frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 November 2021
doi: 10.3389/fpls.2021.699589

Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks

Cathy C. Westhues[1,2]*, *Gregory S. Mahone*[3], *Sofia da Silva*[3], *Patrick Thorwarth*[3], *Malthe Schmidt*[3], *Jan-Christoph Richter*[3], *Henner Simianer*[2,4] and *Timothy M. Beissinger*[1,2]

[1] *Division of Plant Breeding Methodology, Department of Crop Sciences, University of Goettingen, Goettingen, Germany,* [2] *Center for Integrated Breeding Research, University of Goettingen, Goettingen, Germany,* [3] *Kleinwanzlebener Saatzucht (KWS) SAAT SE, Einbeck, Germany,* [4] *Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Goettingen, Germany*

The development of crop varieties with stable performance in future environmental conditions represents a critical challenge in the context of climate change. Environmental data collected at the field level, such as soil and climatic information, can be relevant to improve predictive ability in genomic prediction models by describing more precisely genotype-by-environment interactions, which represent a key component of the phenotypic response for complex crop agronomic traits. Modern predictive modeling approaches can efficiently handle various data types and are able to capture complex nonlinear relationships in large datasets. In particular, machine learning techniques have gained substantial interest in recent years. Here we examined the predictive ability of machine learning-based models for two phenotypic traits in maize using data collected by the Maize Genomes to Fields (G2F) Initiative. The data we analyzed consisted of multi-environment trials (METs) dispersed across the United States and Canada from 2014 to 2017. An assortment of soil- and weather-related variables was derived and used in prediction models alongside genotypic data. Linear random effects models were compared to a linear regularized regression method (*elastic net*) and to two nonlinear gradient boosting methods based on decision tree algorithms (*XGBoost*, *LightGBM*). These models were evaluated under four prediction problems: (1) tested and new genotypes in a new year; (2) only unobserved genotypes in a new year; (3) tested and new genotypes in a new site; (4) only unobserved genotypes in a new site. Accuracy in forecasting grain yield performance of new genotypes in a new year was improved by up to 20% over the baseline model by including environmental predictors with gradient boosting methods. For plant height, an enhancement of predictive ability could neither be observed by using machine learning-based methods nor by using detailed environmental information. An investigation of key environmental factors using gradient boosting frameworks also revealed that temperature at flowering stage, frequency and amount of water received during the vegetative and grain filling stage, and soil organic matter content appeared as important predictors for grain yield in our panel of environments.

Keywords: machine learning, genotype-by-environment interactions, gradient boosting, maize, yield, genomic prediction, plant breeding

Edited by:
Fernando H. Toledo,
International Maize and Wheat Improvement Center, Mexico

Reviewed by:
Diego Jarquin,
University of Nebraska-Lincoln, United States
Germano Costa-Neto,
Cornell University, United States
Sergio Perez-Elizalde,
Colegio de Postgraduados (COLPOS), Mexico

***Correspondence:**
Cathy C. Westhues
cathy.jubin@uni-goettingen.de

Specialty section:
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

Received: 23 April 2021
Accepted: 15 October 2021
Published: 11 November 2021

Citation:
Westhues CC, Mahone GS, da Silva S, Thorwarth P, Schmidt M, Richter J-C, Simianer H and Beissinger TM (2021) Prediction of Maize Phenotypic Traits With Genomic and Environmental Predictors Using Gradient Boosting Frameworks. Front. Plant Sci. 12:699589. doi: 10.3389/fpls.2021.699589

1. INTRODUCTION

The development of environmental sensing technologies, including local weather stations, soil and crop sensors has progressively enabled field-level climate data to be incorporated into the analysis of plant breeding experiments (Tardieu et al., 2017; Ersoz et al., 2020; Crossa et al., 2021). When used to enhance genomic prediction, climate data can be useful to estimate the differential response of genotypes to new environmental conditions, i.e., genotype-by-environment interactions (GxE), almost omnipresent in multi-environment trial (MET) experiments (Cooper and DeLacy, 1994; Chenu, 2015). In plant breeding, an environment generally refers to the set of growing conditions associated with a given location in a given year. Various statistical models, such as factorial regression methods, have been developed to model genotype sensitivity to continuous environmental covariates (ECs) (van Eeuwijk et al., 1996; Malosetti et al., 2004) or even to simple geographic coordinates (Costa-Neto et al., 2020b) capturing primarily genotype-by-location interaction effects explained by crop management or soil characteristics.

Before the emergence of environmental data in breeding, large whole-genome marker datasets, generated by high-throughput genotyping platforms, have progressively enabled the routine implementation of genomic prediction (GP) methods (Haley and Visscher, 1998; Meuwissen et al., 2001). GP allows to predict performance of untested genotypes based on their genetic similarity, estimated with marker data, with other phenotyped genotypes. GP has since been expanded to achieve predictions in a multi-environment context, for instance by implementing a multivariate GBLUP approach (Burgueño et al., 2012) to use genetic correlations among environments. Despite the overall success of genomic prediction, a lingering challenge has regularly been to incorporate interactions between high-dimensional genomic data and high-dimensional environmental data. A solution proposed by Jarquín et al. (2014) is to use reaction norm models, where markers and environmental effects are modeled using covariance structures. Interactions between markers and environmental covariates are computed with the Hadamard product which avoids the need to fit all first-order interaction terms. This extension of the GBLUP GxE mixed effects models has been applied on a large number of datasets in different species (Pérez-Rodríguez et al., 2015; Pérez-Rodríguez et al., 2017; Jarquín et al., 2017; Sukumaran et al., 2017, 2018; Monteverde et al., 2019; Rincent et al., 2019; De Los Campos et al., 2020). Several studies have also focused on the integration of crop growth models in genomic prediction to better model the differential impact of abiotic stress depending on the crop developmental stage (Heslot et al., 2014a; Rincent et al., 2017, 2019). Rincent et al. (2019) proposed a method to select the optimal subset of ECs from the output of a crop growth model on the basis of the correlation between the environmental covariance matrix, which is based on ECs, and the covariance matrix between GxE interactivity of environments obtained by AMMI decomposition. Overall, many studies have found that using quantitative environmental information in genomic prediction models in the form of additional covariates can result in an enhancement of prediction accuracies (Heslot et al., 2014b; Jarquín et al., 2014; Malosetti et al., 2016; Millet et al., 2019; Monteverde et al., 2019; Costa-Neto et al., 2020a) and a better characterization of the genotype-by-environment interaction effects (Rogers et al., 2021).

However, modeling interaction effects with nonlinear techniques is a crucial topic that has not been conclusively explored for genomic prediction in MET. In particular, machine learning techniques have gained attention over the last two decades due to their ability to handle nonlinear effects (Hastie et al., 2009) and to uncover higher-order interactions between predictor variables (Lampa et al., 2014; Behravan et al., 2018). With machine learning algorithms, the mapping function linking input variables to the outcome—i.e., a phenotypic trait—is learned from training data and no strong assumptions about its form need to be explicitly formulated beforehand. Hence, these methods represent relatively flexible frameworks for data-driven integration of different data types. Among these new techniques, ensembles of trees, such as methods based on bagging (e.g., random forests), or on boosting (e.g., gradient boosted trees) have become increasingly popular. Ensemble methods designate predictive modeling techniques which aggregate the predictions of a group of base learners, and thereby generally allow better predictions than by using only the single best learner (Friedman, 2001; Hastie et al., 2009; Géron, 2019). Broad applications of these approaches include human disease prediction (Fukuda et al., 2013; Romagnoni et al., 2019; Yu et al., 2019; Kopitar et al., 2020), bioinformatics (Yu et al., 2019), ecology (Moisen et al., 2006; Elith et al., 2008) and agricultural forecasting (Fukuda et al., 2013; Delerce et al., 2016; Jeong et al., 2016; Crane-Droesch, 2018; Shahhosseini et al., 2020). In the field of genomic prediction, ensemble methods have progressively been used, as they appear especially interesting for capturing non-additive effects such as epistasis or dominance effects, which can be important for predicting complex phenotypic traits (Ogutu et al., 2011; González-Recio et al., 2013; Azodi et al., 2019; Abdollahi-Arpanahi et al., 2020). Abdollahi-Arpanahi et al. (2020) concluded from results obtained on both a real animal and simulated datasets that gradient boosting was the best predictive modeling approach when the genetic architecture included non-additive effects. While these new predictive modeling approaches can also potentially enable superior prediction results, special attention must be paid to an appropriate optimization of hyperparameters during the training phase in order to prevent overfitting on new test data (Friedman, 2001; Hastie et al., 2009; Géron, 2019).

In addition, these new predictive modeling frameworks, coupled with large volumes of environmental data, can provide powerful data mining opportunities to identify critical environmental factors affecting economically important phenotypic traits in the field. Much research has already been done to examine the expected impact of climate change on the vulnerability of major staple food crops. Extreme weather events are expected to happen at a higher frequency in the future, characterized for instance by heat waves or prolonged drought periods according to various climate scenarios (Rahmstorf et al., 2012; Trnka et al., 2014). When occurring at crucial crop developmental stages, risks for important yield losses are augmented. Different studies on maize have for instance reported a physiological sensitivity to higher temperatures, heightened during the reproductive phase, which often results in grain yield reduction when a certain threshold is exceeded (Cicchino et al., 2010; Butler and Huybers, 2015; Lizaso et al., 2018). In addition, nonlinear effects of environmental covariates, especially of temperature and precipitation on maize plants, have also been regularly described in the literature (Schlenker and Roberts, 2009; Mushore et al., 2017). Therefore, machine learning techniques break new ground to get an extended comprehension of the effect—both in direction and magnitude—of environmental conditions in the context of breeding for abiotic stress resilience.

Motivated by previous studies emphasizing the benefit of nonlinear methods, we tested two machine learning ensemble methods, based on gradient boosted trees, which, to our knowledge, have never been examined for data-driven predictions and interpretation using MET experimental datasets from the Maize Genomes to Fields initiative. The Maize Genomes to Fields (G2F) initiative (www.genomes2fields.org) includes yearly evaluations of inbred and hybrid maize across a large range of climatically-distinct regions in North America. The project makes publicly available phenotypic and genotypic (genotyping-by-sequencing datasets relating to the inbred lines) information, as well as weather (field weather stations), agronomic practices and soil data (Falcon et al., 2020; McFarland et al., 2020). The large number of phenotypic observations, and the assortment of various data types makes the application of machine learning models here particularly relevant to evaluate their performance, as well as their usefulness to disentangle hidden relationships. Our objectives in this study were (1) to evaluate recent gradient boosting methods for prediction of two phenotypic traits (plant height and grain yield) across four different cross-validations, and compare them to traditional prediction models classically used for multi-environment trials; (2) to examine if the use of environmental information, in addition to genomic predictor variables, could lead to a gain of predictive ability of genotype performance based on these various prediction models; and (3) to better understand the influence of some environmental factors on maize grain yield using tools derived from the machine learning framework.

2. MATERIALS AND METHODS

2.1. Phenotypic Data Cleaning and Analysis

Phenotypic datasets (years 2014–2017) were downloaded from the official website of the Genomes to Fields project. The full dataset represents a large collection of trials located on the North-American continent run by different principal investigators and institutions, but the experimental design used for most of the hybrid trials was a randomized complete block design with two replications per environment. A total number of 71 trial experiments remained for further analysis (**Supplementary Figure 1**; **Supplementary Table 1**) after having eliminated environments with critical missing information, such as flowering time (**Supplementary Table 2**). Plots with low phenotypic quality, as interpreted by the researcher groups who collected field data, were removed before within-experiment analysis. Replicates within a same ID experiment but planted seven or more days apart were considered as different environments and treated as unreplicated environments, due to the difference in the weather conditions they experienced at their respective phenological stages.

Each environment (Year-Site combination) was independently analyzed to obtain best linear unbiased estimates (BLUEs) for each hybrid in each environment for grain yield, plant height and silking date. We performed this analysis with the *lme4* package (Bates et al., 2015) in R version 3.6.0 (R Core Team, 2019) based on the following model:

$$Y_{ij} = \mu + G_i + R_j + \varepsilon_{ij},$$

where Y_{ij} is the observed phenotypic response variable of the i-th hybrid genotype (G) in the j-th replicate (R), μ is the general mean, G_i is the effect of the i-th hybrid genotype, R_j is the effect of the j-th replicate and ε_{ij} is the error associated with the observation Y_{ij}. We treated genotype as a fixed effect and replicate as a random effect.

Phenotypic observations with absolute studentized conditional residuals greater than three were identified as potential outliers and removed from the dataset. The plant material and phenotypic datasets are described in more details in previous publications (AlKhalifah et al., 2018; McFarland et al., 2020) and on the project website (https://www.genomes2fields.org/home/). Ultimately, 18,325 and 16,951 phenotypic observations for grain yield and plant height, respectively, with available silking date, genotypic and environmental data, were used as target response variable in the prediction models.

2.2. Genotypic Data

Genotype-by-sequencing (GBS) data of inbred lines used in Genomes to Fields hybrid experiments were downloaded on CyVerse. SNPs with more than two observed alleles were removed before analysis. Taxa with less than 70% site coverage and more than 8% heterozygosity were discarded. Monomorphic markers were removed, as were those missing or heterozygous in more than 5% of the parental lines. These filtering analyses were performed with TASSEL 5 (Bradbury et al., 2007). After filtering, 246,818 SNPs remained for analysis. These were imputed using the software LinkImpute (Money et al., 2015). The genotype matrix was coded as the number of minor alleles at each locus (0, 1, or 2). Markers with minor allele frequency less than 2% and in high linkage Disequilibrium (LD) were further removed using the pruning function of Plink (Purcell et al., 2007) with a window of size 100 markers, a step of 5, and a LD threshold of 0.99. *In silico* genotypes of maize hybrids, for which phenotypic data had been analyzed, were constructed based on the processed genotypes of parental lines, and a final minor allele frequency filtering of 2% was applied. The final hybrid genotype dataset contained 107,399 SNPs characterizing 2,033 hybrids. Additional details regarding the genotype-by-sequencing procedure implemented by the Genomes to Fields project has been previously published (Gage et al., 2017).

2.3. Weather Data

All field experiment locations in the Genomes to Fields project had a WatchdogTM Model 2700 weather station (Spectrum Technologies Inc., East-Plainfield, Illinois, 60585, USA) on-site. Weather records were recorded every 30 min during the growing season. Measurements for air temperature (°C), relative humidity (%), rainfall (mm), solar radiation (W/m2) and wind speed (m/s) were specifically analyzed. In-field weather station measurements provide climatic information of a very localized scale in comparison to weather service stations. Therefore, we prioritized the use of weather-station data whenever data quality criteria were fulfilled and the proportion of missing data was reasonable. When quality criteria were not met, weather data was acquired from nearby weather service stations.

In the first step, we summarized the hourly or semi-hourly records for each climatic variable on a daily basis using various quality control criteria (consistent number of weather records per day; threshold tests; persistence tests, i.e., flagging observations with null variability during the day; internal consistency tests, i.e., verification of the relation between measured variables). These criteria were applied based on the recommendations from the official published guidelines on quality control procedures for data acquired from weather stations (Zahumenský, 2004; Estévez et al., 2011) and are detailed in **Supplementary Table 3**. Data from the field weather station were compared against weather data obtained from public climate summaries to check for possible large data divergences and to fill out missing values. Data from the Global Historical Climatology Network (GHCN) and from the Global Surface Summary of the Day (GSOD) were retrieved from the National Oceanic and Atmospheric Administration (NOAA) website to investigate American locations, while data for Canadian locations were downloaded from the Environment and Climate Change Canada (ECCC) website, based each time on a 70-kilometer radius from the geographic coordinates for each field experiment. In case data from the field weather station data were missing or assigned as erroneous, data from the closest publicly accessible weather station were used, if it was located less than 2 km from the field. If the distance to the nearest station was large, interpolation by spatio-temporal kriging or inverse distance weighting was performed using the R package *gstat* to impute the missing data (Pebesma, 2004; Gräler et al., 2016). For wind data, we only used results obtained from inverse distance weighting because of the consistency regarding the standard height measurement obtained from GSOD data. Similarly, in-field weather stations solar radiation data were characterized by a high percentage of missing values and inconsistencies; we used instead the R package nasapower (Sparks, 2018), which enables an easy access to NASA POWER surface solar radiation energy data. Some environments were irrigated: for those of which the precise amount was tracked during the growing season, these data were added to the final daily precipitation data.

Hence, the daily weather data consisted of the daily maximum, minimum and mean temperature (average of minimum and maximum daily temperatures), average wind speed, precipitation, humidity, incoming solar radiation. Based on these processed weather data, we were then able to calculate the daily growing degrees (Baskerville and Emin, 1969), the photothermal time (product between GDs and day length in hours, for each day, also referred as an environmental index; Li et al., 2018), the mean vapor pressure deficit, the reference evapotranspiration (ET_0) using FAO-56 Penman-Monteith method (Allen et al., 1998). These latter variables were computed because they incorporate crop physiological parameters which make them sometimes more relevant than the initial weather data.

2.4. Derivation of Environmental Variables per Hybrid Growth Stage

The next step was to obtain pertinent environmental predictors from daily weather summaries for the predictive modeling framework. The objective was to relate each hybrid phenotypic performance (e.g., yield) in a particular environment, individually characterized by its specific flowering dates, to the corresponding weather series during the growing season. To develop a unified framework across the different growing season lengths, which varied throughout locations and years, we used three critical maize growth stages, as was performed in previous similar work for other crops (Heslot et al., 2014b; Delerce et al., 2016; Gillberg et al., 2019; Monteverde et al., 2019). This approach was needed to account for the differential impact of weather-based variables according to the crop developmental stage. Each intermediate plant developmental stage could not be precisely determined since visual scoring for all stages is in practice highly time-consuming and expensive. However, the sowing date and the flowering date, i.e., when 50% of plants in a plot have visible silk, were recorded for each hybrid kept after phenotypic data analysis. Based on these known dates, three hybrid maize growth periods could be estimated: vegetative (from the planting date to 1 week before the 50% silking date); flowering (from 1 week before 50% silking date to 2 weeks after that date, which corresponds approximately to the end of the pollination period); and the grain filling stage (from the end of the flowering period to 65 days after, after which maturity should be reached). By definition, these three periods do not overlap. The typical duration of the grain filling stage varies according to the hybrid and the environment; nonetheless, based on literature and agronomic knowledge, the corn plant is normally at physiological maturity (R6) about 55–65 days after silking (Ritchie et al., 1993).

Based on these dates, 13 weather-based environmental predictor variables were computed for each phenological stage and therefore were both environment- and hybrid-specific (**Table 1**). We included stress covariates related to heat, as it is expected that an excess of heat can be detrimental, especially during the flowering stage, and results in a lower yield. To examine the presence of clusters of environments based on climatic similarity, a principal component analysis on the weather-based covariates using the R package factoextra (Kassambara and Mundt, 2017) was applied.

In addition to climatic variables, our framework accommodates four soil-based environmental variables: soil

TABLE 1 | Environmental predictor variables used in the prediction models.

Acronym	General description
P.V, P.F, P.G	Accumulated precipitation + irrigation (mm) by growth period
FreqP5.V, FreqP5.F, FreqP5.G	Frequency of days with more than 5 mm precipitation by growth period
MeanT.V, MeanT.F, MeanT.G	Average of daily mean temperature (°C) by growth period
MinT.V, MinT.F, MinT.G	Average of minimum daily temperature (°C) by growth period
MaxT.V, MaxT.F, MaxT.G	Average of maximum daily temperature (°C) by growth period
GDD.V, GDD.F, GDD.G	Cumulative growing degree days, Base 10°C (°C) by growth period
Photothermal.Time.V, Photothermal.Time.F, Photothermal.Time.G	Cumulative photothermal time (GDD x Day Length) by growth period
FreqMaxT30.V, FreqMaxT30.F, FreqMaxT30.G	Frequency of days with maximum temperature above 30°C by growth period
FreqMaxT35.V, FreqMaxT35.F, FreqMaxT35.G	Frequency of days with maximum temperature above 35°C by growth period
St30.V, St30.F, St30.G	Sum of the daily maximal temperatures above 30°C (°C)
CumSumET0.V, CumSumET0.F, CumSumET0.G	Accumulated reference evapotranspiration (mm), under standard conditions, according to the FA0-56 Penman-Monteith methodology for each growth period
CumDailyWaterBalance.V, CumDailyWaterBalance.F, CumDailyWaterBalance.G	Cumulative daily water balance, i.e., daily precipitation + irrigation - daily reference evapotranspiration (mm)
Sdrad.V, Sdrad.F, Sdrad.G	Accumulated incoming daily solar radiation (MJ m-2 day-1) by growth period
SandProp.SC	Sand composition (%)
Silt.Prop.SC	Silt composition (%)
ClayProp.SC	Clay composition (%)
OM.SC	Percentage of organic matter (%)

The suffixes refer to: V, vegetative period; F, flowering period; G, grain fill period; SC, soil covariate.

quality types (percentages of sand, silt, and clay composition) and percentage of soil organic matter. The majority of the soil information originates from the soil samples realized at each G2F field location; otherwise, when the location presented missing information, we defined an area of interest based on field geographical coordinates using the Web Soil Survey application for American locations, and the web mapping application Agricultural Information Atlas for Canadian locations, and retrieved the aforementioned data of interest. In the rest of the paper, the abbreviation "W" refers to the set of weather-based and soil-based environmental covariates. For the trait plant height, weather-based covariates from the grain filling stage were not used as explanatory variable for prediction, since this trait was usually measured shortly after flowering time.

2.5. Prediction Models Implemented

2.5.1. Linear Random Effects Models (LRE Models)

In multi-environment trial analysis and plant breeding experiments, linear random effects models, abbreviated to LRE models thereafter, are often used as genomic prediction models and were compared in this study with machine learning techniques, according to the models outlined in Jarquín et al. (2014). In particular, GxE can be modeled with a covariance function equal to the product of two random linear functions of markers and of environmental covariates, which is equivalent to a reaction norm model (Jarquín et al., 2014). An environment always refers to a Site x Year combination.

Main effects models

(1) *Model G + E: Marker + Environment Main Effects (baseline model)*

The response variable is modeled as the sum of an overall mean (μ), plus random deviations due to the environment E_i and to the genotypic random effect of the jth hybrid genotype g_j based on marker covariates (G-BLUP component), plus an error term ε_{ij}:

$$y_{ij} = \mu + E_i + g_j + \varepsilon_{ij}, \tag{1}$$

where $E_i \overset{IID}{\sim} N(0, \sigma_E^2)$, $\mathbf{g} \overset{IID}{\sim} N(\mathbf{0}, \mathbf{G}\sigma_g^2)$ and $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, and N(.,.) denotes a normally distributed random variable, IID stands for independent and identically distributed, and σ_E^2, σ_g^2 are the corresponding environmental and genomic variances, respectively.

g_j corresponds to a regression on marker covariates of the form $g_j = \sum_{m=1}^{p} x_{jm} b_m$, linear combination of p markers and their respective marker effects. Marker effects were regarded as IID draws from normal distributions of the form $b_m \overset{IID}{\sim} N(0, \sigma_b^2)$, $m = 1,...,p$. The vector **g=Xb** follows a multivariate normal density with null mean and covariance-matrix $Cov(\mathbf{g}) = \mathbf{G}\sigma_g^2$, where $G = \frac{XX'}{p}$ is the genomic relationship matrix, X representing the centered and standardized genotype matrix and p is the total number of markers.

(2) *Model G + S: Marker + Site Main Effects*

The present model allows to gain information from a site evaluated over several years, as it includes the site effect:

$$y_{kj} = \mu + S_k + g_j + \varepsilon_{kj} \tag{2}$$

Here y_{kj} corresponds to the phenotypic response of the jth genotype in the kth site with $S_k \overset{IID}{\sim} N(0, \sigma_S^2)$, $k = 1,...,K$.

(3) *Model G+E+W: Marker + Ennvironment + Environmental Covariates Main Effects*

This model incorporates additionally the main effect of the environmental covariates (including the longitude and latitude coordinates). We can model the environmental effects by a random regression on the ECs (**W**), that represents the environmental conditions experienced by each hybrid in each environment: $w_{ij} = \sum_{q=1}^{Q} W_{ijq}\gamma_q$, where W_{ijq} is the value of the qth EC evaluated in the ijth environment x hybrid combination, γ_q is the main effect of the corresponding EC, and Q is the total number of ECs. We considered the effects of the ECs as IID draws from normal densities, i.e., $\gamma_q \sim N(0, \sigma_\gamma^2)$. Consequently, the vector $\mathbf{w} = \mathbf{W}\boldsymbol{\gamma}$ follows a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{\Omega}\sigma_w^2$, where $\boldsymbol{\Omega} \propto \mathbf{WW}'$, and the matrix $\mathbf{W}$, which is centered and standardized, contains the values of the ECs. The model becomes then:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + \varepsilon_{ij} \tag{3}$$

with $\mathbf{w} \sim N(\mathbf{0}, \Omega\sigma_w^2)$.

In this model, as explained in Jarquín et al. (2014), environmental effects are subdivided in two components, one that originates from the regression on numeric environmental variables, and one due to deviations from the Year-Site combination effect which cannot be accounted for by the ECs. Indeed, the environmental variables might not be able to fully explain the differences across environments. The modeling of the covariance matrices $\boldsymbol{\Omega}$ and $\mathbf{G}$ allows to borrow information between environments and between hybrid genotypes, respectively.

Models with interaction

(4) Model G+E+GxE: main effects G+E with Genomic x Environment Interaction

The model G+E was extended by including the interaction term between environments and markers (GxE):

$$y_{ij} = \mu + E_i + g_j + gE_{ij} + \varepsilon_{ij} \tag{4}$$

with $\mathbf{gE} \sim N(\mathbf{0}, [\mathbf{Z_g G Z_g'}] \circ [\mathbf{Z_E Z_E'}]\sigma_{gE}^2)$, $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_g}$ and $\mathbf{Z_E}$ are the design matrices that connect the phenotype entries with hybrid genotypes and with environments, respectively; σ_{gE}^2 is the variance component of the gE_{ij} interaction term; and $\circ$ denotes the Hadamard product between two matrices.

(5) Model G+S+GxS: main effects G+S with Genomic x Site Interaction

Similar to the previous model, this model extends model G+S by including the interaction term between sites and markers (GxS):

$$y_{kj} = \mu + S_k + g_j + gS_{kj} + \varepsilon_{kj} \tag{5}$$

where $\mathbf{gS} \sim N(\mathbf{0}, [\mathbf{Z_g G Z_g'}] \circ [\mathbf{Z_S Z_S'}]\sigma_{gS}^2)$, $\varepsilon_{kj} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_S}$ and σ_{gS}^2 are the design matrix for sites and the associated variance component for this interaction, respectively.

(6) Model G+E+S+Y+GxS+GxY+GxE: main effects G+E+S+Y with Genomic x Environment Interaction, Genomic x Site Interaction and Genomic x Year Interaction

This model corresponds to the most complete model using only basic GxE information (year and site information) about environments:

$$y_{jkm} = \mu + g_j + S_k + Y_m + E_{km} + gS_{jk} + gY_{jm} + gE_{jkm} + \varepsilon_{jkm} \tag{6}$$

where $\mathbf{gY} \sim N(\mathbf{0}, [\mathbf{Z_g G Z_g'}] \circ [\mathbf{Z_Y Z_Y'}]\sigma_{gY}^2)$, $\varepsilon_{kj} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_Y}$ and σ_{gY}^2 are the design matrix for years and the associated variance component for this interaction, respectively.

(7) Model G+E+W+GxW: main effects G+E+W with interactions between markers and environmental covariates

The model G+E+W was extended by adding the interaction between genomic markers and environmental covariates. Jarquín et al. (2014) demonstrated that this interaction term induced by the reaction-norm model can be described by a covariance structure which corresponds, under standard assumptions, to the Hadamard product of two covariance structures: one characterizing the relationships between lines based on markers information (e.g., $\mathbf{G}$), and one describing the environmental resemblance based on ECs (e.g., $\boldsymbol{\Omega}$). The vector of random effects, denoted **gw** represents the interaction terms between markers and ECs, is assumed to follow a multivariate normal distribution with null mean and covariance structure $[\mathbf{Z_g G Z_g'}] \circ \Omega$. The model can be expressed as follows:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + \varepsilon_{ij}, \tag{7}$$

with $\mathbf{gw} \sim N(\mathbf{0}, [\mathbf{Z_g G Z_g'}] \circ \Omega\sigma_{gw}^2)$.

(8) Model G+E+W+GxW+GxE: main effects G+E+W with Genomic x Environment Interaction and Genomic x Environmental Covariates Interaction

The interaction term gE_{ij} is incorporated in this model, because some GxE might not be completely captured by the interaction term gw_{ij}, and the model becomes:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + gE_{ij} + \varepsilon_{ij} \tag{8}$$

Main and interactions effects included in the different models described above are summarized in **Supplementary Table 5**. Models using W, i.e., the matrix of environmental covariates, were tested with and without longitude and latitude data included. Additional combinations of main effects and interactions not detailed here were also evaluated and results are presented as **Supplementary Material**. These models were implemented in a Bayesian framework using the R package BGLR (Pérez and de Los Campos, 2014), for which the MCMC algorithm was run for 42,000 iterations and the first 2000 cycles were removed as burn-in with thinning equal to 5.

2.5.2. Machine Learning Based-Methods Used

The potential of machine learning models was explored using the following three algorithms: the linear regularized Elastic Net (Zou and Hastie, 2005), XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017). All the machine learning regression models were conducted in R version 3.6.1 (R Core Team, 2019) using the tidymodels framework (Kuhn and Wickham, 2020) and wrapper functions of treesnip (https://github.com/curso-r/treesnip/). Elastic net is a regularized linear regression method that has proven to be useful with datasets characterized by multicollinearity to identify the most relevant predictor variables as well as reducing the computing time (Zou and Hastie, 2005). It corresponds to a linear combination of two penalty terms: the lasso (L1 regularization), noted $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ and the ridge (L2 regularization), noted $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$. While the L2 penalty tends to contract the coefficients of highly correlated features toward each other, the L1 penalty supports a sparse solution, as many coefficients are zeroed. However, this method does not account for interactions between features.

Originally introduced by Friedman (2001), gradient boosting approach sequentially builds an ensemble of decision trees, with each new tree improving the predictions of the previous one by fitting on its residual errors. Two implementations of gradient boosting of decision trees (GBDT) for regression were used: Light Gradient Boosting Machine (LightGBM) and eXtreme Gradient Boosting (XGBoost). The two GBDT frameworks stand out from other similar boosting algorithms regarding their efficiency, which can be achieved by their common implementation of a histogram-based method for split finding, which groups continuous features into discrete bins. Hence, the algorithm does not iterate through all feature values, which is extremely time-consuming, but instead performs splitting on the bins. This speeds up training for very large datasets, as well as reducing memory usage. LightGBM, developed more recently, incorporates additional features, among others a downsampling during the training on basis of gradients. GBDT frameworks can handle well various types of data (binary, continuous data), and they are relatively robust to the effects of outliers among predictor variables (Hastie et al., 2009). Decision trees can capture, by construction, higher-order interactions between features, as well as nonlinear relationships between predictors and response variable (Friedman, 2001). Hence, interactions do not need to be explicitly provided as input data, since new splits are built conditional on preceding splits made on other predictors.

2.5.3. Data Pre-processing for Machine Learning-Based Models

For data processing, we used the R package recipes (Kuhn and Wickham, 2020). To reduce genomic data dimensionality, we did not input SNP data into our prediction models directly. Instead, we used the top 275 or 350 principal components (PCs) of SNP data, for the traits grain yield and plant height, respectively. This set of PCs was chosen after evaluation of the predictive ability using different sets of top PCs explaining a various proportion of the variance in the data. Covariates which had no variance were removed using the step_nzv function. Retained covariates were standardized to zero mean and unit variance. As for linear random effect models, we tested the influence on prediction of longitude and latitude data by including and removing them as predictor variables across the different cross-validation scenarios. The year was also included as an input variable as a predictor variable in some models to account for environmental variation not fully captured by environmental covariates. In that case, the factor variable was converted into four new variables corresponding to each level of the original predictor. To model the site effect in models without numerical environmental information, we used the simple geographic coordinates of each location instead of using its label. Indeed, in decision trees, the use of a categorical predictor with a high number of levels can lead to overfitting (Hastie et al., 2009).

2.5.4. Optimization of Hyperparameters and Hyperparameter Importance for Machine Learning-Based Models

Bayesian optimization using an iterative Gaussian process was used for hyperparameter tuning. It represents a much faster approach than grid search while allowing more flexibility in how the parameter space is covered. The Gaussian process builds a probability model based on an initial set of performance metrics obtained for various hyperparameter combinations during an initialization step, and predicts new tuning hyperparameters to test based on these previous results (Williams and Rasmussen, 2006; Snoek et al., 2012). Bayesian optimization incorporates prior assumptions on model parameter distribution and update it after each iteration, seeking to minimize the root mean square error (RMSE). Hyperparameter tuning was evaluated with 30 iterations under resampling based on a fivefold cross-validation (CV) with two repeats on the training set. **Supplementary Table 4** indicates the set of hyperparameters tuned for each method during this optimization step. This set of hyperparameters was then used to fit the whole training data and predict the test set, which was unused during the optimization of hyperparameters. The general procedure for this nested cross-validation is illustrated in **Figure 1**. Fine-tuning of hyperparameters is required in order to prevent overfitting and to achieve the best prediction accuracy and representation of the data.

In addition, we examined the role of each hyperparameter on the overall model performance. This analysis provide insights into the most important hyperparameters to primarily tune in order to yield accurate models. We focus here on the LightGBM algorithm and XGBoost. A method based on random forests and functional ANOVA (fANOVA) was proposed by Hutter et al. (2014) to quantify the marginal contribution of each hyperparameter and pairwise interaction effects. Briefly, we used the output table of performance metrics of each algorithm with different hyperparameter combinations, which was obtained during the optimization step. The metric (root mean square error) is then used as target variable while hyperparameters represent the explaining variables to fit a random forest algorithm. fANOVA is then applied to evaluate the importance of each hyperparameter used in the grid search.

2.5.5. Assessment of Prediction Accuracy for New Environments

In order to mimic real plant breeding problems, we considered four different cross-validation strategies aiming at predicting genotypes in environments that were never tested before, namely CV0-Year, CV0-Site, CV00-Year, and CV00-Site, described in Jarquín et al. (2017). The CV0 cross-validation scheme allows to borrow information in the training set about the performance of predicted genotypes in other tested environments, while the CV00 cross-validation scheme consists of the prediction of newly developed genotypes. This means that for implementation of the CV00 cross-validation, any observation from a genotype included in the test set (i.e., new environments) was removed from the training set. Predictions of untested genotypes can be achieved by exploiting information from marker data on

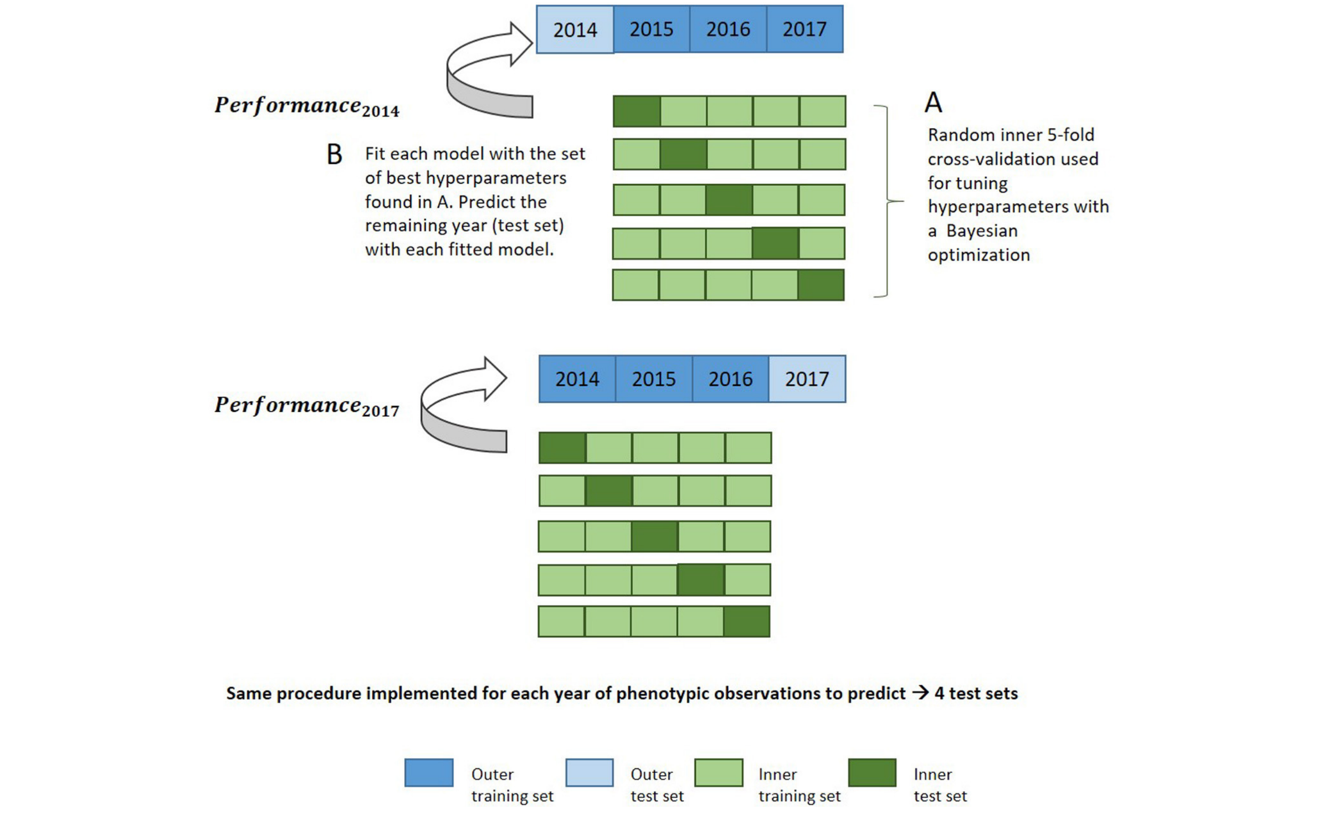
2014
2015
2016
2017
Performance2014
A
Random inner 5-fold cross-validation used for tuning hyperparameters with a Bayesian optimization
B
Fit each model with the set of best hyperparameters found in A. Predict the remaining year (test set) with each fitted model.
Performance2017
2014
2015
2016
2017
Same procedure implemented for each year of phenotypic observations to predict → 4 test sets
Outer training set
Outer test set
Inner training set
Inner test set

FIGURE 1 | Nested cross-validation diagram for evaluation of model performance in the leave-1-year-out CV scheme with a machine learning approach.

genetic similarities between genotypes from the training set and from the test set. Four scenarios in total were examined, which differ according to whether site or year were used to build the test set, and to the degree of relationship between training and test set: (1) CV0-Year, where phenotypic information about the performance of genotypes evaluated in the same year was masked; (2) CV00-Year, where phenotypic information about the performance of any genotypes present in the test set in other years was additionally masked; (3) CV0-Site, where phenotypic information about the performance of genotypes evaluated in the same site was masked and (4) CV00-Year, where phenotypic information about the performance of any genotypes present in the test set in other sites was additionally masked. In this procedure, the number of observations contained in each outer fold is not the same, due to the unbalanced character of the dataset. This approach reflects a common issue arising in multi-environment plant breeding trials, as all selection candidates cannot be grown in all environments. However, we can ensure a fair model comparison by having the same data splits across tested models.

Regarding evaluation metrics, we define the prediction accuracy as the Pearson correlation between the predicted and the observed performance in a given environment, i.e., correlations were computed on a trial basis.

In order to take into account the difference in sample sizes between environments, we evaluated the weighted average predictive ability across environments according to Tiezzi et al. (2017), for each combination of prediction model, predictor variables and trait, as following:

$$r_w = \frac{\sum_{j=1}^{J} \frac{r_j}{V(r_j)}}{\sum_{j=1}^{J} \frac{1}{V(r_j)}},$$

with r_j the Pearson's correlation between predicted and observed values at the j^{th} environment, $V(r_j)= \frac{1-r_j^2}{n_j-2}$ its sampling variance and n_j the total number of phenotypic observations in the j^{th} environment.

2.6. Variable Importance and Partial Dependence Plots for Grain Yield

We used the gain metric to quantify the feature importance in the XGBoost model fitted to the full dataset. This metric corresponds to the relative contribution of the variable to the ensemble model, calculated by considering each variable's contribution for each boosting iteration. A superior value of the gain for one feature compared to another feature means that this feature is more important to generate a‘prediction.

Overall partial dependence plots (PDPs) were computed using the R package DALEX (Biecek, 2018) using the four trained datasets from the CV0-Year scheme and the full dataset. PDPs are relevant to study how the predicted outcome of a machine learning model is partially influenced by a subset of explanatory variables of interest, by marginalizing over the values of all other variables.

The partial dependence profile of *f(X)* is defined as following by Friedman (2001):

$$f_S(X_S) = E_{X_C} f(X_S, X_C),$$

where the X_S represents the set of input predictor variables for which the effect on the prediction is analyzed, and X_C represent the complement set of other predictor variables used in the model. The following partial function can be used as an estimator:

$$\overline{f_S}(X_S) = \frac{1}{N} \sum_{i=1}^{N} f(X_S, x_{iC}),$$

where $x_{1C}, x_{2C}, ..., x_{NC}$ are the values of X_C observed in the training data. This means that we estimate this expected value as the average of the model predictions, over the joint distribution of variables in X_C, when the set of joint values in X_S is fixed. As emphasized by Hastie et al. (2009), partial dependence functions represent hence the influence of X_S on *f(X)*, after taking into account the average effects of the other variables X_C on *f(X)*.

2.7. Code Availability

A Github repository containing the various R scripts and Bash scripts used for phenotypic analysis, processing of weather data, spatio-temporal interpolation of missing weather data, and predictive modeling is available: https://github.com/cjubin/G2F_data.

3. RESULTS

3.1. Variability of Climatic Conditions in the Panel of Environments

Figure 2 reveals a partitioning of environments into clusters corresponding mostly to different US climate zones. It suggests that the sample of environments was broad enough to cover a large spectrum of environmental conditions across the North-American continent. The first two principal components explained more than 55% of total variation among environments on the basis of weather-based environmental covariates. The loading plot shows that MinT.F and GDD.F, FreqMaxT30.G, which are covariates related to temperature during flowering and grain filling stage, strongly influenced the first principal component (PC1). Environments from the South/Southeast (Arkansas, Texas, Georgia) showed positive PC1 and PC2 scores, which can be explained by a common humid subtropical climate, according to the Köppen climate type classification (Köppen and Geiger, 1930). One exception was one location in Texas (denoted 2014_TXH2), associated with more semi-arid climatic conditions. These results indicate that a closer geographical distance does not necessarily imply similar environmental conditions, based on climate types. For instance, environments from Delaware were closer to environments from the Midwest than Northeastern environments. Environments from the Midwest, associated with a humid continental climate, were situated mostly around the origin of the plot, and environments further north or in Canada exhibited the lowest temperatures among this set of sampled environments and presented a negative PC1 score.

3.2. Hyperparameter Importance for Gradient Boosting Approaches

Computing by fANOVA the marginal contribution of each tuned hyperparameter, using the performance data gathered during the hyperparameter optimization step on the different training sets, highlights large differences regarding their respective impact on model performance (**Supplementary Figure 3**). For the two gradient boosting algorithms, the learning rate (named eta in XGBoost) and the maximum depth of the tree were the most relevant algorithm parameters, as well as their interaction. The number of boosting iterations did not play a major role in model performance. We also found an advantage of using the hyperparameter feature_fraction and colsample_bytree, implemented in LightGBM and XGBoost, respectively, as it allowed an important reduction of the training time without having any observed negative effect on the accuracy of the predictions. It should be emphasized that we did not fully explore the influence of all possible hyperparameters implemented in these algorithms because of computational limitations, and therefore many of these were fixed during the hyperparameter optimization step.

3.3. Comparison of Model Performance Across Two Traits and Four Different CV Scenarios

CV0-Year

When the aim was to predict yield performance of already tested hybrids in new environments, the weighted average correlation of the baseline LRE model (G+E) was 0.356 (**Figure 3**; **Supplementary Table 6**). When the GxE term was added, the average correlation improved to 0.362. The model that included all interactions (G+E+W+GxW+GxE) was the best LRE model, while using only interactions between environmental covariates and genomic information (model G+E+W+GxW) slightly decreased the predictive ability of the baseline model to 0.347. In this prediction scenario, the two GBDT methods outperform all LRE models; model XGBoost-G+W+Y+Lon+Lat improved upon the baseline model by 18%. In addition, a small increase of predictive ability could be observed when environmental covariates were included as features for the machine learning-based frameworks. Furthermore, models that included geographical coordinates as predictor variables resulted in better prediction accuracies, and this revealed true across all prediction problems; therefore, **Figures 4**, **5** display results from LRE models using W as including longitude and latitude as predictor variables. For plant height, the baseline

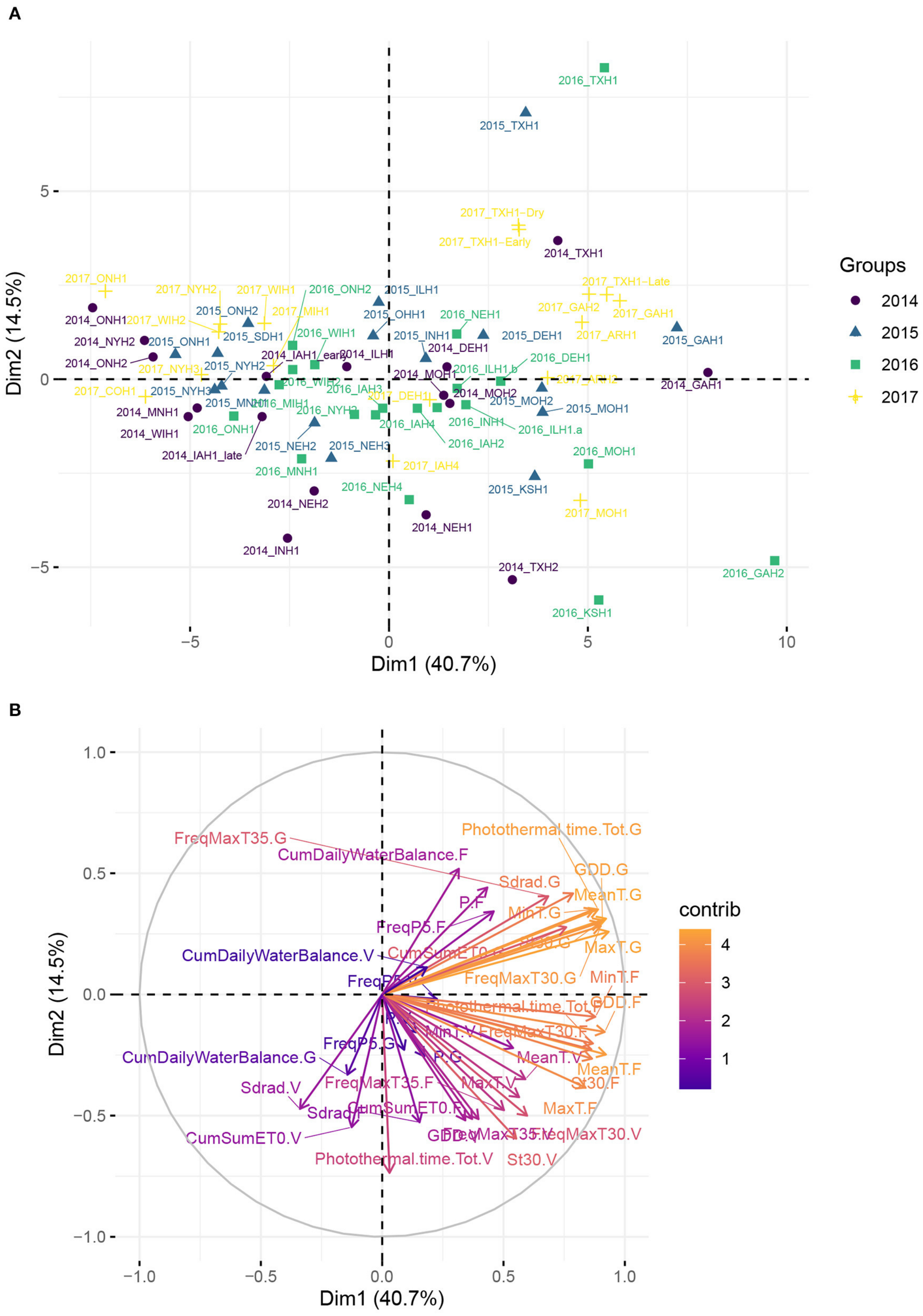

FIGURE 2 | Principal component analysis (PCA) plot of environmental data from the 71 environments, using the median flowering date as reference in each environment. **(A)** Maize trial experiments located in the US and in Canada used in analyses. Name of the locations and their geographical position are given in **Supplementary Table 1**. **(B)** Correlation plot of the weather-based covariates used in the PCA.

model performed best (**Figure 4**; **Supplementary Table 8**), and gradient boosting models incorporating environmental predictor variables performed consistently worse than models based only on genotypic data, geographical data and year information.

CV00-Year

CV00-Year produced lower average correlation coefficients for the two traits and for all models compared to CV0-Year, which illustrates that genomic prediction in multi-environment trials achieves better results when the training set includes information from the same genotypes evaluated in other environments. Regarding the trait grain yield (**Figure 3**; **Supplementary Table 6**), modeling the effect of sites instead of environments resulted in a small improvement of the predictive ability (4% better than the G+E model). Adding the GxE term to the LRE baseline model also positively affected the predictive ability (8% better than the G+E model). However, the LRE model with main site and genotype-by-site interaction effects (G+S+GxS) outperformed LRE models based on the modeling of year-location (E) effects. Overall the best predictive model for this trait was again the GBDT model XGBoost-G+W+Y+Lon+Lat, which displayed an average correlation of 0.301 (20% higher than the baseline model). GBDT models incorporating W performed between 6 and 13% better than GBDT models excluding W, which demonstrates the usefulness of environmental data for prediction of yield performance of new genotypes in an untested year. Among LRE models, the LRE model with all interactions and using enviromental data was the best model and resulted in an improvement of 17% over the baseline model. Regarding the trait plant height (**Supplementary Table 8**), the best predictive model was the baseline LRE model with an average weighted correlation of 0.604. Among LRE and GBDT models, models which did not include any environmental data performed better than those using these. An explanation for this lack of improvement with environmental data for plant height in this prediction problem can be that year and geographical position are appropriate and sufficient data to efficiently characterize environments for prediction of plant height, while using all environmental variables might generate noise here.

CV0-Site

The prediction of already tested genotypes in all environments associated with a common site revealed higher predictive abilities than with the CV0-Year prediction problem (**Figures 3**, **4**; **Supplementary Tables 7**, **9**). Indeed, based on our dataset, which covers many different sites across the US (see **Supplementary Figure 1**), the leave-one-site-out CV strategy generates large ratios across all training/test splits. This greater amount of data available to predict environments from one site can explain why this CV scheme obtained higher predictive abilities than the CV0-Year strategy. For the trait grain yield (**Figure 3**; **Supplementary Table 7**), the XGBoost-G+Lon+Lat+Y outperformed other models, showing an increase of 9% compared to the baseline LRE model. LightGBM models showed also better predictive abilities than LRE models. Only for LRE models did the use of environmental data yield a very small increase in predictive ability; the best result within this type of statistical approach was obtained by the model including all interactions (0.477, 3% higher than the baseline model). However, for the trait plant height (**Figure 4**; **Supplementary Table 9**), LRE models performed better than machine learning-based methods, with the model G+E+S+Y+GxS+GxY+GxE, which uses only basic information on environments, showing a mean correlation of 0.742. LightGBM and XGBoost methods with geographical and year information predicted reasonably well compared to the latter model (average r between 0.7 and 0.72), and again, the addition of environmental covariates decreased the predictive ability of GBDT models G+Lon+Lat+Y.

CV00-Site

As expected, the prediction of new genotypes in new sites resulted in lower mean correlations than CV0-Site for the two traits under study across predictive models. This highlights again the importance of the relationship between training and test sets. For the trait grain yield (**Figure 3**; **Supplementary Table 7**), the weighted average predictive ability of the reference model (G+E) was 0.248, and the model using sites instead of environment main effect was slightly better with a mean correlation of 0.265 (7% over G+E model). When the GxE term was added to the baseline model, the weighted average predictive ability was improved to 0.269 (8% over G+E model). It is worth to underline that models incorporating genotype-by-site effects performed even better (10% and 11% higher than the reference model). Modeling the interaction between ECs and genotypes and between environments and genotypes (model G+E+W+GxW+GxE) yielded an improvement of the baseline model by 19% (average r = 0.296), which was closely followed by the LightGBM and XGBoost models incorporating environmental covariates (between 11 and 16 % increase over the baseline model). As for the CV0-Year and CV00-Year CV schemes, the use of environmental data slightly increased the average predictive ability for grain yield. For the trait plant height (**Figure 4**; **Supplementary Table 9**), the baseline model with interactions by environment (G+E+GxE) outperformed other models. As for the previous prediction problems, environmental data decreased predictive abilities over all implemented models for the trait plant height.

When comparing the predictive abilities across traits, grain yield was the trait showing the lowest predictive ability across all CV schemes. Across all CV schemes, Elastic Net was the worst predictive modeling approach, which can be related to the absence of interactions between predictors in this model, if these are not explicitly provided as new features.

Figure 5; **Supplementary Tables 10**, **11** display the detailed within-environment correlation results for grain yield for two (CV0-Year and CV0-Site) cross-validation schemes. If a predicted environment is over the identity line, this means that there was an increment of the predictive ability by using environmental information. For CV0-Year, the machine learning-based model including environmental data outperformed the model only using geographical and year information in 44 of the 71 considered environments. For CV0-Site, however, the model with environmental features was better

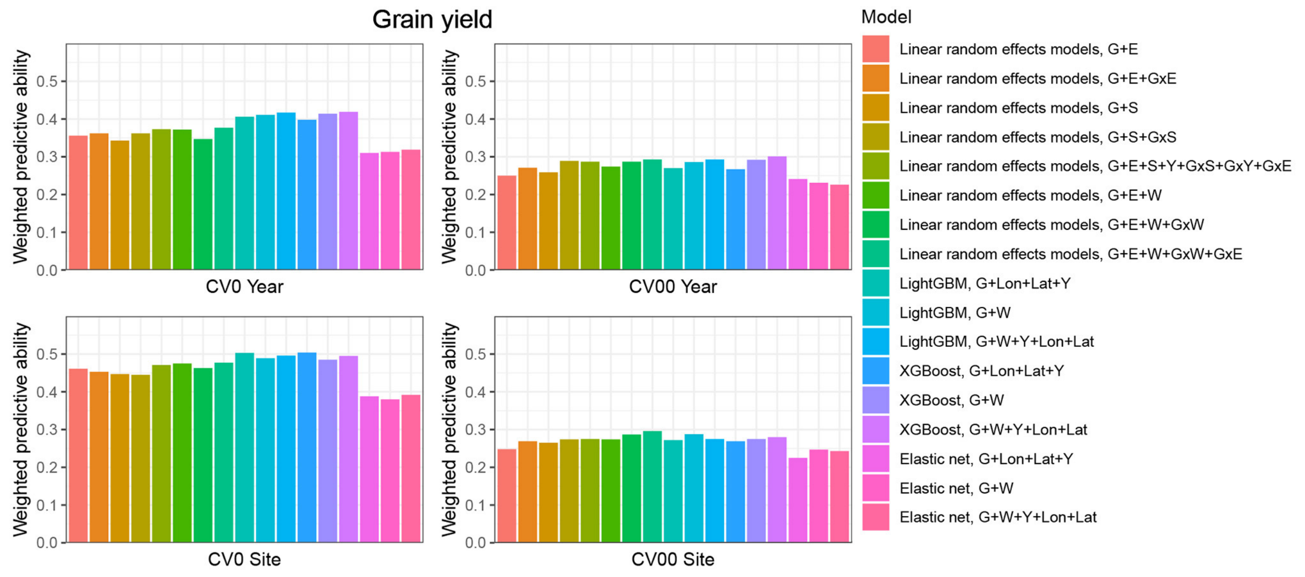

FIGURE 3 | Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait grain yield. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; GxS, genotype-by-site interaction; E, environment effect; GxY, genotype-by-year interaction; GxS, genotype-by-site interaction; GxE, genotype-by-environment interaction; GxW, interaction between W and SNPs; Lon, longitude; Lat, latitude; W, effect of weather- and soil-based covariates. For linear random effects models, results with models including longitude and latitude data in the matrix W are depicted here.

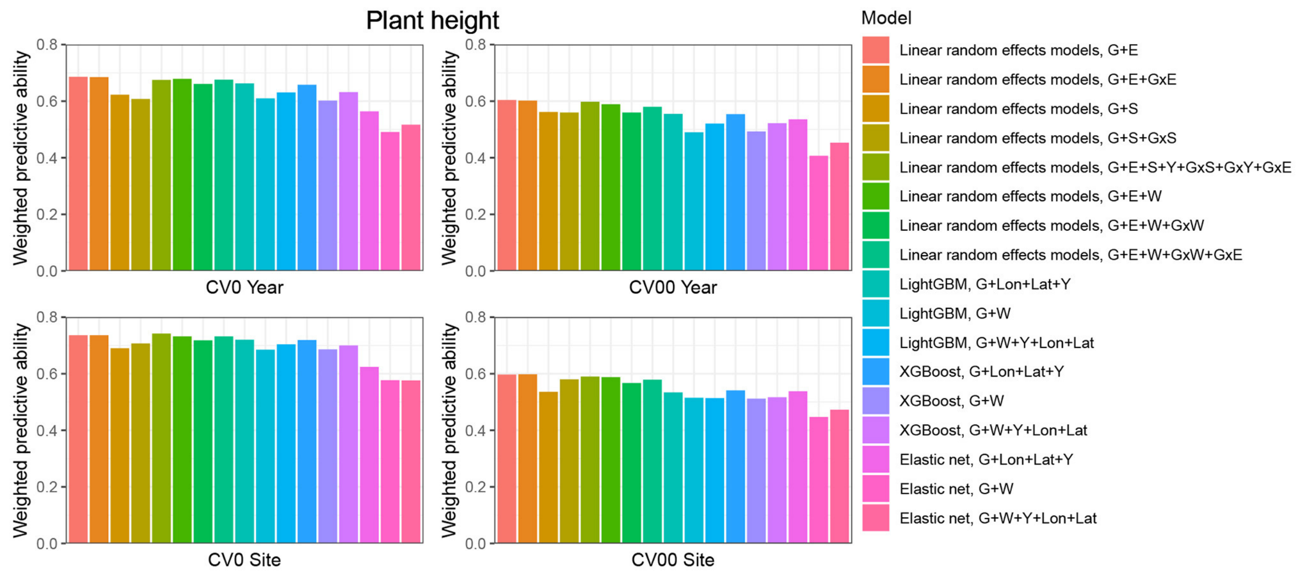

FIGURE 4 | Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait plant height. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; GxS, genotype-by-site interaction; E, environment effect; GxY, genotype-by-year interaction; GxS, genotype-by-site interaction; GxE, genotype-by-environment interaction; GxW, interaction between W and SNPs; Lon, longitude; Lat, latitude; W, effect of weather- and soil-based covariates. For linear random effects models, results with models including longitude and latitude data in the matrix W are depicted here.

than the less complex one in only 34 environments. This can be interpreted as a failure to explain a large part of the GxE by the computed ECs, and by a more efficient representation of environmental effects by simple geographic information.

3.4. Variable Importance

Regarding the trait grain yield, many of the identified top variables were related to temperature, such as the average minimum temperature during the flowering stage, or the

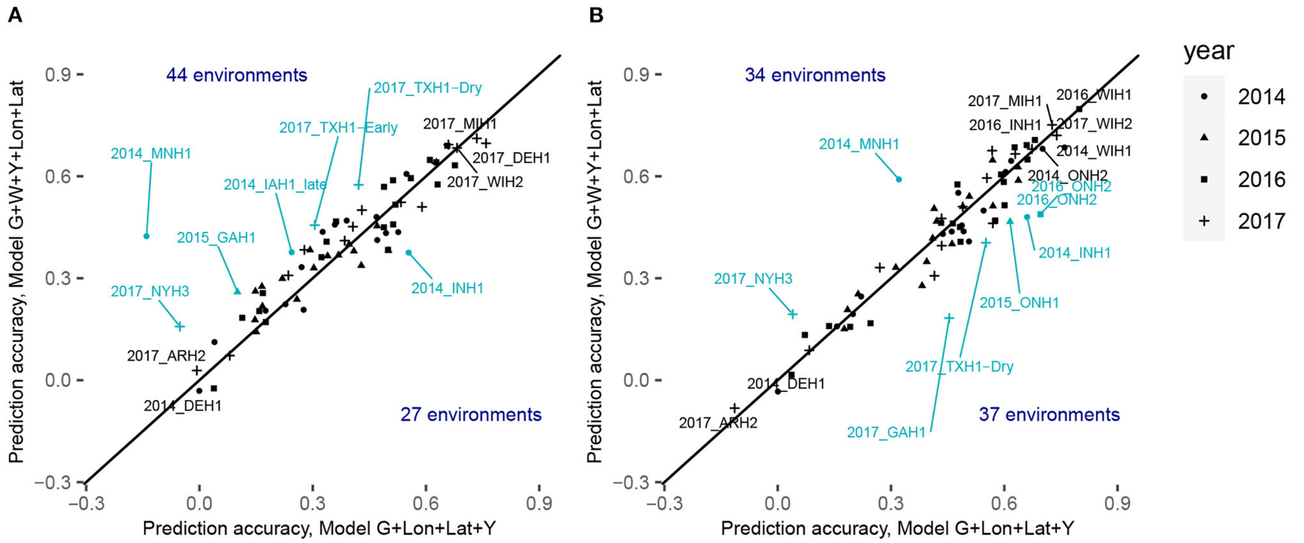

FIGURE 5 | Comparison of the within-environment predictive ability with different sets of predictors for the trait grain yield for XGBoost **(A)** with the CV0-Year scenario and **(B)** CV0-Site scenario. The x-axis corresponds to the within-environment correlation obtained with the model incorporating PCs derived from SNPs, year and geographical coordinates. The y-axis corresponds to the within-environment correlation obtained with the model incorporating PCs, year, W (i.e., weather- and soil-based covariates) and geographical coordinates. The line indicates the identity. Blue-colored points with a label indicate environments for which the absolute difference between the two predictive abilities was superior to 0.13. Black-colored points with a label indicate the least and the most accurately predicted environments.

frequency of days during which the maximum temperature was above 35°C (**Figure 6**). Organic soil matter concentration was the third most important feature, which demonstrates that fields with fertile soils were associated with higher yields. The amount of water received by the field (P.V) during the vegetative and grain filling stage was also a major feature for the model, as well as the frequency of days during the vegetative stage for which the amount of water was greater than 5 mm. Regarding the trait plant height, variables based on soil information played a major role for trait prediction, as they likely affect the crop shoot architecture. The amount of water received during the vegetative stage was also an important explanatory variable for plant height.

Partial dependence plots (**Figure 7**) show that minimum temperature at flowering stage was strongly impacting yield from approximately 20°C onwards. Maximum temperature during the vegetative stage had a detrimental effect on yield, suggesting that very elevated temperatures can impair a normal plant growth, eventually required to achieve optimal grain yield, although it tended to have a more gradual effect than minimum temperature at flowering stage. The relationship with yield of the total amount of precipitation during the vegetative stage was positive, before reaching a plateau. A high soil organic matter content yielded in superior yield predicted values.

4. DISCUSSION

Breeders, working on the development of climate resilient cultivars, risk making incorrect selection decisions if genotype-by-location and genotype-by-year interactions are not properly accounted for (Jarquín et al., 2017; De Los Campos et al., 2020). By incorporating environmental variables in our models, we assessed the value of these predictor variables for genomic prediction of complex phenotypes across four cross-validation scenarios. Gradient boosting frameworks based on decision trees have demonstrated high prediction performance for traits affected by non-additive effects (Abdollahi-Arpanahi et al., 2020), as well as model interpretability to extract important insights from the model's decision making process (Shahhosseini et al., 2020). Thus, a second objective was to evaluate these new prediction methods on the basis of prediction accuracies and for identification of the most relevant environmental variables.

4.1. Comparison of Prediction Methods Across the Two Traits

We observed that GBDT frameworks produced a slightly improved predictive ability for grain yield compared to the linear random effects models in three (CV0-Year, CV00-Year, and CV0-Site) out of the four CV schemes. However, no advantage was observed when GBDT was used to predict plant height. Overall, GBDT methods were competitive to LRE models, but we did not find any case where these machine learning-based methods considerably exceeded the predictive ability of LRE models. Previous studies have suggested that machine learning-based approaches can provide superior accuracy for prediction of phenotypic traits characterized by substantial non-additive effects. For instance, results from Zingaretti et al. (2020) in strawberries suggest that traits, exhibiting large epistatic effects, can be better predicted by convolutional neural networks (CNN), than by Bayesian penalized linear models. On the other hand, for

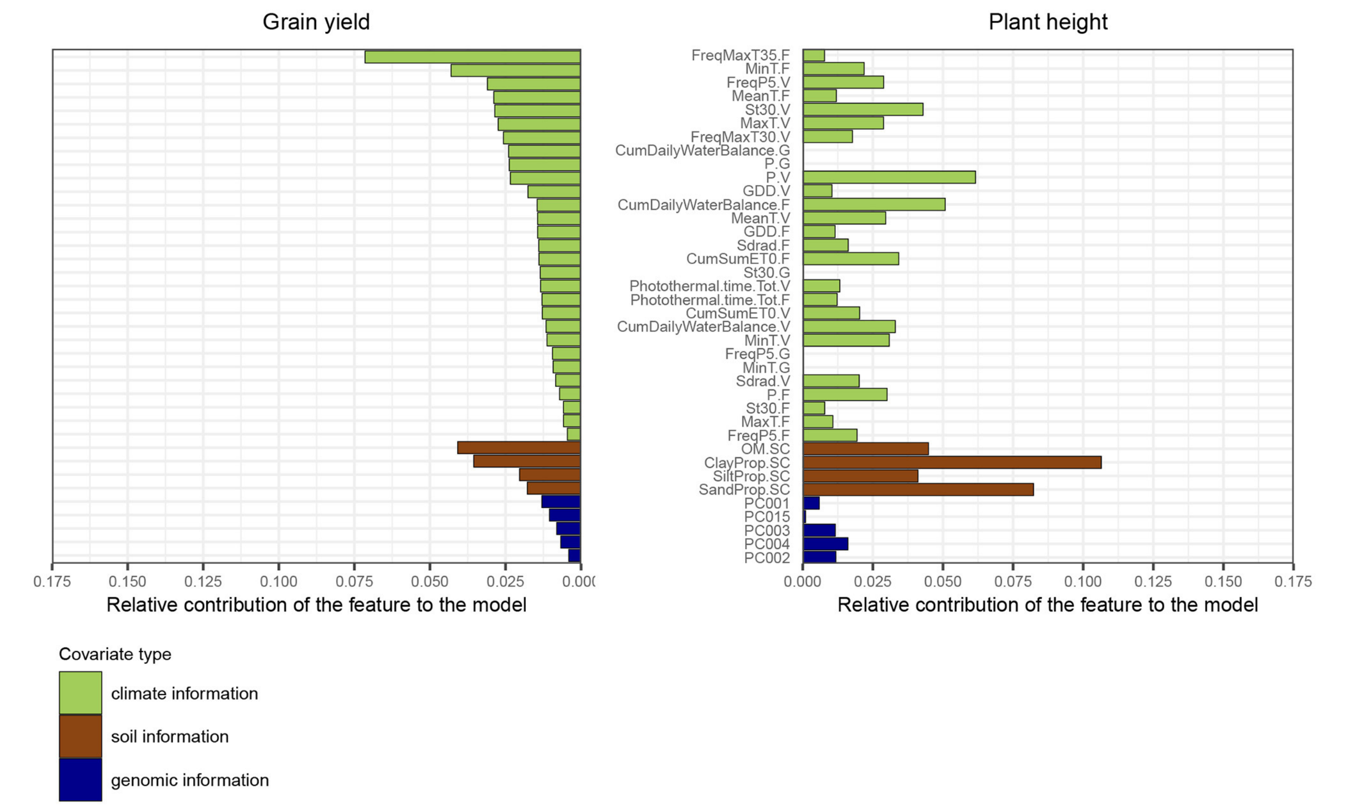

FIGURE 6 | Feature importance ranking based on the average relative gain per feature obtained with the model XGBoost-G+W, for the two traits grain yield and plant height. The metric was estimated using a model fitted on the full dataset. The gain represents the improvement in accuracy when using a feature for splitting, across all trees in the model. The order of features is based on feature performance within covariate class for the trait grain yield. The sum of all feature contributions is equal to 1. Weather-based variables from the grain filling stage were not used to predict plant height.

moderately to highly heritable traits, no real advantage of using machine learning-based methods was observed in their study. Bellot et al. (2018) pointed out that human height, a trait with a prevailing additive component and a polygenic architecture, was better predicted by linear methods than by CNNs. For other traits they examined in their study, a deep learning approach did not significantly outperform other methods in terms of prediction accuracy. Similar conclusions were drawn by Azodi et al. (2019) who reported an inconsistency of performance for non-linear machine learning-based algorithms in comparison with linear algorithms, according to the trait under study.

In our study, we incorporated not only genomic-based, but also environmental-based predictor variables. Yield component traits are controlled by numerous physiological processes under the influence of environmental factors, which can explain the large contribution of the GxE variance component for the phenotypic variance of grain yield, while for plant height, the proportion of variance explained by GxE is generally much lower than the proportion of variance related to genetic effects (Olivoto et al., 2017; Rogers et al., 2021). Nonlinear relationships between some environmental factors, such as temperature or rainfall amounts, and grain yield are well-known in the field of ecology and agriculture (Troy et al., 2015; Li et al., 2019). Hence, the slightly better prediction performance for grain yield with GBDT frameworks might originate from their ability to model nonlinear effects of environmental predictor variables, as observed with the partial dependence plots, as well as interactions with other predictor variables like genomic-based principal components. This asset was also described by Heslot et al. (2014b) when implementing soft rule fit (a modified ensemble method) capturing nonlinear interactions between markers and environmental stress covariates. Additional studies are required to validate this hypothesis using other phenotypic traits showing various genetic architectures. Moreover, it should be noted that we used only linear kernels in the reaction norm models to model genetic and environmental similarities. This means that we did not account for the specific combining ability (i.e., nonlinear genetic effects, due to dominance or epistasis, of specific hybrid combinations) which can influence the magnitude of yield heterosis in maize hybrids. Alternative approaches exist to model additive and dominant genetic effects, as well as environmental relatedness with nonlinear kernels (Bandeira e Sousa et al., 2017; Cuevas et al., 2018; Costa-Neto et al., 2020a). Bandeira e Sousa et al. (2017) and Cuevas et al. (2018) obtained better predictive abilities when using a Gaussian kernel rather than a linear GBLUP kernel with multi-environment G–E interactions models.

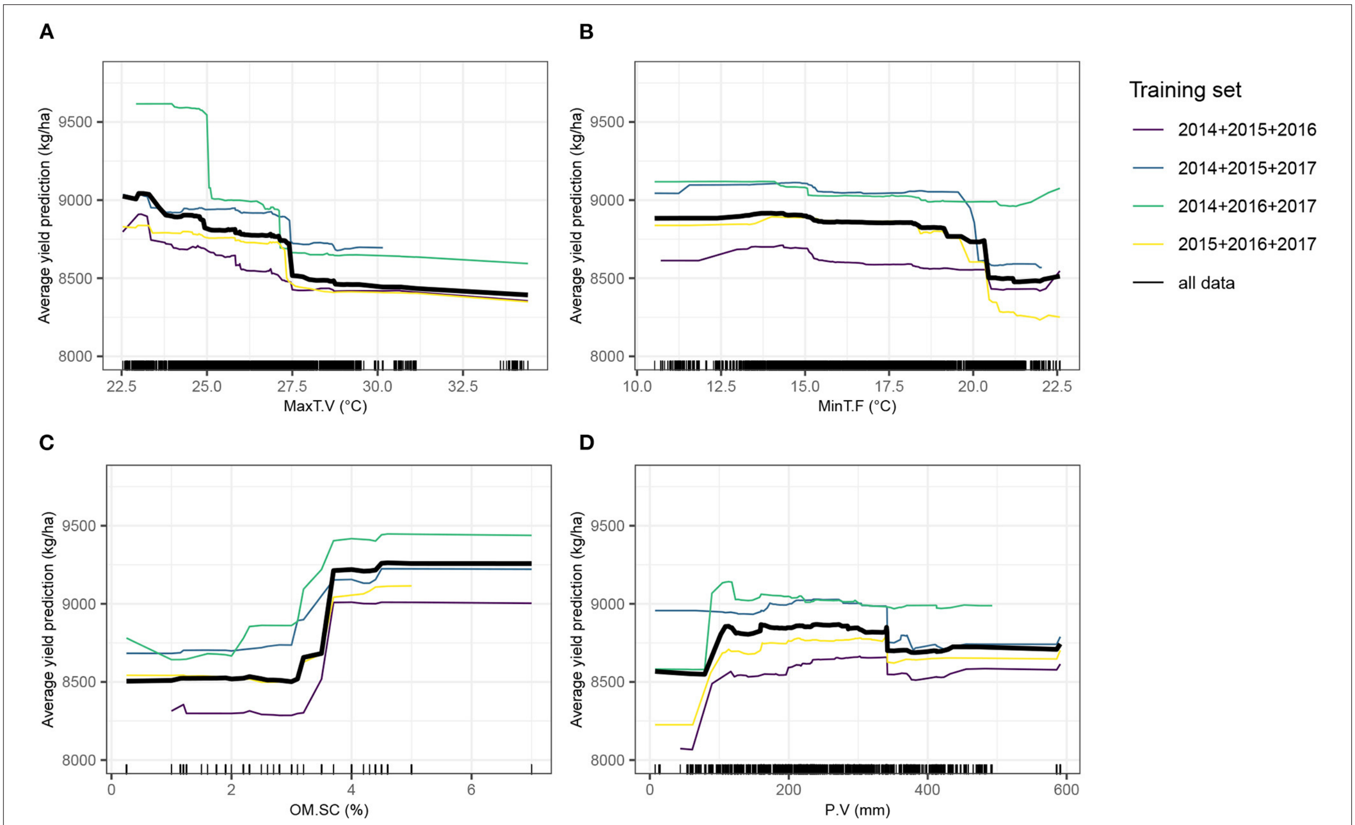

FIGURE 7 | Partial dependence plots (PDPs) showing the behavior of the expected value of predicted yield as a function of four top-ranked predictor variables. The Y-axis value of a PDP is calculated average of all model predictions obtained from the training dataset, when the value of the predictor variable is equal to X. The four training sets from the leave-1-year-out cross-validation scheme (CV0-Year) and the full dataset, separately trained with XGBoost, were used. Tick marks indicate individual observations. **(A)** MaxT.V, maximum temperature during the vegetative stage; **(B)** MinT.F, minimum temperature during the flowering stage; **(C)** OM.SC, percentage of soil organic matter; **(D)** P.V, Amount of precipitation and irrigation during the vegetative stage.

More recently, Costa-Neto et al. (2020a) implemented Gaussian and arc-cosine kernels-based approaches on both genomic and environmental datasets from a MET maize dataset, and noted an improvement in prediction accuracy using these methods across various cross-validation strategies. These results highlight the potential of nonlinear methods to better unravel nonlinear relationships existing in the input space.

4.2. Model Performance Under Various Prediction Problems

The four cross-validation schemes we evaluated represent challenging prediction problems. They seeked to assess the ability of the models to predict the effect of unknown combinations of environmental stresses on the studied phenotypic traits in a new year (CV0-Year and CV00-Year) or in a new site (CV0-Site and CV00-Site). Previously published work has revealed somewhat similar ranges of prediction accuracies for this trait in maize (Costa-Neto et al., 2020a; Jarquin et al., 2020). In winter wheat, Jarquín et al. (2017) and Sukumaran et al. (2017) reported the predictions of yield performance in future years (CV0-Year) as the most challenging prediction problem on the basis of results obtained for various cross-validation schemes, and results of Sukumaran et al. (2018) showed that modeling site effect instead of environment effect based on basic information about the environments (year and location) had a positive effect on predictive ability with CV0-Year, as we could also observe for CV0-Year, CV00-Year, and CV00-Site in our results. Indeed, this type of models allows to exploit information from the same site tested across several years. Another factor which is important to take into account in multi-year breeding data, as emphasized by Bernal-Vasquez et al. (2017), is the degree of genetic relatedness between the training and validation sets. Hence, CV00-Year and CV00-Site were more challenging prediction problems than CV0-Year and CV0-Site, respectively, and yielded lower weighted mean correlations across all models.

Regarding the usefulness of environmental information, the best model for grain yield based on mean predictive ability included these data for three (CV0-Year, CV00-Year, and CV00-Site) out of the four CV schemes. In addition, it must be taken into account that much less phenotypic observations were masked for CV0-Site (1/28, about 3.6% on average, with some sites being present more often than others across years in our dataset) than for CV0-Year (1/4, about 25% as the dataset is unbalanced). Hence, we can consider CV0-Year and

CV00-Year as more challenging prediction problems than CV0-Site and CV00-Site in our study. The improvement due to the incorporation of environmental data was however less remarkable and less consistent across CV schemes than expected, which was in contrast with previous results. Monteverde et al. (2019) also implemented a leave-1-year-out scenario, with one unique location present in the dataset, and the best prediction accuracies for grain yield were always reached by the models integrating environmental predictors alongside genomic predictors. Findings from Costa-Neto et al. (2020a) also show a significant increase of prediction accuracy with the linear GB kernel incorporating environmental data in a CV0 scheme, but the authors additionally modeled dominant genetic effects, which were not accounted for in our study. On the other hand, Jarquin et al. (2020) also used the same Genomes to Fields dataset and reported a lack of enhancement when using a model that solely incorporated interactions between genotype and environmental covariates (i.e., without using the environment label). The best predictive models for the CV0 and CV00 schemes, that they implemented, included both genotype-by-environment and genotype-by-EC interactions, similarly to our results (**Supplementary Tables 6–9**). In agreement with the reasons invoked by the authors of this study, we argue that environmental data are especially relevant for predictions when a larger number of environments is used, e.g., by testing sites within a limited geographical range with relatively similar environmental conditions across multiple years. This was for example achieved in the study of De Los Campos et al. (2020), where 16 sites located in France were tested over 16 years. A reasonable hypothesis is that historical weather data obtained across multiple years for a specific geographical area can lend the model reliable information on the effect of year-to-year climatic variation on phenotypic performance, in addition to site-based factors (soil and geographical position). A finding supporting this hypothesis is that the environments, which showed the best prediction accuracies with an environmental model, corresponded generally to the sites which were repeated across years, like Madison (WI) or College Station (TX) (**Supplementary Tables 10**, **11**). Interestingly, 2014_TXH2, a location for which data were only included for a single year, showed a moderate prediction accuracy with the XGBoost model without environmental information in CV0-Year ($r = 0.28$; **Supplementary Table 10**), which was superior to the model with environmental covariates ($r = 0.21$ with all environmental covariates included). We can suppose that the inclusion of environmental information, when predicting a new environment with properties that are very different from environments covered by the training set, is not useful to enhance the predictive ability of the model using basic predictors, such as the year factor and geographic coordinates. Extreme weather events can make some environments very unpredictable. 2017_ARH1 and 2017_ARH2 exhibited a very low prediction accuracy for grain yield (< 0 for 2017_ARH2) in both CV0-Year and CV0-Site (**Supplementary Table 11**), which is likely to be related to the effect of the tropical storm Harvey at the end of August 2017, which caused substantial lodging due to wind and excessive rainfall affecting the yield, and was reported by collaborators in the metadata.

4.3. Incorporation of Weather-Covariates in the Predictive Models

The use of environmental information yielded a small gain in average prediction accuracy for many models tested on grain yield, but did not lead to any improvement for plant height. For this latter trait, the large influence of soil-based variables, illustrated by the variable importance ranking (**Figure 6**), can also possibly explain why prediction models using only geographical coordinates outperformed more elaborate models. For this trait, latitude and longitude data might indirectly capture information which is site-specific and repeatable across years, e.g., related to the quality of soil. For instance, environments from the Corn Belt, which were present in our dataset, usually exhibited fertile soils with much higher organic soil matter content than environments located in other US regions. Costa-Neto et al. (2020b) highlighted that simple geographic-related information, such as longitude and latitude data, can also efficiently represent environmental patterns that are specific to a site (for instance related to soil characteristics), and hence capture well genotype-by-site interaction while using only two variables.

In general, the lack of real enhancement of predictive ability may result from the way we incorporated developmental stages into our models, as we defined only three main developmental stages (i.e., vegetative, flowering and grain filling stages). Trial data often lack a rigorous collection of phenological data due to phenotyping costs. A possible solution to predict plant developmental stages can be to use crop models, such as APSIM (Holzworth et al., 2014) or SiriusQuality (Keating et al., 2003), as done in related studies (Heslot et al., 2014b; Rincent et al., 2017, 2019; Bustos-Korts et al., 2019). In our case, we did not implement a crop model since we aimed at estimating the flowering stage at the hybrid level as accurately as possible, as it is known to be a critical period for the determination of yield-related components. Therefore, we based our environmental characterization on available field data (sowing date and silking date scored) in order to derive environmental covariates for three main developmental stages, similarly to Monteverde et al. (2019) in rice. The reported variability among crop growth models (CGM) in simulating temperature response can complicate the task of choosing the most appropriate one (Bassu et al., 2014). In addition, the task of integrating genetic variation for earliness in crop growth models can also be rather challenging, with the risk that the predicted developmental crop stages might not appropriately reflect the plant developmental stages observed in the field if the model does not properly account for genotype-specific parameters (Rincent et al., 2019). Technow et al. (2015) developed a complex framework combining both CGM and whole-genome prediction, where the CGM is used to predict grain yield as a function of several physiological traits and of weather and management data. Genotype-specific physiological parameters were estimated in this study by running a Bayesian algorithm which models them as linear functions of the effects of genomic features. It would be of high interest to apply CGM models on this dataset by taking advantage of the flowering time data that are available. We should also mention that other types of input data could be incorporated in future analyses, such as the type of field management, the field disease pressure, preceding crop, or the presence of external treatments (organic, nitrogen fertilizers).

4.4. Prerequisites to Use Machine Learning-Based Models and Their Usefulness to Understand Significant Environmental Factors

Specific techniques should be employed to ensure an efficient application of machine learning-based models. These can provide better results when expert knowledge is incorporated (Kagawa et al., 2017; Roe et al., 2020; Brock et al., 2021). Here, we restricted weather information to the duration of the growing season, transformed some raw weather information into new variables (evapotranspiration) and built stress indices besides typical climate covariates based on previous biological knowledge (e.g., detrimental temperature thresholds for maize (Greaves, 1996; Schlenker and Roberts, 2009; Lobell et al., 2014; Zhu et al., 2019; Mimić et al., 2020). Prior understanding of the role of input features can help mitigate the risk of using irrelevant information in the model. As expected, the correlation matrix between environmental covariates (**Supplementary Figure 2**) showed that numerous predictor variables were highly correlated with each other, especially those related to temperature and heat stress. We did not perform feature selection based on the Pearson correlation coefficients between environmental covariates, because of the risk of dropping highly predictive variables, since the metric ignores the relationship to the output variable. In addition, methods based on decision trees can perform internal feature selection, making them robust to the inclusion of irrelevant input variables and to multicollinearity (Hastie et al., 2009; Kuhn et al., 2013). If two variables are strongly correlated, the decision tree will pick either one or the other when deciding upon a split, which should not eventually affect prediction results. Another approach to reduce the number of features and reduce training time is to apply feature extraction, as we did by deriving principal components from the genotype matrix and use these as new predictor variables in the machine learning-based models. This procedure did not seem to affect model performance.

Machine learning models often require an elaborated hyperparameter optimization strategy, implying for example a nested cross-validation approach which can be computationally expensive (Varma and Simon, 2006), since it involves a series of train/validation/test set splits to prevent data leakage. Inadequate model tuning can result in a suboptimal performance of the algorithm. Here, we found that the hyperparameters such as the learning rate or tree depth were relevant regularization parameters to reduce the model complexity, thereby dealing with overfitting. In accordance with these results, other authors had also reported these two hyperparameters as the most important ones for another gradient boosting library similar to LightGBM, Adaboost (Van Rijn and Hutter, 2018). In general, lower values of the learning rate (< 0.01) are recommended to reach the best optimum (Ridgeway, 2007). Nonetheless, as the learning rate is decreased, more iterations are needed to get to the optimum, which implies an increase of the computation time and of additional memory (Ridgeway, 2007; Kuhn et al., 2013). With regard to the tree depth, a relatively low maximal depth generally helped to prevent overfitting, and better results were generally obtained with our data using a tree depth lower than to 8. The deeper a tree is, the more splits it contains, resulting in very complex models which do not generalize well on new data. Knowledge regarding the most important hyperpararmeters to tune is useful if limited computational resources hamper the investigation of numerous hyperparameter combinations during the training phase. Our results demonstrated similar predictive abilities of LightGBM and XGBoost, with a clear speed advantage for LightGBM, which ran often more than twice as fast. This asset relies in particular on a feature implemented in LightGBM, the gradient-based one-side sampling method (GOSS), which implies that not all data actually contribute equally to training. Training instances with large training error (i.e., larger gradients) should be re-trained, while data instances with small gradients are closer to the local minima and indicate that data is well-trained. Hence, this new sampling approach focuses on data points with large gradients and keeps them, while randomly sampling from those with smaller gradient values. A drawback of this method is the risk of biased sampling which might change the distribution of data, but this issue is mitigated in LightGBM by increasing the weight of training instances with small gradients. The main advantage is that it makes LightGBM much faster with comparable accuracy results. Another crucial aspect when applying machine learning models is the adequacy of the dataset for machine learning applications, which should be large enough to allow the algorithm to learn from the data (Géron, 2019). In our case, we benefited from a very large training dataset and a low feature-to-instance ratio (316/18,325).

In our study, on top of prediction applications, tree-based methods were also used to obtain estimates of feature importance, and thereby contributed to a better understanding of key abiotic factors driving the response of the tested genotypes. Feature importance rankings and partial dependence profiles showed that the minimal temperatures and indices related to prolonged heat stress, or to amounts of water received in the field, especially at the flowering stage, ranked among the most important variables for grain yield. When comparing these results with established agronomic knowledge, it was reported that, above a certain threshold, high minimum temperature can lead to an increase of the rate of senescence and reduce the ability of the plant to produce grain across many plant species (Hatfield et al., 2011; Hatfield and Prueger, 2015). Previous research also revealed that increases in average night temperatures were associated with a reduction of grain yield in maize (Millet et al., 2019) and in rice (Welch et al., 2010). In an alternative study on rice cultivars in Colombia, Delerce et al. (2016) identified high minimum temperature (above 22.7°C) as one of the most important environmental factors negatively impacting grain yield by using a machine learning approach based on conditional inference trees. Exposure to temperatures exceeding 35°C during the flowering stage was also a key factor in our study (best predictor variable for grain yield), which can be related to a loss of pollen viability, and consequently to a reduced final kernel set (Hatfield et al., 2011). In our study, water availability at vegetative and grain-filling stages appeared to affect yield, in accordance with the literature outlining that any water deficit during these growth stages can impact grain yield (Denmead and Shaw, 1960; Cakir, 2004), with a more significant impact when water stress occurs during the grain-filling stage (Cakir, 2004). Caution should nonetheless be taken regarding feature importance ranking due to the important correlations between some environmental variables. Furthermore, only 4 years of field trials were used in our analyses, therefore variable importances could be refined with additional data from following years, to mitigate the influence of some environments characterized by adverse climatic conditions and potentially acting as outliers.

4.5. Applications

The usefulness of medium to high prediction accuracies, when predicting the performance in a new environment, must always be related to our predictability of the environmental variation. If the weather fluctuates considerably year to year, then the environmental predictors used to compute these predictions might be very different from the true value in the corresponding year. In addition, even if more precise climate change models were available to improve upon the precision of environmental predictors, predictions of observations falling outside the applicability domain, i.e., the range of predictor space in the training set for which the model can give relativey accurate predictions (Netzeva et al., 2005), might not be trustworthy and should be used cautiously (Kuhn et al., 2013). The degree of similarity of the new test set to the training set should hence always be carefully considered.

While some environmental factors are repeatable from year to year, such as the soil type or agronomic practices, a large part of the GxE variation is attributable to weather patterns. Hence, the success of this type of prediction scenario depends on the relative stability of the climate in the targeted regions across years. Nonetheless, we posit that our approach presents two key advantages to predict performance in future years. First, because they are fundamentally data-directed, the tree-based models can take into account new phenotypic data in the training set in a more flexible manner than classical mixed models, without the need to explicitly specify interactions for example. The development of high-throughput phenotyping technologies announces a future enhancement of rapid and accurate training data (Juliana et al., 2019). The predictive frameworks we presented here can make use of new information to refine the estimated effects of the predictor variables. Secondly, we were able to predict a quantitative phenotype in a new environment by using a novel configuration of genotypic and environmental predictors describing it. A point of interest relates to resource allocation and the possibility to select more efficiently candidates to test in field trials. Based on the exploration of different plausible climatic scenarios—within a range of conditions experienced by the training set—these models can help to evaluate which genotypes might be more adapted to which range of environmental conditions. For regions or target population of environments presenting relatively stable climatic conditions across years, the probability of success of this type of predictive modeling approach is heightened.

5. CONCLUSIONS

Encouraged by the effectiveness of machine learning-based frameworks reported in the recent literature across various research fields, we compared two popular ensemble models with linear random effects models implemented in a Bayesian framework and a regularized linear model. In three CV schemes with the trait grain yield, the use of gradient boosting models resulted in a slight improvement of the average predictive ability but not for plant height. This finding indicates that machine learning-based approaches can be envisaged for genomic prediction but their efficiency may vary according to the trait under study and its degree of responsiveness to environmental variation. For a trait strongly under the influence of environmental factors, machine learning-based models could provide predictive abilities similar or slightly superior to linear random effects, and could additionally be used for interpretation of feature ranking and to build partial dependence plots detailing relationships between predictor variables and outcome. Provided further efficiency gains in machine learning algorithms, as well as the standardization and harmonization of large-scale environmental data, new opportunities in the field of predictive modeling for developing climate resilient varieties appear forthcoming.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. Raw genotypic, phenotypic, weather, and soil data from the Genomes to Fields Initiative can be found at: https://datacommons.cyverse.org/browse/iplant/home/shared/commons_repo/curated/GenomesToFields_2014_2017_v1.

AUTHOR CONTRIBUTIONS

CW analyzed the data and wrote the manuscript. TB and HS supervised research. CW, TB, HS, GM, and PT designed the study. TB, HS, GM, SdS, and PT supported with statistical advice. CW, TB, HS, GM, SdS, PT, MS, and J-CR participated in the interpretation of results and contributed to discussion. All authors contributed to the writing of the final draft and approved the manuscript.

FUNDING

Financial support for CW was provided by KWS SAAT SE by means of a Ph.D. fellowship. Additional financial support was provided by the University of Göttingen and by the Center for Integrated Breeding Research. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

ACKNOWLEDGMENTS

The authors would like to thank the G2F Consortium for making data publicly available, sharing them among collaborators, and for their constructive feedback with this study. The authors acknowledge support by the computing center for the university of Göttingen (GWDG) through the use of their High Performance Computing resources. We would like to thank two reviewers for their thoughtful ideas and comments that improved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.699589/full#supplementary-material

REFERENCES

Abdollahi-Arpanahi, R., Gianola, D., and Peagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel. Evolut*. 52, 12. doi: 10.1186/s12711-020-00531-z

AlKhalifah, N., Campbell, D. A., Falcon, C. M., Gardiner, J. M., Miller, N. D., Romay, M. C., et al. (2018). Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. *BMC Res. Notes* 11:452. doi: 10.1186/s13104-018-3508-1

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-Fao Irrigation and Drainage Paper 56, Vol. 300*. Rome: Fao. D05109.

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de Los Campos, G., and Shiu, S.-H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3* 9, 3691–3702. doi: 10.1534/g3.119.400498

Bandeira e Sousa, M., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., et al. (2017). Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. *G3* 7, 1995–2014. doi: 10.1534/g3.117.042341

Baskerville, G. L., and Emin, P. (1969). Rapid estimation of heat accumulation from maximum and minimum temperatures. *Ecology* 50, 514–517. doi: 10.2307/1933912

Bassu, S., Brisson, N., Durand, J.-L., Boote, K., Lizaso, J., Jones, J. W., et al. (2014). How do various maize crop models vary in their responses to climate change factors? *Glob. Chang Biol*. 20, 2301–2320. doi: 10.1111/gcb.12520

Bates, D., Mchler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw. Articles* 67, 1–48. doi: 10.18637/jss.v067.i01

Behravan, H., Hartikainen, J. M., Tengström, M., Pylkäs, K., Winqvist, R., Kosma, V.-M., et al. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: a case study in Finnish cases and controls. *Sci. Rep*. 8, 1–13. doi: 10.1038/s41598-018-31573-5

Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298

Bernal-Vasquez, A.-M., Gordillo, A., Schmidt, M., and Piepho, H.-P. (2017). Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. *BMC Genet*. 18:51. doi: 10.1186/s12863-017-0512-8

Biecek, P. (2018). Dalex: Explainers for complex predictive models in r. *J. Mach. Learn. Res*. 19, 3245–3249.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brock, J., Lange, M., Tratalos, J. A., More, S. J., Graham, D. A., Guelbenzu-Gonzalo, M., et al. (2021). Combining expert knowledge and machine-learning to classify herd types in livestock systems. *Sci. Rep*. 11, 1–10. doi: 10.1038/s41598-021-82373-3

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype environment interaction using pedigree and dense molecular markers. *Crop Sci*. 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Bustos-Korts, D., Boer, M. P., Malosetti, M., Chapman, S., Chenu, K., Zheng, B., et al. (2019). Combining crop growth modeling and statistical genetic modeling to evaluate phenotyping strategies. *Front. Plant Sci*. 10:1491. doi: 10.3389/fpls.2019.01491

Butler, E. E., and Huybers, P. (2015). Variations in the sensitivity of US maize yield to extreme temperatures by region and growth phase. *Environ. Res. Lett*. 10, 034009. doi: 10.1088/1748-9326/10/3/034009

Cakir, R. (2004). Effect of water stress at different development stages on vegetative and reproductive growth of corn. *Field Crops Res*. 89, 1–16. doi: 10.1016/j.fcr.2004.01.005

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chenu, K. (2015). Characterising the crop environment – Nature, significance and applications. In: *Crop Physiology. Applications for Genetic Improvement and Agronomy*, eds Sadras V. and Calderini D. London: Elsevier, 321–348. doi: 10.1016/B978-0-12-417104-6.00013-3

Cicchino, M., Edreira, J. I. R., Uribelarrea, M., and Otegui, M. E. (2010). Heat stress in field-grown maize: response of physiological determinants of grain yield. *Crop Sci*. 50, 1438–1448. doi: 10.2135/cropsci2009.10.0574

Cooper, M., and DeLacy, I. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet*. 88, 561–572. doi: 10.1007/BF01240919

Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2020a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106. doi: 10.1038/s41437-020-00353-1

Costa-Neto, G. M. F., Júnior, O. P. M., Heinemann, A. B., de Castro, A. P., and Duarte, J. B. (2020b). A novel gis-based tool to reveal spatial trends in reaction norm: upland rice case study. *Euphytica* 216, 1–16. doi: 10.1007/s10681-020-2573-4

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

Crossa, J., Neto, R.-F., Montesinos-López, O. A., Costa-Neto, G. M. F., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: optimising the use of genomics, phenomics and enviromics data. *Front. Plant Sci*. 12:332. doi: 10.3389/fpls.2021.651480

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., Bandeira e Sousa, M., et al. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3* 8, 1347–1365. doi: 10.1534/g3.117.300454

De Los Campos, G., Pérez-Rodríguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars performances under uncertain weather conditions. *Nat. Commun*. 11, 1–10. doi: 10.1038/s41467-020-18480-y

Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Pati, V. H., et al. (2016). Assessing weather-yield relationships in rice at local scale using data mining approaches. *PLoS ONE* 11:e0161620. doi: 10.1371/journal.pone.0161620

Denmead, O., and Shaw, R. H. (1960). The effects of soil moisture stress at different stages of growth on the development and yield of corn 1. *Agron. J*. 52, 272–274. doi: 10.2134/agronj1960.00021962005200050010x

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *J. Anim. Ecol*. 77, 802–813. doi: 10.1111/j.1365-2656.2008.01390.x

Ersoz, E. S., Martin, N. F., and Stapleton, A. E. (2020). On to the next chapter for crop breeding: convergence with data science. *Crop Sci*. 60, 639–655. doi: 10.1002/csc2.20054

Estévez, J., Gavilán, P., and Giráldez, J. V. (2011). Guidelines on validation procedures for meteorological data from automatic weather stations. *J. Hydrol*. 402, 144–154. doi: 10.1016/j.jhydrol.2011.02.031

Falcon, C. M., Kaeppler, S. M., Spalding, E. P., Miller, N. D., Haase, N., AlKhalifah, N., et al. (2020). Relative utility of agronomic, phenological, and morphological traits for assessing genotype-by-environment interaction in maize inbreds. *Crop Sci*. 60, 62–81. doi: 10.1002/csc2.20035

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232. doi: 10.1214/aos/1013203451

Fukuda, S., Spreer, W., Yasunaga, E., Yuge, K., Sardsud, V., and Müller, J. (2013). Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agric. Water Manage*. 116, 142–150. doi: 10.1016/j.agwat.2012.07.003

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., et al. (2017). The effect of artificial selection on phenotypic plasticity in maize. *Nat. Commun*. 8, 1–11. doi: 10.1038/s41467-017-01450-2

Géron, A. (2019). *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.

Gillberg, J., Marttinen, P., Mamitsuka, H., and Kaski, S. (2019). Modelling G–E with historical weather information improves genomic prediction in new environments. *Bioinformatics* 35, 4045–4052. doi: 10.1093/bioinformatics/btz197

González-Recio, O., Jiménez-Montero, J., and Alenda, R. (2013). The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *J. Dairy Sci*. 96, 614–624. doi: 10.3168/jds.2012-5630

Gräler, B., Pebesma, E., and Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *R J*. 8, 204–218. doi: 10.32614/RJ-2016-014

Greaves, J. A. (1996). Improving suboptimal temperature tolerance in maize-the search for variation. *J. Exp. Bot*. 47, 307–323. doi: 10.1093/jxb/47.3.307

Haley, C., and Visscher, P. (1998). Strategies to utilize marker-quantitative trait loci associations. *J. Dairy Sci*. 81, 85–97. doi: 10.3168/jds.S0022-0302(98)70157-2

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2nd Edn*. New York, NY: Springer.

Hatfield, J. L., Boote, K. J., Kimball, B., Ziska, L., Izaurralde, R. C., Ort, D., et al. (2011). Climate impacts on agriculture: implications for crop production. *Agron. J*. 103, 351–370. doi: 10.2134/agronj2010.0303

Hatfield, J. L., and Prueger, J. H. (2015). Temperature extremes: effect on plant growth and development. *Weather Climate Extremes* 10, 4–10. doi: 10.1016/j.wace.2015.08.001

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

Heslot, N., Akdemir, D., Sorrells, M. E., and Jannink, J.-L. (2014b). 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

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

Hutter, F., Hoos, H., and Leyton-Brown, K. (2014). "An efficient approach for assessing hyperparameter importance," in *International Conference on Machine Learning* (PMLR), 754–762.

Jarquin, D., De Leon, N., Romay, M. C., Bohn, M. O., Buckler, E. S., Ciampitti, I. A., et al. (2020). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front. Genet*. 11:1819. doi: 10.3389/fgene.2020.592769

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., et al. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype× environment interactions in kansas wheat. *Plant Genome* 10, 1–15. doi: 10.3835/plantgenome2016.12.0130

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet*. 127, 595–607. doi: 10.1007/s00122-013-2243-1

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., et al. (2016). Random forests for global and regional crop yield predictions. *PLoS ONE* 11:e0156571. doi: 10.1371/journal.pone.0156571

Juliana, P., Montesinos-López, O. A., Crossa, J., Mondal, S., Pérez, L. G., Poland, J., et al. (2019). Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theor. Appl. Genet*. 132, 177–194. doi: 10.1007/s00122-018-3206-3

Kagawa, R., Kawazoe, Y., Ida, Y., Shinohara, E., Tanaka, K., Imai, T., et al. (2017). Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach. *J. Diabetes Sci. Technol*. 11, 791–799. doi: 10.1177/1932296816681584

Kassambara, A., and Mundt, F. (2017). Package factoextra. Extract and visualize the results of multivariate data analyses 76.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "Lightgbm: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 30, 3146–3154.

Keating, B. A., Carberry, P. S., Hammer, G. L., Probert, M. E., Robertson, M. J., Holzworth, D., et al. (2003). An overview of apsim, a model designed for farming systems simulation. *Eur. J. Agron*. 18, 267–288. doi: 10.1016/S.1161-0301(02)00108-9

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., and Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep*. 10, 1–12. doi: 10.1038/s41598-020-68771-z

Köppen, W., and Geiger, R. (1930). *Handbuch der Klimatologie, Vol. 1*. Gebrüder Borntraeger Berlin.

Kuhn, M., and Johnson, K. (2013). *Applied Predictive Modeling, Vol. 26*. New York, NY: Springer.

Kuhn, M., and Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Available online at: https://www.tidymodels.org

Lampa, E., Lind, L., Lind, P. M., and Bornefalk-Hermansson, A. (2014). The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environ. Health* 13:57. doi: 10.1186/1476-069X-13-57

Li, B., Zhang, N., Wang, Y.-G., George, A., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Front. Genet*. 9:237. doi: 10.3389/fgene.2018.00237

Li, Y., Guan, K., Schnitkey, G. D., DeLucia, E., and Peng, B. (2019). Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the united states. *Glob. Chang Biol*. 25, 2325–2337. doi: 10.1111/gcb.14628

Lizaso, J., Ruiz-Ramos, M., Rodríguez, L., Gabaldon-Leal, C., Oliveira, J., Lorite, I., et al. (2018). Impact of high temperatures in maize: phenology and yield components. *Field Crops Res*. 216, 129–140. doi: 10.1016/j.fcr.2017.11.013

Lobell, D. B., Roberts, M. J., Schlenker, W., Braun, N., Little, B. B., Rejesus, R. M., et al. (2014). Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. *Science* 344, 516–519. doi: 10.1126/science.1251423

Malosetti, M., Bustos-Korts, D., Boer, M. P., and van Eeuwijk, F. A. (2016). Predicting responses in multiple environments: issues in relation to genotype environment interactions. *Crop Sci*. 56, 2210–2222. doi: 10.2135/cropsci2015.05.0311

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., and Van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying qtl by environment interaction. *Euphytica* 137, 139–145. doi: 10.1023/B:EUPH.0000040511.46388.ef

McFarland, B. A., AlKhalifah, N., Bohn, M., Bubert, J., Buckler, E. S., Ciampitti, I., et al. (2020). Maize genomes to fields (g2f): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Res. Notes* 13, 1–6. doi: 10.1186/s13104-020-4922-8

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet*. 51, 952–956. doi: 10.1038/s41588-019-0414-y

Mimić, G., Brdar, S., Brkić, M., Panić, M., Marko, O., and Crnojević, V. (2020). engineering meteorological features to select stress tolerant hybrids in maize. *Sci. Rep*. 10, 1–10. doi: 10.1038/s41598-020-60366-y

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., and Edwards Jr, T. C. (2006). Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecol. Modell*. 199, 176–187. doi: 10.1016/j.ecolmodel.2006.05.021

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., and Myles, S. (2015). Linkimpute: fast and accurate genotype imputation for nonmodel organisms. *G3* 5, 2383–2390. doi: 10.1534/g3.115.021667

Monteverde, E., Gutierrez, L., Blanco, P., Prez de Vida, F., Rosas, J. E., Bonnecarrre, V., et al. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa L.*) grown in subtropical areas. *G3* 9, 1519–1531. doi: 10.1534/g3.119.400064

Mushore, T., Manatsa, D., Pedzisai, E., Muzenda-Mudavanhu, C., Mushore, W., and Kudzotsa, I. (2017). Investigating the implications of meteorological indicators of seasonal rainfall performance on maize yield in a rain-fed agricultural system: case study of mt. darwin district in zimbabwe. *Theor. Appl. Climatol*. 129, 1167–1173. doi: 10.1007/s00704-016-1838-2

Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T., Gramatica, P., et al. (2005). Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: the report and recommendations of ecvam workshop 52. *Alternat. Lab. Anim*. 33, 155–173. doi: 10.1177/026119290503300209

Ogutu, J. O., Piepho, H.-P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc*. 5, 1–5. doi: 10.1186/1753-6561-5-S3-S11

Olivoto, T., Nardino, M., Carvalho, I., Follmann, D., Ferrari, M., Szareski, V., et al. (2017). Reml/blup and sequential path analysis in estimating genotypic values and interrelationships among simple maize grain yield-related traits. *Genet. Mol. Res*. 16, 1–19. doi: 10.4238/gmr16019525

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers Geosci.* 30, 683–691. doi: 10.1016/j.cageo.2004.03.012

Pérez, P., and de Los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pérez-Rodríguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., and de los Campos, G. (2015). A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. *Crop Sci*. 55, 1143–1151. doi: 10.2135/cropsci2014.08.0577

Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., et al. (2017). Single-step genomic and pedigree genotype× environment interaction models for predicting wheat lines in international environments. *Plant Genome* 10:plantgenome2016-09. doi: 10.3835/plantgenome2016.09.0089

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81, 559–575. doi: 10.1086/519795

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rahmstorf, S., Foster, G., and Cazenave, A. (2012). Comparing climate projections to observations up to 2011. *Environ. Res. Lett*. 7, 044035. doi: 10.1088/1748-9326/7/4/044035

Ridgeway, G. (2007). Generalized boosted models: a guide to the gbm package. *Update Univ S C Dep Music. 1, 2007*.

Rincent, R., Kuhn, E., Monod, H., Oury, F.-X., Rousset, M., Allard, V., et al. (2017). Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet*. 130, 1735–1752. doi: 10.1007/s00122-017-2922-4

Rincent, R., Malosetti, M., Ababaei, B., Touzy, G., Mini, A., Bogard, M., et al. (2019). Using crop growth model stress covariates and ammi decomposition to better predict genotype-by-environment interactions. *Theor. Appl. Genet*. 132, 3399–3411. doi: 10.1007/s00122-019-03432-y

Ritchie, S. W., Hanway, J. J., Benson, G. O., Herman, J. C., and Lupkes, S. J. (1993). *How a Corn Plant Develops. Iowa State University Cooperative*. Extension Special report 48.

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., et al. (2020). Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. *PLoS ONE* 15:e0231300. doi: 10.1371/journal.pone.0231300

Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3*. 11:jkaa050. doi: 10.1093/g3journal/jkaa050

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., and Hugot, J.-P. (2019). Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. *Sci. Rep*. 9, 1–18. doi: 10.1038/s41598-019-46649-z

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

Shahhosseini, M., Hu, G., and Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *Front. Plant Sci*. 11:1120. doi: 10.3389/fpls.2020.01120

Snoek, J., Larochelle, H., and Adams, R. P. (2012). "Practical bayesian optimization of machine learning algorithms," in *Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 2, NIPS'12* (Red Hook, NY: Curran Associates Inc.), 2951–2959.

Sparks, A. (2018). nasapower: a nasa power global meteorology, surface solar energy and climatology data client for r. *J. Open Source Softw*. 3:1035. doi: 10.21105/joss.01035

Sukumaran, S., Crossa, J., Jarquín, D., and Reynolds, M. (2017). Pedigree-based prediction models with genotype× environment interaction in multienvironment trials of cimmyt wheat. *Crop Sci*. 57, 1865–1880. doi: 10.2135/cropsci2016.06.0558

Sukumaran, S., Jarquin, D., Crossa, J., and Reynolds, M. (2018). Genomic-enabled prediction accuracies increased by modeling genotype× environment interaction in durum wheat. *Plant Genome* 11, 1–11. doi: 10.3835/plantgenome2017.12.0112

Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., and Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Curr. Biol.* 27, R770–R783. doi: 10.1016/j.cub.2017.05.055

Technow, F., Messina, C. D., Totir, L. R., and Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate bayesian computation. *PLoS ONE* 10:e0130855. doi: 10.1371/journal.pone.0130855

Tiezzi, F., de Los Campos, G., Gaddis, K. P., and Maltecca, C. (2017). Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. *J. Dairy Sci*. 100, 2042–2056. doi: 10.3168/jds.2016-11543

Trnka, M., Rtter, R. P., Ruiz-Ramos, M., Kersebaum, K. C., Olesen, J. E., Ealud, Z., et al. (2014). Adverse weather conditions for european wheat production will become more frequent with climate change. *Nat. Clim. Chang* 4, 637–643. doi: 10.1038/nclimate2242

Troy, T. J., Kipgen, C., and Pal, I. (2015). The impact of climate extremes and irrigation on us crop yields. *Environ. Res. Lett*. 10:054013. doi: 10.1088/1748-9326/10/5/054013

van Eeuwijk, F. A., Denis, J. B., and Kang, M. S. (1996). "Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables." in *Genotype-by-Environment Interaction,* eds M. S. Kang and H. G. Gauch (Boca Raton, FL: CRC Press Inc.), 15–50.

Van Rijn, J. N., and Hutter, F. (2018). "Hyperparameter importance across datasets," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, 2367–2376.

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7:91. doi: 10.1186/1471-2105-7-91

Welch, J. R., Vincent, J. R., Auffhammer, M., Moya, P. F., Dobermann, A., and Dawe, D. (2010). Rice yields in tropical/subtropical asia exhibit large but opposing sensitivities to minimum and maximum temperatures. *Proc. Natl. Acad. Sci. U.S.A*. 107, 14562–14567. doi: 10.1073/pnas.1001222107

Williams, C. K., and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning,* Vol. 2. Cambridge, MA: MIT Press.

Yu, J., Shi, S., Zhang, F., Chen, G., and Cao, M. (2019). Predgly: predicting lysine glycation sites for homo sapiens based on xgboost feature optimization. *Bioinformatics* 35, 2749–2756. doi: 10.1093/bioinformatics/bty1043

Zahumenský, I. (2004). *Guidelines on Quality Control Procedures for Data From Automatic Weather Stations*. World Meteorological Organization.

Zhu, P., Zhuang, Q., Archontoulis, S. V., Bernacchi, C., and Müller, C. (2019). Dissecting the nonlinear response of maize yield to high temperature stress with model-data integration. *Glob. Chang Biol*. 25, 2470–2484. doi: 10.1111/gcb.14632

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci*. 11:25. doi: 10.3389/fpls.2020.00025

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc*. 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare 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.