RESEARCH PAPER

Climate and genetic data enhancement using deep learning analytics to improve maize yield predictability

Parisa Sarzaeim[1], Francisco Muñoz-Arriola[1,2,*,] and Diego Jarquín[3]

[1] Department of Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE 68583-0726 USA
[2] School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE 68583-0996 USA
[3] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583-0915 USA

* Correspondence: fmunoz@unl.edu

Abstract

Despite efforts to collect genomics and phenomics ('omics') and environmental data, spatiotemporal availability and access to digital resources still limit our ability to predict plants' response to changes in climate. Our goal is to quantify the improvement in the predictability of maize yields by enhancing climate data. Large-scale experiments such as the Genomes to Fields (G2F) are an opportunity to provide access to 'omics' and climate data. Here, the objectives are to: (i) improve the G2F 'omics' and environmental database by reducing the gaps of climate data using deep neural networks; (ii) estimate the contribution of climate and genetic database enhancement to the predictability of maize yields via environmental covariance structures in genotype by environment (G×E) modeling; and (iii) quantify the predictability of yields resulting from the enhancement of climate data, the implementation of the G×E model, and the application of three trial selection schemes (i.e. randomization, ranking, and precipitation gradient). The results show a 12.1% increase in predictability due to climate and 'omics' database enhancement. The consequent enhancement of covariance structures evidenced in all train–test schemes indicated an increase in maize yield predictability. The largest improvement is observed in the 'random-based' approach, which adds environmental variability to the model.

Keywords: Climate data science, deep neural network (DNN), genotype by environment (G×E) model, Genomes to Fields (G2F), maize yield predictability, train–test schemes.

Introduction

Global crop production is required to rise by 100–110% to meet the demands of the growing population by 2050, and, specifically, this value needs to increase 70% in the case of cereal yield (Tilman *et al.*, 2011; Alexandratos and Bruinsma, 2012; Matthews *et al.*, 2013). At the local scale, weather and climate impact on crop production, leading to positive and negative production trends across the globe (Ray *et al.*, 2015).

Furthermore, data availability and missing values may constrain our ability to diagnose and predict complex crop traits, mainly yields subject to different environmental conditions (Hoogenboom, 2000). This study aims to improve crop yield predictability in multienvironments by enhancing climate data and genomics and phenomics data ('omics') structures for better crop phenotypic responses.

Worldwide efforts have been made to predict phenotypes of major crops, mainly yield under current and future climate variations in a range of spatial and temporal resolution scales (Stehfest *et al.*, 2007). Despite improvements in crop model performance (biophysical and statistical models), climatic drivers remain an unclear factor in the diagnostics and prognostics of crop productivity. Unlike the biophysical modeling efforts for yield prediction (Olesen *et al.*, 2000; Mbungu *et al.*, 2015; Raoufi and Soufizadeh, 2020), the statistical models provide an opportunity to analyze genetic variation in the modeling procedure, such as factorial regression (Baril *et al.*, 1995), the Finlay–Wilkinson model (Finlay and Wilkinson, 1963), quantitative trait locus-based models (Hayes *et al.*, 1993), and genomic selection models (Meuwissen *et al.*, 2001). These methods facilitate quantifying interactive effects of genes and environments called genotype by environment (G×E) interactions across environments (Meuwissen *et al.*, 2001; Jarquín et al., 2014, 2017; Crossa *et al.*, 2017). In particular, genomic selection models use all molecular markers for phenotypic prediction (Crossa *et al.*, 2017), enabling the use of complex environmental data. Jarquín *et al.* (2014) improved a genomic selection model incorporating G×E interactions between markers and environmental factors via covariance structures.

Several studies have investigated the impacts of climate variables on yield, such as minimum and maximum temperature, solar radiation, and rainfall (Hoogenboom, 2000; Tao *et al.*, 2008). Ray *et al.* (2015) found that climate variability explains about one-third of crop yield variability, following spatial patterns. Also, Lobell *et al.* (2009) reported global average yields of major crops at ~80% of the potential yield in most irrigated fields, suggesting improvements of cropping systems through climate adaptation actions. These improvements require a deeper understanding of how genotypes interact with environments and how phenotypes respond to key climate covariables, turning the selection of superior lines in each environment into sound breeding, a decrease in the yield gap between the current and the potential yields, and food production sufficiency (Osei *et al.*, 2014). Duvick (2005) highlighted that 50–60% of maize yields rely on genetic improvements, indicating the critical role of genetics in breeding tolerant varieties in response to biotic and abiotic stresses. However, the net climate effects on the predictability of phenotypes, including environmental variables such as temperature, dew point, relative humidity, precipitation, and their interactions with genomics, remain unclear. The elucidation of the role that climate drivers play in phenotypic responses can advance our understanding of how crop phenotypes respond to weather and climate changes across regions and scales.

Several national and international projects have conducted breeding trials for the major crops to record the genotypic, phenotypic, and environmental datasets at field scale for training and testing multienvironment statistical trait simulation efforts. Some of these are the International Maize and Wheat Improvement Center (genomics.cimmyt.org), ARVALIS (arvalis-infos.fr),

Genomes to Fields (G2F) initiative (genomes2fields.org), and SOYNAM (soybase.org). The G2F's 'Genotype by Environment' project records, synthesizes, and releases large-scale, multiyear, and multienvironment data of maize breeding trials across North America (Lawrence-Dill *et al.*, 2019; McFarland *et al.*, 2020). However, several environmental data gaps exist in the recorded environmental time series due to technological, logistic, and experimental design complications, limiting their use when fitting G×E models.

High-dimensional databases have been created to manage and harness an increasing availability of data from novel, advanced, and low-cost technologies to traditional digital products (Shekhar *et al.*, 2017, Preprint; Quiñones *et al.*, 2021). Expressions of these high-dimensional digital products are the 'omics' databases, which have contributed to improving the diagnostics and predictability yields through statistical methodologies such as the covariance structures (Howard *et al.*, 2014; Jarquín *et al.*, 2021) and machine learning (Long *et al.*, 2011). The consolidation of such complex databases can benefit from the expansion of climatological stations, the development of gridded products and models, and the enhancement of other 'omics' digital resources. The enhancement of 'omics' and climate data contributes to better understanding of the propagation of errors from the climate data to the creation of covariance structures and the same predictability of phenotypes (Sarzaeim *et al.*, 2020). Machine learning techniques such as artificial neural networks, support vector machines, and deep neural networks (DNNs) have been widely used to improve environmental data gaps of multiple complexities. Some experiences include improvements in daily precipitation (Hernández *et al.*, 2016; Kumar et al., 2019, 2021), solar radiation (Ghimire *et al.*, 2019), temperature (Amato *et al.*, 2020), and those that include complex integrated climate, agricultural, and hydrologic processes (Amaranto et al., 2018, 2019, 2020). A DNN is a sophisticated supervised and multilayer artificial neural network that potentially outperforms more traditional machine learning techniques such as the support vector machine method. DNNs enable the recognition of complicated unknown mathematical relationships between a large number of input(s) and output data more efficiently. Here we take advantage of the capability of DNNs to impute the missing values and improve the database for yield predictability.

The hypotheses associated with this study are formulated as follows: (i) the G×E predictability of maize yield values increases by using a DNN-enhanced environmental covariance matrix interacting with genetic markers; and (ii) the contribution of environmental variability will be evidenced in G×E prediction skill by the selection of train–test structure based on randomness, ranked covariance matrix, and the gradient of single climate driver values (e.g. precipitation).

To test the hypotheses, the objectives of this research are defined as: (i) to improve the G2F 'omics' and environmental database by reducing the gaps in climate data using a DNN, enabling a larger sample of high-throughput markers and

tested phenotypes; (ii) to estimate the contribution of database enhancement (from objective i) to the predictability of maize yield based on the interactions of the enhanced genetic molecular markers with the environment through the environmental covariance structures in G×E modeling; and (iii) to identify the contribution and the possible attribution to the predictability of phenotypes by enhancing climate data (via environmental covariance matrices) to G×E model predictability by applying three trial selection schemes (random, ranked, and gradient) to provide evidence for the role played by the randomization, ranking of the environmental covariance matrix, and precipitation gradient.

This study is structured as follows: first, the G2F database including 'omics' and environmental (i.e. climate) data and their limitations is described. Then, the methodology of the G2F database for improvement of the evaluation–improvement pipeline to categorize G2F experiments and fill in the missing values is described in detail. Next, the G×E model and the equations employed incorporating the main and interactive effects of climate and genetics are explained to evaluate the improved data effect on predictability skills. To evidence the contribution of climate to predictability enhancement, three train–test design approaches based on random selection, ranked environmental covariance, and precipitation gradient schemes are discussed. Finally, the results, discussion, concluding remarks, and proposed future work for subsequent efforts in maize phenotypic predictability analysis and improvements are also presented.

Study area and data

The G2F initiative, established in 2014, is one of the most comprehensive public-accessible maize breeding databases (Lawrence-Dill *et al.*, 2019). The G2F initiative has operated several maize field trial plots across the USA and Ontario in Canada since 2014, assembling a unique large-scale, multiyear, and multienvironment data source for detailed and accurate maize breeding research. The database includes maize inbreeds' genetic molecular markers (G), phenotypic measurements (P) during and at the end of the growing season, and environmental data (E), mainly climatic variables captured during the crop development in every single experimental trial. The G2F data enable agricultural researchers, engineers, and economists to understand the main and interaction effects of 'omics' and environmental drivers on maize phenotypic responses, which helps to develop climate-adaptive and resilient maize cropping systems for economically important traits such as yield.

In the current study, we downloaded, processed, and used the G2F data collected between 2014 and 2017 for maize hybrid yield prediction. In this period, 98 hybrid experiments have been tested, including >46 000 maize replicated individual hybrids selected from >3300 unique maize lines. The G2F experiments are distributed across more than 17 states in the USA and the province of Ontario in Canada (Fig. 1). More than one hybrid trial has been implemented in multiple cases in one experimental field. The name of each hybrid experimental trial is constituted of 'year', 'state', and 'number
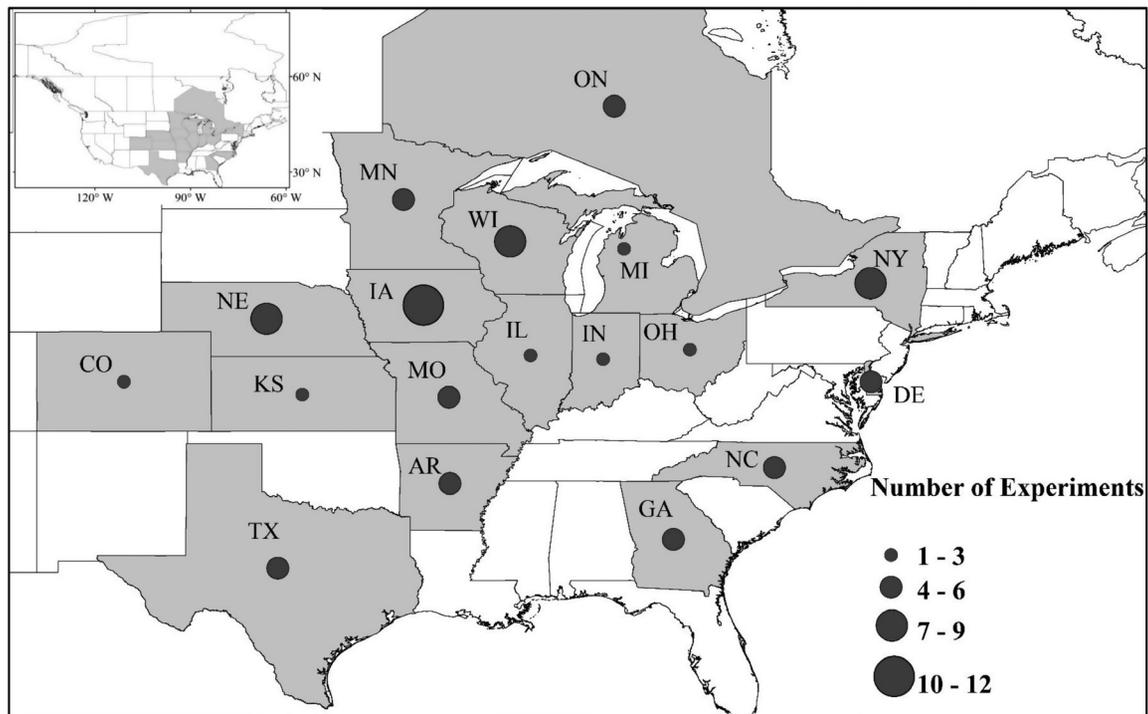
Fig. 1. The spatial distribution of Genomes to Fields (G2F) experiments between 2014 and 2017. The size of the markers represents the accumulative number of experiments in each state.

of hybrid experiment in the field', respectively (for instance, '2014TXH1', '2014TXH2', or '2015ILH1'). For each experimental field, the genotypic, phenotypic, and environmental data are available through the G2F initiative website.

The G2F's 'G×E' project aims to integrate and provide accurate 'omics' and environmental data to boost knowledge on the predictability of maize hybrid traits under diverse environmental conditions. Yet, the main limitation with the G2F database are the several missing values in existing genotypes, phenotypes, and weather time series records, which precludes the use of the associated trials, limiting the simulation process. Therefore, to take the most advantage of the G2F database and involve a larger number of trials in the maize yield simulation procedure, we first need to fill in the missing values. The details of each dataset are explained below.

Environmental data (G2F-E)

The G2F has implemented and collected environmental data in 22 experiments in 2014, 26 experiments in 2015, 25 experiments in 2016, and 25 experiments in 2017 in multiple states (Fig. 1). In each experimental field, several trials have been implemented with different maize hybrid cultivars. During the growing season, eight environmental variables were recorded every 30 min, and these are temperature [T (°C)], dew point [DP (°C)], relative humidity [RH (%)], solar radiation [SR (W m^{-2})], rainfall [R (mm)], wind speed [WS (m s^{-1})], wind direction [WD (°)], and wind gust [WG (m s^{-1})] by a weather station located in the field. As mentioned before, despite the efforts made to record and integrate a comprehensive environmental database, there are several gaps and missing values in the released time series in G2F-E datasets.

Cured environmental databases

For missing data imputation, three other publicly available databases were used: (i) the National Solar Radiation Database (NSRDB), modeling and integrating a half-hourly gridded meteorological dataset in the nation developed by the U.S. Department of Energy (Sengupta *et al.*, 2018); (ii) DayMet, daily surface weather and climatological summaries developed by Thornton *et al.* (2018); and (iii) The Automated Surface Observing Systems (ASOS), developed by the National Weather Service (NWS) which is a station-based program containing daily and sub-daily historical and forecasting hydroclimate data. The variables included in the NSRDB, DayMet,

and NWS databases, along with their spatiotemporal resolution, and sources are listed in Table 1.

Genomic data (G2F-G)

Single nucleotide polymorphism (SNP) sequences, as genetic DNA markers, have been generated by the genotyping-by-sequence (GBS) technique (McFarland *et al.*, 2020). The data are stored in a hierarchical data format file and released by G2F through their portal to represent genomic information of 1576 maize (*Zea mays* L.) lines. The raw data in hierarchical data format have been processed by TASSEL 5 software (TASSEL, 2021) to recognize the genotype in each allele for all the hybrids and convert the genetic codes to numerical genotypes. The numerical genotype refers to the probability of a major allele being selected randomly in a site marker. Genotypes are converted to the probability that an allele selected at random at a site is the major allele; in other words, homozygous major is 1.0, homozygous minor is 0.0, and heterozygous is 0.5 (TASSEL, 2019). For marker quality control, the hybrid lines with >20% missing SNPs have been removed. We also considered only the markers with minor allele frequency >0.03 in the G2F-G. In the updated SNP data, the missing values in each marker have been imputed by the average of the numerical genotypes of the non-missing values in the same site marker.

Phenotype data (G2F-P)

The growing stages of the maize crops were monitored, and various phenotypic categories have been recorded during and at the end of the growing season (at the maturity stage), among which are plant morphology [e.g. plant height (cm)], ear morphology [e.g. ear height (cm), width (cm), and length (cm)], and plant productivity [e.g. grain moisture (%) and yield (bu A^{-1})]. In this study, the target phenotypic variable for simulation and prediction purposes is yield measured in [bushels per acre (bu A^{-1})].

Materials and methods

G2F-E database evaluation and improvement

We designed a pipeline to find data gaps in the G2F-E time series and impute them. This pipeline evaluates G2F-E time series for each experiment to find any data gaps over the growing season. Also, it categorizes

Table 1. Summary of hydroclimatic variables features from NSRDB, DayMet, and NWS databases

Database	Variable (unit)	Spatial resolution	Temporal resolution	Source
NSRDB	T (°C), DP (°C), RH (%), SR (W m^{-2}), WS (m s^{-1}), WD (°), PW (mm), P (mbar)	4×4 km^2	30 min	https://nsrdb.nrel.gov
DayMet	T$_{min}$ (°C), T$_{max}$ (°C), R (mm), SR (W m^{-2}), P (Pa)	1×1 km^2	Daily	https://daymet.ornl.gov/getdata
NWS	T (°C), DP (°C), RH (%), R (mm), WS (m s^{-1}), WD (°), WG (m s^{-1})	ASOS network	Subdaily	https://mesonet.agron.iastate.edu

PW, precipitable water, P, pressure. The other acronyms are defined in the text.

them into 'complete', 'empty', and 'incomplete' experiments, and finally fills the data gaps of the 'empty' and 'incomplete' datasets to circumvent the missing records (Fig. 2). Suppose the time series for the *m*th variable in each G2F experiment is available completely (i.e. 'complete' dataset) over the growing season; in that case, it is directly stored in the final G2F–E database (see Fig. 2). Otherwise, the time series is either 'empty' if the time series for variable *m*th has not been recorded during the entire growing season or 'incomplete' if the time series for variable *m*th has been collected, but there are still some gaps in the recorded time series. These gaps are required to be filled before transfer to the final G2F–E database for further simulation. In summary, the raw G2F–E is the primary input of the pipeline, and the improved fulfilled G2F–E is the ultimate output of that.

To fill the data gaps and enhance the 'empty' and 'incomplete' G2F–E categories, we earlier proposed the application of three other hydroclimate data sources: option (i) NSRDB; option (ii) DayMet; and option (iii) NWS. We first need to find the most consistent option with each G2F–E-*m* time series based on the root mean squared error (RMSE) metric and then use the selected option to replace the empty records in 'empty' experiments. Then, the fulfilled experiment is stored in the final G2F–E database. The same approach is applied to the 'incomplete' experiments to find the best option at each location for a given variable *m*. Then, the selected option and the 'incomplete' G2F time series for the given variable *m* are fed into the DNN model as predictors to simulate the missing samples. Finally, the improved G2F–E-*m* time series is

transferred to the final improved G2F–E database. More details of the applied gap-filling methodology are explained in Sarzaeim *et al.* (2022) at Zenodo (doi:10.5281/zenodo.6299090).

Phenotype modeling

Recently there have been successful developments of statistical phenotype modeling by employing G×E interactions for technological advances in genotyping, phenotyping, and envirotyping (Van Eeuwijk *et al.*, 2016). The G×E concept describes how different genotypes may respond differently to similar environmental changes (Fig. 3). In other words, the phenotypes are not only influenced by genetic information, but they are outcomes of the complex gene and environment interactions (Fig. 3A). Figure 3B symbolizes hypothetical examples of the phenotypic responses from cultivars 1 to *m* exposed to the same environmental changes from environments 1 to *n* without conceptualization of G×E interactions. On the other hand, Fig. 3C indicates the phenotypic responses of cultivars with diverse G×E interactions. The complexity of modeling G×E stems from the high non-linear (see Fig. 3C) and high-dimensional (genotype 1 to *m* and environment 1 to *n*) nature of the G×E interactions. Several statistical modeling efforts have been developed to increase the prediction skill of the crop phenotype models by incorporating high-dimensional genotypes and environmental information to capture G×E interactions (for a review, see Van Eeuwijk *et al.*, 2016). Jarquín *et al.* (2014) showed that incorporating genetic and environmental covariance ma-
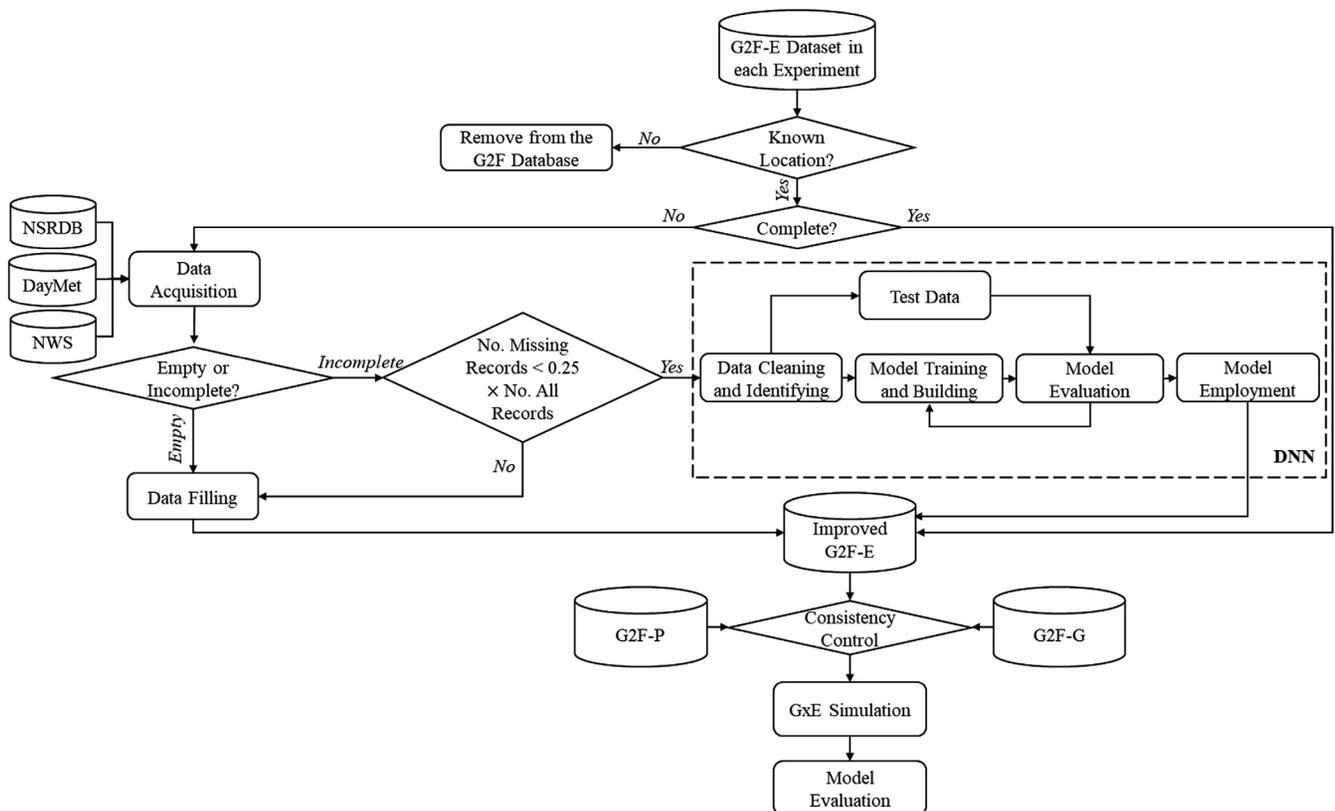

Fig. 2. Flowchart of the methodology for the G2F–E evaluation–improvement pipeline. The pipeline (i) categorizes the G2F experiments, and (ii) fulfills and simulates 'empty' and 'incomplete' G2F–E time series. This methodology has been implemented for each of the 15 G2F–E variables, namely minimum temperature (T_{min}), average temperature (T_{mean}), maximum temperature (T_{max}), minimum dew point (DP_{min}), average dew point (DP_{mean}), maximum dew point (DP_{max}), minimum relative humidity (RH_{min}), average relative humidity (RH_{mean}), maximum relative humidity (RH_{max}), minimum solar radiation (SR_{min}), average solar radiation (SR_{mean}), maximum solar radiation (SR_{max}), accumulative rainfall (R_{acc}), average wind speed (WS_{mean}), and average wind direction (WD_{mean}).
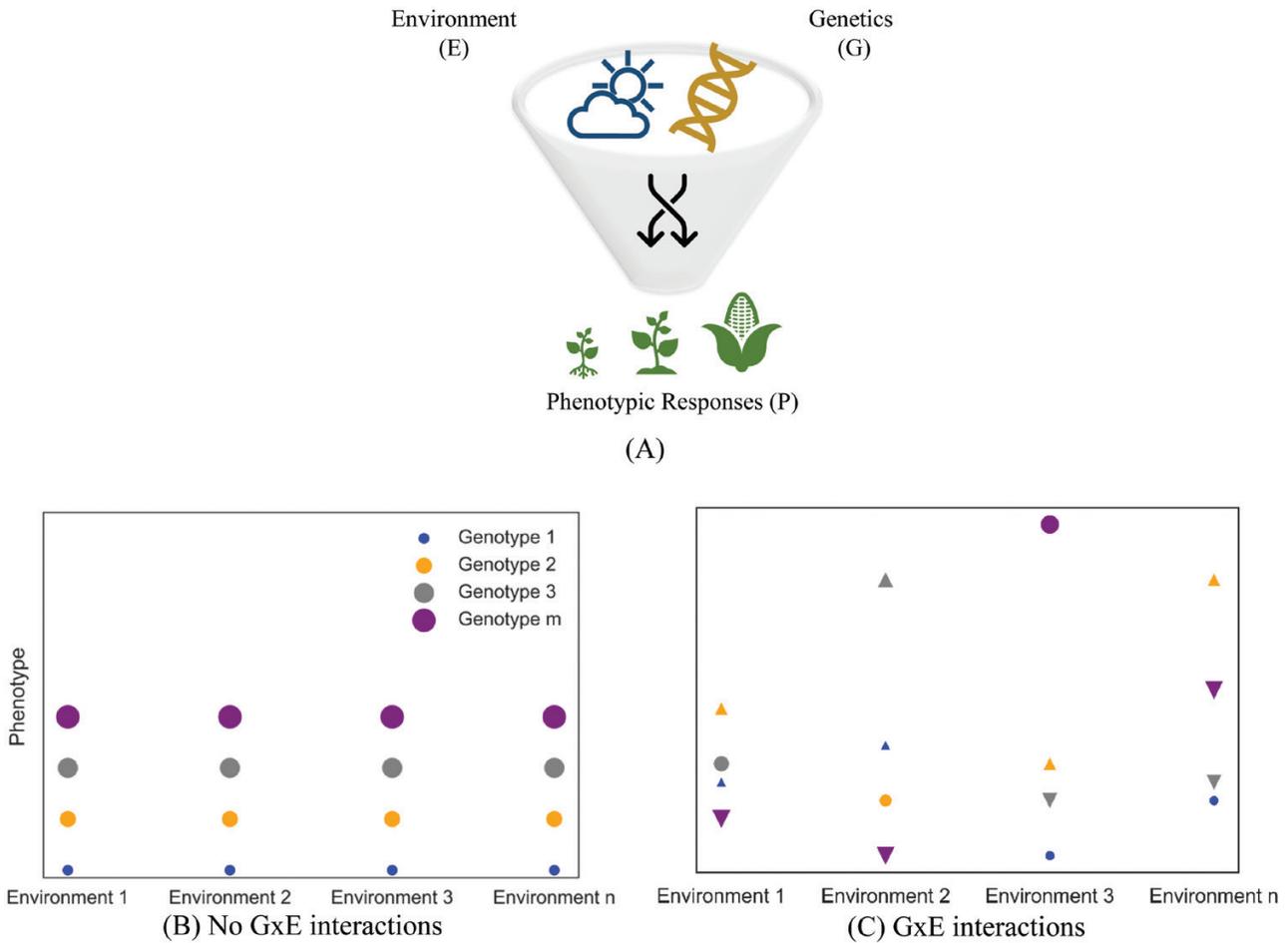
Fig. 3. (A) Conceptualization of G×E interactions, and the visualization of phenotypic responses of (B) genotype 1 to *m* not interacting with environment 1 to *n*, and (C) of genotype 1 to *m* interacting with environment 1 to *n*. Note that the rank of the superior genotype varies because of the G×E interactions across the environments. The upright and inverted triangle symbols in (C) indicate the change in ranks. If there is no change in the genotype's rank, the symbol remains a circle. In (B), genotype *m* is the superior genotype across environments with the lack of G×E interactions. In (C), the rank of genotype *m* varies across environments and remains superior only in environment 3.

trices improved crop performance predictability. The covariance matrices provide the genetic similarity criterion between each pair of genotypes (in a genetic covariance matrix) and the environmental similarity criterion between each pair of environments (in an environmental covariance matrix). Thus, in this study, we use the model developed by Jarquín *et al.* (2014) to predict maize yields using the multienvironment G2F datasets. The following paragraphs present the development process of modeling traits of complex crops, the main and interaction effects, and developed equations in more detail.

As mentioned earlier, crop yield is generally affected by the crop's information, environmental conditions, as well as the complex interactions between gene and environment (Jarquín *et al.*, 2014; Van Eeuwijk *et al.*, 2016; Bustos-Korts *et al.*, 2018). Therefore, the simple baseline equation to model the phenotypes is as in Equation 1:

$$P_{mn} = \mu + E_n + G_m + (G \times E)_{mn} + \epsilon_{mn}$$
$$m = 1, \cdots, M \text{ and } n = 1, \cdots, N$$

(1)

where P_{mn} is the mean of phenotypic response (i.e. yield) of genotype *m* in environment *n*, μ is the overall phenotypic response from all observations, G_m is the main random effect of genome *m*, E_n is the main random

effect of environment *n*, $(G \times E)_{mn}$ is the interaction effect of genome *m* and environment *n*, and ϵ_{mn} is residual random error.

Genotyping technological advances and innovative sequencing methods have enabled the extraction of the high-intensity SNP genetic variation of maize lines and analysis of segregation effects, which enhances the understanding of the genotype–phenotype relationship (Cobb *et al.*, 2013). These advances led to incorporation of high-intensity SNP variations of parent cultivars for maize hybrid *m* in the phenotypic simulation. Therefore, the G_m term in Equation 1 can be broken down into two terms, $G_{P1,m}$ and $G_{P2,m}$. Then the updated equation is:

$$P_{mn} = \mu + E_n + G_{P1,m} + G_{P2,m}$$
$$+ (G_{P1,m} \times G_{P2,m}) + (G_{P1,m} \times E_n)$$
$$+ (G_{P2,m} \times E_n) + (G_{P1,m} \times G_{P2,m} \times E_n) + \epsilon_{mn}$$

(2)

where, $G_{P1,m}$ and $G_{P2,m}$ are the random genomic main effects corresponding to SNPs of parent 1 and parent 2 of genotype *m*, respectively; $G_{P1,m} \times G_{P2,m}$ is the specific combining ability of crossing parent 1 and parent 2 of genome *m*, $G_{P1,m} \times E_n$ is the interaction effect of parent 1 and the *n*th environment, $G_{P2,m} \times E_n$ is the interaction effect of parent 2 and

the *n*th environment, and $G_{P1,m} \times G_{P2,m} \times E_n$ is the interaction effect of the combined SNPs of parents and the *n*th environment.

By introducing covariance matrices, Jarquín *et al*. developed, validated, and recommended Equation 3 to predict a crop's yield:

$$
\begin{aligned}
P_{mn} = {} & \mu + E_n + G_{P1,m} + G_{P2,m} + W_{mn} \\
& + \left(G_{P1,m} \times G_{P2,m} \right) + \left(G_{P1,m} \times E_n \right) + \left(G_{P2,m} \times E_n \right) \\
& + \left(G_{P1,m} \times G_{P2,m} \times E_n \right) + \left(G_{P1,m} \times W \right) \\
& + \left(G_{P2,m} \times W \right) + \left(G_{P1,m} \times G_{P2,m} \times W_n \right) + \epsilon_{mn}
\end{aligned}
\tag{3}
$$

where W is the main effect of the environmental factors modeled by using environmental covariables (ECs), $G_{P1,m} \times W_n$ is the interaction effect of parent 1 and the environmental factors, $G_{P2,m} \times W_n$ is the interaction effect of parent 2 and the environmental factors, $G_{P1,m} \times G_{P2,m} \times W$ is the interaction effect of the combined SNPs of parents and the environmental factors, with the following assumptions:

$$
E \sim N(0, \ \sigma_E^2)
\tag{4}
$$

$$
W \sim N\left(0, \ \cdot \sigma_w^2\right), \quad \text{where } \cdot = \frac{1}{Q} WW'
\tag{5}
$$

$$
G_{P1} \sim N\left(0, \ G\sigma_g^2\right), \quad \text{where } \mathbf{G} = \frac{1}{P_1} XX'
\tag{6}
$$

$$
G_{P2} \sim N\left(0, \ G\sigma_g^2\right), \quad \text{where } G = \frac{1}{P_2} XX'
\tag{7}
$$

$$
\varepsilon \sim N(0, \sigma_\varepsilon^2)
\tag{8}
$$

where Ω is the covariance matrix describing the environmental similarities between pairs of environments using ECs, Q is total number of ECs, W is random regression on ECs, G is the covariance matrix describing the similarities between maize lines using molecular markers, P_1 and P_2 are the total number of maize molecular markers of parent 1 and parent 2 (in our case $P_1 = P_2$), respectively, and X is the genomic value of each maize marker. N (,) denotes a normal distribution.

For model fitting and simulation, the G2F database trials split into training and testing ensembles. The training set is observations used to train the G×E model, while the testing set is unobserved samples used to test the accuracy of the constructed model in the training step. The procedure for the G×E train and test for the simulation of yields uses the BGLR (Bayesian Generalized Linear Regression) R-package. BGLR has analyzed highly dimensional data where a predictand (here; maize yield) needs to be regressed on a large number of predictors (here; genotypes, environments, and their interactions) (Pérez and Campos, 2014).

G×E train–test set selection design

The accuracy of the statistical model's performance relies on training datasets (Lobell and Burke, 2010; Gianola, 2021). Thus, we have designed three different trial selection schemes to evaluate the change rate of G×E predictive performance for the train–test set size in each scheme. Figure 4 provides a step-by-step flowchart of these three approaches of selection of train–test experiments.

The first approach is 'random-based' experiment selection, through which the model is implemented in several iterations. In each iteration, the number of the test set is increased by five random environments. The rest of the trials become the training set. This procedure will continue until all experiments have been covered in the test dataset, and at least one experiment remains in the training set. After each iteration, the G×E model is implemented, and its efficiency is evaluated.

The following strategy for trial selection is built upon the ranked environmental similarities among the trials. As described earlier in Equation 3, the environmental covariance values computed based on DNN-enhanced weather covariates between each pair of the experiments are considered as environmental similarities criteria. The covariance is calculated as shown in Equation 9:

$$
\text{cov} \quad \left(EC_x, EC_\gamma \right) = \frac{\sum \left(EC_{x,t} - \overline{EC}_x \right) \left(EC_{\gamma,t} - \overline{EC}_\gamma \right)}{T}
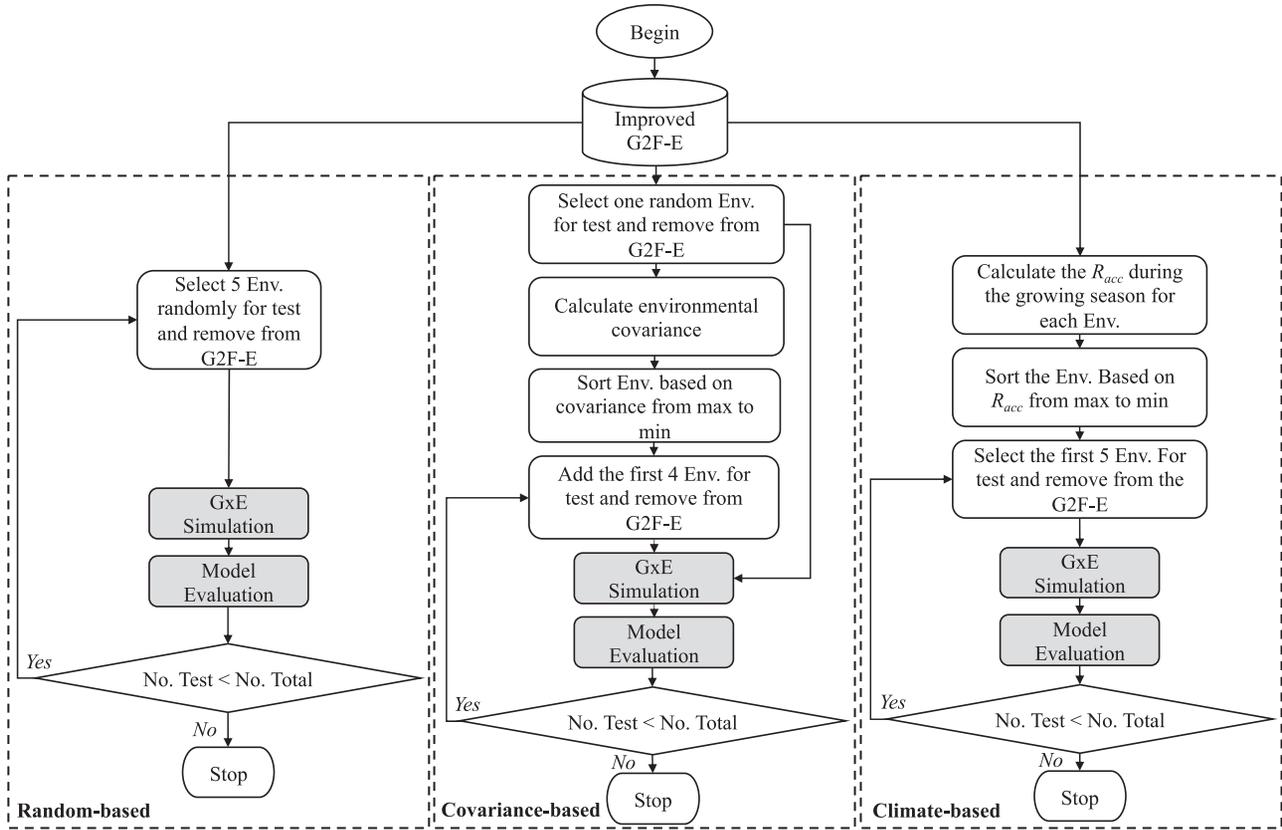\tag{9}
$$

where, $\text{cov}(EC_x, EC_\gamma)$ is the environmental covariance value of EC time series between experiment x and γ, $EC_{x,t}$ is the standardized values of ECs in experiment x on day t, $EC_{\gamma,t}$ is the standardized values of ECs in experiment γ on day t, $\overline{EC}_x$ is the average of standardized ECs in experiment x, $\overline{EC}_\gamma$ is the average of standardized ECs in experiment γ, and T is the total number of time steps in day.

For the 'covariance-based' trial selection scheme, first we select one random trial and then calculate the environmental covariance between the chosen trial and the other remaining experiments: the more covariance values, the stronger the environmental relationship (i.e. similarities). In the next step, experiments are sorted by the calculated environmental covariance values by size. In the first model-run iteration, the first four trials with the largest covariances (i.e. the first four most similar trials with the selected experiment) are chosen and assigned as the test set along with the first randomly selected experiment (i.e. five experiments), and the remaining experiments are kept as the training set. In the next iteration, the following five similar experiments are added on to the test set, and the remaining experiments are allocated to the training set. This process will continue until it covers all experiments as a test set and at least one in the training set.

The last train–test selection scheme focuses on the selection trials based on one single climate gradient in each experiment. Here we chose accumulative rainfall (R_{acc}) during the growing season at each G2F trial location. In this approach, the experiments will first be sorted in decreasing order based on the growing season R_{acc} values. Then, in each training G×E iteration, five sorted experiments with larger R_{acc} values will be added to the test set. The remaining experiments will be considered as the training set. In other words, the size of the test and train samples will increase and decrease by five experiments, respectively, based on R_{acc} values in the arranged experiments until they cover all experiments as the test set and at least one in the training set. This approach aims to identify the effect of a single climate driver such as R_{acc} on G×E performance and the predictability of yields.

To evaluate the G×E model simulation on the test set in each iteration, four performance metrics will be discussed, which are coefficient of determination (R^2), RMSE, mean squared error (MSE), and mean absolute error (MAE), each of which is formulated as below (Willmott, 1982; Chai and Draxler, 2014):

$$
R^2_{\,G \times E, n} = \frac{\sum_{l=1}^{L} \left(\gamma_{obs_{n}, \ m} - \gamma_{\text{sim}_{n}, \ m} \right)^2}{\sum_{l=1}^{L} \left(\gamma_{obs \ n, \ m} - \hat{\gamma}_{n,m} \right)^2}
\tag{10}
$$

Fig. 4. Trial selection design for 'random-based', 'covariance-based', and 'climate-based' approaches. Note: the total number of experiments used for G×E simulation is 84, which is explained in the text. Env., environments; No., number.

$$\mathrm{RMSE}_{\mathrm{G\times E},n} = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\gamma_{\mathrm{obs}n,m} - \hat{\gamma}_{\mathrm{sim}\ n,m}\right)^2} \tag{11}$$

$$\mathrm{MSE}_{G\times E,n} = \frac{1}{L}\sum_{l=1}^{L}\left(\gamma_{\mathrm{obs}n,m} - \hat{\gamma}_{sim\ n.m}\right)^2 \tag{12}$$

$$\mathrm{MAE}_{\mathrm{G\times E},n} = \frac{1}{L}\sum_{l=1}^{L}\left|\gamma_{\mathrm{obs}n,m} - \hat{\gamma}_{\mathrm{sim}\ n,m}\right| \tag{13}$$

where, $R^2_{\mathrm{G\times E},n}$, $\mathrm{RMSE}_{\mathrm{G\times E},n}$, $\mathrm{MSE}_{\mathrm{G\times E},n}$, and $\mathrm{MAE}_{\mathrm{G\times E},n}$ are calculated G×E model R^2, RMSE, MSE, and MAE for environment n, respectively; $\gamma_{\mathrm{obs}_{n,m}}$ and $\hat{\gamma}_{\mathrm{sim}_{n,m}}$ is observed and simulated yield values for recorded individual genotype m in environment n, respectively; and L is total number of recorded m in environment n.

Results and discussion

G2F-E improvement

Among 112 G2F hybrid experiments in 2014–2017, the location information (e.g. latitude and longitude) of 15 experi-mental fields is missing. In addition, yield values of 11 trials have not been recorded. The G2F–E datasets from the remaining experiments (i.e. 86) are integrated into the evaluation–improvement sequence. Figure 5 illustrates the proportion of each category for each G2F-E variable. Among the environmental variables, G2F-E-RH is the most complete, with 80% completeness (i.e. 69 experiments), while there are several missing records in the G2F–E-SR published time series, with 49% completeness (i.e. 42 experiments).

The constructed evaluation–improvement pipeline shown in Fig. 2 categorizes and improves the environmental datasets, enabling the advantage of an enhanced three-dimensional database consisting of improved environmental data and genotypes and phenotypes from a more significant number of G2F experiments for yield prediction. The pipeline found that 32 out of 86 G2F experiments contain all the climate variables completely assigned to the 'complete' category, meaning that 480 datasets (=32 experiments×15 variables) are transferred to the final improved G2F-E database directly. The variables included are minimum temperature (T_{min}), average temperature (T_{mean}), maximum temperature (T_{max}), minimum dew point (DP_{min}), average dew point (DP_{mean}), maximum dew point (DP_{max}), minimum relative humidity (RH_{min}), average relative humidity (RH_{mean}), maximum relative humidity (RH_{max}), minimum solar radiation (SR_{min}), average solar radiation (SR_{mean}),
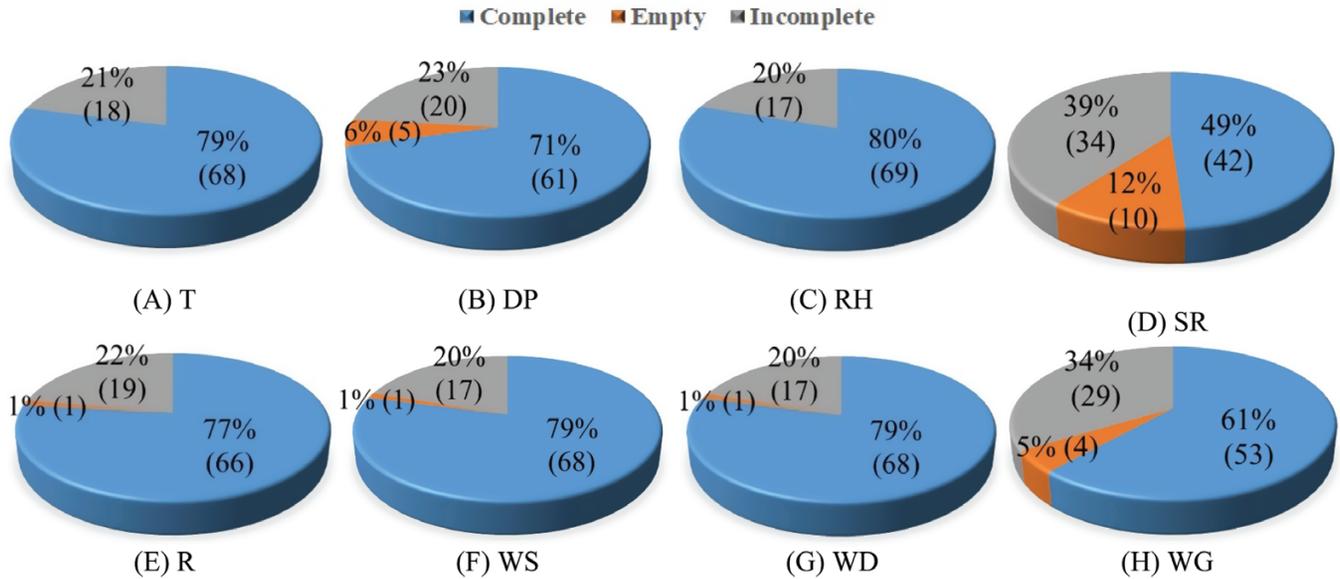
Fig. 5. The rounded portion of completeness, incompleteness, and emptiness in G2F-E. The numbers in parentheses indicate the absolute number of datasets in each category for each variable: temperature (T), dew point (DP), relative humidity (RH), solar radiation (SR), rainfall (R), wind speed (WS), wind direction (WD), and wind gust (WG).

maximum solar radiation (SR_{max}), accumulative rainfall (R_{acc}), average wind speed (WS_{mean}), and average wind direction (WD_{mean}). Since the NSRDB, DayMet, and NWS do not provide wind gust data, the improvement of the G2F-E-WG datasets was not performed and was not considered for further processing. In some cases, imperfect wind gust values are reported in the NWS database. However, they cannot be used in this study due to several missing values in NWS. Among the remaining G2F datasets associated with other experiments (86–32=54), 442 more 'complete' datasets are directly transferred to the final database. The DNN has used additional 181 'incomplete' datasets for data gaps simulation (see the Materials and methods and Sarzaeim *et al.*, 2022), and the subsequent datasets are 'empty' which have been filled as described earlier (Sarzaeim *et al.*, 2022).

To complete the 'empty' experiments, the RMSE metric has been calculated (Sarzaeim *et al.*, 2022) between each of the observed non-empty time series (including 'complete' and 'incomplete') and the associated time series of available sources for variable *m*. For illustration, since DP values are accessible from NSRDB and NWS sources, the RMSE has been calculated between observed G2F-E-DP and each of NSRDB-DP and NWS-DP measurements at each G2F trial location. Then, the kernel density of the calculated pairwise RMSE values is plotted to recognize the best-fitted source to the G2F-E-DP measurements (see Fig. 6). The lower RMSE values (i.e. the *x*-axis) with higher probability (i.e. the *y*-axis), the better the source is fitted to the observed G2F–E-DP. By selecting the option with a greater probability of smaller deviation from G2F-E, we minimize the error introduced to the enhanced G2F-E database by other source replacements. Therefore, the

source with more probable, lower RMSE values is selected to replace the data gaps in 'empty' G2F-E-DP. The same procedure has been applied for other variables as well. Figure 6 illustrates the kernel density for the calculated RMSE between G2F-E and each NSRDB, DayMet, and NWS option for given climatic variable *m*. Based on the probability of RMSE values, the NSRDB was selected for gap fulfillment in 'empty' DP, SR, WS, and WD datasets. In contrast, DayMet was only selected to fulfill the R dataset, and NWS was not selected for any variable. This selection process based on the RMSE criterion indicates that observed G2F-E datasets are more consistent with the NSRDB database in general.

From the RMSE probability plots in Fig. 6, it is also worth mentioning that although G2F weather station devices located at experimental fields are initially calibrated with the nearest ASOS-NWS stations by G2F collaborators (McFarland *et al.*, 2020), and low RMSE values are expected between G2F and NWS time series; the NWS databases (station-based data) have remarkable deviations from the observed G2F-E time series compared with NSRDB and DayMet (gridded data). In other words, the gridded NSRDB and DayMet databases are in better agreement with G2F-E in terms of RMSE values (Fig. 6). It is noteworthy that G2F is a station-based database. Lowering the deviation between NWS and G2F records through regular calibration is required due to the sensitivity of the devices for accurate long-term measurements (Bojanowski *et al.*, 2014).

For 'incomplete' time series improvement, we implemented DNNs for 58 SR, 30 T, 30 DP, 30 RH, 13 WS, 11 R, and 10 WD datasets. Table 2 shows the DNN's performance based on the RMSE metric for the test DNN population in the associated G2F experiment. The average, minimum, maximum, and
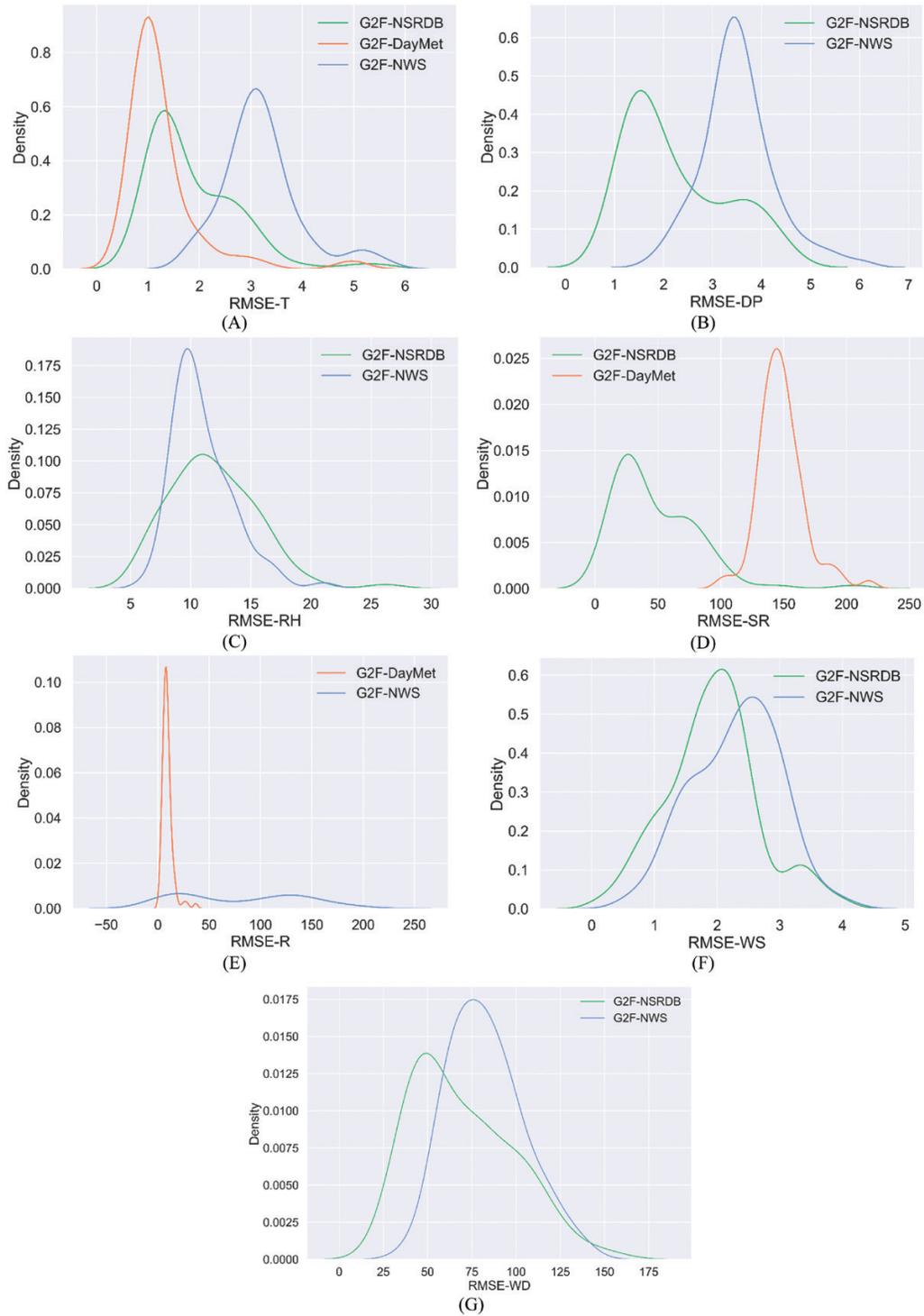
Fig. 6. Kernel density of pairwise calculated RMSE values between observed G2F-E-*m* and NSRDB-*m*, DayMet-*m*, and NWS-*m* for variable *m*. Note: temperature (T) and relative humidity (RH) presented no data gaps, but the RMSE values are presented together with those for dew point (DP), solar radiation (SR), rainfall (R), wind speed (WS), wind direction (WD), and wind gust (WG).

SD of the RMSE performance metric have also been calculated and are presented in Table 2 for the given variable *m*. The calculated statistics shown in Table 2 indicate the low RMSE values in the DNN test population. The constrained diver-

gence indicates a high capability of DNN for the imputation of G2F-E missing records, particularly in T_{max}, DP_{mean}, RH_{mean}, and SR_{mean} time series regarding their associated average performance metric values. The percentage improvement in

Table 2. RMSE values for the performance of DNNs on the test dataset in the 'incomplete' time series for each variable

G2F incomplete experiment	T_{mean} (°C)	T_{min} (°C)	T_{max} (°C)	DP_{mean} (°C)	DP_{min} (°C)	DP_{max} (°C)	RH_{mean} (%)	RH_{min} (%)	RH_{max} (%)	SR_{mean} (Wm⁻²)	SR_{max} (W m⁻²)	R_{acc} (mm)	WS_{mean} (m s⁻¹)	WD_{mean} (°)
2014_NEH3	0.381	0.252	2.5e-5	0.304	1.880	1.899	0.015	0.051	0.308	–	–	9.668	0.005	26.725
2015_KSH1	0.007	0.349	0.132	0.128	2.637	0.022	0.002	0.241	78.210	0.2e-4	0.3e-4	5.657	0.530	0.026
2015_NYH3	1.1e-5	5.7e-6	2.1e-5	5.9e-5	2.365	0.079	1.740	5.401	4.8e-5	–	–	4.004	0.002	23.290
2016_ARH1	1.8e-5	1.2e-5	1.89e-5	0.007	2.085	0.437	2.111	5.155	0.954	–	–	2.913	0.465	0.2e-4
2016_DEH1	0.001	1.2e-5	0.006	1.9e-5	3.554	0.004	0.027	0.001	0.954	–	–	2.040	0.059	0.007
2016_NCH1	1.682	1.951	0.022	0.626	0.784	0.034	0.008	3.3e-6	0.205	42.197	40.580	17.974	0.4e-4	0.007
2017_ARH1	1.483	1.464	1.178	0.003	0.003	0.006	0.001	0.005	0.284	–	–	11.160	0.563	7.3e-5
2017 IAH3	1.211	1.1e-5	9.9e-6	7.9e-6	0.012	1.4e-5	1.203	1.3e-5	0.291	28.459	39.287	0.814	0.002	27.514
2017_NYH1	1.629	0.003	2.4e-5	0.180	0.069	0.212	5.4e-5	1.8e-5	3.9e-5	19.218	36.574	1.175	0.048	22.378
2017_NYH2	2.13	0.37e-4	1.6e-5	0.033	0.014	0.006	1.216	3.413	2.9e-5	0.001	0.5e-4	23.831	1.5e-7	0.1e-4
2014_IAH1	–	–	–	–	–	–	–	–	–	0.2e-4	3.3e-4	–	–	–
2016_ILH1	–	–	–	–	–	–	–	–	–	48.694	4.1e-4	–	–	–
2016_KSH1	–	–	–	–	–	–	–	–	–	0.2e-4	5.4e-4	–	–	–
2016_MIH1	–	–	–	–	–	–	–	–	–	0.1e-4	98.834	–	–	–
2016_NYH2	–	–	–	–	–	–	–	–	–	1.5e-4	5.0e-4	–	–	–
2016_OHH1	–	–	–	–	–	–	–	–	–	1.3e-4	4.0e-4	–	–	–
2016_WIH2	–	–	–	–	–	–	–	–	–	0.002	4.8e-4	–	–	–
2017_IAH1	–	–	–	–	–	–	–	–	–	9.2e-5	22.627	–	–	–
2017_IAH2	–	–	–	–	–	–	–	–	–	1.0e-4	70.555	–	–	–
2017_MNH1	–	–	–	–	–	–	–	–	–	0.013	0.56	–	–	–
2017_NCH1	–	–	–	–	–	–	–	–	–	5.3e-5	35.976	–	–	–
2017_OHH1	–	–	–	–	–	–	–	–	–	2.9e-4	5.0e-4	–	–	–
2017_WIH1	–	–	–	–	–	–	–	–	–	0.0003	6.0e-4	–	–	–
2017_WIH2	–	–	–	–	–	–	–	–	–	30.540	0.001	–	–	–
2014_IAH4	–	–	–	–	–	–	–	–	–	–	–	5.242	–	–
2014_NEH2	–	–	–	–	–	–	–	–	–	–	–	–	0.002	–
2015_DEH1	–	–	–	–	–	–	–	–	–	–	–	–	0.009	–
2017_ARH2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Average	0.853	0.402	0.133	0.128	1.340	0.270	0.632	1.426	8.025	8.901	18.131	7.680	0.173	9.995
Minimum	1.14e-5	5.7e-6	9.9e-6	7.9e-6	0.003	1.4e-5	5.4e-5	3.3e-5	2.9e-5	5.3e-5	3.3e-4	0.814	1.5e-7	7.3e-5
Maximum	2.13	1.951	1.178	0.626	3.554	1.899	2.111	5.401	78.210	48.694	98.834	23.831	0.568	27.514
SD	0.855	0.708	0.369	0.202	1.321	0.588	0.844	2.287	24.662	16.258	28.707	7.418	0.250	12.976

increasing the number of complete trials before and after the improvement process is 26% for T, 41% for DP, 25% for RH, 105% for SR, 30% for R, 26% for WS, and 26% for WD. The number of complete observations after the improvement is 86 for each variable, and the associated numbers of complete experiments before the improvement are listed in Fig. 5.

By applying the G2F-E evaluation–improvement methodology described above (and in Sarzaeim *et al.*, 2022), the number of total 'complete' G2F-E experiments increased from 32 to 86 experiments. This means that we increased the environmental degrees of freedom (DF) by 54 (DF_{after}–DF_{before}=54). Now it is possible to feed a further 54 experimental trials and their datasets, including genomic, phenotypic, and environmental information, into the G×E model simulation process. In other words, the environment DF improvement is ~169%. Increasing the number of experiments by gap filling of 'empty' and 'incomplete' environmental time series provides the opportunity of using not only the enhanced G2F-E but also G2F-G and G2F-P associated with those time series in the simulation process, which were not capable of being processed in the simulation previously. Before the improvement process, we had 32 environments, 372 genotypes, and 3169 phenotypic observations, which could be fed into the simulation. After improving the dataset, we increased the number of environments, genotypes, and phenotypic observations up to 86, 376, and 8171, respectively. Below, we will evaluate the effect of this significant improvement in the number of total experiments in a multidimensional database reflected on G×E predictability performance.

### G×E simulation and predictability evaluation

The enhanced DNN-G2F-E database (consisting of 86 experiments) along with G2F-G molecular markers and G2F-P yield measurements are consistently controlled in terms of data availability in all three types of datasets for each trial before feeding into the G×E model and simulation process (see Fig. 2). After this consistency control, we have 84 experiments capable of being employed as G×E's model input. The criterion is to keep those experiments in the simulation process that all required that G2F-E, G2F-G, and G2F-P data are available in the database. In other words, there are two experiments that either their associated G2F-G or G2F-P information for all tested cultivars are not recorded in the G2F database.

The G×E model has been simulated by feeding 32 complete experiments from the not improved (N.I.) database and 84 DNN-improved (I.) environments. For model yield predictability evaluation, four model performance metrics, namely R^2, RMSE, MSE, and MAE, have been calculated for each environment. The performance of the results of 32 common 'complete' experiments' are compared in Fig. 7 from N.I. and I. G×E implementations. The results generally show more accurate G×E predictability from I. in comparison with N.I. (see

the dashed horizontal lines). The mean R^2 in all 32 common experiments is improved ~12.1%. The RMSE, MSE, and MAE measurements are also improved by 2.2, 11.4, and 1.4%, respectively (Table 3). The improvement in the predictability of yields proves the hypothesis that G×E predictability increases with the DF of the environment data (in this case, from DF_{before}=32-1 to DF_{after}=84-1). While the efficiency metrics in Table 3 are spatiotemporal aggregates, Figs 7–10 illustrate how, for example, R^2 varies across the study area. The improvement in the G×E model performance has been reported by Jarquín *et al.* (2014), who showed that the average model R^2 increased 12.9% when the interaction genotype by environment is included. Acosta-Pech *et al.* (2017), Crossa *et al.* (2017), and Lopez-Cruz *et al.* (2015) have reported improvements in R^2 between 5% and 29% for maize yield prediction. These findings support the thesis that enhancing climate data integrated into the environmental covariance matrices increases the predictive skill of G×E models. Further, a non–parametric *t*-test is applied to evaluate the metrics of the model's performance between N.I. and I. The associated null hypothesis expresses that there is no significant difference between performance metrics of N.I. and I. implementations. The operation of the *t*-test evaluation uses the 'stat.ttest_rel' function from Python, setting a default *P*-value of 0.05. For an estimated significance level larger than the *P*-value, the null hypothesis is accepted. In other words, it can be concluded that there is no significant difference between N.I. and I. performance metrics. The calculated *t*-value and *P*-value for each performance metric are illustrated in Fig. 7. Since the calculated *P*-value is larger than the significance level (0.05) in all metrics, it is concluded that the enhancement is not statistically significant, despite predictability improvement in I. compared with N.I. scenarios. The possible reason for this insignificant predictability improvement may rely on the unbalanced G2F experimental design. This experimental unbalancing refers to the ununiform distribution of experimental sample units (observations) among the trials. The number of observations varies between five (in 2014GAH1, 2014NCH1, 2014TXH1, and 2014TXH2 trials) and 201 (in the 2014ONH2 trial). Thus, the model fitting is more affected by the trials with a larger number of observations.

The geospatial distribution of the I. scenario predictability is represented by the R^2 metric in Fig. 8. The map shows that the G×E predictive skill is more accurate (represented by point size) in the northern part of the G2F layout, where, in most cases, the number of observations (represented by point color) is large.

In summary, the improvement in the environmental dataset provides the opportunity to increase the number of experimental sites and their associated 'omics' information consideration for simulation procedure. Although this three-dimensional data improvement led to a marginal enhancement in G×E predictability, it revealed that the database improvement is reflected by more accurate G×E yield prediction skills in terms of all R^2, RMSE, MSE, and MAE
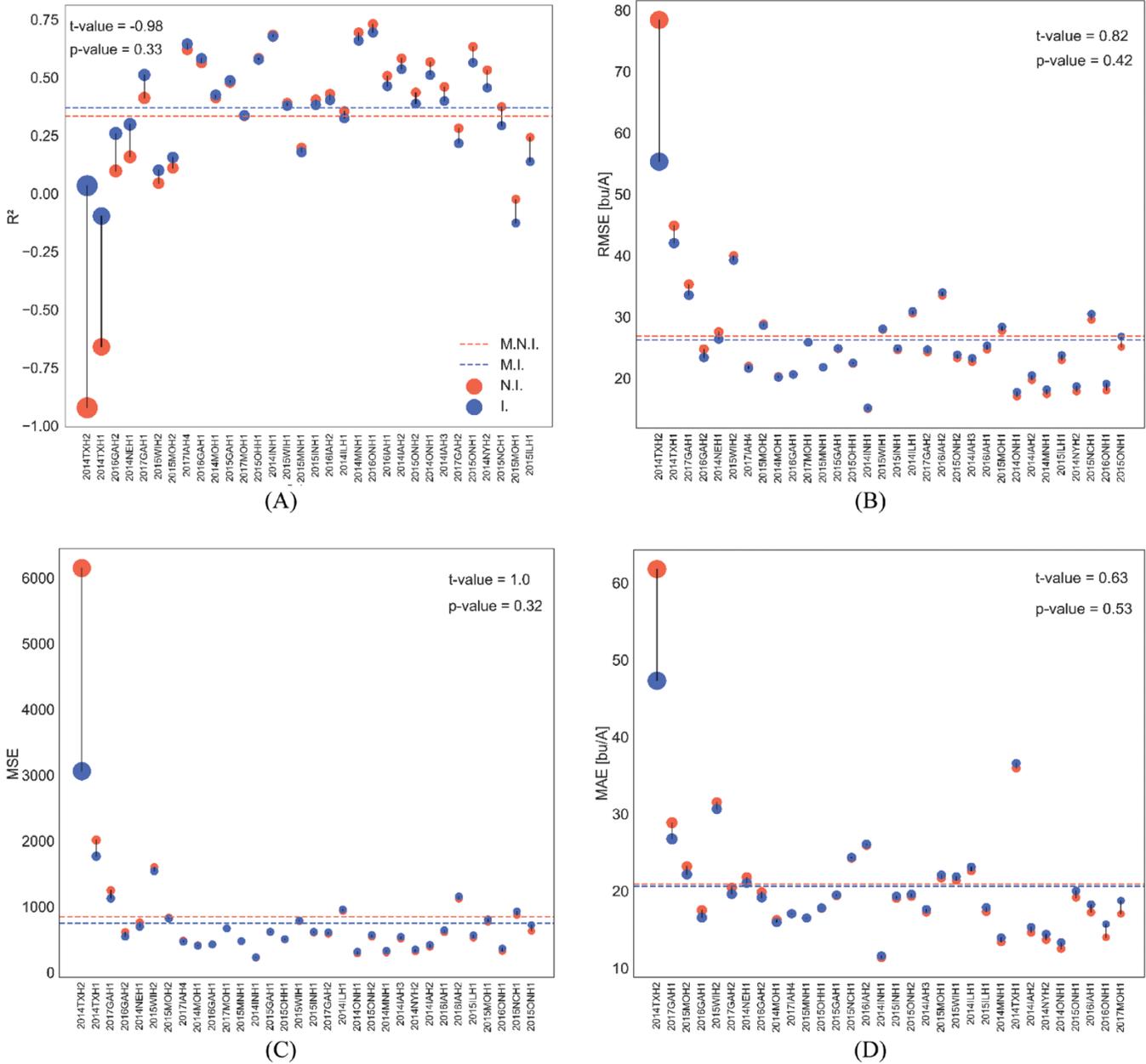
Fig. 7. G×E-based predictability comparing the initial 32 (N.I.=not improved) experiments and the completed 84 (I.=improved) experiments after filling all data gaps. Both N.I. and I. databases have been evaluated based on the efficiency indices: (A) R^2, (B) RMSE, (C) MSE, and (D) MAE metrics. The vertical black lines show the difference between the efficiency indices before and after the improved data. The size of the markers is directly proportional to the improvement level for R^2, and indirectly proportional to the other metrics. The *t*-value and *P*-value are based on the *t*-test statistic and significance level, respectively. Note: M.N.I.=mean of not improved (dashed orange line), and M.I.=mean of improved (dashed blue line).

performance metrics. This result suggests the G×E can predict the phenotypes more accurately by expanding the number of observed sites. This result is also in line with previous findings regarding the increase in 'omics' sample size, which led to better model performance (Auinger *et al.*, 2021; Lopez-Cruz and de los Campos, 2021). The data enhancement enables the robustness of the G×E model to capture the gene by environment interactions more effectively.

G×E predictability in selection of train–test trial scenarios

We provide evidence of the contribution of the DNN-driven enhancement of climate data to the improvement of 'omics' databases and the predictability of maize yields. Three trial selection designs, called 'random-based', 'covariance-based', and 'climate-based', are implemented to quantify and identify the

Table 3. The mean G×E model performance metrics in the N.I. and I. datasets, the minimum and maximum difference of G×E model performance metrics between N.I. and I. datasets, and the mean improvement percentage of the G×E model performance metrics

Performance metric	Mean in N.I.	Mean in I.	Minimum difference	Maximum difference	Mean percentage change (%)
R^2	0.33	0.37	−0.10	0.95	12.1
RMSE	26.86	26.25	−23.09	1.75	2.2
MSE	848.72	751.58	−3089.18	91.16	11.4
MAE	20.93	20.63	−14.49	1.73	1.4

Note that the negative values for RMSE, MSE, and MAE show the improvement of the G×E model performance.
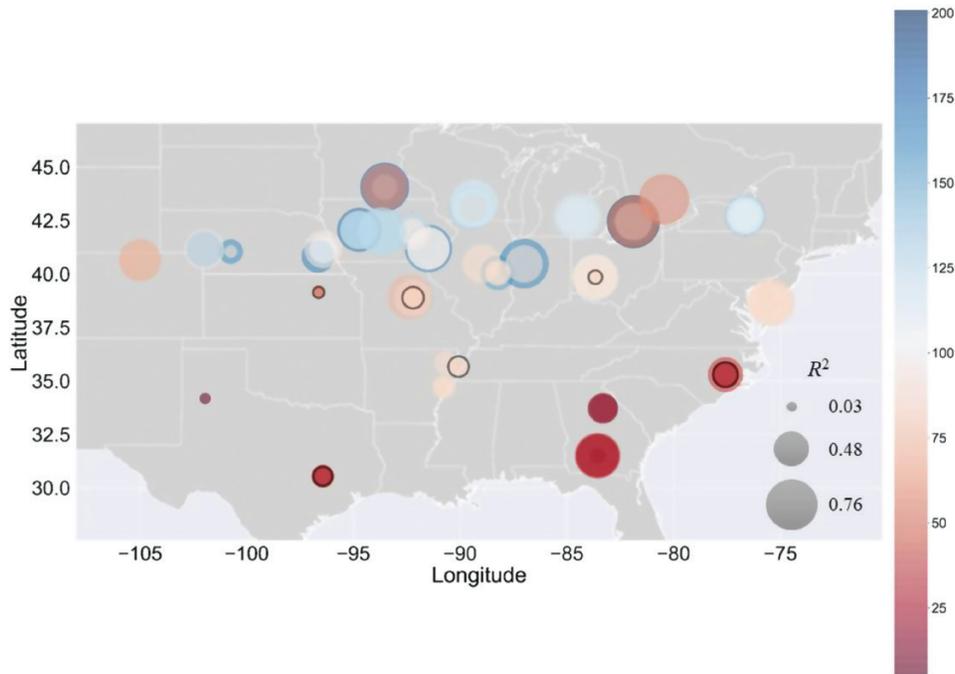
Fig. 8. Spatial distribution of G×E model predictability (R^2) based on 84 DNN-based improvements in the G2F-E experiments. The color gradient represents the number of observations, and the size of the circles represents the R^2 at each trial. The colored circles are transparent to distinguish multiple trials at the same location. Note: the circles with a black perimeter represent negative correlation values. No perimeter means positive correlation.

factors that control the increase in phenotype predictability. Varying iterations formed the scenarios that emerged from a dominant test scenario to a prevailing training scenario. The scenarios gradually increase by five test sets (where the number of training set observations decreases by the same number) (see Fig. 4).

In the first iteration (iteration 1), five selective trials were allocated for testing, and the rest (84−5=79) remained in the training set. The largest training set is in iteration 1, and the smallest training set is associated with the last iteration (iteration 16); the train and test set size is 4 and 80, respectively.

The average R^2 estimations for the tested experiments (the black line) in each iteration were calculated and illustrated in Fig. 9 for each selection scenario. Additionally, the average predictability of newly selected sets in an iteration through the last iteration is represented by the colored lines. The longest line is representative of the average of the first five selected test experiments remaining in the test set from

iteration 1 through iteration 16. The next longest set is associated with the five newly selected test experiments in iteration 2 through iteration 16. Note that in iteration 2, there are 10 test experiments in total: five from iteration 1 and five from iteration 2. A similar explanation applies to the other lines in the plot. The model performance estimated by R^2 shows improvement from iteration 16 to 1 in all test sample sizes in all selection scenarios. This result suggests that the model predictability improves when more experiments are allocated to the training sets, from the smallest training set size in iteration 16 (i.e. $Size_{train}=4$) to the largest training set size in iteration 1 (i.e. $Size_{train}=79$). Thus, a larger number of training samples leads to higher environmental variability in the simulations, leading to the improvement of the model's predictability. These observations are also in line with general results from above; a larger number of observed environment sets can provide a larger training population size and, eventually, G×E predictive skill enhancement.
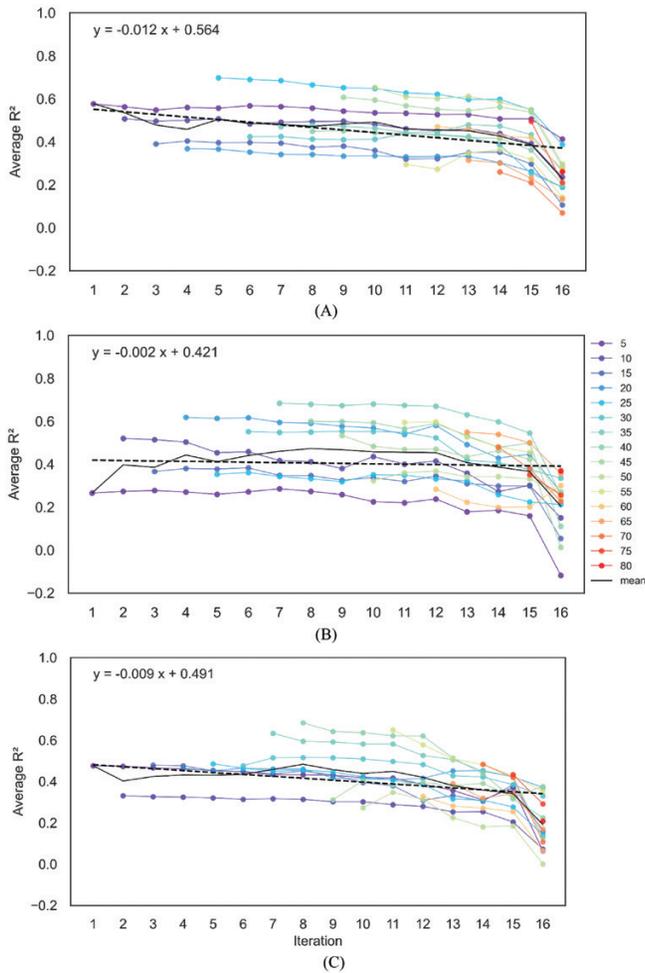
Fig. 9. The number of iterations (1–16) versus the averages of the coefficient of determination, R^2, for test ensembles from 5 (in purple) to 80 (in red). The three sampling schemes were (A) 'random-based', (B) 'covariance-based', and (C) 'climate-based.' The number of iterations is explained as follows: if the number of samples equals 5, five experiments belong to the test ensemble, and the rest belong to the training set. The dashed line results from applying a linear regression model fitted to the total average values.

In addition to the effect of changing training sets, the variability inherent in the environmental covariance structures in the training sets can be affected by the train and test selection strategy (i.e. random versus ranking versus the precipitation gradient). For generally comparing the performance of selection scenarios in improvement of predictability, the linear model follows the averaged R^2 values from iteration 1 to iteration 16. The largest absolute values of the slope (*m*) indicate the largest improvement rate over the iterations. In other words, over Size$_{train}$, the largest improvement is reached in 'random-based' (*m*=0.012), 'climate-based' (*m*=0.009), and 'covariance-based' (*m*=0.002) scenarios, respectively. The reason behind that stems from the different levels of captured environmental variability in training ensembles. In both 'climate-based' and 'covariance-based' scenarios, the selection strategy removes the most similar

experiments from the training set and sums them to the test population. These strategies are built upon experiments with similar ranked R$_{acc}$ values in 'climate-based' and similar ranked environmental covariance values in 'covariance-based' designs. Also, in both approaches, the environmental variation among the retained experiments in the training set decreases gradually in response to possible climate patterns embedded in the complex structure of the environmental covariance matrices. The climate information is aggregated, and no causality between climate and model performance is evident. These results agree with previous studies by Messina *et al*. (2018). These authors attributed the poor crop yield model performance to the low environmental variation, which concurs with that proposed here in the second objective. In another study, Rogers and Holland (2022) showed that decreasing environmental similarity between train and test sets leads to a decrease in the G×E model predictability. Thus, the difference in the G×E predictive performance as evidenced by the 'random-based', 'covariance-based', and 'climate-based' scenarios can be attributed to the skill gained by the training set; also found by Gianola (2021).

The 'random-based' and the 'covariance-based' scenarios indicate improved predictability due to the consolidation of the climate and 'omics' databases. However, these improvements cannot build causality or create climate-driven phenotypic diagnostics or prognostics, such as geospatial or temporal attributions of maize phenotypes to climate variability and change. The 'climate-based' scenario introduces climate variability by selecting environmental covariance structures by a gradient of precipitation. This simple construction addresses the randomness of the 'random-based', the incremental sequence of the 'covariance-based' scenarios, and a gradual increase in rainfall, the main driver of the hydrologic cycle and an element of both the weather and climate systems. Through this scenario design, we suggest that areas and times with high rainfall influence the phenotypic prediction. Rojas *et al*. (2019) predicted that declining rainfall in the coming decades will lead to 1–14% variability of global cereal production. Maltais-Landry and Lobell (2012) associate a high rainfall variability with an increase in the inter-annual variability of yields. The Intergovernmental Panel on Climate Change (IPCC, 2001) has suggested the scenario of increasing variability in rainfall for >25 years. However, it is unclear if crops will have the ability to adapt at a higher pace.

The 'climate-based' scenario illustrates the opportunity to quantify the improvements in predictability as larger variations in rainfall occur. Andresen *et al*. (2001) identified the rainfall and its variability as the major driver of maize inter-annual yield variability. The results in Fig. 9C illustrate a relatively consistent improving pattern from Size$_{test}$=10 to Size$_{test}$=40, which implies that the prediction accuracy of G×E is higher in experiments with larger R$_{acc}$ during the growing season. The consistent improvement in the predictability is only observed in the 'climate-based' scenario. The R$_{acc}$ records ranked decreasingly are presented in Fig. 10. The vertical dashed line separates
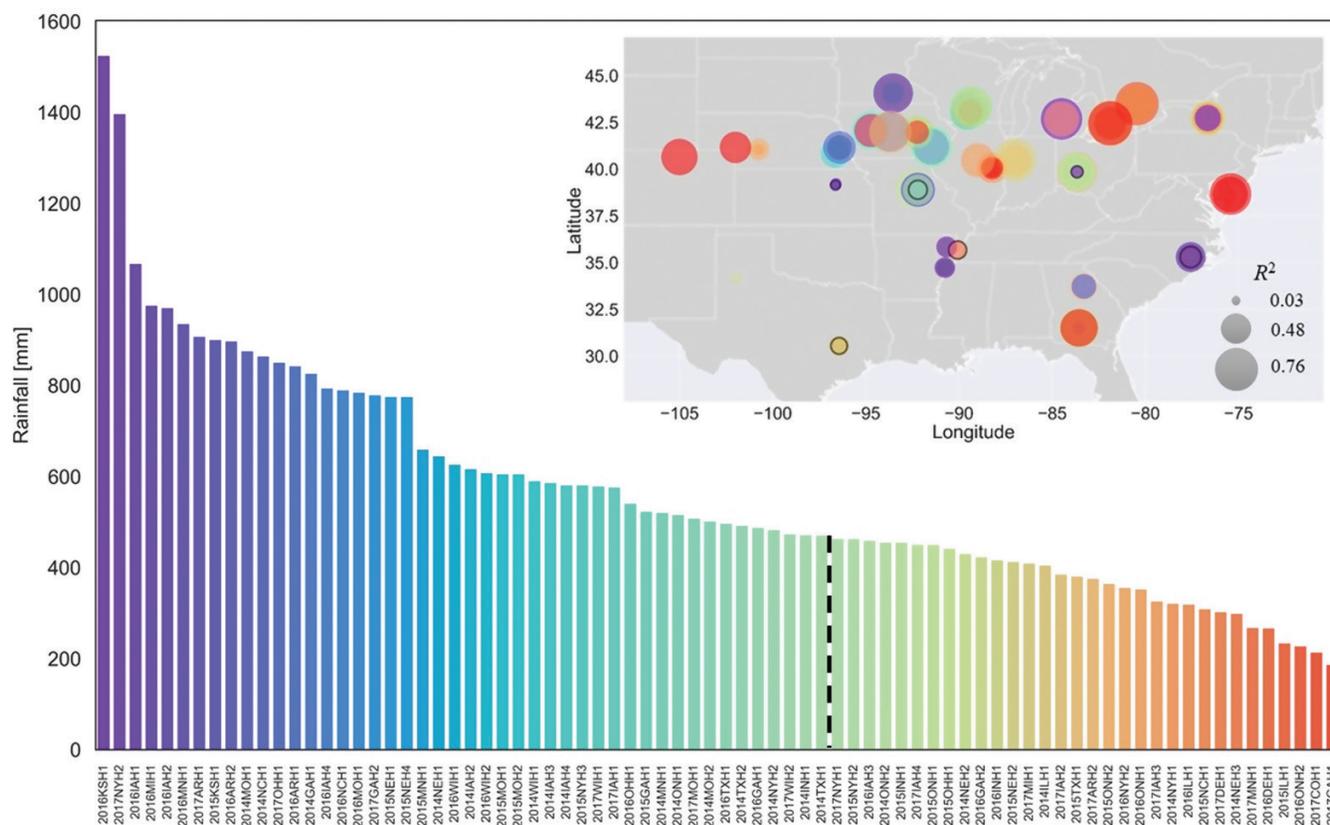
Fig. 10. Spatial distribution and decreasingly sorted R_{acc} (colored gradients) and R^2 (same as in Fig. 8). The vertical dashed line separates the experiments with a consistent improvement pattern as observed in Fig. 9C. The color code (mm) in the lower panel matches the color code of the circles in the map. The colored circles are transparent to distinguish multiple trials at the same location. Note: as in Fig. 8, the circles with a black perimeter represent negative correlation values. No perimeter means positive correlation.

the experiments with Size$_{test}$ ≤40 from the rest. The range of R_{acc} is between max(R_{acc})=1525 mm in the '2016KSH1' and min(R_{acc})=497 mm in the '2014TXH2' experiment. The model's consistent improvement in performance is reached with the accumulative rainfall value >497 mm. The R_{acc} standard deviation for experiments groups with R_{acc} ≥497 mm, and R_{acc} <497, is 229, and 84 mm, respectively. These results show improved G×E predictability by introducing more variability in the training set. The change rate of R^2 obtained by a rainfall gradient and the pattern of consistent improvement imply the contribution of a climatic variable (here, rainfall) in prediction enhancement, which suggests the exploration of other major climate covariates for consideration in the simulation procedure. In Fig. 10, it can also be seen how geospatial rainfall variability is reflected in model predictability. For example, model predictability in South Georgia experiments with smaller rainfall values outperforms that in the northern portion of the state with higher rainfall. This difference in model predictive skill in experiments with different rainfall values is another reason for the necessity for sensitivity analysis in order to find the variables with the most influence on model performance. While it is beyond the scope of this study to indicate the climate and genetic attributions of such variation, it reveals an opportunity

to elucidate those attributions as potential sources of predictability.

The generally better accuracy achieved by the 'climate-based' than the 'covariance-based' scenario illustrates a sequential and consistent improvement in the predictability of maize yields. The former shows the effect of the error introduced when all the environmental covariables are introduced in the simulation. In other words, model performance does not improve just by adding climatic variables. Based on the elements of the discussion above, the environmental heterogeneity strengthens the model performance, but it will be more effective when the sensitivity of crop yields comes from environmental variables such as rainfall. As explained earlier in this research and other studies, crop yield is affected by climatic variables at different levels (Meyer *et al.*, 1991; Romay *et al.*, 2010; Zhao *et al.*, 2015). Yet, this study introduces what climate variability across spatiotemporal scales and emerging patterns may offer to the diagnostics and prognostics of a plant's ability to adapt to changes in climate.

Conclusions

In this study, we assessed the performance change of the statistical G×E model for maize yield prediction for (i) the train–

test population size and (ii) the train–test structure designs. The G2F initiative releases a comprehensive database consisting of genetic, phenotypic, and environmental data of several maize hybrids for G×E modeling and simulation. Despite the efforts in recording databases without missing elements, there are several gaps and missing records throughout the G2F-E time series. The missing data affect all the three-dimensional data coverage, including environmental and 'omics' information in modeling and simulation procedures. To tackle this problem, we designed an evaluation–improvement pipeline to categorize the environmental datasets into 'complete', 'empty', and 'incomplete' groups, then replaced the data gaps in the 'empty' datasets and simulated missing samples in the 'incomplete' datasets by employing DNNs. Thus, the number of total observed experiments for G×E simulation upgrades from 32 to 84, enabling the improvement of genotypes from 372 to 376 and yield observations from 3169 to 8171. This improvement in three datasets resulted in a 12.1% improvement in predictability of maize yield values in terms of R^2 measurement. The RMSE, MSE, and MAE measurements improved by 2.2, 11.4, and 1.4%, respectively. The data can be found in Sarzaeim *et al*. (2022).

In conclusion, any improvement in environmental or 'omics' databases will lead to a better G×E predictive skill. Additionally, the statistically insignificant G×E predictability enhancement suggests more DFs for remarkable simulation accuracy. On the other hand, we examined the G×E prediction skill in three train–test selection experiments scenarios which are called 'random-based', 'covariance-based', and 'climate-based' schemes. The model performance is enhanced with a larger Size$_{train}$ in all mentioned approaches due to more variability to which the model is exposed in the training set structure. The 'random-based' scenario achieved the best improvement rate of 0.012. This improvement indicates that environmental variability is a key driver of phenotype predictability within the training set experiments as we compared 'climate-based' and 'covariance-based' scenarios. Thus, larger Size$_{train}$ in the 'random-based' scenario leads to a larger environmental variation to the G×E model because of the built train–test selection strategies. Also, the consistent improvement pattern observed in the 'climate-based' scenario, in which the trials are selected based on ranked accumulative rainfall, evidences the different levels of effectiveness of climate variables in yield predictive skill and the environmental variation introduced to the model by them. Environmental variability contributes to enhancing predictability through the use of more effective variables such as rainfall. The results suggest that environmental variables such as rainfall (i.e. soil moisture) will enhance environmental covariates, leading to improvements in phenotype predictability.

Future work

The results found in this study can nurture further efforts to improve G×E analytics and predictability by enhancing

hydroclimate analytics as follows: (i) further analyze multiple levels of digital climate and hydrologic data to improve model predictability, coupling a Global Sensitivity Analysis with G×E analytics; such a coupling will contribute to elucidating the environmental variables that increase the predictability of maize yields and to what extent; (ii) identify geospatial patterns of variability, designing geospatial visualization and aggregation analyses based on climate selection schemes to provide regional G×E yield predictability improvements; and (iii) develop robust response software architectures for multidimensional database management and visualization, and multi-lead-time phenotype predictive systems.

Author contributions

PS: data curation; PS and FM-A: formal analysis, methodology, visualization, writing—original draft; FM-A: conceptualization, funding acquisition, project administration, supervision, writing—review & editing; DJ: supervision, methodology, writing—review & editing.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.6299090 (Sarzaeim *et al*., 2022).

References

Acosta-Pech R, Crossa J, de los Campos G, Teyssèdre S, Claustres B, Pérez-Elizalde S, Pérez-Rodríguez P. 2017. Genomic models with genotype × environment interaction for predicting hybrid performance: an application in maize hybrids. Theoretical and Applied Genetics **130**, 1431–1440.

Alexandratos N, Bruinsma J. 2012. World Agriculture towards 2030/2050: The 2012 Revision. ESA Working paper No. 12-03. Rome: FAO.

Amaranto A, Munoz-Arriola F, Corzo G, Solomatine DP, Meyer G. 2018. Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland. Journal of Hydroinformatics **20**, 1227–1246.

Amaranto A, Munoz-Arriola F, Solomatine DP, Corzo G. 2019. A spatially enhanced data-driven multimodel to improve semiseasonal groundwater forecasts in the high plains aquifer, USA. Water Resources Research **55**, 5941–5961.

Amaranto A, Pianosi F, Solomatine DP, Corzo G, Muñoz-Arriola F. 2020. Sensitivity analysis of data-driven groundwater forecasts to hydroclimatic controls in irrigated croplands. Journal of Hydrology **587**, 124957.

Amato F, Guignard F, Robert S, Kanevski M. 2020. A novel framework for spatio-temporal prediction of environmental data using deep learning. Scientific Reports **10**, 22243.

Andresen JA, Alagarswamy G, Alan Rotz C, Ritchie JT, LeBaron AW. 2001. Weather impacts on maize, soybean, and alfalfa production in the Great Lakes region, 1895–1996. Agronomy Journal **93**,1059–1070.

Auinger HJ, Lehermeier C, Gianola D, Mayer M, Melchinger AE, Silva S, Knaak C, Ouzunova M, Schön CC. 2021. Calibration and validation of predicted genomic breeding values in an advanced cycle maize population. Theoretical and Applied Genetics **134**, 3069–3081.

Baril CP, Denis JB, Wustman R, van Eeuwijk FA. 1995. Analysing genotype by environment interaction in Dutch potato variety trials using factorial regression. Euphytica **82**, 149–155.

Bojanowski JS, Vrieling A, Skidmore AK. 2014. A comparison of data sources for creating a long-term time series of daily gridded solar radiation for Europe. Solar Energy **99**, 152–171.

Bustos-Korts D, Romagosa I, Borràs-Gelonch G, Casas AM, Slafer GA, van Eeuwijk F. 2018. Genotype by environment interaction and adaptation. In: Meyers R, ed. Encyclopedia of sustainability science and technology. New York: Springer.

Chai T, Draxler RR. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development **7**, 1247–1250.

Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. 2013. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theoretical and Applied Genetics **126**, 867–887.

Crossa J, Pérez-Rodríguez P, Cuevas J, et al. 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends in Plant Science **22**, 961–975.

Duvick DN. 2005. Genetic progress in yield of United States maize (*Zea mays* L.). Maydica **50**,193–202.

Finlay KW, Wilkinson GN. 1963. The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research **14**, 742–754.

Ghimire S, Deo RC, Raj N, Mi J. 2019. Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction. Energies **12**, 2407.

Gianola D. 2021. Opinionated views on genome-assisted inference and prediction during a pandemic. Frontiers in Plant Science **12**, 717284.

Hayes PM, Liu BH, Knapp SJ, et al. 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. Theoretical and Applied Genetics **87**, 392–401.

Hernández E, Sanchez-Anguix V, Julian V, Palanca J, Duque N. 2016. Rainfall prediction: a deep learning approach. In: Martínez-Álvarez F., Troncoso A., Quintián H., Corchado E, eds. Hybrid artificial intelligent systems. HAIS 2016. Lecture notes in computer science, vol. **9648**. Cham: Springer, 151–162.

Hoogenboom G. 2000. Contribution of agrometeorology to the simulation of crop production and its applications. Agricultural and Forest Meteorology **103**, 137–157.

Howard R, Carriquiry AL, Beavis WD. 2014. Parametric and non-parametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3 Genes, Genomes, Genetics **4**,1027–1046.

IPCC. 2001. Climate change 2001: the scientific basis. Cambridge: Cambridge University Press.

Jarquín D, Crossa J, Lacaze X, et al. 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics **127**, 595–607.

Jarquín D, da Silva CL, Gaynor RC, Poland J, Fritz A, Howard R, Battenfield S, Crossa J. 2017. Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. The Plant Genome **10**, doi:10.3835/plantgenome2016.12.0130.

Jarquín D, de Leon N, Romay C, et al. 2021. Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Frontiers in Genetics **11**, 592769.

Kumar A, Ramsankaran R, Brocca L, Munoz-Arriola F. 2019. A machine learning approach for improving near-real-time satellite-based rainfall estimates by integrating soil moisture. Remote Sensing **11**, 2221.

Kumar A, Ramsankaran R, Brocca L, Muñoz-Arriola F. 2021. A simple machine learning approach to model real-time streamflow using satellite inputs: demonstration in a data scarce catchment. Journal of Hydrology **595**, 126046.

Lawrence-Dill CJ, Schnable PS, Springer NM. 2019. Idea factory: the maize genomes to fields initiative. Crop Science **59**, 1406–1410.

Lobell DB, Burke MB. 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology **150**, 1443–1452.

Lobell DB, Cassman KG, Christopher BF. 2009. Crop yield gaps: their importance, magnitudes, and causes. Annual Review of Environment and Resources **34**, 179–204.

Long N, Gianola D, Rosa GJM, Weigel KA. 2011. Application of support vector regression to genome-assisted prediction of quantitative traits. Theoretical and Applied Genetics **123**, 1065–1074.

Lopez-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink JL, Singh RP, Autrique E, de los Campos G. 2015. Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3: Genes, Genomes, Genetics **5**, 569–582.

Lopez-Cruz M, de los Campos G. 2021. Optimal breeding-value prediction using a sparse selection index. Genetics **218**, iyab030.

Maltais-Landry G, Lobell DB. 2012. Evaluating the contribution of weather to maize and wheat yield trends in 12 U.S. counties. Agronomy Journal **104**, 301–311.

Matthews RB, Rivington M, Muhammed S, Newton AC, Hallett PD. 2013. Adapting crops and cropping systems to future climates to ensure food security: the role of crop modelling. Global Food Security **2**, 24–28.

Mbungu WB, Mahoo HF, Tumbo SD, Kahimba FC, Rwehumbiza FB, Mbilinyi BP. 2015. Using climate and crop simulation models for assessing climate change impacts on agronomic practices and productivity. In: Lal R., Singh B., Mwaseba D., Kraybill D., Hansen D., Eik L, eds. Sustainable intensification to advance food security and enhance climate resilience in Africa. Cham: Springer, 201–2019

McFarland BA, Alkhalifah N, Bohn M, et al. 2020. Maize Genomes to Fields (G2F): 2014–2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. BMC Research Notes **13**, 71.

Messina CD, Technow F, Tang T, Totir R, Gho C, Cooper M. 2018. Leveraging biological insight and environmental variation to improve phenotypic prediction: integrating Crop Growth Models (CGM) with Whole Genome Prediction (WGP). European Journal of Agronomy **100**, 151–162.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**, 1819–1829.

Meyer SJ, Hubbard KG, Wilhite DA. 1991. The relationship of climatic indices and variables to corn (maize) yields: a principal components analysis. Agricultural and Forest Meteorology **55**, 59–84.

Olesen JE, Jensen T, Petersen J. 2000. Sensitivity of field-scale winter wheat production in Denmark to climate variability and climate change. Climate Research **15**, 221–238.

Osei MK, Asante MD, Agyeman A, Adebayo MA, Adu-Dapaah H. 2014. Plant breeding: a tool for achieving food sufficiency. In: Nandwani D, ed. Sustainable horticultural systems. sustainable development and biodiversity, vol 2. Cham: Springer, 253–274.

Pérez P, de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics **198**, 483–495.

Quiñones R, Munoz-Arriola F, Choudhury SD, Samal A. 2021. Multi-feature data repository development and analytics for image cosegmentation in high-throughput plant phenotyping. PLoS One **16**, e0257001.

Raoufi RS, Soufizadeh S. 2020. Simulation of the impacts of climate change on phenology, growth, and yield of various rice genotypes in humid sub-tropical environments using aquacrop-rice. International Journal of Biometeorology **64**, 1657–1673.

Ray DK, Gerber JS, MacDonald GK, West PC. 2015. Climate variation explains a third of global crop yield variability. Nature Communications **6**, 5989.

Rogers AR, Holland JB. 2022. Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. G3 Genes, Genomes, Genetics **12**, jkab440.

Rojas M, Lambert F, Ramirez-Villegas J, Challinor AJ. 2019. Emergence of robust precipitation changes across crop production areas in the 21st century. Proceedings of the National Academy of Sciences, USA **116**, 6673–6678.

Romay MC, Malvar RA, Campo L, Alvarez A, Moreno-González J, Ordás A, Revilla P. 2010. Climatic and genotypic effects for grain yield in maize under stress conditions. Crop Science **50**, 51–58.

Sarzaeim P, Jarquin D, Muñoz-Arriola F. 2020. Analytics for climate-uncertainty estimation and propagation in maize-phenotype predictions. 2020 ASABE Annual International Meeting, Omaha, NE, USA, July 12–15 doi: 10.13031/aim.202000884.

Sarzaeim P, Muñoz-Arriola F, Jarquin D. 2022. Large-scale and multi-dimensional climate, genetics, and phenotypes database for maize yield predictability in the U.S. and Canada (version 1) [Dataset]. Zenodo doi:10.5281/zenodo.6299090.

Sengupta M, Xie Y, Lopez A, Habte A, Maclaurin G, Shelby J. 2018. The National Solar Radiation Data Base (NSRDB). Renewable and Sustainable Energy Reviews **89**, 51–60.

Shekhar S, Colleti J, Munoz-Arriola F, Ramaswamy L, Krintz C, Varshney L, Richardson D. 2017. Intelligent infrastructure for smart agriculture: an integrated food, energy and water system. arXiv:1705.01993. 2017arXiv170501993S. [Preprint].

Stehfest E, Heistermann M, Priess JA, Ojima DS, Alcamo J. 2007. Simulation of global crop production with the ecosystem model DayCent. Ecological Modelling **209**, 203–219.

Tao F, Yokozawa M, Liu J, Zhang Z. 2008. Climate–crop yield relationships at provincial scales in China and the impacts of recent climate trends. Climate Research **38**, 83–94.

TASSEL. 2019. Maizegenetics Tassel 5 Source. https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual/NumericalGenotype/NumericalGenotype. Accessed April 2021.

TASSEL. 2021. Tasseladmin/ Tassel 5 Source/ Wiki/ Home — Bitbucket. https://bitbucket.org/tasseladmin/tassel-5-source/wiki/Home. Accessed October 2021.

Thornton PE, Thornton MM, Mayer BW, Wei Y, Devarakonda R, Vose RS, Cook RB. 2018. Data from: Daymet: daily surface weather data on a 1-km grid for North America,. https://daac.ornl.gov/DAYMET/guides/Daymet_Daily_V4.html

Tilman D, Balzer C, Hill J, Befort BL. 2011. Global food demand and the sustainable intensification of agriculture. Proceedings of the National Academy of Sciences, USA **108**, 20260–20264.

Van Eeuwijk FA, Bustos-Korts DV, Malosetti M. 2016. What should students in plant breeding know about the statistical aspects of genotype × environment interactions? Crop Science **56**, 2119–2140.

Willmott CJ. 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society **63**, 1309–1313.

Zhao J, Guo J, Mu J. 2015. Exploring the relationships between climatic variables and climate-induced yield of spring maize in Northeast China. Agriculture, Ecosystems & Environment **207**, 79–90.