve the prediction of temporal nitrous oxide emissions from corn and soybean in the Midwest of the United States.

In this paper, we present a data-driven crop model for maize in an attempt to combine the strengths of process-based models with those of data-driven models and overcome their limitations. The proposed model attempts to provide explanatory crop yield predictions with the available historical data over both temporal and spatial dimensions without the need for experimental calibration. The proposed model uses a crop model to describe how crop yield is determined by genotype, environment, and their interactions; data-driven techniques are used to calibrate model parameters from historical data. Figure 1 illustrates how the data-driven crop model (subfigure c) conceptually differs from a process-based model (subfigure a) and a data-driven model (subfigure b). Similar to the process-based model, the data-driven crop model also describes plant phenotype as a result of genotype, environment and their interactions throughout the crop growth process, preserving the advantage of being scientifically explainable and insightful. There are three major differences between the proposed data-driven crop model and other existing crop models in the literature. First, the data-driven crop model defines the genetic properties as parameters for each crop variety. In contrast, some parameters used in conventional crop models (e.g., LAR and LAI in APSIM) are jointly determined by genotype and environment. Being independent from environmental effects, the genotypic parameters in the data-driven crop model are transferable to other environments, whereas the parameters for other crop models may need to be re-calibrated when the same varieties are grown in a different environment. Second, the data-driven crop model is designed to be a flexible framework that consists of a number of modules to reflect the crop growth process. The composition of these modules depends on the availability of data. Conventional crop models have a fixed requirement of datasets; as a result, missing or unavailable data must be imputed or assumed before the modeling can be used[54]. Third, rather than relying on large amount of field experiments for parameter calibration, the data-driven crop model employs machine learning methods to train the parameters to best fit historical data within reasonable ranges.

Result

In order to demonstrate the effectiveness of the data-driven crop approach, we applied the descriptive and predictive models to the

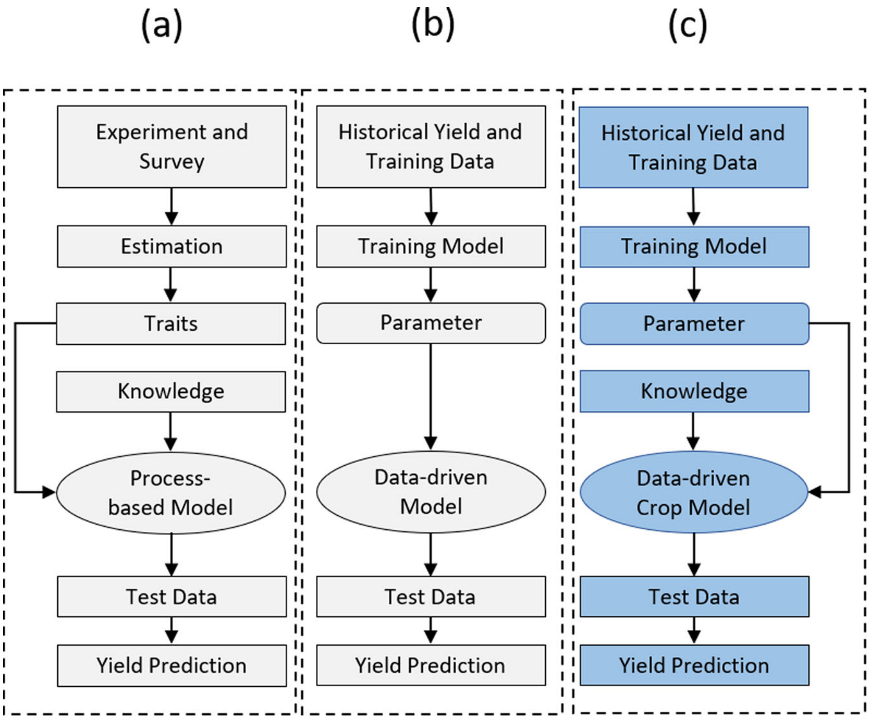

Fig. 1 Comparison of process-based, data-driven, and the proposed data-driven crop models. a Process-based models are built with human knowledge on plant physiology with explicit assumptions about genotype by environment interactions; numerous traits (modeling parameters) need to be estimated using experiments or survey of the literature. **b** Data-driven models rely on historical data to approximate the complex relationship between input and output. **c** The proposed Data-driven Crop model combines the strengths of two types of models.

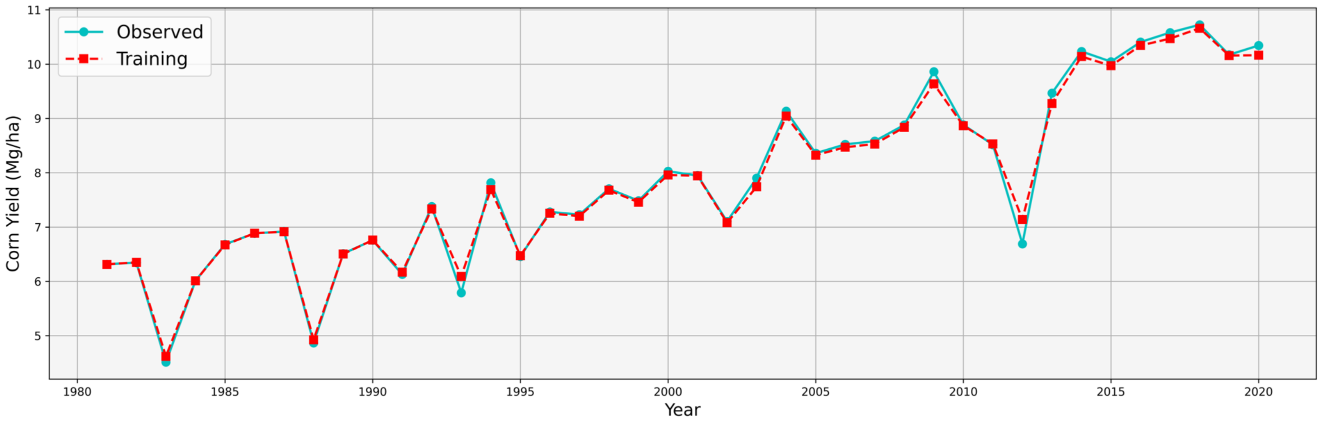

Fig. 2 Training performance of proposed model. The cyan and red curves are, respectively, observed yield and fitted yield using training data between 1981 and 2020, averaged across all counties in the Corn Belt.

dataset described in next Method section. Computational experiments were conducted using the Python on a laptop with an Intel i7-10750H processor running at 2.60 GHz with 16 GB of RAM.

Training accuracy. We were able to calibrate the genotypic parameters to achieve an RMSE of 0.74 Mg/ha for the training data; with respect to the average yield in 2020 in the Corn Belt (10.34 Mg/ha), the relative RMSE (or RRMSE) was 7.16%. Figure 2 shows the observed and fitted yields between 1981 and 2020. The overall training accuracy in the last decade was slightly higher than the first three; low accuracy years were often accompanied by extreme weather, such as the great flood in 1993 and the drought in 2012.

To benchmark the modeling performance, we found two deep learning models published in 2019[10] and 2020[11] using similar Corn Belt datasets. Their training RMSEs were 0.67 Mg/ha and 0.72 Mg/ha, respectively. In a more recent study[52], a new model was proposed that combined machine learning and APSIM models, and their training RMSE was 0.69 Mg/ha using a similar dataset. Therefore, the data-driven crop model demonstrated its capability to reach a comparable prediction accuracy with state-of-the-art models in the literature.

Spatial extrapolation. To evaluate the predictive performance of a trained data-driven crop model on an unseen location, we conducted thirteen experiments. In each experiment, we first select a county *c* in the test state, carving out all data of county *c* from the training data and using them as test data. After obtaining the predictive performance of the previously unseen county *c*, we move to the next county in the test state until the process is complete for

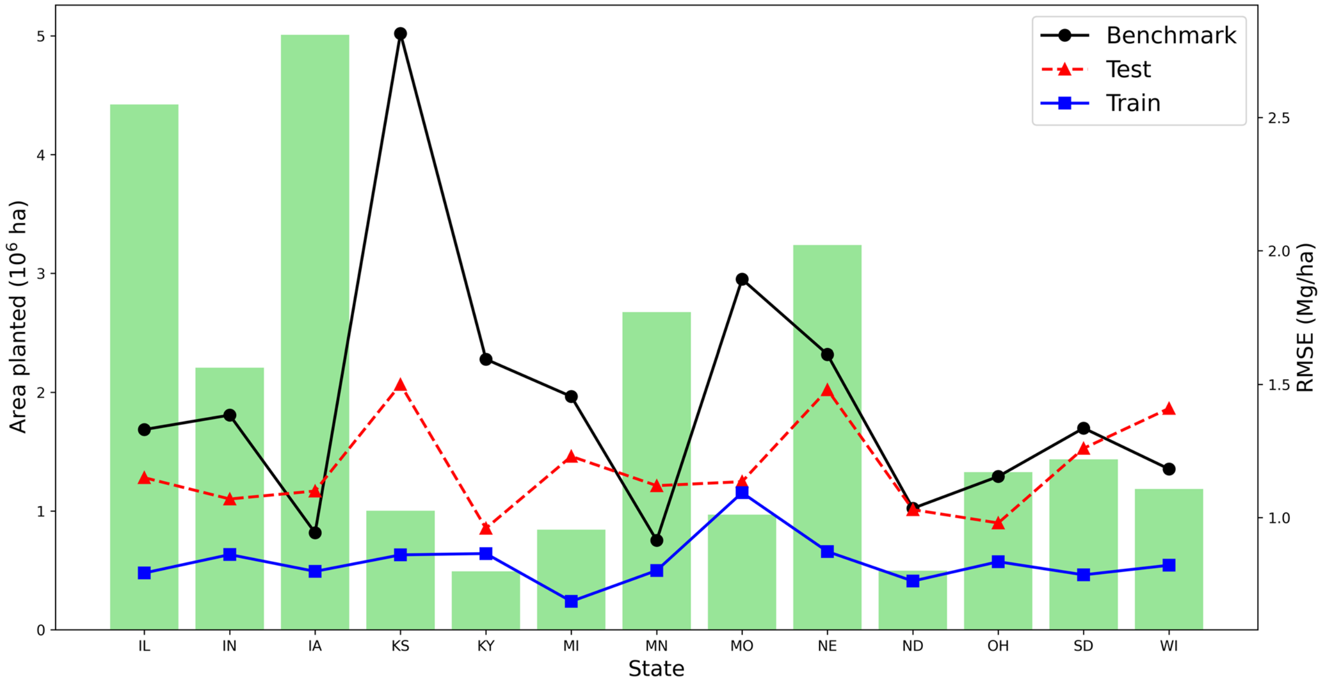

Fig. 3 Spatial extrapolation results. Green bars are average planted areas from 1981 to 2020. The three curves represent the benchmark nearest-county approach on the test data, and the data-driven crop model on test data and training data.

all counties in the test state. The nearest-county approach was used as a benchmark prediction strategy: the historical yield for the nearest county to county c in year t was used as the predicted yield for the unseen county c in year t; the planted area weighted average predicted yield for all counties from 1981 to 2020 in the test state was then used to compare with the observed average yield in the test state. Figure 3 plots the RMSEs of the benchmark approach, the data-driven crop prediction on the test data and training data, as well as the planted areas.

The average RMSE and RRMSE for spatial extrapolation were 1.17 Mg/ha and 11.32%, respectively. In contrast, the benchmark RMSE and RRMSE were 1.44 Mg/ha and 13.93%, respectively; the training RMSE and RRMSE were 0.83 Mg/ha and 8.03%, respectively. Nebraska and Kansas had the highest RMSEs, which may be partly due to a lack of irrigation data. The descriptive modeling assumes zero irrigation, given that many states in the Corn Belt are rainfed and no irrigation data are available. However, Nebraska and Kansas were among the most irrigated states in the Corn Belt, which could lead to higher prediction errors. These results suggest that the predictive performance of the model could be further improved with additional irrigation data.

Temporal extrapolation. Similar to spatial extrapolation, we also evaluated the temporal extrapolation of the data-driven crop model. We carried out forty experiments, each time carving out all data for one year between 1981 and 2020 from the training data and using them as test data. The nearest-year approach was used as a benchmark prediction strategy: the average historical yield for county c in years $t-1$ and $t+1$ was used as the predicted yield for county c in the unseen year t; the planted area weighted average predicted yield for all counties in the test year was then used to compare with the observed average yield in the test year. Figure 4 plots the RMSEs of the benchmark approach, the data-driven crop model prediction on the test data and training data, as well as the planted areas.

The average RMSE and RRMSE for temporal extrapolation were 1.15 Mg/ha and 11.12%, respectively. In contrast, the benchmark RMSE and RRMSE were 1.55 Mg/ha and 14.99%, respectively; the training RMSE and RRMSE were 0.71 Mg/ha and 6.87%, respectively. The benchmark approach struggled in drought (1983, 1988, 2012) or flood (1993) years. The data-driven crop model performance improved during the aforementioned years, although 1993 was still more challenging than other years. These results suggest the direction of improving predictive performance by refining the design of the stress module in the descriptive model.

Genotype by environment interactions. Since the genotypic parameters in the data-driven crop model were defined to be solely determined by the genotype and independent of environmental effects, the model is able to answer "what-if" questions regarding genotype by environment interactions.

In this experiment we explore the hypothetical scenarios of growing all the historical seeds under all historical weather conditions. To estimate yields in all of these scenarios, we extracted genotypic parameters for all states and all years and combined them with environmental and management data to produce the predicted yield for the desired combination. For example, the predicted yield of growing genotype from year t_1 in the environmental conditions of year t_2 in county c is calculated from function $f(W_{t_2,c}, M_{t_2,c}, S_c, g_{t_1,c}, s_c)$.

Results for this analysis were presented in Fig. 5, where the horizontal axis is the environmental conditions (weather, soil, management) from 1981 to 2020 averaged across all counties in the Corn Belt, and the vertical axis is the genotypic parameters from 1981 to 2020 averaged across all Corn Belt counties. Each colored square indicates the predicted yield of growing a given genotype under a given set of environmental conditions averaged over all counties in the Corn Belt. Diagonal thick squares represented the actual observed historical scenarios with genotype and environments belonging to the same years, while the other colored squares represented predicted yields of other hypothetical combinations. The lower triangle answers the question of "what if historically available seeds were grown in subsequent years?", which could potentially have been carried out given sufficient

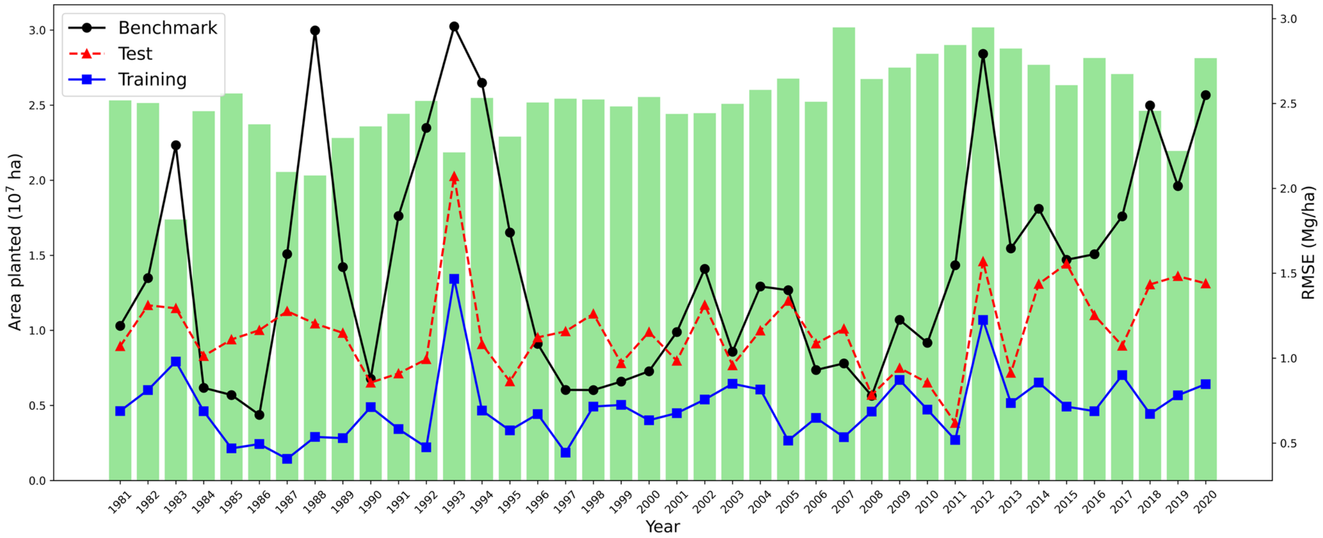

Fig. 4 Temporal extrapolation results. Green bars are average planted areas from 1981 to 2020. The three curves represent the benchmark nearest-year approach, data-driven crop model on test data and training data.

resources; whereas the upper triangle answers the question of "what if future seeds were brought back and grown in historical years?", which would not be physically possible without a time machine. The answers to both types of what-if questions provide insights into the evolution of seed genotype, environmental conditions, and their interactions over the past four decades. For instance, the 2012 drought was so devastating that no seeds in the past four decades could have produced much better; whereas the genotype of the seeds since 2009 have improved so much that they would have resulted in much higher yields if the same environmental conditions from 1981 to 2018 were to be repeated.

Yield improvement from optimal seed selection. In this experiment, we demonstrate the potential yield improvement from optimal seed selection. For the growing season in year *t* in county *c*, suppose all seeds for all counties in the Corn Belt from 1981 to *t* are available, then results from training can be used to select the optimal seed to maximize the yield. Here we consider two scenarios, one that assumes a complete knowledge of the weather in year *t* at the time of seed selection, representing a more optimistic scenario, and another that assumes zero additional knowledge of weather in year *t* beyond historical weather data, which is a more realistic scenario.

The seed selection problem for county *c* in year *t* under known weather can be formulated as the following optimization model:

$$\max_{d,r} f(W_{t,c}, M_{t,c}, S_c, g_{r,d}, s_c) \tag{1}$$

$$d \in \mathcal{C} \tag{2}$$

$$r \in \{1981, 1982, \ldots, t\}, \tag{3}$$

where $\mathcal{C}$ is the set of all counties in the Corn Belt. The objective function (1) is to maximize the predicted yield in county *c* in year *t* by selecting the optimal genotype from county *d* in a historical year *r*, which was simulated using Eq. (7) with detailed definitions in Supplementary Note 3. Constraints (2) and (3) set the limits on county *d* and historical year *r*, respectively.

The seed selection problem for county *c* in year *t* under unknown weather can be formulated as the following optimization model:

$$\max_{d,r} \frac{1}{t-1981} \sum_{\tau=1981}^{t-1} f(W_{\tau,c}, M_{\tau,c}, S_c, g_{r,d}, s_c) \tag{4}$$

$$d \in \mathcal{C} \tag{5}$$

$$r \in \{1982, 1983, \ldots, t\}, \tag{6}$$

Here, the objective function (4) is to maximize the expected predicted yield in county *c* in year *t* under all historical weather conditions by selecting the optimal genotype from county *d* in a historical year *r*, with the ranges of *d* and *r* being specific by Constraints (5) and (6), respectively.

Results are presented in Fig. 6, with the observed annual yield (averaged across all counties in the Corn Belt), improved yield with known weather using model (1) and (2), and improved yield with unknown weather using model (3) and (4) are all plotted in the same figure. The overall yield benefit trend is increasing over time due to the increased pool of historically available genotype since 1981. The average observed yield from 2011 to 2020 across all counties in the Corn Belt was 9.72 Mg/ha, whereas optimal seed selection would have been able to achieve an additional 0.38 Mg/ha, which was 3.91% of the average observed yield. With perfect meteorological insight, such yield improvement would have become 1.73 Mg/ha and 17.59%.

Results from this experiment demonstrated the potential value of the data-driven crop model for prescriptive analysis, which would not have been possible without its descriptive ability to separate the genotypic and environmental effects of crop yield and its predictive capability to answer what-if questions.

Discussion

In an attempt to combine the complementary strengths of process-based models and data-driven models and overcome their limitations, we proposed a data-driven crop model for maize yield prediction; this model has several salient features. First, its descriptive modeling framework adopts a crop model structure without the need for experimental calibration. As such, the modeling results are scientifically insightful and explainable. Second, its predictive modeling framework is able to extract knowledge from historical data without using a blackbox modeling structure. Since all modeling parameters are biologically meaningful, the training process is less sensitive to the quantity and quality of the training dataset. Third, the model is capable of providing prescriptive insight due to the clear separation of

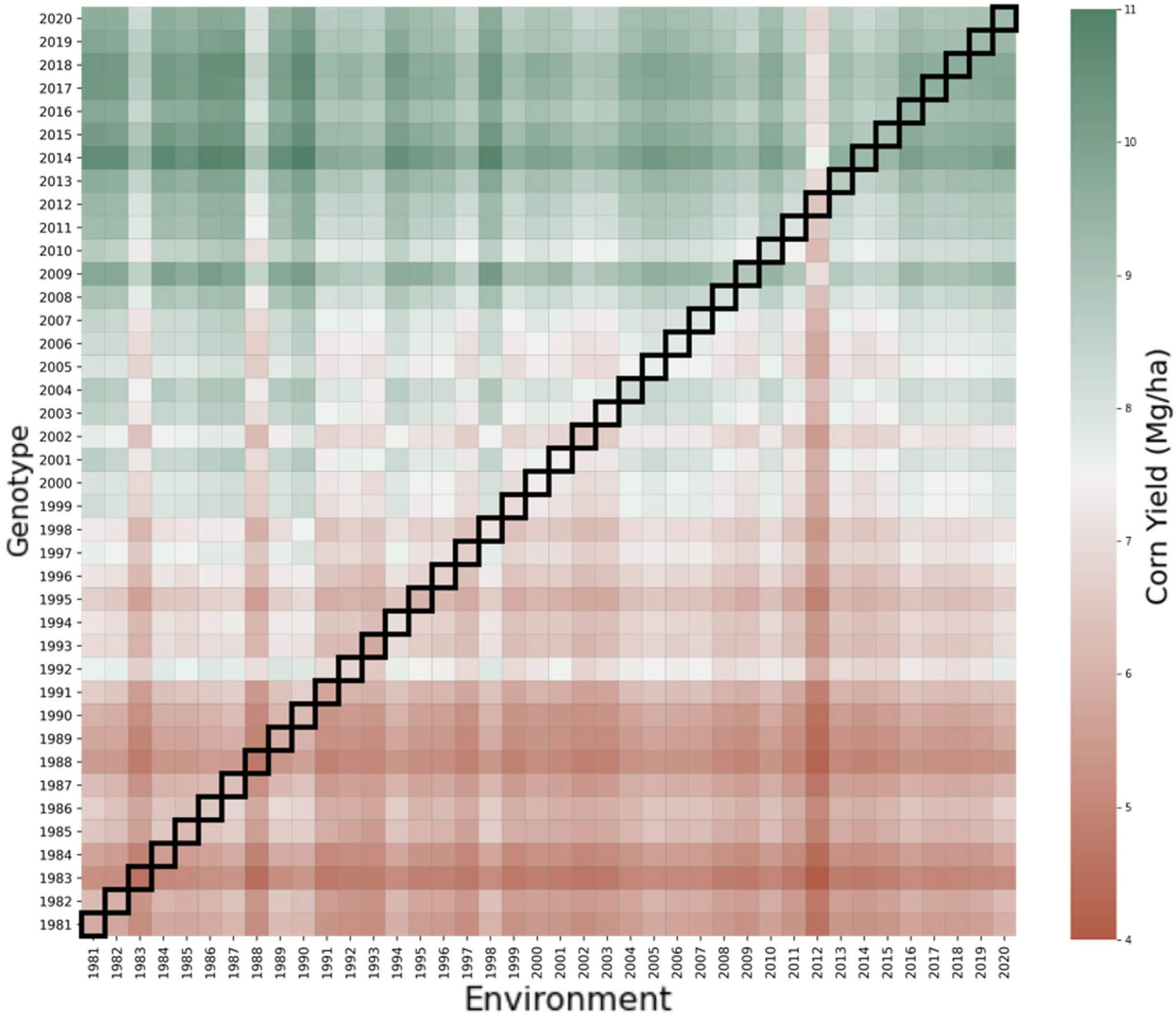

Fig. 5 Genotype by environment interactions result. Each colored square in the heat map indicates the predicted yield using the data-driven crop model when the genotype in the year from its corresponding vertical axis was grown in the environmental conditions in the year from its corresponding horizontal axis.

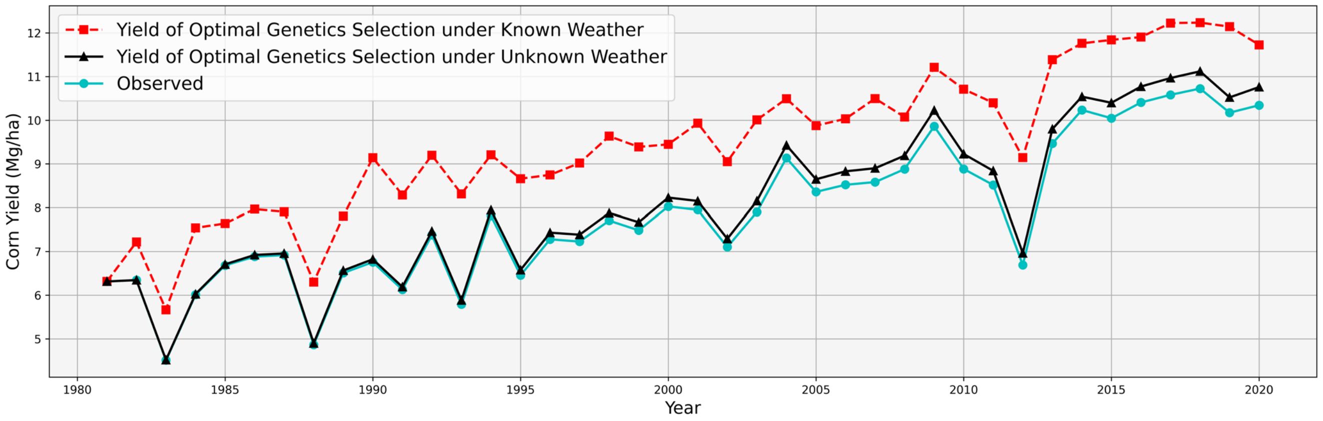

Fig. 6 Comparison of observed corn yield with improved yield from optimal seed selection under known and unknown weather scenarios. In both scenarios, optimal genotype were selected from all seeds in all counties in the Corn Belt that were historically available at the time of selection.

genotypic parameters from environmental variables and explicit descriptions of their interactions.

A comprehensive county-level dataset for the Corn Belt was used to demonstrate the performance of the data-driven crop model in our computational experiments. Many factors (such as waterlogging) are assumed to be uniform across the county. Results showed that the model was able to fit the historical data with a 7.16% RRMSE, and its spatial and temporal extrapolation RRMSEs are 11.32% and 11.12%, respectively. These predictive performances are competitive against the state-of-the-art crop yield prediction models. The data-driven crop model also predicted the yield of all combinations of historically available genotype and environmental conditions using insights from genotype by environment interactions. Additionally, the model demonstrated its prescriptive value in maximizing predicted returns through optimal seed selection. Our results indicated that optimal seed selection would have increased the average yield between 2011 and 2020 by 17.59% and 3.91%, respectively, with and without perfect weather predictions, under the optimistic assumption that all historically available seeds would be available in all counties in all subsequent years.

The proposed model is not without its limitations. For example, prediction errors were particularly large under extreme weather years such as 1983, 1988, 1993, and 2012. The transferability of a modeling structure from one crop species to another is low, since each crop has its unique physiological properties that need to be reflected by a carefully designed new modeling structure. Furthermore, the model relies on some data (such as irrigation and fertilization) that are hard to find or only available at reduced resolution (such as plant population density, planting and harvesting times).

Several future research directions are worth pursuing. First, the data-driven crop model framework needs to be developed and validated for other crop species. Second, more comprehensive case studies should be conducted using a more complete and higher resolution dataset. Third, results from optimal seed selection need to be validated experimentally. Fourth, results from the data-driven crop model, such as the genotypic parameters, may provide useful information for plant breeders.

Method

In this section, we describe the data-driven crop model for maize yield prediction. The modeling framework, however, may apply to other crop species with an appropriately selected crop model for such species and available data.

Data. We collected data for the US Corn Belt, which is an important agricultural region, accounting for approximately 87% of the total US corn production and 31% of global production in 2021[55]. Here, we briefly describe the data in different categories. More details are provided in Supplementary Note 1.

Yield and geographic data. County-level corn yield in the Corn Belt area from 1981 to 2020 were collected from USDA[56]. After excluding missing values, 47,710 county-year combinations yield data were recorded. Shape files of counties were collected from the National Weather Service[57]. This information was used to determine the membership of counties in crop reporting districts and states. Shape files were also used to locate weather stations and soil map units for calculating average weather and soil variables within each county.

Weather data. Daily surface weather data on a 1-km grid from 1981 to 2020 were collected from Daymet[58].

Management data. All management data for counties in the Corn Belt area were collected from USDA[56]. The plant and harvest dates were derived from the data from the state-level crop growth process taking into account the agricultural districts. The corn plant population density (number of plants per acre) data was also at the state level with over 50% of missing value. We utilized the mean of non-missing data (e.g., other years for the same state, if available) for data imputation.

Soil data. Soil data were collected from the latest version of the Gridded Soil Survey Geographic (gSSURGO) Database released in July 2020[59].

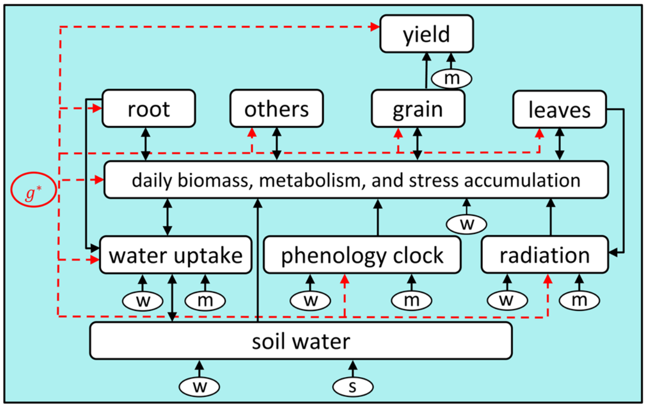

Fig. 7 Illustration of a simplified maize growth model. Here "w", "s", "m" represent weather, soil, and management variables, respectively, and "g" represents the set of parameters that are determined solely by the genotype and independent of the environment. The arrows indicate how environmental variables and genetic parameters influence different modules and eventually determine crop yield.

The descriptive modeling framework. Here we present a data-driven crop model for maize, which is tailored to the available weather, soil, and management data. Several major simplifying assumptions were necessary to account for data that were either lacking or only available at low resolution. First, due to unavailable genotype data, we assume that all seeds in each county each year were collectively represented by a unique genotype. As such, these genotypic parameters shed light on temporal and spatial trends in the average genetic performance of commercially available seeds. Second, due to lack of fertilization and irrigation data, we assume that crops were grown without irrigation but under the appropriate fertilizer availability. It is worth noting that the modeling framework does have the ability to incorporate genotype, irrigation, and fertilization data into the crop model should they become available.

Figure 7 illustrates the major modules in the corn crop model, which are briefly described as follows. More details are provided in Supplementary Note 3.

- **Soil water:** Daily soil water levels are affected by precipitation, runoff, crop water uptake, and evaporation.
- **Water uptake:** Daily amount of water uptake is proportional to root mass and atmospheric vapor pressure deficit.
- **Radiation interception:** Daily amount of solar radiation interception is proportional to LAI.
- **Phenology clock:** The growth process of maize can be separated into two growth stages: vegetative and reproductive. The transition occurs when a hybrid specific growing degree daily threshold has been reached.
- **Daily biomass and metabolism:** Daily biomass accumulation is determined by water uptake, solar radiant and leaf weight. Daily metabolism is influenced by crop weight and stress.
- **Stress:** Heat, drought, and flooding stresses are considered. Water deficits caused by heat and drought stresses reduce the amount of soil water available for plant uptake and transpiration, radiation use efficiency, and eventually growth will also be reduced.
- **Crop organs:** In the vegetative stage, certain proportions of daily biomass accumulation are allocated to leaves, roots, and other plant organs; during the reproductive stage, grains begin to fill and leaves and roots cease to grow.

The predictive modeling framework. We use the following function to represent the descriptive model:

$$\hat{y}_{t,c} = f(W_{t,c}, M_{t,c}, S_c, g_{t,c}, s_c). \tag{7}$$

Here,

- $\hat{y}_{t,c}$ is the predicted yield for county c in year t,
- $W_{t,c}$ is the weather data for county c in year t,
- S_c is the soil data for county c, which is assumed to be static over time,
- $M_{t,c}$ is the management data for county c in year t,
- $g_{t,c}$ is the genotypic parameter for county c in year t,
- s_c is the soil parameter for county c, and
- $f(\cdot)$ is the complex function defined in Supplementary Note 3 that describes the complex relationship between input (genotype, weather, soil, management) and output (corn yield), which was hypothesized based on human knowledge in plant physiology and our simplifying assumptions. Detailed variable definitions can be found in Supplementary Note 2.

A key component of the predictive modeling framework is the calibration of $g_{t,c}$. Rather than experimentally estimating such parameters as most traditional crop models do, the data-driven crop model uses historical data to identify the optimal set of genotypic parameters to produce the best fit between predicted yield and observed yield. The calibration of genotypic parameter g can be formulated as the following optimization problem. A heuristic algorithm that describes how to solve the data-driven crop model is presented at the end of Supplementary Note 3. Also, definitions of other variables are located in Supplementary Note 2.

$$\min_{g,s} \sqrt{\frac{\sum_{(t,c)} \left(M^{\text{area}}_{t,c}\right)^2 \left(y_{t,c} - \hat{y}_{t,c}\right)^2}{\sum_{(t,c)} \left(M^{\text{area}}_{t,c}\right)^2}} \tag{8}$$

$$\hat{y}_{t,c} = f(W_{t,c}, M_{t,c}, S_c, g_{t,c}, s_c) \tag{9}$$

$$g_{t_1,c} \leq 1.25 g_{t_2,c} \qquad \forall c, t_1, t_2 \tag{10}$$

$$g_{t,c_1} \leq 1.25 g_{t,c_2} \qquad \forall c_1, c_2, t. \tag{11}$$

The objective function (8) is to minimize the root-mean square error (RMSE) between predicted and observed yields weighted by planting areas. Equation (9) defines the complex function that produces the predicted yield. Constraints (10) and (11) limit, respectively, the temporal and spatial ranges of the genotypic parameters, which not only help avoid overfitting but also better reflect the fact that changes in genotype are usually gradual. The upper bound ratio of 1.25 between any two counties or years was arbitrary, yet our computational results have shown that the model is insensitive to such ratio.

The optimization model (8)–(11) serves as a data-driven training process, which not only removes the need for experimental calibration of the genotypic parameters (like typical process-based models have) but also enhances the predictive performance of the model, as will be shown in the next section.

Although the training process is similar with that of machine learning models, the data-driven crop model takes a fundamentally and philosophically different learning approach from conventional neural networks. Neural networks use a generic modeling structure with a large number of parameters and rely almost exclusively on data to learn the input-output relationship without preset underlying assumptions. This approach has the potential to capture extremely subtle and insightful knowledge beyond the comprehension of human intelligence. Along with this potential benefit come two disadvantages. The first is the risk of data deficiency, either quantitatively or qualitatively, which could mislead the model into collecting biased or false knowledge and offsetting the potential benefit. The second disadvantage is the large number of parameters, which are necessary to achieve a universal approximation capability, but they make the model not only prone to overfitting but also hard to explain.

On the other hand, the structure of the data-driven crop model is determined according to human knowledge of plant physiology, which is advanced enough to qualitatively describe the crop growth process; historical data were used only to calibrate a small number of biologically meaningful parameters. For example, the fact that radiation contributes to photosynthesis is incorporated in the structure of the model, whereas historical data were used to quantitatively determine the exact rate of radiation contribution to photosynthetic yield. These genotypic parameters are independent of environmental influences, thus can be used to identify genetic characteristics of unique genotype.

Statistics and reproducibility. The corn yield data from 1981 to 2020 in the Corn Belt area downloaded from USDA-NASS contains many missing values in different states. 47,710 county-year combinations of yield data remain after we excluded the missing values. We also utilized the means of non-missing data to impute the missing value in plant population density data.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this manuscript were openly available in public domain, and the sources of these data can be found in the Method Section. Supplementary Data 1 contains the source data behind the Figs. 2, 3, 4 and 6 in the paper.

Code availability

Code for the data-driven crop model built in this paper is available at https://doi.org/10.5281/zenodo.7792271.

Received: 8 August 2022; Accepted: 10 April 2023;
Published online: 21 April 2023

References

1. Marko, O. et al. Soybean varieties portfolio optimisation based on yield prediction. *Comput. Electron. Agric.* **127**, 467–474 (2016).
2. Messina, C., Podlich, D., Dong, Z., Samples, M. & Cooper, M. Yield–trait performance landscapes: from theory to application in breeding maize for drought tolerance. *J. Exp. Bot.* **62**, 855–868 (2010).
3. Penning de Vries, F. W., Van Laar, H. & Kropff, M. *Simulation and Systems Analysis for Rice Production (SARP)* (PUDOC, Wageningen, The Netherlands, 1991).
4. Bouman, B., Van Keulen, H., Van Laar, H. & Rabbinge, R. The 'school of de wit'crop growth simulation models: a pedigree and historical overview. *Agric. Syst.* **52**, 171–198 (1996).
5. van Ittersum, M. K. et al. On approaches and applications of the wageningen crop models. *Eur. J. Agron.* **18**, 201–234 (2003).
6. Verón, S. R., De Abelleyra, D. & Lobell, D. B. Impacts of precipitation and temperature on crop yields in the pampas. *Clim. Change* **130**, 235–245 (2015).
7. Hatfield, J. L. & Walthall, C. L. Meeting global food needs: realizing the potential via genetics × environment × management interactions. *Agron. J.* **107**, 1215–1226 (2015).
8. Battisti, R. et al. Assessment of soybean yield with altered water-related genetic improvement traits under climate change in southern brazil. *Eur. J. Agron.* **83**, 1–14 (2017).
9. MacCarthy, D. S., Adiku, S. G., Freduah, B. S., Gbefo, F. & Kamara, A. Y. Using ceres-maize and enso as decision support tools to evaluate climate-sensitive farm management practices for maize production in the northern regions of ghana. *Front. Plant Sci.* **8**, 31 (2017).
10. Khaki, S. & Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **10**, 621 (2019).
11. Khaki, S., Wang, L. & Archontoulis, S. V. A CNN-RNN framework for crop yield prediction. *Front. Plant Sci.* **10**, 1750 (2020).
12. Agnolucci, P. et al. Impacts of rising temperatures and farm management practices on global yields of 18 crops. *Nat. Food* **1**, 562–571 (2020).
13. Gul, F. et al. Use of crop growth model to simulate the impact of climate change on yield of various wheat cultivars under different agro-environmental conditions in khyber pakhtunkhwa, pakistan. *Arab. J. Geosci.* **13**, 1–14 (2020).
14. Cooper, M. et al. Integrating genetic gain and gap analysis to predict improvements in crop productivity. *Crop Sci.* **60**, 582–604 (2020).
15. Lesk, C. et al. Stronger temperature–moisture couplings exacerbate the impact of climate warming on global crop yields. *Nat. Food* **2**, 683–691 (2021).
16. Elahi, E., Khalid, Z., Tauni, M. Z., Zhang, H. & Lirong, X. Extreme weather events risk to crop-production and the adaptation of innovative management strategies to mitigate the risk: a retrospective survey of rural punjab, pakistan. *Technovation* **117**, 102255 (2021).
17. Keating, B. A. et al. An overview of apsim, a model designed for farming systems simulation. *Eur. J. Agron.* **18**, 267–288 (2003).
18. Malone, R. W. et al. Evaluating and predicting agricultural management effects under tile drainage using modified apsim. *Geoderma* **140**, 310–322 (2007).
19. Balboa, G. R. et al. A systems-level yield gap assessment of maize-soybean rotation under high-and low-management inputs in the western us corn belt using apsim. *Agric. Syst.* **174**, 145–154 (2019).
20. Jones, J. W. et al. The dssat cropping system model. *Eur. J. Agron.* **18**, 235–265 (2003).
21. Jones, J. W. et al. Estimatingdssat cropping system cultivar-specific parameters using Bayesian techniques. In *Methods of Introducing System Models Into Agricultural Research,* vol **2**, 365–393 (Wiley, 2011).
22. Corbeels, M., Chirat, G., Messad, S. & Thierfelder, C. Performance and sensitivity of the dssat crop growth model in simulating maize yield under conservation agriculture. *Eur. J. Agron.* **76**, 41–53 (2016).
23. Hunt, J. et al. Yield prophet®: an online crop simulation service. In *Proc 13th Australian Agronomy Conference*, 10–14 (The Australian Society of Agronomy, 2006).
24. Folberth, C. et al. Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. *Nat. Commun.* **7**, 1–13 (2016).
25. Ramirez-Villegas, J., Koehler, A.-K. & Challinor, A. J. Assessing uncertainty and complexity in regional-scale crop model simulations. *Eur. J. Agron.* **88**, 84–95 (2017).
26. Folberth, C. et al. Parameterization-induced uncertainties and impacts of crop management harmonization in a global gridded crop model ensemble. *PLoS ONE* **14**, e0221862 (2019).
27. Ramesh, D. & Vardhan, B. V. Analysis of crop yield prediction using data mining techniques. *Int. J. Res. Eng. Technol.* **4**, 47–473 (2015).
28. Foster, A., Kakani, V. & Mosali, J. Estimation of bioenergy crop yield and n status by hyperspectral canopy reflectance and partial least square regression. *Precis. Agric.* **18**, 192–209 (2017).
29. Jeong, J. H. et al. Random forests for global and regional crop yield predictions. *PLoS ONE* **11**, e0156571 (2016).

30. Sakamoto, T. Incorporating environmental variables into a modis-based crop yield estimation method for united states corn and soybeans through the use of a random forest regression algorithm. *ISPRS J. Photogramm. Remote Sens.* **160**, 208–228 (2020).
31. Sun, J., Di, L., Sun, Z., Shen, Y. & Lai, Z. County-level soybean yield prediction using deep cnn-lstm model. *Sensors* **19**, 4363 (2019).
32. Bhojani, S. H. & Bhatt, N. Wheat crop yield prediction using new activation functions in neural network. *Neural Comput. Appl.* **32**, 13941–13951 (2020).
33. Wang, X., Huang, J., Feng, Q. & Yin, D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. *Remote Sens.* **12**, 1744 (2020).
34. Reeves, M. C., Zhao, M. & Running, S. W. Usefulness and limits of modis gpp for estimating wheat yield. *Int. J. Remote Sens.* **26**, 1403–1421 (2005).
35. Kogan, F., Gitelson, A. A., Zakarin, E., Spivak, L. & Lebed, L. Avhrr-based spectral vegetation index for quantitative assessment of vegetation state and productivity: calibration and validation. *Photogrammetric Engineering and Remote Sensing* **69**, 899–906 (2003).
36. Becker-Reshef, I., Vermote, E., Lindeman, M. & Justice, C. A generalized regression-based model for forecasting winter wheat yields in kansas and ukraine using modis data. *Remote Sens. Environ.* **114**, 1312–1323 (2010).
37. Esquerdo, J., Zullo Júnior, J. & Antunes, J. Use of ndvi/avhrr time-series profiles for soybean crop monitoring in brazil. *Int. J. Remote Sens.* **32**, 3711–3727 (2011).
38. Gusso, A., Ducati, J. R., Veronez, M. R., Arvor, D. & Silveira Junior, L. G. d. Spectral model for soybean yield estimate using modis/evi data. *Int. J. Geosci.* **4**, 1233–1241 (2013).
39. Kouadio, L., Newlands, N. K., Davidson, A., Zhang, Y. & Chipanshi, A. Assessing the performance of modis ndvi and evi for seasonal crop yield forecasting at the ecodistrict scale. *Remote Sens.* **6**, 10193–10214 (2014).
40. Kuwata, K. & Shibasaki, R. Estimating crop yields with deep learning and remotely sensed data. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 858–861 (IEEE, 2015).
41. Fernandes, J. L., Ebecken, N. F. F. & Esquerdo, J. C. D. M. Sugarcane yield prediction in brazil using ndvi time series and neural networks ensemble. *Int. J. Remote Sens.* **38**, 4631–4644 (2017).
42. You, J., Li, X., Low, M., Lobell, D. & Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In *Proc AAAI conference on Artificial Intelligence*, vol. 31 (KP Publishing Services Network, 2017).
43. Haghverdi, A., Washington-Allen, R. A. & Leib, B. G. Prediction of cotton lint yield from phenology of crop indices using artificial neural networks. *Comput. Electron. Agric.* **152**, 186–197 (2018).
44. Wang, X., Huang, J., Feng, Q. & Yin, D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. *Remote Sens.* **12**, 1744 (2020).
45. Khaki, S., Pham, H. & Wang, L. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Sci. Rep.* **11**, 11132 (2021).
46. Dang, C., Liu, Y., Yue, H., Qian, J. & Zhu, R. Autumn crop yield prediction using data-driven approaches:-support vector machines, random forest, and deep neural network methods. *Can. J. Remote Sens.* **47**, 162–181 (2021).
47. Ansarifar, J., Wang, L. & Archontoulis, S. V. An interaction regression model for crop yield prediction. *Sci. Rep.* **11**, 1–14 (2021).
48. Martinez-Feria, R. A., Licht, M. A., Antonio-Ordoñez, R. A., Hatfield, J. L. & Archontoulis, S. V. An improved algorithm to predict in-field dry-down of maize and soybean grains and genotype-by-environment analysis. In *ASA, CSSA, and CSA International Annual Meeting (2018)* (ASA-CSSA-SSSA, 2018).
49. Mourtzinis, S. et al. Sifting and winnowing: analysis of farmer field data for soybean in the us north-central region. *Field Crops Res.* **221**, 130–141 (2018).
50. Huang, X., Huang, G., Yu, C., Ni, S. & Yu, L. A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. *Field Crops Res.* **211**, 114–124 (2017).
51. Feng, P. et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agric. For. Meteorol.* **285**, 107922 (2020).
52. Shahhosseini, M., Hu, G., Huber, I. & Archontoulis, S. V. Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt. *Sci. Rep.* **11**, 1–15 (2021).
53. Saha, D., Basso, B. & Robertson, G. P. Machine learning improves predictions of agricultural nitrous oxide (n2o) emissions from intensively managed cropping systems. *Environ. Res. Lett.* **16**, 024004 (2021).
54. Peng, B. et al. Towards a multiscale crop modelling framework for climate change adaptation assessment. *Nat. Plants* **6**, 338–348 (2020).
55. USDA-NASS. *Crop Production 2021 Summary (February 2022)* (USDA-NASS, Washington, DC, 2022).
56. USDA-NASS. United states department of agriculture national agricultural statistics service. https://www.nass.usda.gov/Quick_Stats/ (2022).
57. National-Weather-Service. U.S. counties. https://www.weather.gov/gis/Counties (2020).
58. Thornton, P. et al. Daymet: Daily surface weather data on a 1-km grid for North America, version 3. https://doi.org/10.3334/ORNLDAAC/1328 (2020).
59. USDA. The gridded soil survey geographic. https://www.nrcs.usda.gov/wps/portal/nrcs/site/soils/home (2020).

Acknowledgements

This work was partially supported by NSF and USDA (#1830478, #1842097, #2021-67021-35329) and by the Plant Sciences Institute at Iowa State University. The authors are grateful to the Editors and Reviewers for their insightful and constructive feedback, which greatly improved the quality of this manuscript. We also thank Dr. Silvia Cianzio and Dr. Maria Salas Fernandez for helpful discussions about plant physiology.

Author contributions

Y.C., J.L., M.L., and L.W. were involved in brainstorming the idea through numerous discussions. Y.C. and L.W. designed the model and collected the data; Y.C. conducted coding and computational experiments; Y.C. and L.W. wrote the manuscript. Y.C., J.L., M.L., and L.W. revised, proofread and approved the final version.

Competing interests

L.W. is a co-founder of Crop Convergence LLC. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-023-04833-y.

Correspondence and requests for materials should be addressed to Lizhi Wang.

Peer review information *Communications Biology* thanks Roger Lawes and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary handling editors: Jonathan Touboul and Gene Chong.

Reprints and permission information is available at http://www.nature.com/reprints

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