Investigation of machine learning approaches to predict quantitative traits using environmental and genomic information

Dissertation

to obtain the Doctor of Philosophy (Ph.D.) degree

at the Faculty of Agricultural Sciences,

Georg-August University Göttingen, Germany

Submitted by Cathy Colette Westhues, born Jubin

Born on June 4, 1994 in Beaupréau, France

Göttingen, April 2022

1st Referee: Prof. Dr. Timothy Mathes Beissinger

Plant Breeding Methodology

Department of Crop Sciences

Georg-August-University Göttingen

2nd Referee: Prof. Dr. Henner Simianer

Animal Breeding and Genetics Group

Department of Animal Sciences

Georg-August-University Göttingen

3rd Referee: Prof. Dr. Alexander Edward Lipka

Department of Crop Sciences

University of Illinois, Urbana-Champaign

4rd Referee: Prof. Dr. Diego Jarquin

Department of Agronomy

University of Florida, Gainesville

Date of oral examination: 18th of May, 2022

Declaration

1. I hereby declare that this work has not been submitted to other examination authorities in the same or a similar form. I further declare that I have not applied for a doctorate at any other university.

2. I hereby declare under oath that this dissertation was prepared independently and without unauthorized assistance.

Einbeck, 01.04.2022

Cathy Westhues

Mathematics is the science of patterns, and nature exploits just about every pattern that there is.

Ian Stewart (2008). "Nature's Numbers: The Unreal Reality Of Mathematics"

Table of Contents

List of Abbreviations

ANN	Artificial Neural Network
AWS	Automatic Weather Station
CV	Cross-validation
DTW	Dynamic Time Warping
EC	Environmental covariate
FW	Finlay-Wilkinson
G2F	Genomes to Fields
GBDT	Gradient boosted decision trees
GDD	Growing Degree Days
G × E	Genotype-by-environment
G × L	Genotype-by-location
G × Y	Genotype-by-year
G × Y × L	Genotype-by-year-by-location interaction
GP	Genomic prediction
GWAS	Genome-wide Association Study
L	Location effect
MET	Multi-environmental trials
ML	Machine learning
PCA	Principal component analysis
PDP	Partial dependence plot

QTL	Quantitative trait locus
LME	Linear mixed effects (model)
SNP	Single nucleotide polymorphism
TRN	Training set
TST	Test set
Y	Year effect

1. General Introduction

1.1 Background

Developing new crop cultivars resilient to future climatic conditions presents a mounting challenge for plant breeding, as weather patterns become increasingly volatile and extreme due to climate change in many regions of the world. In the last decades, many studies have underpinned the expected aftermath of detrimental weather conditions, with dramatically negative effects on yields for major cereal species, such as wheat, rice and maize (Gammans et al., 2017; Trnka et al., 2014). These three staple grains represented on their own more than 41% of the total calories for human consumption in 2019 (FAOSTAT, 2020). Zhao et al. (2017) reported, based on different analytical methods, that every degree-Celsius increase in mean temperature could result in a reduction of global yields on average by 7.4% and by 6%, for maize and wheat, respectively. Yield variability can often be attributed to major environmental stresses (deficits of soil water, high temperatures) occurring at critical plant developmental stages (Lizaso et al., 2018). Hence, current and future genetic improvement of crop varieties need to equip genotypes with increased phenotypic plasticity, characterized by an improved tolerance for prolonged periods of drought and heat stresses, in order to mitigate serious yield losses and subsequent economical consequences.

Yet, the task of predicting genotype performance in future environments, investigated in this thesis, is further compounded by the presence of complex genotype-by-environment interactions (GE). G $\times$ E refers to a pervasive observation in plant breeding that the relative ranking of genotypes generally changes conditional on the considered growing environment (Allard and Bradshaw, 1964; Cooper and DeLacy, 1994). Dissecting G $\times$ E effects is only possible by analyzing multi-environment trials (MET) data. The recent advancements in the field of environmental sensing technologies, of high-throughput genotyping and phenotyping have enabled generating considerable amounts of data in the context of plant breeding programs across MET. Harnessing these various types of information, often described by a large number of variables, is demanding and requires sophisticated modeling approaches. While whole-genome prediction of complex quantitative traits has become a major tool in modern plant breeding (Crossa et al., 2017), weather and/or soil data are still not routinely used alongside genomic information to select promising candidate genotypes. Pedoclimatic data

can help define more precisely environmental conditions encountered in multi-environment trial (MET) datasets; however their incorporation also raises statistical and computational issues with regards to the high dimensionality of the corresponding G × E component.

In addition, genotypic responses to environmental gradients are often complex and nonlinear (Heslot et al., 2014; Malosetti et al., 2006). Machine learning techniques offer high potential to capture these nonlinear relationships and to explore untapped sources of environmental data. Nonetheless, capitalizing on these predictive modeling approaches requires taking into account typical pitfalls, among which (1) inappropriate pre-processing of the data, (2) inadequate size of training data to efficiently learn G × E patterns, (3) over-fitting the model to the training data, i.e. the model uncovers the random noise rather than true patterns within the data, and is therefore not able to generalize well on a new TST. We explored in this thesis recent data mining methods, with the objective of evaluating the predictive ability of genotype performance across various cross-validation schemes relevant for plant breeders. In the following sections, characteristics related to the statistical approaches for genome-based prediction in multi-environment trials, as well as major hurdles when dealing with environmental datasets and machine learning methods are reviewed and discussed.

1.2 Genotype-by-environment interactions in plant breeding

1.2.1 Strategies to deal with G × E interactions

Before getting officially released as new varieties on the market, genotypes need to be assessed across multiple years and locations in order to obtain reliable estimates of their performance for quantitative phenotypic traits, such as yield, to assess their stability and plasticity. In this thesis, we designate an environment as a specific year-location, sometimes even also year-location-management, combination.

Multiple-environment trials (MET) are useful to identify relevant genotype-by-environment (GE) patterns, such as G × E cross-over interactions, which imply that the best genotype in an environment might perform less well than others in another environment. G × E can be further partitioned into genotype-by-location (G × L), genotype-by-year (G × Y) and genotype-by-year-by-location (G × Y × L) interactions using analyses of variance (ANOVA). When the G × L term prevails, i.e. when repeatable G × E components, such as soil and/or management factors largely influence the trait of interest, it becomes possible to make recommendations for genotypes specifically adapted to subsets of geographical regions with homogeneous environmental conditions (Bernardo, 2002). These groups of environments can be defined using unsupervised learning methods, such as clustering analyses or principal component analysis (PCA). On the other hand, when G × Y and G × Y × L terms are dominant, the unpredictability of weather conditions, especially in the current

context of climate change, further complicates selection decisions.

A key concept related to G × E interactions is the phenotypic plasticity. A genotype can be characterized by its ability to respond phenotypically to modifications in the environmental conditions, and this level of phenotypic plasticity is in general represented by the form of its reaction norm. The creation of cultivars exhibiting a high level of plasticity (e.g. reaction norm with a large slope) might be suitable for local adaptation to specific environmental conditions; however, it also means that the variety presents less potential for broad adaptation and a precise understanding of the factors underlying the G × E interactions is oftentimes necessary. Thus, rather than defining the best genotype for a specific environment, a more commonly applied strategy is to release cultivars with the best average performance across all environments included in the MET, which is supposed to be a proxy for the target population of environments (TPE). In this latter case, G × E effects are hence not exploited (Bernardo, 2002).

1.2.2 Classical parametric approaches to describe G × E interactions

In classical plant breeding, the use of linear regression models to model G × E effects remains a standard and relatively efficient practice. Joint-regression analyses (Eberhart and Russell, 1966; Finlay and Wilkinson, 1963; Yates and Cochran, 1938), which consist in a genotype-specific regression on the environmental mean - the latter being generally simply calculated as the mean of all genotypes in the considered environment - have become very popular as a method to identify stable genotypes across environments. It should be noted that the objectives in terms of stability can differ considerably according to the breeding regions. In this model, a given genotype i with a large intercept (μ_i), i.e. a high average performance, and a slope (b_i) close to 1, i.e. its performance follows the mean response of other cultivars across environments, might be defined as a suitable genotype according to the concept of dynamic stability (type II) (Lin et al., 1986).

An improvement of G × E modeling was provided by the development of the additive main effects and multiplicative interaction (AMMI) (Gauch Jr et al., 1992). This model assumes that the G × E interaction effects can be decomposed into more than one multiplicative term, while the genotype and environmental effects are further treated as additive main effects and estimated with ANOVA. Specifically, a principal component analysis is applied on the residuals of the additive model, i.e. the G × E matrix with dimensions $n \times m$, where n represents the number of genotypes and m the number of environments. This yields two matrices: one giving the genotype scores (dimensions $n \times \min(n, m)$) and another one representing the environment scores (dimensions $m \times \min(n, m)$). The objective of the AMMI analysis is to obtain improved estimates of genotype performance in various environments, which are corrected for the random, spurious G × E present in MET data, and which represent the true G × E patterns captured by the first principal component axes. The noisy component is often contained within the G × E term, as it exhibits the highest degrees of freedom within the data (Gauch Jr, 2006; Malosetti et al., 2013). The AMMI model

capitalizes on the complete dataset to obtain a good approximation of repeatable G × E interaction effects, instead of simply considering the average estimate of the genotype performance for a given environment.

Neither the standard Finlay-Wilkinson approach, nor AMMI allow for the inclusion of genetic relationships among individuals based on pedigree or genomic data. Another limitation of these models is that they use as proxy for the environmental index the average effect of the environment over all genotypes and therefore cannot account for specific environmental factors. Hence, they cannot shed light on crucial environmental stress covariates, which might provoke linear or non-linear genotype responses along their gradients. Nonenetheless, these two approaches remain very relevant in predictive breeding today, as we will show by providing further application examples with these methods in the **chapter 5**.

1.3 Collecting and processing environmental data for incorporation in predictive models

Characterizing the quality of growing environments with relevant environmental indices implies to have access to environmental data. Various environmental variables can influence cultivar performance, among them soil physical, chemical and health properties, climatic variables, or factors related to disease pressure (Bernardo, 2002). While various sensors or satellite-based systems can give us access to this information, figuring out the most appropriate data source and meaningfully processing raw hourly or daily weather data are important preprocessing steps.

1.3.1 Weather data sources

In recent years, monitoring and collecting climate information using automatic weather stations closely located to fields in MET experiments have been more widely put into practice, as demonstrated by the Genomes to Fields Initiative (AlKhalifah et al., 2018; McFarland et al., 2020) or other large MET experiments (Ly et al., 2018; Rincent et al., 2019). These sensors generally collect data for temperature, rainfall, relative humidity, dewpoint, solar radiation, wind speed and wind direction at 30-min or 15-min intervals during the growing season. The resulting meteorological time series need to be processed with quality control procedures, which we detail in the supplemental information of **chapter 2**, in order to deliver high quality climatic data. In particular, partial (e.g. rainfall gauge) or complete breakdowns of weather stations can occur during the growing season, leading to missing data which need to be imputed. Using data retrieved from public surface observing systems (e.g. records from the Global Historical Climatology Network (GHCN) or from the Global Surface Summary of the Day (GSOD) in the US), different solutions can be applied to infer missing values, such as (1) simple and fast deterministic methods like inverse distance weighting, which basically estimates the missing daily weather variable value by taking the aver-

age of all nearby sites within a particular radius and assigning greater weights to closer locations; (2) geo-statistical methods, such as spatio-temporal kriging that exploits both a spatial and temporal covariance matrix (Pebesma and Heuvelink, 2016; Pebesma, 2004); or (3) using data-driven machine learning algorithms, e.g. random forest or artifical neural networks (ANN) (Hengl et al., 2018; Kashani and Dinpashoh, 2012; Mital et al., 2020). In **chapter 2**, we implemented (1) and (2) approaches to replace missing values in the G2F weather dataset.

While automatic weather stations offer the possibility to get accurate local measurements, they must be situated close enough to the field, require maintenance, application of a control quality procedure and imputation of missing weather values. Alternatively, satellite- and model-based databases, such as the National Aeronautics and Space Administration's POWER database (NASA, `https://power.larc.nasa.gov/cgi-bin/cgiwrap/solar/agro.cgi`), produced by the NASA Langley Research Center POWER Project, can be easily downloaded when surface measurements are not available. An R package has also been recently developed to facilitate the retrieval of solar and weather data from NASA POWER database (Sparks, 2018). As outlined by the NASA POWER (NP) database documentation, this type of data offers two advantages: it has global coverage, with a 0.5 x 0.625 degree latitude and longitude spatial resolution for meteorology, and 1 x 1 degree latitude and longitude for solar parameters. In addition, it does not include data gaps.

1.3.2 Developing crop-based environmental indices

Once the weather data are cleaned, another issue arises: how to reduce the number of variables, and condense large amounts of daily weather information without loss of crucial information? Although some studies make use of complete daily weather data (Washburn et al., 2021a; Widener et al., 2021), a more common practice is to summarize it over specific time windows. Thereby, the number of variables used in subsequent prediction models can be reduced considerably. When the variability in crop growing season lengths across environments is low, a simple approach can be to use non-overlapping sliding day-windows of fixed lengths throughout the growing season, over which the environmental variables are calculated (Jarquin et al., 2021; Rogers et al., 2021). This approach does not require any knowledge about the phenology of the crop under study. However, defining environmental covariates (ECs) using crop developmental stages may better account for the fact that climatic conditions can have a variable impact depending on the plant's phenological stage. In particular, when plants are subjected to environmental stresses that line up with critical phenological stages, for instance at flowering or during grain filling, major consequences on important agronomic traits can often be observed, e.g. a decrease of grain yield and quality (Tardieu et al., 2018).

Nevertheless, determining the timing of development is highly challenging and theoretically requires to have extended phenotypic information about the phenological development of each genotype within each field experiment, which is currently infeasible from a management and cost perspec-

tive. A relatively easy solution is to use agronomic knowledge to approximate plant developmental stages from day after planting onwards, based on a standard crop growing season, as illustrated by studies with maize hybrids (Costa-Neto et al., 2021), with rice (Delerce et al., 2016) and with cotton (Pérez-Rodríguez et al., 2015). The major drawback of these methods is that differences in crop development among environments, mainly explained by variability in weather conditions, are not considered. Millet et al. (2019) proposed a more precise method, exploiting information coming from phenotyping platforms and from field observations. Based on repeated measurements obtained via advanced phenotyping platforms of leaf phenological progression over a panel of maize genotypes, and on field records for silking date, these authors defined per hybrid the timing of different phenological events (e.g. floral transition, silk initiation, end of abortion and grain maturity) for MET field experiments in Europe on the basis of temperature records for these environments. Thermal times weres also used by Boer et al. (2007) to approximate maize developmental stages and to derive ECs from weather data. When flowering time is scored within each field experiment and for each genotype, a more straightforward method, that we applied in **chapter 2**, is to define crucial hybrid-specific phenological stages (i.e. vegetative, flowering, and grain fill stages) and to derive climatic and stress covariates over these growth stages, which was done similarly in the work of Monteverde et al. (2019).

When few or no other phenotypic data than the final trait of interest is available, more elaborate process-based simulation frameworks, such as crop growth models (CGM), have been of interest to predict plant developmental stages (Heslot et al., 2014). These models rely on a set of linear and nonlinear equations to model crop physiological processes in response to environmental conditions (e.g. accumulation of thermal time), and take as input genotype-specific physiological parameters, weather and soil data. Once the duration of these developmental stages is known, climatic and/or stress covariates can be calculated over these time periods (Heslot et al., 2014; Rincent et al., 2019). Two possibilities to integrate crop growth models in association with genomic-based predictive models are outlined by Heslot et al. (2014). As done by these authors, the output of CGM models, run using only one or a few genotypes as representative cultivars, can be directly incorporated as ECs in genomic prediction models. The second approach requires to first estimate genetic parameters, accounting for differences among genotypes for some phenological traits (e.g. phenological traits generally showing a higher heritability than the final trait), and then to run the CGM, for each cultivar using its genetic parameters along with environmental data, to predict the integrated trait, for instance grain yield. To calibrate the model and accurately estimate these genetic parameters, one possibility is to exploit measurements of the underlying phenological traits for genotypes included in the TRN, which can be obtained for instance with high-throughput phenotyping capabilities (e.g. drones). Thus, whole-genome prediction (WGP) models can be implemented to estimate marker effects, which in turn can be used to predict these physiological traits for newly developed genotypes. Nonetheless, as mentioned before, it can be

rather complicated to have access to this highly detailed phenotypic data, even for the TRN only, and especially for the plant breeding industry for which the number of candidate genotypes is generally high. Hence, a more straightforward approach is to consider these physiological traits as hidden variables (Technow et al., 2015) and to infer them using Bayesian algorithms in so-called WGP-CGM (Messina et al., 2018; Technow et al., 2015), or CGM coupled with marker-assisted selection (CGM-MAS) (Rincent et al., 2017). The procedures described by Technow et al. (2015) and Messina et al. (2018) are of particular interest, since molecular markers effects are used to connect genotypic information to these latent physiological parameters, with the goal of being able to predict genotype-specific parameters in the test set (TST) on the basis of genotypic data only.

A meaningful feature engineering strategy is to compute or to extract from CGM additional variables based on science-based equations of crop ecophysiology in response to weather conditions, such as stress covariates (for instance, number of days above or below a certain temperature threshold likely associated with heat or frost waves events, respectively); daily crop evapotranspiration (Allen et al., 1998), that describes the two processes of evaporation and transpiration by which water moves from land surface to atmosphere and essential to understand crop water use; or radiation use efficiency (Monteith, 1977; Russell et al., 1989), which quantifies the conversion of intercepted radiation to biomass. Numerous publications used these types of more elaborate ECs (Costa-Neto et al., 2021; Heslot et al., 2014; Ly et al., 2018; Monteverde et al., 2019; Rincent et al., 2019), which can better reflect plant physiological responses to drought and heat stresses. Stress covariates were also shown to be superior to basic climatic variables and better related to the GE covariance matrix derived from AMMI analysis in the study by Rincent et al. (2019). Some popular process-based models used to generate these integrative physiological ECs are SiriusQuality (Martre et al., 2006) for small grain cereals, such as wheat, APSIM (Hammer et al., 2010; Holzworth et al., 2018, 2014), and WOFOST (Van Diepen et al., 1989; de Wit et al., 2019).

1.4 Accounting for genotype-by-environment interactions in the genomic prediction framework

1.4.1 Genomic prediction

The development of high-throughput genotyping techniques has enabled the large and successful deployment of genomic prediction (GP) in animal breeding first, and later adapted in plant breeding, starting with the RidG $\times$ ERegression BLUP model Meuwissen et al. (2001). Considering a training population of individuals that are characterized with genome-wide molecular markers and phenotyped (named training set = TRN), a prediction model linking genomic data to phenotypic traits is fitted and subsequently used to predict the performance of a testing population, that has also been genotyped but has no phenotypic records. This model relies on the statistical assump-

tion that most complex traits are controlled by a very large number of loci with very small effects, hence the variance explained by each marker is very small and equal for all loci. In these models, marker effects are considered as random, while all other effects, including environments, are generally treated as fixed. VanRaden (2008) proposed an equivalent model, the so-called GBLUP, which presents the advantage of reducing the dimensionality of the problem and of making it computationally more affordable, by calculating a genomic relationship matrix denoted **G** from marker data. GP has been applied with success, leading to a decrease of the length of breeding cycles, and proved superior to pedigree-based approaches for a broad range of plant and animal species, and offers relatively low computational times (Crossa et al., 2017; Heslot et al., 2012a; Riedelsheimer et al., 2012).

However, a first limitation of these models is that they do not allow markers to have larger or no effects, although some previous genetic studies, such as genome-wide association studies (GWAS) or quantitative trait locus (QTL) analyses, can often inform us with a prior understanding of the genetic architecture of the trait under study. Indeed, it is often the case that some particular genes might explain a larger part of the genetic variance than other genomic regions. For instance, qualitative and quantitative resistance genes have been identified for many crop diseases, such as Htn1 for northern corn leaf blight (Hurni et al., 2015). This problem can be partially solved by using known large effects QTL as fixed covariates in the GP models (Bernardo, 2014; Rice and Lipka, 2019). Another issue arising from these models is that many complex phenotypic traits actually result from gene networks involving epistatic and dominance effects, i.e. non-additive genetic effects can also contribute to the observed phenotypic variation for these traits. For instance, epistatic effects were shown to be a major component of heterosis for selfing species like bread wheat (Jiang et al., 2017). As we will explain in **Section 1.5**, machine learning approaches can provide solutions to model nonlinear genomic effects.

An additional constraint of the original GP framework is that it does not account for genotype-by-environment interactions, hence preventing the development of cultivars on the basis of their specific adaptation to a given region.

1.4.2 Development of GP models accounting for GE

Burgueño et al. (2012) first proposed a multivariate multi-environment extension of the GBLUP model, where genomic and residual correlations among environments were modelled with various covariance functions (i.e. diagonal, identity and factor analytic). This model allows for the exploitation of genetic correlations between environments, thereby improving across-environment predictive ability compared to single-environment pedigree and genomics prediction models that ignored G $\times$ Eeffects.

Rather than using across-environment covariance structures, other approaches have focused on

modelling marker-by-environment (M × E) interactions (Crossa et al., 2016; Lopez-Cruz et al., 2015; Malosetti et al., 2008). The method developed by Lopez-Cruz et al. (2015) implies that marker effects are decomposed into a main component, which reflects stable effects across environments, and environment-specific deviations. While the original model was tested with a shrinkage approach (i.e. Bayesian ridge regression), Crossa et al. (2016) noted an increase in predictive ability by instead implementing Bayes B as a variable selection method. The M × E perspective offers several advantages relative to the covariance-based approach, that we previously mentioned, among them the possibility to identify markers presenting highly environment-specific effects, and are hence potentially involved in GE. Nonetheless, Lopez-Cruz et al. (2015) specified an important condition for a successful application: as the covariance among environments is restricted to be positive and constant, it is more reasonable to apply this approach for a set of environments with positive correlations that are confined to lie within a similar range.

1.4.3 A step further with the introduction of environmental information

Introducing environmental covariates into the GP statistical framework further increases the dimensionality of the problem, already introduced with whole-genome based predictions. On the other hand, these models allow to predict genotype performance when facing new environmental conditions, and to evaluate the impact of some specific stress covariates, for instance related to heat, drought or nitrogen deficiency or excesses. A class of models, namely factorial regression models, has frequently been used to integrate differential genotypic sensitivies to explicit environmental characteristics into predictive models (Heslot et al., 2014; Ly et al., 2018; Malosetti et al., 2004; Millet et al., 2019). This type of model follows the general expression:

$$\mu_{ij} = \mu + G_i + E_j + \sum_{k=1}^{K} \beta_{ik} Z_{jk},$$

where μ_{ij} is the mean of genotype i in environment j, μ is the overall mean, G_i is the random genotypic main effect of genotype i, E_j is the random environmental main effect of environment j, β_{ik} is the genotypic sensitivity for genotype i to the k-th EC and Z_{jk} is the k-th EC characterizing environment j. Heslot et al. (2014) tackled the aforementioned issue of data dimensionality by performing variable selection to retain only a subset of markers, which showed the most variable effects across environments, while the main genotype effect was modelled with the complete marker dataset. This approach avoids fitting all combinations of markers with ECs, which could be computationally highly challenging.

Jarquín et al. (2014) proposed another approach to incorporate interactions between large environmental and genomic datasets in a Bayesian reaction norm mixed model, where the random main effects of markers and of ECs are estimated using the covariance functions $\mathbf{G}$ and Ω, respectively, and the interaction between these two components is modelled by their Hadamard product. This model allows reducing the dimensionality of the M × E problem and to benefit from information

on the strength of correlations among environments and among genotypes. It has been widely used for various multi-environment crop datasets, demonstrating the benefit from using environmental information in GP models (Basnet et al., 2019; Costa-Neto et al., 2021; De Los Campos et al., 2020; Monteverde et al., 2019; Rincent et al., 2019; Sukumaran et al., 2017). We also implemented this statistical framework in **chapter 2** and in **chapter 4** of this thesis.

1.5 Machine learning methods to harness environmental and genomic data

1.5.1 The potential of machine learning with complex datasets

With the exception of the study of Heslot et al. (2014), where genotype sensitivities to particular ECs are modelled with regression trees, it should be noted that the methods described in the previous section assume that the interactions between molecular markers and environmental covariates are linear, and thus, do not adequately reflect the biological phenotypic plasticity of genotypes. In fact, QTL responses to environmental stress factors can be fit better with nonlinear methods, as illustrated by a study in potato (Malosetti et al., 2006) using logistic curves to model senescence progression over time. While the crop growth models described above simulate nonlinear responses to stress covariates, they cannot directly integrate genomic data, but require a previous step of estimating genotype-specific parameters. Hence, more flexible predictive modeling frameworks, able to combine various data sources (e.g. genomic, environmental and phenotypic data) and to accommodate nonlinear patterns and complex relationships among predictor variables, are of particular interest to overcome some limitations of mixed linear models.

Machine learning methods are particularly appealing in this context, because numerous learning algorithms are inherently nonlinear in nature, which means that no precise assumptions regarding the form of the nonlinear function needs to be specified before training. In the last decade, these statistical learning techniques have been more closely investigated in genetic studies for their potential ability to capture non-additive genetic effects, which can result in a viable improvement over traditional methods for some complex traits (Abdollahi-Arpanahi et al., 2020; Azodi et al., 2019; Gianola et al., 2011; González-Recio and Forni, 2011; Heslot et al., 2012b; Ogutu et al., 2011). Additionally, machine learning algorithms are expected to be able to learn complex interactions between genes and environmental conditions, without explicit modeling of all interaction terms Khaki and Wang (2019); Shook et al. (2021).

Nonetheless, an efficient usage of these methods remains challenging, notably because the explosion of environmental, high-throughput genotyping and phenotyping data leads to the well-known 'large p, small n' issue, meaning that the number of predictor variables (also called features), characterizing the data, is much larger than the sample size (Libbrecht and Noble, 2015). Multicollinearity,

overfitting and computational time represent some of the problems that need to receive attention when applying advanced machine learning techniques.

Machine learning techniques are generally classified as either supervised or unsupervised learning algorithms. Supervised learning methods make use of labels, such as classes (for classification problem) or numeric values (regression problem), that characterize each training instance and are used by the algorithm to learn the relationship linking the predictor variables and the label. Unsupervised learning techniques aim at recognizing patterns and at discovering groups of similar instances with unlabelled training data. Considering the diversity of existing machine learning algorithms, we will focus next on ensemble models for regression, since these models were applied in the thesis. We will also describe methods which can be used in general to prevent overfitting.

1.5.2 Ensemble models

Methods based on an ensemble of trees are averaging techniques that have been developed to capitalize on the advantages of single decision trees. A decision tree divides the predictor space into disjoint regions, by asking a series of questions to the data, as shown in **Figure 1.1**. The respective regions, defined by the terminal nodes of the tree, show a homogeneous response to the predictor variables. In this simple example, the outcome is given by a numeric value, but in more advanced models, the terminal node corresponds to a more complex function of the predictor variables. One of the most widely used method is the *classification and regression tree* (CART) conceived by Breiman et al. (1984). Decision trees can be used with different types of predictor variables (continuous, categorical, ordinal, etc), are able to perform feature selection with irrelevant variables, and can handle missing data by using surrogates (Hastie et al., 2009). However, single trees are relatively instable, meaning that when the data is even slightly modified, it might result in a substantial change in the structure of the decision tree, hence they are prone to overfitting resulting in larger prediction errors and are often referred as "weak learners".

Different techniques exist to combine a set of individual decision trees, among them *bagging*, *boosting* and *stacking*. Bagging (bootstrap aggregation) generates bootstrap samples from the original dataset, which are used to create a classifier or regression model, for instance based on a decision tree. The final prediction is given by the average of all single prediction models. Although this technique allows to reduce the overall variance by introducing a random component into the tree building procedure, the main issue is the correlation among decision trees because the same set of predictor variables is used at each split. A solution using the same fundamental principles as bagging has been proposed by Breiman (2001) with the random forests algorithm. In random forests, a random subset of m_{try} of the original p predictor variables is sampled within each bagged tree at each node to split, thus contributing to a reduction of the correlation among trees.

Boosting is an ensemble technique which aims at reducing both bias and variance. Let us consider

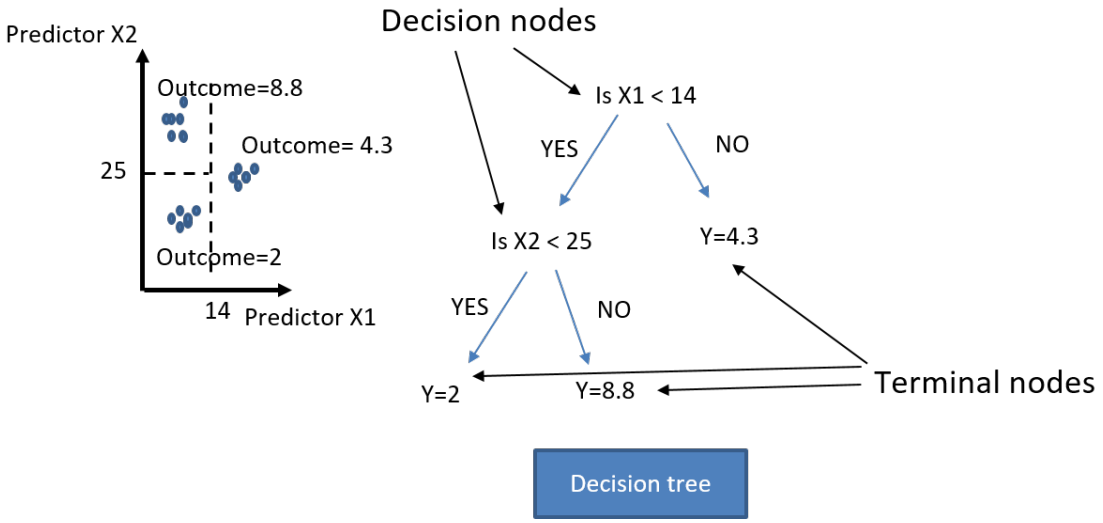


Figure 1.1: A regression tree fitting the average response for training instances within each rectangle of the predictor space.

here only the case of a regression task. Boosting is a forward and stagewise procedure, developed by Friedman (2001), and is described in **Algorithm 1**. At each iteration, a new tree is fit to the residuals of the previous trees, rather than to the original response variable, with the objective of minimizing the loss function using gradient descent. The gradient is calculated as the partial derivative of the loss function and is useful to help improve model parameters in order to obtain a better fit and a reduced error in the next iteration. After a new tree is fit, the residuals are updated by calculating the residuals from the model containing the ensemble of previous trees with the newly constructed one.

Thereby, it becomes clear that an emphasis is made at each iteration on the observations that are poorly predicted by the model, which explains why the boosted trees approach is different from random forests, for which the trees are independently built from each other. The final prediction is the sum of all trees weighted by the learning rate λ, also called shrinkage parameter. λ constitutes a hyperaparameter, which can be optimized using cross-validation. Other critical tuning hyperparameters with boosted trees are the number of trees and the tree depth, which is an important regularization parameter corresponding to the number of splits within each tree, and thus, it controls the interaction order within the data. Gradient boosting allows flexibility regarding the loss function to use. Mean squared error (MSE) is the most commonly used loss function for regression tasks, while hinge loss or logarithmic loss are frequently used for classification tasks. Popular gradient boosting libraries developed for R and Python are XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017), that we employed in **chapter 2**.

Other ensemble methods, such as stacked generalization, consist of combining the output obtained from different individual models. In a first layer of models, each individual base-learner (e.g. multiple linear regression, random forest model, etc) generates predictions, which are then directly

Algorithm 1 Boosting for Regression Trees from James et al. (2021)

Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the TRN

for b = 1, 2,..., B, **do** :

(a) Fit a tree $\hat{f}^b$ with d splits (d + 1 terminal nodes) to the TRN (X,r).

(b) Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

(c) Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

used as input to train a meta-learner (e.g. Lasso). Using cross-validation, optimal weights are assigned to each base learner by this meta-learner, that are used to obtain the final prediction. More details are given about stacked generalization in **chapter 3**.

1.5.3 Mitigating overfitting with machine learning techniques

1.5.3.1 Hyperparameter optimization

Hyperparameters are parameters of machine learning algorithms for which no analytical formula exists to estimate them directly from the TRN. Many of these parameters influence the complexity of the model and should be carefully chosen to regulate the extent to which the model adapts to very specific and noisy patterns in the training data. The classical approach, named *grid search*, implies the definition of a grid of candidate values, to test all possible combinations of hyperparameters via cross-validation and to finally select the combination of hyperparameters resulting in the lowest mean squared error. As emphasized by Kuhn et al. (2013) and Azodi et al. (2019), this tuning step should be performed using only the training data, while the TST is used to obtain an unbiased estimate of model performance. Azodi et al. (2019) demonstrated the importance of choosing adequate hyperparameters for nonlinear methods, showing for instance that shallower trees (low interaction depth) in random forests generally improve model performance. For artificial neural networks, the penalty method (L_1 or L_2), activation function and network architecture can considerably influence the model's predictive ability (Azodi et al., 2019; Bellot et al., 2018; Pérez-Enciso and Zingaretti, 2019).

Despite its simplicity, the grid search method is computationally intensive with advanced algorithms necessitating tuning numerous hyperparameters. Another possibility is to implement randomized search, which enables the evaluation of a user-defined number of random combinations,

with random sampling of each hyperparameter value at each iteration. In **chapter 2**, we describe another method we implemented to optimize hyperparameters based on a Gaussian process model.

1.5.3.2 Resampling the data via cross-validation

Cross-validation (CV) is the most widely used technique to estimate the average generalization error on an independent test sample. In genomic selection, the evaluation of the models is conducted within the set of genotyped and phenotyped individuals, that is randomly partitioned into a number of folds (k) of equal size. Data from (k-1) folds is used to train a predictive model, that is subsequently used to predict the remaining fold (i.e. the TST). Assessment metrics are calculated for each fold serving as a TST. This k-fold CV procedure has been the standard approach in genomic selection studies over the last two decades.

However, with the increasing application of machine learning, more advanced methodologies are required to prevent data leakage during the steps of data preprocessing/hyperaparameter optimization and of model assessment. Following recommendation made by Varma and Simon (2006) and Krstajic et al. (2014), who showed that nested CV procedures can significantly reduce the bias of error estimates, we apply this method in **chapter 2**, which involves the usage of two layers of resampling to separate the hyperparameter tuning procedure from the model assessment procedure, thus ensuring that the loss estimates are unbiased.

Additionally, different cross-validations might be of particular interest in the specific context of plant breeding. It is of utmost interest to reproduce real prediction problems that the breeders generally encounter. Therefore, using the same terminology as many related publications Burgueño et al. (2012); Jarquín et al. (2014, 2017), in **chapter 2** we investigated how well genotype performance can be predicted in a new year (i.e. no phenotypic data for the corresponding year is included in the TRN, CV0-year), or at a new location (i.e. no phenotypic data for the corresponding location is included in the TRN, CV0-location), which represent challenging prediction scenarios. The same scenarios were considered when the hybrids have never been evaluated before (CV00-year and CV00-location). In **chapter 4**, prediction of genotype performance in new year-location combinations (CV0-environment), of newly developed genotypes in already tested environments (CV1) and of incomplete field trials (CV2) were implemented. Results obtained by Jarquín et al. (2017) with wheat grown in Kansas indicate that moderately accurate results can be achieved to predict tested genotypes in new environments, but the task of predicting in new years, especially for new genotypes, was still infeasible. Yet, it should be noted that this study did not incorporate weather data, which can potentially lend the model useful information about the relationships between environments.

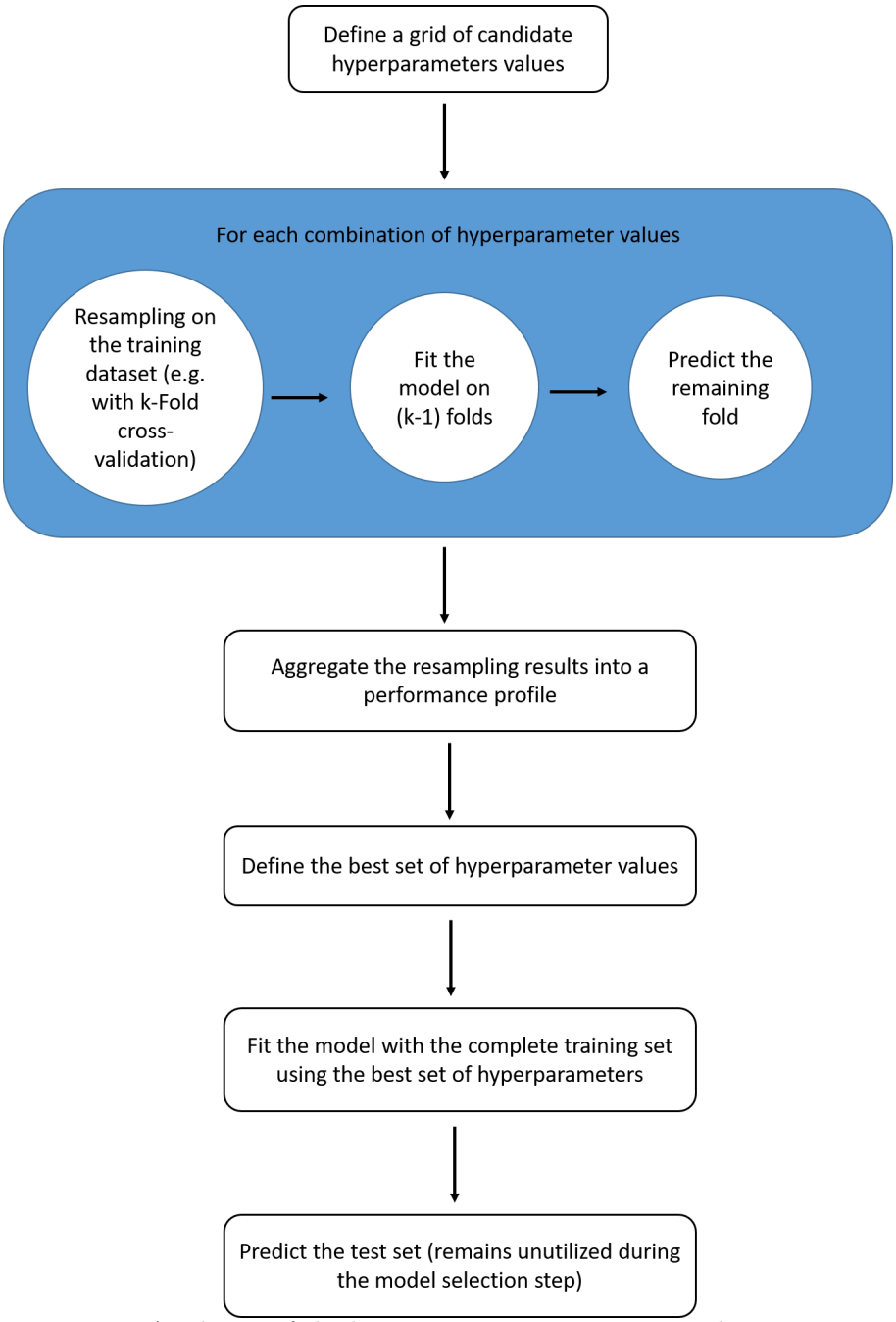


Figure 1.2: A scheme of the hyperparameter tuning procedure preventing data leakage, i.e. no test data is used for model selection (modified from Kuhn et al. (2013))

1.5.3.3 Variable selection and feature extraction

As the amount of gathered data is steadily growing, variable selection is becoming increasingly useful for facilitating the interpretation of the outputs of the machine learning models and to reduce data dimensionality, thereby accelerating training. From a statistical perspective, it is also advantageous to use fewer predictor variables explaining a larger variance of the outcome. Machine learning algorithms do not perform equally well with regard to multicollinearity. Tree-based methods are relatively robust to the inclusion of irrelevant variables (Hastie et al., 2009; James et al., 2021), while artificial neural networks cannot efficiently handle extremely large genotypic datasets (Azodi et al., 2019; Pérez-Enciso and Zingaretti, 2019), hence a step of feature selection should systematically be carried out. A range of methods, among which are filter, wrappers and embedded methods, can be employed to reduce the number of features in the model (Kuhn et al., 2013). Filter methods, such as Pearson's correlation coefficient, Spearman's rank correlation, Chi-Square or entropy-based features, generate a score for each variable and do not require using any learning algorithm, while providing stable results (Piles et al., 2021).

Principal component analysis (PCA) can be used to handle collinearity, which occurs when input variables are highly correlated with each other. The orthogonal new variables can then be employed as inputs for machine learning algorithms. This procedure can be of particular interest to reduce the dimensionality of the model's genomic components, and we applied it in **chapter 2** as well as in different models proposed in the package *learnMET* described in **chapter 3**.

1.6 Objectives of the thesis

Integrating genotypic and environmental data to predict phenotypes across multiple environments appears as a valuable strategy to make informed breeding decisions, and makes even more sense as we expect the occurrence of abiotic stresses to considerably increase in the next decades. Jarquín et al. (2014) proposed a reaction norm model based on the estimation of an environmental covariance matrix, providing an efficient computational framework in a LME framework. However, this predictive approach relies on strong statistical assumptions, which are: i) common variance for all markers based on the G-BLUP model, ii) common variance of the slopes of the reaction norms for all environmental covariates. Therefore, the model does not allow any heterogeneity regarding the variance among slopes of reaction norms. These statistical conditions can explain why the use of additional environmental data does not always result in substantial increases of predictive ability (De Los Campos et al., 2020; Jarquin et al., 2021), although this data does explain a larger proportion of the phenotypic variance.

While machine learning models are expected to capture implicit physical and biological relationships among these diverse predictor variables, only few studies, focusing on deep neural networks

(Khaki and Wang, 2019; Washburn et al., 2021b), investigated these techniques for the prediction of complex phenotypes. Furthermore, the latter rely on very little feature engineering prior to model training. Therefore, attention needs to be given to the performance of additional machine learning techniques and to the impact of processing environmental data with regards to ecophysiological knowledge.

Within the framework of the thesis, several public datasets have been utilized. In particular, we extensively used the publicly available datasets of the Genomes to Fields maize Initiative (AlKhalifah et al., 2018; McFarland et al., 2020). Briefly, a large number of unique maize hybrids ($\approx$ 2,000) was generated from a set of diverse inbred lines (e.g. recently expired plant variety protection elite lines, recombinant inbred lines) and their performance was measured at various locations across North America for several phenotypic traits (grain-related traits, plant height, ear height, and phenological traits). A particular emphasis of the project is put on the understanding of key G $\times$ Ecomponents, thus automatic weather stations were placed at each location. Starting in 2014, the project also aims at regularly releasing genotypic, environmental and phenotypic datasets characterizing the experimental trials managed by collaborators on their website (`https://www.genomes2fields.org/`).

Chapter 2 investigates the performance of two gradient boosting (GB) algorithms and linear random effects models as a benchmark for prediction of phenotypic traits, based on four different cross-validation scenarios mimicking concrete plant breeding prediction problems implemented with the G2F datasets. To borrow information from correlated environments due to marker $\times$ environmental covariates interaction effects, the latter were explicitly modeled in random effects models, while it was assumed that these G $\times$ E interactions can be inherently captured by the two GB algorithms. We also explored some machine learning tools to better understand the respective importance of the predictor variables to build predictions in the gradient boosting models.

Chapter 3 presents an R package which provides flexible pipelines to apply various types of machine learning methods for genomic prediction using multi-environment trial data. In particular, learnMET enables environmental characterization via the retrieval and aggregation of daily weather data, and different cross-validation schemes are proposed.

Chapter 4 examines a multivariate nonlinear method to build an environmental distance matrix on the basis of raw daily weather data characterizing the crop growing season within each environment. Using a wheat and a maize multi-environment dataset, we explore the potential of this method for environmental clustering and for predictive purposes with the model proposed by (Jarquín et al., 2014).

Chapter 5 includes a general discussion on factors which can influence the gain obtained by using environmental data along with genomic data for the prediction of genotype performance across environments.

1.7 Bibliography

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution 52(1):1–15

AlKhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh CT, Bohn M, et al. (2018) Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC Research Notes 11(1):1–5

Allard RW, Bradshaw AD (1964) Implications of genotype-environmental interactions in applied plant breeding 1. Crop science 4(5):503–508

Allen RG, Pereira LS, Raes D, Smith M, et al. (1998) Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. Fao, Rome 300(9):D05,109

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics 9(11):3691–3702

Basnet BR, Crossa J, Dreisigacker S, Pérez-Rodríguez P, Manes Y, Singh RP, Rosyara U, Camarillo-Castillo F, Murua M (2019) Hybrid wheat prediction using genomic, pedigree, and environmental covariables interaction models

Bellot P, de Los Campos G, Pérez-Enciso M (2018) Can deep learning improve genomic prediction of complex human traits? Genetics 210(3):809–819

Bernardo R (2002) Breeding for quantitative traits in plants, vol 1. Stemma press Woodbury, MN

Bernardo R (2014) Genomewide selection when major genes are known. Crop Science 54(1):68–75

Boer MP, Wright D, Feng L, Podlich DW, Luo L, Cooper M, van Eeuwijk FA (2007) A mixed-model quantitative trait loci (qtl) analysis for multiple-environment trial data using environmental covariables for qtl-by-environment interactions, with an example in maize. Genetics 177(3):1801–1813

Breiman L (2001) Random forests. Machine learning 45(1):5–32

Breiman L, Friedman J, Olshen R, Stone C (1984) Cart. Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science 52(2):707–719

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Cooper M, DeLacy I (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theoretical and Applied Genetics 88(5):561–572

Costa-Neto G, Fritsche-Neto R, Crossa J (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106

Crossa J, de los Campos G, Maccaferri M, Tuberosa R, Burgueño J, Pérez-Rodríguez P (2016) Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Science 56(5):2193–2209

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, De Los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, et al. (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends in plant science 22(11):961–975

De Los Campos G, Pérez-Rodríguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nature communications 11(1):1–10

Delerce S, Dorado H, Grillon A, Rebolledo MC, Prager SD, Patiño VH, Garcés Varón G, Jiménez D (2016) Assessing weather-yield relationships in rice at local scale using data mining approaches. PloS one 11(8):e0161,620

Eberhart St, Russell W (1966) Stability parameters for comparing varieties 1. Crop science 6(1):36–40

FAOSTAT (2020) Food and agriculture organization of the United Nations. FAOSTAT statistical database. `http://faostat.fao.org`. Accessed on December 22, 2021.

Finlay K, Wilkinson G (1963) The analysis of adaptation in a plant-breeding programme. Australian journal of agricultural research 14(6):742–754

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics pp 1189–1232

Gammans M, Mérel P, Ortiz-Bobea A (2017) Negative impacts of climate change on cereal yields: statistical evidence from france. Environmental Research Letters 12(5):054,007

Gauch Jr H, et al. (1992) Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier Science Publishers

Gauch Jr HG (2006) Statistical analysis of yield trials by ammi and gge. Crop science 46(4):1488–1500

Gianola D, Okut H, Weigel KA, Rosa GJ (2011) Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. BMC genetics 12(1):1–14

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution 43(1):1–12

Hammer GL, van Oosterom E, McLean G, Chapman SC, Broad I, Harland P, Muchow RC (2010) Adapting apsim to model the physiology and genetics of complex adaptive traits in field crops. Journal of experimental botany 61(8):2185–2202

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics, Springer, New York, NY

Hengl T, Nussbaum M, Wright MN, Heuvelink GB, Gräler B (2018) Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012a) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52(1):146, DOI 10.2135/cropsci2011.06.0297

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012b) Genomic selection in plant breeding: a comparison of models. Crop science 52(1):146–160

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics 127(2):463–480

Holzworth D, Huth N, Fainges J, Brown H, Zurcher E, Cichota R, Verrall S, Herrmann N, Zheng B, Snow V (2018) Apsim next generation: Overcoming challenges in modernising a farming systems model. Environmental Modelling Software 103:43–51, DOI https://doi.org/10.1016/j.envsoft.2018.02.002, URL https://www.sciencedirect.com/science/article/pii/S1364815217311921

Holzworth DP, Huth NI, deVoil PG, Zurcher EJ, Herrmann NI, McLean G, Chenu K, van Oosterom EJ, Snow V, Murphy C, et al. (2014) Apsim–evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62:327–350

Hurni S, Scheuermann D, Krattinger SG, Kessel B, Wicker T, Herren G, Fitze MN, Breen J, Presterl T, Ouzunova M, et al. (2015) The maize disease resistance gene htn1 against northern corn leaf blight encodes a wall-associated receptor-like kinase. Proceedings of the National Academy of Sciences 112(28):8780–8785

James G, Witten D, Hastie T, Tibshirani R (2021) Statistical learning. In: An introduction to statistical learning, Springer, pp 15–57

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics 127(3):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 x environment interactions in kansas wheat

Jarquin D, De Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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 p 1819

Jiang Y, Schmidt RH, Zhao Y, Reif JC (2017) A quantitative genetic framework highlights the role of epistatic effects for grain-yield heterosis in bread wheat. Nature genetics 49(12):1741–1746

Kashani MH, Dinpashoh Y (2012) Evaluation of efficiency of different estimation methods for missing climatological data. Stochastic Environmental Research and Risk Assessment 26(1):59–71

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30:3146–3154

Khaki S, Wang L (2019) Crop yield prediction using deep neural networks. Frontiers in Plant Science 10, DOI 10.3389/fpls.2019.00621, URL https://www.frontiersin.org/article/10.3389/fpls.2019.00621

Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of cheminformatics 6(1):1–15

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332, DOI 10.1038/nrg3920

Lin CS, Binns MR, Lefkovitch LP (1986) Stability analysis: where do we stand? 1. Crop science 26(5):894–900

Lizaso J, Ruiz-Ramos M, Rodríguez L, Gabaldon-Leal C, Oliveira J, Lorite I, Sánchez D, García E, Rodríguez A (2018) Impact of high temperatures in maize: Phenology and yield components. Field Crops Research 216:129–140

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(4):569–582

Ly D, Huet S, Gauffreteau A, Rincent R, Touzy G, Mini A, Jannink JL, Cormier F, Paux E, Lafarge S, et al. (2018) Whole-genome prediction of reaction norms to environmental stress in bread wheat (triticum aestivum l.) by genomic random regression. Field Crops Research 216:32–41

Malosetti M, Voltas J, Romagosa I, Ullrich S, Van Eeuwijk F (2004) Mixed models including environmental covariables for studying qtl by environment interaction. Euphytica 137(1):139–145

Malosetti M, Visser R, Celis-Gamboa C, Van Eeuwijk F (2006) Qtl methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. Theoretical and applied genetics 113(2):288–300

Malosetti M, Ribaut JM, Vargas M, Crossa J, Van Eeuwijk FA (2008) A multi-trait multi-environment qtl mixed model with an application to drought and nitrogen stress trials in maize (zea mays l.). Euphytica 161(1):241–257

Malosetti M, Ribaut JM, van Eeuwijk FA (2013) The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Frontiers in physiology 4:44

Martre P, Jamieson PD, Semenov MA, Zyskowski RF, Porter JR, Triboi E (2006) Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. European Journal of Agronomy 25(2):138–154

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, 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(1):1–6

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 TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019) Genomic prediction of maize yield across european environmental conditions. Nature genetics 51(6):952–956

Mital U, Dwivedi D, Brown JB, Faybishenko B, Painter SL, Steefel CI (2020) Sequential imputation of missing spatio-temporal precipitation data using random forests. Frontiers in Water 2:20

Monteith JL (1977) Climate and the efficiency of crop production in britain. Philosophical Transactions of the Royal Society of London B, Biological Sciences 281(980):277–294

Monteverde E, Gutierrez L, Blanco P, Pérez de Vida F, Rosas JE, Bonnecarrère V, Quero G, McCouch S (2019) Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (oryza sativa l.) grown in subtropical areas. G3: Genes, Genomes, Genetics 9(5):1519–1531

Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. In: BMC proceedings, BioMed Central, vol 5, pp 1–5

Pebesma E, Heuvelink G (2016) Spatio-temporal interpolation using gstat. RFID Journal 8(1):204–218

Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Computers & geosciences 30(7):683–691

Pérez-Enciso M, Zingaretti LM (2019) A guide on deep learning for complex trait genomic prediction. Genes 10(7):553

Pérez-Rodríguez P, Crossa J, Bondalapati K, De Meyer G, Pita F, Campos Gdl (2015) A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. Crop Science 55(3):1143–1151

Piles M, Bergsma R, Gianola D, Gilbert H, Tusell L (2021) Feature selection stability and accuracy of prediction models for genomic prediction of residual feed intake in pigs using machine learning. Frontiers in genetics 12:137

Rice B, Lipka AE (2019) Evaluation of rr-blup genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. The Plant Genome 12(1):180,052, DOI https://doi.org/10.3835/plantgenome2018.07.0052, URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.3835/plantgenome2018.07.0052, https://acsess.onlinelibrary.wiley.com/doi/pdf/10.3835/plantgenome2018.07.0052

Riedelsheimer C, Technow F, Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC genomics 13(1):1–9

Rincent R, Kuhn E, Monod H, Oury FX, Rousset M, Allard V, Le Gouis J (2017) Optimization of multi-environment trials for genomic selection based on crop models. Theoretical and Applied Genetics 130(8):1735–1752

Rincent R, Malosetti M, Ababaei B, Touzy G, Mini A, Bogard M, Martre P, Le Gouis J, van Eeuwijk F (2019) Using crop growth model stress covariates and ammi decomposition to better predict genotype-by-environment interactions. Theoretical and Applied Genetics 132(12):3399–3411

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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(2):jkaa050

Russell G, Jarvis P, Monteith J, et al. (1989) Absorption of radiation by canopies and stand growth. Plant canopies: their growth, form and function pp 21–39

Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK (2021) Crop yield prediction integrating genotype and weather variables using deep learning. Plos one 16(6):e0252,402

Sparks AH (2018) nasapower: a nasa power global meteorology, surface solar energy and climatology data client for r

Sukumaran S, Crossa J, Jarquin D, Lopes M, Reynolds MP (2017) Genomic prediction with pedigree and genotype× environment interaction in spring wheat grown in south and west asia, north africa, and mexico. G3: Genes, Genomes, Genetics 7(2):481–495

Tardieu F, Simonneau T, Muller B (2018) The physiological basis of drought tolerance in crop plants: a scenario-dependent probabilistic approach. Annual review of plant biology 69:733–759

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate bayesian computation. PloS one 10(6):e0130,855

Trnka M, Rötter RP, Ruiz-Ramos M, Kersebaum KC, Olesen JE, Žalud Z, Semenov MA (2014) Adverse weather conditions for european wheat production will become more frequent with climate change. Nature Climate Change 4(7):637–643

Van Diepen Cv, Wolf Jv, Van Keulen H, Rappoldt C (1989) Wofost: a simulation model of crop production. Soil use and management 5(1):16–24

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414–4423

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BMC bioinformatics 7(1):1–8

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES (2021a) Predicting phenotypes from genetic, environment, management, and historical data using cnns. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 134(12):3997—4011, DOI 10.1007/s00122-021-03943-7, URL https://doi.org/10.1007/s00122-021-03943-7

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES (2021b) Predicting phenotypes from genetic, environment, management, and historical data using CNNs. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 134(12):3997–4011, DOI 10.1007/s00122-021-03943-7, place: Germany

Widener S, Graef G, Lipka AE, Jarquin D (2021) An assessment of the factors influencing the prediction accuracy of genomic prediction models across multiple environments. Frontiers in Genetics 12

de Wit A, Boogaard H, Fumagalli D, Janssen S, Knapen R, van Kraalingen D, Supit I, van der Wijngaart R, van Diepen K (2019) 25 years of the wofost cropping systems model. Agricultural Systems 168:154–167

Yates F, Cochran WG (1938) The analysis of groups of experiments. The Journal of Agricultural Science 28(4):556–580

Zhao C, Liu B, Piao S, Wang X, Lobell DB, Huang Y, Huang M, Yao Y, Bassu S, Ciais P, et al. (2017) Temperature increase reduces global yields of major crops in four independent estimates. Proceedings of the National Academy of Sciences 114(35):9326–9331

2. 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, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[2]Center for Integrated Breeding Research, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[3]Kleinwanzlebener Saatzucht (KWS) SAAT SE, Einbeck, Germany

[4]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Albrecht-Thaer-Weg 3, 37075, Goettingen, Germany

* Email: cathy.jubin@uni-goettingen.de

Published in *Frontiers in Plant Science*
https://doi.org/10.3389/fpls.2021.699589

2.1 Abstract

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

2.2 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 (Crossa et al., 2021; Ersoz et al., 2020; Tardieu et al., 2017). 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 (G×E), almost omnipresent in multi-environment trial (MET) experiments (Chenu, 2015; Cooper and

DeLacy, 1994). 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., 2020) 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 G×E mixed effects models has been applied on a large number of datasets in different species (De Los Campos et al., 2020; Jarquín et al., 2017; Monteverde et al., 2019; Pérez-Rodríguez et al., 2015, 2017; Rincent et al., 2019; Sukumaran et al., 2017, 2018). 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., 2014; 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 G×E 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 (Costa-Neto et al., 2021; Heslot et al., 2014; Jarquín et al., 2014; Malosetti et al., 2016; Millet et al., 2019; Monteverde et al., 2019) 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 (Behravan et al., 2018; Lampa et al., 2014). 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; Géron, 2019; Hastie et al., 2009). Broad applications of these approaches include human disease prediction (Fukuda et al., 2013; Kopitar et al., 2020; Romagnoni et al., 2019), bioinformatics (Yu et al., 2019), ecology (Elith et al., 2008; Moisen et al., 2006) and agricultural forecasting (Crane-Droesch, 2018; Delerce et al., 2016; Fukuda et al., 2013; Jeong et al., 2016; 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 (Abdollahi-Arpanahi et al., 2020; Azodi et al., 2019; González-Recio et al., 2013; Ogutu et al., 2011). 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; Géron, 2019; Hastie et al., 2009).

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 (Butler and Huybers, 2015; Cicchino et al., 2010; 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 (Mushore et al., 2017; Schlenker and Roberts, 2009). 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.

2.3 Material and Methods

2.3.1 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 S2.1, Supplementary Table S2.1) after having eliminated environments with critical missing information, such as flowering time (Supplementary Table S2.2). 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.

2.3.2 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).

2.3.3 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 (Estévez et al., 2011; Zahumenskỳ, 2004) and are detailed in Supplementary Table S2.3. 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 (Gräler et al., 2016; Pebesma, 2004). 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), the mean vapor pressure deficit, the reference evapotranspiration (ET0) 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.

2.3.4 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 (Delerce et al., 2016; Gillberg et al., 2019; Heslot et al., 2014; 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 2.1). 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.

Table 2.1: Environmental predictor variables used in the prediction models. The suffixes refer to: V, vegetative period; F, flowering period; G, grain fill period; SC, soil covariate.

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 (%)

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

2.3.5 Prediction Models Implemented

2.3.5.1 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, G×E 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 × Year combination.

Main effects models

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

The response variable is modelled 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}:

$$y_{ij} = \mu + E_i + g_j + \varepsilon_{ij}, \tag{2.3.1}$$

where $E_i \overset{IID}{\sim} N(0, \sigma^2_E)$, $\mathbf{g} \overset{IID}{\sim} N(\mathbf{0}, \mathbf{G}\sigma^2_g)$ and $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2_\varepsilon)$, and N(.,.) denotes a normally distributed random variable, IID stands for independent and identically distributed, and σ^2_E, σ^2_g 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 $b_m \overset{IID}{\sim} N(0, \sigma^2_b)$, *m = 1,...,p* . The vector $\mathbf{g}$=$\mathbf{Xb}$ follows a multivariate normal density with null mean and covariance-matrix $Cov(\mathbf{g}) = \mathbf{G}\sigma^2_g$, 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:

$$y_{kj} = \mu + S_k + g_j + \varepsilon_{kj} \tag{2.3.2}$$

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

(3) Model G+E+W: Marker + Environment + 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 ($\mathbf{W}$), 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. $\gamma_q \sim N(0, \sigma^2_\gamma)$. Consequently, the vector $\mathbf{w} = \mathbf{W}\boldsymbol{\gamma}$ follows a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{\Omega}\sigma^2_w$, where $\boldsymbol{\Omega} \propto \mathbf{W}\mathbf{W}'$, and the matrix $\mathbf{W}$, which is centered and standardized, contains the values of the ECs. The model becomes then:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + \varepsilon_{ij} \tag{2.3.3}$$

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

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 Ω and $\mathbf{G}$ allows to borrow information between environments and between hybrid genotypes, respectively.

Models with interaction

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

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

$$y_{ij} = \mu + E_i + g_j + gE_{ij} + \varepsilon_{ij} \tag{2.3.4}$$

with $\mathbf{gE} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g'}] \circ [\mathbf{Z_E}\mathbf{Z_E'}]\sigma^2_{gE}), \varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2_\varepsilon)$, 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^2_{gE} is the variance component of the gE_{ij} interaction term; and $\circ$ denotes the Hadamard product between two matrices.

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

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

$$y_{kj} = \mu + S_k + g_j + gS_{kj} + \varepsilon_{kj} \tag{2.3.5}$$

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

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

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

$$y_{jkm} = \mu + g_j + S_k + Y_m + E_{km} + gS_{jk} + gY_{jm} + gE_{jkm} + \varepsilon_{jkm} \tag{2.3.6}$$

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

(7) Model G+E+W+G× W: 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. $\mathbf{\Omega}$). The vector of random effects, denoted $\mathbf{gw}$ represents the interaction terms between markers and ECs, is assumed to follow a multivariate normal distribution with null mean and covariance structure $[\mathbf{Z_g G Z_g'}] \circ \mathbf{\Omega}$. The model can be expressed as follows:

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + \varepsilon_{ij}, \tag{2.3.7}$$

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

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

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

$$y_{ij} = \mu + E_i + g_j + w_{ij} + gw_{ij} + gE_{ij} + \varepsilon_{ij} \tag{2.3.8}$$

Main and interactions effects included in the different models described above are summarized in Table S5. 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 data. 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.

2.3.5.2 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.

2.3.5.3 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).

2.3.5.4 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 (Snoek et al., 2012; Williams and Rasmussen, 2006). 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 S2.4 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 2.1. 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.

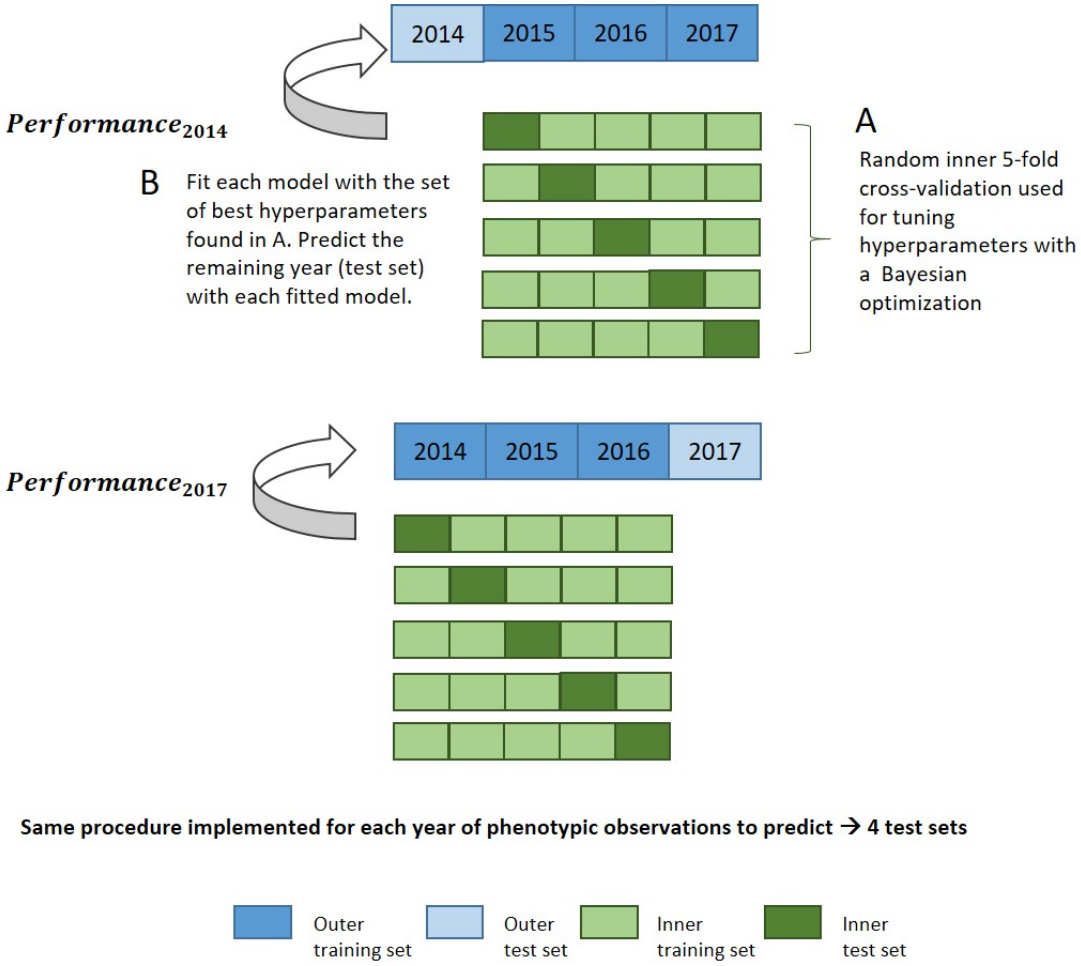


Figure 2.1: Nested cross-validation diagram for evaluation of model performance in the leave-1-year-out CV scheme with a machine learning approach.

2.3.5.5 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 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.

2.3.6 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 represents the complement set of other predictor variables used in the model.

The following partial function can be used as an estimator:

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

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 XS 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)*.

2.3.7 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`.

2.4 Results

2.4.1 Variability of Climatic Conditions in the Panel of Environments

Figure 2.2 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.

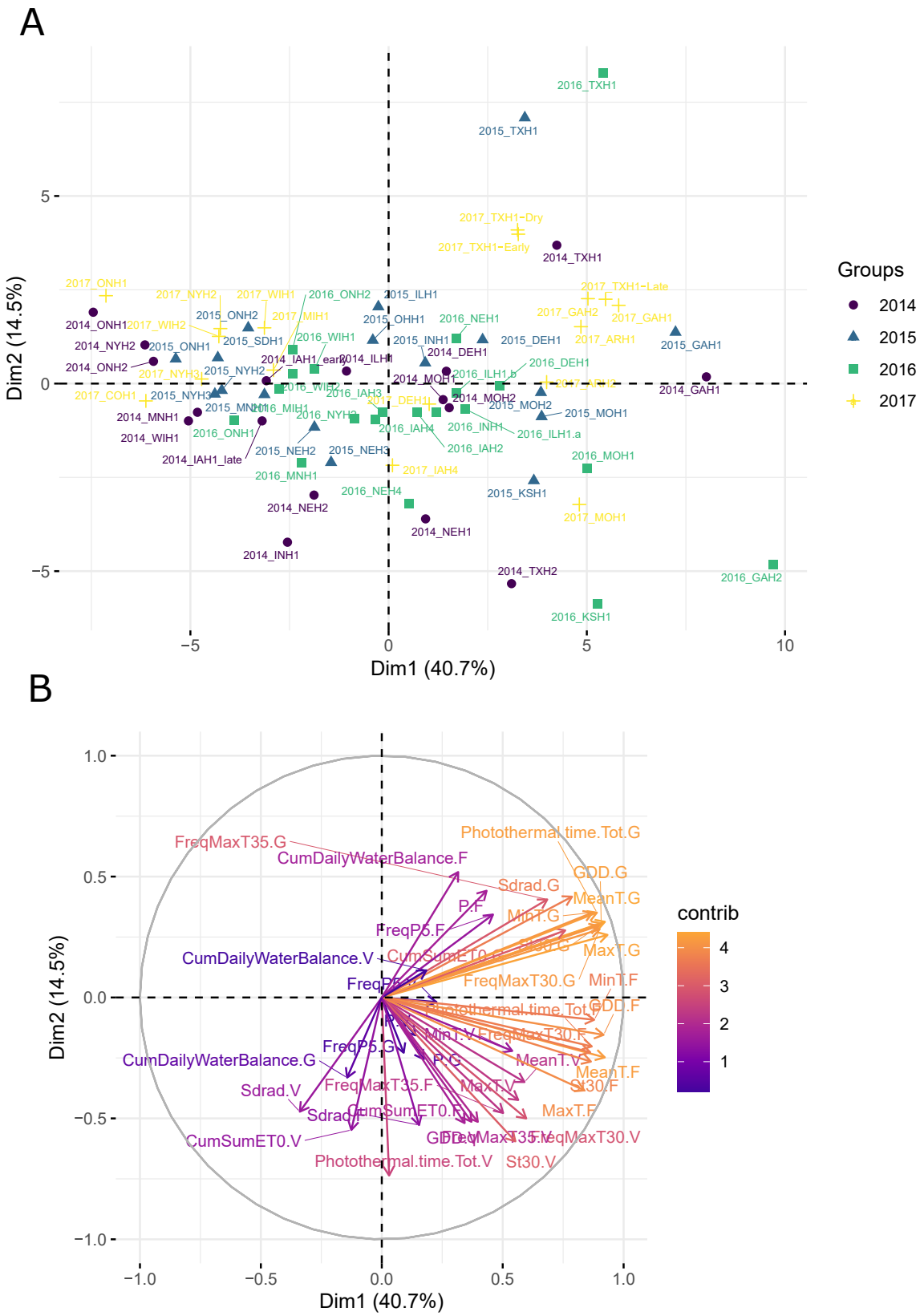


Figure 2.2: Principal component analysis (PCA) plot of environmental data from the 71 environments, using the median flowering date as reference in each environment. **(A)** Maize trial experiments located in the US and in Canada used in analyses. Name of the locations and their geographical position are given in Supplementary Table S2.1. **(B)** Correlation plot of the weather-based covariates used in the PCA.

2.4.2 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 S2.3). 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.

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

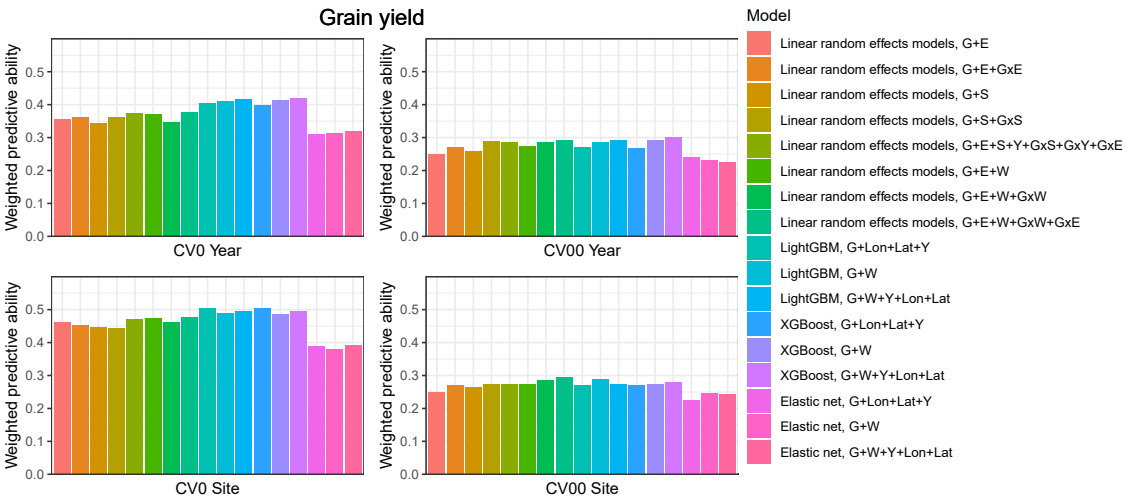


Figure 2.3: Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait grain yield. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; G×S, genotype-by-site interaction; E, environment effect; G×Y, genotype-by-year interaction; G×S, genotype-by-site interaction; G×E, genotype-by-environment interaction; G×W, 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.

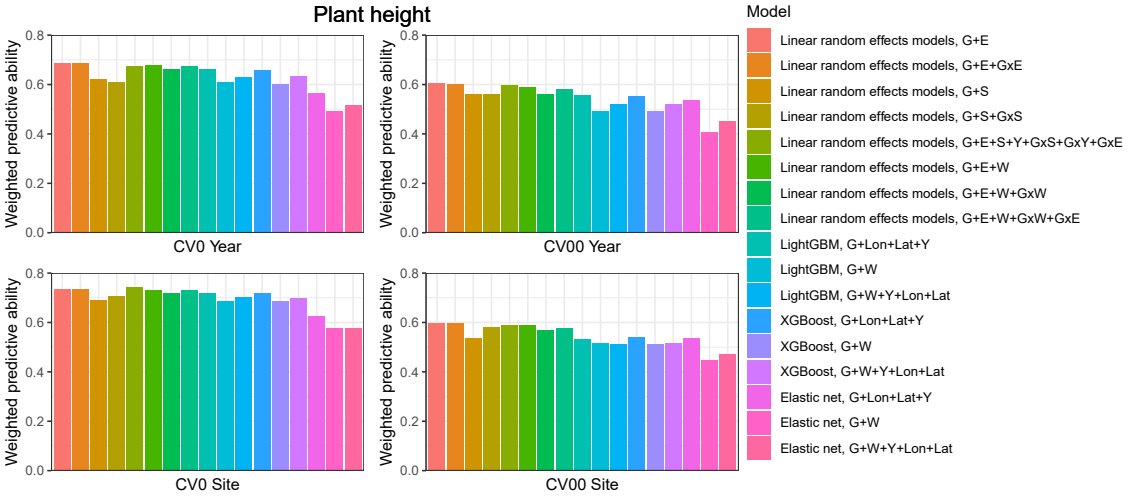


Figure 2.4: Weighted average predictive ability across 71 environments obtained for four cross-validation schemes and 16 models for the trait plant height. G, main effect of SNPs markers (genomic relationship matrix for LRE models; principal components derived from marker matrix for machine learning-based approaches); Y, year effect; S, site effect; G×S, genotype-by-site interaction; E, environment effect; G×Y, genotype-by-year interaction; G×S, genotype-by-site interaction; G×E, genotype-by-environment interaction; G×W, 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.

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 2.3; Supplementary Table S2.6). When the G×E term was added, the average correlation improved to 0.362. The model that included all interactions (G+E+W+G×W+G×E) was the best LRE model, while using only interactions between environmental covariates and genomic information (model G+E+W+G×W) 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 model performed best (Figure 2.4; Supplementary Table S2.8), 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 com-

pared 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 2.3; Supplementary Table S2.6), modeling the effect of sites instead of environments resulted in a small improvement of the predictive ability (% better than the G+E model). Adding the G×E 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+G×S) 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 S2.8), 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 2.3 and 2.4; Supplementary Tables S2.7 and S2.9). Indeed, based on our dataset, which covers many different sites across the US (see Supplementary Figure S2.1), 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 2.3; Supplementary Table S2.7), 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 2.4; Supplementary Table S2.9), LRE models performed better than machine learning-based methods, with the model G+E+S+Y+G×S+G×Y+G×E, 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 2.3; Supplementary Table S2.7), 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 G×E 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+G×W+G×E) yielded an improvement of the baseline model by % (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 2.4; Supplementary Table S2.9), the baseline model with interactions by environment (G+E+G×E) 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 2.5; Supplementary Tables S2.10 and S2.11 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 than the less complex one in only 34 environments. This can be interpreted as a failure to explain a large part of the G×E by the computed ECs, and by a more efficient representation of environmental effects by simple geographic information.

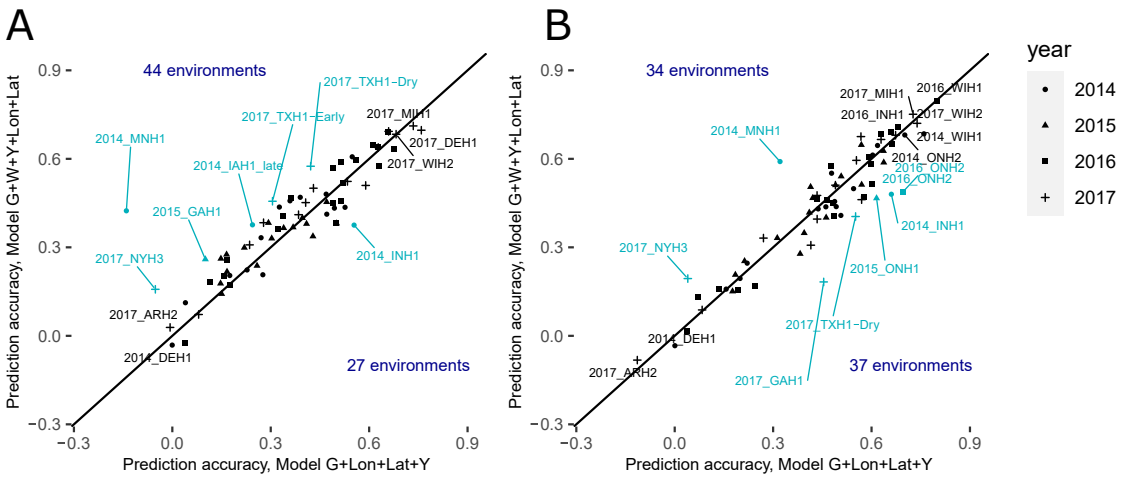


Figure 2.5: | Comparison of the within-environment predictive ability with different sets of predictors for the trait grain yield for XGBoost **(A)** with the CV0-Year scenario and **(B)** CV0-Site scenario. The x-axis corresponds to the within-environment correlation obtained with the model incorporating 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.

2.4.4 Variable Importance

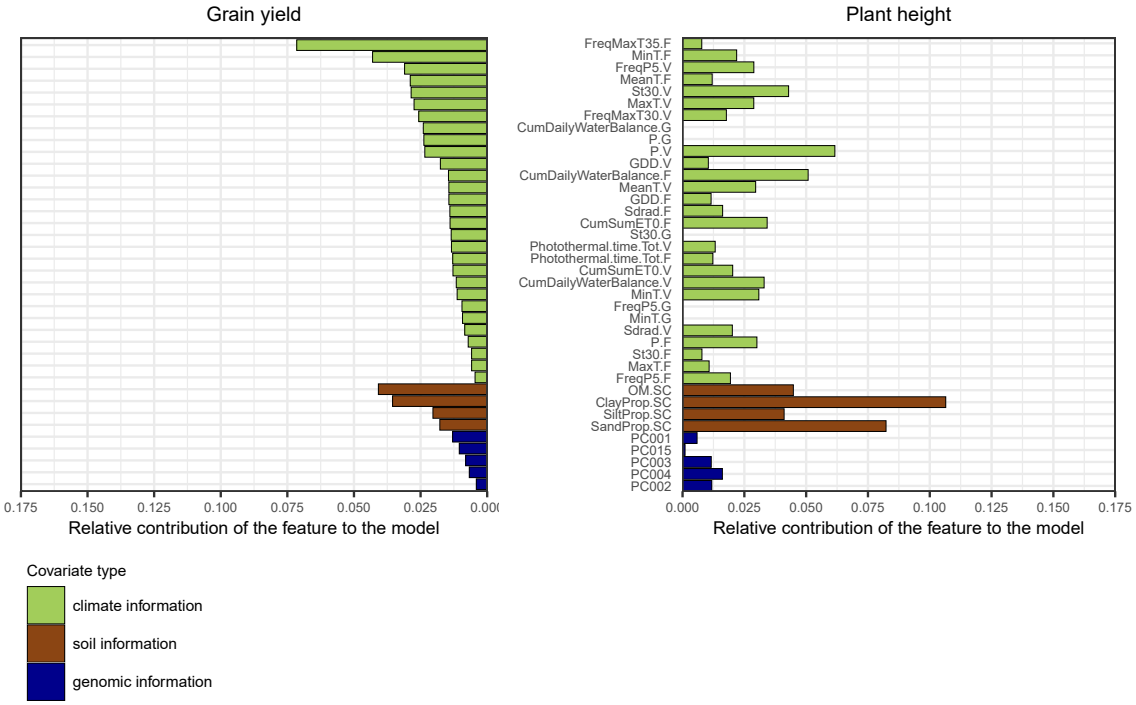


Figure 2.6: Feature importance ranking based on the average relative gain per feature obtained with the model XGBoost-G+W, for the two traits grain yield and plant height. The metric was estimated using a model fitted on the full dataset. The gain represents the improvement in accuracy when using a feature 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.

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 frequency of days during which the maximum temperature was above 35°C (Figure 2.6). 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.

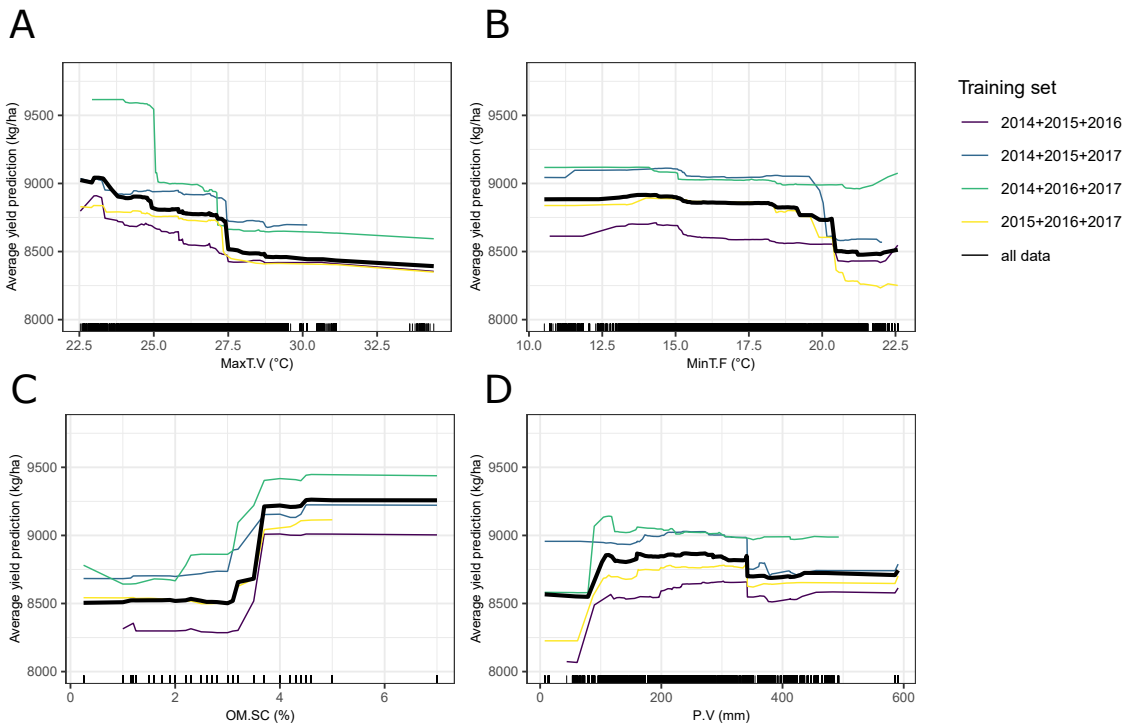


Figure 2.7: Partial dependence plots (PDPs) showing the behavior of the expected value of predicted yield as a function of four top-ranked predictor variables. The Y-axis value of a PDP is calculated average of all model predictions obtained from the training dataset, when the value of the predictor 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.

Partial dependence plots (Figure 2.7) 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.

2.5 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 (De Los Campos et al., 2020; Jarquín et al., 2017). 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.

2.5.1 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 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 G×E variance component for the phenotypic variance of grain yield, while for plant height, the proportion of variance explained by G×E 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 (Li et al., 2019; Troy et al., 2015). 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. (2014) 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 (Costa-Neto et al., 2021; Cuevas et al., 2018; Bandeira e Sousa et al., 2017). 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. More recently, Costa-Neto et al. (2021) 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.

2.5.2 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. (2021); 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. (2021) 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 S2.6, S2.7, S2.8, S2.9). 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 S2.10, S2.11). 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 S2.10), 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 (Supplementary Table S2.10) and CV0-Site (Supplementary Table S2.11), 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.

2.5.3 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 2.6), 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. (2020) 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 (Bustos-Korts et al., 2019; Heslot et al., 2014; Rincent et al., 2017, 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).

2.5.4 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 (Brock et al., 2021; Kagawa et al., 2017; Roe et al., 2020). 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; Lobell et al., 2014; Mimić et al., 2020; Schlenker and Roberts, 2009; Zhu et al., 2019). 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 S2.2) 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 (Kuhn et al., 2013; Ridgeway, 2007). 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 and Prueger, 2015; Hatfield et al., 2011). 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 (Cakir, 2004; Denmead and Shaw, 1960), 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.

2.5.5 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 G×E 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.

2.6 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.

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.

Acknowledgements

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

Figure S2.1: Maps of the experimental trials used in this study (from original Genomes To Fields Initiative datasets). Sample size designates the number of phenotypic observations for grain yield. Some points gather several experiments at very close distance from each other.

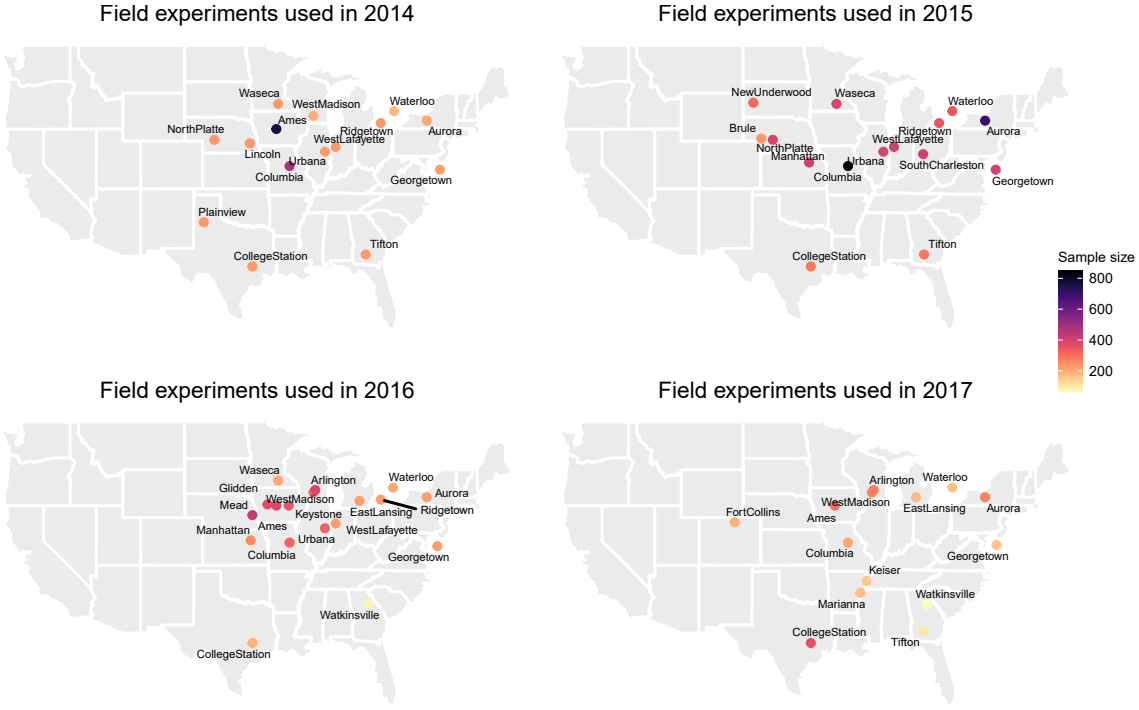


Figure S2.2: Pearson's coefficients of correlation within and across weather- and soil-based environmental predictors across 71 environments and 18,325 phenotypic observations. Non-significant coefficients ($P < 0.01$) were left blank; please refer to Table 1 for the abbreviations of the environmental predictors.

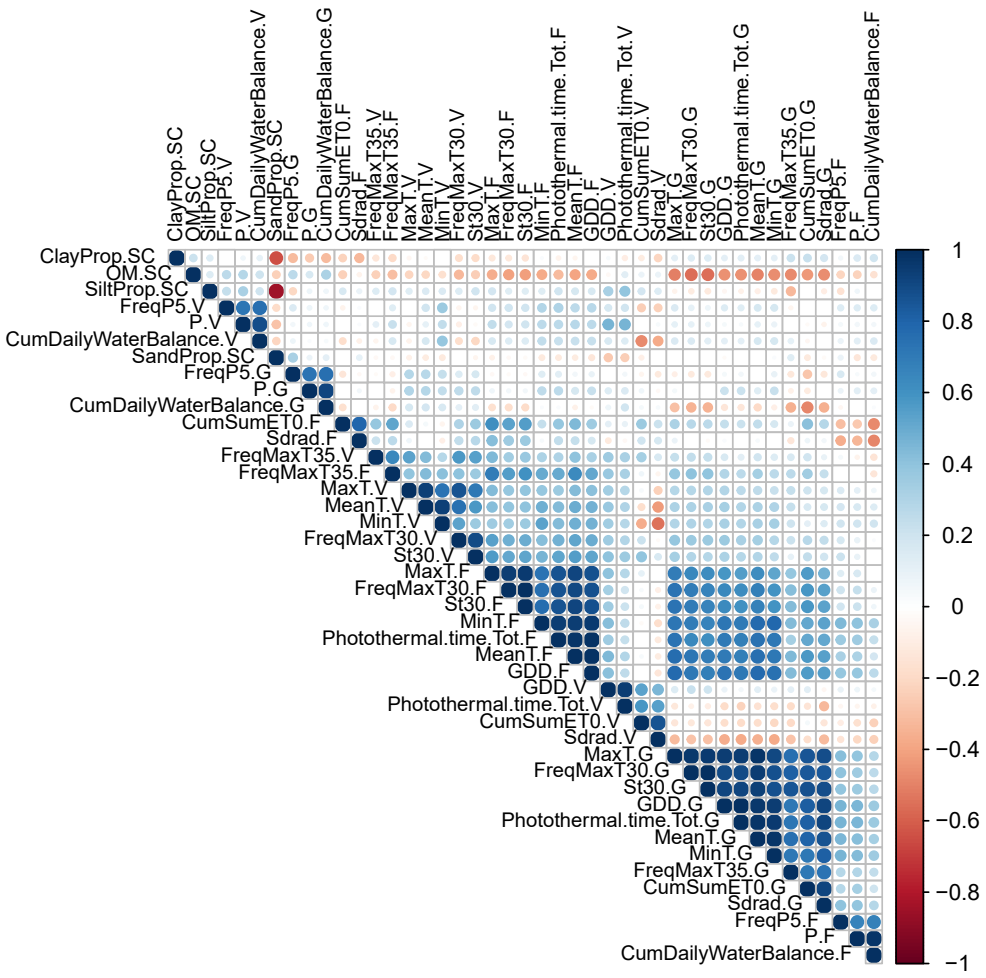


Figure S2.3: Boxplots showing the most important hyperparameters and interactions between them for **(A)** LightGBM and **(B)** XGBoost, using performance metrics obtained in the leave-one-year-out CV schemes (CV0-Year and CV00-Year). Hyperparameter importance values were obtained from fitting a random forest model on performance data obtained with various hyperparameter settings, followed by a functional ANOVA analysis.

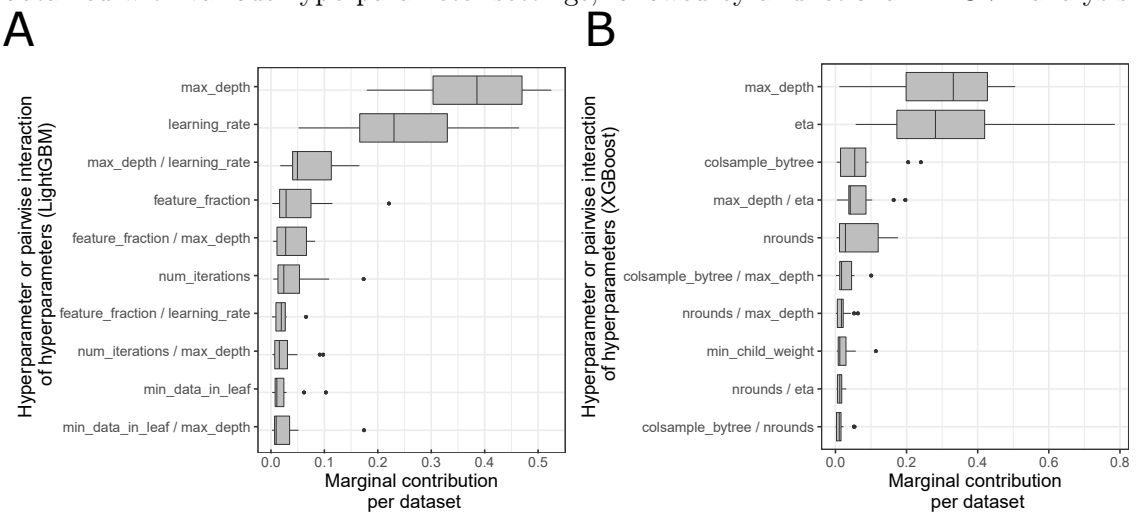


Table S2.1: Seventy-one trial experiments from the Genomes to Fields datasets used in analyses, with the location, the university factor (used to determine the training/test splits in the CV0-Site and CV00-Site CV schemes), longitude (Lon), latitude (Lat), the planting and harvest dates (in day of year) and the irrigation status of the field experiment. Not all of the environments originally present in the Genomes to Fields datasets were eventually incorporated in the prediction analyses (see Table S2.2).

ID_Experiment	Location, US State	Site/University factor	Lon	Lat	Planting Date (DOY)	Harvest Date (DOY)	Irrigated (and tracked)	number of yield records used	number of plant height records used
2014_DEH1	Georgetown1, DE	UniversityDelaware	-75.20	38.64	125	272	yes	224	224
2014_GAH1	Tifton, GA	UniversityGeorgia1	-83.56	31.51	94	254	yes	237	236
2014_IAH1_early	Ames, IA	IowaStateUniversity1	-93.70	42.00	129	293	no	131	131
2014_IAH1_late	Ames, IA	IowaStateUniversity1	-93.70	42.00	137	294	no	634	619
2014_ILH1	Urbana, IL	UniversityIllinois	-88.23	40.06	126	280	no	233	229
2014_INH1	WestLafayette, IN	PurdueUniversity	-87.01	40.49	145	322	no	237	236
2014_MNH1	Waseca, MN	UniversityMinnesota	-93.53	44.07	136	289	no	232	230
2014_MOH1	Columbia, MO	UniversityMissouri	-92.21	38.90	127	295	yes	244	243
2014_MOH2	Columbia HinksonCreek, MO	UniversityMissouri	-92.35	38.93	125	317	no	240	239
2014_NEH1	Lincoln, NE	UniversityNebraska	-96.66	40.83	136	295	no	242	242
2014_NEH2	North Platte, NE	UniversityNebraska3	-100.75	41.05	135	309	no	237	234
2014_NYH2	Aurora, NY	CornellUniversity	-76.65	42.73	148	336	no	212	208
2014_ONH1	Waterloo, ON	GuelphUniversity2	-80.43	43.50	139	308	no	181	179
2014_ONH2	Ridgetown, ON	GuelphUniversity1	-81.88	42.45	147	333	no	240	239
2014_TXH1	CollegeStation, TX	TexasAMUniversity1	-96.43	30.55	60	233	yes (not tracked)	236	235
2014_TXH2	Plainview, TX	TexasAMUniversity2	-101.95	34.18	113	273	yes (not tracked)	243	242
2014_WIH1	WestMadison, WI	UniversityWisconsin1	-89.53	43.06	129	301	no	200	197
2015_DEH1	Georgetown2, DE	UniversityDelaware	-75.47	38.63	119	257	yes	407	407
2015_GAH1	Tifton, GA	UniversityGeorgia1	-83.56	31.51	91	238	yes	295	294
2015_ILH1	Urbana, IL	UniversityIllinois	-88.23	40.06	120	275	no	393	392
2015_INH1	WestLafayette, IN	PurdueUniversity	-87.00	40.48	134	288	no	414	414
2015_KSH1	Manhattan1, KS	KansasStateUniversity	-96.61	39.22	113	264	no	408	
2015_MNH1	Waseca, MN	UniversityMinnesota	-93.53	44.07	139	314	no	395	394
2015_MOH1	Columbia, MO	UniversityMissouri	-92.21	38.90	124	279	no	411	410
2015_MOH2	Columbia, MO	UniversityMissouri	-92.21	38.90	125	275	no	411	409
2015_NEH2	North Platte, NE	UniversityNebraska3	-100.75	41.05	113	299	no	394	
2015_NEH3	Brule, NE	UniversityNebraska2	-101.99	41.16	161	348	no	242	
2015_NYH2	Aurora, NY	CornellUniversity	-76.65	42.73	127	320	no	347	344
2015_NYH3	Aurora, NY	CornellUniversity	-76.66	42.72	143	322	no	348	348
2015_OHH1	South Charleston, OH	OhioStateUniversity	-83.66	39.86	141	290	no	413	412
2015_ONH1	Waterloo, ON	GuelphUniversity2	-80.45	43.50	120	288	no	345	345
2015_ONH2	Ridgetown, ON	GuelphUniversity1	-81.88	42.45	127	285	no	350	349
2015_SDH1	New Underwood, SD	SouthDakotaUniversity	-102.93	44.21	142	301	no	317	316
2015_TXH1	College Station, TX	TexasAMUniversity1	-96.43	30.55	66	209	no	290	287
2016_DEH1	Georgetown2, DE	UniversityDelaware	-75.45	38.65	116	258	no	227	225
2016_GAH2	Watkinsville, GA	UniversityGeorgia2	-83.31	33.72	146	286	no	84	84
2016_IAH2	Glidden, IA	IowaStateUniversity2	-94.73	42.07	116	285	no	375	372
2016_IAH3	Keystone, IA	IowaStateUniversity3	-92.26	41.99	115	280	no	352	351
2016_IAH4	Ames, IA	IowaStateUniversity1	-93.70	42.00	117	291	no	388	384
2016_ILH1.a	Urbana, IL	UniversityIllinois	-88.23	40.06	127	283	no	125	125
2016_ILH1.b	Urbana, IL	UniversityIllinois	-88.23	40.06	117	283	no	201	201
2016_INH1	WestLafayette, IN	PurdueUniversity	-86.99	40.48	140	280	no	224	221
2016_KSH1	Manhattan2, KS	KansasStateUniversity	-96.63	39.14	106	271	no	262	262

2016_MIH1	EastLansing Powrline, MI	MichiganStateUniversity	-84.30	42.41	145	321	no	227	227
2016_MNH1	Waseca, MN	UniversityMinnesota	-93.53	44.07	138	294	no	216	
2016_MOH1	Columbia, MO	UniversityMissouri	-92.21	38.89	144	281	no	328	326
2016_NEH1	Mead, NE	UniversityNebraska4	-96.42	41.17	127	313	no	225	225
2016_NEH4	Mead, NE	UniversityNebraska4	-96.42	41.17	159	314	no	211	211
2016_NYH2	Aurora, NY	CornellUniversity	-76.66	42.73	131	343	no	228	226
2016_ONH1	Waterloo, ON	GuelphUniversity2	-80.45	43.50	125	289	no	207	206
2016_ONH2	Ridgetown, ON	GuelphUniversity1	-81.88	42.45	132	306	no	223	222
2016_TXH1	CollegeStation, TX	TexasAMUniversity1	-96.43	30.55	64	218	no	197	193
2016_WIH1	WestMadison, WI	UniversityWisconsin1	-89.53	43.06	130	288	no	303	300
2016_WIH2	Arlington, WI	UniversityWisconsin2	-89.34	43.33	145	299	no	395	392
2017_ARH1	Marianna, AR	ArkansasStateUniversity1	-90.76	34.73	107	254	yes	178	175
2017_ARH2	Keiser, AR	ArkansasStateUniversity2	-90.07	35.67	115	259	yes	160	159
2017_COH1	FortCollins, CO	ColoradoStateUniversity	-105.00	40.65	151	326	no	199	195
2017_DEH1	Georgetown2, DE	UniversityDelaware	-75.43	38.67	118	251	no	163	162
2017_GAH1	Tifton, GA	UniversityGeorgia1	-83.56	31.51	94	250	no	107	106
2017_GAH2	Watkinsville, GA	UniversityGeorgia2	-83.30	33.73	122	251	no	72	72
2017_IAH4	Ames, IA	IowaStateUniversity1	-93.69	41.99	127	290	no	310	307
2017_MIH1	EastLansing, MI	MichiganStateUniversity	-84.49	42.68	142	293	no	180	179
2017_MOH1	Columbia, MO	UniversityMissouri	-92.20	38.89	135	292	no	217	216
2017_NYH2	Aurora, NY	CornellUniversity	-76.65	42.73	138	328	no	187	185
2017_NYH3	Aurora, NY	CornellUniversity	-76.65	42.73	138	328	no	88	88
2017_ONH1	Waterloo, ON	GuelphUniversity2	-80.43	43.50	137	304	no	171	169
2017_TXH1-Dry	CollegeStation, TX	TexasAMUniversity1	-96.43	30.55	62	206	no	112	111
2017_TXH1-Early	CollegeStation, TX	TexasAMUniversity1	-96.43	30.55	62	212	no	146	146
2017_TXH1-Late	CollegeStation, TX	TexasAMUniversity1	-96.43	30.55	96	222	no	109	107
2017_WIH1	WestMadison, WI	UniversityWisconsin1	-89.53	43.06	125	292	no	266	261
2017_WIH2	Arlington, WI	UniversityWisconsin2	-89.34	43.32	131	310	no	279	276

Table S2.2: List of discarded environments from the original G2F phenotypic datasets, with the reason justifying their removal

Year_Exp	Reason
2014_IAH2	no flowering date
2014_IAH3	no flowering date
2014_IAH4	no flowering date
2014_NCH1	no flowering date
2014_NEH3	no flowering date
2014_NYH1	no silking date
2015_NCH1	no flowering date
2015_NEH1	no silking date
2015_NEH4	no silking date
2015_NYH1	disease field treatment
2015_TXH2	no location nor geographical coordinates in metadata file
2015_WIH1	no flowering date
2015_WIH2	no flowering date
2016_ARH1	many recorded dates with field irrigation but no amount tracked in the metadata
2016_ARH2	many recorded dates with field irrigation but no amount tracked in the metadata
2016_GAH1	many recorded dates with field irrigation but no amount tracked in the metadata
2016_IAH1	no flowering date
2016_NCH1	no flowering date
2016_NYH1	disease field treatment
2016_NYH3	no yield data
2016_OHH1	no flowering date
2016_SCH1	no yield data
2016_TXH2	no location nor geographical coordinates in metadata file
2017_IAH1	no flowering date
2017_IAH2	no flowering date
2017_IAH3	no flowering date
2017_ILH1	no location nor geographical coordinates in metadata file
2017_INH1	no location nor geographical coordinates in metadata file
2017_MNH1	based on comments regarding low phenotypic quality from metadata file
2017_NCH1	no flowering date
2017_NEH3	no flowering date
2017_NEH4	no flowering date
2017_NYH1	disease field treatment
2017_OHH1	no flowering date
2017_ONH2	strange values for flowering date (FW 30 days after sowing date?)
2017_SCH1	no yield data
2017_TXH2	no location nor geographical coordinates in metadata file

Table S2.3: Quality control applied to daily and semi-hourly weather data. Data which did not meet the specified requirements were flagged and assigned as missing values.

Validation procedure	**Air temperature (°C)**	**Precipitation (mm)**	**Relative humidity (%)**
Range test	-40 $\leq$ T $\leq$ 60	0 $\leq$ P_{sh} $\leq$ 120 (Estévez et al., 2011) 0 $\leq$ P $\leq$ 250	0 $\leq$ RH $\leq$ 100
Persistence tests	var(T_{sh}) $\neq$ 0		var(RH_{sh}) $\neq$ 0
Number of records per day	n(T_{sh})= {24,48,72,96}	n(P_{sh})= {24,48,72,96}	n(RH_{sh})= {24,48,72,96}
Percentage of missing data	r(T) $\geq$ 0.9	r(P) $\geq$ 0.9	r(RH) $\geq$ 0.9
Internal consistency tests	T_{max} $\geq$ T_{mean} $\geq$ T_{min} T_{max}(d) $\geq$ T_{min}(d-1) T_{min}(d) $\leq$ T_{max}(d-1) (Estévez et al., 2011)		RH_{max} $\geq$ RH_{mean} $\geq$ RH_{min} (Estévez et al., 2011)

T: daily or semi-hourly temperature (°C); T_{sh}: semi-hourly temperature (°C); T_{max}: daily maximum temperature (°C); T_{min}: daily minimum temperature (°C); T_{mean}: daily average temperature (°C); n(T_{sh}): total number of temperature records per day; r(T): ratio of non-missing semi-hourly or hourly temperature records to the total number of temperature records per day; P: daily precipitation (mm); P_{sh}: semi-hourly precipitation (mm); n(P_{sh}): total number of temperature records per day; r(P): ratio of non-missing semi-hourly or hourly precipitation records to the total number of precipitation records per day; RH: mean, maximum or minimum daily relative humidity (%); RH_{sh}: semi-hourly relative humidity (%) ; RH_{max} : daily maximum relative humidity (%); RH_{min} : daily minimum relative humidity (%); RH_{mean} : daily mean relative humidity (%); n(RH_{sh}): total number of relative humidity records per day; r(RH): ratio of non-missing semi-hourly or hourly relative humidity records to the total number of relative humidity records per day; d: day d; d-1: day before day d.

Note: exceptions were tolerated regarding the number of daily records throughout the season, if it was shown that the device correctly recorded a sufficient amount of weather data per day. These exceptions were decided on a case-by-case examination.

Table S2.4: Hyperparameters tuned with Bayesian optimization for regression models implemented.

Algorithm (R package)	**Hyperparameter**	**Meaning**	**Min**	**Max**
glmnet	alpha	elastic net mixing parameter	0	1
	lambda	penalty	0	1
XGBoost	nrounds	maximum number of iterations	4000	7000
	min_child_weight	minimum number of samples required to create a new node	5	18
	colsample_bytree	subsample fraction of features to use when constructing each tree	0.4	0.8
	max_depth	maximum depth of a tree	2	12
	eta	learning rate	3e-4	0.01
LightGBM	num_iterations	maximum number of iterations	4000	7000
	min_data_in_leaf	minimum number of samples in a leaf	5	18
	feature_fraction	subsample fraction of features to use when constructing each tree	0.4	0.8
	max_depth	maximum depth of a tree	2	12
	learning_rate	learning rate	3e-4	0.01

Table S2.5: Linear random effects (LRE) models evaluated in four cross-validation scenarios (CV0-Year, CV00-Year, CV0-Site, CV00-Site).
E, environment (YearxSite combination); G, SNPs markers;Y, year; S, site; W, environmental covariates + longitude + latitude; G×E, interactions between environments and markers; G×S, interactions between sites and markers; G×Y, interactions between years and markers; G×W, interactions between markers and environmental covariates.

Model abbreviation	Effects included								
	Main effects					Interaction terms			
	G	E	S	Y	W	G×E	G×S	G×Y	G×W
G+E	X	X							
G+S	X		X						
G+Y	X			X					
G+E+G×E	X	X				X			
G+S+G×S	X		X				X		
G+Y+G×Y	X			X				X	
G+E+S+Y+G×S+G×Y+G×E	X	X	X	X		X	X	X	
G+W	X				X				
G+E+W	X	X			X				
G+W+G×W	X				X				X
G+E+W+G×W	X	X			X				X
G+E+W+G×W+G×E	X	X			X	X			X

Table S2.6: Weighted average correlation between predicted and observed values across 71 environments for the trait grain yield for two cross-validation schemes leaving one year out (CV0-Year, CV00-Year) and for four types of statistical models (XGBoost, LightGBM, Elastic net and linear random effects model, i.e. LRE model) tested with different combinations of predictor variables. The best model for each cross-validation scheme is written in bold.

Type of statistical model	Predictors used	Weather- and soil- based variables included	Longitude and latitude included	CV0: Leave-one-year-out	CV00: Leave-one-year-out, new genotypes
XGBoost	G+W+Y+Lon+Lat	Y	Y	**0.419**	**0.301**
XGBoost	G+W	Y	N	0.414	0.292
XGBoost	G+Lon+Lat+Y	N	Y	0.398	0.267
LightGBM	G+W+Y+Lon+Lat	Y	Y	0.417	0.293
LightGBM	G+W	Y	N	0.411	0.286
LightGBM	G+Lon+Lat+Y	N	Y	0.406	0.27
Elastic net	G+W+Y+Lon+Lat	Y	Y	0.319	0.226
Elastic net	G+W	Y	N	0.313	0.231
Elastic net	G+Lon+Lat+Y	N	Y	0.31	0.241
LRE model	G+E	N	N	0.356	0.25
LRE model	G+E+G×E	N	N	0.362	0.271
LRE model	G+S	N	N	0.343	0.259
LRE model	G+S+GS	N	N	0.362	0.289
LRE model	G+Y	N	N	0.32	0.193
LRE model	G+Y+G×Y	N	N	0.313	0.199
LRE model	G+E+S+Y+G×S+G×Y+G×E	N	N	0.373	0.287
LRE model	G+W	Y	N	0.341	0.256
LRE model	G+E+W	Y	N	0.371	0.273
LRE model	G+W+G×W	Y	N	0.316	0.258
LRE model	G+E+W+G×W	Y	N	0.347	0.281
LRE model	G+E+W+G×W+G×E	Y	N	0.377	0.291
LRE model	G+W	Y	Y	0.35	0.26
LRE model	G+E+W	Y	Y	0.372	0.274
LRE model	G+W+G×W	Y	Y	0.323	0.267
LRE model	G+E+W+G×W	Y	Y	0.347	0.287
LRE model	G+E+W+G×W+G×E	Y	Y	0.377	0.293

Table S2.7: Weighted average correlation between predicted and observed values across 71 environments for the trait grain yield for two cross-validation schemes leaving one site out (CV0-Site, CV00-Site) and for four types of statistical models (XGBoost, LightGBM, Elastic net and linear random effects model, i.e. LRE model) tested with different combinations of predictor variables. The best model for each cross-validation scheme is written in bold.

Type of statistical model	Predictors used	Weather- and soil- based variables included	Longitude and latitude included	CV0: Leave-one-site-out	CV00: Leave-one-site-out, new genotypes
XGBoost	G+W+Y+Lon+Lat	Y	Y	0.495	0.28
XGBoost	G+W	Y	N	0.485	0.275
XGBoost	G+Lon+Lat+Y	N	Y	**0.504**	0.269
LightGBM	G+W+Y+Lon+Lat	Y	Y	0.496	0.275
LightGBM	G+W	Y	N	0.489	0.288
LightGBM	G+Lon+Lat+Y	N	Y	0.503	0.272
Elastic net	G+W+Y+Lon+Lat	Y	Y	0.392	0.243
Elastic net	G+W	Y	N	0.38	0.247
Elastic net	G+Lon+Lat+Y	N	Y	0.388	0.225
LRE model	G+E	N	N	0.461	0.248
LRE model	G+E+G×E	N	N	0.453	0.269
LRE model	G+S	N	N	0.447	0.265
LRE model	G+S+GS	N	N	0.445	0.274
LRE model	G+Y	N	N	0.392	0.177
LRE model	G+Y+G×Y	N	N	0.403	0.185
LRE model	G+E+S+Y+G×S+G×Y+G×E	N	N	0.471	0.275
LRE model	G+W	Y	N	0.423	0.264
LRE model	G+E+W	Y	N	0.475	0.274
LRE model	G+W+G×W	Y	N	0.379	0.242
LRE model	G+E+W+G×W	Y	N	0.46	0.287
LRE model	G+E+W+G×W+G×E	Y	N	0.475	0.294
LRE model	G+W	Y	Y	0.437	0.276
LRE model	G+E+W	Y	Y	0.475	0.274
LRE model	G+W+G×W	Y	Y	0.4	0.257
LRE model	G+E+W+G×W	Y	Y	0.463	0.287
LRE model	G+E+W+G×W+G×E	Y	Y	0.477	**0.296**

Table S2.8: Weighted average correlation between predicted and observed values across 71 environments for the trait plant height for two cross-validation schemes leaving one year out (CV0-Year, CV00-Year) and for four types of statistical models (XGBoost, LightGBM, Elastic net and linear random effects model, i.e. LRE model) tested with different combinations of predictor variables. The best model for each cross-validation scheme is written in bold.

Type of statistical model	Predictors used	Weather- and soil- based variables included	Longitude and latitude included	CV0: Leave-one-year-out	CV00: Leave-one-year-out, new genotypes
XGBoost	G+W+Y+Lon+Lat	Y	Y	0.632	0.522
XGBoost	G+W	Y	N	0.602	0.493
XGBoost	G+Lon+Lat+Y	N	Y	0.658	0.554
LightGBM	G+W+Y+Lon+Lat	Y	Y	0.631	0.521
LightGBM	G+W	Y	N	0.61	0.49
LightGBM	G+Lon+Lat+Y	N	Y	0.663	0.555
Elastic net	G+W+Y+Lon+Lat	Y	Y	0.517	0.453
Elastic net	G+W	Y	N	0.491	0.407
Elastic net	G+Lon+Lat+Y	N	Y	0.564	0.536
LRE model	G+E	N	N	**0.686**	**0.604**
LRE model	G+E+G×E	N	N	0.685	0.602
LRE model	G+S	N	N	0.623	0.562
LRE model	G+S+GS	N	N	0.608	0.56
LRE model	G+Y	N	N	0.487	0.426
LRE model	G+Y+G×Y	N	N	0.476	0.429
LRE model	G+E+S+Y+G×S+G×Y+G×E	N	N	0.675	0.598
LRE model	G+W	Y	N	0.499	0.431
LRE model	G+E+W	Y	N	0.678	0.59
LRE model	G+W+G×W	Y	N	0.41	0.377
LRE model	G+E+W+G×W	Y	N	0.66	0.556
LRE model	G+E+W+G×W+G×E	Y	N	0.674	0.58
LRE model	G+W	Y	Y	0.513	0.458
LRE model	G+E+W	Y	Y	0.679	0.589
LRE model	G+W+G×W	Y	Y	0.416	0.396
LRE model	G+E+W+G×W	Y	Y	0.661	0.56
LRE model	G+E+W+G×W+G×E	Y	Y	0.676	0.58

Table S2.9: Weighted average correlation between predicted and observed values across 71 environments for the trait plant height for two cross-validation schemes leaving one site out (CV0-Site, CV00-Site) and for four types of statistical models (XGBoost, LightGBM, Elastic net and linear random effects model, i.e. LRE model) tested with different combinations of predictor variables. The best model for each cross-validation scheme is written in bold.

Type of statistical model	Predictors used	Weather- and soil- based variables included	Longitude and latitude included	CV0: Leave-one-site-out	CV00: Leave-one-site-out, new genotypes
XGBoost	G+W+Y+Lon+Lat	Y	Y	0.700	0.517
XGBoost	G+W	Y	N	0.686	0.512
XGBoost	G+Lon+Lat+Y	N	Y	0.719	0.541
LightGBM	G+W+Y+Lon+Lat	Y	Y	0.704	0.514
LightGBM	G+W	Y	N	0.685	0.515
LightGBM	G+Lon+Lat+Y	N	Y	0.72	0.534
Elastic net	G+W+Y+Lon+Lat	Y	Y	0.576	0.473
Elastic net	G+W	Y	N	0.577	0.447
Elastic net	G+Lon+Lat+Y	N	Y	0.624	0.538
LRE model	G+E	N	N	0.736	0.597
LRE model	G+E+G×E	N	N	0.736	**0.598**
LRE model	G+S	N	N	0.69	0.536
LRE model	G+S+GS	N	N	0.707	0.58
LRE model	G+Y	N	N	0.554	0.38
LRE model	G+Y+G×Y	N	N	0.554	0.369
LRE model	G+E+S+Y+G×S+G×Y+G×E	N	N	**0.742**	0.59
LRE model	G+W	Y	N	0.591	0.46
LRE model	G+E+W	Y	N	0.732	0.588
LRE model	G+W+G×W	Y	N	0.486	0.363
LRE model	G+E+W+G×W	Y	N	0.717	0.565
LRE model	G+E+W+G×W+G×E	Y	N	0.731	0.579
LRE model	G+W	Y	Y	0.59	0.469
LRE model	G+E+W	Y	Y	0.732	0.588
LRE model	G+W+G×W	Y	Y	0.476	0.384
LRE model	G+E+W+G×W	Y	Y	0.718	0.567
LRE model	G+E+W+G×W+G×E	Y	Y	0.732	0.579

Table S2.10: Pearson's correlations between predicted and observed values computed within each environment in the CV0-Year prediction problem, using XGBoost with and without environmental data (results ordered by year).

Year_Exp	Pearson's correlation between predicted and observed values - Model XGBoost-G+Lon+Lat+Y	Pearson's correlation between predicted and observed values - Model XGBoost-G+W+Y+Lon+Lat
2014_DEH1	-0.00038	-0.03141
2014_GAH1	0.039908	0.112209
2014_IAH1_early	0.470778	0.412316
2014_IAH1_late	0.244271	0.376256
2014_ILH1	0.326263	0.436469
2014_INH1	0.554022	0.3753
2014_MNH1	-0.14013	0.42382
2014_MOH1	0.35867	0.457647
2014_MOH2	0.389949	0.469562
2014_NEH1	0.175078	0.204411
2014_NEH2	0.227811	0.22328
2014_NYH2	0.494023	0.432915
2014_ONH1	0.270531	0.332591
2014_ONH2	0.526447	0.43561
2014_TXH1	0.468913	0.480222
2014_TXH2	0.275946	0.20704
2014_WIH1	0.548343	0.606907
2015_DEH1	0.428647	0.337348
2015_GAH1	0.100642	0.259561
2015_ILH1	0.292991	0.382949
2015_INH1	0.33941	0.365498
2015_KSH1	0.166681	0.217401
2015_MNH1	0.219478	0.298932
2015_MOH1	0.147063	0.177479
2015_MOH2	0.258198	0.237618
2015_NEH2	0.150283	0.142226
2015_NEH3	0.148109	0.261604
2015_NYH2	0.368203	0.367911
2015_NYH3	0.396606	0.399689
2015_OHH1	0.302886	0.330156
2015_ONH1	0.500552	0.381253
2015_ONH2	0.470438	0.453759
2015_SDH1	0.165344	0.275692
2015_TXH1	0.409079	0.379781
2016_DEH1	0.676664	0.632206
2016_GAH2	0.167941	0.256443
2016_IAH2	0.361859	0.467092
2016_IAH3	0.323082	0.361847
2016_IAH4	0.513045	0.45837
2016_ILH1.a	0.15853	0.20329
2016_ILH1.b	0.113304	0.183547
2016_INH1	0.610272	0.64807
2016_KSH1	0.038255	-0.02406
2016_MIH1	0.630822	0.57595
2016_MNH1	0.499454	0.383406
2016_MOH1	0.519051	0.516576
2016_NEH1	0.559879	0.594455
2016_NEH4	0.336597	0.407543
2016_NYH2	0.17573	0.170988
2016_ONH1	0.512765	0.588222
2016_ONH2	0.488428	0.449989
2016_TXH1	0.488807	0.569038
2016_WIH1	0.656356	0.689881
2016_WIH2	0.627307	0.640923
2017_ARH1	0.080175	0.072308
2017_ARH2	-0.00686	0.028488
2017_COH1	0.384681	0.410604
2017_DEH1	0.759031	0.696925
2017_GAH1	0.235513	0.308306
2017_GAH2	0.406465	0.45164
2017_IAH4	0.659361	0.693042
2017_MIH1	0.734516	0.711509
2017_MOH1	0.589471	0.509746
2017_NYH2	0.430153	0.500422
2017_NYH3	-0.05178	0.157664
2017_ONH1	0.533839	0.523437

2017_TXH1-Dry	0.421615	0.574875
2017_TXH1-Early	0.305317	0.456222
2017_TXH1-Late	0.278046	0.383466
2017_WIH1	0.627729	0.638422
2017_WIH2	0.682127	0.683406

Table S2.11: Pearson's correlations between predicted and observed values computed within each environment in the CV0-Site prediction problem, using XGBoost with and without environmental data (results ordered by site).

Year_Exp	Pearson's correlation between predicted and observed values - Model XGBoost-G+Lon+Lat+Y	Pearson's correlation between predicted and observed values - Model XGBoost-G+W+Y+Lon+Lat
2017_ARH1	0.083618	0.087805
2017_ARH2	-0.11405	-0.08229
2017_COH1	0.434181	0.395753
2014_NYH2	0.618258	0.645201
2015_NYH2	0.420689	0.46726
2015_NYH3	0.414204	0.504762
2016_NYH2	0.071884	0.13314
2017_NYH2	0.489832	0.510642
2017_NYH3	0.039865	0.194063
2014_ONH2	0.701346	0.680824
2015_ONH2	0.636242	0.626894
2016_ONH2	0.696007	0.488337
2014_ONH1	0.545165	0.498874
2015_ONH1	0.615251	0.466422
2016_ONH1	0.59863	0.583291
2017_ONH1	0.553925	0.59485
2014_IAH1_early	0.492282	0.43794
2014_IAH1_late	0.437323	0.430194
2016_IAH4	0.576048	0.469782
2017_IAH4	0.629009	0.665542
2016_IAH2	0.463735	0.4613
2016_IAH3	0.482565	0.450599
2015_KSH1	0.312603	0.331673
2016_KSH1	0.036547	0.015976
2016_MIH1	0.661474	0.649279
2017_MIH1	0.726516	0.751137
2015_OHH1	0.569874	0.512396
2014_INH1	0.660406	0.479942
2015_INH1	0.410292	0.41781
2016_INH1	0.681134	0.706826
2015_SDH1	0.382275	0.278046
2014_TXH1	0.487586	0.455357
2015_TXH1	0.461609	0.40085
2016_TXH1	0.484127	0.407383
2017_TXH1-Dry	0.551344	0.404471
2017_TXH1-Early	0.415016	0.307181
2017_TXH1-Late	0.270513	0.331329
2014_TXH2	0.50653	0.408112
2014_DEH1	0.000496	-0.03348
2015_DEH1	0.568924	0.647196
2016_DEH1	0.475822	0.576064
2017_DEH1	0.567695	0.675415
2014_GAH1	0.15681	0.158373
2015_GAH1	0.637245	0.587707
2017_GAH1	0.454122	0.182706
2016_GAH2	0.245837	0.167543
2017_GAH2	0.433908	0.475884
2014_ILH1	0.478114	0.551237
2015_ILH1	0.507431	0.540645
2016_ILH1.a	0.191753	0.156859
2016_ILH1.b	0.136002	0.159364
2014_MNH1	0.321004	0.590724
2015_MNH1	0.490254	0.510987
2016_MNH1	0.60069	0.514643
2014_MOH1	0.46074	0.437299
2014_MOH2	0.60346	0.612721
2015_MOH1	0.176224	0.150778
2015_MOH2	0.394699	0.348521
2016_MOH1	0.591315	0.605213
2017_MOH1	0.56974	0.461632
2014_NEH1	0.220867	0.246266
2015_NEH3	0.184895	0.207049
2014_NEH2	0.199181	0.194448
2015_NEH2	0.212923	0.253392
2016_NEH1	0.659393	0.691755
2016_NEH4	0.432744	0.463022
2014_WIH1	0.759711	0.684768

2016_WIH1	0.798492	0.797723
2017_WIH1	0.673311	0.680053
2016_WIH2	0.627422	0.685192
2017_WIH2	0.738874	0.720376

2.7 Bibliography

Abdollahi-Arpanahi R, Gianola D, Peñagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genetics Selection Evolution 52(1):12, DOI 10.1186/s12711-020-00531-z, URL https://doi.org/10.1186/s12711-020-00531-z

AlKhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh CT, Bohn M, et al. (2018) Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC research notes 11(1):452

Allen RG, Pereira LS, Raes D, Smith M, et al. (1998) Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. Fao, Rome 300(9):D05,109

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics 9(11):3691–3702

Baskerville GL, Emin P (1969) Rapid estimation of heat accumulation from maximum and minimum temperatures. Ecology 50(3):514–517

Bassu S, Brisson N, Durand JL, Boote K, Lizaso J, Jones JW, Rosenzweig C, Ruane AC, Adam M, Baron C, et al. (2014) How do various maize crop models vary in their responses to climate change factors? Global change biology 20(7):2301–2320

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software, Articles 67(1):1–48, DOI 10.18637/jss.v067.i01, URL https://www.jstatsoft.org/v067/i01

Behravan H, Hartikainen JM, Tengström M, Pylkäs K, Winqvist R, Kosma VM, Mannermaa A (2018) Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. Scientific Reports 8(1):1–13, DOI 10.1038/s41598-018-31573-5, URL https://www.nature.com/articles/s41598-018-31573-5, number: 1 Publisher: Nature Publishing Group

Bellot P, de Los Campos G, Pérez-Enciso M (2018) Can deep learning improve genomic prediction of complex human traits? Genetics 210(3):809–819

Bernal-Vasquez AM, Gordillo A, Schmidt M, Piepho HP (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC genetics 18(1):51

Biecek P (2018) Dalex: Explainers for complex predictive models in r. Journal of Machine Learning Research 19(84):1–5, URL http://jmlr.org/papers/v19/18-416.html

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) Tassel: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635

Brock J, Lange M, Tratalos JA, More SJ, Graham DA, Guelbenzu-Gonzalo M, Thulke HH (2021) Combining expert knowledge and machine-learning to classify herd types in livestock systems. Scientific Reports 11(1):1–10

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Science 52(2):707–719, DOI 10.2135/cropsci2011.06.0299, URL http://doi.wiley.com/10.2135/cropsci2011.06.0299

Bustos-Korts D, Boer MP, Malosetti M, Chapman S, Chenu K, Zheng B, Van Eeuwijk FA (2019) Combining crop growth modeling and statistical genetic modeling to evaluate phenotyping strategies. Frontiers in Plant Science 10

Butler EE, Huybers P (2015) Variations in the sensitivity of US maize yield to extreme temperatures by region and growth phase. Environmental Research Letters 10(3):034,009, DOI 10.1088/1748-9326/10/3/034009, URL https://doi.org/10.1088%2F1748-9326%2F10%2F3%2F034009, publisher: IOP Publishing

Cakir R (2004) Effect of water stress at different development stages on vegetative and reproductive growth of corn. Field Crops Research 89(1):1–16

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Chenu K (2015) Characterizing the crop environment–nature, significance and applications. Crop physiology pp 321–348

Cicchino M, Edreira JIR, Uribelarrea M, Otegui ME (2010) Heat Stress in Field-Grown Maize: Response of Physiological Determinants of Grain Yield. Crop Science 50(4):1438–1448, DOI 10.2135/cropsci2009.10.0574, URL https://dl.sciencesocieties.org/publications/cs/abstracts/50/4/1438, publisher: Crop Science Society of America

Cooper M, DeLacy I (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theoretical and Applied Genetics 88(5):561–572

Costa-Neto G, Fritsche-Neto R, Crossa J (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106

Costa-Neto GMF, Júnior OPM, Heinemann AB, de Castro AP, Duarte JB (2020) A novel gis-based tool to reveal spatial trends in reaction norm: upland rice case study. Euphytica 216(3):1–16

Crane-Droesch A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters 13(11):114,003, DOI 10.1088/1748-9326/aae159, URL https://doi.org/10.1088%2F1748-9326%2Faae159, publisher: IOP Publishing

Crossa J, Neto RF, Montesinos-López OA, Costa-Neto GMF, Dreisigacker S, Montesinos-Lopez A, Bentley AR (2021) The modern plant breeding triangle: Optimising the use of genomics, phenomics and enviromics data. Frontiers in Plant Science 12:332

Cuevas J, Granato I, Fritsche-Neto R, Montesinos-Lopez OA, Burgueño J, Bandeira e Sousa M, Crossa J (2018) Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. G3: Genes, Genomes, Genetics 8(4):1347–1365

De Los Campos G, Pérez-Rodríguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nature communications 11(1):1–10

Delerce S, Dorado H, Grillon A, Rebolledo MC, Prager SD, Patiño VH, Garcés Varón G, Jiménez D (2016) Assessing Weather-Yield Relationships in Rice at Local Scale Using Data Mining Approaches. PLOS ONE 11(8):e0161,620, DOI 10.1371/journal.pone.0161620, URL https://dx.plos.org/10.1371/journal.pone.0161620

Denmead O, Shaw RH (1960) The effects of soil moisture stress at different stages of growth on the development and yield of corn 1. Agronomy journal 52(5):272–274

van Eeuwijk F, Kang M, Denis J (1996) Incorporating Additional Information on Genotypes and Environments in Models for Two-way Genotype by Environment Tables. In: Gauch HG, Kang M (eds) Genotype-by-Environment Interaction, CRC Press, pp 15–49, DOI 10.1201/9781420049374.ch2, URL http://www.crcnetbase.com/doi/10.1201/9781420049374.ch2

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77(4):802–813

Ersoz ES, Martin NF, Stapleton AE (2020) On to the next chapter for crop breeding: Convergence with data science. Crop Science 60(2):639–655

Estévez J, Gavilán P, Giráldez JV (2011) Guidelines on validation procedures for meteorological data from automatic weather stations. Journal of Hydrology 402(1):144–154, DOI 10.1016/j.jhydrol.2011.02.031, URL http://www.sciencedirect.com/science/article/pii/S0022169411001594

Falcon CM, Kaeppler SM, Spalding EP, Miller ND, Haase N, AlKhalifah N, Bohn M, Buckler ES, Campbell DA, Ciampitti I, et al. (2020) Relative utility of agronomic, phenological, and

morphological traits for assessing genotype-by-environment interaction in maize inbreds. Crop Science 60(1):62–81

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5):1189–1232, DOI 10.1214/aos/1013203451, URL https://projecteuclid.org/euclid.aos/1013203451, publisher: Institute of Mathematical Statistics

Fukuda S, Spreer W, Yasunaga E, Yuge K, Sardsud V, Müller J (2013) Random Forests modelling for the estimation of mango (Mangifera indica L. cv. Chok Anan) fruit yields under different irrigation regimes. Agricultural Water Management 116:142–150, DOI 10.1016/j.agwat.2012.07.003, URL https://linkinghub.elsevier.com/retrieve/pii/S0378377412001874

Gage JL, Jarquin D, Romay C, Lorenz A, Buckler ES, Kaeppler S, Alkhalifah N, Bohn M, Campbell DA, Edwards J, et al. (2017) The effect of artificial selection on phenotypic plasticity in maize. Nature communications 8(1):1–11

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, Kaski S (2019) Modelling G×E with historical weather information improves genomic prediction in new environments. Bioinformatics 35(20):4045–4052, DOI 10.1093/bioinformatics/btz197, URL https://academic.oup.com/bioinformatics/article/35/20/4045/5448861

González-Recio O, Jiménez-Montero J, Alenda R (2013) The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. Journal of dairy science 96(1):614–624

Gräler B, Pebesma EJ, Heuvelink GB (2016) Spatio-temporal interpolation using gstat. R J 8(1):204

Greaves JA (1996) Improving suboptimal temperature tolerance in maize-the search for variation. Journal of experimental botany 47(3):307–323

Haley C, Visscher P (1998) Strategies to utilize marker-quantitative trait loci associations. Journal of Dairy Science 81:85–97

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics, Springer, New York, NY

Hatfield JL, Prueger JH (2015) Temperature extremes: Effect on plant growth and development. Weather and climate extremes 10:4–10

Hatfield JL, Boote KJ, Kimball B, Ziska L, Izaurralde RC, Ort D, Thomson AM, Wolfe D (2011) Climate impacts on agriculture: implications for crop production. Agronomy journal 103(2):351–370

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics 127(2):463–480

Holzworth DP, Huth NI, deVoil PG, Zurcher EJ, Herrmann NI, McLean G, Chenu K, van Oosterom EJ, Snow V, Murphy C, et al. (2014) Apsim–evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62:327–350

Hutter F, Hoos H, Leyton-Brown K (2014) An efficient approach for assessing hyperparameter importance. In: International conference on machine learning, PMLR, pp 754–762

Jarquín D, Lemes da Silva C, 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(2):1–15

Jarquin D, De Leon N, Romay MC, Bohn MO, Buckler ES, Ciampitti IA, Edwards JW, Erlt D, Flint-Garcia S, Gore MA, et al. (2020) Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Frontiers in genetics 11:1819

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics 127(3):595–607, DOI 10.1007/s00122-013-2243-1, URL http://link.springer.com/10.1007/s00122-013-2243-1

Jeong JH, Resop JP, Mueller ND, Fleisher DH, Yun K, Butler EE, Timlin DJ, Shim KM, Gerber JS, Reddy VR, Kim SH (2016) Random Forests for Global and Regional Crop Yield Predictions. PLoS ONE 11(6), DOI 10.1371/journal.pone.0156571, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4892571/

Juliana P, Montesinos-López OA, Crossa J, Mondal S, Pérez LG, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S, et al. (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. Theoretical and Applied Genetics 132(1):177–194

Kagawa R, Kawazoe Y, Ida Y, Shinohara E, Tanaka K, Imai T, Ohe K (2017) Development of type 2 diabetes mellitus phenotyping framework using expert knowledge and machine learning approach. Journal of diabetes science and technology 11(4):791–799

Kassambara A, 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, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, pp 3146–3154

Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, Huth NI, Hargreaves JN, Meinke H, Hochman Z, et al. (2003) An overview of apsim, a model designed for farming systems simulation. European journal of agronomy 18(3-4):267–288

Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G (2020) Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific reports 10(1):1–12

Köppen W, Geiger R (1930) Handbuch der klimatologie, vol 1. Gebrüder Borntraeger Berlin

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

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Lampa E, Lind L, Lind PM, Bornefalk-Hermansson A (2014) The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environmental Health: A Global Access Science Source 13:57, DOI 10.1186/1476-069X-13-57

Li Y, Guan K, Schnitkey GD, DeLucia E, Peng B (2019) Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the united states. Global change biology 25(7):2325–2337

Lizaso J, Ruiz-Ramos M, Rodríguez L, Gabaldon-Leal C, Oliveira J, Lorite I, Sánchez D, García E, Rodríguez A (2018) Impact of high temperatures in maize: Phenology and yield components. Field Crops Research 216:129–140, DOI 10.1016/j.fcr.2017.11.013, URL https://linkinghub.elsevier.com/retrieve/pii/S0378429017310183

Lobell DB, Roberts MJ, Schlenker W, Braun N, Little BB, Rejesus RM, Hammer GL (2014) Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest. Science 344(6183):516–519, DOI 10.1126/science.1251423, URL https://science.sciencemag.org/content/344/6183/516, publisher: American Association for the Advancement of Science Section: Report

Malosetti M, Voltas J, Romagosa I, Ullrich S, Van Eeuwijk F (2004) Mixed models including environmental covariables for studying qtl by environment interaction. Euphytica 137(1):139–145

Malosetti M, Bustos-Korts D, Boer MP, van Eeuwijk FA (2016) Predicting Responses in Multiple Environments: Issues in Relation to Genotype × Environment Interactions. Crop Science 56(5):2210–2222, DOI 10.2135/cropsci2015.05.0311, URL http://doi.wiley.com/10.2135/cropsci2015.05.0311

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, 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(1):1–6

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829

Millet EJ, Kruijer W, Coupel-Ledru A, Alvarez Prado S, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019) Genomic prediction of maize yield across European environmental conditions. Nature Genetics 51(6):952–956, DOI 10.1038/s41588-019-0414-y, URL http://www.nature.com/articles/s41588-019-0414-y

Mimić G, Brdar S, Brkić M, Panić M, Marko O, Crnojević V (2020) engineering meteorological features to select stress tolerant hybrids in maize. Scientific reports 10(1):1–10

Moisen GG, Freeman EA, Blackard JA, Frescino TS, Zimmermann NE, Edwards Jr TC (2006) Predicting tree species presence and basal area in utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecological modelling 199(2):176–187

Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong GY, Myles S (2015) Linkimpute: fast and accurate genotype imputation for nonmodel organisms. G3: Genes, Genomes, Genetics 5(11):2383–2390

Monteverde E, Gutierrez L, Blanco P, Pérez de Vida F, Rosas JE, Bonnecarrère V, Quero G, McCouch S (2019) Integrating Molecular Markers and Environmental Covariates To Interpret Genotype by Environment Interaction in Rice (*Oryza sativa* L.) Grown in Subtropical Areas. G3: Genes|Genomes|Genetics 9(5):1519–1531, DOI 10.1534/g3.119.400064, URL http://g3journal.org/lookup/doi/10.1534/g3.119.400064

Mushore T, Manatsa D, Pedzisai E, Muzenda-Mudavanhu C, Mushore W, 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. Theoretical and Applied Climatology 129(3):1167–1173

Netzeva TI, Worth AP, Aldenberg T, Benigni R, Cronin MT, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, 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. Alternatives to Laboratory Animals 33(2):155–173

Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. In: BMC proceedings, BioMed Central, vol 5, pp 1–5

Olivoto T, Nardino M, Carvalho I, Follmann D, Ferrari M, Szareski V, Souza Vd (2017) Reml/blup and sequential path analysis in estimating genotypic values and interrelationships among simple maize grain yield-related traits. Genetics and Molecular Research 16(1):1–19

Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Computers & geosciences 30(7):683–691

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the bglr statistical package. Genetics 198(2):483–495

Pérez-Rodríguez P, Crossa J, Bondalapati K, De Meyer G, Pita F, Campos Gdl (2015) A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. Crop Science 55(3):1143–1151

Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, Legarra A, Autrique E, Campos Gdl, Burgueño J, Dreisigacker S (2017) Single-step genomic and pedigree genotype× environment interaction models for predicting wheat lines in international environments. The plant genome 10(2):plantgenome2016–09

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. The American journal of human genetics 81(3):559–575

R Core Team (2019) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Rahmstorf S, Foster G, Cazenave A (2012) Comparing climate projections to observations up to 2011. Environmental Research Letters 7(4):044,035, DOI 10.1088/1748-9326/7/4/044035, URL https://doi.org/10.1088%2F1748-9326%2F7%2F4%2F044035, publisher: IOP Publishing

Ridgeway G (2007) Generalized boosted models: A guide to the gbm package. Update 1(1):2007

Rincent R, Kuhn E, Monod H, Oury FX, Rousset M, Allard V, Le Gouis J (2017) Optimization of multi-environment trials for genomic selection based on crop models. Theoretical and Applied Genetics 130(8):1735–1752

Rincent R, Malosetti M, Ababaei B, Touzy G, Mini A, Bogard M, Martre P, Le Gouis J, van Eeuwijk F (2019) Using crop growth model stress covariates and ammi decomposition to better

predict genotype-by-environment interactions. Theoretical and Applied Genetics 132(12):3399–3411

Ritchie SW, Hanway JJ, Benson GO, Herman JC, Lupkes SJ (1993) How a corn plant develops. Special report (48)

Roe KD, Jawa V, Zhang X, Chute CG, Epstein JA, Matelsky J, Shpitser I, Taylor CO (2020) Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. PloS one 15(4):e0231,300

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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 Genes| Genomes| Genetics

Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP (2019) Comparative performances of machine learning methods for classifying crohn disease patients using genome-wide genotyping data. Scientific reports 9(1):1–18

Schlenker W, Roberts MJ (2009) Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences 106(37):15,594–15,598, DOI 10.1073/pnas.0906865106, URL https://www.pnas.org/content/106/37/15594, publisher: National Academy of Sciences Section: Social Sciences

Shahhosseini M, Hu G, Archontoulis SV (2020) Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science 11:1120

Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2, Curran Associates Inc., Red Hook, NY, USA, NIPS'12, p 2951–2959

Bandeira e Sousa M, Cuevas J, de Oliveira Couto EG, Pérez-Rodríguez P, Jarquín D, Fritsche-Neto R, Burgueño J, Crossa J (2017) Genomic-enabled prediction in maize using kernel models with genotype× environment interaction. G3: Genes, Genomes, Genetics 7(6):1995–2014

Sukumaran S, Crossa J, Jarquín D, Reynolds M (2017) Pedigree-based prediction models with genotype× environment interaction in multienvironment trials of cimmyt wheat. Crop Science 57(4):1865–1880

Sukumaran S, Jarquin D, Crossa J, Reynolds M (2018) Genomic-enabled prediction accuracies increased by modeling genotype× environment interaction in durum wheat. The Plant Genome 11(2):1–11

Tardieu F, Cabrera-Bosquet L, Pridmore T, Bennett M (2017) Plant phenomics, from sensors to knowledge. Current Biology 27(15):R770–R783

Technow F, Messina CD, Totir LR, Cooper M (2015) Integrating crop growth models with whole genome prediction through approximate bayesian computation. PloS one 10(6):e0130,855

Tiezzi F, de Los Campos G, Gaddis KP, Maltecca C (2017) Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. Journal of dairy science 100(3):2042–2056

Trnka M, Rötter RP, Ruiz-Ramos M, Kersebaum KC, Olesen JE, Žalud Z, Semenov MA (2014) Adverse weather conditions for european wheat production will become more frequent with climate change. Nature Climate Change 4(7):637–643, DOI 10.1038/nclimate2242, URL http://www.nature.com/articles/nclimate2242

Troy TJ, Kipgen C, Pal I (2015) The impact of climate extremes and irrigation on us crop yields. Environmental Research Letters 10(5):054,013

Van Rijn JN, Hutter F (2018) Hyperparameter importance across datasets. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 2367–2376

Varma S, 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, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1397873/

Welch JR, Vincent JR, Auffhammer M, Moya PF, Dobermann A, Dawe D (2010) Rice yields in tropical/subtropical asia exhibit large but opposing sensitivities to minimum and maximum temperatures. Proceedings of the National Academy of Sciences 107(33):14,562–14,567

Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning, vol 2. MIT press Cambridge, MA

Yu J, Shi S, Zhang F, Chen G, Cao M (2019) Predgly: predicting lysine glycation sites for homo sapiens based on xgboost feature optimization. Bioinformatics 35(16):2749–2756

Zahumenskỳ I (2004) Guidelines on quality control procedures for data from automatic weather stations. World Meteorological Organization, Switzerland

Zhu P, Zhuang Q, Archontoulis SV, Bernacchi C, Müller C (2019) Dissecting the nonlinear response of maize yield to high temperature stress with model-data integration. Global change biology 25(7):2470–2484

Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, Whitaker VM, Pérez-Enciso M (2020) Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Frontiers in plant science 11:25

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2):301–320

3. learnMET: an R package to apply machine learning methods for genomic prediction using multi-environment trial data

Cathy C. Westhues[1,3,*], Henner Simianer[2,3] and Timothy M. Beissinger[1,3]

[1]Division of Plant Breeding Methodology, Department of Crop Sciences, University of Goettingen, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[2]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Albrecht-Thaer-Weg 3, 37075, Goettingen, Germany

[3]Center for Integrated Breeding Research, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[*] Email: cathy.jubin@uni-goettingen.de

Published in *G3 Genes|Genomes|Genetics*
https://doi.org/10.1093/g3journal/jkac226

3.1 Abstract

We introduce the R-package *learnMET*, developed as a flexible framework to enable a collection of analyses on multi-environment trial (MET) breeding data with machine learning-based models. *learnMET* allows the combination of genomic information with environmental data such as climate and/or soil characteristics. Notably, the package offers the possibility of incorporating weather data

from field weather stations, or to retrieve global meteorological datasets from a NASA database. Daily weather data can be aggregated over specific periods of time based on naive (for instance, non-overlapping 10-day windows) or phenological approaches. Different machine learning methods for genomic prediction are implemented, including gradient boosted decision trees, random forests, stacked ensemble models, and multi-layer perceptrons. These prediction models can be evaluated via a collection of cross-validation schemes that mimic typical scenarios encountered by plant breeders working with MET experimental data in a user-friendly way. The package is published under a MIT license and accessible on GitHub.

Keywords: multienvironment trials, machine learning, genotype x environment interaction, genomic prediction, R software

3.2 Introduction

Large amounts of data from various sources (phenotypic records from field trials, genomic or omics data, environmental information) are regularly gathered as part of multi-environment trials (MET). The efficient exploitation of these extensive datasets has become of utmost interest for breeders to address essentially two objectives: (1) accurately predicting genotype performance in future environments; (2) untangling complex relationships between genetic markers, environmental covariables (ECs) and phenotypes to better understand the pervasive phenomenon of genotype-by-environment (G x E) interaction.

Many R packages have recently been developed that allow to implement genomic prediction models accounting for G x E effects using mixed models: BGLR (Pérez and de Los Campos, 2014), sommer (Covarrubias-Pazaran, 2016), Bayesian Genomic Genotype × Environment Interaction (BGGE) (Granato et al., 2018), Bayesian Multi-Trait Multi-Environment for Genomic Selection (BMTME) (Montesinos-López et al., 2019), bWGR (Xavier et al., 2019), EnvRtype (Costa-Neto et al., 2021b) and MegaLMM (Runcie et al., 2021). BGGE presents a speed advantage compared to BGLR, that is explained by the use of an optimization procedure for sparse covariance matrices, while BMTME additionally exploits the genetic correlation among traits and environments to build linear G x E models. EnvRtype further widens the range of opportunities in Bayesian kernel models with the possibility to use non-linear arc-cosine kernels aiming at reproducing a deep learning approach (Costa-Neto et al., 2021a; Cuevas et al., 2019), and to harness environmental data retrieved by the package.

While Bayesian approaches have been successful at dramatically improving predictive ability in multi-environment breeding experiments (Costa-Neto et al., 2021b; Cuevas et al., 2017, 2019), data-driven machine learning algorithms represent alternative predictive modeling techniques with increased flexibility with respect to the form of the mapping function between input and output variables. In particular, non-linear effects including gene x gene and genotype x environment (G

x E) interactions can be captured with machine learning models (Crossa et al., 2019; McKinney et al., 2006; Ritchie et al., 2003; Westhues et al., 2021). G x E interactions are of utmost interest for plant breeders, especially when they present a crossover-type, because the latter implies a change in the relative ranking of genotypes across different environments. Breeders generally cope with G × E by either (1) focusing their program on wide adaptation of cultivars over a target population of environments, from which follows that the developed varieties are not the best ones for a given environment, and positive G × E interactions are not exploited, or (2) identifying varieties that are the best adapted to specific environments (Bernardo, 2002). Enhancing the modeling of genotype-by-environment interactions, by the inclusion of environmental covariates related to critical developmental stages, also resulted in an increase of predictive ability in many studies using MET datasets (Costa-Neto et al., 2021a; Heslot et al., 2012; Monteverde et al., 2019; Rincent et al., 2019).

In this article we describe the R-package learnMET and its principal functionalities. learnMET provides a pipeline to (1) facilitate environmental characterization and (2) evaluate and compare different types of machine learning approaches to predict quantitative traits based on relevant cross-validation schemes for MET datasets. The package offers flexibility by allowing to specify the sets of predictors to be used in predictions, and different methods to process genomic information to model genetic effects.

To validate the predictive performance of the models, different cross-validation (CV) schemes are covered by the package, that aim at addressing concrete plant breeding prediction problems with multi-environment field experiments. We borrow the same terminology as in previous related studies (see Burgueño et al. (2012); Jarquín et al. (2014, 2017)), as follows: (1) CV1: predicting the performance of newly developed genotypes (never tested in any of the environments included in the MET); (2) CV2: predicting the performance of genotypes that have been tested in some environments but not in others (also referred to as field sparse testing); (3) CV0: predicting the performance of genotypes in new environments, i.e. the environment has not been tested; and (4) CV00: predicting the performance of newly developed genotypes in new environments, i.e. both environment and genotypes have not been observed in the training set. For CV0 and CV00, four configurations are implemented: leave-one-environment-out, leave-one-site-out, leave-one-year-out and forward prediction.

3.3 Methods

3.3.1 Installation and dependencies

Using the devtools package (Wickham et al., 2021), learnMET can be easily installed from GitHub and loaded (Box 1).

Box 1: Install learnMET

```
> devtools::install_github("cjubin/learnMET")
> library(learnMET)
```

Dependencies are automatically installed or updated when executing the command above.

3.3.2 Real multi-environment trial datasets

Three toy datasets are included with the learnMET package to illustrate how input data should be provided by the user and how the different functionalities of the package can be utilized.

Rice datasets: The datasets were obtained from the INIA's Rice Breeding Program (Uruguay) and were used in previous studies (Monteverde et al., 2018, 2019). Two breeding populations of rice (*indica*, composed of 327 elite breeding lines; and *japonica*, composed of 320 elite breeding lines) were phenotyped for four traits. The two populations were evaluated at a single location (Treinta y Tres, Uruguay) across multiple years (2010-2012 for *indica* and 2009-2013 for *japonica*) and were genotyped using genotyping-by-sequencing (GBS) (Monteverde et al., 2019). Environmental covariables, characterizing three developmental stages throughout the growing season, were directly available. More details about the dataset are given in Monteverde et al. (2018).

Maize datasets: A subset of phenotypic and genotypic datasets, collected and made available by the G2F initiative (`www.genomes2fields.org`), were integrated into learnMET. Hybrid genotypic data were computed *in silico* based on the GBS data from inbred parental lines. For more information about the original datasets, please refer to AlKhalifah et al. (2018) and McFarland et al. (2020). In total, phenotypic data, collected from 22 environments covering 4 years (2014 to 2017) and 6 different locations in American states and Canadian provinces, are included in the package.

3.3.3 Running learnMET

learnMET is implemented as a three-step pipeline. These are described below.

3.3.4 Step 1: specifying input data and processing parameters

The first function in the learnMET pipeline is *create_METData()* (Box 2). The user must provide genotypic and phenotypic data, as well as basic information about the field experiments (e.g. longitude, latitude, planting and harvest date). Missing genotypic data should be imputed beforehand. Climate covariables can be directly provided as day-interval aggregated variables, using the argument *climate_variables*. Alternatively, in order to compute weather-based covariables, based on daily weather data, the user can set the *compute_climatic_ECs* argument to TRUE, and two possibilities are given. The first one is to provide raw daily weather data (with the *raw_weather_data* argument), which will undergo a quality control with the generation of an output file with flagged values. The second possibility, if the user does not have weather data avail-

able from measurements (e.g. from an in-field weather station), is the retrieval of daily weather records from the NASA's Prediction of Worldwide Energy Resources (NASA POWER) database (`https://power.larc.nasa.gov/`), using the package nasapower (Sparks, 2018). Spatio-temporal information contained in the *info_environments* argument is required. Note that the function also checks which environments are characterized by in-field weather data in the *raw_weather_data* argument, in order to retrieve satellite-based weather data for the remaining environments without in field weather stations. An overview of the pipeline is provided in Figure 3.1.

Some covariates are additionally computed, based on the daily weather data, such as vapor pressure deficit or the reference evapotranspiration using the Penman-Monteith (FAO-56) equation. The aggregation of daily information into day-interval based values is also carried out within this function. Four methods are available and should be specified with the argument *method_ECs_intervals*: (1) default: use of a definite number of intervals across all environments (i.e. the window length varies according to the duration of the growing season); (2) use of day-windows of fixed length (i.e. each window spans a given number of days, which remains identical across environments), that can be adjusted by the user; (3) use of specific day intervals according to each environment provided by the user, which should correspond to observed or assumed relevant phenological intervals; and (4) based on the estimated crop growth stage within each environment using accumulated growing degree-days in degrees Celsius.

Besides weather-based information, soil characterization for each environment can also be provided given the *soil_variables* argument. The output of *create_METData()* is a list object of class *METData*, required as input for all other functionalities of the package.

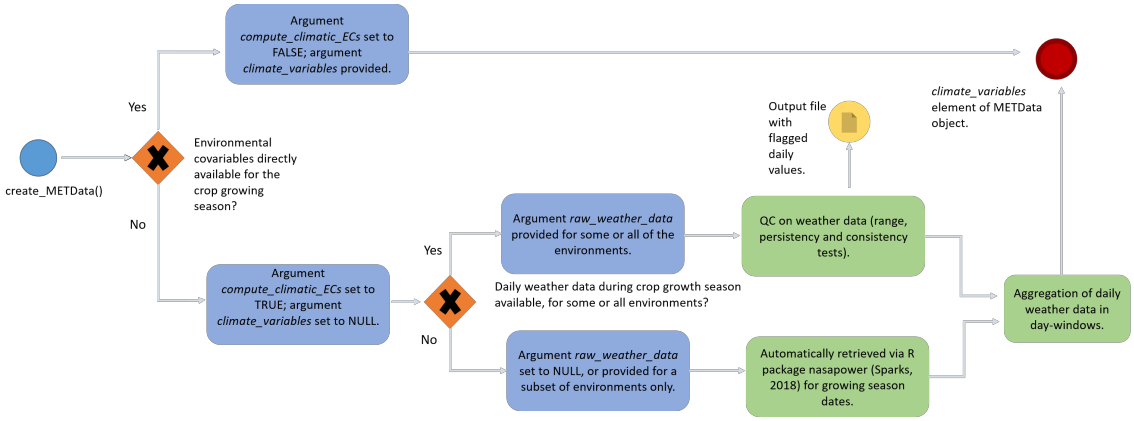


Figure 3.1: Overview of the pipeline regarding integration of weather data using the function *create_METData()* within the learnMET package. The blue circle signals the first step of the process, when the function is initially called. The blue boxes indicate how the arguments of the function should be characterized, according to the type of datasets available to the user. The green boxes indicate a task which is run in the pipeline via internal functions of the package. The red circle signals the final step, when the METData object is created. Details on the quality control tests implemented on daily weather data are provided at `https://cjubin.github.io/learnMET/reference/qc_raw_weather_data.html`, and on the methods to build ECs based on aggregation of daily data at `https://cjubin.github.io/learnMET/reference/get_ECs.html`.

Box 2: Integration of input data in a METData list object

***Case 1*: ECs directly provided by the user**

```
> library(learnMET)
> data(geno_indica)
> data(map_indica)
> data(pheno_indica)
> data(info_environments_indica)
> data(env_data_indica)
> METdata_indica <- create_METData(
geno = geno_indica,
map = map_indica,
pheno = pheno_indica,
climate_variables = climate_variables_indica,
info_environments = info_environments_indica,
compute_climatic_ECs = FALSE,
path_to_save = " /learnMET_analyses/indica")
```

***Case 2*: daily climate data automatically retrieved and ECs calculated via the package**

```
> data(geno_G2F)
> data(pheno_G2F)
> data(map_G2F)
> data(info_environments_G2F)
> data(soil_G2F)
> METdata_g2f <- create_METData(
geno = geno_G2F,
pheno = pheno_G2F,
map = map_G2F,
climate_variables = NULL,
raw_weather_data = NULL,
compute_climatic_ECs = TRUE,
info_environments = info_environments_G2F,
soil_variables = soil_G2F,
path_to_save = " /learnMET_analyses/G2F" )
```

3.3.5 Machine learning-based models implemented

Different machine learning-based regression methods are provided as S3 classes in an object oriented programming style. These methods are called within the pipeline of the *predict_ trait_ MET_ cv()* function, that is presented in the following section. In particular, the XGBoost gradient boosting library (Chen and Guestrin, 2016), the Random Forest algorithm (Breiman, 2001), stacked ensemble models with Lasso regularization as meta-learners (Van der Laan et al., 2007), and multi-layer perceptrons using Keras (Chollet et al., 2015) are implemented as prediction methods. In this section, we briefly present how these machine learning algorithms work.

Gradient boosted decision trees (GBDT) can be seen as an additive regression model, where the final model is an ensemble of weak learners (i.e. a regression tree in this case), in which each base learner is fitted in a forward sequential manner. Considering a certain loss function (e.g. mean squared error for regression), a new tree is fitted to the residuals of the prior model (i.e. an ensemble of trees) to minimize this loss function. Then, the previous model is subsequently updated with the current model. From this definition, it becomes clear that GBDT and Random Forest models strongly differ from each other, since for GBDT, trees are built conditional on past trees, and the trees contribute unequally to the final model (Kuhn et al., 2013).

In contrast, in Random Forest algorithms, trees are created independently from each each other, and results from each tree are only combined at the end of the process. The concept of GBDT was originally developed by Friedman (2001). In *learnMET*, a set of prediction models, denoted *xgb_ reg* and *rf_ reg*, is proposed that use the XGBoost algorithm or the Random Forest algorithm, respectively, with different input variables.

A multi-layer perceptron (MLP) consists of one input layer, one or more hidden layers, and one output layer. Each layer, with the exception of the final output layer, includes a bias neuron (i.e. a constant value that acts like the intercept in a linear equation and is used to adjust the output) and is fully connected to the next layer. Here, the first hidden layer receives the marker genotypes and the ECs as input, computes a weighted linear summation of these inputs (i.e. $z = \mathbf{W}^\intercal \cdot X + b$, where X represent the input features, $\mathbf{W}^\intercal$ the vector of weights, and b the bias), and transforms the latter with a non-linear activation function $\mathbf{f}(\mathbf{z})$, yielding the output of the given neuron. In the next hidden layers, each neuron (also named node) in one layer connects with a given weight to each neuron in the consecutive layer. The last hidden layer is generally connected with a linear function to the output layer, that consists of a single node. In MLP, learning is done via backpropagation: the network makes a prediction for each training instance, calculates the error associated with this prediction, estimates the error contribution from each connection at each hidden layer by iterating backward from the last layer (reverse pass), and finally changes the connection weights to decrease this error, usually using gradient descent step (Géron, 2019). For more details about deep learning methods in genomic prediction, we refer to the review written by Pérez-Enciso and

Zingaretti (2019). In *learnMET*, a set of prediction models named *DL_ reg*, are proposed that apply multi-layer perceptrons models with different input variables.

Stacked models can be understood as an ensemble method that exploits the capabilities of many well-working models (called base learners) on a classification or regression task. The theoretical background of this method was originally proposed by Breiman (1996), and further developed by Van der Laan et al. (2007). In the first step, different individual base learners are fitted to the same training set resamples (typically generated via cross-validation), and potentially using different sets of predictor variables or different hyperparameter settings. Then, the predictions of the base learners are used as input to predict the output by fitting a regularization method, such as Lasso, on the cross-validated predictions. Hence, the final model has learned how to combine the first-level predictions of the base learners, and this stacked ensemble is expected to achieve similar or better results than any of the base learners (Van der Laan et al., 2007). This implies also that some weak learners, trained in the first stage, are generally excluded by variable selection from the resulting ensemble model if their predictions are highly correlated with other models, or irrelevant for predicting the trait of interest. In *learnMET*, prediction models named *stacking_ reg* apply stacked ensemble models with different base learners and input variables. For instance, *stacking_ reg_ 3* combines a support vector machine regression model fitted to the ECs, an elastic net model fitted to the SNPs data, and a XGBoost model using as features the 40 genomic-based PCs and the ECs. The stacked model was designed to embrace individual learners as diverse as possible, in order to improve the likelihood that the predictions of the different models are different from each other, and that the meta learning algorithm really benefits from combining these first-level predictions. Regularized regression methods are widely used for genomic selection (de los Campos et al., 2013; Zou and Hastie, 2005), thus our choice to incorporate Elastic Net as an individual learner to estimate the SNPs effects.

3.3.6 Step 2: model evaluation through cross-validation

The second function in a typical workflow is *predict_ trait_ MET_ cv()* (Box 3). The goal of this function is to assess a given prediction method with a specific CV scenario, that mimic concrete plant breeding situations.

When *predict_ trait_ MET_ cv()* is executed, a list of training/test splits is constructed according to the CV scheme chosen by the user. Each training set in each sub-element of this list is processed (e.g. standardization and removal of predictors with null variance, feature extraction based on principal component analysis), and the corresponding test set is processed using the same transformations. Performance metrics are computed on the test set, such as the Pearson correlation between predicted and observed phenotypic values (always calculated within the same environment, regardless of how the test sets are defined according to the different CV schemes), and the root mean square error. Analyses are fully reproducible given that seed and tuned hy-

perparameters are stored with the output of *predict_ trait_ MET_ cv()*. Note that, if one wants to compare models using the same CV partitions, specifying the seed and modifying the model would be sufficient.

The function applies a nested CV to obtain an unbiased generalization performance estimate. After splitting the complete dataset using an outer CV partition (based on either CV1, CV2, CV0 or CV00 prediction problems), an inner CV scheme is applied to the outer training dataset for optimization of hyperparameters. Subsequently, the best hyperparameters are selected and used to train the model using all training data. Model performance is then evaluated based on the predictions of the unseen test data using this trained model. This procedure is repeated for each training-test partition of the outer CV assignments. Table 3.1 shows the different arguments that can be adjusted when executing the cross-validation evaluation.

Note that the classes we developed for pre-processing data and for fitting machine learning-based methods use functions from the tidymodels collection of R packages for machine learning (Kuhn and Wickham, 2020), such as Bayesian optimization to tune hyperparameters (function *tune_ bayes()*) or the package *stacks*. For models based on XGBoost, the number of boosting iterations, the learning rate and the depth of trees represent important hyperparameters that are automatically tuned. Ranges of hyperparameter values are pre-defined based on expert knowledge. Bayesian optimization techniques use a surrogate model of the objective function in order to select better hyperparameter combinations based on past results (**?**). As more combinations are assessed, more data become available from which this surrogate model can learn to sample new combinations from the hyperparameter space that are more likely to yield an improvement. This technique allows a reduction of the number of model settings tested during the hyperparameter tuning.

Box 3: evaluation of a prediction method using a CV scheme (i.e. METData object with phenotypic data)

```
> res_cv0_indica <- predict_trait_MET_cv(
METData = METdata_indica,
trait = "GC",
prediction_method = "xgb_reg_1",
cv_type = "cv0",
cv0_type = "leave-one-year-out",
seed = 100,
path_folder = " /project1/indica_cv_res/cv0" )
```

3.3.7 Extracting evaluation metrics from the output

Once a model has been evaluated with a CV scheme, various results can be extracted from the returned object, as shown in Box 4, and plots for visualization of results are also saved in the *path_ folder*.

Table 3.1: **Description of the main arguments used with the function predict_trait_MET_cv()**

Function argument	Description
METData	An object created by the initial function of the package create_METData().
trait	Name of the trait to predict.
prediction_method	String to name the trait to predict.
lat_lon_included	Logical to use longitude and latitude as predictor variables. FALSE by default.
year_included	Logical to use year effect as dummy variable. FALSE by default.
cv_type	String indicating the cross-validation scheme to use among "cv0" (prediction of genotypes in new environments), "cv00" (prediction of new genotypes in new environments), "cv1" (prediction of new genotypes) or "cv2" (prediction of incomplete field trials). Default is "cv0".
cv0_type	String indicating the type of cv0 scenario, among "leave-one-environment-out", "leave-one-site-out","leave-one-year-out" and "forward-prediction". Default is "leave-one-environment-out".
nb_folds_cv1	Integer for the number of folds to use in the cv1 scheme, if selected.
repeats_cv1	Integer for the number of repeats in the cv1 scheme, if selected.
nb_folds_cv2	Integer for the number of folds to use in the cv2 scheme, if selected.
repeats_cv2	Integer for the number of repeats in the cv2 scheme, if selected.
include_env_predictors	Logical to indicate if ECs should be used in predictions. TRUE by default.
list_env_predictors	Vector of character strings with the names of the environmental predictors which should be used in predictions. NULL by default, which means that all environmental predictor variables are used.
seed	Integer with the seed value. Default is NULL, which implies that a random seed is generated, used in the other stages of the pipeline, and given as output for reproducibility.
save_processing	Logical to save the processing steps used to build the model in a RDS file. Default is FALSE.
path_folder	String to indicate the full path where the RDS file with results and plots generated during the analysis should be saved.
num_pcs	Optional argument. Integer to indicate the number of principal components to derive from the genotype matrix or from the genomic relationship matrix (encouraged to speed up cross-validation with large datasets).
save_model	Logical indicating whether the fitted model for each training-test partition should be saved. Default is FALSE.

Box 4: Extraction of results from returned object of class *met_ cv*

```
# Extract predictions for each test set in the CV scheme:
> pred_2010 <- res_cv0_indica$list_results_cv[[1]]$prediction_df
> pred_2011 <- res_cv0_indica$list_results_cv[[2]]$prediction_df
> pred_2012 <- res_cv0_indica$list_results_cv[[3]]$prediction_df

# The length of the list_results_cv sub-element is equal to the number of train/test sets partitions.

# Extract Pearson correlation between predicted and observed values for 2010:
> cor_2010 <- res_cv0_indica$list_results_cv[[1]]$cor_pred_obs

# Extract root mean square error between predicted and observed values for 2011:
> rmse_2011 <- res_cv0_indica$list_results_cv[[2]]$rmse_pred_obs


# Get the seed used:
> seed <- res_cv0_indica$seed_used
```

3.3.8 Step 3: prediction of performance for a new test set

The third module in the package aims at implementing predictions for unobserved configurations of genotypic and environmental predictors using the function *predict_ trait_ MET()* (Box 5). The user needs to provide a table of genotype IDs (e.g. name of new varieties) with their growing environments (i.e. year and location) using the argument *pheno* in the function *create_ METData()*. Genotypic data of the selection candidates to test within this test set should all be provided using the *geno* argument. Regarding characterization of new environments, the user can either provide a table of environments, with longitude, latitude and growing season dates, or can directly provide a table of ECs that should be consistent with the ECs provided for the training set. Environmental variables for the unobserved test set should be provided or computed with the same aggregation method (i.e. same *method_ ECs_ intervals*) as for the training set. To build an appropriate model with learning parameters, able to generalize well on new data, a hyperparameter optimization with cross-validation is conducted on the entire training dataset when using the function *predict_ trait_ MET()*.

This function can potentially be applied to harness historical weather data and to obtain predictions across multiple years at a set of given locations (de Los Campos et al., 2020), or to conjecture about the best selection candidates to assess in field trials at specific locations. However, we emphasize the importance of both environmental and genetic similarity between training and test sets. If the selection candidates within the test set are not strongly genetically related to the genotypes included in the training set, or if the climatic conditions experienced in the test set differ too much from the feature space covered within the training set, the prediction results might not be trustworthy for decision making.

The function *analysis_ predictions_ best_ genotypes()*, takes directly the output of *predict_ trait_ MET()* and can be used to visualize the predicted yield of the best performing genotypes at each of the locations across years included in the test set.

Box 5: prediction of new observations using a training set and a test set (i.e. phenotypic data not required)

```
# Create a training set composed of years 2014, 2015 and 2016:
> METdata_G2F_training <-
create_METData(
geno = geno_G2F,
pheno = pheno_G2F[pheno_G2F$year %in% c(2014,2015,2016), ],
map = map_G2F,
climate_variables = NULL,
compute_climatic_ECs = TRUE,
et0 = T, # Possibility to calculate reference evapotranspiration with the package (if TRUE, elevation data should be preferably added as a column in info_environments)
info_environments = info_environments_G2F[info_environments_G2F$year %in% c(2014,2015,2016), ],
soil_variables = soil_G2F[soil_G2F$year %in% c(2014,2015,2016), ],
path_to_save = " /project1/g2f_trainingset") # path where daily weather data and plots are saved

# Create a prediction set (same default method to compute ECs as above):
> METdata_G2F_new <-
create_METData(
geno = geno_G2F,
pheno = as.data.frame(pheno_G2F[pheno_G2F$year %in% 2017, ] %>% dplyr::select(-pltht, -yld_bu_ac, -earht)),
map = map_G2F,
et0 = T,
climate_variables = NULL,
compute_climatic_ECs = TRUE,
info_environments = info_environments_G2F[info_environments_G2F$year %in% 2017, ],
soil_variables = soil_G2F[soil_G2F$year %in% 2017, ],
path_to_save = " /project1/g2f_testset",
as_test_set = T) # in order to provide only predictor variables (no phenotypic data for the test set available) in pheno argument.

# Fitting the model to the training set and predicting the test set
> results_list <- predict_trait_MET(
METData_training = METdata_G2F_training,
METData_new = METdata_G2F_new,
trait = "yld_bu_ac",
prediction_method = "xgb_reg_1",
use_selected_markers = F,
save_model = TRUE,
# save_model set to TRUE in order to retrieve subsequently variable importance
lat_lon_included = F,
year_included = F,
num_pcs = 200,
include_env_predictors = T,
seed = 100,
path_folder = " /project1/g2f_results_year_2017"
)
```

3.3.9 Interpreting ML models

Compared to parametric models, ML techniques are often considered as black-boxes implementations, that complicate the task of understanding the importance of different factors (genetic, environmental, management or their respective interactions) driving the phenotypic response. Therefore, various methods have recently been proposed to aid the understanding and interpretation of the output of ML models. Among these techniques, some are model-specific techniques (Molnar, 2022), in the sense that they are only appropriate for certain types of algorithms. For instance, the Gini importance or the gain-based feature importance measures can only be applied for tree-based methods (e.g. decision trees, Random Forests, gradient boosted trees), since it calculates how much a predictor variable can reduce the sum of squared errors in the child nodes, compared to the parent node, across all splits for which this given predictor was used. Feature importances are in this case scaled between 0 and 100.

Other model-agnostic interpretation techniques have been developed, that provide the advantage of being independent from the original machine learning algorithm applied, thereby allowing straightforward comparisons across models (Molnar, 2022). After shuffling the values of a given predictor variable, the value of the loss function (e.g. root mean square error in regression problems), estimated using the predictions of the shuffled data and the observed values, can be used to obtain an estimate of the permutation-based variable importance. Fisher et al. (2019) formally defined the permutation importance for a variable j as follows: $vip^{j}_{diff} = L(\hat{y}, X_{permuted}, y) - L(\hat{y}, X_{original}, y)$, where $L(\hat{y}, X, y)$ is the loss function evaluating the performance of the model, $X_{original}$ is the original matrix of predictor variables and $X_{permuted}$ is the matrix obtained after permuting the variable j in $X_{original}$. The reason behind this approach is that, if a predictor contributes strongly to a model's predictions, shuffling its values will result in increased error estimates. On the other hand, if the variable is irrelevant for the fitted model, it should not affect the prediction error. It is recommended to repeat the permutation process to obtain a more reliable average estimate of the variable importance (Fisher et al., 2019; Molnar, 2022). Another interesting aspect of permutation-based variable importance is the possibility to calculate it using either the training or the unused test set. Computing variable importance using unseen data is useful to evaluate whether the explanatory variables, identified as relevant for prediction during model training, are truly important to deliver accurate predictions, and whether the model does not overfit. However, in the latter case, one needs to ensure that the training and test set are sufficiently related. New data might behave very differently from the data used for training without implying that the trained model is fundamentally wrong. The function *variable_importance()* enables retrieving variable importance, either with a model-specific method (via the package vip (Greenwell et al., 2020)), when available, or based on a permutation-based method (argument *type*, see Box 6), and the calculation is made by default using the training set, but can be achieved for the test set by setting the argument *unseen_data* to TRUE.

Box 6: retrieving variable importance using the fitted model and the training data

```
> fitted_split <- results_list$list_results[[1]]

# Model-specific: variable importance based on the gain as importance metric from the XGBoost model (via vip package)
> variable_importance <- variable_importance(
object = fitted_split,
path_plot = " /project1/variable_imp_trset",
type = "model_specific")

# Model-agnostic: variable importance based on 10 permutations
> variable_importance <- variable_importance(
object = fitted_split,
path_plot = " /project1/variable_imp_trset",
type = "model_agnostic",
permutations = 10)

# Model-agnostic: accumulated local effects plot
> ALE_plot_split(fitted_split,
path_plot = " /project1/ale_plots",
variable ="freq_P_sup10_2")
```

Accumulated local effects (ALE) plots, also model-agnostic, allow to examine the influence of a given predictor variable on the model prediction, conditional on the predictor value (Apley and Zhu, 2020). Compared to partial dependence (PD) plots, they provide the advantage of addressing the bias that emerges when features are correlated. While predictions are computed over the marginal distribution of predictor variables in the case of PD plots (i.e. meaning that predictions of unrealistic instances are considered), ALE plots offer a solution to this issue by considering the conditional distribution, thus avoiding to use predictions of unrealistic training observations. To build an ALE plot, the range of the explanatory variable is first split into equally-sized small windows, such as quantiles. For each window, the ALE method only considers observations, that show for this feature a value falling within the interval. Then, it computes model predictions for the upper limit and for the lower limit of the interval for these data instances, and calculates the difference in predictions. The changes of predictions are averaged within each interval, which allows to block the impact of other features. These average effects are then accumulated across all intervals and centered at 0. The function *ALE_plot_split()* yields the ALE plot for a given predictor variable. An example is provided in Box 6.

3.4 Results and Discussion

To illustrate the use of learnMET with multi-environment trials datasets, we provide here two example pipelines, both of which are available in the official package documentation. The first one demonstrates an implementation that requires no user-provided weather data, while the second pipeline shows prediction results obtained based on user-provided environmental data.

3.4.0.1 Retrieving meteorological data from NASA POWER database for each environment

When running the commands for step 1 (Box 1, Case 2) on the maize dataset, a set of weather-based variables (see documentation of the package) is automatically calculated using weather data retrieved from the NASA POWER database. By default, the method used to compute ECs uses a fixed number of day-windows (10) that span the complete growing season within each environment. This optional argument can be modified via the argument *method_ ECs_ intervals* (detailed information about the different methods can be found at `https://cjubin.github.io/learnMET/reference/get_ECs.html`). The function *summary()* provides a quick overview of the elements stored and collected in this first step of the pipeline (Box 7).

Box 7: Summary method for class METData

```
> summary(METdata_g2f)
```

Clustering analyses, that can help to identify groups of environments with similar climatic conditions and to identify outliers, were generated based on (a) only climate data; (b) only soil data (if available); and (c) all environmental variables together, for a range of values for K = 2 to 10 clusters (Figure 3.2).

3.4.1 Benchmarking two prediction methods from *learnMET* and a linear reaction norm model

Phenotypic yields were predicted by the reaction norm model proposed by Jarquín et al. (2014), thereafter denoted as G-W-GxW, that account for the random linear effects of the molecular markers (G), of the environmental covariates (W) and of the interaction term (GxW), under the following assumptions:

$$y_{ij} = \mu + g_i + w_j + gw_{ij} + \varepsilon_{ij},$$

with $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, where $\mathbf{G} = \mathbf{X}\mathbf{X}'/p$ (with p being the number of SNPs and X the scaled and centered marker matrix), $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Omega}\sigma_w^2)$, where $\mathbf{\Omega} = \mathbf{W}\mathbf{W}'/q$ (with q being the number of ECs and W the scaled and centered matrix that contains the ECs), $\mathbf{gw} \sim N(\mathbf{0}, [\mathbf{Z_g}\mathbf{G}\mathbf{Z_g}'] \circ \mathbf{\Omega}\sigma_{gw}^2)$ where $\circ$ denotes the Hadamard product (cell by cell product), $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

For additional details about the benchmark model, we refer to the original publication of Jarquín

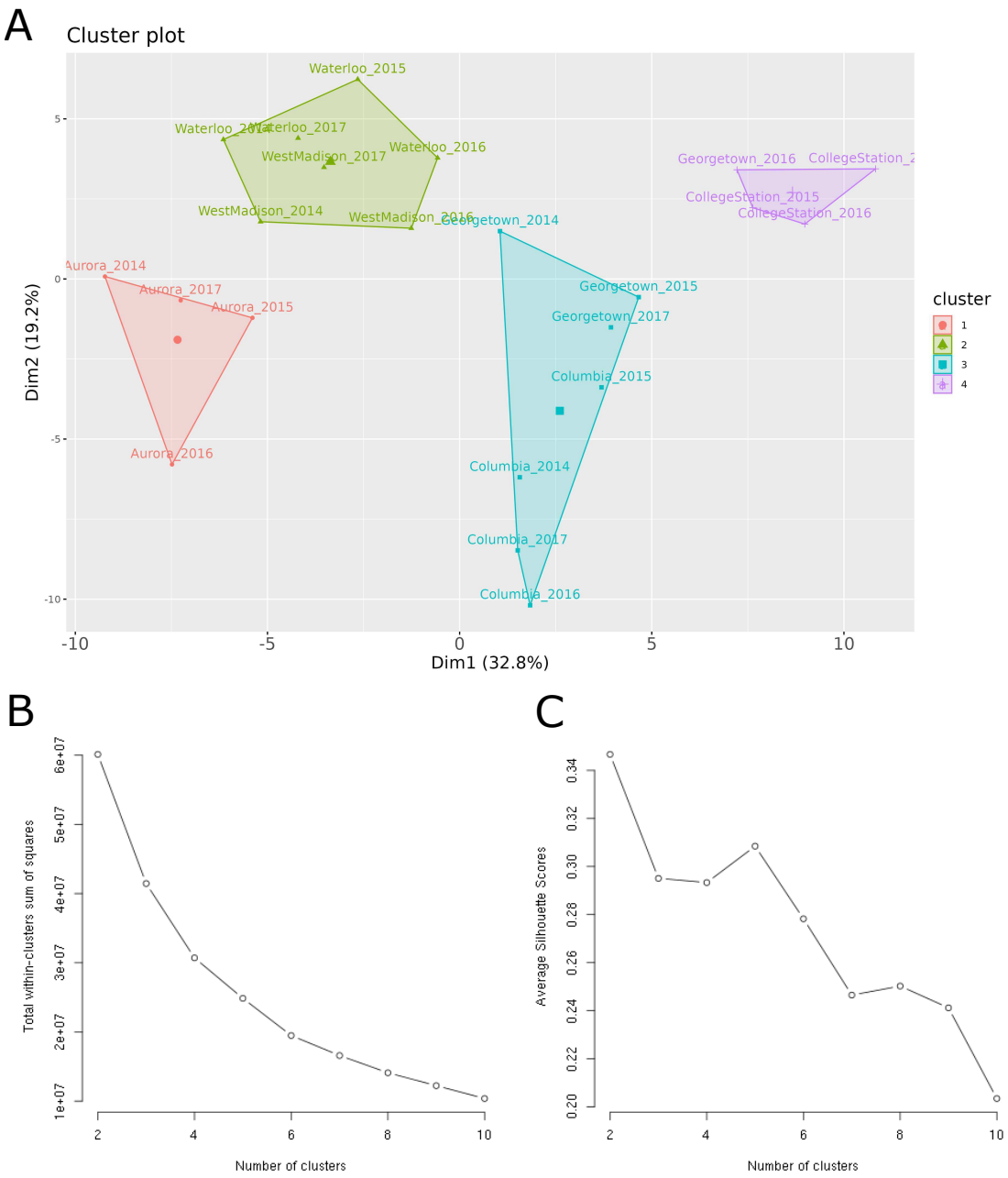


Figure 3.2: Output results from the *create_METData()* function. (A) Cluster analysis using K-means algorithm (K=4) to identify groups of similar environments based on climate and soil data. (B) Total within-cluster sum of squares as a function of the number of clusters. (C) Average Silhouette score as a function of the number of clusters. These methods can help users decide on the optimal number of clusters. Data used here is a subset of the Genomes to Fields maize dataset (AlKhalifah et al., 2018; McFarland et al., 2020). Weather data were retrieved from NASA POWER database via the package nasapower Sparks (2018). Plots are saved in the directory provided in the *path_to_save* argument.

et al. (2014). We implemented this model using BGLR (Pérez and de Los Campos, 2014), for which the MCMC algorithm was run for 20,000 iterations and the first 2000 iterations were removed as burn-in using a thinning equal to 5.

Two prediction models proposed in *learnMET* were tested: (i) *xgb_reg_1*, which is an XGBoost model that uses a certain number of principal components (PCs) derived from the marker matrix and ECs, as features and (ii) *stacking_reg_3*. Although computationally more expensive than parametric methods, we paid attention to reasonable computational time (e.g. maximum of 13.3 hours to fit *stacking_reg_3* model to n = 4,587 training instances with 10 CPUs).

We conducted a forward CV0 cross-validation scheme, meaning that future years were predicted when using only past years as the training set. For the rice datasets, at least two years of data were used to introduce variation in the EC matrix characterizing the training set (only one location was tested each year). Year, location or year-location effects were not incorporated in any of the linear and machine learning models, because we focused our evaluation on how the different models could efficiently capture the effects of SNPs and ECs, and of SNP × EC interaction effects.

Results from the benchmarking approach are presented in Figure 3.3 and in Figure 3.4. We have observed that the machine learning models are competitive with the linear reaction norm approach and tend to outperform it, albeit not consistently, as the training set size increases. Applied to small training set sizes, sophisticated prediction models are likely not able to capture informative patterns related to SNP × EC interactions, and linear models perform better. Similarly, the root mean square error was generally reduced with the machine learning methods as the training set increased (Figure 3.4). Machine learning also performed better with the G2F data, that integrated multiple locations per year and was therefore larger and probably more relevant to learn G × E patterns than with the rice dataset. Therefore, we encourage users to first evaluate whether their datasets are sufficiently large to leverage the potential of the advanced techniques proposed in this package and whether the latter provide satisfying predictive abilities in cross-validation settings.

3.4.2 Model interpretation from a gradient boosted model fitted to the maize dataset

Figure 3.5.A illustrates the permutation-based approach on the maize dataset, and Figure 3.5. Figures B and C describe how two environmental variables (sum of photothermal time and frequency of rainfall) influence the average prediction of maize grain yield using accumulated local effects plots. We should stress that the size of the dataset employed here is likely too small to make real inferences about the relationship between the predictor variables and the outcome (sharp drops observed at some feature values). Our goal here is essentially to illustrate how these functions can be used to gain insights into a model's predictions using the package.

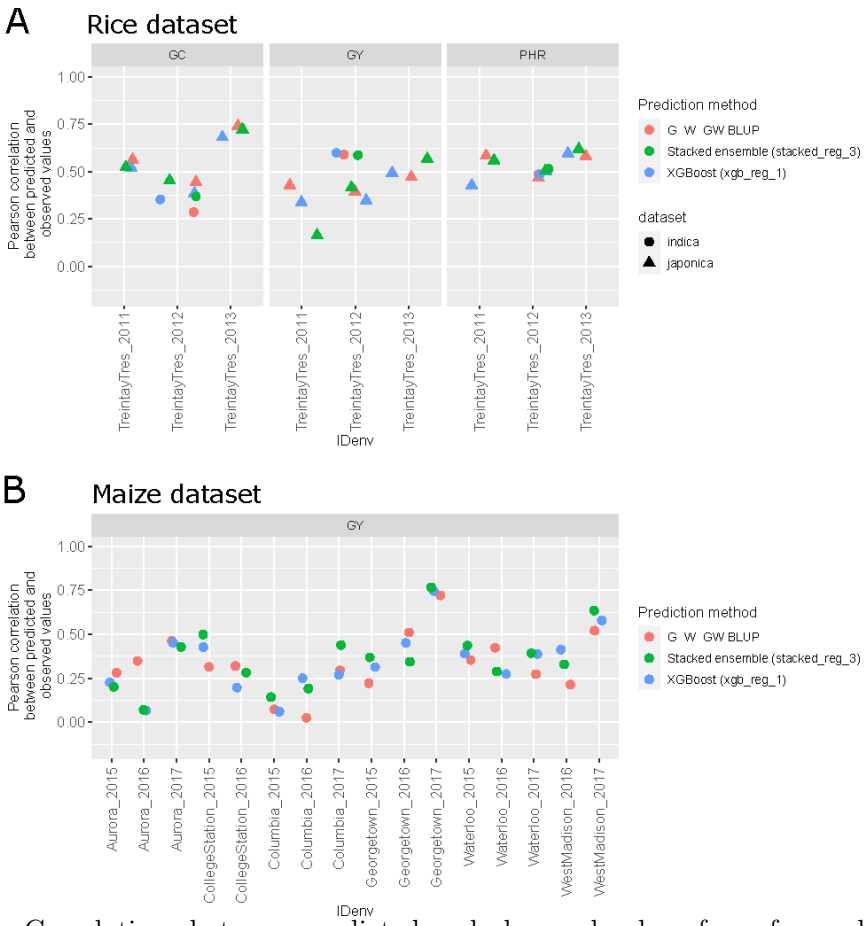


Figure 3.3: Correlations between predicted and observed values for a forward prediction scenario using two machine learning models and a linear reaction norm approach. **A** Three traits predicted for two rice populations. Each year is predicted based on at least two past years of phenotypic data (one single location). **B** Grain yield predicted for the G2F dataset. GC (rice data): percentage of chalky kernels; GY (rice data): grain yield (kg/ha); PHR (rice data): percentage of head rice recovery; GY (G2F): bushels per acre.

3.5 Concluding remarks and future developments

learnMET was developed to make the integration of complex datasets, originating from various data sources, user-friendly. The package provides flexibility at various levels: (1) regarding the use of weather data, with the possibility to provide on-site weather station data, or to retrieve external weather data, or a mix of both if on-site data are only partially available; (2) regarding how time intervals for aggregation of daily weather data are defined; (3) regarding the diversity of non-linear machine learning models proposed; (4) regarding options to provide manually specified subsets of predictor variables (for instance for environmental features via the argument *list_env_predictors* in *predict_trait_MET_cv()*).

To allow analyses on larger datasets, future developments of the package should include parallel processing to improve the scalability of the package and to best harness high performance computing resources. Improvements and extensions of stacked models and deep learning models are

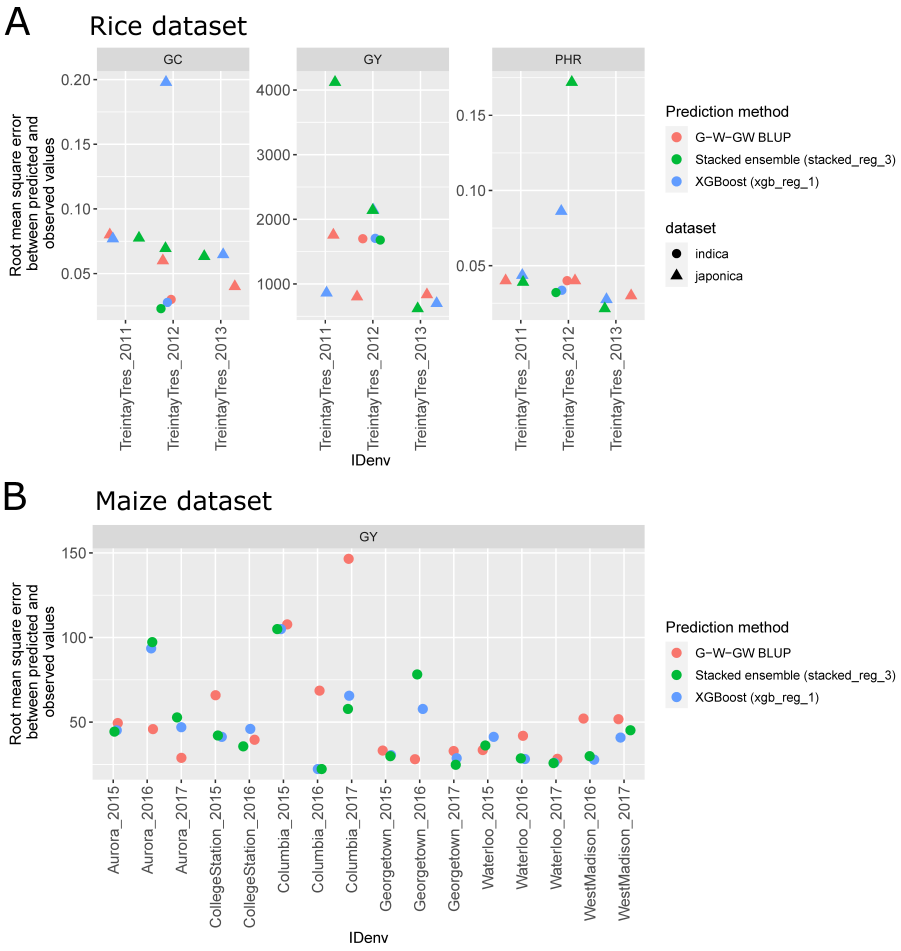


Figure 3.4: Root mean square error between predicted and observed values for a forward prediction scenario using two machine learning models and a linear reaction norm approach. **A** Three traits predicted for two rice populations. Each year is predicted based on at least two past years of phenotypic data (one single location). **B** Grain yield predicted for the G2F dataset.
GC (rice data): percentage of chalky kernels; GY (rice data): grain yield (kg/ha); PHR (rice data): percentage of head rice recovery; GY (G2F): bushels per acre.

also intended, as we did not investigate in-depth the network architecture (e.g. number of nodes per layer, type of activation function, type of optimizer), nor other types of deep learning models, that might perform better (e.g. convolutional neural networks). Finally, the package could be extended to allow genotype-specific ECs, because the timing of developmental stages differs across genotypes (e.g. due to variability in earliness) and should ideally be taken into account.

3.5.1 Data Availability

The software is available on GitHub at `https://github.com/cjubin/learnMET`. Documentation and vignettes are provided at `https://cjubin.github.io/learnMET/`. All scripts used to obtain the results presented in this paper can be found on GitHub at `https://github.com/cjubin/learnMET/tree/main/scripts_publication`.

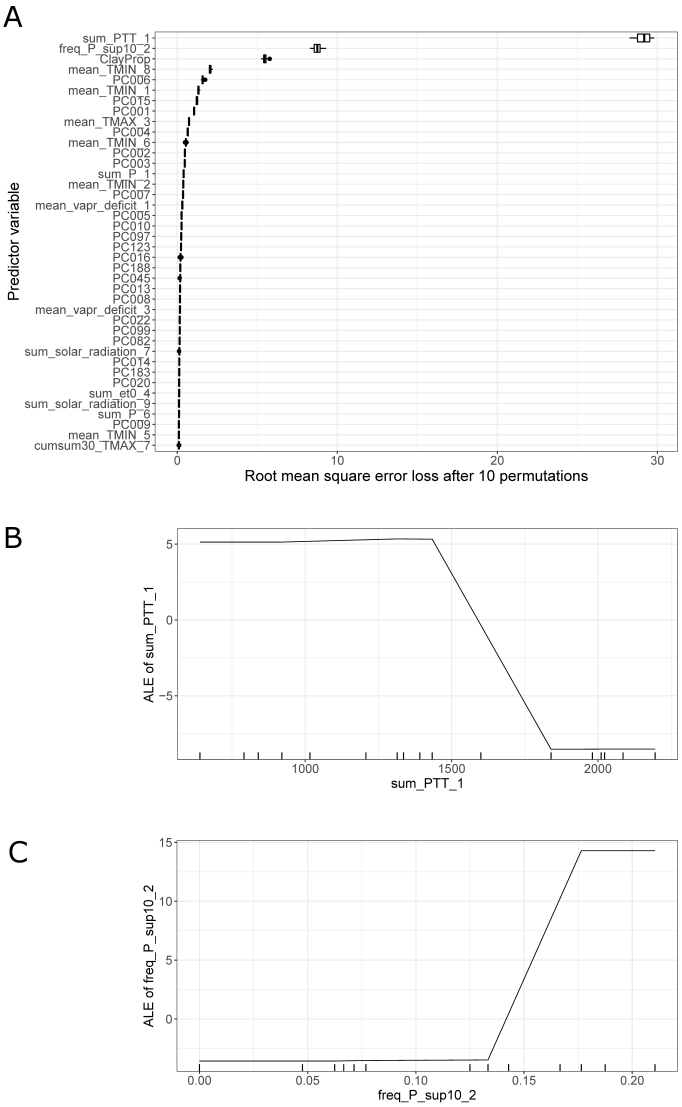


Figure 3.5: Model interpretation methods applied on the model fitted to a subset of the G2F dataset from years 2014 to 2016 (17 environments included) with *xgb_reg_1* for the trait grain yield. **A** Model-agnostic variable importance using 10 permutations. The top 40 most important predictor variables are displayed, and the table containing results across all permutations for all variables is returned. Accumulated local effects (ALE) plots for **B** sum of photothermal time during the 1^{st} day-interval of the growing season, and **C** the frequency of days with an amount of precipitation above 10 mm during the 2^{nd} day-interval of the growing season.

3.6 Acknowledgments

This work used the Scientific Compute Cluster at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

The authors would like to thank the G2F Consortium for collecting data and making these publicly available.

The authors are grateful to Eliana Monteverde for her useful input regarding the rice dataset, and also thank the National Institute of Agricultural Research (INIA-Uruguay) and technical staff from the experimental station from Treinta y Tres (Uruguay) for collecting the data.

In this work, data from the NASA POWER database were used. These data were obtained from the NASA Langley Research Center POWER Project funded through the NASA Earth Science Directorate Applied Science Program.

3.7 Funding

Financial support for C.C.W. 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.

3.8 Bibliography

AlKhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh CT, Bohn M, et al. (2018) Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC Research Notes 11(1):1–5

Apley DW, Zhu J (2020) Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(4):1059–1086

Bernardo R (2002) Breeding for quantitative traits in plants, vol 1. Stemma press Woodbury

Breiman L (1996) Stacked regressions. Machine learning 24(1):49–64

Breiman L (2001) Random forests. Machine learning 45(1):5–32

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science 52(2):707–719

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics 193(2):327–345, DOI 10.1534/genetics.112.143313, URL https://doi.org/10.1534/genetics.112.143313, https://academic.oup.com/genetics/article-pdf/193/2/327/42120942/genetics0327.pdf

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Chollet F, et al. (2015) Keras. https://keras.io

Costa-Neto G, Fritsche-Neto R, Crossa J (2021a) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106

Costa-Neto G, Galli G, Carvalho HF, Crossa J, Fritsche-Neto R (2021b) Envrtype: a software to interplay enviromics and quantitative genomics in agriculture. G3 11(4):jkab040

Covarrubias-Pazaran G (2016) Genome-assisted prediction of quantitative traits using the r package sommer. PloS one 11(6):e0156,744

Crossa J, Martini JW, Gianola D, Pérez-Rodríguez P, Jarquin D, Juliana P, Montesinos-López O, Cuevas J (2019) Deep kernel and deep learning for genome-based prediction of single traits in multienvironment breeding trials. Frontiers in genetics 10:1168

Cuevas J, Crossa J, Montesinos-López OA, Burgueño J, Pérez-Rodríguez P, de Los Campos G (2017) Bayesian genomic prediction with genotype× environment interaction kernel models. G3: Genes, Genomes, Genetics 7(1):41–53

Cuevas J, Montesinos-López O, Juliana P, Guzmán C, Pérez-Rodríguez P, González-Bucio J, Burgueño J, Montesinos-López A, Crossa J (2019) Deep kernel for genomic and near infrared predictions in multi-environment breeding trials. G3: Genes, Genomes, Genetics 9(9):2913–2924

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Annals of statistics pp 1189–1232

Géron A (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc."

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, Burgueño J, Fritsche-Neto R (2018) Bgge: a new package for genomic-enabled prediction incorporating genotype× environment interaction models. G3: Genes, Genomes, Genetics 8(9):3039–3047

Greenwell BM, Boehmke BC, Gray B (2020) Variable importance plots-an introduction to the vip package. R J 12(1):343

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop science 52(1):146–160

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics 127(3):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 x environment interactions in kansas wheat

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

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Van der Laan MJ, Polley EC, Hubbard AE (2007) Super learner. Statistical applications in genetics and molecular biology 6(1)

de Los Campos G, Pérez-Rodríguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nature communications 11(1):1–10

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, 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(1):1–6

McKinney BA, Reif DM, Ritchie MD, Moore JH (2006) Machine learning for detecting gene-gene interactions. Applied bioinformatics 5(2):77–88

Molnar C (2022) Interpretable Machine Learning, 2nd edn. URL `christophm.github.io/interpretable-ml-book/`

Montesinos-López OA, Montesinos-López A, Luna-Vázquez FJ, Toledo FH, Pérez-Rodríguez P, Lillemo M, Crossa J (2019) An r package for bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. G3: Genes, Genomes, Genetics 9(5):1355–1369

Monteverde E, Rosas JE, Blanco P, Pérez de Vida F, Bonnecarrère V, Quero G, Gutierrez L, McCouch S (2018) Multienvironment models increase prediction accuracy of complex traits in advanced breeding lines of rice. Crop Science 58(4):1519–1530

Monteverde E, Gutierrez L, Blanco P, Pérez de Vida F, Rosas JE, Bonnecarrère V, Quero G, McCouch S (2019) Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (oryza sativa l.) grown in subtropical areas. G3: Genes, Genomes, Genetics 9(5):1519–1531

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the bglr statistical package. Genetics 198(2):483–495

Pérez-Enciso M, Zingaretti LM (2019) A guide on deep learning for complex trait genomic prediction. Genes 10(7), DOI 10.3390/genes10070553, URL https://www.mdpi.com/2073-4425/10/7/553

Rincent R, Malosetti M, Ababaei B, Touzy G, Mini A, Bogard M, Martre P, Le Gouis J, van Eeuwijk F (2019) Using crop growth model stress covariates and ammi decomposition to better predict genotype-by-environment interactions. Theoretical and Applied Genetics 132(12):3399–3411

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH (2003) Optimizationof neural network architecture using genetic programming improvesdetection and modeling of gene-gene interactions in studies of humandiseases. BMC bioinformatics 4(1):1–14

Runcie DE, Qu J, Cheng H, Crawford L (2021) Megalmm: Mega-scale linear mixed models for genomic predictions with thousands of traits. Genome biology 22(1):1–25

Sparks AH (2018) nasapower: a nasa power global meteorology, surface solar energy and climatology data client for r

Westhues CC, Mahone GS, da Silva S, Thorwarth P, Schmidt M, Richter JC, Simianer H, Beissinger TM (2021) Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. Frontiers in Plant Science 12:2529, DOI 10.3389/fpls.2021.699589, URL https://www.frontiersin.org/article/10.3389/fpls.2021.699589

Wickham H, Hester J, Chang W, Hester MJ (2021) Package 'devtools'

Xavier A, Muir WM, Rainey KM (2019) bWGR: Bayesian whole-genome regression. Bioinformatics 36(6):1957–1959, DOI 10.1093/bioinformatics/btz794, URL https://doi.org/10.1093/bioinformatics/btz794, https://academic.oup.com/bioinformatics/article-pdf/36/6/1957/32915348/btz794.pdf

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology) 67(2):301–320

4. Using dynamic time warping for genomic prediction in multi-environment trials

Cathy C. Westhues[1,3,*], Johannes W. R. Martini[4], Henner Simianer[2,3] and Timothy M. Beissinger[1,3]

[1]Division of Plant Breeding Methodology, Department of Crop Sciences, University of Goettingen, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[2]Animal Breeding and Genetics Group, Department of Animal Sciences, University of Goettingen, Albrecht-Thaer-Weg 3, 37075, Goettingen, Germany

[3]Center for Integrated Breeding Research, Carl-Sprengel-Weg 1, 37075, Goettingen, Germany

[4]Aardevo B.v, Johannes Postweg 8, 8308 PB Nagele, Netherlands

* Email: cathy.jubin@uni-goettingen.de

Manuscript in preparation.

4.1 Abstract

Inter-annual variability and uncertainty of climatic conditions can strongly impact the phenotypic performance of crop cultivars in multi-environment trials and complicates accurate estimation of genotype × environment interaction (G × E) effects. The recent development of large-scale environmental data is expected to enable precise quantification of environmental contributions to

phenotypic variation, and to allow the partitioning of the full dataset into homogeneous groups of environments with reduced G × E. A second advantage of using climatic data relates to the prediction of genotype response in a new environment using as predictor variables only marker data and a set of environmental variables, thus to also assess *in silico* potential superior candidates in the context of limiting weather conditions. Yet, some questions remain regarding the best approach to reduce weather data dimensionality. Especially when the amount of phenotypic data is limited to the target trait such as grain yield, deriving environmental covariates (ECs) per growth stage is a challenging task. We introduce a new method based on dynamic time warping distance to cluster year-location cropping events in two multi-environment datasets for two major crops (wheat lines, with n = 6,716, and maize hybrids, with n = 18,325). DTW appeared as an efficient method to subdivide the MET into clusters of environments with pronounced climatic similarity, leading to a reduction of the G × E interaction variance within cluster in the wheat dataset. This result is encouraging to develop cultivars adapted to specific types of environmental conditions, more precise than based on geographical information. We then estimated an environmental relationship matrix, K_{DTW}, derived from DTW distance and used it in reaction-norm models derived from G-BLUP. Across three cross-validation (CV) schemes, we tested a serie of models characterized by increasing levels of complexity. In particular, we examined the impact of modeling dominance genetic effects in the hybrid dataset, and of integrating climatic data incorporated with either ECs, or with K_{DTW} in the two datasets. In the maize dataset, the most complex model that integrated dominance genetic effects and interactions between K_{DTW} and genomic components yielded the best predictive abilities across all CV schemes, with the maximum gain of 3.5% observed in the leave-one-environment-out (CV0) prediction scheme, while in the wheat dataset, the full model with DTW and additive genetic effects led to an improvement of ∼10% in the same CV. Our results also show the importance of considering genotype × environment interaction effects, either using climatic data or by disentangling the effects of year and location, in CV1 and CV2 with these two datasets.

Keywords: genotype-by-environment interaction, dynamic time warping, environmental covariates, reaction norm model, genomic prediction

4.2 Introduction

Global demands for food are expected to increase in the coming decades. At the same time, adverse environmental conditions are occurring more frequently. The occurrence and magnitude of heat waves or shortages of precipitation during the crop growing season will be heightened in the future, and severe yield losses for major staple crops, such as wheat (Semenov et al., 2014), maize (Hawkins et al., 2013), represent a major concern for food security in many regions of the world. Boosting the development of crop cultivars resilient to drought and heat stresses, and characterized by yield

stability under limited agricultural inputs, constitutes an effective solution to mitigate the impact of climate change and bolster agricultural production.

The breakthrough in ground environmental monitoring technologies, such as automatic weather stations (AWS), and in high-resolution satellite remote sensing technologies in agriculture have enabled to generate important amount of meteorological data. Harnessing this information in the context of plant breeding can be achieved in two principal manners. First, climatic data can be retrieved for field experiments composing the multi-environment trials (MET) network, in which candidate genotypes are grown and evaluated for their performance and stability, and subsequently used to identify homogeneous subgroups of environments. Classically, methods aiming to group environments, such as the Additive Main Effect and Multiplicative Interaction (AMMI) decomposition, exploit phenotypic data rather than environmental information. However, an important requirement underlying its implementation, is that performance data should characterize a common set of genotypes across all year-location combinations composing the MET. Thus, it is not always possible to derive meaningful distances among growing environments, when phenotypic datasets are not connected, or only weakly connected, with common genotypes across years and locations. On the other hand, employing unsupervised learning techniques, such as hierarchical clustering or k-means using as input data available climatic datasets represents a more straightforward strategy. However, the crucial question remains as to whether cultivar's response is consistent with these *a priori* clustering techniques, and whether they can efficiently capture repeatable G $\times$ E patterns observed within the phenotypic data.

Another application of weather data is to incorporate these in advanced statistical models to predict and assess the variability of the phenotypic performance for candidate genotypes under new environmental conditions. This is a prediction problem of utmost interest for plant breeders, as many breeding programs are not able to evaluate all selection candidates across all environments, and genotypes are generally advanced to further experimental trials on the basis of their performance in a very limited sample of the total target population of environments (TPE). Widely used, classical genomic prediction (GP) models capitalize on realized relationships calculated using dense molecular genotypic information, between individuals forming the training set, evaluated in field trials, and those in the prediction set, for which no phenotypic records are available. While GP models have allowed breeders to realize more rapid genetic gains for many plant and animal species, the original methodological framework of ignoring G $\times$ E interaction effects needs to be adapted in the context of MET data. Coping with G $\times$ E interactions in GP models has been carried out with the integration of explicit environmental covariates (e.g. sum of precipitation, average temperatures during the crop season) in factorial regression models (Heslot et al., 2014), or by modeling G $\times$ E with linear (Basnet et al., 2019; Jarquín et al., 2014, 2017; Jarquin et al., 2021a) and non-linear covariance functions in reaction norm models (Costa-Neto et al., 2021). Dealing with very large genomic and environmental datasets represents an additional challenge, tackled

by Heslot et al. (2014) with a variable selection approach seeking at identifying the most variable markers across environments to use for the G × E interaction term, thus allowing a reduction of the total number of interactions used in the model.

The choice of method to summarize daily weather data into ECs for incorporation in genomic prediction models is not trivial for two main reasons. First, the duration of the growing season can differ significantly across environments, as was illustrated in the Genomes to Fields dataset (AlKhalifah et al., 2018; McFarland et al., 2020). Second, some plant developmental stages have been shown to be particularly vulnerable to abiotic stresses, such as reproductive stages like silking stage in maize (Dong et al., 2021) or heading stage in wheat (Kazan and Lyons, 2016). For this reason, it can be pertinent to derive ECs based on the growth stage timing, rather than by simply segmenting the total crop growing season into day periods of equal length, has often been outlined as a useful strategy to efficiently reduce weather data dimensionality and to enhance predictive ability (Jarquin et al., 2021b; Millet et al., 2019; Westhues et al., 2021a). Nonetheless, this method requires some information about the occurrence of some phenological stages, and can be inexact when dates corresponding to key biological events (e.g. flowering dates) are not available. To our knowledge, no studies have proposed a comparison of predictive abilities obtained with an approach based on ECs derived from predicted phenological stages, versus an approach based on naive day periods.

Nonetheless, a loss of information regarding day-to-day variability necessarily arises when summarizing weather data using ECs. In this study, we propose the application of a feature engineering method, called dynamic time warping (DTW), that takes advantage of raw meteorological time series data to quantify nonlinear climatic similarity among field trials experiments. The method was proposed by Delerce et al. (2016) to group similar rice on-farm cropping events based on daily weather patterns, but has never been used, to our knowledge, in the frame of multi-environment genomic prediction. Dynamic Time Warping (DTW) is a nonlinear dynamic warping algorithm that computes a dissimilarity measure between two time series. This measure was originally developed for speech recognition (Sakoe and Chiba, 1978) and has been since then widely implemented in the field of remote sensing for analysis of satellite image time series (Petitjean et al., 2012), with applications for crop mapping (Csillik et al., 2019; Guan et al., 2016; Zhao et al., 2020), but also in finance (Fu et al., 2001), in medicine (Wismüller et al., 2002), and in bioinformatics (Aach and Church, 2001). In the context of shape-based time series similarity, DTW is a commonly used distance measure (Aghabozorgi et al., 2015). The algorithm on which its calculation is based, attempts at finding the optimal alignment between two temporal sequences. It takes into account local distortions, such as stretched or compressed time series, generally associated with differences in time scaling, time shifts or noise Lhermitte et al. (2011). Hence, the method is able to match temporal patterns that do not occur at the exact same time point in the two time series. On the other hand, the Euclidean distance or Manhattan distance, two widely used distance measures,

perform one-to-one alignment of time series, which can be brittle and fail at considering flexible similarities. It has also been shown in several studies that DTW can provide better accuracy than Euclidean distance (Aach and Church, 2001; Chu et al., 2002; Keogh and Pazzani, 2000), and can be used to compare time series of different lengths (Aghabozorgi et al., 2015). This can be advantageous when dealing with weather data acquired from weather stations which might vary in length.

The objectives of the present study were to investigate in two independent multi-environment datasets whether DTW was beneficial in the context of MET data analyses. In particular, we tested DTW for hierarchical clustering of crop growing environments, and compared the classification results with the established Köppen-Geiger climate classification and with the USDA plant hardiness classification. Then, we evaluated the new environmental similarity matrix derived from the DTW distance across various cross-validation scenarios in a suite of prediction models and compared results obtained by using environmental covariates to quantify environmental similarity.

4.3 Material & Methods

4.3.1 Genotypic, phenotypic and environmental datasets used

Grain yield phenotypic records of two large multi-environment datasets for two cereal species (wheat and maize) were collected from publicly available databases. The wheat dataset consisted of a diversity panel from advanced spring wheat lines developed at the International Maize and Wheat Improvement Center (CIMMYT). This dataset has been described and used in previous publications on genomic prediction (Li et al., 2021; Sukumaran et al., 2017) and on genome-wide association studies for flowering traits (Li et al., 2021; Sukumaran et al., 2016). The lines have been evaluated in different locations in South and Western Asia, Mexico, and North Africa. Phenotypic data from 2009-2010 and 2010-2011, as well as genomic data of the lines (21,321 markers), were downloaded from the CIMMYT shared research data repository at `https://hdl.handle.net/11529/10714`. Geographical coordinates and planting dates for 23 year-site combinations (i.e. environments), were retrieved from the supplemental data of Sukumaran et al. (2016) and Li et al. (2021). Harvest dates were not precise, so we considered an approximate date defining the end of each growing season (Table S4.1). This dataset was balanced across environments and comprised in total 6,716 phenotypic observations. Daily weather data were obtained with the package *nasapower* from a satellite-based weather system (Sparks, 2018), which was called in the pipeline of the *learnMET* package (Westhues et al., 2021b). We discarded information related to water stress patterns due to a lack of information regarding the amount and dates of irrigation.

The large multi-environment hybrid maize dataset from the Genomes to Fields initiative (AlKhalifah et al., 2018; McFarland et al., 2020), curated as described in Westhues et al. (2021a), was used.

Briefly, it consisted of 18,325 phenotypic observations across a set of 71 environments, spread across Southern Canada, Northeastern, Midwestern, and Southern US sites in years 2014-2016 (Table S4.2). Weather data collected using automatic weather stations, phenotypic, and molecular marker data were publicly available and retrieved at `https://www.genomes2fields.org/resources/`. Irrigation data were also integrated in the daily weather data. A genotypic matrix for 2,073 hybrids was built *in silico* from inbred parental line genotypes using 106,414 single nucleotide polymorphisms (SNPs). The G2F dataset is an unbalanced dataset for two primary reasons. First, the tested hybrids were assigned to the northern, midwestern, and southern locations on the basis of their maturity group. Second, different types of specific experiments were conducted across years. More details about the project can be found at `https://www.genomes2fields.org/about/project-overview/`. It should be noted that the quality control on phenotypic data allowed for the presence of outlier environments in the final datasets. The motivation behind this was to observe whether these outlier environments, characterized for instance by particular low or high average yields, and often reported by collaborators in the metadata file, could also be identified with the environmental clustering results.

4.3.2 Using dynamic time warping (DTW) distance as a dissimilarity measure among environments

Let us define two matrices, Q and R, that can present a different number of rows, m and n, respectively. Since we are considering multivariate time series, each matrix is defined by a set of V vectors of equal length, where each vector corresponds to a daily weather variable, for a given field experiment during the growing season. Maximum and minimum temperature, total precipitation (only for the maize dataset), average relative humidity, vapour pressure deficit and solar incoming radiation on daily scale were used. The time series were preprocessed to obtain z-scores by subtracting the mean from the variable and dividing by the standard deviations.

Hence, we can write q_i^v to denote the i-th element of the v-th variable of Q, and r_i^v to denote the i-th element of the v-th variable of R. Note that all time series being compared must have the same number of variables. The first step in the algorithm implies the creation of a local cost matrix, denoted *lcm*, with n x m dimensions. This matrix is estimated for each pair of time series (i.e. each pair of environments), and can be written as follows, when using the l_1 norm between two points (q,r) as local cost function:

$$lcm(i,j) = \sum_{v=1}^{V} |q_i^v - r_j^v|$$

In the next step, the DTW algorithm finds the optimum warping path over all potential warping paths:

$$DTW(Q,R) = \arg\min_{W \in \mathcal{P}} \sum_{(i,j) \in \mathcal{W}} lcm(i,j)$$

The optimum warping path is the one minimizing the cumulative distance between the two time series. As mentioned by Berndt and Clifford (1994), the number of possible warping paths $\mathcal{P}$ is prohibitively high. Hence, constraints can be used to reduce the search space for the set $\mathcal{P}$ (Sakoe and Chiba, 1978):

1. Monotonicity condition: temporal ordering in time series must be respected, $i_{k-1} \leq i_k$ and $j_{k-1} \leq j_k$
2. Continuity condition: the steps in the matrix are confined to neighbouring points, i_k - i_{k-1} ≤ 1 and j_k - $j_{k-1} \leq 1$
3. Boundary conditions: the first elements of R and Q must match, $i_1 = 1$ and $j_1 = 1$.

An example illustrating how to compute DTW between two temporal sequences is provided in Figure 4.1. To compute the DTW pairwise distances, we used the R package dtwclust (Sardá-Espinosa, 2017) using the method *dtw_basic*, with L_1 norm and the *symmetric1* step pattern, that ensures to obtain a symmetric matrix. We note D_{DTW} the matrix containing the pairwise DTW distances based on climatic data between growing environments.

4.3.3 Clustering analyses

Hierarchical, agglomerative clustering was applied to group field experiments using the DTW distance as a dissimilarity measure between the pairwise observations. As the name suggests, hierarchical clustering algorithm produces clusters at each level of the hierarchy by joining clusters from the next lower level (Hastie et al., 2009), and has been widely used for time-series clustering (Keogh and Pazzani, 1998; Oates et al., 2000). The average linkage clustering was chosen as an intergroup dissimilarity measure. Clustering analyses were conducted in R (R Core Team, 2021) with the function *hclust* and dendogram visualizations were achieved with the R package *dendextend* (Galili, 2015).

4.3.4 Evaluation of clusters

To verify that the groups defined by the clustering step presented reliable patterns in accordance with knowledge about climatic zones, we compared the DTW-based clustering with two well-known climatic classification systems. The first one was the Köppen–Geiger (KG) classification approach, which is a heuristic rule-based system, where different climate classes are determined using criteria based on temperature and observations on different vegetation types. We used the re-analyzed Köppen-Geiger digital worldwide high resolution map available at `http://koeppen-geiger.vu-wien.ac.at/present.htm` to define the classes based on data across the period 1986-2010 (Kottek et al., 2006; Rubel et al., 2017). The second system was the plant hardiness zone classification, developed by the United States Department of Agriculture (USDA), which defines 13 regions based on annual extreme minimum temperature and is available for the US at

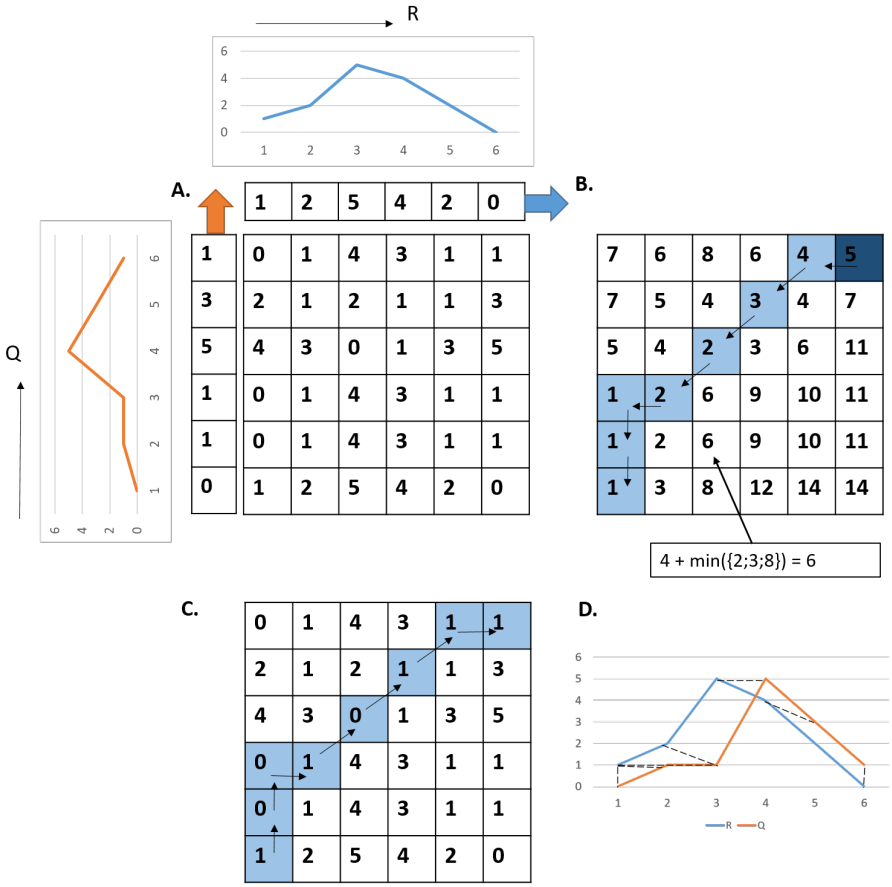


Figure 4.1: Example of calculation of the DTW distance between two time series Q and R, both of length 6.
A. The distance matrix (or local cost matrix) is computed, in which each entry is the Manhattan distance (l_1 norm) between datapoints from sequences Q and R, i.e. $d(x_i, y_j) = |(x_i - y_j)|$
B. Accumulated cost matrix.Calculation of the DTW matrix (accumulated cost matrix) as follows: $DTW(i,j) = d(x_i, y_j) + min(DTW(i-1, j-1), DTW(i, j-1), DTW(i-1, j))$.
C. The optimal (minimum) warping path is given by the arrow directing along the blue boxes. Restrictions on the warping path: it starts from (1,1) and ends at (6,6). One step is taken at a time. A warping path aligns the two temporal sequences, such that all datapoints are matched with at least one datapoint of the other sequence. The final DTW distance, computed with all the points falling on the optimum path, is 5 in this example. The Manhattan norm between R and Q moves along the main diagonal (one-to-one matching) and has a cost equal to 9 in this example.
D. DTW alignment.

`https://planthardiness.ars.usda.gov/`. Hence, this latter classification was used only for the G2F maize dataset.

The V-measure score was used to assess how similar are two completely independent clustering approaches (Rosenberg and Hirschberg, 2007) and has been used by Netzel and Stepinski (2016) to estimate the degree of similarity between two climatic clustering methods. This score presents several advantages: the two label assignment strategies do not need to share the same number of groups, and the score is not related to the absolute values of the labels. Rosenberg and Hirschberg (2007) define two concepts for homogeneity and completeness that one needs in order to calculate to estimate the V-measure score. Given ground truth labels (here, either the KG or the hardiness map labels), the homogeneity criteria is fulfilled when the algorithm assigns within a cluster *only* data points that belong to a single class (e.g. the true label). The homogeneity is defined as follows:

$$h = \begin{cases} 1 & \text{if } H(C,K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{else} \end{cases} \tag{4.3.1}$$

where K denotes a set of clusters obtained via hierarchical clustering using the DTW distance, and C a set of classes from the Köppen-Geiger or USDA hardiness classification systems. The completeness criteria is fulfilled when *all* data points that belong to a same unique class are assigned to a same unique cluster. The completeness is defined as follows:

$$c = \begin{cases} 1 & \text{if } H(K,C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{else} \end{cases} \tag{4.3.2}$$

The V-measure score can then be calculated as the weighted harmonic mean of homogeneity and completeness:

$$V_\beta = \frac{(1+\beta) * h * c}{\beta * h + c} \tag{4.3.3}$$

We used the default value of 1 for β.

In addition, for the wheat dataset, we computed pairwise Euclidean distances between environments using marker effects as environment predictors, similar to the approach proposed by Heslot et al. (2013) for environmental characterization without using environmental data. The marker effects were estimated with ridge regression, a marker-homogeneous shrinkage method, in BGLR (Pérez and de Los Campos, 2014). We decided not to conduct the same analysis on the maize dataset, since the genetic composition of hybrids tested across the environments depended on major environmental differences (e.g. different lengths of the crop growing season), and the marker-effects

based correlation could therefore be affected by the amount of genetic relatedness between environments.

4.3.5 Computing environmental covariates

Our objective was to compare the performance of the DTW-based environmental similarity matrix, that we described in the previous section, with classical approaches that estimate an environmental similarity matrix based on a set of environmental covariates (ECs) related to abiotic stresses calculated over day periods. A total of 11 climatic covariates were considered using the pipeline implemented in the package *learnMET* (Westhues et al., 2021b) (Table 4.1).

Although we had access to flowering data recorded for each genotype in the maize dataset, we decided to consider a more general and simplified situation, where one only has knowledge about the approximate required growing degree units (GDUs) to flowering or to physiological maturity for each environment. This method allows to calculate calendar dates for relevant distinct developmental phases based on the daily accumulated GDUs within each environment. GDUs were calculated considering a base temperature of 0 °C and 10 °C, and a maximum temperature of 35 °C and 30 °C, for wheat and corn, respectively. These temperatures were chosen based on crop physiology standards (Bauer et al., 1984; Swan et al., 1987). To derive the phenologically-informed set of ECs, the main function *create_METData* of the package *learnMET* was used, with the argument *method_ECs_intervals* set to *GDD*. For the maize dataset, we determined three sets of growth stages based on observed differences among environments for thermal time to silking (Figure S4.2). Regarding the wheat dataset, two groups of maturity were similarly defined, by exploiting information related to thermal time to heading from the study of Sukumaran et al. (2016). Thus, for each species dataset, the calendar dates coinciding with the beginning of a new developmental stage based on GDUs accumulation could be subsequently estimated according to the estimated GDUs requirements within each environment (Table S4.3, Table S4.4). By applying this method, 6 and 9 development periods were determined for the maize and wheat datasets, respectively, and ECs were calculated for each of them according to Table 4.1.

To obtain the second naive set of ECs, the argument *method_ECs_intervals* was set to *fixed_nb_windows_across*
to divide the growing season into 10 windows of equal length within each environment. For instance, if the growing season spanned 155 days in a given environment, 10 consecutive non-overlapping windows of 15 days were generated. Note that a maximum of 9 days could remain unused in ECs calculations. Thus, for the two crop datasets, each environment was characterized by 10 intervals covering a certain number of days.

Table 4.1: List of 13 environmental covariates used in the study.

Acronym	Description	Dataset (wheat: W or maize: M)
mean_TMIN	Average minimum temperature for the respective day-window or predicted phenological stage (°C)	W, M
mean_TMAX	Average maximum temperature for the respective day-window or predicted phenological stage(°C)	W, M
mean_TMEAN	Average mean temperature for the respective day-window or predicted phenological stage(°C)	W, M
freq_TMAX_sup30	Frequency of days with a maximum temperature > 30 °C for the respective day-window or predicted phenological stage	W, M
freq_TMAX_sup35	Frequency of days with a maximum temperature > 35 °C for the respective day-window or predicted phenological stage	W, M
freq_TMAX_sup40	Frequency of days with a maximum temperature > 40 °C for the respective day-window or predicted phenological stage	W
cumsum30_TMAX	Sum of the daily maximum temperature > 30 °C for the respective day-window or predicted phenological stage	W
sum_PTT	Cumulative photothermal time (daily growing degree-days × day length in hours) for the respective day-window or predicted phenological stage	W, M
sum_P	Precipitation (with irrigation, if indicated) for the respective day-window or predicted phenological stage	M
freq_P_sup10	Frequency of days with total precipitation > 10 mm for the respective day-window or phenological stage	M
sum_solar_radiation	Accumulated incoming global solar radiation ($MJ.m^{-2}.d^{-1}$) for the respective day-window or predicted phenological stage	W, M
mean_vapr_deficit	Average vapour pressure deficit (kPa) for the respective day-window or predicted phenological stage	W, M
freq_TMIN_inf_minus5	Frequency of days with a minimum temperature < -5 °C for the respective day-window or predicted phenological stage	M

4.3.6 Statistical prediction models

Prediction models with different components were fitted for the maize and the wheat datasets, respectively, and are presented in the following section. The different components without weather data included E, environment (field trial) main effect; Y, year effect; L, location effect; A, additive genetic effect; D, dominance genomic effect; A × E, additive × environment interaction effect; A × Y, additive × year interaction effect; A × L, additive × location interaction effect and DE, dominance × environment interaction effect.

To specify how the climatic data were integrated in the models, the following acronyms will be used: $\mathbf{K_{DTW}}$, DTW-based environmental similarity matrix; $\mathbf{W_{ECs\_stages}}$, ECs based on estimated phenological stages and $\mathbf{W_{ECs\_windows}}$, ECs based on the 10 windows within each environment. Covariance between environments derived from the DTW distance was computed as: $K_{DTW} = 1_E - \frac{D_{DTW}}{max(D_{DTW})}$, where 1_E is a matrix of size $(N_E \times N_E)$, with N_E the total number of environments in the MET dataset, and $N_E = 23$ and $N_E = 71$ for the wheat and maize datasets, respectively. $\mathbf{W_{ECs\_stages}}$ and $\mathbf{W_{ECs\_windows}}$ have dimensions $N_E \times q$, with q the number of ECs that depended on the method used to summarize the daily weather data and on the species dataset (see section 4.3.5). Importantly, $\mathbf{W_{ECs\_stages}}$ and $\mathbf{W_{ECs\_windows}}$ were centered and scaled using the training data, and the same transformation was applied to the respective test set, before fitting one of the following statistical models to the training data and predicting the test set. Covariates with null variance were removed.

Interaction terms incorporating weather data were denoted as follows in the models: AW, additive

× climate ECs interaction effect; DW, dominance × climate ECs interaction effect; A × K_{DTW}, additive × DTW-based environmental kinship and D × K_{DTW} DTW-based environmental kinship.

4.3.6.1 Models with main effects (without weather data)

M1: Additive genetic and environment main effects (A + E)

The phenotypic response (y_{ij}) of the j^{th} genotype (which is a maize hybrid or a wheat line in our study) in the i^{th} environment was defined as:

$$y_{ij} = \mu + E_i + a_j + \varepsilon_{ij}, \tag{4.3.4}$$

where μ is the grand mean, E_i is the random effect of the i^{th} environment, a_j is the random effect of the additive genetic value of the j^{th} genotype linked to genomic markers, and ε_{ij} is the random error term. All random effects follow an independent and identically distributed (iid) multivariate normal distribution: $E_i \overset{IID}{\sim} N(0, \sigma_E^2)$, $\mathbf{a} \tilde{N} (\mathbf{0}, \mathbf{A}\sigma_a^2)$ and $\varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$.

Specifically, the additive genetic random effect of the genotype j^{th}, noted a_j, is expressed as the regression on marker data $a_j = \sum_{m=1}^{p} x_{jm} b_m$, where x_{jm} is the genotype of the j^{th} genotype at the m^{th} molecular marker, and b_m is the random effect of the m^{th} marker such that $b_m \overset{IID}{\sim} N(0, \sigma_b^2)$ (m=1,...,p), with σ_b^2 the variance of the marker effects. Based on the assumptions of the G-BLUP model (VanRaden, 2008), the vector $\mathbf{a} = (a_1, ..., a_J)'$ contains the additive genetic values of all the genotypes and follows a multivariate normal density with zero mean and with a covariance matrix $Cov(\mathbf{a}) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the additive genomic relationship matrix computed as $\mathbf{A} = \mathbf{X}\mathbf{X}'/p$. Here, $\mathbf{X}$ is the standardized genotype matrix containing molecular markers coded as counts of the minor allele (0, 1, 2).

This baseline model allows to borrow information across genotypes based on their degree of additive genetic relationship derived from marker data, but no information related to correlations among environments is used.

M2: Additive genetic, location, year, and environment main effects (A + L + Y + E)

The model M1 assumes that field trials correspond to independent year-location combinations (E). In M2, the main year (Y) and location (L) effects were added, under the assumptions $Y_s \overset{IID}{\sim} N(0, \sigma_S^2)$, and $L_k \overset{IID}{\sim} N(0, \sigma_K^2)$:

$$y_{skj} = \mu + Y_s + L_k + E_{sk} + a_j + \varepsilon_{skj}, \tag{4.3.5}$$

Compared to model 4.3.4, this model enables exploiting information about genotypes tested at a same location or in a common year. E was still incorporated to model potential trial-specific effects, that could not be accounted for by year nor location effects (e.g. field management).

M3: Additive and dominance genetic effects, with environment main effect (A + D + E)

This baseline model includes the dominance-deviation effects, as follows:

$$y_{ij} = \mu + E_i + a_j + d_j + \varepsilon_{ij}, \tag{4.3.6}$$

where d_j is the random effect of the dominance genetic value of the j^{th} genotype linked to genomic markers, with $\mathbf{d} \sim N(\mathbf{0}, \mathbf{D}\sigma_d^2)$, where $\mathbf{D}$ is the genomic matrix of dominance deviations, calculated as $\mathbf{D} = \mathbf{SS}'$, where $\mathbf{S}$ is the centered and scaled marker allele matrix for dominance effects, obtained by assigning all homozygous genotypes to 0 in the original hybrid genotypic matrix, while heterozygous genotypes remain coded as 1. This model was only implemented for the maize hybrids dataset.

4.3.6.2 Models with interactions (without weather data)

M4: Additive main genetic effect, environment main effect, and additive × environment interaction (A + E + A × E)

The baseline model 4.3.4 was extended to include the interaction term between environments and molecular markers (G ×E) using covariance structures, as demonstrated by Jarquín et al. (2014):

$$y_{ij} = \mu + E_i + a_j + aE_{ij} + \varepsilon_{ij} \tag{4.3.7}$$

with $\mathbf{aE} \sim N(\mathbf{0}, [\mathbf{Z_a A Z_a'}] \circ [\mathbf{Z_E Z_E'}]\sigma_{aE}^2), \varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma_\varepsilon^2)$, where $\mathbf{Z_a}$ and $\mathbf{Z_E}$ are the design matrices that connect the phenotype entries with genotypes (wheat lines or maize hybrids) and with environments, respectively; σ_{aE}^2 is the variance component of the aE_{ij} interaction term; and $\circ$ denotes the cell-by-cell product (Hadamard) between two matrices. The component aE_{ij} corresponds to the interaction between each marker, from the additive genomic relationship, and each environment.

M5: Additive main genetic effect, year, location and environment main effect, additive × year interaction, additive × location interaction, and additive × environment interaction (A + Y + L + E + A × Y + A × L + A × E)

The model 4.3.5 was extended to include genotype × year, genotype × location, and genotype × environment interaction effects:

$$y_{jsk} = \mu + a_j + Y_s + L_k + E_{sk} + aY_{jsk} + aL_{jsk} + aE_{jsk} + \varepsilon_{jsk}, \tag{4.3.8}$$

M6: Additive and dominance main genetic effects, environment main effect, additive × environment interaction, dominance × environment interaction effects (A + D + E + A × E + D × E)

Many authors reported an increase of predictive ability in maize by including dominance effects with a genomic relationship matrix modeling dominance deviations (Costa-Neto et al., 2021; Jarquin et al., 2021a; Rogers and Holland, 2021). As proposed by Costa-Neto et al. (2021), model 4.3.7 was extended to incorporate dominance main effect and dominance × environment interaction effect, as follows:

$$y_{ij} = \mu + E_i + a_j + d_j + aE_{ij} + dE_{ij} + \varepsilon_{ij} \tag{4.3.9}$$

with $\mathbf{dE} \sim N(\mathbf{0}, [\mathbf{Z_d D Z_d'}] \circ [\mathbf{Z_E Z_E'}]\sigma^2_{dE}), \varepsilon_{ij} \overset{IID}{\sim} N(0, \sigma^2_\varepsilon)$, where $\mathbf{Z_d}$ is the design matrix that connects the phenotype entries with hybrid genotype, and σ^2_{dE} is the variance component of the dE_{ij} interaction term. The component dE_{ij} corresponds to the interaction between each dominance-deviation and each environment. Only the maize dataset was tested with this model.

4.3.6.3 Models with main effects (with weather data)

M7: Additive main genetic effect, environment main effect, and ECs main effect (A + E + W)

Environmental data extracted from daily weather data was included as a main effect as a random regression on the ECs, $w_i = \sum_{q=1}^{Q} W_{iq}\gamma_q$, where W_{iq} is the value of the q^{th} EC evaluated in the i^{th} environment, γ_q is the main effect of the corresponding EC, and Q is the total number of ECs. In this study, W_i was replaced with $\mathbf{W_{ECs\_stages}}$ or $\mathbf{W_{ECs\_windows}}$. Here, W_i was only dependent on the environment (and not on the genotype), because we considered that all genotypes encountered the same environmental conditions in an environment. Because the climatic data we used cannot not fully characterize differences across environments (no soil information included in this study, and only linear interactions are considered), we still included the location-year (environment) effect:

$$y_{ij} = \mu + E_i + a_j + w_i + \varepsilon_{ij}, \tag{4.3.10}$$

where the vector $\mathbf{w} = \mathbf{W}\boldsymbol{\gamma}$ follows a multivariate normal distribution with null mean and covariance matrix $\boldsymbol{\Omega}\sigma^2_w$, where $\boldsymbol{\Omega} \propto \mathbf{WW'}/q$, i.e. the entries in Ω are calculated in the same way as those of the $\mathbf{G}$ matrix.

M8: Additive main genetic effect, environment main effect, environment effect derived from DTW distance, K_{DTW} (A + E + K_{DTW})

The environmental relationship matrix obtained from the DTW distance was used as a main random effects:

$$y_{ij} = \mu + E_i + a_j + k_i + \varepsilon_{ij}, \tag{4.3.11}$$

with $k_i \tilde{N} (0, K_{DTW}\sigma^2_{DTW})$

M9: Additive and dominance main genetic effect, environment main effect, and ECs main effect (A + D + E + W)

Model 4.3.10 was extended to incorporate dominance effects, for the maize dataset:

$$y_{ij} = \mu + E_i + a_j + d_j + w_i + \varepsilon_{ij} \tag{4.3.12}$$

M10: Additive and dominance main genetic effect, environment main effect, and environment effect derived from DTW distance, K_{DTW} (A + D + E + K_{DTW})

Model 4.3.10 was extended to incorporate dominance effects, for the maize dataset:

$$y_{ij} = \mu + E_i + a_j + d_j + k_i + \varepsilon_{ij} \tag{4.3.13}$$

4.3.6.4 Models with interactions (with weather data)

M11: Additive main genetic effect, environment main effect, ECs main effect, and A × W interaction effect (A + E + W + A × W)

This model added the interaction between genetic markers (coded for additive effects) and environmental covariates. Jarquín et al. (2014) showed that predictive ability could be enhanced by including this interaction term, described by a covariance structure equal to the Hadamard product of the entries of $\mathbf{\Omega}$, the covariance-variance matrix corresponding to the relationship between environments and $\mathbf{A}$, the covariance-variance matrix of additive genetic relationship between genotypes. This first-order multiplicative component induces a reaction norm model. Thus,

$$y_{ij} = \mu + E_i + a_j + w_i + aw_{ij} + \varepsilon_{ij}, \tag{4.3.14}$$

with $\mathbf{aw} \sim N(\mathbf{0}, [\mathbf{Z_a A Z'_a}] \circ \mathbf{\Omega}\sigma^2_{aw})$. The vector $\mathbf{aw}$ represents the random effect of interaction terms between markers and ECs, and is assumed to follow a multivariate normal distribution with null mean and covariance structure $[\mathbf{Z_a A Z'_a}] \circ \mathbf{\Omega}$.

M12: Additive main genetic effect, environment main effect, ECs main effect, A × E interaction effect, and A × W interaction effect (A + E + W + A × E + A × W)

We extend the model 4.3.14 by including the A × E deviations, which account for genotype-by-environment interactions which are not captured by the interactions between genotype and weather-based covariates:

$$y_{ij} = \mu + E_i + a_j + w_i + aw_{ij} + aE_{ij} + \varepsilon_{ij}, \tag{4.3.15}$$

with $\mathbf{ak} \sim N(\mathbf{0}, [\mathbf{Z_a A Z'_a}] \circ \mathbf{K_{DTW}}\sigma^2_{ak})$.

M13: Additive main genetic effect, environment main effect, environment main effect from K_{DTW}, and $\mathbf{A} \times K_{DTW}$ interaction effect ($\mathbf{A} + \mathbf{E} + K_{DTW} + \mathbf{A} \times K_{DTW}$)

We replaced in 4.3.14 the covariance matrix Ω by the environmental relationship matrix defined based on DTW, $\mathbf{K_{DTW}}$:

$$y_{ij} = \mu + E_i + a_j + k_i + ak_{ij} + \varepsilon_{ij}, \tag{4.3.16}$$

with $\mathbf{ak} \sim N(\mathbf{0}, [\mathbf{Z_a A Z_a'}] \circ \mathbf{K_{DTW}} \sigma^2_{ak})$.

M14: Additive and dominance main genetic effect, environment main effect, ECs main effect, $\mathbf{A} \times \mathbf{W}$ and $\mathbf{D} \times \mathbf{W}$ interaction effects using ECs ($\mathbf{A} + \mathbf{D} + \mathbf{E} + \mathbf{W} + \mathbf{A} \times \mathbf{W} + \mathbf{D} \times \mathbf{W}$)

The model 4.3.13 was extended to incorporate dominance $\times$ environment interaction effects for the maize dataset:

$$y_{ij} = \mu + E_i + a_j + w_i + aw_{ij} + dw_{ij} + \varepsilon_{ij}, \tag{4.3.17}$$

with $\mathbf{dw} \sim N(\mathbf{0}, [\mathbf{Z_d D Z_d'}] \circ \mathbf{\Omega} \sigma^2_{dw})$.

M15: Additive and dominance main genetic effect, environment main effect, environment main effect from on K_{DTW}, $\mathbf{A} \times K_{DTW}$ and $\mathbf{D} \times K_{DTW}$ interaction effects using DTW distance ($\mathbf{A} + \mathbf{D} + \mathbf{E} + K_{DTW} + \mathbf{A} \times K_{DTW} + \mathbf{D} \times K_{DTW}$)

The model 4.3.15 was extended to incorporate dominance $\times$ environment interaction effects for the maize dataset:

$$y_{ij} = \mu + E_i + a_j + k_i + ak_{ij} + dk_{ij} + \varepsilon_{ij}, \tag{4.3.18}$$

with $\mathbf{dk} \sim N(\mathbf{0}, [\mathbf{Z_d D Z_d'}] \circ \mathbf{K_{DTW}} \sigma^2_{dk})$.

In the suite of models presented above, environmental main effects were modeled using two terms, one related to the weather data (using ECs or DTW distance), and one due to deviations from the Year-Location combination effect which cannot be accounted for by the weather data only.

4.3.7 Implementation of the models

The R package BGLR (Pérez and de Los Campos, 2014) was used to implement the models, for which the MCMC algorithm was run for 20,000 iterations and the first 2000 iterations were removed as burn-in using a thinning equal to 5.

4.3.8 Evaluation of the predictive ability

Following Burgueño et al. (2012) and Jarquín et al. (2017), three types of cross-validation (CV) schemes were considered to evaluate the model's predictive ability (PA), that aimed at simulating real prediction schemes relevant for plant breeders in the context of multi-environment trials. CV1, aimed at predicting newly developed genotypes, meaning that no phenotypic records of these lines was included in the training set. Training and test sets were obtained by assigning lines to 5 folds, so that approximately 20% of the genotypes were contained in each fold, and with all phenotypic records of a given genotype assigned to the same fold. The second CV scheme, CV2, was used to assess the ability of each model to predict the performance of a certain proportion of genotypes, using as training data phenotypic records of the same genotypes evaluated in other environments, and from related genotypes in the same environment. In this prediction scenario, the complete dataset of location-year yield BLUEs was randomly divided into five folds, which means that phenotypic evaluations of a given genotype were potentially assigned to different folds. For both CV1 and CV2, each fold was independently predicted using the four remaining folds, and the process was repeated 10 times, yielding a total of 50 training-test set partitions. With regard to the maize data, due to the highly unbalanced structure of the G2F dataset (i.e. not each genotype was repeated in each Year-Location combination), it was not possible to obtain equal sample sizes within each fold for each environment. For this dataset, the average number of observations per environment, within one testing set, was equal to 52 in the two CV schemes.

The third CV (CV0) corresponded to a leave-one-environment-out prediction problem. Therefore, the number of training-test partitions was equal to the number of environments in each MET, i.e. 71 and 23 for the maize and wheat datasets, respectively.

For all CV schemes, PA was calculated as the Pearson correlation between predicted test set values and observed yield within each year-location IDs. For the CV1 and CV1, the average within-environment correlation from the 50 training-testing set partitions was calculated. For CV0, each environment represented one fold (i.e. one test set), so only one value per environment was obtained. Importantly, the exact same training-test partitions were used to evaluate all models presented in section 4.3.6.

To evaluate statistical differences between prediction models, we applied a paired t-test to Fisher's transformed (r) Pearson correlation coefficient. Following the approach described by de Los Campos et al. (2020), we calculated the Fisher's transformation as $z = \frac{\sqrt{n_{ij}-3}}{2} log(\frac{1+r}{1-r})$, n_{ij} being the number of records in the given year-location, for each testing set, per model and per environment. We used the function orderPValue from the R package *agricolae* (Felipe de Mendiburu and Muhammad Yaseen, 2020) to group models that were significantly different from each other ($\alpha = 5\%$).

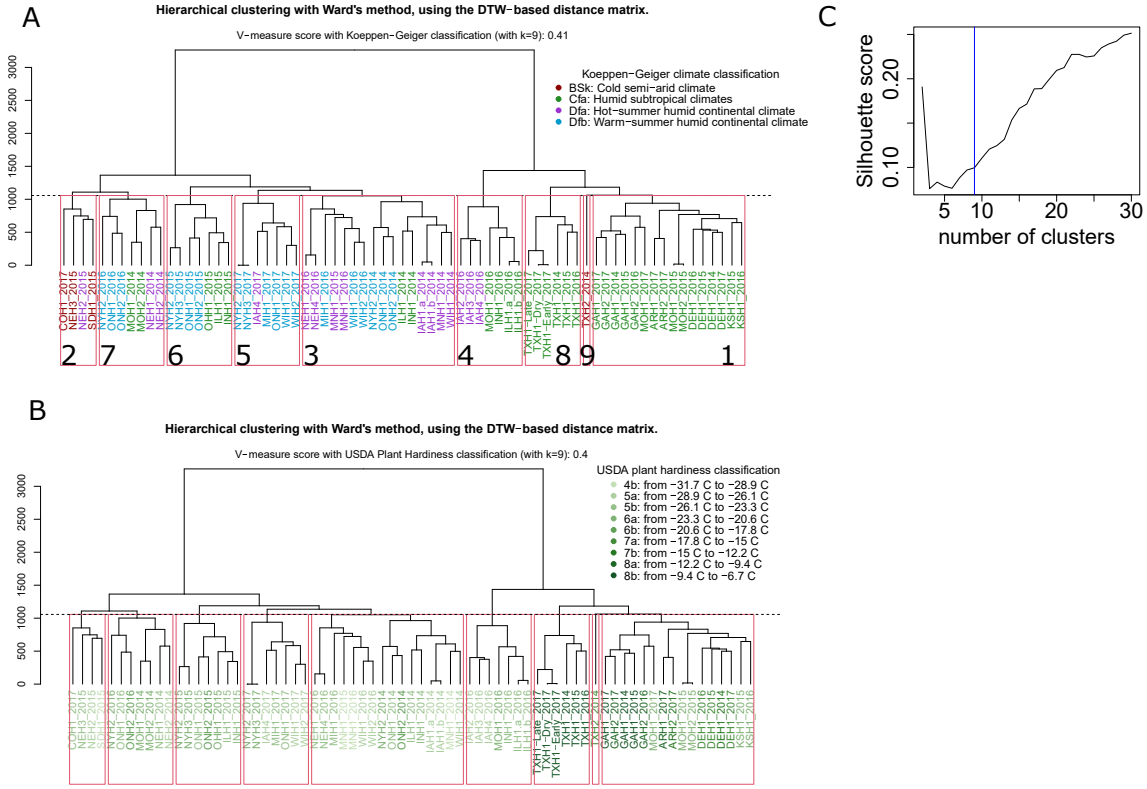


Figure 4.2: Dendogram of environments based on DTW distance calculated using weather time series. **A** Ward's hierarchical clustering given DTW pairwise distances. Environments are colored according to their class in the Koeppen-Geiger climate classification. **B** Dendogram on DTW distance with environments colored according to the class in the USDA plant hardiness system. **C** The silhouette score (Rousseeuw, 1987) was used to determine the optimal number of groups.

4.3.9 Code availability

All scripts and public datasets used in our analyses are available at `https://github.com/cjubin/DTW_paper`.

4.4 Results

4.4.1 Environmental clustering using dynamic time warping identifies groups consistent with classical climate classification systems

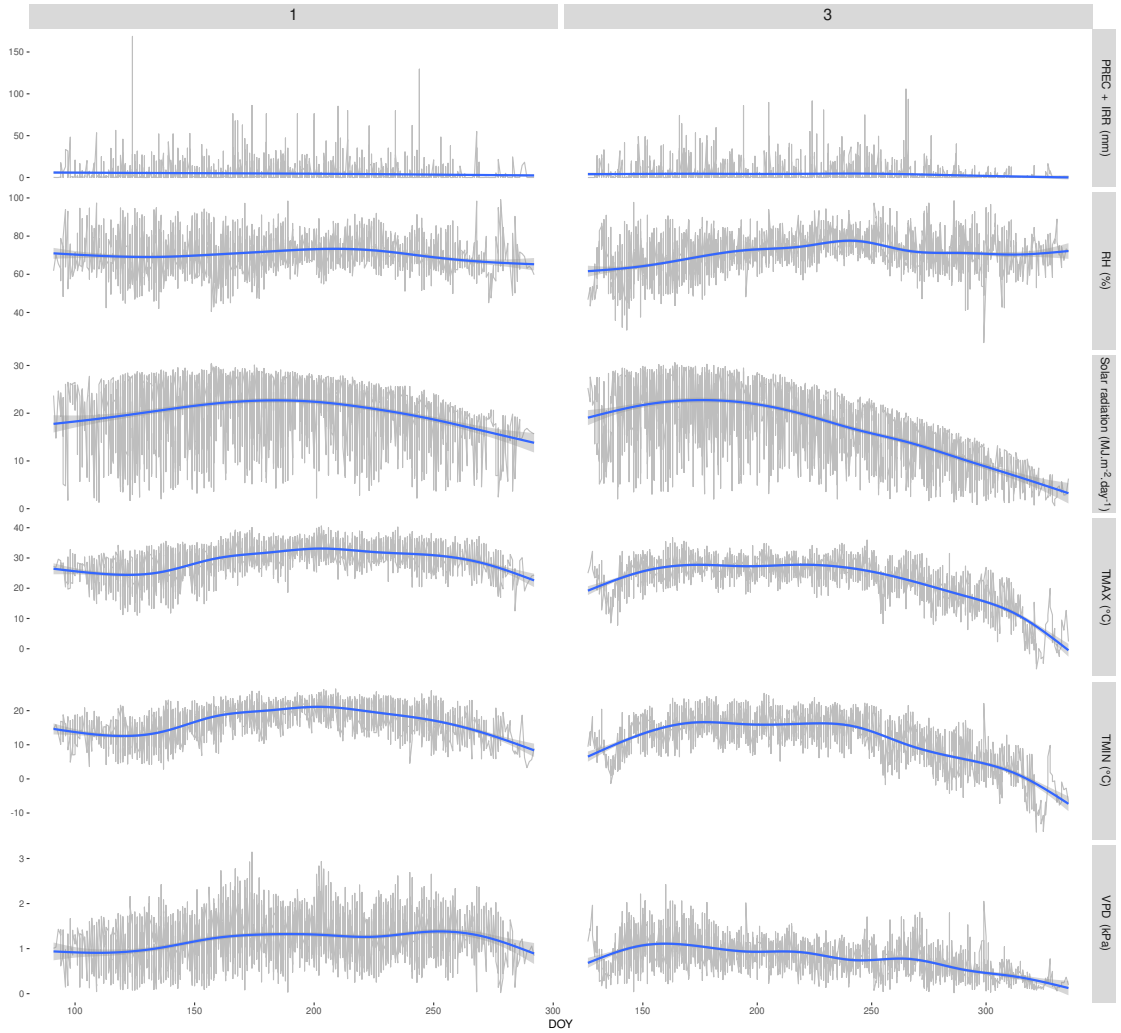


Figure 4.3: Daily weather time series for 6 climatic variables (in rows) characterizing the clusters 1 and 3 (in columns) in the maize dataset. Individual time series for each environment are depicted in grey, and the blue line represents the loess smoothing function to help seeing patterns.

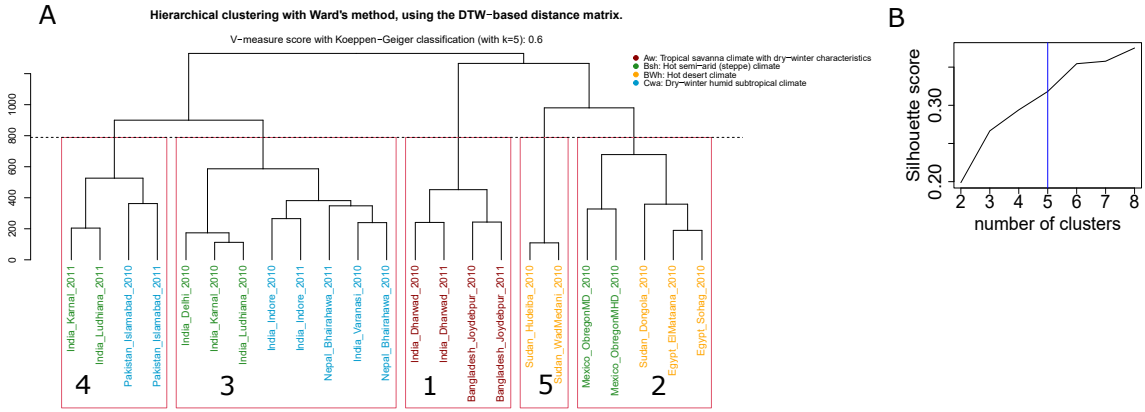


Figure 4.4: Dendogram of environments based on DTW distance calculated using weather time series. **A** Ward's hierarchical clustering given DTW pairwise distances. Environments are colored according to their class in the Koeppen-Geiger climate classification. **B** The silhouette score (Rousseeuw, 1987) was used to determine the optimal number of groups.

The 71 environments of the G2F dataset and the 23 environments of the wheat dataset were clustered based on Ward's hierarchical clustering technique (Ward Jr, 1963) using the dynamic time warping (DTW) distance (Figure 4.2, Figure 4.4) calculated using daily weather data. The number of groups was determined according to the Silhouette score. For the two datasets, the V-measure score computed to estimate the goodness of the DTW output, by taking as reference the Koeppen-Geiger climate classification system, was relatively moderate (about 0.4 in the G2F dataset and 0.6 in the wheat dataset). For the wheat dataset, some environments associated with different Koeppen-Geiger classes clustered together, such as India_Karnal_2011 and Pakistan_Islamabad_2010, which can be explained by the fact that water supply (precipitation + irrigation) had not been included for this dataset. In general, for the two datasets, the largest differences among clusters were attributed to geographic patterns, but at smaller height cutoff values, year-to-year variability could explain the year-wise clustering of environments, especially with the G2F dataset. Within the clusters 2, 3, 4, 5, 6 and 7, at least two Koeppen-Geiger classes were represented, and up to three classes in clusters 3 and 7. It should however be noted that some locations were almost at the border between two classes, such as locations in Illinois or Indiana, which can explain why these Midwest locations present more climatic similarity with the northern locations, in some years, than with southern locations. Interestingly, some sites relatively far from each other were grouped together like SDH1_2015 and COH1_2017 in the maize dataset.

In the maize dataset, the two largest clusters comprised both 16 environments. Cluster 3 included environments from northern and midwestern locations in North America, while environments from cluster 1 were mostly located in southern US regions (Figure 4.2). As illustrated in Figure 4.3, cluster 3 was characterized by lower light, lower maximum and minimum temperatures, especially in late stages of the growing season, and lower vapour pressure deficit (VPD) values than cluster 1. VPD is an indicator of how plants respond to humidity in their growing environment. Very high VPD can yield water stress, and results in a reduction of evapotranspiration via stomata, impeding plant growth. Characteristic time series for other clusters of the maize dataset are provided in S4.6, and for wheat in S4.7.

For the wheat dataset, we also compared these results with a heatmap of pairwise environment distances calculated from marker effects estimated with ridge regression (Figure S4.3), e.g. by considering only phenotypic and marker data without weather information. Two environments appear as striking outliers, namely India_Indore_2010 and Pakistan_Islamabad_2011, which revealed to be two high-yielding environments (Figure S4.1). These environments could not be recognized as outliers on the basis of the DTW clustering, presumably because some critical environmental or management information was lacking, for the reasons previously advanced. Overall, some similarities between the marker-effects heatmap and the DTW-based clustering could be identified. For instance, Sudan_Hudeiba_2010 and Sudan_WadMedani_2010 presented a very small Euclidean distance (< 0.02), and both belonged to the same group based on the hierarchical clus-

tering with DTW distances (Figure 4.4). The same observation could be done with environments Bangladesh_Joydebpur_2010, Bangladesh_Joydebpur_2011 and India_Dharwad_2010.

4.4.2 Partitioning of variance components in the wheat dataset

Figure 4.5**A** show the proportion of variance components obtained from full data analysis. Environment was the random effect explaining about 80% of grain yield variance in the basic models M1 and M2, while the amount of variance attributed to main additive genetic effects (A) was about 3% (S4.7). In M4, approximately 19% and 31% of the across-environment variance was explained by location (L) and year (Y) factors, respectively, while year-location (E) factor remained the most important environmental component (between 37 and 42%). In M5, the A × E, A × Y and A × L interactions captured only a very small amount of variance (∼3 %), but M5 helped to reduce residual error by 38 %.

In models including main effects of weather data, when comparing with M1, the estimated variance attributed to environments (E) was reduced by approximately 72% in M7 using $W_{ECs\_windows}$, and by 46% in M7 with W_{ECs_stages}, which both incorporated weather data as covariates. Using K_{DTW} in M8, E variance component was reduced by ∼78%. This highlights that $W_{ECs\_windows}$, W_{ECs_stages} and especially K_{DTW} were efficient to model differences among environments due to weather conditions experienced by the crop. Interestingly, the inclusion of interaction terms between ECs and SNPs in model M11 did not have the same impact according to the methodology used to derive ECs. In M11 with $W_{ECs\_windows}$, we could notice a strong decrease of the proportion of variance explained by the main effect of ECs (∼62%), while in the case of W_{ECs_stages}, the decrease was more modest (∼22%).

M13 was the model for which the modeling of weather data and of their interaction with genetic additive effects captured the largest proportion of grain yield variance across environments (∼69% with K_{DTW} main effect and A × K_{DTW}). In comparison, main effects of ECs and their interactions with SNPs captured between ∼28% and ∼34% of grain yield variance in models M11 and M12 for $W_{ECs\_windows}$ and W_{ECs_stages}, respectively. Additionally, M13 was the model revealing the smallest fraction of residual error among all models. We still included E effects in models M7, M11, M12, M8 and M13, because the weather data used here could not fully capture all environmental differences between field experiments, due for instance to precipitation patterns, irrigation conditions or soil factors.

We applied M6 within the clusters defined using DTW and presented in 4.4 to verify whether these groups of environment based on weather conditions yielded a reduction of the total G × E variance (Figure 4.7 **B**). Across all clusters, the fraction of phenotypic variation captured by A and A × E within cluster exceeds the same fraction estimated using the full wheat dataset. In 4 out of 5 clusters, the ratio of A on A × E is also increased, by up to 71% in cluster 3.

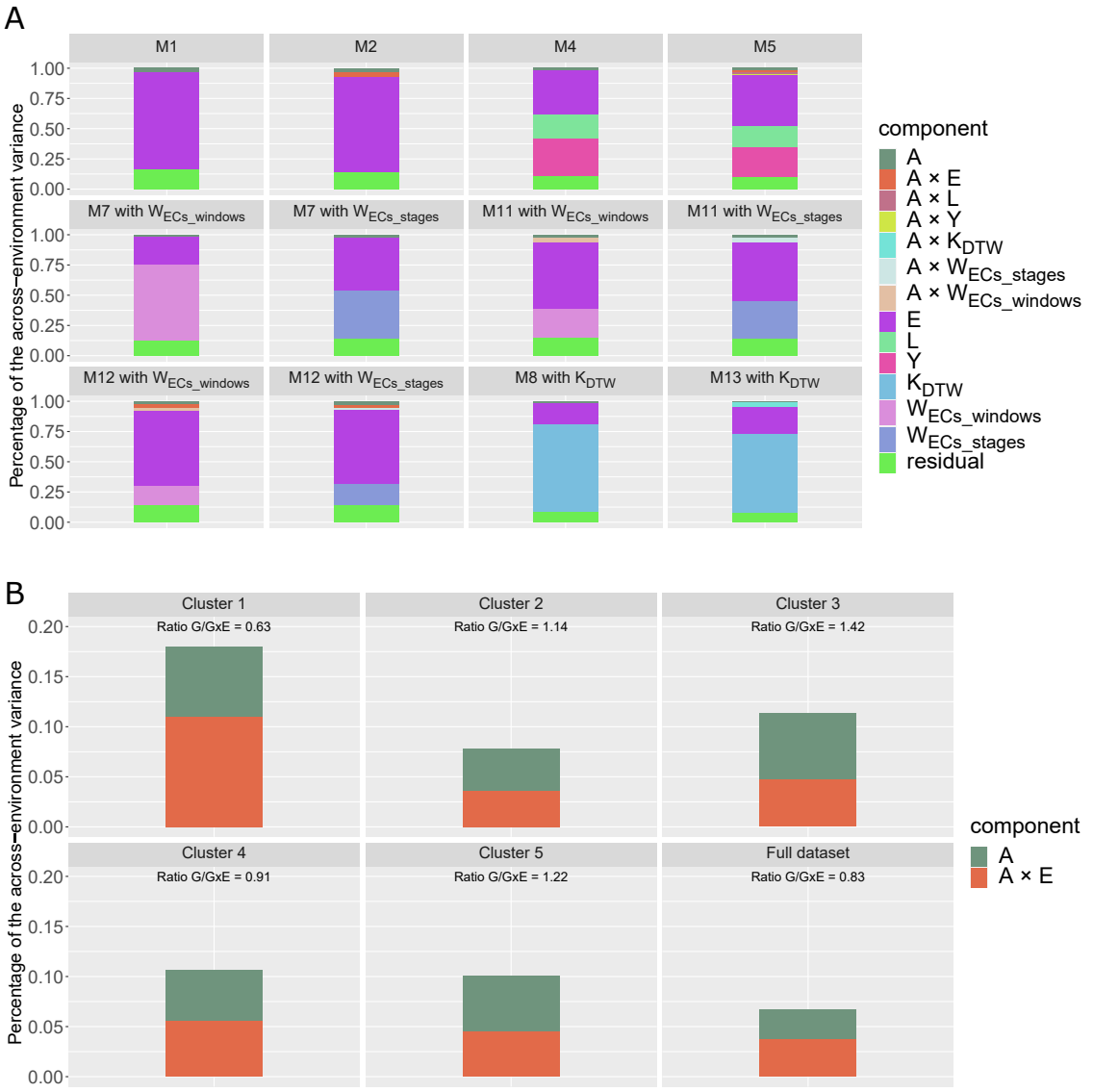


Figure 4.5: **A** Proportion of total variance explained by different genetic and environmental components, by their respective interactions and by residual error, in the different models implemented with the full wheat dataset.
B Proportion of total variance explained by G and G × E within each cluster defined based on DTW distance and with the full dataset in the model M2.
E, environment (field trial) main effect; Y, year effect; L, location effect; A, additive genetic effect; A × E,additive × environment interaction effect; A × Y,additive × year interaction effect; A × L,additive × location interaction effect; K_{DTW}, DTW-based environmental similarity matrix; W_{ECs_stages}, ECs based on estimated phenological stages; $W_{ECs\_windows}$, ECs based on the 10 windows within each environment; AW, additive × ECs interaction effect; A × K_{DTW}, additive × DTW-based environmental kinship.

4.4.3 Partitioning of variance components in the maize dataset

In the maize G2F dataset, the estimated variance due to additive genetic effects (A) ranged between 12 and 20% in the models M1, M2, M4 and M5, in which dominance effects were not incorporated (Figure 4.6 **A**, Table S4.8). When dominance effects were added (in M3, M6, M9, M14, M10 and M15), they explained a larger part of the phenotypic variation than their additive counterpart, and we could also observe a decrease of the proportion of the additive genomic variance component in these models, ranging between 2% and 6% of the total across-environment genomic variance (Figure 4.6 **A**). Adding dominant effects reduced the genetic source of variation by 32%, from M1 to M3.

Using the model M2, the environmental and A × E variance component terms accounted for about 53% and 10%, respectively, of the across-environment phenotypic variation. Year and location effects contributed to a substantial proportion of the across-environment variance in the main effect model M4 (∼39%), thus reducing by the E component by 64% compared with M1. Main effects models that included weather data with ECs (i.e. M7 and M9) captured between 42% and 51% of the total proportion of phenotypic variance due to environmental effects (ratio $W/(\text{E} + W)$). Interestingly, the ratio $K_{DTW}/(\text{E} + K_{DTW})$ was even superior (∼79% and ∼81% in M8 and M10 respectively, thus indicating that K_{DTW} efficiently recovered phenotypic variation due to weather patterns. The full model M15, followed by M13, based on K_{DTW}, were both more efficient than other models to reduce the residual variance, similarly to the wheat dataset. In addition, reaction-norm model M15 was also better at capturing additive × environment and dominance × environment variance (∼17%) than the two other reaction norm-models M14 with W_{ECs_stages} (∼10%) and $W_{ECs\_windows}$ (∼11%). Overall, these models that incorporated interactions between climatic data and both dominance and additive genetic terms (i.e. M14 and M15) captured a superior amount of phenotypic variation than the model using A × E and D × E, i.e. only the environment label (∼6%).

We applied M6 within the clusters defined using DTW and presented in 4.2. The proportion of yield phenotypic variance explained by genetic components (A, D, A × E and D × E) was larger within clusters than within the full dataset, with the exception of cluster 1 (Figure 4.6 **B**). The ratio of G (i.e. A + D) on G × E (i.e. A × E and D × E) increased substantially for clusters 5 (264%), 6 (216%) and 8 (207%), but was reduced in the remaining clusters, thereby suggesting that the G × E variance component remained important within clusters, and/or that the mere genetic variance must explain on its own a large variability of grain yield variation with the full dataset.

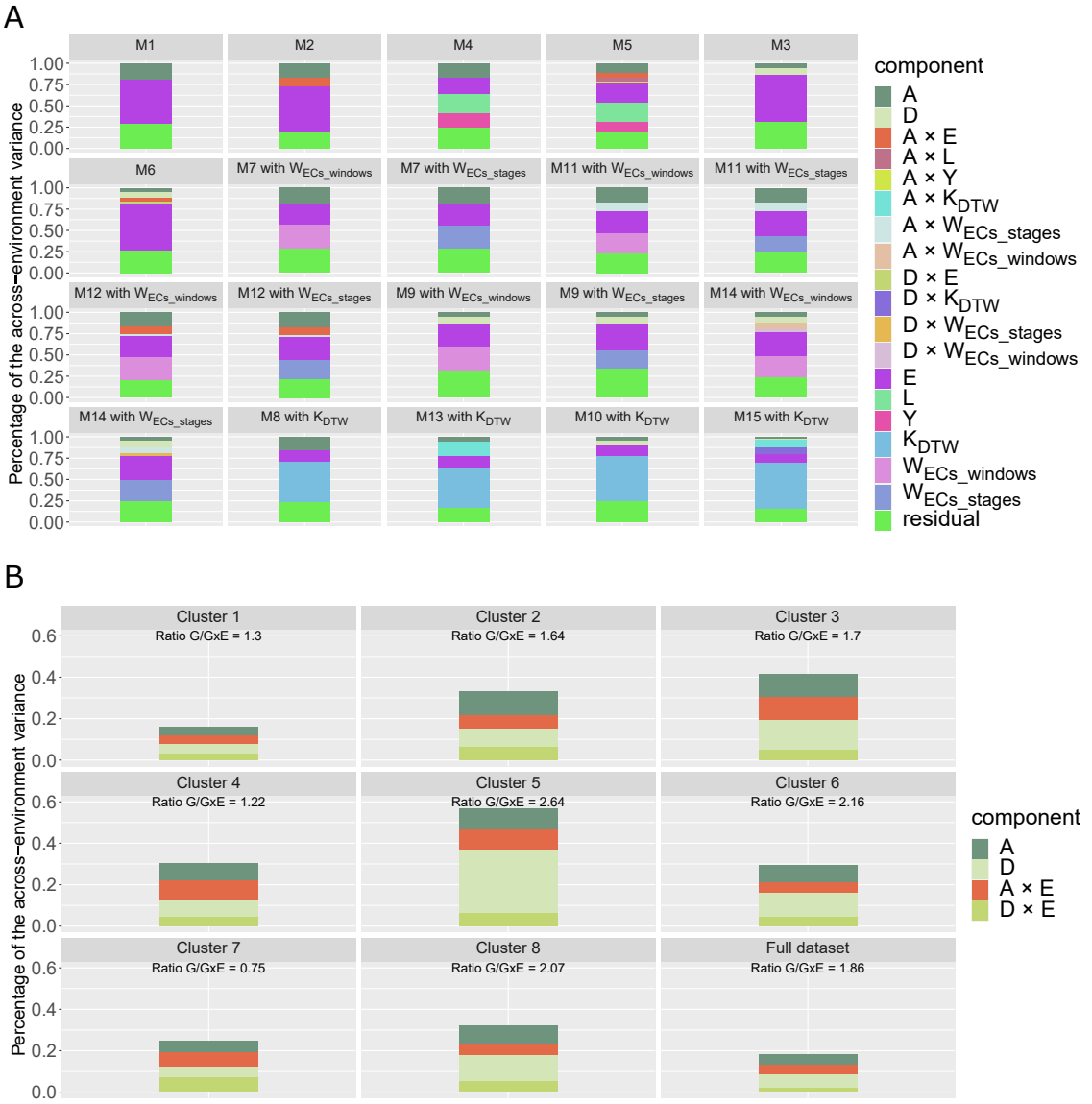


Figure 4.6: **A** Proportion of total variance explained by different genetic and environmental components, by their respective interactions and by residual error, in the different models implemented with the maize dataset.
B Proportion of total variance explained by G and G × E within each cluster defined based on DTW distance and with the full dataset in the model M6.
E, environment (field trial) main effect; Y, year effect; L, location effect; A, additive genetic effect; D, dominance genomic effect; A × E,additive × environment interaction effect; A × Y,additive × year interaction effect; A × L,additive × location interaction effect; DE, dominance × environment interaction effect; K_{DTW}, DTW-based environmental similarity matrix; W_{ECs_stages}, ECs based on estimated phenological stages; $W_{ECs\_windows}$, ECs based on the 10 windows within each environment; AW, additive × ECs interaction effect; DW, dominance × ECs interaction effect; A × K_{DTW}, additive × DTW-based environmental kinship; D × K_{DTW}, dominance × DTW-based environmental kinship.

Table 4.2: Average within-environment correlations between predicted and observed values for grain yield using 9 linear random effects models, three different methods to incorporate weather data, and three different cross-validation schemes (CV0, CV1 and CV2) for the wheat dataset in 23 environments. Letters show groups that are significantly different according to a paired t-test ($\alpha = 5\%$).

	Without weather data				**With weather data**							
					Models based on W						Models based on DTW	
					With $W_{ECs\_windows}$			With W_{ECs_stages}				
CV type	M1[a]	M2	M4	M5	M7	M11	M12	M7	M11	M12	M8	M13
CV0	0.306 (a)	0.305 (a)	0.307 (a)	0.320 (a)	0.306 (a)	0.324 (ab)	0.336 (ab)	0.306 (a)	0.348 (ab)	**0.352** (b)	0.306 (a)	0.351 (b)
CV1	0.218 (a)	0.220 (bc)	0.288 (fg)	**0.289** (g)	0.220 (c)	0.272 (d)	0.288 (fg)	0.219 (b)	0.274 (d)	0.287 (f)	0.219 (b)	0.283 (e)
CV2	0.302 (a)	0.302 (a)	0.376 (c)	0.382 (d)	0.302 (a)	0.369 (b)	0.386 (e)	0.302 (a)	0.384 (d)	**0.392** (f)	0.302 (a)	0.388 (e)

E, environment (field trial) main effect; Y, year effect; L, location effect; A, additive genetic effect; A × E,additive × environment interaction effect; A × Y,additive × year interaction effect; A × L,additive × location interaction effect; K_{DTW}, DTW-based environmental similarity matrix; W_{ECs_stages}, ECs based on estimated phenological stages; $W_{ECs\_windows}$,

[a]ECs based on the 10 windows within each environment; A × W, additive × ECs interaction effect; A × K_{DTW}, additive × DTW-based environmental kinship.

Models: M1, A + E; M2, A + L + Y + E; M4, A + E + A × E; M5, A + Y + L + E + A × Y + A × L + A × E; M7, A + E + W; M11, A + E + W + A × W; M12, A + E + W + A × E + A × W; M8, A + E + K_{DTW}; M13, A + E + K_{DTW} + A ×K_{DTW}

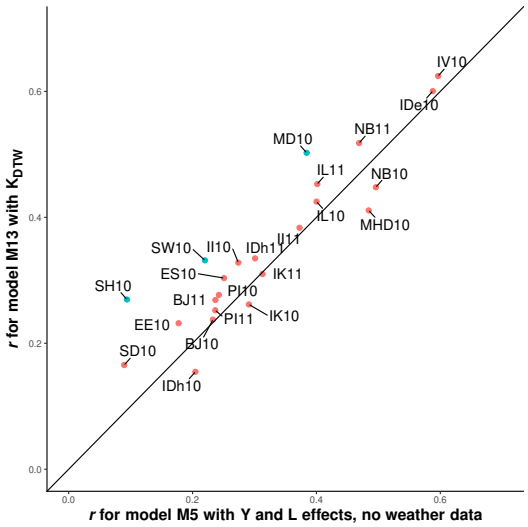


Figure 4.7: Comparison of the predictive ability in each environment of the wheat dataset in the CV0 scheme (leave-one-environment-out) between model M13 (full model with interactions between K_{DTW} and additive genetic effects) and model M5 (full model with interactions between year and location effects with additive genetic components, without weather data). The line indicates the identity, and blue points show an environment for which the absolute difference of predictive ability between the two models was superior to 0.10. Environment labels are precised in Table S4.1.

4.4.4 Predictive ability in the wheat dataset

Table 4.2 summarizes the results obtained with the three cross-validation schemes (CV0, CV1 and CV2) for each model tested with the wheat dataset. Across all CV schemes we evaluated, the incorporation of G × E terms (modeled with either year, location and/or environment label or with weather information) in addition to the the main effect terms, yielded a substantial improvement of predictive ability, by 15% in CV0 (from M7 to M12 with W_{ECs_stages}), 32% in CV1 (from M1 to M4) and by 29% in CV2 (from M7 to M12 with W_{ECs_stages}).

An advantage of using weather data was observed with CV0 and CV2, where the best model M12 with W_{ECs_stages} outperformed the most complex model without weather data (i.e. M5), by ∼10% in CV0, and by ∼3% in CV2. Models M12 with W_{ECs_stages} and M13 with K_{DTW} provided very similar results with the CV0 scheme. Figure 4.7 shows the comparison between two interaction models, M13 (with weather data modeled with K_{DTW}) and M5 (environment modeled with year and location labels), where the diagonal line corresponds to the case where the environment is predicted exactly the same with the two models. It highlights that 18 over 23 environments were better predicted with the interaction model M13, that uses K_{DTW} (i.e. weather data), than with the interaction model 5, that incorporates year and location effects, in CV0. The gain was particularly important for two environments located in Sudan, 2010 and in Mexico.

The prediction of new genotypes (CV1) resulted in the lowest average predictive abilities compared with CV2 and CV0 (Table 4.2). Figure 4.8 shows the range of predictive abilities obtained across 50 test partitions by environment in the CV1 (panel A) and in the CV2 (panel B) prediction

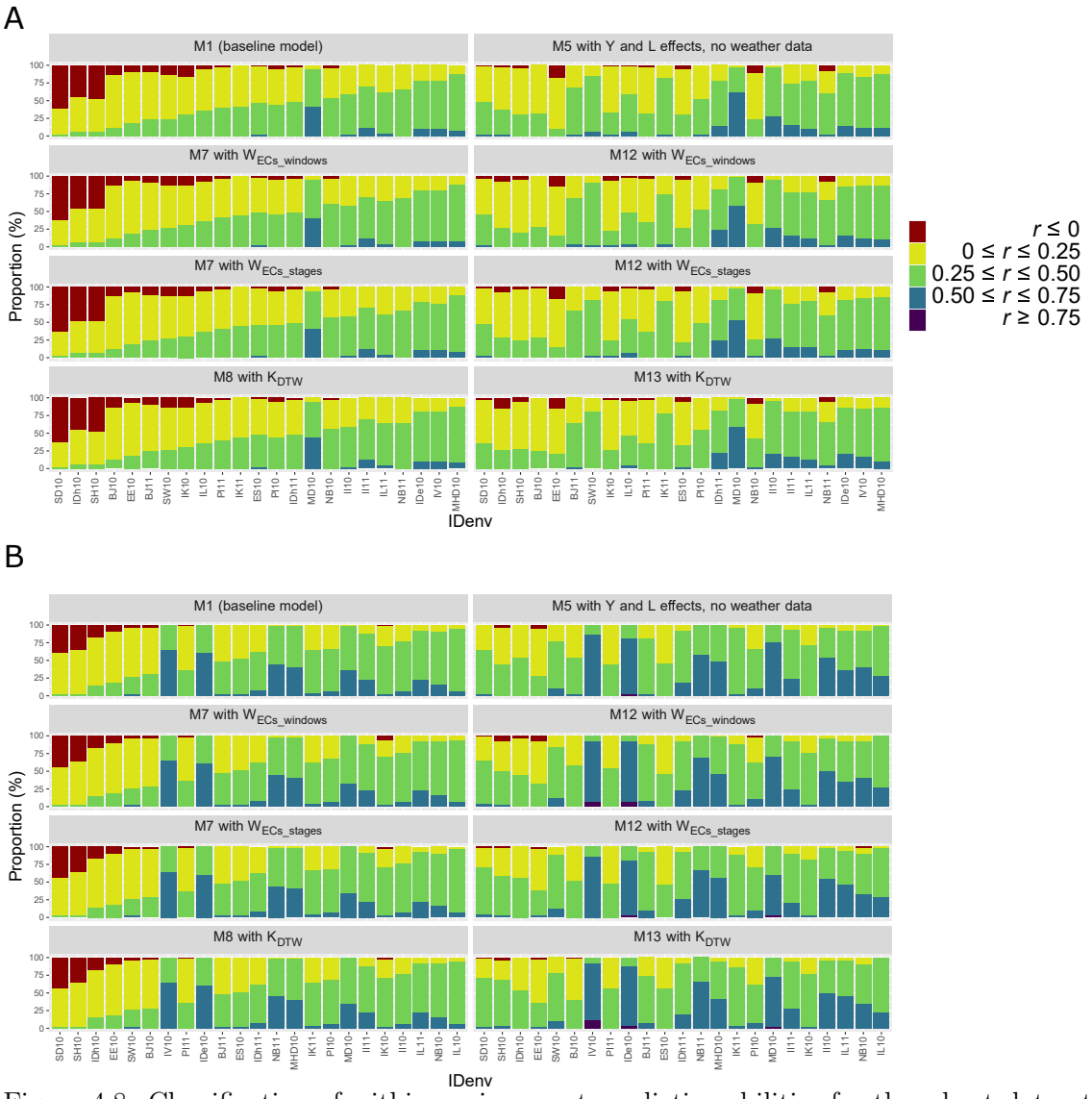


Figure 4.8: Classification of within-environment predictive abilities for the wheat dataset for a set of implemented predictive models. **A** CV1 prediction scheme (prediction of new genotypes), evaluated across 50 different testing sets. **B** CV2 prediction scheme (incomplete field trials), evaluated across 50 different testing sets.
Each facet corresponds to a predictive model. Each column corresponds to an environment, sorted according to the green category in model M1 for each panel A and B. Models in the left column correspond to main effects models, while models in the right column are interaction effects models. Environment labels are precised in Table S4.1.

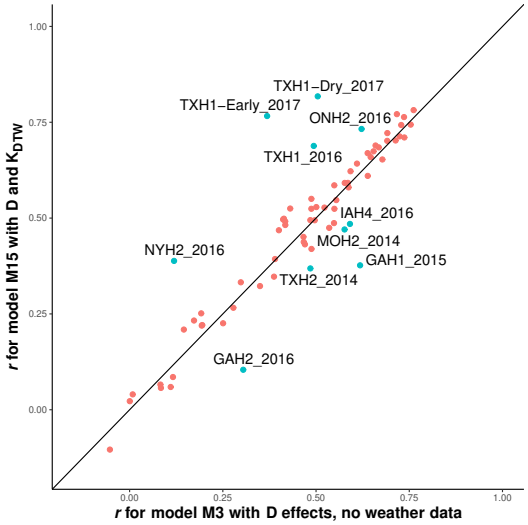


Figure 4.9: Comparison of the predictive ability in each environment of the maize dataset in the CV0 scheme (leave-one-environment-out) between model M15 (full model with interactions between K_{DTW} and both additive and dominance genetic effects) and model M3 (model with main effects of additive, dominance genetic effects and environment, without weather data). The line indicates the identity, and blue points show an environment for which the absolute difference of predictive ability between the two models was superior to 0.10. Environment labels are precised in Table S4.2.

schemes. Predictions for some environments, such as India_Varanasi_2010, India_Delhi_2010 or Nepal_Bhairahawa_2011 could be considerably improved in CV2 compared with CV1. It highlights that some lines could not be well predicted if completely absent from the training set, suggesting that the amount of genetic relatedness was not sufficient between the training and test sets. It can be related to the fact that the set of lines included in the wheat WAMI panel is genetically diverse, as reported in a previous study (Lopes et al., 2015). For environments SD10, SH10 and IDh10 in particular, adding G × E terms in CV1 and CV2 had a strong impact on the range of the predictive abilities (Figure 4.8), leading to an almost disappearance of negative predictive abilities using M12 with W_{ECs_stages} and M13 with K_{DTW} in CV2. Regarding the modeling of ECs, a slight benefit of using W_{ECs_stages} over $W_{ECs\_windows}$ was observable for CV1 and CV2 but remained less than 5%.

4.4.5 Predictive ability in the maize dataset

Table 4.3: Average within-environment correlations between predicted and observed values for grain yield using 6 linear random effects models including both additive and dominance genetic effects, three different methods to incorporate weather data, and three different cross-validation schemes (CV0, CV1 and CV2) for the maize dataset in 71 environments. Letters show groups that are significantly different according to a paired t-test ($\alpha = 5\%$).

	Without weather data						**With weather data**													
							Models based on W										Models based on DTW			
							With $W_{ECs\_windows}$					With W_{ECs_stages}								
					With D					**With D**					**With D**					**With D**	
CV type	M1[a]	M2	M4	M5	M3	M6	M7	M11	M12	M9	M14	M7	M11	M12	M9	M14	M8	M13	M10	M15
CV0	0.419 (b)	0.418 (b)	0.402 (ab)	0.446 (cd)	0.462 (de)	0.455 (cde)	0.418 (b)	0.378 (a)	0.418 (b)	0.461 (de)	0.429 (bc)	0.418 (b)	0.392 (a)	0.412 (b)	0.461 (de)	0.441 (c)	0.418 (b)	0.432 (bc)	0.461 (de)	**0.478** (e)
CV1	0.344 (a)	0.344 (b)	0.469 (g)	0.470 (g)	0.409 (c)	**0.501** (k)	0.344 (b)	0.435 (g)	0.471 (h)	0.410 (d)	0.477 (j)	0.344 (b)	0.428 (e)	0.469 (g)	0.410 (d)	0.472 (i)	0.344 (a)	0.469 (g)	0.409 (c)	**0.501** (k)
CV2	0.421 (c)	0.421 (a)	0.519 (h)	0.523 (m)	0.464 (e)	0.548 (o)	0.421 (b)	0.494 (g)	0.521 (k)	0.464 (de)	0.525 (n)	0.421 (b)	0.486 (f)	0.519 (i)	0.464 (d)	0.519 (j)	0.421 (b)	0.523 (l)	0.464 (d)	**0.552** (p)

E, environment (field trial) main effect; Y, year effect; L, location effect; A, additive genetic effect; D, dominance genomic effect; A × E, additive × environment interaction effect; A × Y, additive × year interaction effect; A × L, additive × location interaction effect; D × E, dominance × environment interaction effect; K_{DTW}, DTW-based environmental similarity matrix; W_{ECs_stages}, ECs based on estimated phenological stages; $W_{ECs\_windows}$, ECs based on the 10 windows within each environment; A × W, additive × ECs interaction effect; [a]D × W, dominance × ECs interaction effect; A × K_{DTW}, additive × DTW-based environmental kinship; D × K_{DTW}, dominance × DTW-based environmental kinship.

Models: M1, A + E; M2, A + L + Y + E; M3, A + D + E; M4, A + E + A × E; M5, A + Y + L + E + A × Y + A × L + A × E; M6, A + D + E + A × E + D × E; M7, A + E + W; M8, A + E + K_{DTW}; M9, A + D + E + W; M10, A + D + E + K_{DTW}; M11, A + E + W + A × W; M12, A + E + W + A × E + A × W; M13, A + E + K_{DTW} + A ×K$_{DTW}$; M14, A + D + E + W + A ×W + $D \times W$; $M15, A + D + E$+K$_{DTW}$ + A ×K$_{DTW}$ + D ×K$_{DTW}$

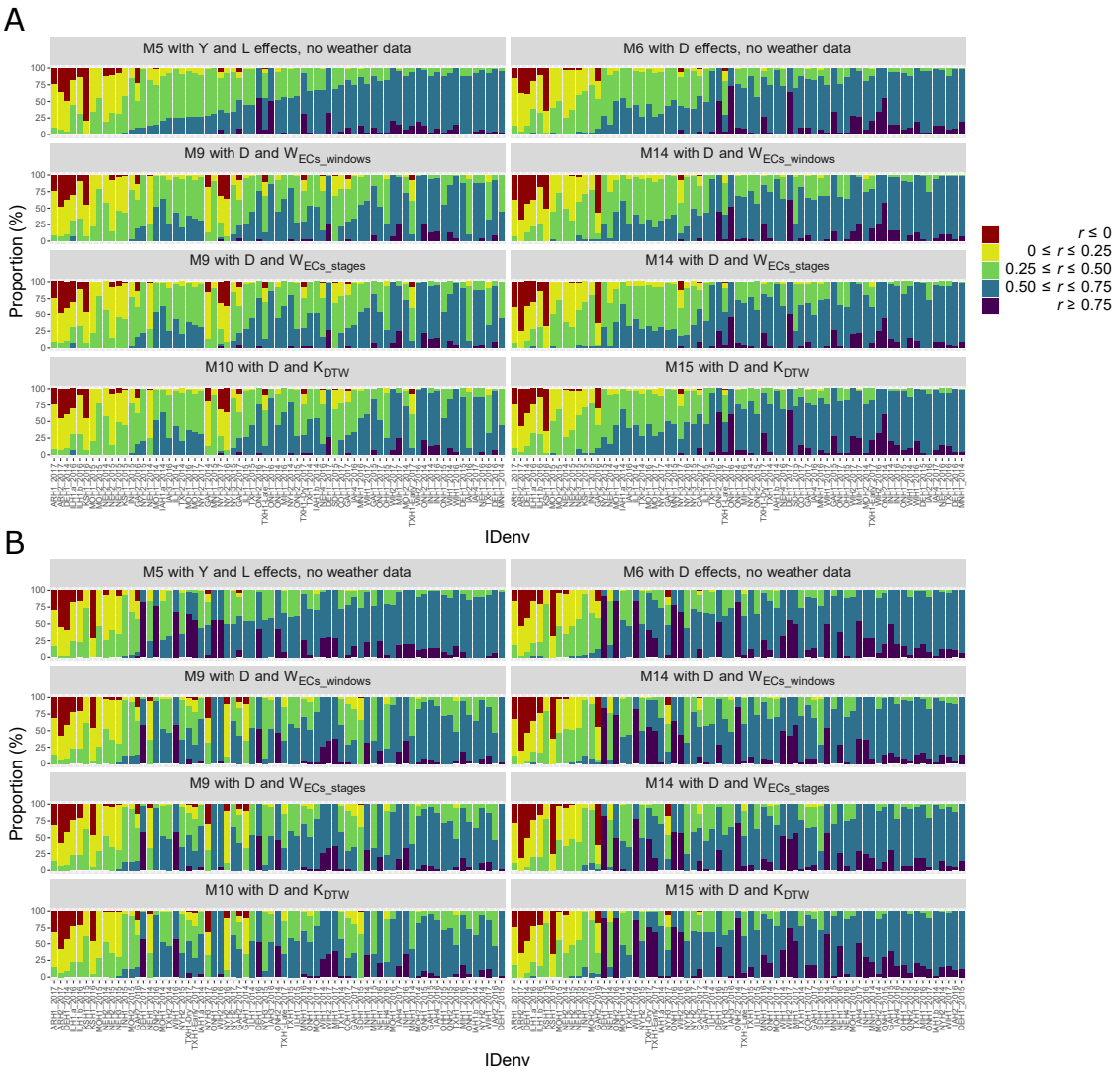


Figure 4.10: Classification of within-environment predictive abilities for the maize dataset for a set of implemented predictive models. **A** CV1 prediction scheme (prediction of new genotypes), evaluated across 50 different testing sets.**B** CV2 prediction scheme (incomplete field trials), evaluated across 50 different testing sets.
Each column corresponds to an environment, sorted according to the blue category in model M1 for each panel A and B. Each facet corresponds to a predictive model. Models in the left column correspond to main effects models, while models in the right column are interaction effects models. Environment labels are precised in Table S4.2.

Table 4.3 summarizes the results obtained with the three cross-validation schemes (CV0, CV1 and CV2) for each model tested with the maize dataset. The prediction of new environments (CV0) resulted in the lowest average predictive abilities compared with CV2 and CV1 (Table 4.3, Figure 4.10). This difference from results obtained with the wheat dataset can be explained by the fact that within CV1, the training set still comprised a large number of maize hybrids due to the size of the experiments and it often shared at least one parent with hybrids assigned to the test set. Thus, the training set in CV1 generally still contained hybrids with relatively high genetic similarity to those included in the test set. In CV0, the use of model M5, that considers the year and location main effects together with their respective interactions with the additive genetic effects, yielded an important improvement over the M4 model, that considers interactions between environment label and additive genetic effects (∼11%). This corroborated the assumption that year and location effects are beneficial to model when dealing with large multi-environment datasets characterized by a number of common locations tested across several years. Thereby, model M5 allows to recover information from the same site observed in different years and from the same year considered at different sites within the leave-one-environment-out CV prediction scheme. On the other hand, model M4 does not allow to borrow information based on the level of climatic similarity due to geographical patterns or year trends among environments.

Across all CV schemes (CV0, CV1, and CV2) we implemented, models incorporating main additive and dominance genetic effects, i.e. M3, M9 and M10, yielded a better predictive ability than their counterpart modeling only main additive effects (M1, M7 and M8). Considering CV1 and CV2, the modeling of additive-by-environment and of dominance-by-environment interactions effects (either with A × E and D × E, or with A × W and D × W, or with A × K_{DTW} and D × K_{DTW}), as proposed in models M6, M14 and M15, yielded an improved average performance compared with the main effects models.

The inclusion of weather data led to a maximum improvement of 3.5% in CV0, considering the model M15 versus the best model without weather data, i.e. M3. Figure 4.9 shows a comparison of predictive abilities in the CV0 scheme between the two interaction models M3 and M15. 41 out of 71 environments were better predicted with the full model incorporating weather data with K_{DTW} than with M3. Environments corresponding to locations very well represented in the total dataset, i.e. from northern or midwestern regions, did not display major improvements, with the exception of NYH2_2016 and ONH2_2016. A location which was present only once in our dataset, TXH1_2014, and differed significantly from all other environments in terms of weather conditions, was better predicted without weather data, suggesting that the degree of climatic similarity between training and testing sets also plays a major role with models integrating weather data.

For CV1 and CV2, adding weather data yielded non or negligible effects on within-environment average predictive abilities, and M6 was competitive with M15 (Figure 4.10). An interesting result was the fact that the K_{DTW} method in M15 was more efficient to capture additive-by-environment

and dominance-by-environment interactions than models based on $W_{ECs\_windows}$ or W_{ECs_stages}, and this observation was consistent across all CV prediction schemes.

4.5 Discussion

Summarizing efficiently large amounts of weather records for integration in prediction models remains a complex task, as it implies (1) to determine a set of informative ECs to calculate based on daily climatic data and (2) to define a temporal resolution over which the ECs are computed. In this study, we presented a data-driven method, utilizing a dissimilarity measure between weather time series, to cluster crop growing season events and to derive an environmental relationship matrix between environments, which can be used for instance in reaction-norm models. We compared a large series of models to evaluate how the modeling of weather conditions impacted the proportion of variance explained by the environmental effects in prediction models, as well as the predictive ability in different CV schemes. In addition, we assessed the benefit of including interaction effects between genomic and environmental components in these models, as well as of considering dominance deviations for the maize dataset.

4.5.1 Modeling dominance effects and its impact on the predictive ability in the hybrid maize dataset

The correlation between elements of the dominance D and additive A genomic relationship matrices was equal to 0.89 in our study, which was close to the value of 0.83 reported by Rogers et al. (2021), and this dependency indicates that partitioning dominance from additive genetic effects was not easily achieved. In contrast to Rogers et al. (2021), we found a larger contribution of the dominance effects compared to additive effects using the full dataset. Yet, different marker subsets were used across the two studies, complicating direct comparisons of variance component estimates. Jarquin et al. (2021a) also found a larger portion of phenotypic variation explained by the specific combining ability (SCA), which is associated with nonadditive genetic effects, than by the individual parental main additive effects, modeled through the general combining abilities (GCA) components.

In our study, the main D effects and the D $\times$ E interactions, using the environment label, explained about 9% of the phenotypic variation for the trait grain yield, and this proportion was slightly increased when modeling E with ECs in M14 ($\sim$10%). Some other studies found a greater range of variance attributed to D effects. Costa-Neto et al. (2021) showed that D main effects and their interactions with E explained up to 40% of the phenotypic variation for the same trait, but we used a different formula to define D, the dominance relationship matrix, as similarly applied in other studies (Ferrão et al., 2020; Granato et al., 2018; Rogers and Holland, 2021). The parametrization method for D they carried out was suggested by Vitezica et al. (2013), and would potentially result in a different partition of the genetic variance, and on a different interpretation of genetic effects

of the markers. Another interesting fact from our results is that, as D effects were included in the model, the proportion of phenotypic variation due to additive genetic effects strongly decreased (∼69% from M1 to M3), as it could also be observed in the study of Muñoz et al. (2014), that reports a decrease of narrow-sense heritability up to 40% and of the broad-sense heritability of 9%, when dominance effects are included in marker-based models. About 47% of the decrease in additive variance was captured by the dominance variance, illustrating that when dominance is not included, estimated additive variance can account for it.

Consistent with previous studies conducted with the same G2F dataset, analyzed from 2014 to 2015 (Jarquin et al., 2021a), and from 2014 to 2016 (Rogers and Holland, 2021; Rogers et al., 2021), we found that modeling dominance effects in the G2F dataset yielded an important increase in predictive ability across all CV schemes we studied. The gain in predictive ability was the largest in the CV1 prediction problem (∼ 19%), which can be interpreted by the fact that accurate prediction of newly released F_1 hybrids might necessitate the inclusion of dominance effects. Since dominance variance (σ^2_D) depends on the product $p^I(1-p^I)p^{II}(1-p^{II})$ (Griffing, 1962), where p_I and p_{II} designate the allele frequencies in two heterotic groups, reduced dominance variance is expected when the cross occurs between two strongly divergent populations, such as Dent × Flint (Westhues et al., 2017). Thus, the presence of substantial dominance effects that we observed in the present dataset is fully concordant with the observation made by Jarquin et al. (2021a) that the F_1 crosses examined in this study were realized within and between heterotic groups, making the modeling of dominance effects all the more beneficial.

4.5.2 Clustering environments in MET datasets with DTW distance

We decided to classify the two MET datasets based on a multivariate method using climate time series. The groups of environments that we identified were coherent with classical climate classification systems, such as the Koppen-Geiger classification. The DTW-based clustering can be useful to elucidate why some environments appear as outliers based on phenotypic data, e.g. particular low or high yielding environments, in case that no meta-data from breeders' observations is available. Nonetheless, this method is not flawless, since extreme weather events, such as heavy storms which can impact final yield due to stem lodging, delayed harvest, or emergence of disease related to wet conditions (Rötter et al., 2015), could probably not be detected as wind speed was not included in the weather time series. Therefore, marker-effects clustering (Figure S4.3) should also be systematically combined to help interpretation, provided that a variable amount of genetic relatedness among environments does not affect the marker-effects clustering (Heslot et al., 2013). Recognition and subsequent removal of atypical year-location combinations can be of interest, because these environments might not appropriately represent the target population of environments (Heslot et al., 2013).

Furthermore, it is crucial to ensure that relationships among environments based on climatic

similarity estimated with DTW can be related to true G × E patterns exploitable with breeding, hence to examine the relationship with the target trait grain yield. First, regarding the complete G2F dataset, we showed that a substantial larger fraction of grain yield variance could be captured by incorporating A × K_{DTW} and D × K_{DTW} in M15, than by using year-location ID as in model M6. Additionally, full models incorporating interactions between K_{DTW} and genomic components were associated with the least residual error across all models in the two datasets. Thus, a superior proportion of yield variance could be explained by quantitative environmental information in the full model that used K_{DTW} compared to all other models.

Secondly, given the fact that (1) the exact same set of lines was assessed in the different environments and (2), that the ratio of G to G× E interaction variances was substantially enhanced in 4 out of 5 clusters within the wheat dataset, the interest of our method to reduce G × E interaction, thereby increasing heritability, is demonstrated for the wheat dataset. This partitioning of environments in homogeneous subgroups can be useful to identify cultivars specifically suitable for each of these subgroups on the basis of climatic conditions impacting grain yield. With regard to the maize dataset, the same ratio was increased after clustering in only 3 out of 9 clusters. However, several characteristics related to the composition of the G2F dataset can be advanced to explain this observation. Jarquin et al. (2021a) conducted a study with the data of the G2F Initiative from 2014 to 2015 and reported the confounding effect of environmental conditions with genetic background of the tested hybrids, due to an unbalanced allocation of hybrids to growing environments, partially based on maturity groups. In addition, Rogers et al. (2021) highlighted the important level of diverse genetic backgrounds identifiable within this material, as illustrated by the high within-cluster heterogeneity observed after using 10 clusters determined by principal component analysis on the GBS data on hybrids and parental lines. Hence, it could be hypothesized that, considering the full dataset, the proportion of phenotypic variation due to pure genetic components is larger than when performing within-cluster analyses, where the genetic material is already more adapted to the environment type. If the amount of genetic relatedness among environments is generally heightened within clusters, thus it can possibly explain why we observe that the G × E interaction contributes more significantly than mere genetics to yield variation in the case of large clusters, where substantial environmental differences still occur, such as within clusters 1 or 3. To solve this problem, a solution could be to increase the number of groups to further partition some of the clusters. In general, though, the exact number of clusters cannot be exactly determined, and clustering based on DTW cannot guarantee that all clusters have the same level of climate homogeneity.

A supplemental reason for a lack of improvement of the G to G× E interaction variance ratio can also be that other environmental features, that we did not take into account here, might have influenced grain yield in some clusters, for instance management aspects (type and amount of fertilizers, precedent crop, that potentially affect nitrogen availability), soil factors or even disease

pressure (Heslot et al., 2014; Ly et al., 2017; Touzy et al., 2019). In the wheat dataset, the lack of information regarding irrigation prevented to integrate explicitly water stress as a potential abiotic stress factor in the clustering analysis.

4.5.3 Incorporating climatic information with ECs in prediction models

Overall, the impact of the temporal resolution used to summarize weather data into environmental covariates, i.e. by predicting growth stages based on GDD or by simply dividing the total crop growing season into a fixed number of day windows, did not seem to substantially impact predictive abilities for either dataset. As a consequence, we could not find any advantage for prediction purposes by using approximated growth stages to compute ECs. This could be attributed to the fact that we used the average flowering time to define three (two) main groups of maturity and subsequently assigned each environment to one of these groups for the maize (wheat) dataset. Dates corresponding to several developmental stages, predicted on the basis of accumulated growing degree days, were obtained within each environment, without taking into account variable flowering time among wheat lines or maize hybrids. As shown by Figure S4.5, the variability among silking dates within environment could span more than two weeks, and the average root mean square error was equal to 4.7 days. In our previous results (Westhues et al., 2021a), we harnessed flowering time data to derive hybrid-specific ECs for three main developmental stages, thereby ensuring that the variables used in our models perfectly reflected real environmental conditions occurring at this crucial developmental stage for each genotype. Here, we chose to make the assumption of sparse phenotypic data (i.e. only grain yield data per genotype and an average flowering time within the environment), in order to mimic large-scale plant breeding programs, which are unlikely to collect detailed phenotypic observations for each genotype in each environment. For new selection candidates, this method would also imply to estimate their flowering time with marker-based approaches, such as CGM coupled with marker-assisted selection (Rincent et al., 2017), because this information would be required as input variables to predict grain yield performance. Utilizing ECs derived from crop growth models might yield better predictive abilities in some cases (Heslot et al., 2014; Ly et al., 2017; Rincent et al., 2017), provided that the predictive stages can be predicted with accuracy, which was difficulty achievable in our study S4.5 for the reasons explained above.

Across the two datasets, we observed that the inclusion of weather data in the form of ECs was especially useful in the CV0 prediction problem for the wheat dataset, leading to a gain in predictive ability of up to $\sim 10\%$, compared with the best model that uses the respective effects of year and location (M5). It is worth mentioning that the CV0 prediction scheme, which was a leave-one-environment-out CV scheme, yielded a large training set consisting of the remaining environments, which was supposedly enough related to the test set, from an environmental perspective, to provide the model with reliable estimates of interaction effects between marker data and ECs for this

specific year-location combination. Regarding the maize dataset, the range of average values we obtained with the CV0 scheme (0.402 to 0.482) was close to the results reported by Rogers and Holland (2021) (0.47 to 0.48), that utilized more than 370 environmental covariates accounting for climatic and soil information. The more challenging scenarios explored by Rogers and Holland (2021) with the 2014-2016 G2F dataset, for instance by leaving out 1 year and related hybrids from the training set, or by stratification-by-environment clusters, did not lead to any improvement by adding G × E interactions based on environmental data. The same authors also examined which temporal resolution was the most adequate to aggregate weather data, using a slightly different approach than ours with $W_{ECs\_windows}$ by considering a fixed number of days included in each window, which necessarily leads to a heterogeneous number of windows across environments as crop growing season lengths differ. The temporal resolution leading to the best average predictive ability was 5-day window, which was reduced compared to the temporal resolution used in our study with $W_{ECs\_windows}$; the minimum number of days used within a window was equal to 12 in our case. A previous study conducted with the G2F data (Jarquin et al., 2021a) considering all hourly values from the weather stations yielded a reduced gain in predictive ability, compared with the study of Rogers and Holland (2021) and our study, which both initially aggregated the hourly weather records into daily datasets. For CV1 and CV2, we noticed, however, that the mere individual modeling of year and location effects yielded a leap in accuracy that was comparable with the results obtained with the integration of weather data. Nonetheless, this advantage might be observed because the dataset exhibited repeated field trials across multiple years in geographically close locations. On the other hand, one key advantage of using an environmental kinship matrix, either based on ECs or DTW distance, is that predictions can be achieved for an untested location in a new year that has never been tested before.

4.5.4 Advantages of using dynamic time warping (DTW) in prediction models

The use of an environmental kinship matrix based on DTW for prediction purposes appeared especially useful for the maize dataset, for which it outperformed methods based on ECs, and did not demonstrate any important drop in accuracy with the wheat dataset. Figure 4.9 shows that substantial improvements were observed to predict TXH1-Early_2017, TXH1-Dry_2017 and TXH1_2016 environments, while adding weather information worsened predictive ability in GAH2_2016. Similar impacts of the impact of adding G × E for TXH1_2016 and GAH2_2016 were reported by Rogers and Holland (2021).

The method we propose allows to conserve the precision provided by daily weather records and to operate an efficient reduction of weather data dimensionality (Delerce et al., 2016) with little feature engineering. The main advantage of DTW over Euclidean distance is that this measure is able to synchronize two weather time series, even though variations in the weather data do not

occur at the exact same time point (Netzel and Stepinski, 2016). For instance, the method can also be of interest when weather stations start collecting data from the planting date onwards, which differ among environments. In this case, DTW is able to match two time series that would present very similar patterns overall, but present a time shift regarding their set first day, that cannot be detected by the Euclidean distance. Compared with the methods we described based on ECs, applying DTW does not require to determine a temporal resolution for summarizing daily weather data, and has relevance when phenotypic data on crop phenological and physiological development is scarce, hindering the possibility to use precise crop growth models.

In the wheat dataset, although time series characterizing water supply was not included, it is frequently observed that weather variables follow a coincident path throughout the growing period, as for example maximum temperature is often correlated with drought periods (Sadras et al., 2012). Therefore, as noted in other studies (Touzy et al., 2019), water stress patterns, if they are present, might potentially be captured by the maximum temperature time series. Including the vapour pressure deficit (VPD) in the multivariate time series analysis, besides raw climatic time series, seems appropriate, given the fact that different authors recommended to integrate limitation on transpiration via stomatal closure in response to high VPD as a trait of interest in cereal breeding programs for selecting genotypes with improved drought tolerance (Medina et al., 2019; Sinclair et al., 2005; Yang et al., 2012).

4.5.5 Future directions

In the present study, we used a reaction norm model that makes the assumption of equal variance of all reaction norm slopes associated with different environmental covariates, and of no correlation among these slopes. Thereby, no differential impact of the environmental variables is allowed by this model. However, it is likely that some specific stresses might explain a larger proportion of phenotypic variation than others, as highlighted by Ly et al. (2018) considering drought stress at flowering stage in wheat. Rogers and Holland (2021) used LASSO to model flexible G $\times$ E effects and to enable shrinkage of irrelevant covariates effects. Nonetheless, as mentioned above, no strong increase of predictive ability could yet be observed across the different CV schemes that were implemented in their study. The ability of machine learning approaches, such as tree-based methods or artificial neural networks, to capture nonlinear responses of genes to environmental stresses, represents therefore a promising strategy to improve predictive performance (Delerce et al., 2016; Washburn et al., 2021; Westhues et al., 2021a,b).

Although it can be envisaged to employ the the DTW distance instead of the Euclidean distance in some classical kernel functions, predictive frameworks such as support vector machines necessitate positive definite (PD) kernels (**??**). Since the DTW distance is not in general symmetric, unless a local constraint is set that restricts the directions when advancing in the local cost matrix as the cost is being calculated (Sardá-Espinosa, 2017), the use of DTW distances with kernel methods is

not straightforward. Support vector machines often exploit as PD kernels simple similarity measures, such as inner products, that ensure the existence of a feature space. Deriving admissible PD kernels from DTW dissimilarity measures is a non-trivial issue, but has received attention in recent years. A solution proposed by Kate (2016) is to use a variable-based representation of the DTW distance, and these features can be used as input data for various machine learning approaches. Further investigation would also be needed to know whether deriving separate DTW-based similarity matrices for each type of weather input variable (e.g. temperature, water supply, solar radiation, etc), and calculating interactions with the genotype component for each of these kernels, would lead to a potential gain in predictive abilities compared with the multivariate approach we used here. If the data acquired in the field allow to track variation at the genotype level for some environmental indices, such as water stress index or hyperspectral vegetation indices, it might even be possible to estimate DTW pairwise distances among hybrid-environment training instances.

It should also be noted that important pedological and management variables (e.g soil type, organic soil matter content, amounts of nitrogen fertilization) were not taken into account in the present study, but we would recommend to do so whenever this information is available, as different studies demonstrated their significant weights in neural networks (Khaki and Wang, 2019; Washburn et al., 2021), in LASSO models with G $\times$ E interactions (Rogers et al., 2021) and in gradient boosted trees models (Westhues et al., 2021a).

Authors' contribution

CW, JWM, TB and HS designed the study. CW proposed the use of DTW in MET analyses and JWM proposed the use of DTW to estimate environmental kinship matrices. CW performed the analyses and interpreted the results with input from JWM, TB and HS. CW wrote the original manuscript, which was reviewed by JWM, HS and TB. All authors agree to the final version of 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.

Acknowledgments

We authors thank the Genomes to Fields Consortium for making the genotypic, phenotypic and environmental data available for a large panel of maize hybrids. We thank the CIMMYT for the access to the wheat dataset via the data repository. This work used the Scientific Compute Cluster

at GWDG, the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen.

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary Material

Table S4.1: Description of the wheat multi-environment dataset for the WAMI panel used in the study (Sukumaran et al., 2016, 2017) and downloaded at `https://hdl.handle.net/11529/10714`. Harvest dates were not provided, so an approximate date was used to retrieve weather data.

year	Location	Code environment	Labels	Latitude	Longitude	Planting Date	Harvest Date
2010	WadMedani, Sudan	Sudan_WadMedani_2010	SW10	14.24	33.29	2009.11.20	2010.05.10
2010	Hudeiba, Sudan	Sudan_Hudeiba_2010	SH10	14.4	33.5	2009.11.20	2010.05.10
2010	Dharwad, India	India_Dharwad_2010	IDh10	15.26	75.07	2009.12.07	2010.05.10
2011	Dharwad, India	India_Dharwad_2011	IDh11	15.26	75.07	2010.12.06	2011.05.10
2010	Dongola, Sudan	Sudan_Dongola_2010	SD10	19.1	30.4	2009.12.14	2010.05.10
2010	Indore, India	India_Indore_2010	II10	22.37	75.5	2009.12.10	2010.05.10
2011	Indore, India	India_Indore_2011	II11	22.37	75.5	2010.12.07	2011.05.10
2010	Joydebpur, Bangladesh	Bangladesh_Joydebpur_2010	BJ10	23.46	90.23	2009.12.10	2010.05.10
2011	Joydebpur, Bangladesh	Bangladesh_Joydebpur_2011	BJ11	23.46	90.23	2010.12.20	2011.05.10
2010	Varanasi, India	India_Varanasi_2010	IV10	25.26	82.98	2009.11.22	2010.05.10
2010	ElMataana, Egypt	Egypt_ElMataana_2010	EE10	25.5	32.6	2009.12.22	2010.05.10
2010	Sohag, Egypt	Egypt_Sohag_2010	ES10	27.17	31.32	2009.12.20	2010.05.10
2010	ObregonMD, Mexico	Mexico_ObregonMD_2010	MD10	27.24	-109.56	2009.11.30	2010.05.10
2010	ObregonMHD, Mexico	Mexico_ObregonMHD_2010	MHD10	27.24	-109.56	2010.02.24	2010.06.20
2010	Bhairahawa, Nepal	Nepal_Bhairahawa_2010	NB10	27.32	83.25	2009.11.26	2010.05.10
2011	Bhairahawa, Nepal	Nepal_Bhairahawa_2011	NB11	27.32	83.25	2010.12.07	2011.05.10
2010	Karnal, India	India_Karnal_2010	IK10	29.43	75.57	2009.11.25	2010.05.10
2011	Karnal, India	India_Karnal_2011	IK11	29.43	75.57	2010.11.14	2011.05.10
2010	Delhi, India	India_Delhi_2010	IDe10	28.24	76.5	2009.11.22	2010.05.10
2010	Ludhiana, India	India_Ludhiana_2010	IL10	30.54	75.48	2009.11.27	2010.05.10
2011	Ludhiana, India	India_Ludhiana_2011	IL11	30.54	75.48	2010.11.22	2011.05.10
2010	Islamabad, Pakistan	Pakistan_Islamabad_2010	PI10	33.43	73.06	2009.11.25	2010.05.10
2011	Islamabad, Pakistan	Pakistan_Islamabad_2011	PI11	33.43	73.06	2010.12.11	2011.05.10

Table S4.2: Description of the subset of environments from the maize Genomes to Fields Initiative database used in the study, downloaded at `https://www.genomes2fields.org/`

year	Location	Code environment	Latitude	Longitude	Planting Date	Harvest Date
2014	Georgetown, DE	DEH1_2014	-75.204	38.63741	5/5/2014	9/29/2014
2014	Tifton, GA	GAH1_2014	-83.555	31.50654	4/4/2014	9/11/2014
2014	Ames, IA	IAH1.a_2014	-93.6962	41.99653	5/9/2014	10/20/2014
2014	Ames, IA	IAH1.b_2014	-93.6962	41.99653	5/17/2014	10/20/2014
2014	Urbana, IL	ILH1_2014	-88.2332	40.06114	5/6/2014	10/7/2014
2014	West Lafayette, IN	INH1_2014	-87.006	40.488	5/25/2014	11/18/2014
2014	Waseca, MN	MNH1_2014	-93.5341	44.06972	5/16/2014	10/16/2014
2014	Columbia, Missouri	MOH1_2014	-92.21	38.8987	5/7/2014	10/22/2014
2014	Columbia, Missouri	MOH2_2014	-92.3522	38.92875	5/5/2014	11/13/2014
2014	Lincoln, NE	NEH1_2014	-96.6567	40.83439	5/16/2014	10/22/2014
2014	North Platte, NE	NEH2_2014	-100.749	41.05298	5/15/2014	11/5/2014
2014	Aurora, NY	NYH2_2014	-76.65	42.73	5/28/2014	12/2/2014
2014	Waterloo, ON	ONH1_2014	-80.427	43.49703	5/19/2014	11/4/2014
2014	Ridgetown, ON	ONH2_2014	-81.8831	42.4542	5/27/2014	11/29/2014
2014	College Station, TX	TXH1_2014	-96.4339	30.54684	3/1/2014	8/21/2014
2014	Plainview, TX	TXH2_2014	-101.949	34.18467	4/23/2014	9/30/2014
2014	Madison, WI	WIH1_2014	-89.531	43.05706	5/9/2014	10/28/2014
2015	Georgetown, DE	DEH1_2015	-75.466	38.62998	4/29/2015	9/14/2015
2015	Tifton, GA	GAH1_2015	-83.555	31.5065	4/1/2015	8/26/2015
2015	Urbana, IL	ILH1_2015	-88.2336	40.06031	4/30/2015	10/2/2015
2015	West Lafayette, IN	INH1_2015	-87.0006	40.47666	5/14/2015	10/15/2015
2015	Manhattan, KS	KSH1_2015	-96.6052	39.21586	4/23/2015	9/21/2015
2015	Waseca, MN	MNH1_2015	-93.5349	44.07099	5/19/2015	11/10/2015
2015	Columbia, MO	MOH1_2015	-92.2083	38.89608	5/4/2015	10/6/2015
2015	Columbia, MO	MOH2_2015	-92.2076	38.89853	5/5/2015	10/2/2015
2015	North Platte, NE	NEH2_2015	-100.747	41.05097	4/23/2015	10/26/2015
2015	Brule, NE	NEH3_2015	-101.988	41.15841	6/10/2015	12/14/2015
2015	Aurora, NY	NYH2_2015	-76.6533	42.73271	5/7/2015	11/16/2015
2015	Aurora, NY	NYH3_2015	-76.6562	42.72351	5/23/2015	11/18/2015
2015	South Charleston, OH	OHH1_2015	-83.6644	39.85542	5/21/2015	10/17/2015
2015	Waterloo, ON	ONH1_2015	-80.4489	43.49735	4/30/2015	10/15/2015
2015	Ridgetown, ON	ONH2_2015	-81.8809	42.45433	5/7/2015	10/12/2015
2015	New Underwood, SD	SDH1_2015	-102.93	44.20888	5/22/2015	10/28/2015
2015	College Station, TX	TXH1_2015	-96.4347	30.54677	3/7/2015	7/28/2015
2016	Georgetown, DE	DEH1_2016	-75.4516	38.64865	4/25/2016	9/14/2016
2016	Tifton, GA	GAH2_2016	-83.3114	33.71736	5/25/2016	10/12/2016

2016	Glidden, IA	IAH2_2016	-94.7275	42.06591	4/25/2016	10/11/2016
2016	Keystone, IA	IAH3_2016	-92.2602	41.98738	4/24/2016	10/6/2016
2016	Ames, IA	IAH4_2016	-93.6999	41.9975	4/26/2016	10/17/2016
2016	Urbana, IL	ILH1.a_2016	-88.2333	40.06119	5/6/2016	10/9/2016
2016	Urbana, IL	ILH1.b_2016	-88.2333	40.06119	4/26/2016	10/9/2016
2016	West Lafayette, IN	INH1_2016	-86.9901	40.47835	5/19/2016	10/6/2016
2016	Manhattan, KS	KSH1_2016	-96.6294	39.14409	4/15/2016	9/27/2016
2016	East Lansing, MI	MIH1_2016	-84.2954	42.4118	5/24/2016	11/16/2016
2016	Waseca, MN	MNH1_2016	-93.5342	44.06616	5/17/2016	10/20/2016
2016	Columbia, MO	MOH1_2016	-92.2075	38.89497	5/23/2016	10/7/2016
2016	Mead, NE	NEH1_2016	-96.4173	41.16636	5/6/2016	11/8/2016
2016	Mead, NE	NEH4_2016	-96.4172	41.16702	6/7/2016	11/9/2016
2016	Aurora, NY	NYH2_2016	-76.6551	42.72543	5/10/2016	12/8/2016
2016	Waterloo, ON	ONH1_2016	-80.452	43.49968	5/4/2016	10/15/2016
2016	Ridgetown, ON	ONH2_2016	-81.8835	42.45275	5/11/2016	11/1/2016
2016	College Station, TX	TXH1_2016	-96.4347	30.54677	3/4/2016	8/5/2016
2016	Madison, WI	WIH1_2016	-89.5317	43.05687	5/9/2016	10/14/2016
2016	Arlington, WI	WIH2_2016	-89.3402	43.32695	5/24/2016	10/25/2016
2017	Marianna, AR	ARH1_2017	-90.76	34.7299	4/25/2017	9/11/2017
2017	Keiser, AR	ARH2_2017	-90.075	35.6747	4/17/2017	9/16/2017
2017	Fort Collins, CO	COH1_2017	-105	40.64786	5/31/2017	11/22/2017
2017	Georgetown, DE	DEH1_2017	-75.4339	38.66956	4/28/2017	9/8/2017
2017	Tifton, GA	GAH1_2017	-83.5592	31.50825	4/4/2017	9/7/2017
2017	Watkinsville, GA	GAH2_2017	-83.2978	33.72686	5/2/2017	9/8/2017
2017	Ames, IA	IAH4_2017	-93.6886	41.99439	5/7/2017	10/17/2017
2017	East Lansing, MI	MIH1_2017	-84.4941	42.6819	5/22/2017	10/20/2017
2017	Columbia, MO	MOH1_2017	-92.2048	38.89238	5/15/2017	10/19/2017
2017	Aurora, NY	NYH2_2017	-76.6533	42.73219	5/18/2017	11/24/2017
2017	Aurora, NY	NYH3_2017	-76.6532	42.73306	5/18/2017	11/24/2017
2017	Waterloo, ON	ONH1_2017	-80.4261	43.49604	5/17/2017	10/31/2017
2017	College Station, TX	TXH1-Dry_2017	-96.4326	30.54535	3/3/2017	7/25/2017
2017	College Station, TX	TXH1-Early_2017	-96.4326	30.54535	3/3/2017	7/31/2017
2017	College Station, TX	TXH1-Late_2017	-96.4326	30.54535	4/6/2017	8/10/2017
2017	Madison, WI	WIH1_2017	-89.5311	43.05718	5/5/2017	10/19/2017
2017	Arlington, WI	WIH2_2017	-89.3358	43.32453	5/11/2017	11/6/2017

Table S4.3: GDD requirements for the three groups of maturity used with the maize dataset. VE, emergence; V7, collar of 7-th leaf visible; V15, collar of 15-th leaf visible; R1, silk emergence (silk visible outside the husk), tassel shedding pollen; R3, milk stage (kernel yellow outside, while inner fluid is milky white due to accumulating starch); R4, dough stage (continued starch accumulation in the endosperm).

Corn growth stages	GDD accumulation in °C (group 1)	GDD accumulation in °C (group 2)	GDD accumulation in °C (group 3)
VE	95	110	120
V7	280	315	350
V15	560	630	700
R1	680	760	850
R3	900	1095	1210
R4	1160	1350	1450

Table S4.4: GDD requirements for the two groups of maturity used with the wheat dataset

Wheat growth stages	GDD accumulation in °C (group 1)	GDD accumulation in °C (group 2)
emergence	105	120
crown root initiation	250	270
leaf initiation	310	330
leaf development	840	900
booting stage	1130	1250
heading	1230	1365
anthesis	1290	1405
milky stage	1560	1650
soft dough	1740	1850

Table S4.5: Groups of environments based on estimated thermal time to flowering (maize dataset)

Group 1	Group 2	Group 3
IAH1.a_2014	DEH1_2014	GAH1_2014
NYH2_2014	IAH1.b_2014	INH1_2014
ONH1_2014	ILH1_2014	NEH2_2014
ONH2_2014	MNH1_2014	TXH2_2014
OHH1_2015	MOH1_2014	KSH1_2015
ONH1_2015	MOH2_2014	MOH1_2015
ONH2_2015	NEH1_2014	MOH2_2015
SDH1_2015	TXH1_2014	NEH2_2015
NEH1_2016	WIH1_2014	KSH1_2016
ONH1_2016	DEH1_2015	MOH1_2016
ONH2_2016	GAH1_2015	ARH1_2017
WIH1_2016	ILH1_2015	
WIH2_2016	INH1_2015	
COH1_2017	MNH1_2015	
DEH1_2017	NEH3_2015	
GAH2_2017	NYH2_2015	
MIH1_2017	NYH3_2015	
NYH2_2017	TXH1_2015	
ONH1_2017	DEH1_2016	
WIH1_2017	GAH2_2016	
WIH2_2017	IAH2_2016	
	IAH3_2016	
	IAH4_2016	
	ILH1.a_2016	
	ILH1.b_2016	
	INH1_2016	
	MIH1_2016	
	MNH1_2016	
	NEH4_2016	
	NYH2_2016	
	TXH1_2016	
	ARH2_2017	
	GAH1_2017	
	IAH4_2017	
	MOH1_2017	
	NYH3_2017	
	TXH1-Dry_2017	
	TXH1-Early_2017	
	TXH1-Late_2017	

Table S4.6: Groups of environments based on estimated thermal time to flowering (wheat dataset)

Group 1	Group 2
Egypt_ElMataana_2010	Bangladesh_Joydebpur_2010
Egypt_Sohag_2010	Bangladesh_Joydebpur_2011
India_Dharwad_2010	India_Dharwad_2011
India_Varanasi_2010	India_Indore_2010
Nepal_Bhairahawa_2010	India_Indore_2011
Nepal_Bhairahawa_2011	India_Karnal_2010
Pakistan_Islamabad_2010	India_Ludhiana_2010
Pakistan_Islamabad_2011	India_Ludhiana_2011
Sudan_Hudeiba_2010	Mexico_ObregonMD_2010
Sudan_WadMedani_2010	Mexico_ObregonMHD_2010
India_Karnal_2011	Sudan_Dongola_2010
India_Delhi_2010	

Table S4.7: Estimates of variance components with the different models with the wheat dataset

Model	component	value	Sum of variance components in the model	Percentage of the across-environment variance
M1	residual	0.6336	3.81132	0.16624274
M1	A	0.12118	3.81132	0.03179383
M1	E	3.05654	3.81132	0.80196343
M2	residual	0.50566	3.66478	0.1379785
M2	A	0.11117	3.66478	0.03033528
M2	E	2.91344	3.66478	0.79498451
M2	A × E	0.1345	3.66478	0.03670171
M4	residual	0.63447	5.60514	0.11319454
M4	A	0.09647	5.60514	0.01721019
M4	Y	1.7491	5.60514	0.31205307
M4	L	1.07431	5.60514	0.19166468
M4	E	2.0508	5.60514	0.36587752
M5	residual	0.49351	4.78691	0.10309564
M5	A	0.08082	4.78691	0.0168842
M5	Y	1.18455	4.78691	0.24745629
M5	L	0.82977	4.78691	0.17334189
M5	E	1.99722	4.78691	0.41722632
M5	A × Y	0.06347	4.78691	0.01325894
M5	A × L	0.06741	4.78691	0.0140829
M5	A × E	0.07015	4.78691	0.01465382
M7 with $W_{ECs\_windows}$	residual	0.63392	4.88295	0.12982365
M7 with $W_{ECs\_windows}$	A	0.10472	4.88295	0.02144574
M7 with $W_{ECs\_windows}$	E	1.09148	4.88295	0.22352972
M7 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	3.05282	4.88295	0.62520089
M7 with W_{ECs_stages}	residual	0.63408	4.35829	0.14548832
M7 with W_{ECs_stages}	A	0.10551	4.35829	0.02421003
M7 with W_{ECs_stages}	E	1.90457	4.35829	0.43699977
M7 with W_{ECs_stages}	W_{ECs_stages}	1.71412	4.35829	0.39330189
M11 with $W_{ECs\_windows}$	residual	0.53522	3.61168	0.14819252
M11 with $W_{ECs\_windows}$	A	0.10388	3.61168	0.02876094
M11 with $W_{ECs\_windows}$	E	1.97545	3.61168	0.54696096
M11 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	0.86208	3.61168	0.23869306
M11 with $W_{ECs\_windows}$	A × $W_{ECs\_windows}$	0.13505	3.61168	0.03739252
M11 with W_{ECs_stages}	residual	0.52953	3.77702	0.14019724
M11 with W_{ECs_stages}	A	0.10343	3.77702	0.0273841
M11 with W_{ECs_stages}	E	1.84674	3.77702	0.48894057
M11 with W_{ECs_stages}	W_{ECs_stages}	1.15336	3.77702	0.30536264
M11 with W_{ECs_stages}	A × W_{ECs_stages}	0.14396	3.77702	0.03811544
M12 with $W_{ECs\_windows}$	residual	0.49275	3.41614	0.14424322
M12 with $W_{ECs\_windows}$	A	0.10011	3.41614	0.02930474
M12 with $W_{ECs\_windows}$	E	2.13147	3.41614	0.62394007
M12 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	0.52509	3.41614	0.15370843
M12 with $W_{ECs\_windows}$	A × $W_{ECs\_windows}$	0.08298	3.41614	0.02429024
M12 with $W_{ECs\_windows}$	A × E	0.08374	3.41614	0.0245133
M12 with W_{ECs_stages}	residual	0.49629	3.57978	0.13863677
M12 with W_{ECs_stages}	A	0.09947	3.57978	0.02778759
M12 with W_{ECs_stages}	E	2.19692	3.57978	0.61370276
M12 with W_{ECs_stages}	W_{ECs_stages}	0.62187	3.57978	0.17371659
M12 with W_{ECs_stages}	A × W_{ECs_stages}	0.082	3.57978	0.02290548
M12 with W_{ECs_stages}	A × E	0.08323	3.57978	0.02325081
M8 with K_{DTW}	residual	0.63412	6.98118	0.09083266
M8 with K_{DTW}	A	0.10517	6.98118	0.01506473
M8 with K_{DTW}	K_{DTW}	5.0272	6.98118	0.72010785
M8 with K_{DTW}	E	1.21469	6.98118	0.17399476
M13 with K_{DTW}	residual	0.5164	6.3915	0.0807956
M13 with K_{DTW}	A	0.07959	6.3915	0.01245264
M13 with K_{DTW}	E	1.40288	6.3915	0.21949149
M13 with K_{DTW}	K_{DTW}	4.16454	6.3915	0.65157432
M13 with K_{DTW}	A × K_{DTW}	0.22809	6.3915	0.03568595

Table S4.8: Estimates of variance components with the different models with the G2F dataset

Model	component	value	Sum of variance components in the model	Percentage of the across-environment variance
M1	residual	625.531	2143.85	0.291779
M1	A	425.2847	2143.85	0.198374
M1	E	1093.034	2143.85	0.509846
M2	residual	446.8275	2205.153	0.202629
M2	A	368.3793	2205.153	0.167054
M2	E	1167.278	2205.153	0.529341
M2	A × E	222.6688	2205.153	0.100977
M3	A	116.4151	1925.983	0.060445
M3	D	143.7168	1925.983	0.07462
M3	E	1078.345	1925.983	0.559893
M3	residual	587.5066	1925.983	0.305042
M4	residual	625.7422	2525.817	0.247739
M4	A	420.661	2525.817	0.166545
M4	Y	412.3418	2525.817	0.163251
M4	L	583.4264	2525.817	0.230985
M4	E	483.6456	2525.817	0.191481
M5	residual	446.9876	2401.479	0.18613
M5	A	292.3829	2401.479	0.121751
M5	Y	323.437	2401.479	0.134682
M5	L	518.7515	2401.479	0.216013
M5	E	557.5511	2401.479	0.23217
M5	A × Y	53.03081	2401.479	0.022083
M5	A × L	90.05956	2401.479	0.037502
M5	A × E	119.2788	2401.479	0.049669
M6	A	107.8048	1964.802	0.054868
M6	D	126.8704	1964.802	0.064572
M6	E	1093.878	1964.802	0.556737
M6	A × E	83.88279	1964.802	0.042693
M6	D × E	42.61663	1964.802	0.02169
M6	residual	509.749	1964.802	0.25944
M7 with W_{ECs_stages}	residual	625.6553	2136.641	0.292822
M7 with W_{ECs_stages}	A	423.2289	2136.641	0.198081
M7 with W_{ECs_stages}	E	531.3797	2136.641	0.248699
M7 with W_{ECs_stages}	W_{ECs_stages}	556.3771	2136.641	0.260398
M7 with $W_{ECs\_windows}$	residual	625.3568	2154.723	0.290226
M7 with $W_{ECs\_windows}$	A	425.895	2154.723	0.197656
M7 with $W_{ECs\_windows}$	E	520.073	2154.723	0.241364
M7 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	583.3982	2154.723	0.270753
M8 with K_{DTW}	residual	625.6075	2643.514	0.236658
M8 with K_{DTW}	A	423.6824	2643.514	0.160272
M8 with K_{DTW}	K_{DTW}	1261.188	2643.514	0.477088
M8 with K_{DTW}	E	333.0355	2643.514	0.125982
M9 with W_{ECs_stages}	A	110.0369	1771.063	0.06213
M9 with W_{ECs_stages}	D	143.2094	1771.063	0.080861
M9 with W_{ECs_stages}	E	536.1242	1771.063	0.302713
M9 with W_{ECs_stages}	W_{ECs_stages}	393.9452	1771.063	0.222434
M9 with W_{ECs_stages}	residual	587.7472	1771.063	0.331861
M9 with $W_{ECs\_windows}$	A	108.576	1904.313	0.057016
M9 with $W_{ECs\_windows}$	D	143.0668	1904.313	0.075128
M9 with $W_{ECs\_windows}$	E	520.6062	1904.313	0.273383
M9 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	544.3449	1904.313	0.285848
M9 with $W_{ECs\_windows}$	residual	587.719	1904.313	0.308625
M10 with K_{DTW}	A	108.503	2409.595	0.04503
M10 with K_{DTW}	D	144.1336	2409.595	0.059817
M10 with K_{DTW}	E	293.8054	2409.595	0.121931
M10 with K_{DTW}	K_{DTW}	1275.31	2409.595	0.529263
M10 with K_{DTW}	residual	587.8432	2409.595	0.243959
M11 with W_{ECs_stages}	residual	511.5894	2068.95	0.24727
M11 with W_{ECs_stages}	A	368.6503	2068.95	0.178182
M11 with W_{ECs_stages}	E	605.6739	2068.95	0.292745
M11 with W_{ECs_stages}	W_{ECs_stages}	381.779	2068.95	0.184528
M11 with W_{ECs_stages}	A × W_{ECs_stages}	201.2571	2068.95	0.097275
M11 with $W_{ECs\_windows}$	residual	504.3659	2165.546	0.232905
M11 with $W_{ECs\_windows}$	A	378.6376	2165.546	0.174846
M11 with $W_{ECs\_windows}$	E	573.1125	2165.546	0.26465
M11 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	492.5156	2165.546	0.227433
M11 with $W_{ECs\_windows}$	A × W_{ECs_stages}	216.9148	2165.546	0.100166

M12 with W_{ECs_stages}	residual	446.6623	2074.212	0.215341
M12 with W_{ECs_stages}	A	360.9783	2074.212	0.174032
M12 with W_{ECs_stages}	E	574.6397	2074.212	0.27704
M12 with W_{ECs_stages}	W_{ECs_stages}	462.6648	2074.212	0.223056
M12 with W_{ECs_stages}	A × W_{ECs_stages}	44.12939	2074.212	0.021275
M12 with W_{ECs_stages}	A × E	185.1374	2074.212	0.089257
M12 with $W_{ECs\_windows}$	residual	447.2762	2161.16	0.206961
M12 with $W_{ECs\_windows}$	A	361.5465	2161.16	0.167293
M12 with $W_{ECs\_windows}$	E	562.449	2161.16	0.260253
M12 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	563.1639	2161.16	0.260584
M12 with $W_{ECs\_windows}$	A × W_{ECs_stages}	45.19264	2161.16	0.020911
M12 with $W_{ECs\_windows}$	A × E	181.5314	2161.16	0.083997
M13 with K_{DTW}	residual	454.1882	2693.111	0.168648
M13 with K_{DTW}	A	145.0687	2693.111	0.053867
M13 with K_{DTW}	E	408.5847	2693.111	0.151715
M13 with K_{DTW}	K_{DTW}	1230.011	2693.111	0.456725
M13 with K_{DTW}	A × K_{DTW}	455.2589	2693.111	0.169046
M14 with W_{ECs_stages}	A	104.2291	1917.596	0.054354
M14 with W_{ECs_stages}	D	130.0033	1917.596	0.067795
M14 with W_{ECs_stages}	E	551.5485	1917.596	0.287625
M14 with W_{ECs_stages}	A × W_{ECs_stages}	138.8656	1917.596	0.072416
M14 with W_{ECs_stages}	D × W_{ECs_stages}	56.17767	1917.596	0.029296
M14 with W_{ECs_stages}	W_{ECs_stages}	468.0495	1917.596	0.244081
M14 with W_{ECs_stages}	residual	468.7226	1917.596	0.244432
M14 with $W_{ECs\_windows}$	A	103.0376	1959.29	0.052589
M14 with $W_{ECs\_windows}$	D	134.2496	1959.29	0.06852
M14 with $W_{ECs\_windows}$	E	573.633	1959.29	0.292776
M14 with $W_{ECs\_windows}$	A × $W_{ECs\_windows}$	154.8289	1959.29	0.079023
M14 with $W_{ECs\_windows}$	D × $W_{ECs\_windows}$	54.44863	1959.29	0.02779
M14 with $W_{ECs\_windows}$	$W_{ECs\_windows}$	476.3663	1959.29	0.243132
M14 with $W_{ECs\_windows}$	residual	462.7256	1959.29	0.23617
M15 with K_{DTW}	A	53.97941	2695.129	0.020029
M15 with K_{DTW}	D	46.54905	2695.129	0.017272
M15 with K_{DTW}	E	262.2352	2695.129	0.0973
M15 with K_{DTW}	K_{DTW}	1482.629	2695.129	0.550114
M15 with K_{DTW}	A × K_{DTW}	252.6501	2695.129	0.093743
M15 with K_{DTW}	D × K_{DTW}	197.0667	2695.129	0.07312
M15 with K_{DTW}	residual	400.0203	2695.129	0.148423

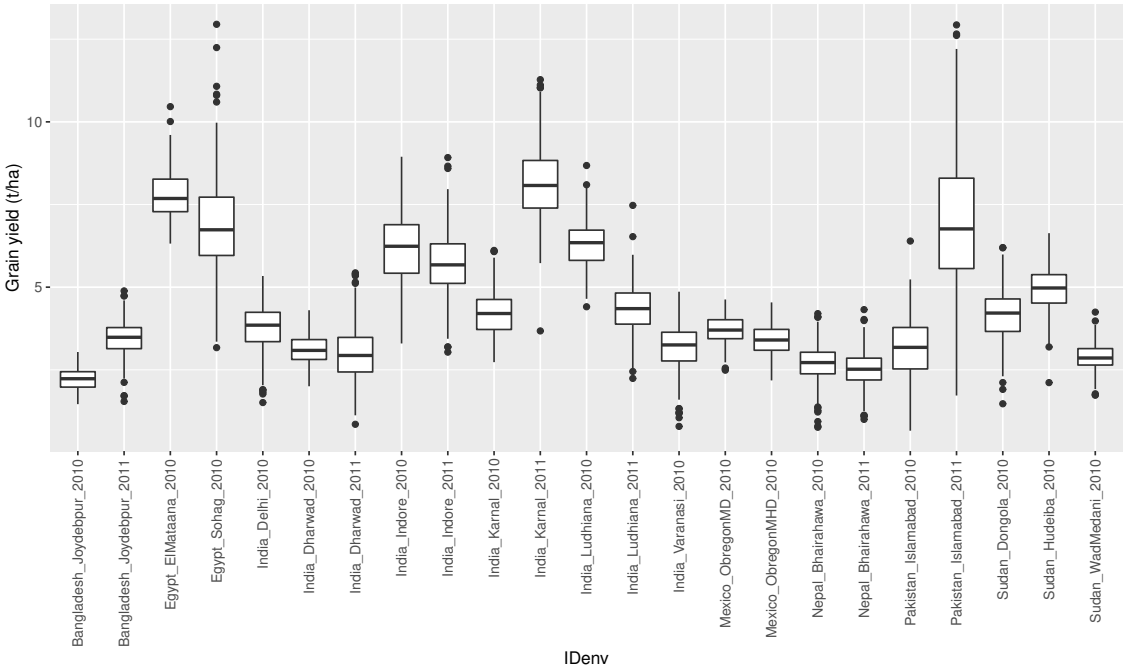


Figure S4.1: Boxplot showing phenotypic values for grain yield in the wheat dataset measured in 23 environments. Whiskers of the boxplots represent 1.5 times the interquartile range of the data, and the middle is marked at the median of the observations.

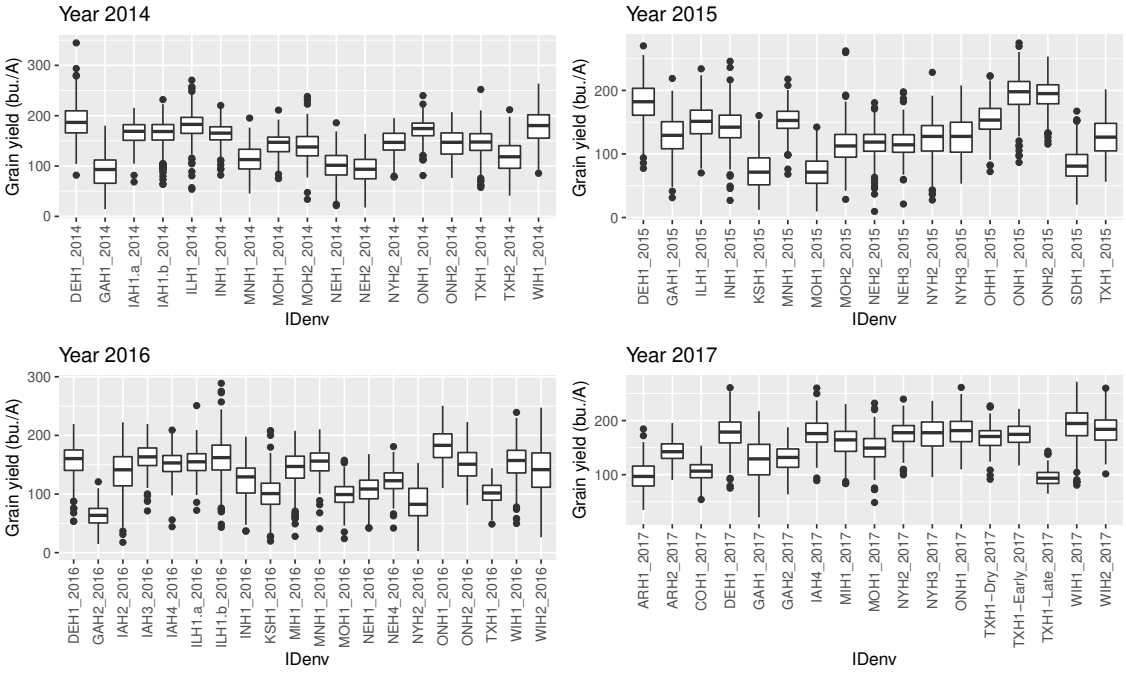


Figure S4.2: Boxplot showing phenotypic values for grain yield in the maize dataset measured in 71 environments. Whiskers of the boxplots represent 1.5 times the interquartile range of the data, and the middle is marked at the median of the observations.

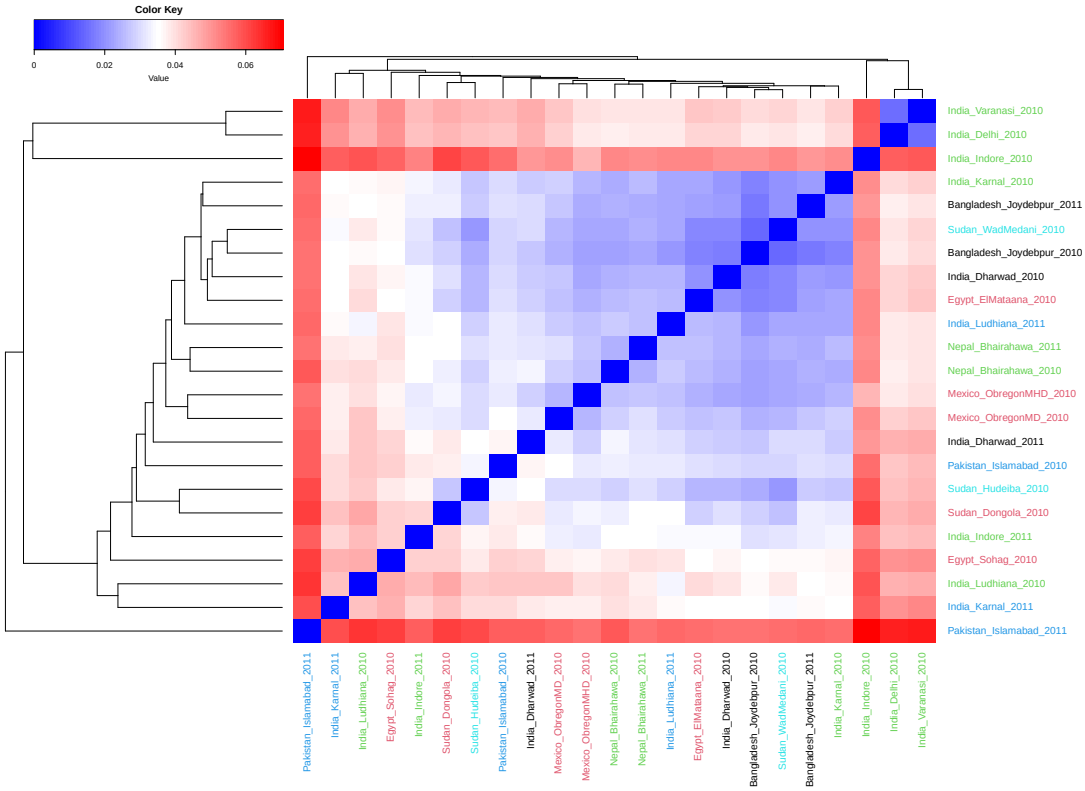


Figure S4.3: Heatmap of environments of the wheat dataset, based on Euclidean distances computed using marker effects estimated with ridge regression. Red color shows larger distance, while blue color indicates a smaller distance between two given environments. The colors in the labels correspond to the colors from the clustering using DTW distance from Figure 4.4.

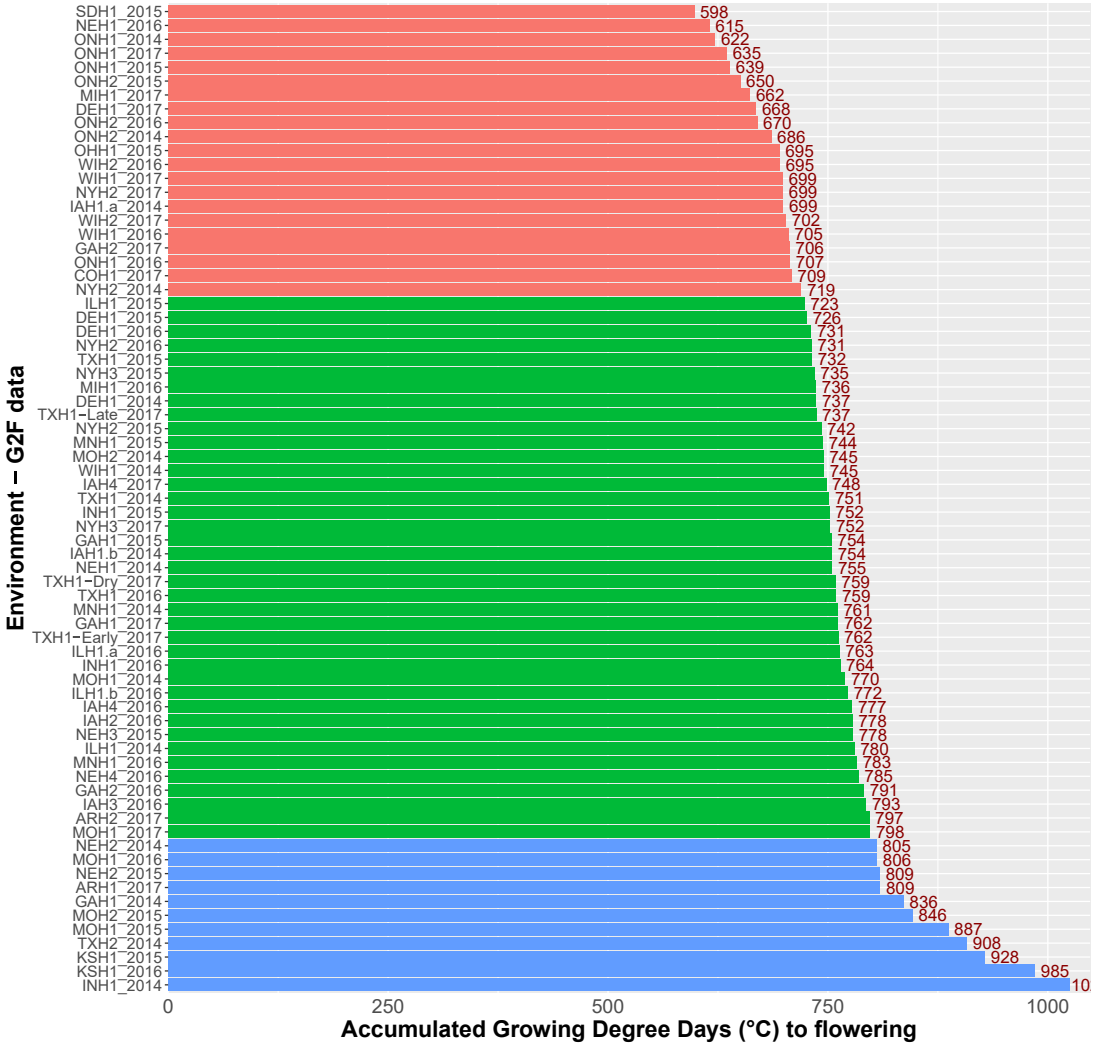


Figure S4.4: Thermal time to flowering (°C) in the Genomes to Fields maize dataset. Three different types of phenological development timing were determined based on these data and used to approximate plant developmental stage within each environment. The orange environments were assigned to group 1, the green environments to group 2 and the blue environments to group 3 in Table S4.4.

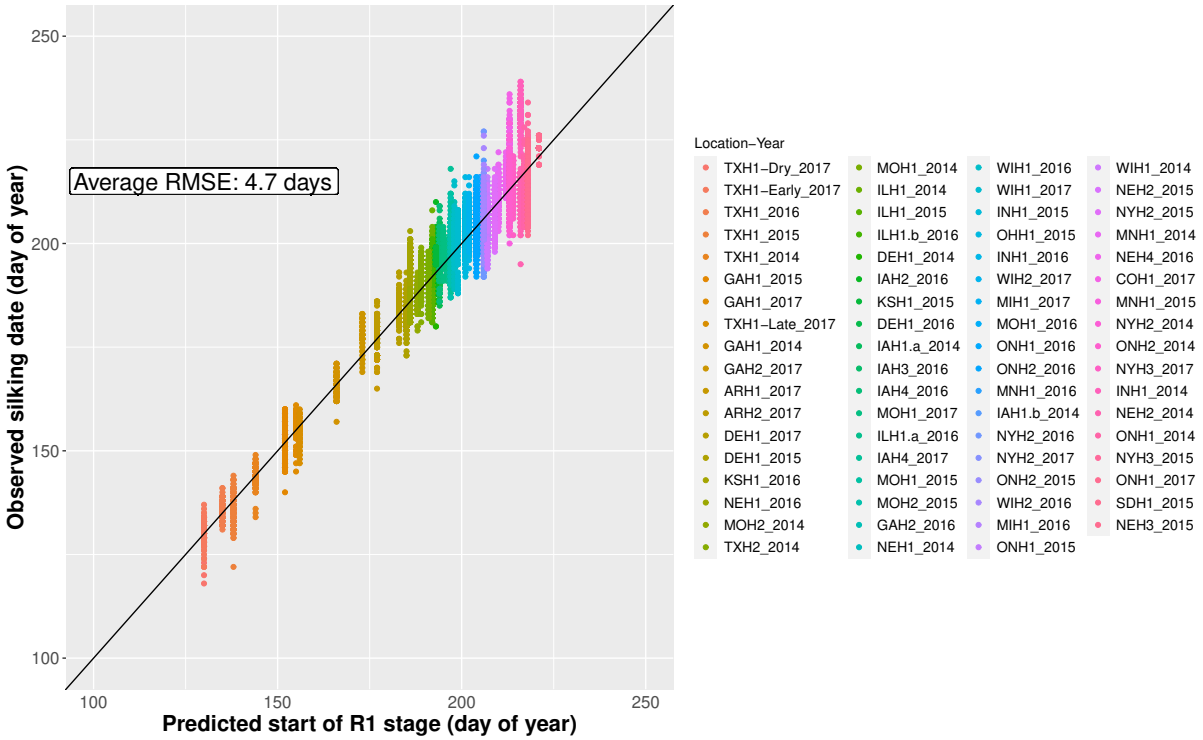


Figure S4.5: Observed hybrid silking date (scored as the day at which 50% of plants within a plot show silk emergence) within each environment against the predicted start of the R1 stage (silk emergence = reproductive corn stage), calculated using estimations of the accumulated thermal time required to reach this stage. Each environment had been beforehand assigned to a maturity group (see S4.4), based on the average estimated GDD at silking date across all hybrids grown in this environment. For sake of simplicity, growth stages were predicted within each environment without accounting for variability of earliness among hybrids. The approximated start of each stage within each environment was then predicted based on the accumulated thermal time (GDDs calculated with a base temperature of 10°C). The colors on the plot along the x-axis appear in the same order as in the legend. Average RMSE was equal to 4.7 days.

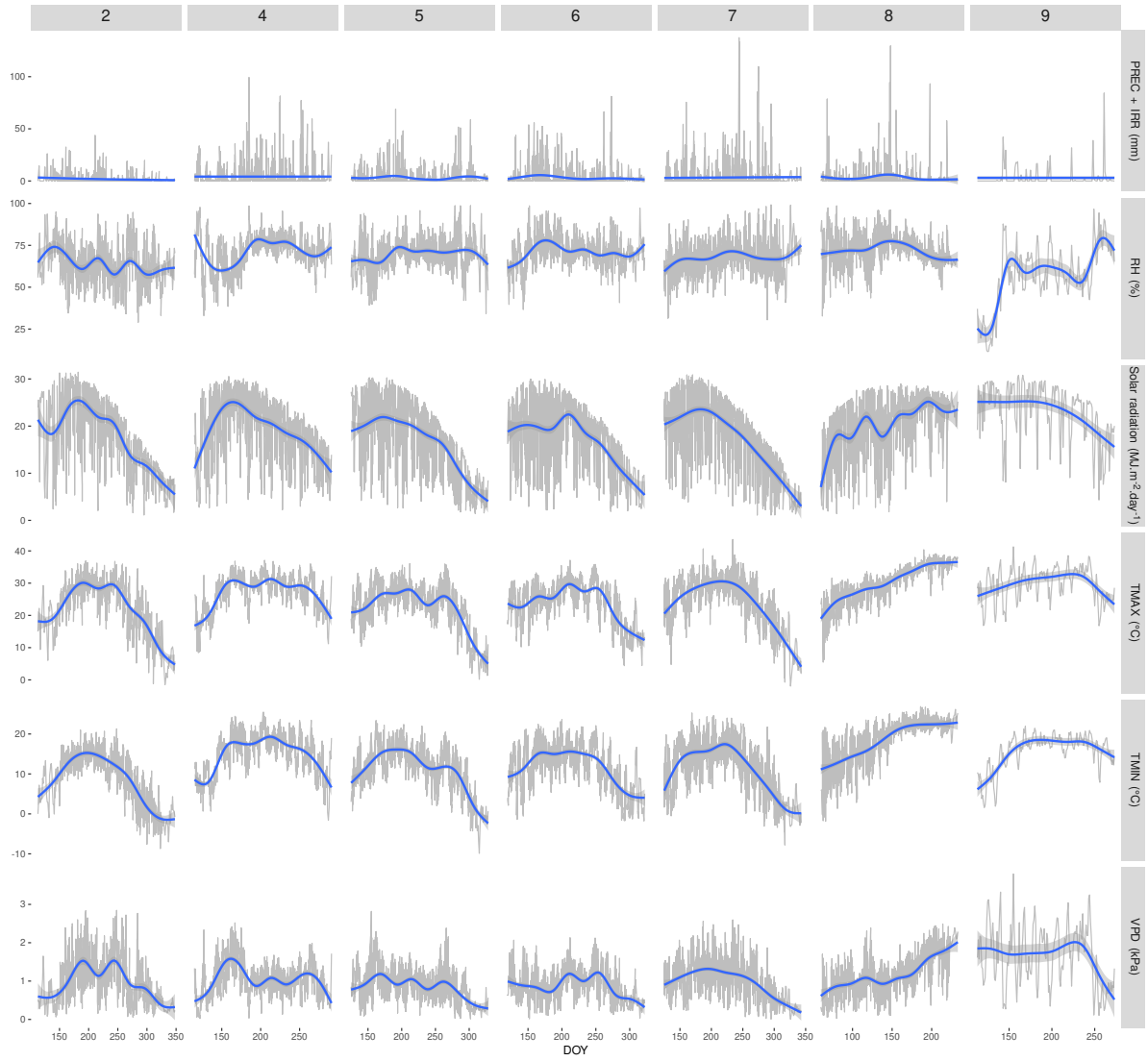


Figure S4.6: Daily weather time series for 6 climatic variables (in rows) characterizing the remaining clusters (in columns) in the maize dataset. Individual time series for each environment are depicted in grey, and the blue line represents the loess smoothing function to help seeing patterns.

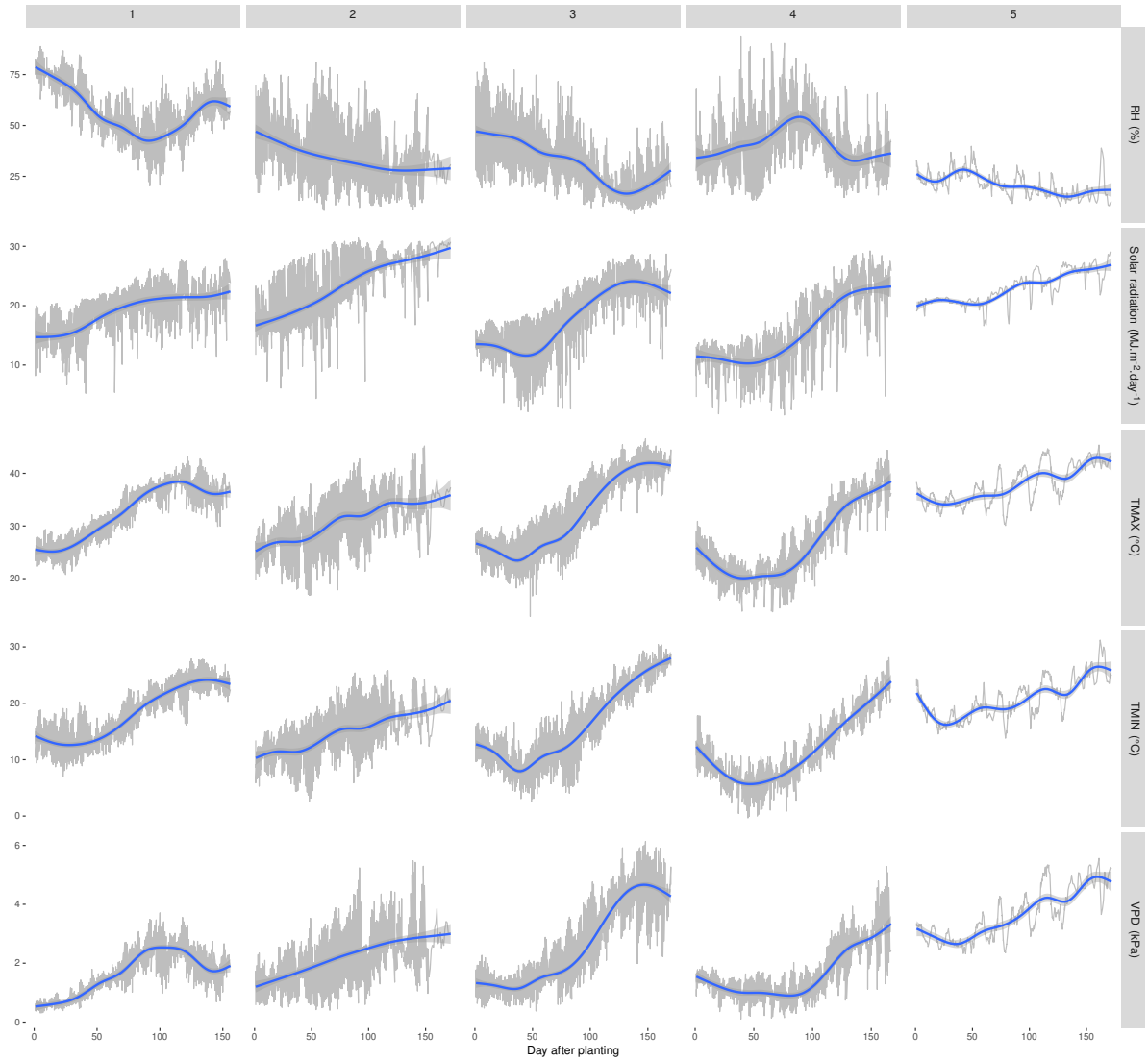
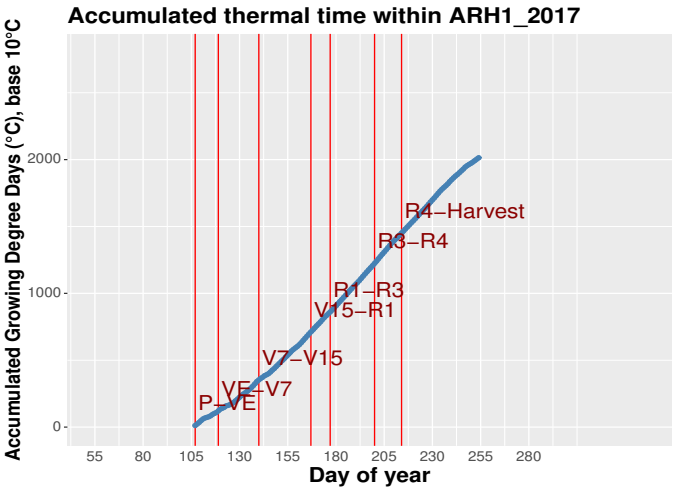
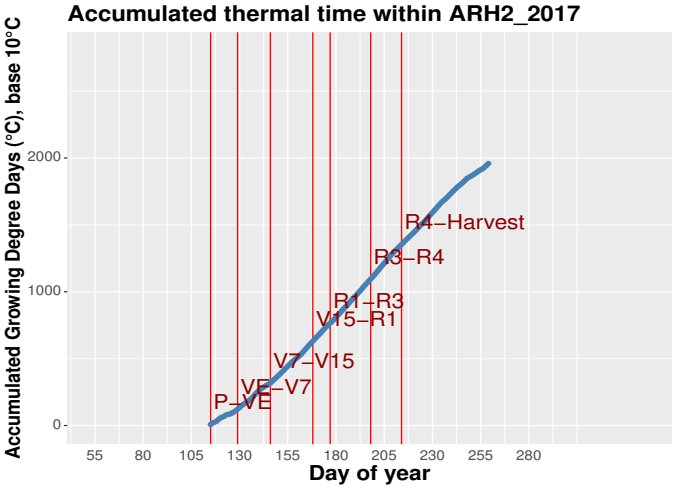
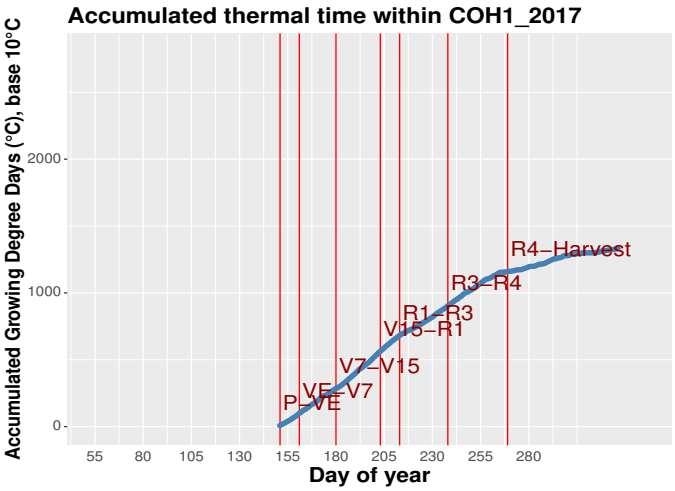


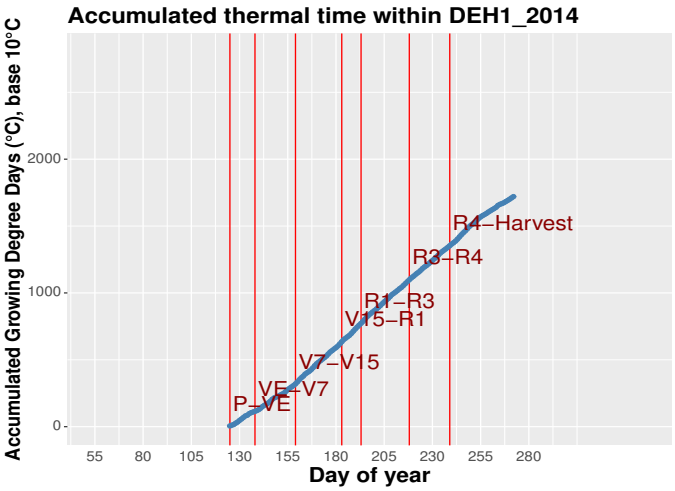
Figure S4.7: Daily weather time series for 5 climatic variables (in rows) characterizing the different clusters (in columns) in the wheat dataset. Individual time series for each environment are depicted in grey, and the blue line represents the loess smoothing function to help seeing patterns.

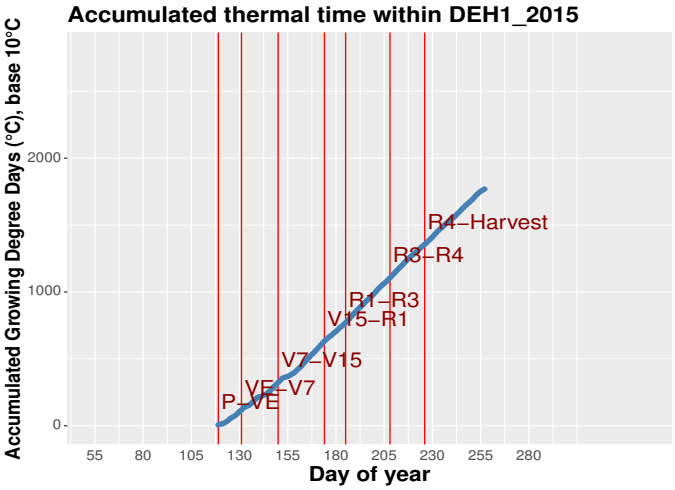
Figure S4.8: Prediction of growth stages within each environment of the maize G2F dataset using total accumulated growing degree days, achieved using *learnMET* with the function *compute_EC_gdd()*. P, planting day; VE, emergence; V7, seven leaf collar (vegetative stage); V15, fifteen collars (vegetative stage); R1, silking (reproductive stage); R3, kernel milk stage (grain fill stage); R4, kernel dough stage (grain fill stage).

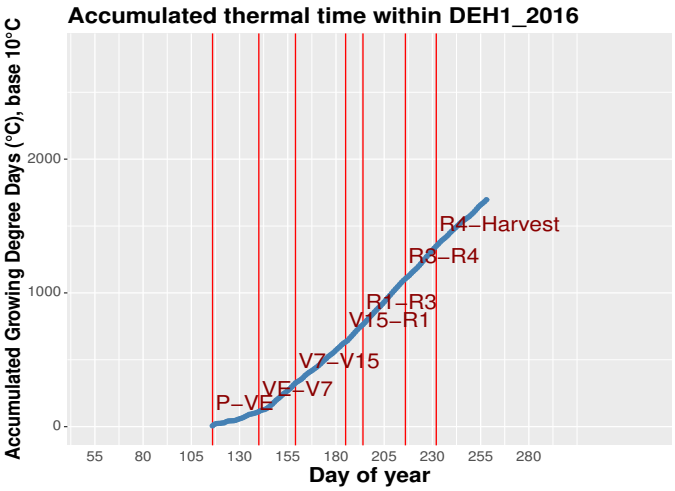


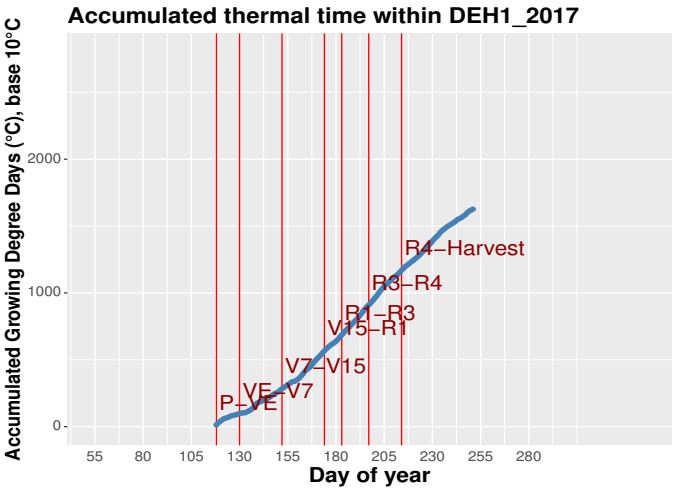


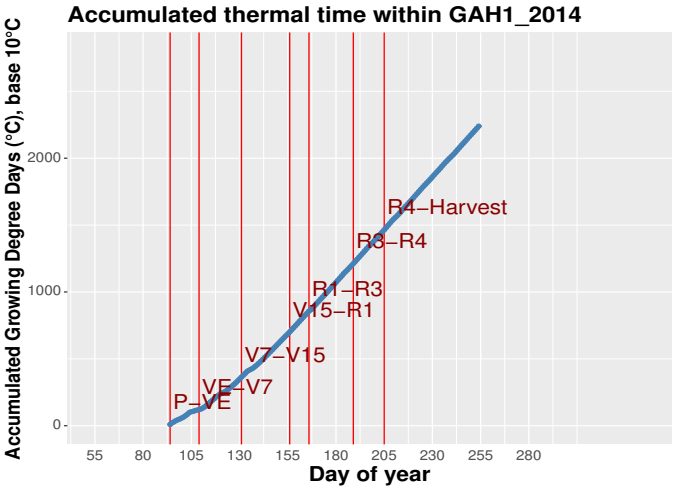

Accumulated thermal time within COH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

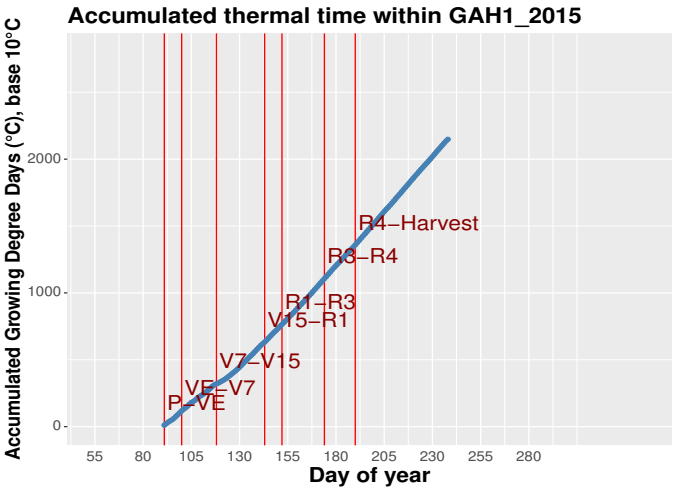

Accumulated thermal time within DEH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

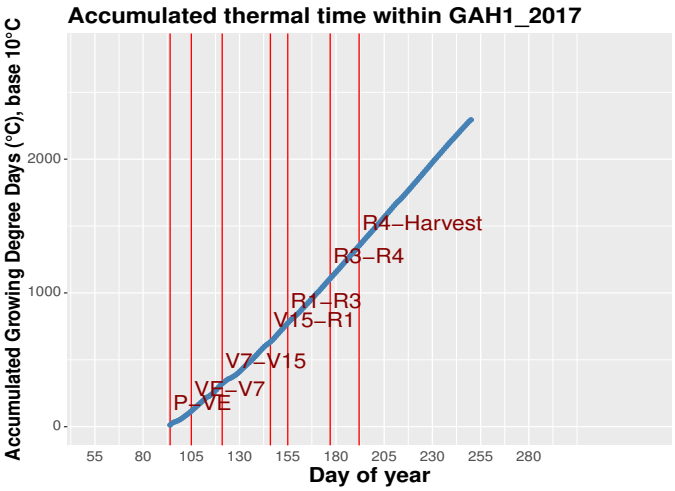

Accumulated thermal time within DEH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

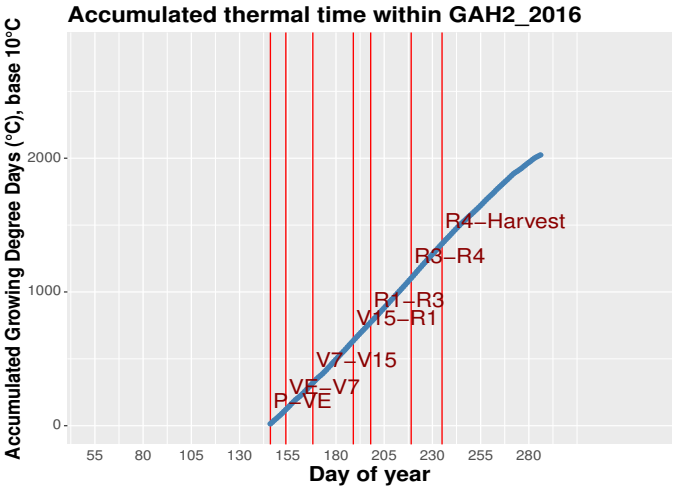

Accumulated thermal time within DEH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

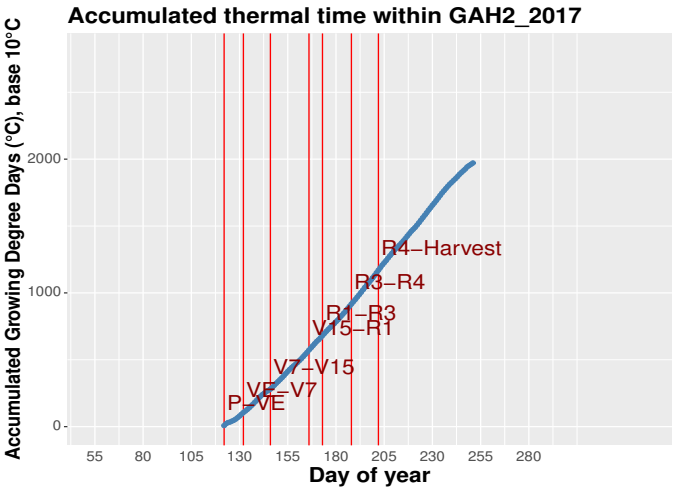

Accumulated thermal time within DEH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

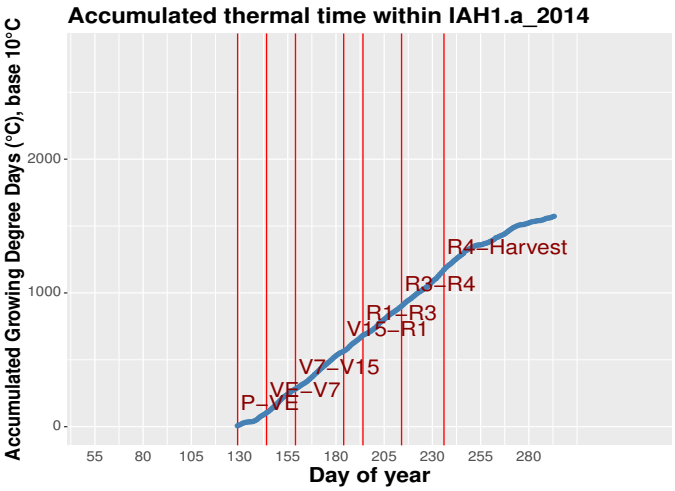

Accumulated thermal time within GAH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

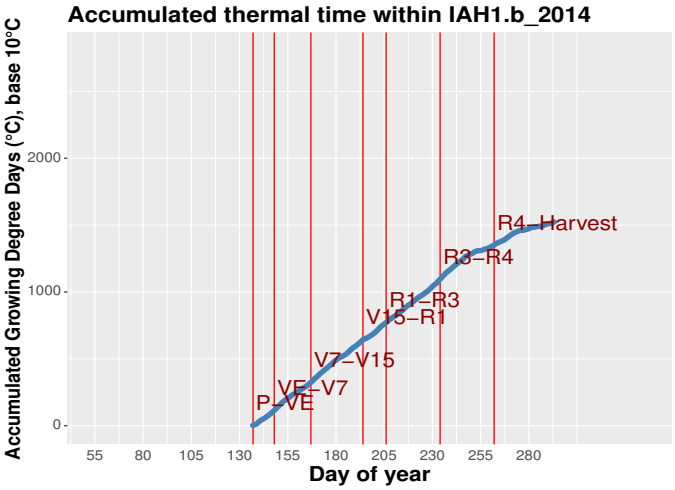

Accumulated thermal time within GAH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

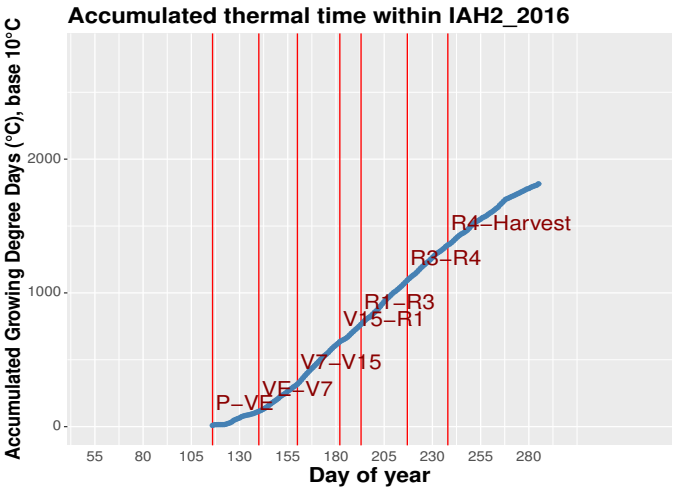

Accumulated thermal time within GAH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

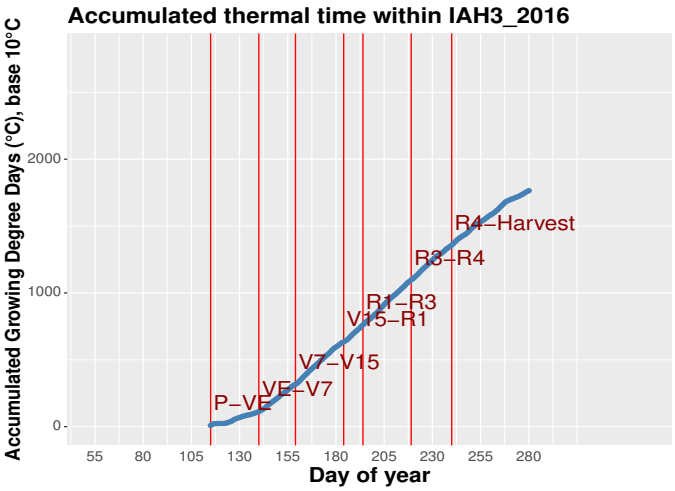

Accumulated thermal time within GAH2_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

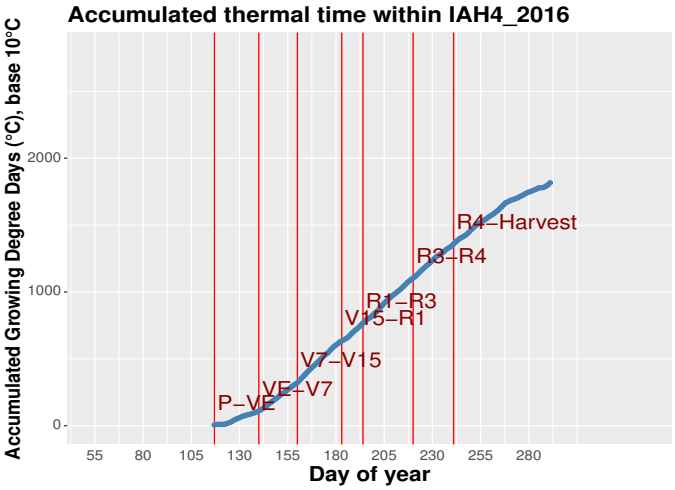

Accumulated thermal time within GAH2_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

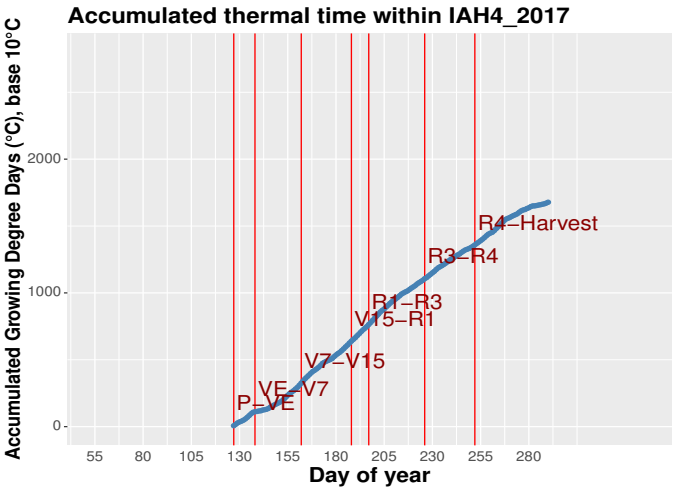

Accumulated thermal time within IAH1.a_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

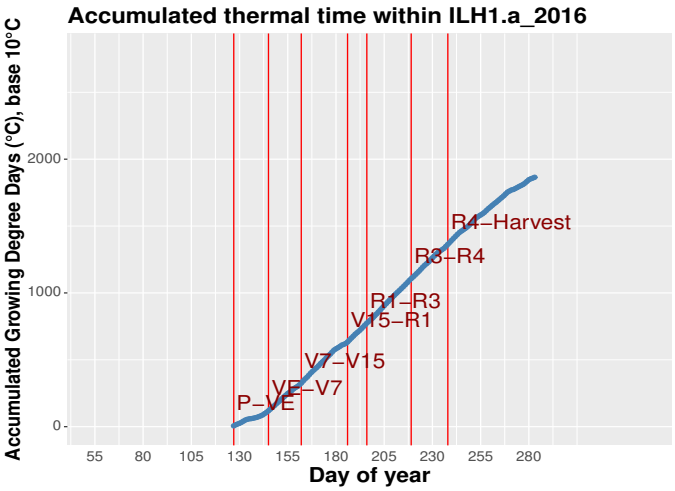

Accumulated thermal time within IAH1.b_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

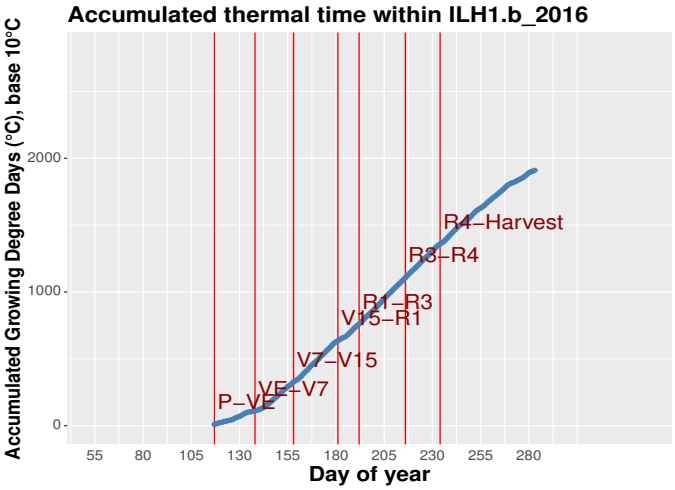

Accumulated thermal time within IAH2_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

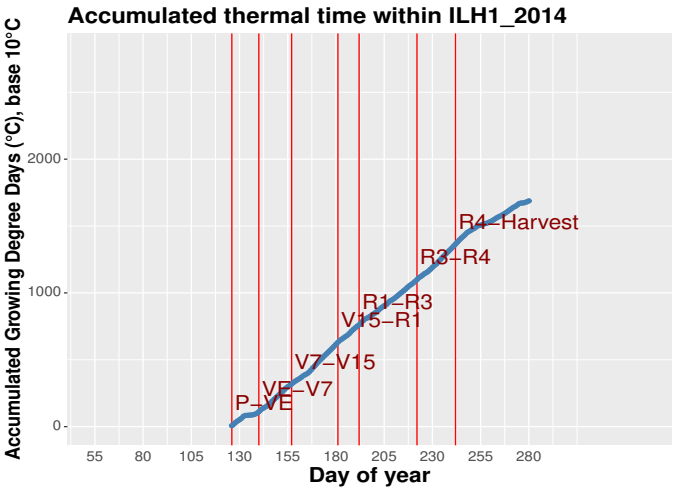

Accumulated thermal time within IAH3_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

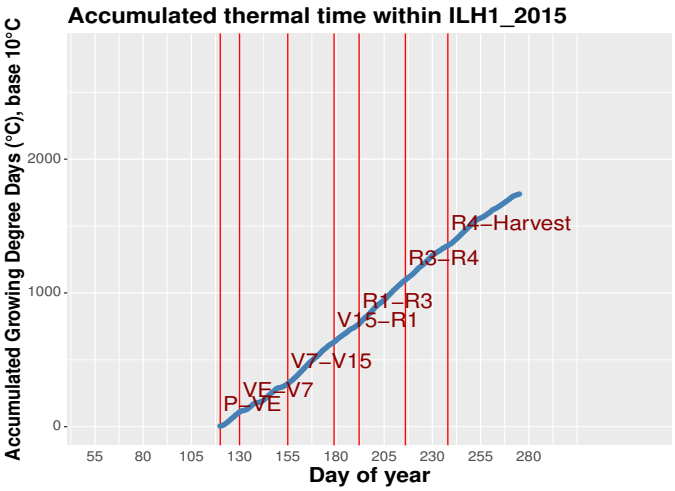

Accumulated thermal time within IAH4_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

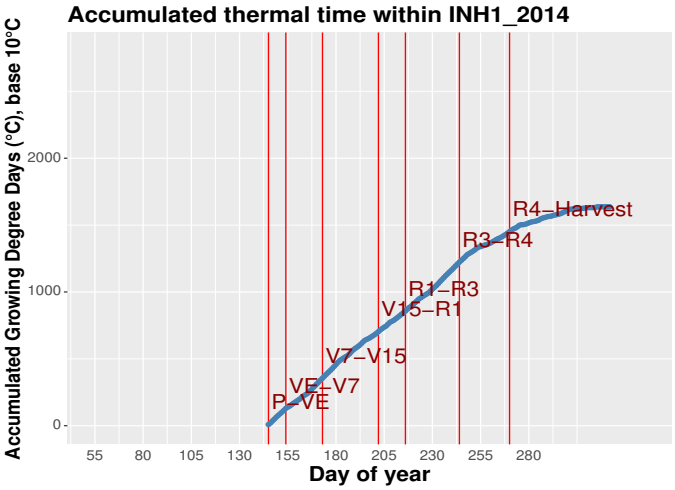

Accumulated thermal time within IAH4_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

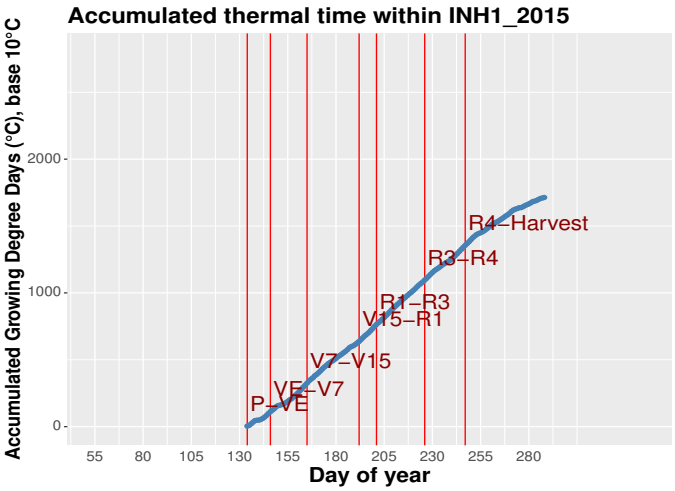

Accumulated thermal time within ILH1.a_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

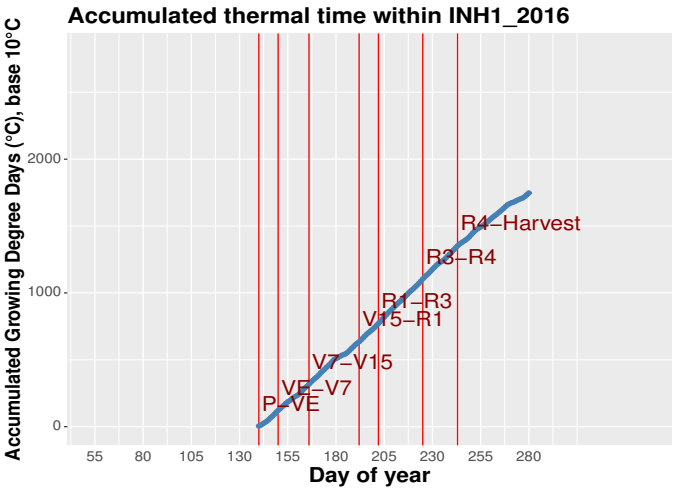

Accumulated thermal time within ILH1.b_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

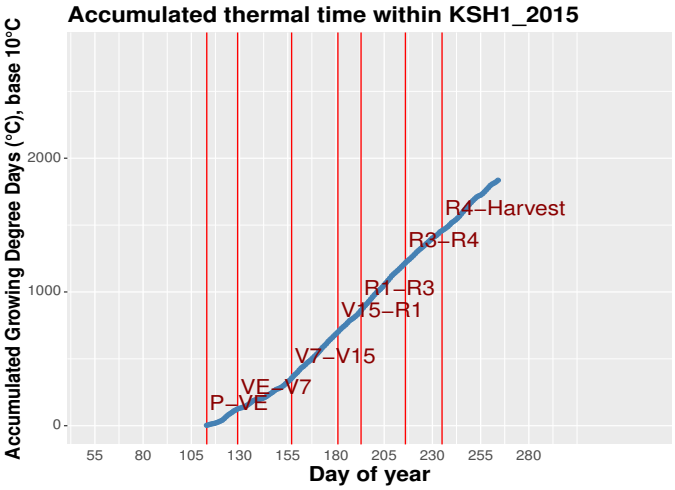

Accumulated thermal time within ILH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

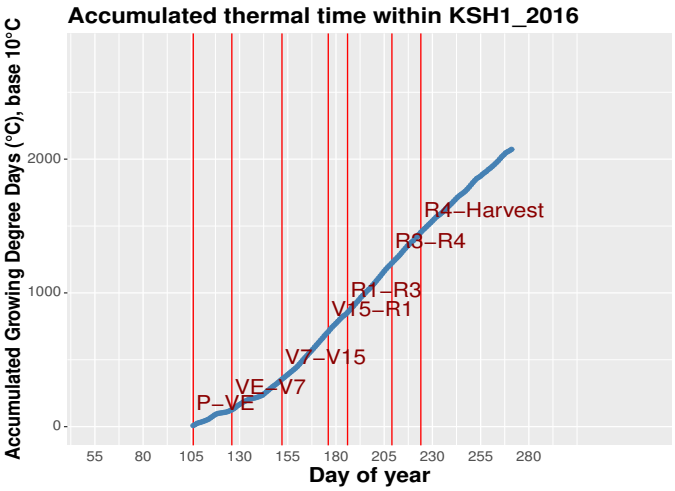

Accumulated thermal time within ILH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

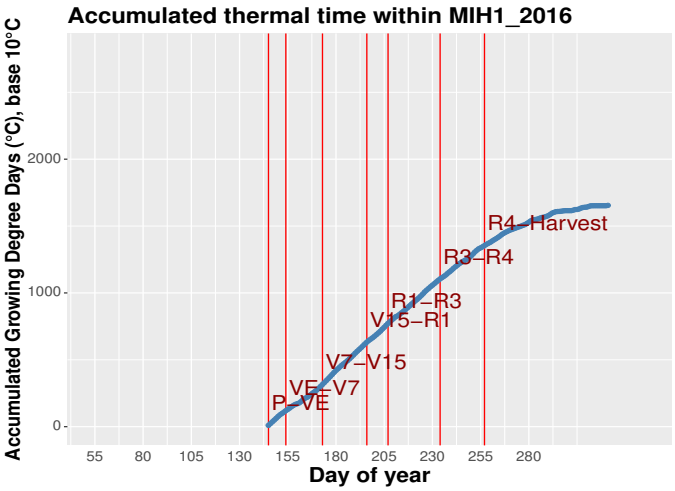

Accumulated thermal time within INH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

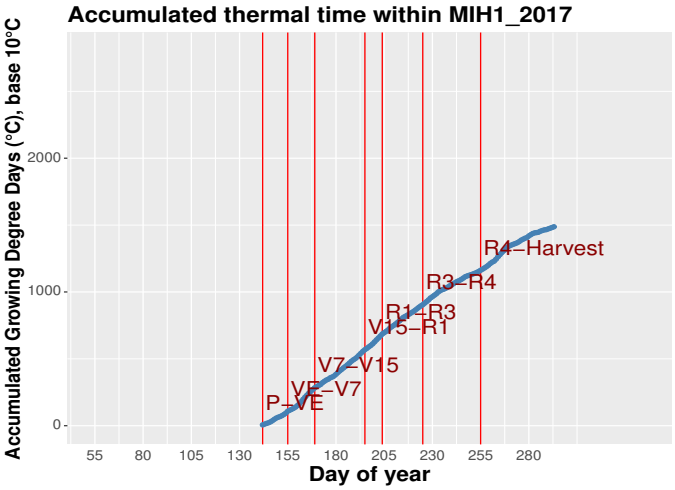

Accumulated thermal time within INH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

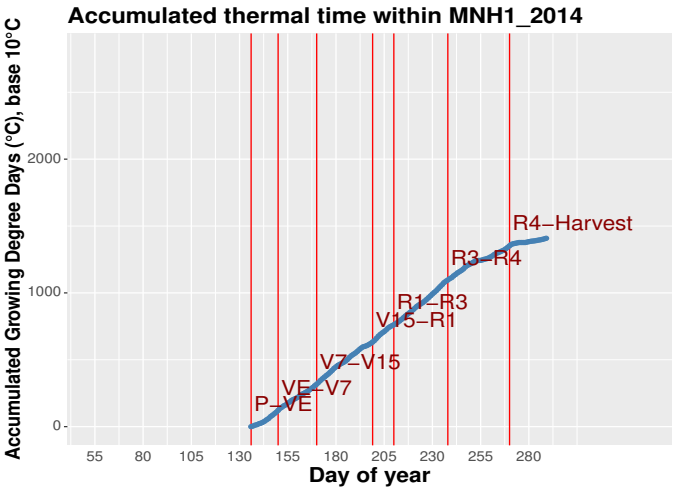

Accumulated thermal time within INH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

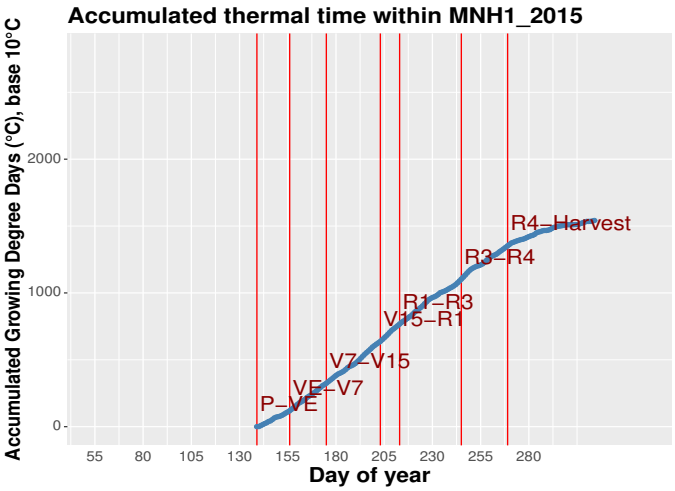

Accumulated thermal time within KSH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

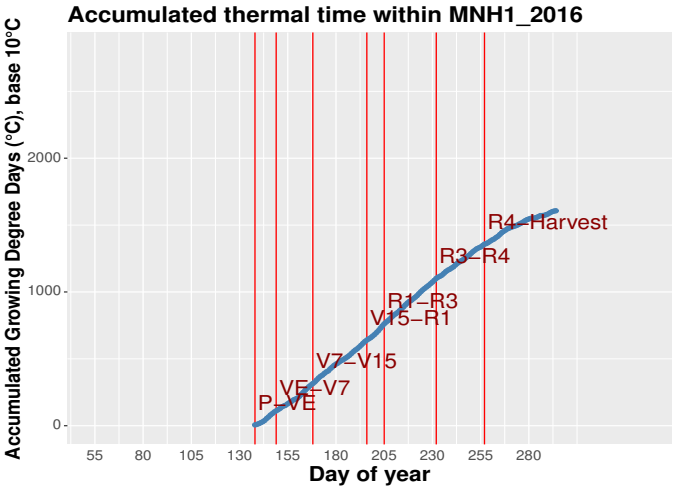

Accumulated thermal time within KSH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

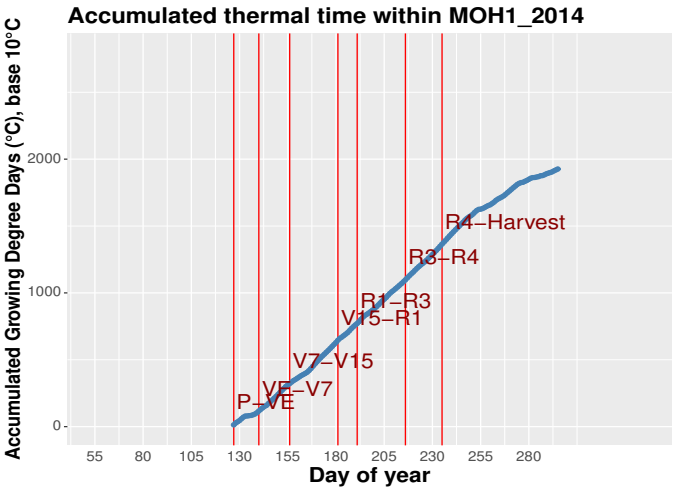

Accumulated thermal time within MIH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

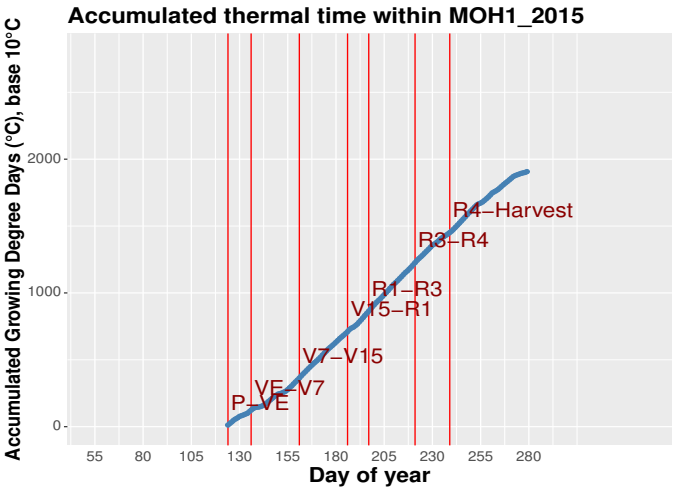

Accumulated thermal time within MIH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

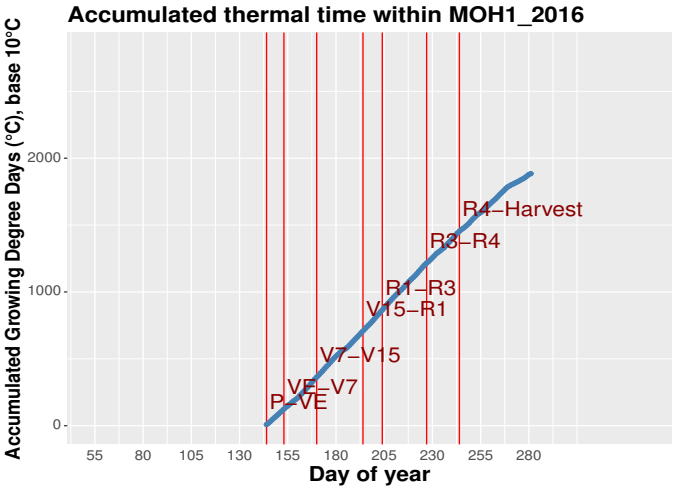

Accumulated thermal time within MNH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

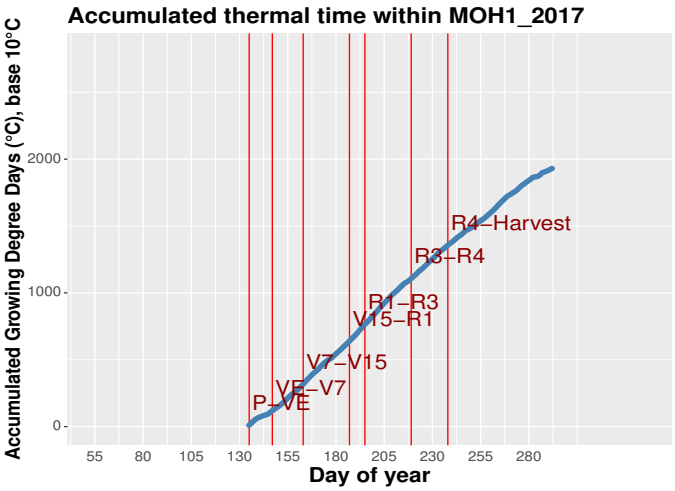

Accumulated thermal time within MNH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

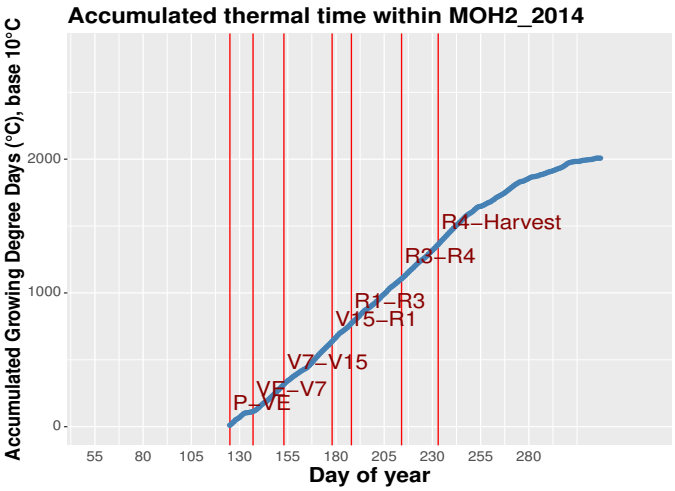

Accumulated thermal time within MNH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

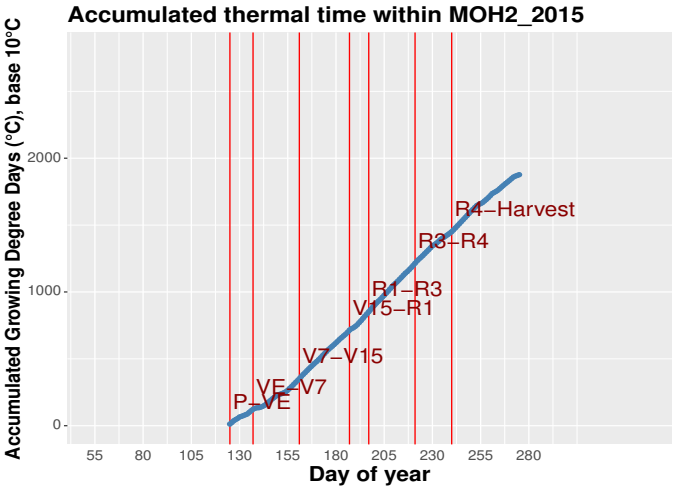

Accumulated thermal time within MOH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

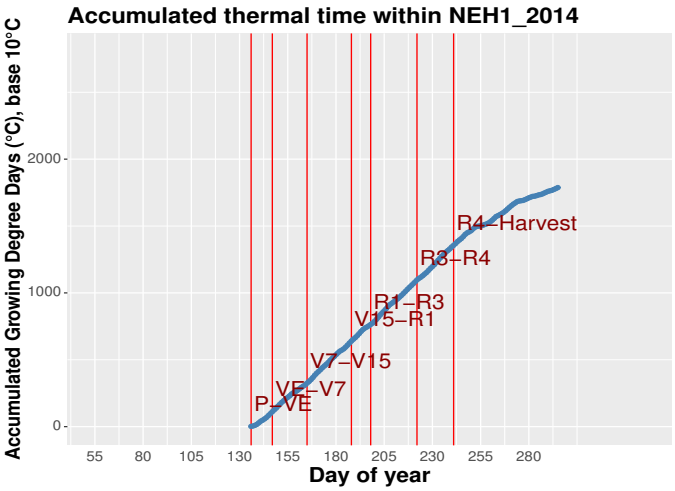

Accumulated thermal time within MOH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

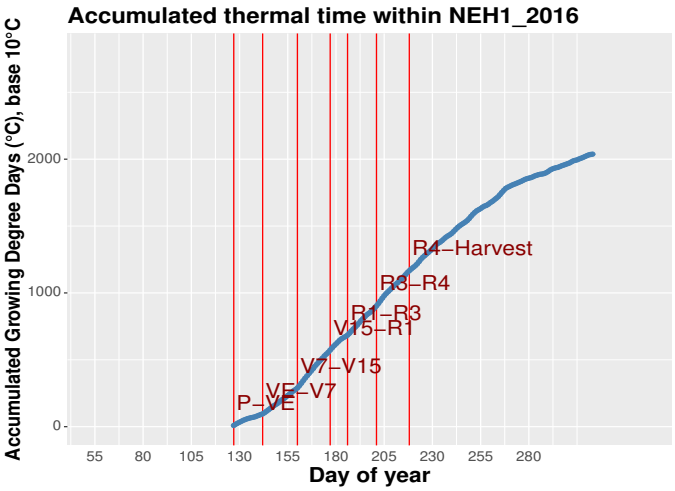

Accumulated thermal time within MOH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

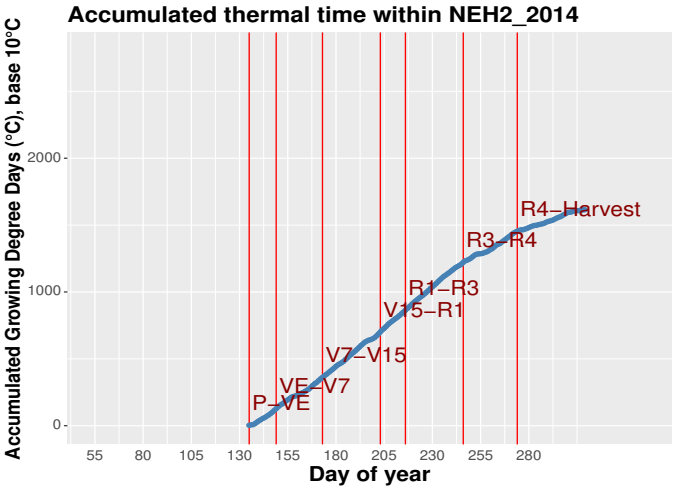

Accumulated thermal time within MOH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

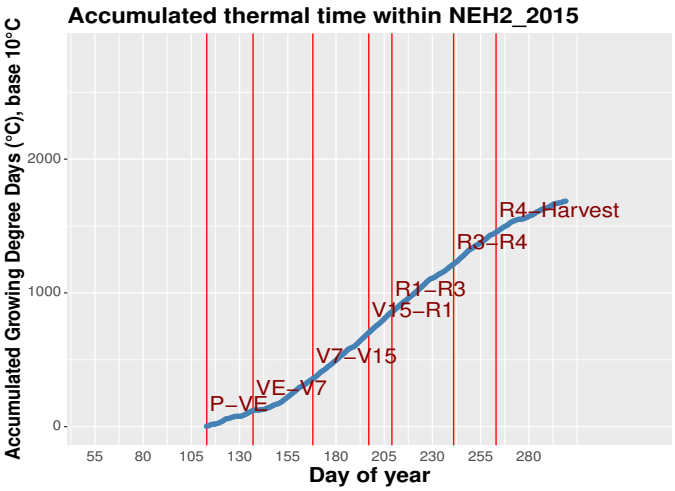

Accumulated thermal time within MOH2_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

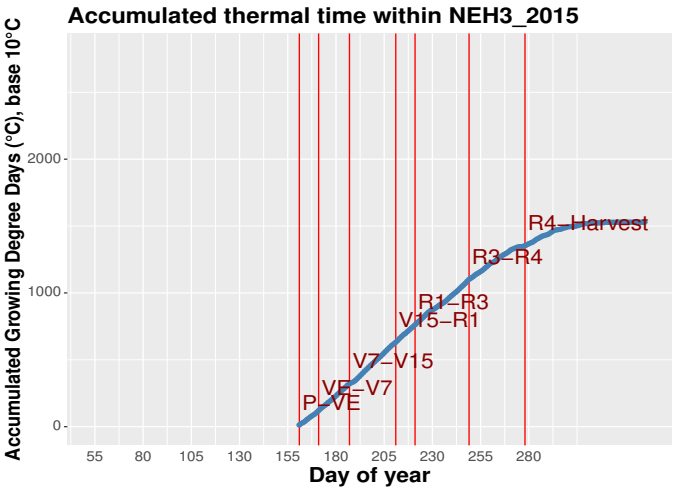

Accumulated thermal time within MOH2_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

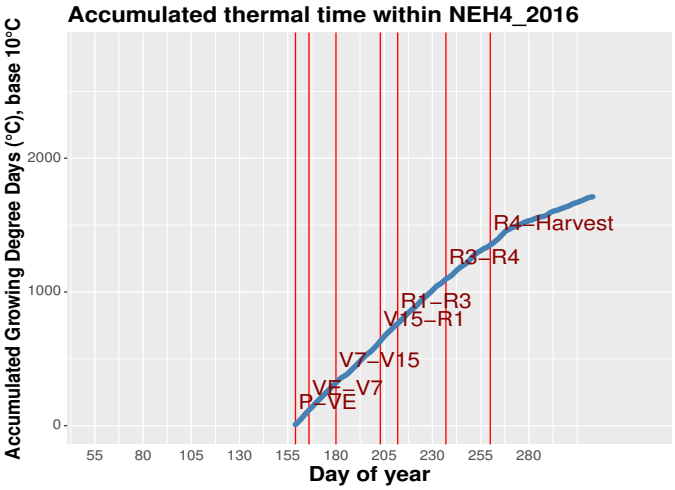

Accumulated thermal time within NEH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

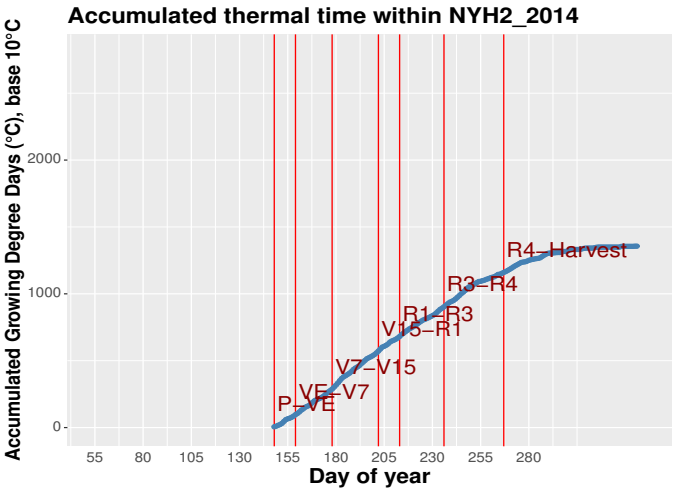

Accumulated thermal time within NEH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

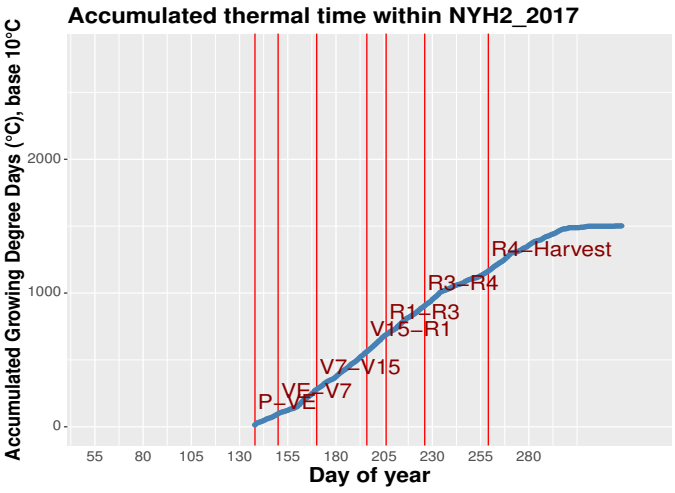

Accumulated thermal time within NEH2_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

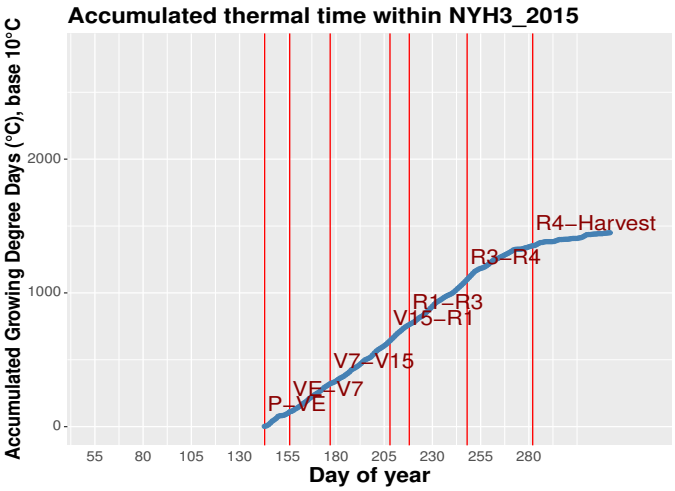

Accumulated thermal time within NEH2_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

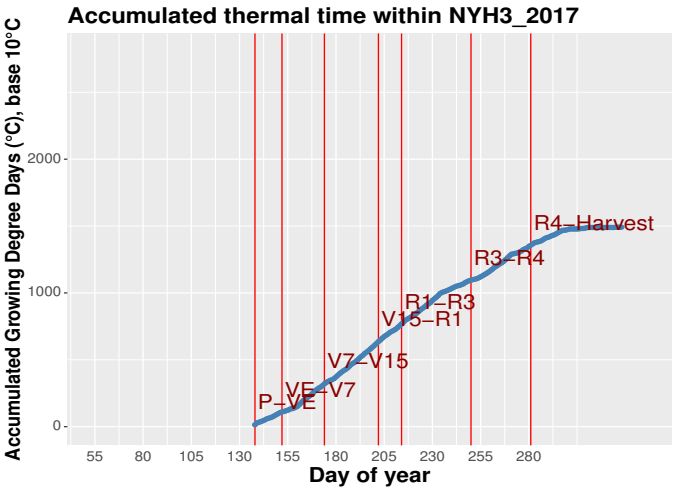

Accumulated thermal time within NEH3_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

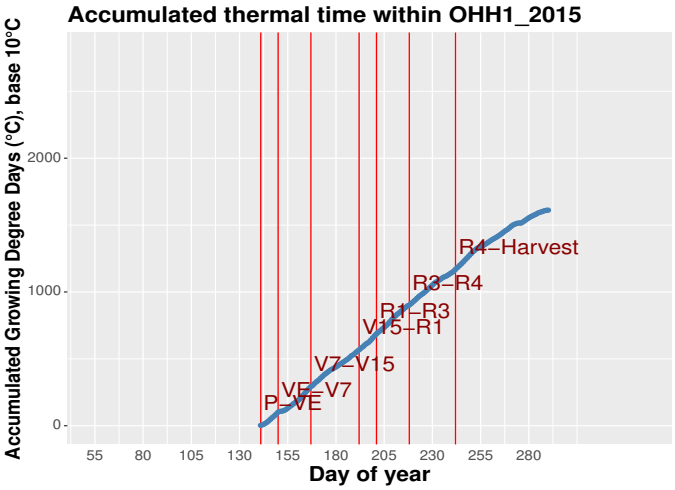

Accumulated thermal time within NEH4_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

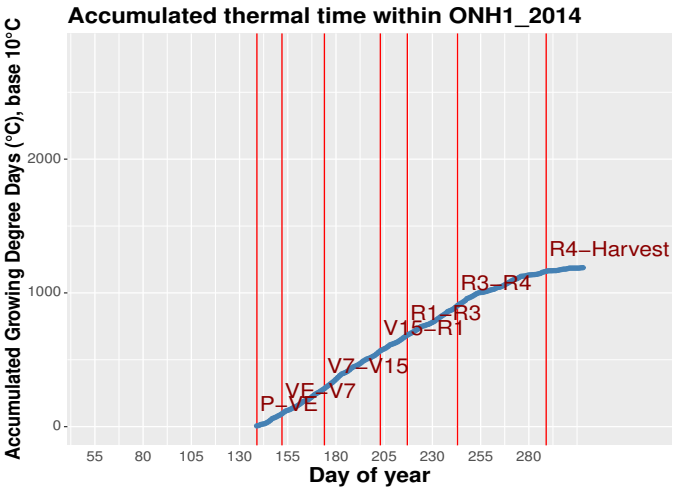

Accumulated thermal time within NYH2_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

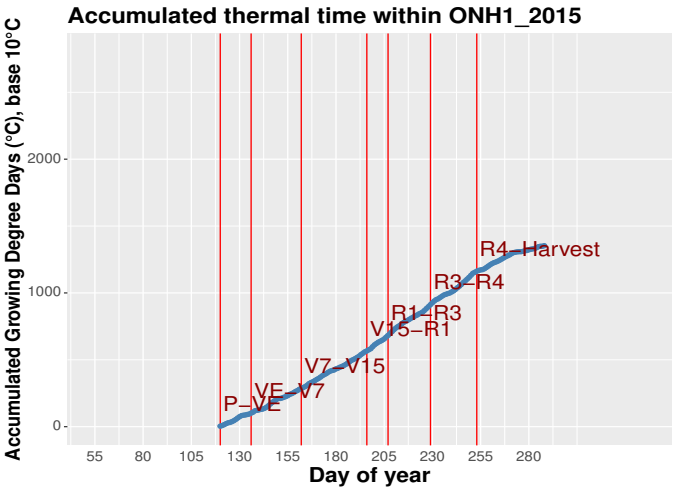

Accumulated thermal time within NYH2_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

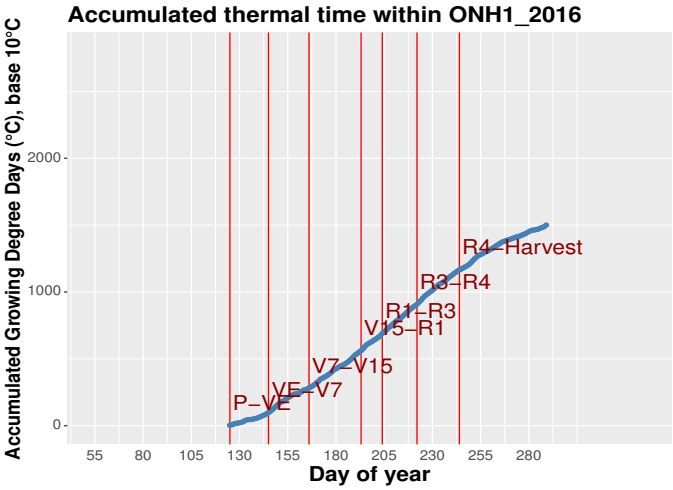

Accumulated thermal time within NYH3_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

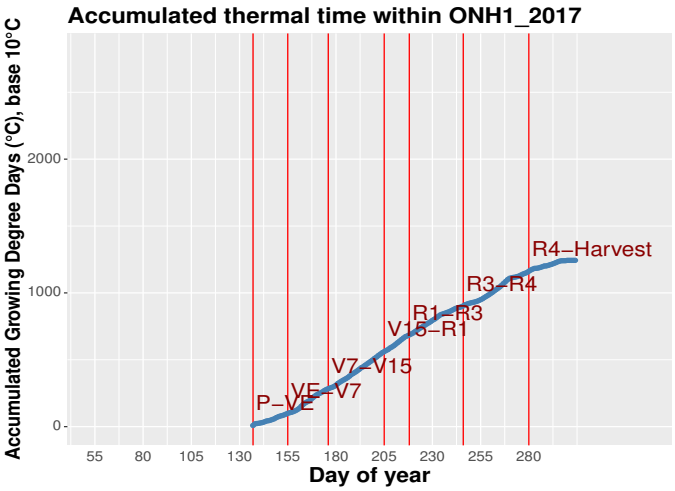

Accumulated thermal time within NYH3_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

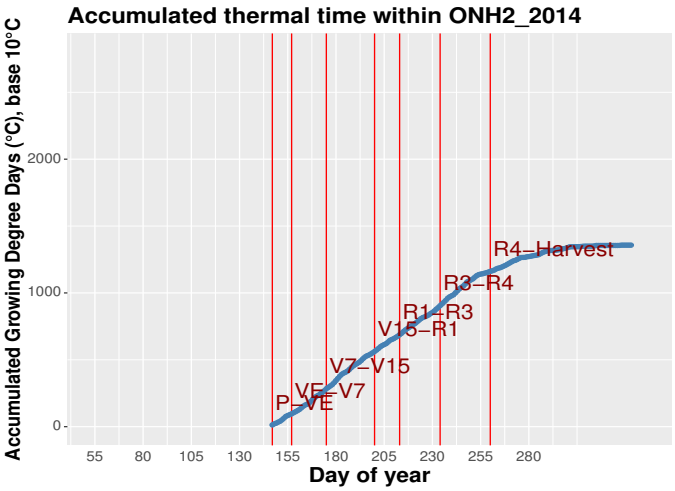

Accumulated thermal time within OHH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

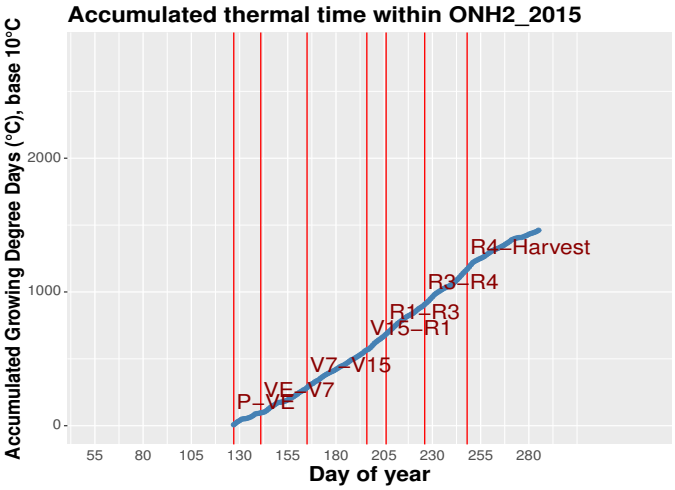

Accumulated thermal time within ONH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

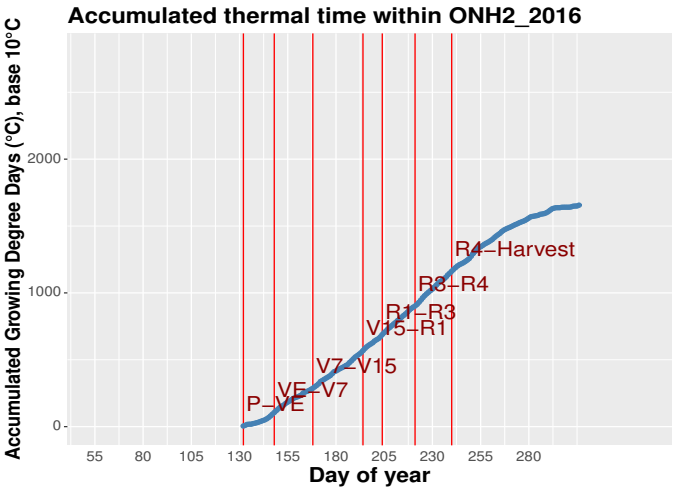

Accumulated thermal time within ONH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

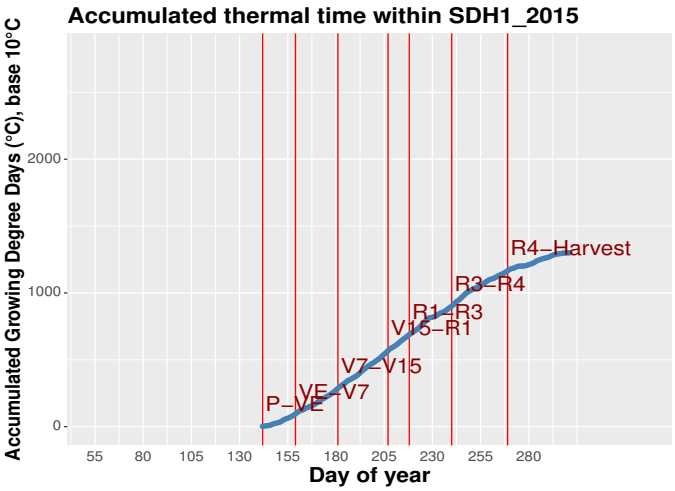

Accumulated thermal time within ONH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

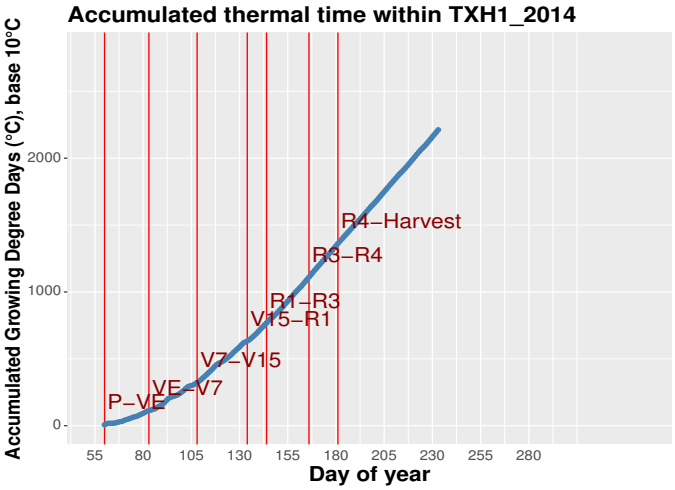

Accumulated thermal time within ONH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

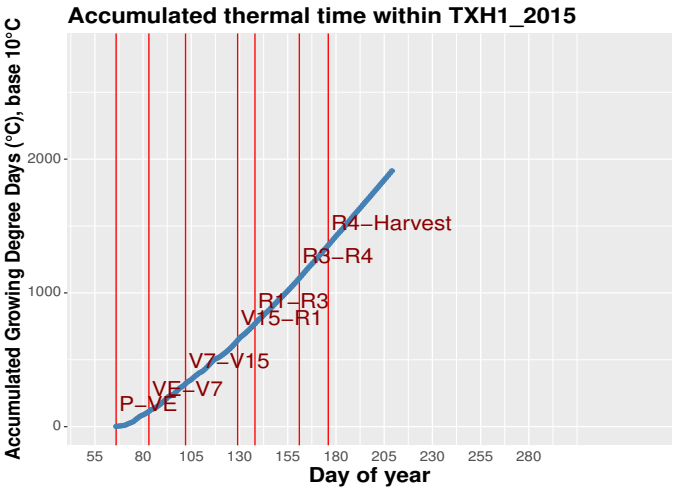

Accumulated thermal time within ONH2_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

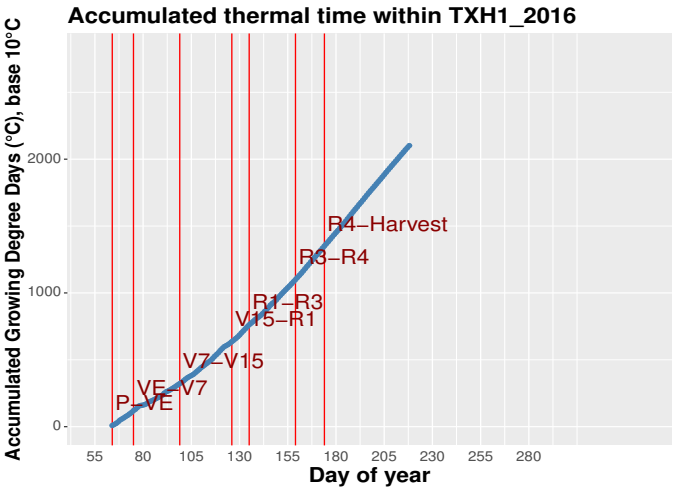

Accumulated thermal time within ONH2_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

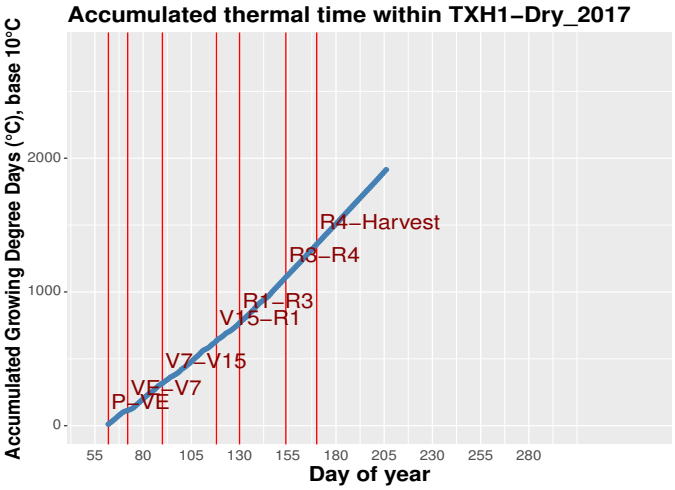

Accumulated thermal time within ONH2_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

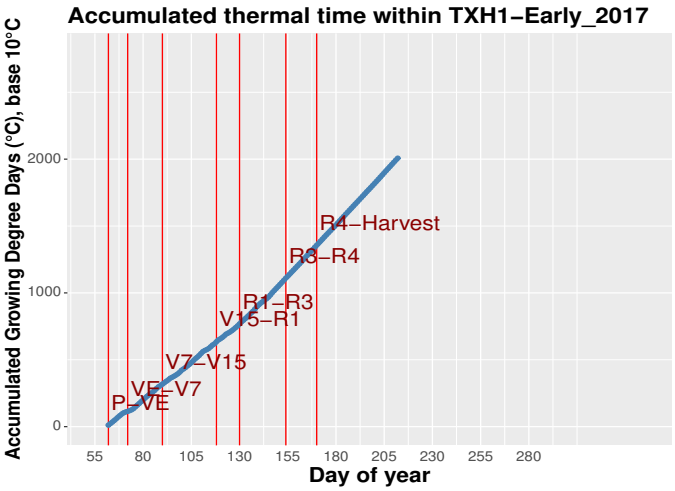

Accumulated thermal time within SDH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

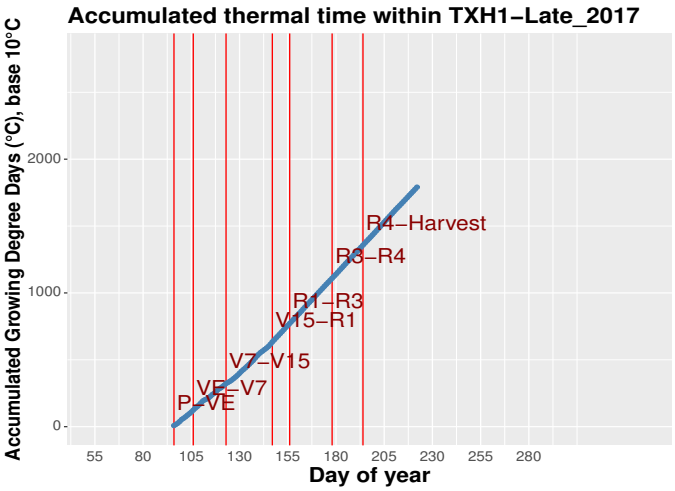

Accumulated thermal time within TXH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest

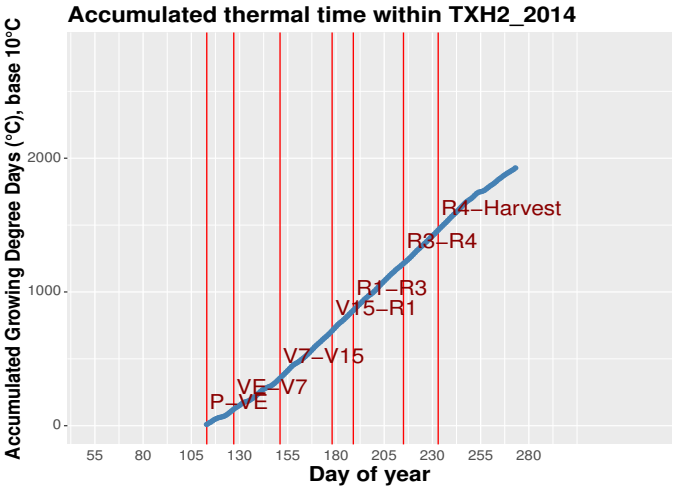

Accumulated thermal time within TXH1_2015
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

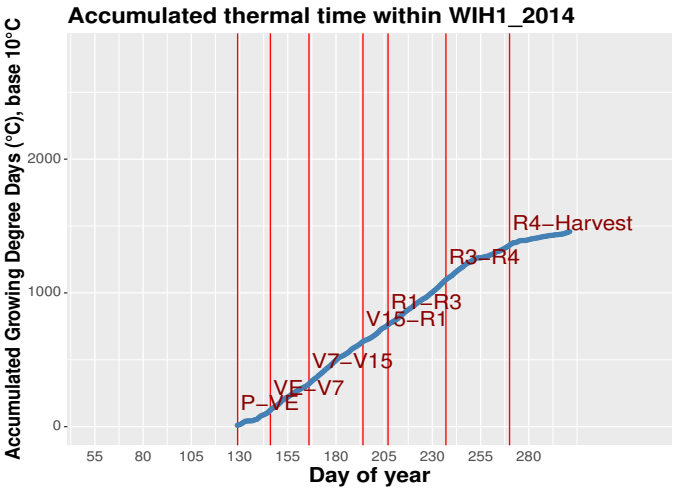

Accumulated thermal time within TXH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

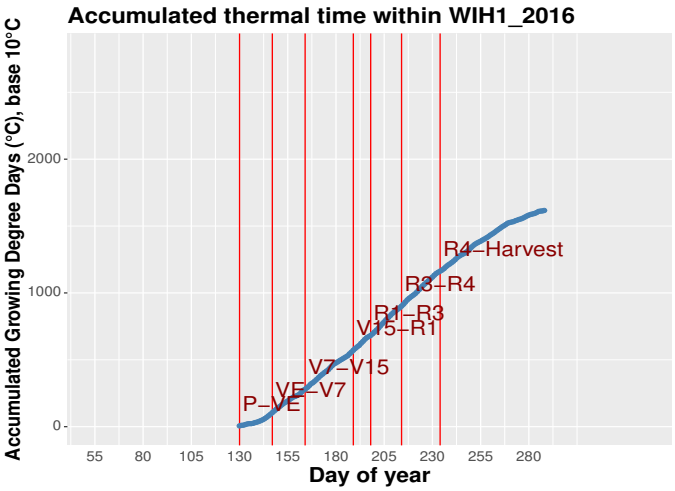

Accumulated thermal time within TXH1–Dry_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE
55
80
105
130
155
180
205
230
255
280
Day of year

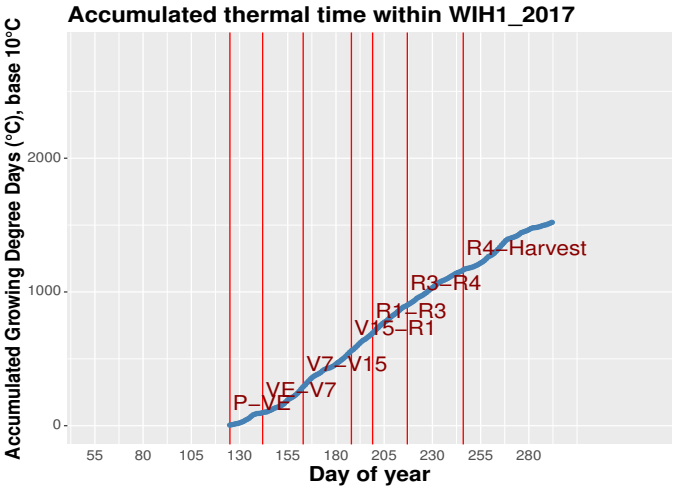

Accumulated thermal time within TXH1–Early_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest
55
80
105
130
155
180
205
230
255
280
Day of year

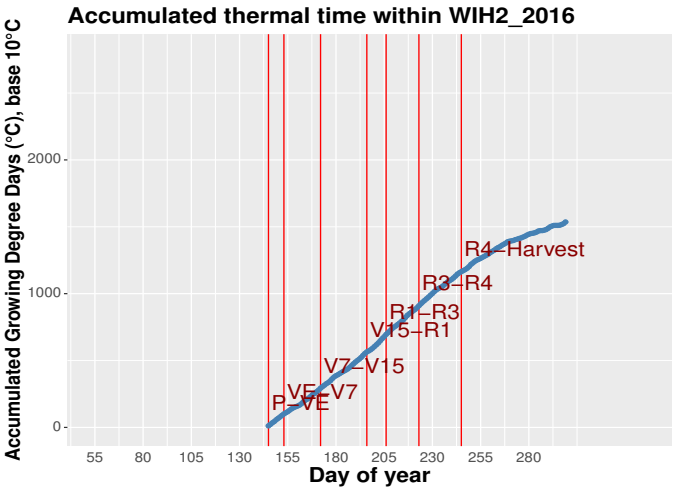

Accumulated thermal time within TXH1–Late_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest
55
80
105
130
155
180
205
230
255
280
Day of year

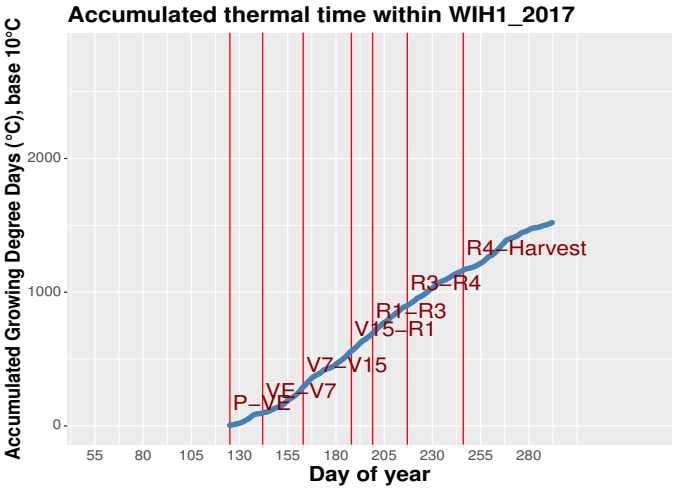

Accumulated thermal time within TXH2_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest
55
80
105
130
155
180
205
230
255
280
Day of year


Accumulated thermal time within WIH1_2014
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest


Accumulated thermal time within WIH1_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest


Accumulated thermal time within WIH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
P–VE
VE–V7
V7–V15
V15–R1
R1–R3
R3–R4
R4–Harvest


Accumulated thermal time within WIH2_2016
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE


Accumulated thermal time within WIH1_2017
Accumulated Growing Degree Days (°C), base 10°C
2000
1000
0
55
80
105
130
155
180
205
230
255
280
Day of year
R4–Harvest
R3–R4
R1–R3
V15–R1
V7–V15
VE–V7
P–VE

4.6 Bibliography

Aach J, Church GM (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17(6):495–508

Aghabozorgi S, Shirkhorshidi AS, Wah TY (2015) Time-series clustering–a decade review. Information Systems 53:16–38

AlKhalifah N, Campbell DA, Falcon CM, Gardiner JM, Miller ND, Romay MC, Walls R, Walton R, Yeh CT, Bohn M, et al. (2018) Maize genomes to fields: 2014 and 2015 field season genotype, phenotype, environment, and inbred ear image datasets. BMC Research Notes 11(1):1–5

Basnet BR, Crossa J, Dreisigacker S, Pérez-Rodríguez P, Manes Y, Singh RP, Rosyara U, Camarillo-Castillo F, Murua M (2019) Hybrid wheat prediction using genomic, pedigree, and environmental covariables interaction models

Bauer A, Fanning C, Enz JW, Eberlein C (1984) Use of growing-degree days to determine spring wheat growth stages. Extension bulletin-North Dakota State University of Agriculture and Applied Science, Cooperative Extension Service (USA)

Berndt DJ, Clifford J (1994) Using dynamic time warping to find patterns in time series. In: KDD workshop, Seattle, WA, USA:, vol 10, pp 359–370

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science 52(2):707–719

Chu S, Keogh E, Hart D, Pazzani M (2002) Iterative deepening dynamic time warping for time series. In: Proceedings of the 2002 SIAM International Conference on Data Mining, SIAM, pp 195–212

Costa-Neto G, Fritsche-Neto R, Crossa J (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106

Csillik O, Belgiu M, Asner GP, Kelly M (2019) Object-based time-constrained dynamic time warping classification of crops using sentinel-2. Remote sensing 11(10):1257

Delerce S, Dorado H, Grillon A, Rebolledo MC, Prager SD, Patiño VH, Garcés Varón G, Jiménez D (2016) Assessing weather-yield relationships in rice at local scale using data mining approaches. PloS one 11(8):e0161,620

Dong X, Guan L, Zhang P, Liu X, Li S, Fu Z, Tang L, Qi Z, Qiu Z, Jin C, et al. (2021) Responses of maize with different growth periods to heat stress around flowering and early grain filling. Agricultural and Forest Meteorology 303:108,378

Felipe de Mendiburu, Muhammad Yaseen (2020) agricolae: Statistical Procedures for Agricultural Research. R package version 1.4.0

Ferrão LFV, Marinho CD, Munoz PR, Resende Jr MF (2020) Improvement of predictive ability in maize hybrids by including dominance effects and marker× environment models. Crop Science 60(2):666–677

Fu Tc, Chung Fl, Ng V, Luk R (2001) Pattern discovery from stock time series using self-organizing maps. In: Workshop Notes of KDD2001 Workshop on Temporal Data Mining, Citeseer, vol 1

Galili T (2015) dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics DOI 10.1093/bioinformatics/btv428, URL https://academic.oup.com/bioinformatics/article/31/22/3718/240978/dendextend-an-R-package-for-visualizing-adjusting, https://academic.oup.com/bioinformatics/article-pdf/31/22/3718/17122682/btv428.pdf

Granato I, Cuevas J, Luna-Vázquez F, Crossa J, Montesinos-López O, Burgueño J, Fritsche-Neto R (2018) Bgge: a new package for genomic-enabled prediction incorporating genotype× environment interaction models. G3: Genes, Genomes, Genetics 8(9):3039–3047

Griffing B (1962) Prediction formulae for general combining ability selection methods utilizing one or two random-mating populations. Australian Journal of Biological Sciences 15(4):650–665

Guan X, Huang C, Liu G, Meng X, Liu Q (2016) Mapping rice cropping systems in vietnam using an ndvi-based time-series similarity measurement based on dtw distance. Remote Sensing 8(1):19

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics, Springer, New York, NY

Hawkins E, Fricker TE, Challinor AJ, Ferro CA, Ho CK, Osborne TM (2013) Increasing influence of heat stress on french maize yields from the 1960s to the 2030s. Global change biology 19(3):937–947

Heslot N, Jannink JL, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Science 53(3):921–933

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics 127(2):463–480

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics 127(3):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 x environment interactions in kansas wheat

Jarquin D, De Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, et al. (2021a) Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Frontiers in genetics p 1819

Jarquin D, de Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, Graham C, Hirsch CN, Holland JB, Hooker D, Kaeppler SM, Knoll J, Lee EC, Lawrence-Dill CJ, Lynch JP, Moose SP, Murray SC, Nelson R, Rocheford T, Schnable JC, Schnable PS, Smith M, Springer N, Thomison P, Tuinstra M, Wisser RJ, Xu W, Yu J, Lorenz A (2021b) 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, DOI 10.3389/fgene.2020.592769, URL https://www.frontiersin.org/article/10.3389/fgene.2020.592769

Kate RJ (2016) Using dynamic time warping distances as features for improved time series classification. Data Mining and Knowledge Discovery 30(2):283–312

Kazan K, Lyons R (2016) The link between flowering time and stress tolerance. Journal of experimental botany 67(1):47–60

Keogh EJ, Pazzani MJ (1998) An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. In: Kdd, vol 98, pp 239–243

Keogh EJ, Pazzani MJ (2000) Scaling up dynamic time warping for datamining applications. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp 285–289

Khaki S, Wang L (2019) Crop yield prediction using deep neural networks. Frontiers in Plant Science 10, DOI 10.3389/fpls.2019.00621, URL https://www.frontiersin.org/article/10.3389/fpls.2019.00621

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the köppen-geiger climate classification updated

Lhermitte S, Verbesselt J, Verstraeten WW, Coppin P (2011) A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote sensing of environment 115(12):3129–3152

Li X, Guo T, Wang J, Bekele WA, Sukumaran S, Vanous AE, McNellie JP, Cortes LT, Lopes MS, Lamkey KR, et al. (2021) An integrated framework reinstating the environmental dimension for gwas and genomic selection in crops. Molecular Plant 14(6):874–887

Lopes M, Dreisigacker S, Peña R, Sukumaran S, Reynolds MP (2015) Genetic characterization of the wheat association mapping initiative (wami) panel for dissection of complex traits in spring wheat. Theoretical and applied genetics 128(3):453–464

de Los Campos G, Pérez-Rodríguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nature communications 11(1):1–10

Ly D, Chenu K, Gauffreteau A, Rincent R, Huet S, Gouache D, Martre P, Bordes J, Charmet G (2017) Nitrogen nutrition index predicted by a crop model improves the genomic prediction of grain number for a bread wheat core collection. Field Crops Research 214:331–340

Ly D, Huet S, Gauffreteau A, Rincent R, Touzy G, Mini A, Jannink JL, Cormier F, Paux E, Lafarge S, et al. (2018) Whole-genome prediction of reaction norms to environmental stress in bread wheat (triticum aestivum l.) by genomic random regression. Field Crops Research 216:32–41

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, 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(1):1–6

Medina S, Vicente R, Nieto-Taladriz MT, Aparicio N, Chairi F, Vergara-Diaz O, Araus JL (2019) The plant-transpiration response to vapor pressure deficit (vpd) in durum wheat is associated with differential yield performance and specific expression of genes involved in primary metabolism and water transport. Frontiers in plant science 9:1994

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019) Genomic prediction of maize yield across european environmental conditions. Nature genetics 51(6):952–956

Muñoz PR, Resende Jr MF, Gezan SA, Resende MDV, de Los Campos G, Kirst M, Huber D, Peter GF (2014) Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198(4):1759–1768

Netzel P, Stepinski T (2016) On using a clustering approach for global climate classification. Journal of Climate 29(9):3387–3401

Oates T, Schmill MD, Cohen PR (2000) A method for clustering the experiences of a mobile robot that accords with human judgments. In: AAAI/IAAI, pp 846–851

Pérez P, de Los Campos G (2014) Genome-wide regression and prediction with the bglr statistical package. Genetics 198(2):483–495

Petitjean F, Inglada J, Gançarski P (2012) Satellite image time series analysis under time warping. IEEE transactions on geoscience and remote sensing 50(8):3081–3095

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Rincent R, Kuhn E, Monod H, Oury FX, Rousset M, Allard V, Le Gouis J (2017) Optimization of multi-environment trials for genomic selection based on crop models. Theoretical and Applied Genetics 130(8):1735–1752

Rogers AR, Holland JB (2021) Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. G3 Genes|Genomes|Genetics DOI 10.1093/g3journal/jkab440, URL https://doi.org/10.1093/g3journal/jkab440, jkab440, https://academic.oup.com/g3journal/advance-article-pdf/doi/10.1093/g3journal/jkab440/42350740/jkab440.pdf

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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(2):jkaa050

Rosenberg A, Hirschberg J (2007) V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pp 410–420

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics 20:53–65

Rubel F, Brugger K, Haslinger K, Auer I, et al. (2017) The climate of the european alps: Shift of very high resolution köppen-geiger climate zones 1800–2100. Meteorologische Zeitschrift 26(2):115–125

Rötter RP, Tao F, Höhn JG, Palosuo T (2015) Use of crop simulation modelling to aid ideotype design of future cereal cultivars. Journal of Experimental Botany 66(12):3463–3476, DOI 10.1093/jxb/erv098, URL https://doi.org/10.1093/jxb/erv098, https://academic.oup.com/jxb/article-pdf/66/12/3463/17129779/erv098.pdf

Sadras V, Lake L, Chenu K, McMurray L, Leonforte A (2012) Water and thermal regimes for field pea in australia and their implications for breeding. Crop and Pasture Science 63(1):33–44

Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing 26(1):43–49

Sardá-Espinosa A (2017) Comparing time-series clustering algorithms in r using the dtwclust package. R package vignette 12:41

Semenov M, Stratonovitch P, Alghabari F, Gooding M (2014) Adapting wheat in europe for climate change. Journal of cereal science 59(3):245–256

Sinclair TR, Hammer GL, Van Oosterom EJ (2005) Potential yield and water-use efficiency benefits in sorghum from limited maximum transpiration rate. Functional Plant Biology 32(10):945–952

Sparks AH (2018) nasapower: a nasa power global meteorology, surface solar energy and climatology data client for r. Journal of Open Source Software 3(30):1035

Sukumaran S, Lopes MS, Dreisigacker S, Dixon LE, Zikhali M, Griffiths S, Zheng B, Chapman S, Reynolds MP (2016) Identification of earliness per se flowering time locus in spring wheat through a genome-wide association study. Crop Science 56(6):2962–2672

Sukumaran S, Crossa J, Jarquin D, Lopes M, Reynolds MP (2017) Genomic prediction with pedigree and genotype× environment interaction in spring wheat grown in south and west asia, north africa, and mexico. G3: Genes, Genomes, Genetics 7(2):481–495

Swan J, Schneider E, Moncrief J, Paulson W, Peterson A (1987) Estimating corn growth, yield, and grain moisture from air growing degree days and residue cover1. Agronomy Journal 79(1):53–60

Touzy G, Rincent R, Bogard M, Lafarge S, Dubreuil P, Mini A, Deswarte JC, Beauchêne K, Le Gouis J, Praud S (2019) Using environmental clustering to identify specific drought tolerance qtls in bread wheat (t. aestivum l.). Theoretical and Applied Genetics 132(10):2859–2880

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414–4423

Vitezica ZG, Varona L, Legarra A (2013) On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. Genetics 195(4):1223–1230, DOI 10.1534/genetics.113.155176, URL https://doi.org/10.1534/genetics.113.155176, https://academic.oup.com/genetics/article-pdf/195/4/1223/42116782/genetics1223.pdf

Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. Journal of the American statistical association 58(301):236–244

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES (2021) Predicting phenotypes from genetic, environment, management, and historical data using CNNs. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 134(12):3997–4011, DOI 10.1007/s00122-021-03943-7, place: Germany

Westhues CC, Mahone GS, Thorwarth P, Schmidt M, Richter J, Simianer H, Beissinger T, et al. (2021a) Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. Frontiers in plant science 12:699,589–699,589

Westhues CC, Simianer H, Beissinger TM (2021b) learnmet: an r package to apply machine learning methods for genomic prediction using multi-environment trial data. bioRxiv

Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A, et al. (2017) Omics-based hybrid prediction in maize. Theoretical and applied genetics 130(9):1927–1939

Wismüller A, Lange O, Dersch DR, Leinsinger GL, Hahn K, Pütz B, Auer D (2002) Cluster analysis of biomedical image time-series. International Journal of Computer Vision 46(2):103–128

Yang Z, Sinclair TR, Zhu M, Messina CD, Cooper M, Hammer GL (2012) Temperature effect on transpiration response of maize plants to vapour pressure deficit. Environmental and Experimental Botany 78:157–162

Zhao Y, Potgieter AB, Zhang M, Wu B, Hammer GL (2020) Predicting wheat yield at the field scale by combining high-resolution sentinel-2 satellite imagery and crop modelling. Remote Sensing 12(6):1024

5. General Discussion

Plant phenotypes result from the combined effects of the genetic makeup of the plant and of the plant's neighbouring environment. The first main focus of this work relate to the incorporation of quantitative environmental data into routinely used genomic prediction (GP) models for analysis of multi-environment trials. The second important facet concerns the application of machine learning model interpretation methods to understand the output of machine learning (ML) models that are trained with genomic and environmental features. In **chapter 2**, different models were evaluated for their ability to predict single-cross maize hybrids, namely linear reaction norms, that belong to the classical toolbox of plant breeders, and gradient boosting frameworks, a ML ensemble method. Secondly, an R package was developed in **chapter 3**, which allows to assess different types of ML prediction methods in various cross-validation (CV) schemes that mimic practical situations faced by plant breeders, and also enables to get some insights into the importance of the predictors. In **chapter 4**, we explored avenues for clustering and predicting phenotypes with a similarity matrix derived from pairwise dynamic time warping distances calculated among growing environments. Our results show that high predictive abilities are mainly related to the incorporation of interaction terms, to the amount of genetic and environmental relatedness between training (TRN) and test sets (TST) and to the method used to summarize the weather data. The following discussion is devoted to important questions arising from our studies, to lessons learned from processing and analyzing environmental data, and to additional topics to investigate.

5.1 Merits of modeling genotype-by-environment, genotype-by-year and genotype-by-location interactions in multi-environment trials

In **chapter 2** and **chapter 4**, we explored different reaction norm models that represent the environment (E) using a single factor corresponding to the year-location combination, as well as by using additional categorical factors separating the effects of year (Y) and location (L). Modelling environments as independent outcomes allows to capture potential trial-specific effects, related for instance to management and crop cultivation practices. However, this approach does not allow to

borrow information among environments, while useful information can be retrieved by specifying in which location and in which year phenotypic observations have been recorded. Modelling site effect appears especially relevant in the G2F dataset, as we expected that the environmental variation experienced by the maize hybrids in the Genomes to Fields Initiative could be largely explained by the localization of field trials, that belonged to various climatic zones, as reported by previous studies (Jarquin et al., 2021; McFarland et al., 2020; Rogers et al., 2021). Our results were coherent with these studies, as we found a larger value of the variance component term for location effect compared to year effect (**chapter 4**), and also a better performance of models that incorporated site effects along marker effects (G+S), than models considering instead only year effect (G+Y) (**chapter 2**).

While aggregated multi-year and multi-location data are beneficial to augment the size of the training data, appropriate modeling of genotype-by-year (GY) effects is essential to separate the main genotypic effects from unpredictable year-to-year weather variations (Bernal-Vasquez et al., 2017; Dias et al., 2020). It is frequently the case that hybrid breeding programs aim at general adaptation and develop varieties capable of maintaining good performance across different environmental conditions. Hence, values for variance components involving years are generally much larger than those involving locations, as observed by Dias et al. (2020) with a Brazilian hybrid maize program. Bernal-Vasquez et al. (2017) demonstrated the interest of performing year-wise analyses of breeding cycles, and to use these directly in GP models to predict in a new year, together with a kinship component estimated from molecular data for modeling G × Y effects, similar to our approach in **chapter 2**. This approach yielded better predictive abilities than models that estimated the year effect only based on common genotype checks used across years (e.g. using the ID of the genotype rather than genomic similarity) that ignored G × Y interaction effects. As noted by Bernal-Vasquez et al. (2017), in hybrid breeding programs, first years of evaluation of general combining ability (GCA) of potential hybrids are generally completely disconnected across years, i.e. no common genotypes are tested across years. Dissecting genotype main effect from G × Y can therefore only be achieved with marker data in such cases and is essential to make predictions that account for the respective effect of previous years on genotype performance. In the Genomes to Fields Initiative, a set of 30 to 50 common check hybrids were used in all locations, as well as regional hybrid checks for the Northern, Central and Southern North-American regions, ensuring that the trials were well connected within a year. Regarding the connectivity across years, 627, 50 and 29 hybrids were tested in exactly 2, 3 and 4 years considering the dataset that we used after data cleaning and pre-processing. Including genotype-by-environment, genotype-by-year and genotype-by-location interaction terms in GP models, as we did in **chapter 2** and **chapter 4** was relevant to improve predictive abilities, as we showed that the best reaction norm model without weather data corresponded to the full model G + E + S + Y + G×S + G×Y + G×E. It should be noted that, when the objective is to predict untested genotypes in already tested environments

(e.g. CV1 or CV2 in **chapter 4**), this latter model performed better or very similar to models with environmental data. Other variance-covariance structures could be explored to model heterogeneous error variance, such as factor-analytical model (Buntaran et al., 2019), especially when the estimated variance components for location and year effects are large.

We also expected that the weather and soil covariates we derived could not capture the complete variation due to the influence of the environment, and for this reason added back the G × E interaction term in the reaction norm models that already incorporated G × W. This assumption appeared fair, as we observed an improvement of the predictive ability for grain yield with the model G+E+W+G×W+G×E, compared to the model G+E+W+G×W, of up to 9% in the CV0-year prediction scheme (**chapter 2**, Supplemental Table S6-S7). We also found an advantage of incorporating the geographical coordinates and the year dummy variables in the gradient boosted decision trees (GBDT) frameworks, as the LightGBM and XGBoost models G+W+Y+Lon+Lat yielded better average predictive abilities than the models G+W, that include solely weather and soil predictor variables (**chapter 2**, Supplemental Table S6-S7). Although these variables have no direct mechanistic effect on yield, other studies also used these as predictors for crop yield prediction (Crane-Droesch, 2018; Guo et al., 2021; Heslot et al., 2014; Huntington et al., 2020; Tiezzi et al., 2017), because they can potentially serve as proxies for numeric variables that are otherwise not included in the models.

In our studies, we more specifically aimed to leverage environmental data, that can be learned from past data. On the other hand, the year effect is *per se* unpredictable, and using it or the location factor in prediction models prevents the possibility of evaluating *in silico* genotype performance under various quantitative pedoclimatic scenarios. In contrast to the above-mentioned studies (Bernal-Vasquez et al., 2017; Dias et al., 2020), our objective was to predict phenotypes, rather than genomic estimated breeding values (i.e. that focus only on additive genetic effects), thus, accounting for both environmental and genomic effects with quantitative data was necessary. Such an approach also allows to envisage development of cultivars adapted for local environmental conditions, rather than global adaptation.

5.2 The applicability and the issues related to the use of quantitative environmental data in prediction models

5.2.1 Attempts at reducing sources of errors by using spatio-temporal interpolation methods for weather data in the G2F dataset

In **chapter 2**, after quality control of the weather data acquired by in-field automatic weather stations, missing or likely erroneous values were flagged within the daily weather dataset. About 20.2% of daily minimum and maximum temperature records and 22.8% of the daily total precipi-

tation records, for the 71 environments used in our analyses, did not pass our quality control and were flagged, thus assigned as missing data. Details of the QC are provided in **chapter 2** (Supplemental Table S2.3), and a pipeline to flag potential erroneous weather values was implemented in *learnMET* (**chapter 3**), when users provide in-field weather data. Subsequently to this quality check, we checked whether the difference between weather records of in-field weather station, and of the closest weather station from the Global Historical Climatology Network (GHCN) was substantial, and if this was the case, the total daily weather records of the given environment were considered as missing, and interpolated as described below.

Whenever a station was found to be very close to the field experiment (less than 2 km) and with reliable and complete data for the growing season, these data were directly used to replace data of the in-field weather station. For many year-location combinations though, it was not possible to find a station at a such high spatial resolution. In these cases, interpolation of flagged daily weather data, based on information extracted from weather stations located in a radius of 70 km from the field trial, was achieved in order to replace missing daily weather data. In addition, the use of multiple weather stations can generate better interpolation data, since single-station data can also exhibit mistakes or missing data.

The fundamental principle of spatio-temporal interpolation originally relates to Waldo Tobler's first law of geography (1969): "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970), that has driven the foundation of kriging and its many variants in spatial statistics. Spatio-temporal kriging is a geostatistical interpolation technique that utilizes the statistical properties of the sample points, both based on geographic and on temporal distances. While adding observations taken at other time points is generally useful to obtain more accurate predictions, adding the temporal dimension in spatial statistical models is not trivial, and specific models are needed that integrate both variability in space and time. In the context of a spatio-temporal random process, Gräler et al. (2016) proposed to adapt the covariance function with the temporal component. The climatic variable is characterized by the location s_i of the weather station where it was observed, and by the time stamp t_i that identifies when the measurement was taken: the spatio-temporal coordinates are noted (s_i,t_j). The spatio-temporal autocorrelation can be modelled using a variogram, that represents the semivariance between any pair of points that are separated by a spatial lag h and a temporal lag u, and can be computed as follows (Sherman, 2011):

$$\gamma(h,u) = \frac{1}{2}\mathbb{E}(\eta(s_i,t_j) - \eta(s_i+h,t_j+u))^2 \tag{5.2.1}$$

where $\mathbb{E}$ denotes the mathematical expectation.

The empirical variogram is used as a first estimate of the variogram model, directly derived from the sample data. Different covariance models are available in the R package *gstat* (Gräler et al., 2016; Pebesma, 2004) to fit the empirical variogram, among which separable, product sum, metric,

sum metric and simple sum metric covariance functions. Based on our CV results, we found an advantage of using metric or sum metric models to interpolate precipitation data. Both metric and sum metric models incorporate a spatio-temporal anisotropy (k) term, thereby more flexibility. From a meteorological standpoint, the motivation behind modelling anisotropy is that rainfall intensity can be affected by wind direction and geomorphological characteristics (Tomczak, 1998). Model selection among these different covariance functions was achieved for each environment to predict, based on 5-fold CV for temperature and humidity, and with leave-one-station-out for precipitation. Interpolation errors were calculated using the root mean square error (RMSE) and interpolation accuracy with the Pearson correlation between predicted and observed daily values during the maize growing season. After selecting the covariance model with the lowest average RMSE, the latter was used to predict in-field weather stations.

With respect to the land-based weather stations used for interpolation, the Global Historical Climatology Network (GHCN) was chosen as an integrated database with a dense network for rainfall and temperature data in the US, for which data can be easily retrieved via the rnoaa R package (Chamberlain, 2021). In particular, we can recommend to use different sources of data in order to impute daily rainfall; we can for example report a gain in cross-validated accuracy by including the "Community Collaborative Rain, Hail and Snow" (also named CoCoRaHs network), which is a network of volunteer weather observers, most likely because the spatiotemporal coverage of precipitation data was thereby increased. It is essential to have enough data to accurately describe the random phenomena and to fit a variogram model to the empirical variogram.

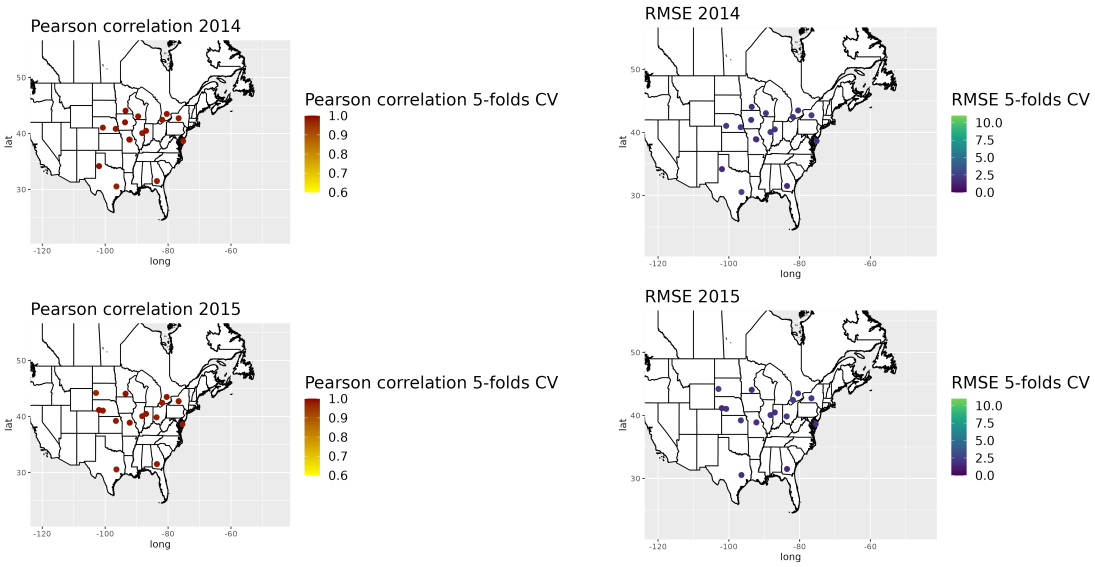


Figure 5.1: Average Pearson correlation coefficient and average root mean square error (RMSE) map for minimum temperature (TMIN) calculated for the different interpolated G2F environments, using 5-fold CV (complete random partition of all spatio-temporal points in the dataset), for years 2014 and 2015.

Figures 5.1, 5.2 and 5.3 show the cross-validated estimates. Very reliable estimates (average r>0.9) could be obtained by the kriging approach to predict temperature data, using a random 5-fold CV.

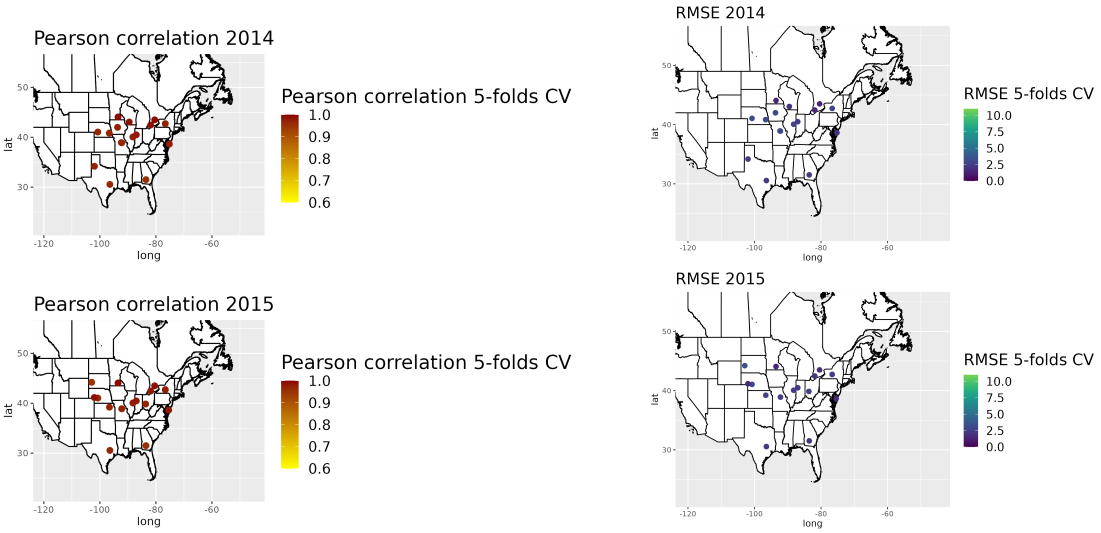


Figure 5.2: Average Pearson correlation coefficient and average root mean square error (RMSE) for maximum temperature (TMAX) calculated for the different interpolated G2F environments, using 5-fold CV (complete random partition of all spatio-temporal points in the dataset), for years 2014 and 2015.

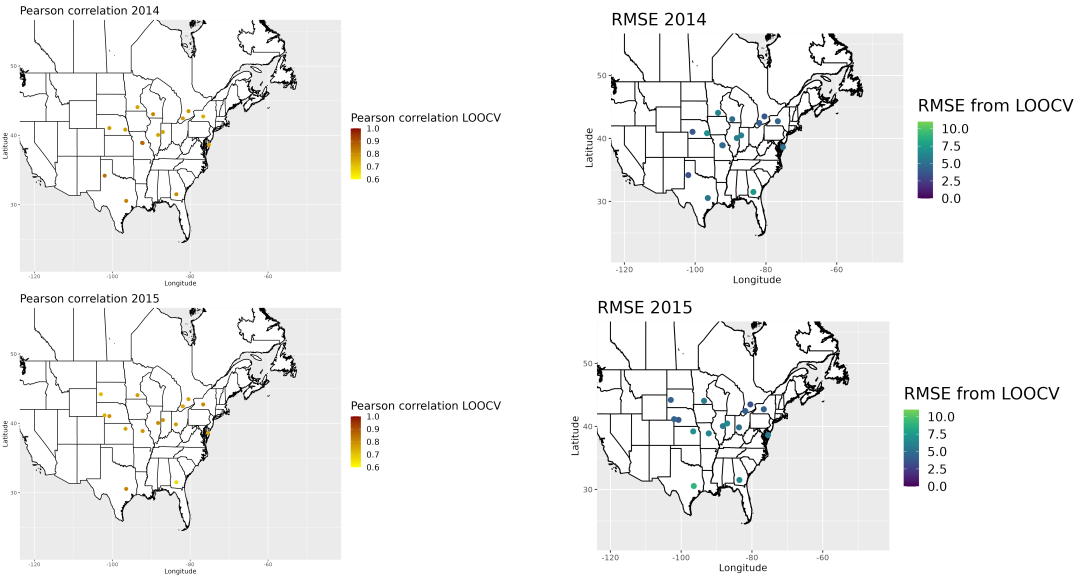


Figure 5.3: Average Pearson correlation coefficient and average root mean square error (RMSE) for precipitation (PRCP) calculated for the different interpolated G2F environments, using leave-one-location-out CV (LOOCV), for years 2014 and 2015.

For precipitation (Figure 5.3), the LOOCV was more challenging, because we wanted to assess the validity of interpolated precipitation data when no data from the exact location was used to fit the model. Overall, accuracy was satisfying and could be used as a proxy for the true rainfall data fallen at the given field experiments.

However, the above-described methodology requires time and additional knowledge to implement geostatistical models, and one could reasonably object that filling-in missing or erroneous values using satellite-based data is much more straightforward (See General Introduction 1.3.1) (Grassini et al., 2015). However, their accuracy to interpolate rainfall and relative humidity data, is regu-

larly questioned, because they do not systematically take into account the topography (Prof. Dr. Reimund P. Rötter, personal communication). Different studies for agro-climatic purposes have investigated the quality of model-based precipitation results from the NASAPOWER database in comparison with rainfall data from surface weather stations (SWS). Some of these publications report significantly better performance of SWS data (Duarte and Sentelhas, 2020), while others indicate encouraging results for NP data for crop yield simulation (Araghi et al., 2021; Monteiro et al., 2018). Accordingly, a crucial question remains: does reducing the measurement error in the environmental predictors (for example, using spatio-temporal kriging) really helps to obtain better predictions of genotype performance in the field? This question could be addressed in further studies.

5.2.2 Attempts at finding the best representation of the climatic data with feature engineering

Adding the environmental dimension in GP frameworks is of utmost interest to better account for the fact that environmental factors are intrinsically interwoven with gene expression, as they influence key physiological processes, for example ear and kernel development in maize. Nonetheless, as we draw a parallel between genomic data and environmental data, similar issues need to be taken care of in order to efficiently utilize environmental parameters in GP models.

First, reduction of the dimensionality of weather data is generally required before using the latter as input in prediction models. By analogy with the use of haplotype block partitioning with marker data, summarizing high-dimensional weather time series over time windows has been a widely used strategy (Heslot et al., 2014; Millet et al., 2019; Rincent et al., 2019; Rogers and Holland, 2021). In **chapter 2**, we generated covariates for only three main maize growth stages, namely vegetative, flowering and grain filling, which were hybrid-specific, and estimated based on the truly observed silking dates of each hybrid. By combining each environmental factor to a specific plant developmental stage, our hope was that informative covariates were generated by this feature engineering step, in order to better reflect the impact of climatic variables on critical physiological events. In contrast, in **chapter 4**, only environment-specific climatic covariates were generated, using two methods proposed in the package *learnMET* (either by considering accumulated thermal time or simple naive day-windows), that both neglected the genetic variation for earliness. This was achieved to mimic the fact that genotype-specific flowering time data are often not available for each environment in large-scale breeding programs.

To evaluate the impact of using hybrid-specific ECs versus using ECs that are common to all hybrids grown in an environment and estimated solely based on accumulated thermal time, we ran the dominance model M14, defined in **chapter 4**, but with the set of weather-based ECs used in **chapter 2**, that also includes reference evapotranspiration as feature (the latter was not computed

in **chapter 4**). As shown in Figure 5.4, an advantage of using hybrid-specific and more complex ECs was demonstrated for most of the predicted environments. The average correlation of the model employing hybrid-specific ECs was 0.466 versus 0.441 for the model tested in **chapter 4**.

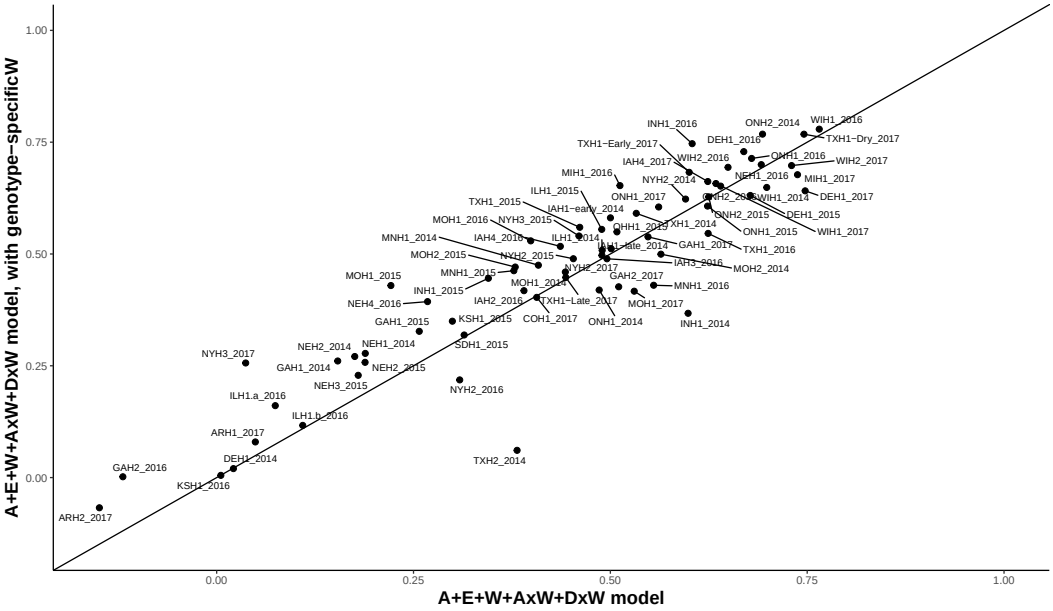


Figure 5.4: Comparison of the predictive ability in each environment of the G2F dataset in the CV0 scheme (leave-one-environment-out) between the dominance model with interactions between SNPs and ECs, where ECs (W_{ECs_stages} from **chapter 4**) are common to all genotypes within an environment (on the x-axis), and the same model, where ECs are genotype-specific (on the y-axis). Only weather-based ECs were used in the W matrix. The line indicates the identity. Environment labels are precised in Table S4.1.

From a biological perspective, the impact of abiotic stress largely depends on the timing of major developmental stages, such as flowering time, and stress covariates also need to account for crop-specific temperature thresholds (Heslot et al., 2014). For instance, different studies reported that high temperatures in maize at flowering time can have an impact on silk and tassel synchrony, and can cause decrease of pollen viability and germination (Dong et al., 2021), which results in kernel abortion, shortening of the duration of grain filling (Edreira and Otegui, 2012) and reduction of grain yield (Dong et al., 2021; Hatfield and Prueger, 2015). Although some genotypes might be more prone to tolerate heat stress, maize can generally flourish with mean temperatures between 28 and 32°C (Sánchez et al., 2014).

Some ML methods have also been applied with very little feature engineering. Exploiting the same original 2014-2017 G2F dataset, Washburn et al. (2021) used gridded estimates of daily weather parameters from DAYMET (Thornton et al., 2016), (e.g. precipitation, minimum and maximum temperature, radiation, vapor pressure), in addition to management and extended soil data, as input data for convolutional neural networks. Apart from cumulative thermal time, no additional climatic-based variable was derived from the original weather or soil dataset, and all daily weather data were provided as input data. Regarding model performance, an average predictive

ability of 0.39 was obtained with a CV scheme named "GEM Practical" (genetic, environmental and management holdout scenario), that was very similar to the CV0-Year prediction scheme we implemented in **chapter 2**, for which our results indicate a weighted average predictive ability of 0.377. Saliency maps were employed in this study to visualize the relative importance of the features, and particular high feature importance scores were observed for precipitation in the first weeks after planting, then diminishing but remaining high throughout the crop growing season. Precipitation during vegetative stage was also an important predictor in GBDT models in **chapter 2**. Nonetheless, we argue that the use of a reasonable number of identified variables that quantify potential environmental stresses (e.g. prolonged heat or drought stress), and are related to a specific growth stage, is more useful to draw general conclusions and to design cultivar ideotypes. An ideotype refers to the set of morphological and physiological phenotypic characteristics that confer a crop a suitable adaptation for a given type of environment (Martre et al., 2015). For example, Ly et al. (2017) used the water deficit, calculated based on the daily rainfall and potential evapotranspiration, to better take into account the drought stress at flowering in reaction norm models, and showed that the reaction norm to drought yielded prediction gains between 2.4% and 12.9%. The season average water stress associated with drought, calculated via the crop growth model APSIM Holzworth et al. (2014), was systematically the top best ranked predictor based on average normalized permutation importance for three test years in the study by Shahhosseini et al. (2021). We also calculated the Penman-Monteith crop evapotranspiration from a reference crop canopy (Allen et al., 1998), as well as water balance estimates with the precipitation and irrigation data (**chapter 2**), and implemented the calculation of this variable in *learnMET* (**chapter 3**).

In **chapter 4**, a very different approach was assessed by using dynamic time warping as a nonlinear method to quantify climatic similarity between two environments on the basis of their respective daily weather time series. This time-series representation of the environments in the reaction norm models demonstrated similar or slightly better predictive abilities than methods based on ECs, and was useful to group environments. A limitation with how we carried out this method, though, is that it does not allow to identify subsequently the weight of some particular environmental features to explain grain yield. To this end, separate kernels for temperature or water patterns could be considered.

On the other hand, reduction of data dimensionality by performing feature extraction on the environmental component should probably be avoided. Rogers and Holland (2021) compared two types of G×E models on the G2F dataset. One of the models proposed by these authors, named PCA(Markers)*Env, used as input PCs derived from a marker dominance relationship matrix and all environmental variables, while the other model, named PCA(Env)*Markers used as input PCs obtained from the environmental data together with all dominance marker data. Rogers and Holland (2021) reported better predictive abilities on average by using the first type of model, and suggested that specific stress indicators could likely explain on their own a larger part of

the phenotypic variation than other variables with almost null effect, as supported by the results obtained by Ly et al. (2018) mentioned above, that outline the specific weight of drought at flowering time in wheat, for instance. Utilizing as input features the principal components (PCs) hinders readability and interpretability, compared to the use of original features.

5.2.3 How to identify the most relevant environmental variables?

Secondly, collinearity among weather and soil-based covariates is also a critical topic with high-dimensional environmental datasets. As we predicted grain yield given climatic and soil variables, we could observe substantial correlations among some of these variables (**chapter 2**, Supplemental Figure S2). Some of these correlations are completely expected, because some variables are connected via mathematical formulas, such as the Penman-Monteith (FAO-56 method) reference evapotranspiration (Allen et al., 1998), that is derived from solar radiation, air temperature, humidity and wind daily data:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\frac{900u_z(e_s - e_a)}{T+273}}{\Delta + \gamma(1 + 0.34u_z)}, \tag{5.2.2}$$

where ET_0 is the reference evapotranspiration rate (mm/day), R_n the net radiation flux (MJ/m^2/day), G the sensible heat flux into the soil (MJ/m^2/day), T the mean air temperature (°C), e_s the mean saturated vapor pressure (kPa) calculated using the daily minimum and maximum air temperatures (°C), e_a the actual vapour pressure derived from relative humidity data (kPa) - e_s-e_a is noted the saturation vapour pressure deficit (kPa) -, u_z the wind speed at 2 m height (m/s), Δ the slope of the saturated vapor pressure curve (kPa/°C) and γ the psychrometric constant (kPa/°C).

Another example is between the different soil, silt and sand fractions, because the percentage of each of these elements depends on the two others to explain the total soil texture. Other correlated distributions are generally explained due to observed climatic phenomena, such as between vapour pressure deficit and temperature. Thus, the question can be asked whether the presence of correlated features affects predictions or variable importance ranking. In the original version of **chapter 2**, we tested a recursive feature elimination for the two GBDT frameworks, by sequentially removing the 10 least important environmental variables and refitting the model at each iteration of feature selection (Kuhn et al., 2013). The relative contribution of each feature to the fitted model is calculated by taking each feature's contribution for each tree in the ensemble of boosted trees. Results with a leave-one-year-out (CV0-year) scheme are presented in Figure 5.5. Even though this method allows to reduce redundancy in the input space, it is computationally burdensome and did not yield major improvements, likely because gradient boosted trees methods are not strongly impacted by removal of correlated variables, as long as they use another feature that carries similar information (Hastie et al., 2009).

Other techniques have also been employed to reduce the number of environmental variables to the ones affecting the most severely the trait of interest. Rincent et al. (2019) used an environmental

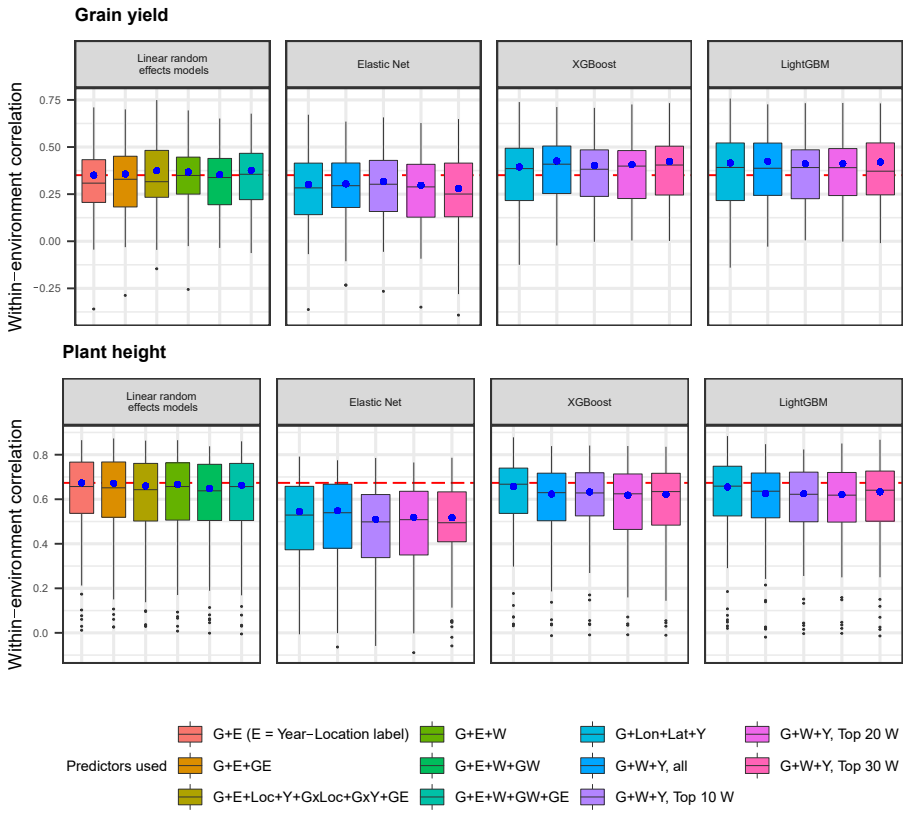


Figure 5.5: Within-environment average predictive abilities for two phenotypic traits (A) grain yield and (B) plant height. For XGBoost and LightGBM, different sizes of best-ranking environmental predictors were considered.
G, SNPs; E, trial label (environment); GxLoc, genotype-by-location interaction; GxY, genotype-by-year interaction; GE, genotype-by-environment interaction; GW, interaction between W and SNPs; Lon, longitude; Lat, latitude; Y, year, W, weather- and soil-based covariates + Longitude + Latitude. The blue dots correspond to the weighted average predictive ability obtained across all predicted environments with the the leave-one-year-out CV scheme for each model. The red dashed line indicates the weighted average predictive ability of the benchmark model (linear random effect model, G+E).

covariance matrix $\mathbf{W_{AMMI}}$, derived from AMMI decomposition and therefore directly linked to the G × E patterns, to extract a subset of covariates, that have a substantial effect on grain yield, from the environmental covariance matrix $\mathbf{W}$ based on all computed ECs. Using a stepwise forward procedure, an optimal subset of ECs, that yielded a maximum increase of correlation with $\mathbf{W_{AMMI}}$, was thereby determined (called $\mathbf{W_{sel}}$), and used for subsequent prediction problems. Importantly, the authors tested this strategy both on a covariance matrix with a subset of ECs defined using the full dataset, or for each TRN only. The latter strategy is preferable because it prevents data leakage from the TST to the TRN, which can result in overestimated prediction accuracy results. Yet, none of these two approaches did lead to any improvement in predictive abilities, and the results were especially disappointing in the most challenging prediction scenarios CV0 (decrease of predictive ability by 12%) and CV00 (decrease by 7%), with the method that only uses the TRN for feature selection. This might be attributed to the fact that Rincent et al. (2019) also applied the reaction norm models proposed by (Jarquín et al., 2014), that assume the same

weight for each marker-EC interaction terms in the prediction model. In contrast, XGBoost and LightGBM models that we used in **chapter 2** could learn, based on decision trees, which SNPs-derived PC-by-EC combinations were particularly relevant to obtain gains in predictive ability. One further disadvantage of performing a decomposition of the G×E matrix with AMMI analyses is that a common set of genotypes must be evaluated across all environments, and this is hardly applicable when the trials are not connected with the same genotypes, as it is often the case in early breeding stages (See Section 5.1). To account for realized genetic similarity based on marker data, we would recommend an approach that uses marker effects to characterize environments and to compute a distance between them, as proposed by Heslot et al. (2013), and that we also applied in **chapter 4** (Supplementary Figure S4.3). This approach also enables to identify QTL with contrasting patterns effects across environments.

Prior to carrying out prediction analyses, Millet et al. (2019) selected three environmental indices (night temperature and soil water potential during flowering time, and the amount of radiation during the vegetative phase) based on their correlation with grain number of maize reference hybrid. This approach can be described as a *filter* method in the ML terminology, because it does not use for variable selection the same method that will be used for predictions. Li et al. (2021) proposed an even more conservative feature selection approach by selecting only one EC-growth period combination, that presented the highest correlation with environmental means for a given phenotypic trait. Four environmental parameters were considered, all related to temperature and/or day length, and the selected environmental index (among these four parameters at a given day-window) was used in both GWAS and GP models. However, we would argue that additional parameters need to be considered to better represent the environmental dimension.

When it comes to model interpretation with ML methods, separating out the individual effects of collinear features on the target variable can cause some issues. Considering Random Forest, the random choice of choosing one of the correlated variables will be performed for each tree, because each tree is independently built from others (random bagging procedure). On the other hand, XGBoost algorithm (Chen et al., 2015) offers a more reliable interpretation of feature importance, due to the fact that this algorithm learns the relationship between a feature and the outcome in the first iteration, and subsequently use it in the next iterations (fundamental principles of boosted trees described in **General Introduction, Section 1.3.2**), rather than randomly picking one of the correlated features in each tree like RF. Consequently, all of the variable importance should be assigned to the feature having a major role to explain the outcome, although this information should be examined together with statistical correlations between features.

5.2.4 Relevance of soil and management data

In **chapter 2**, we identified the organic matter content as the third most important predictor explaining grain yield across environments, directly followed by the proportion of clay. We also

noted that covariates related to soil texture were the top factors to predict the trait plant height (**chapter 2**). The significance of these soil factors was also emphasized in two other studies harnessing the G2F dataset. Rogers et al. (2021) established a factor analysis of environmental data, and found significant negative loadings associated with both sandy and clay proportions for the trait grain yield. In the study of Washburn et al. (2021), the soil-based factors represented about 35% of total importance score, before genetic and weather-based factors, when historical data were additionally used in model training. Organic matter content plays an important role in ensuring adequate soil functionality (e.g. coping with changes in soil acidity) and soil fertility (Tiessen et al., 1994). Although it is not possible to draw from our results direct causalities, it is also an established fact in agronomy that soils with higher clay content are more susceptible to compaction than other soil types in case of heavy precipitation and can result in decreased yields (Soinne et al., 2021).

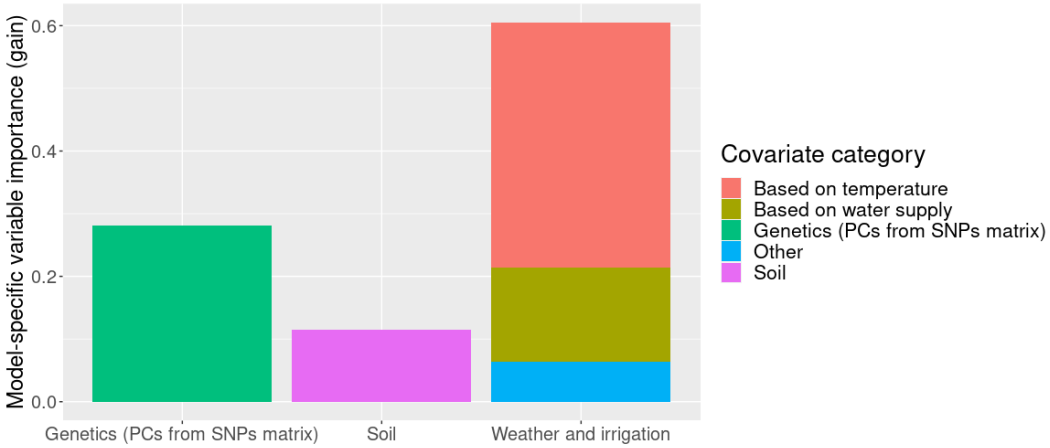


Figure 5.6: Bar plot of the sum of gain scores (model-specific variable importance) by covariate category. The sum of all variable importance scores across predictors is equal to 1.

Summarizing into main categories (weather and irrigation, soil, and genetics), as displayed in Figure 5.6, we found somewhat similar scores to the saliency map scores obtained by Washburn et al. (2021) for soil covariates ($\approx$ 11.5%), when no historical data was used. As proposed by Washburn et al. (2021), other soil characteristics could be relevant to include, such as pH (acid soils are generally associated with lower productivity), soil electrical conductivity, which is an indicator of nutrient availability and of water capacity, or amounts of soil nutrient (nitrogen, phosphorus, potassium). Irrigation was used in their study as a factor, while irrigation data were integrated in our study by directly merging these data with the daily amount of rainfall. Our motivation was to favour the use of quantitative variables, and to avoid mixed data types in our analyses.

Nonlinear interactions between soil types and precipitation, explicitly modelled with an interaction Gaussian kernel, yielded the largest kernel weights associated with G$\times$E in a study with barley in Finland (Gillberg et al., 2019). From an agronomic perspective, interactions between management

factors related to water supply and soil textures have regularly been examined (Fang and Su, 2019; Jalota et al., 2006), and unambiguously adding these interaction components in the prediction models, as implemented by Gillberg et al. (2019), should be considered in further studies. One limit to the feature importance based on gain (displayed in **chapter 2**, Figure 2.6 and in Figure 5.6), is that this measure only indicates the general importance of each individual predictor variable, with interwoven interactions involving this variable. Getting a better level of comprehension of the two-way and three-way interactions, for instance using the Friedman's H-statistic (Friedman and Popescu, 2008) should be considered in further studies.

Management data were not included our study, but some factors were strongly associated with grain yield in the study of Washburn et al. (2021), like plant density. Fertilization levels, crop used as precedent, type of soil preparation, disease pressure, sowing time could also be encoded and integrated in ML models. In a study on crop yield prediction, Shahhosseini et al. (2021) used the simulation crop model APSIM (Holzworth et al., 2014) to generate informative features, that take into account nitrogen limitation. Among others, the season average nitrogen stress (a crop-based APSIM variable) and the nitrogen loss due to leaching and denitrification (soil-based APSIM variable) were integrated in the ML models as input data. Ly et al. (2018) also pointed out the utility of exploiting information on the amounts of nitrogen fertilization and to consider crop nitrogen dynamics by calculating a nitrogen nutrition index (Justes et al., 1994; Lemaire and Meynard, 1997). However, it should be noted that, in the latter study, accuracy gains explained by considering nitrogen stress were very modest (up to 2.4% only) Ly et al. (2018), compared to those obtained by including drought-related stress variables (that we also considered in our analyses). In addition, the accurate estimation of this index in the context of multi-environment plant breeding trials, as proposed by Ly et al. (2018), requires detailed trial information (soil depth, organic and inorganic matter content, precise N management data) and phenotypic data (crop nitrogen at different stages for each genotype). In this thesis, we did not apply sophisticated crop models to obtain more complex crop physiological features (See **General Introduction, Section 1.3.2** about advantages of crop models), due to the lack or incompleteness of the management data in the original metadata file. Besides, most crop growth models need reference genotypic parameters, and their implementation in the context of the G2F hybrid breeding trials, where a large number of unique single-crosses ($>$ 2,000) were evaluated, was not trivial.

5.3 Impact of genetic and environmental similarity across prediction models

5.3.1 Modeling of genetic effects in hybrid predictions

Previous studies have investigated how the genetic relationships between individuals in the TRN and TST impact predictive abilities of GP models (Auinger et al., 2016; Bernal-Vasquez et al., 2017; Habier et al., 2007; Riedelsheimer et al., 2013). In the context of hybrid breeding, including in the TRN crosses with the two parents of hybrids of the TST (also referred as T2 CV scheme (Technow et al., 2014)) was shown to yield better predictive abilities (Riedelsheimer et al., 2013; Technow et al., 2014; Westhues et al., 2017), most likely because the general combining ability of these lines can be learned from these data. In our studies, we did not explicitly model the general combining abilities (GCA) of the parental lines and the specific combining ability of the cross (SCA), but instead directly used *in silico* hybrid genotype data and derived from it either an additive relationship matrix (for reaction norm models) **chapter 2**, or both an additive and dominance relationship matrices **chapter 4**. Examining G2F field trials from 2014-2015, Jarquin et al. (2021) found an average predictive ability of 0.45 with the model that included GCA and SCA, as well as interactions with the environment label covariate, in the CV0 prediction scheme. Although results are not directly comparable because we also included 2016 and 2017 phenotypic data, results from **chapter 4** (Table 4.3) with the model A + D + E + A × E + D × E, with the same CV scheme, show a very similar average predictive ability ($r \approx 0.455$), and a gain of accuracy was systematically observed with models including dominance effects. This can be explained by the within-heterotic group crosses tested in the G2F experiments (See **chapter 4**, Section 4.5.1). In future studies with ML algorithms for hybrid prediction, we would suggest to evaluate models that either directly employ as input features the dominance marker matrix, or that use the first PCs obtained by eigenvalue decomposition of this matrix.

Additionally, our approach in **chapter 2** for the GBDT algorithms is based on using PCs from the marker matrix to capture genetic relationships among maize hybrids. This implies that obtained predictive abilities are not due to linkage disequilibrium between markers and causative QTL for grain yield or plant height, as it should ideally be the case in GP. The reason behind this methodology was the objective of reducing the computational load, of avoiding collinearity issues and to obtain faster results than by using the complete set of SNPs data. It should be also noted that: (i) the predictive models were assessed in various CV schemes, (ii) the number of phenotypic observations was large, (iii) that the environmental component was expected to account for a much larger proportion of the phenotypic variation, and finally, (iv) a rigorous tuning of hyperparameters for all training-test splits was achieved; thus, opting for a dimensionality reduction technique appeared reasonable, and was also proposed by another study with the G2F data (Washburn

et al., 2021). Besides, classical GP prediction methods were also shown to be sensitive to realized genetic relationships among individuals, such as RR-BLUP (Habier et al., 2013). Nonetheless, these methods cannot account for some large effects QTL, whose effects might be sensitive to specific environmental conditions (e.g. the *Ppd-D1 main photoperiod sensitivity locus in winter wheat* (Heslot et al., 2014)). It would be of interest to investigate machine learning models, such as stacked ensemble, that integrate as explanatory variables both markers associated with known QTL (e.g. from previous GWAS studies) along with the PCs representing the main genetic effects. Further, ML methods have already shown their potential to detect interactions between SNPs (epistatic effects) (Azodi et al., 2019; Pook et al., 2020; Zingaretti et al., 2020). The package *learnMET* would be convenient to study additional traits in other species, as it allows to use either SNPs or PCs as predictor variables for a range of ML prediction models, among which gradient boosting, random forest, stacked generalization ensemble models and deep learning models (**chapter 3**).

5.3.2 Importance of relatedness between TRN and TST at the environmental level

In our studies, CV0 schemes consisted of either leaving one year (**chapter 2**), one location (**chapter 2**) or one environment (**chapter 4**) out of the TRN. Among these, predictions for a new year and for new hybrids yielded the lowest predictive abilities, which was consistent with the results of other studies using the same CV scheme with the G2F data (Jarquin et al., 2021; Rogers and Holland, 2021; Washburn et al., 2021). Since the weather represents the most unpredictable component among environmental predictors, these results were expected. Extrapolation, i.e. making predictions for instances in the TST that lie outside the range of values found in the TRN, should in principle be avoided (De Los Campos et al., 2020). To this end, sampling additional years in the given set of locations, thereby augmenting the TRN in order to cover a wide range of environmental variation for all predictors used, would likely help generating predictions for new environments. Especially machine learning approaches benefit from data augmentation, generally associated in MET data with an improved coverage of potential weather conditions: we could notice in **chapter 3** that the performance of the stacked regression model and of XGBoost improved as additional years were used in the TRN set in the forward prediction scenario.

In **chapter 2** and **chapter 4**, we could observe that some environments, characterized by uncommon environmental conditions (e.g. West Texas, dryer weather) compared to the remaining environments, and only sparsely represented in the training set, were not well predicted with GP models including ECs. Widener et al. (2021) also reported negative or lower prediction accuracies with a G $\times$ E-GBLUP model when predicting an extreme environment, and outlined the potential gains that could be achieved by using a more diverse training set. As demonstrated recently by various studies (De Los Campos et al., 2020; Shahhosseini et al., 2020, 2021; Shook et al., 2021;

Washburn et al., 2021), historical data are relevant to enable ML models to really learn interactions among the different genetic, environmental and management components, although extreme weather years generally remain complicated to predict, as reported by Shahhosseini et al. (2021).

On the other hand, CV2, implemented in (**chapter 4**), showed the highest predictive abilities, similar to previous studies (Costa-Neto et al., 2021; Jarquín et al., 2014; Jarquin et al., 2021; Rogers and Holland, 2021), which indicates that when both genotypes and environments from the TST are also included in the TRN, G × E information is well captured by reaction norm models. However, this CV scheme already enables a more efficient allocation of resources, because each genotype does not need to be tested in each environment, but can be predicted by using information from genetically related individuals and correlated environments.

5.4 Outlook and future perspectives

5.4.1 Breeding for target environments?

A very large part of the environmental variation is due to year-to-year weather variability, as shown in **chapter 4**, that are much less predictable than soil or management factors. Evaluating genotypes across multiple years would generate estimates of genotype performance across a larger range of environmental conditions, but also implies more expensive field trials experiments and more time to develop a variety. To develop varieties with a range of various sensitivities to a given environmental stress, it would still be necessary to perform field experiments to generate representative training set data. However, the knowledge generated by data-driven analyses of the form of the phenotypic response to the most critical factors (for instance, heat stress during flowering time, as observed in **chapter 2**, Figure 2.7), could help to further streamline the allocation of resources in multi-environment trials. Concretely, when this phenotypic response is nonlinear, selecting adequately testing environments to better characterize the steepness, key thresholds and possible plateau values for different groups of genotypes could be a meaningful phenotyping strategy. Defining sufficiently dissimilar environments can be achieved using the package *learnMET* (**chapter 3**), that allows to incorporate soil or quantified management data provided by the user and to retrieve satellite-based weather data. In the first function, the package provides plots that allows to assess the similarity among environments using soil and weather data separately, or both types of data jointly.

The machine learning interpretation tools, briefly introduced in **chapter 3**, could theoretically be employed to study how different groups of genotypes (for instance from different families of line crosses) are impacted by specific climatic stress variables across many years. For instance, the accumulated local effects plots could be analyzed with various training sets consisting of different groups of selection candidates, to investigate with a data-driven approach whether some patterns

between abiotic stress resilience and genetic groups are identifiable. If the average yield prediction for a given germplasm is only moderately affected at critical environmental indices related to heat or drought stresses, it can indicate that the developed genotypes might be promising for environments that are likely to experience these stresses in future years. General adaptation traits of some germplasm for certain regions should nonetheless be taken into account, such as photo-thermal units requirement. It can also be envisaged for soil or management factors, in order to gain insights into the sensitivity of genotypes to specific soil types, for example. Once again, we should stress the importance of large training set sizes to harness the full potential of these modeling techniques using field data.

5.4.2 Exploiting historical data to examine yield stability of cultivars

Recently, the Finlay-Wilkinson (FW) model has been revisited by De Los Campos et al. (2020) with the utilization of historical data. The authors performed simulations of genotype performance over 16 years and 16 locations, where locations were sampled in order to represent the main wheat-producing regions in France. This approach allows to consider a broader range of past environmental conditions, rather than only those associated with real past field trials for a few years. From the 143 million simulated grain yield data, De Los Campos et al. (2020), for 28 wheat cultivars, genotype-specific intercepts (i.e. general adaptation) and slopes (level of sensitivity to the quality of the environment) (See General Introduction 1.2.2) were consequently calculated, which can help to determine genotypes that appear the most stable at each site, after smoothing out G $\times$ E effects across many years. *learnMET* (**chapter 3**) could be applied with the same purpose, provided that the training data size is sufficient to leverage machine learning's ability to learn inherently nonlinear interactions. The package allows to retrieve weather data from past years, and predictions with nonlinear algorithms could be obtained for a set of genotypes that has already been grown in a certain region, but for additional years. Thereafter, stability analyses could also be conducted.

5.4.3 Predicting genotype performance in future environments

As highlighted by Trnka et al. (2014), multiple adverse climatic events (e.g. frost stress, late frost, water logging, drought during sowing-anthesis stage, drought during anthesis-maturity stage, heat stress at flowering stage, heat stress at grain fill stage) will increase in frequency in Europe due to climate change. Additionally, the global objective of reducing fertilizer usage by 20% in 2030 in the European Union involves to better take into account nitrogen use efficiency of the developed varieties, and its interaction with weather factors such as precipitation. To account for future potential environmental conditions, a grid of simulated weather time series could be computed by a climate prediction algorithm. Practically, the very popular stochastic weather generator (LARS-WG) (Semenov and Barrow, 1997), has been used in many publications to downscale global climate

models to a smaller resolution, for instance at regional scales (Hashmi et al., 2011; Hassan et al., 2014; Zubaidi et al., 2019). By analyzing the past and future weather time series, the similarity between historical and simulated environments could be estimated using dynamic time warping, as suggested by Netzel and Stepinski (2016). Coupled with historical field trial data, these climate projections could be used to predict *in silico* the performance of the genotypes, with the objective of making better decisions on the most promising varieties to test in field trials.

5.4.4 Potential of phenomics to improve the identification of genotype-specific timing of phenological stages

High-throughput phenotyping technologies can provide efficient solutions to generate in the future larger amounts of plant training phenotypic data, required to train robust predictive models (Yang et al., 2020). In addition, these techniques enable to obtain regular measurements throughout the crop growing season, for instance to precisely characterize the timing of important growth stages, as proposed by Roth et al. (2021) with a modern field phenotyping platform. As we explained earlier, knowing the precise dates of some key phenological events is relevant to obtain accurate values of genotype-specific EC for statistical modeling purposes, and to investigate stress tolerance or avoidance (e.g. early maturing varieties) patterns at development stages that are the most influential on grain yield component traits.

In our studies, we focused on the final integrated trait grain yield. However, examining yield component traits, such as grain number (Millet et al., 2019), or other intermediate phenotypes that could be massively and efficiently generated by the means of high-throughput phenotyping methods, would be of high interest to better understand the extend of G $\times$ E for these specific phenotypic characteristics

Besides, remote sensing data acquired from unmanned aerial vehicles (UAV) can also provide means to identify development stages, as carried out recently in a study with maize (Herrmann et al., 2020), for which high quality spectral data were employed to determine plant phenological stages, as well as irrigation treatments, using drone-mounted superspectral camera.

5.5 Bibliography

Allen RG, Pereira LS, Raes D, Smith M, et al. (1998) Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. Fao, Rome 300(9):D05,109

Araghi A, Jaghargh MR, Maghrebi M, Martinez CJ, Fraisse CW, Olesen JE, Hoogenboom G (2021) Investigation of satellite-related precipitation products for modeling of rainfed wheat production systems. Agricultural Water Management 258:107,222

Auinger HJ, Schönleben M, Lehermeier C, Schmidt M, Korzun V, Geiger HH, Piepho HP, Gordillo

A, Wilde P, Bauer E, et al. (2016) Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (secale cereale l.). Theoretical and applied genetics 129(11):2043–2053

Azodi CB, Bolger E, McCarren A, Roantree M, de Los Campos G, Shiu SH (2019) Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3: Genes, Genomes, Genetics 9(11):3691–3702

Bernal-Vasquez AM, Gordillo A, Schmidt M, Piepho HP (2017) Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC genetics 18(1):1–17

Buntaran H, Piepho HP, Hagman J, Forkman J (2019) A cross-validation of statistical models for zoned-based prediction in cultivar testing

Chamberlain S (2021) rnoaa: 'NOAA' Weather Data from R. URL `https://CRAN.R-project.org/package=rnoaa`, r package version 1.3.0

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, et al. (2015) Xgboost: extreme gradient boosting. R package version 04-2 1(4):1–4

Costa-Neto G, Fritsche-Neto R, Crossa J (2021) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106

Crane-Droesch A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters 13(11):114,003

De Los Campos G, Pérez-Rodríguez P, Bogard M, Gouache D, Crossa J (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nature communications 11(1):1–10

Dias K, Piepho H, Guimarães L, Guimarães PdO, Parentoni S, Pinto MdO, Noda R, Magalhães J, Guimarães C, Garcia A, et al. (2020) Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. Theoretical and Applied Genetics 133(2):443–455

Dong X, Guan L, Zhang P, Liu X, Li S, Fu Z, Tang L, Qi Z, Qiu Z, Jin C, et al. (2021) Responses of maize with different growth periods to heat stress around flowering and early grain filling. Agricultural and Forest Meteorology 303:108,378

Duarte YC, Sentelhas PC (2020) Nasa/power and dailygridded weather datasets—how good they are for estimating maize yields in brazil? International journal of biometeorology 64(3):319–329

Edreira JIR, Otegui ME (2012) Heat stress in temperate and tropical maize hybrids: Differences in crop growth, biomass partitioning and reserves use. Field Crops Research 130:87–98

Fang J, Su Y (2019) Effects of soils and irrigation volume on maize yield, irrigation water productivity, and nitrogen uptake. Scientific reports 9(1):1–11

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. The annals of applied statistics 2(3):916–954

Gillberg J, Marttinen P, Mamitsuka H, Kaski S (2019) Modelling G×E with historical weather information improves genomic prediction in new environments. Bioinformatics 35(20):4045–4052, DOI 10.1093/bioinformatics/btz197, URL https://academic.oup.com/bioinformatics/article/35/20/4045/5448861

Grassini P, van Bussel LG, Van Wart J, Wolf J, Claessens L, Yang H, Boogaard H, de Groot H, van Ittersum MK, Cassman KG (2015) How good is good enough? data requirements for reliable crop yield simulations and yield-gap analysis. Field Crops Research 177:49–63

Gräler B, Pebesma E, Heuvelink G (2016) Spatio-temporal interpolation using gstat. The R Journal 8:204–218, URL https://journal.r-project.org/archive/2016/RJ-2016-014/index.html

Guo Y, Fu Y, Hao F, Zhang X, Wu W, Jin X, Bryant CR, Senthilnath J (2021) Integrated phenology and climate in rice yields prediction using machine learning methods. Ecological Indicators 120:106,935

Habier D, Fernando RL, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397

Habier D, Fernando RL, Garrick DJ (2013) Genomic blup decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607

Hashmi MZ, Shamseldin AY, Melville BW (2011) Comparison of sdsm and lars-wg for simulation and downscaling of extreme precipitation events in a watershed. Stochastic Environmental Research and Risk Assessment 25(4):475–484

Hassan Z, Shamsudin S, Harun S (2014) Application of sdsm and lars-wg for simulating and downscaling of rainfall and temperature. Theoretical and applied climatology 116(1):243–257

Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer series in statistics, Springer, New York, NY

Hatfield JL, Prueger JH (2015) Temperature extremes: Effect on plant growth and development. Weather and climate extremes 10:4–10

Herrmann I, Bdolach E, Montekyo Y, Rachmilevitch S, Townsend PA, Karnieli A (2020) Assessment of maize yield and phenology by drone-mounted superspectral camera. Precision Agriculture 21(1):51–76

Heslot N, Jannink JL, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Science 53(3):921–933

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and applied genetics 127(2):463–480

Holzworth DP, Huth NI, deVoil PG, Zurcher EJ, Herrmann NI, McLean G, Chenu K, van Oosterom EJ, Snow V, Murphy C, et al. (2014) Apsim–evolution towards a new generation of agricultural systems simulation. Environmental Modelling & Software 62:327–350

Huntington T, Cui X, Mishra U, Scown CD (2020) Machine learning to predict biomass sorghum yields under future climate scenarios. Biofuels, Bioproducts and Biorefining 14(3):566–577

Jalota S, Sood A, Chahal G, Choudhury B (2006) Crop water productivity of cotton (gossypium hirsutum l.)–wheat (triticum aestivum l.) system as influenced by deficit irrigation, soil texture and precipitation. Agricultural water management 84(1-2):137–146

Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, et al. (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and applied genetics 127(3):595–607

Jarquin D, De Leon N, Romay C, Bohn M, Buckler ES, Ciampitti I, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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 p 1819

Justes E, Mary B, Meynard JM, Machet JM, Thelier-Huché L (1994) Determination of a critical nitrogen dilution curve for winter wheat crops. Annals of botany 74(4):397–407

Kuhn M, Johnson K, et al. (2013) Applied predictive modeling, vol 26. Springer

Lemaire G, Meynard JM (1997) Use of the nitrogen nutrition index for the analysis of agronomical data. In: Diagnosis of the nitrogen status in crops, Springer, pp 45–55

Li X, Guo T, Wang J, Bekele WA, Sukumaran S, Vanous AE, McNellie JP, Tibbs-Cortes LE, Lopes MS, Lamkey KR, et al. (2021) An integrated framework reinstating the environmental dimension for gwas and genomic selection in crops. Molecular Plant 14(6):874–887

Ly D, Chenu K, Gauffreteau A, Rincent R, Huet S, Gouache D, Martre P, Bordes J, Charmet G (2017) Nitrogen nutrition index predicted by a crop model improves the genomic prediction of grain number for a bread wheat core collection. Field Crops Research 214:331–340

Ly D, Huet S, Gauffreteau A, Rincent R, Touzy G, Mini A, Jannink JL, Cormier F, Paux E, Lafarge S, et al. (2018) Whole-genome prediction of reaction norms to environmental stress in

bread wheat (triticum aestivum l.) by genomic random regression. Field Crops Research 216:32–41

Martre P, Quilot-Turion B, Luquet D, Memmah MMOS, Chenu K, Debaeke P (2015) Model-assisted phenotyping and ideotype design. In: Crop physiology, Elsevier, pp 349–373

McFarland BA, AlKhalifah N, Bohn M, Bubert J, Buckler ES, Ciampitti I, Edwards J, Ertl D, Gage JL, Falcon CM, 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(1):1–6

Millet EJ, Kruijer W, Coupel-Ledru A, Prado SA, Cabrera-Bosquet L, Lacube S, Charcosset A, Welcker C, van Eeuwijk F, Tardieu F (2019) Genomic prediction of maize yield across european environmental conditions. Nature genetics 51(6):952–956

Monteiro LA, Sentelhas PC, Pedra GU (2018) Assessment of nasa/power satellite-based weather system for brazilian conditions and its impact on sugarcane yield simulation. International Journal of Climatology 38(3):1571–1581

Netzel P, Stepinski T (2016) On using a clustering approach for global climate classification. Journal of Climate 29(9):3387–3401

Pebesma EJ (2004) Multivariable geostatistics in S: the gstat package. Computers Geosciences 30:683–691

Pook T, Freudenthal J, Korte A, Simianer H (2020) Using local convolutional neural networks for genomic prediction. Frontiers in genetics 11:1366

Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194(2):493–503

Rincent R, Malosetti M, Ababaei B, Touzy G, Mini A, Bogard M, Martre P, Le Gouis J, van Eeuwijk F (2019) Using crop growth model stress covariates and ammi decomposition to better predict genotype-by-environment interactions. Theoretical and Applied Genetics 132(12):3399–3411

Rogers AR, Holland JB (2021) Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. G3 Genes|Genomes|Genetics DOI 10.1093/g3journal/jkab440, URL https://doi.org/10.1093/g3journal/jkab440, jkab440, https://academic.oup.com/g3journal/advance-article-pdf/doi/10.1093/g3journal/jkab440/42350740/jkab440.pdf

Rogers AR, Dunne JC, Romay C, Bohn M, Buckler ES, Ciampitti IA, Edwards J, Ertl D, Flint-Garcia S, Gore MA, 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(2):jkaa050

Roth L, Rodríguez-Álvarez MX, van Eeuwijk F, Piepho HP, Hund A (2021) Phenomics data processing: A plot-level model for repeated measurements to extract the timing of key stages and quantities at defined time points. Field Crops Research 274:108,314

Sánchez B, Rasmussen A, Porter JR (2014) Temperatures and the growth and development of maize and rice: a review. Global change biology 20(2):408–417

Semenov MA, Barrow EM (1997) Use of a stochastic weather generator in the development of climate change scenarios. Climatic change 35(4):397–414

Shahhosseini M, Hu G, Archontoulis SV (2020) Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science 11:1120

Shahhosseini M, Hu G, Huber I, Archontoulis SV (2021) Coupling machine learning and crop modeling improves crop yield prediction in the us corn belt. Scientific reports 11(1):1–15

Sherman M (2011) Spatial statistics and spatio-temporal data: covariance functions and directional properties. John Wiley & Sons

Shook J, Gangopadhyay T, Wu L, Ganapathysubramanian B, Sarkar S, Singh AK (2021) Crop yield prediction integrating genotype and weather variables using deep learning. Plos one 16(6):e0252,402

Soinne H, Keskinen R, Räty M, Kanerva S, Turtola E, Kaseva J, Nuutinen V, Simojoki A, Salo T (2021) Soil organic carbon and clay content as deciding factors for net nitrogen mineralization and cereal yields in boreal mineral soils. European Journal of Soil Science 72(4):1497–1512

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197(4):1343–1355

Thornton P, Thornton M, Mayer B, Wei Y, Devarakonda R, Vose R, Cook R (2016) Daymet: daily surface weather data on a 1-km grid for north america, version 3. ornl daac, oak ridge, tennessee, usa. In: USDA-NASS, 2019. 2017 Census of Agriculture, Summary and State Data, Geographic Area Series, Part 51, AC-17-A-51

Tiessen H, Cuevas E, Chacon P (1994) The role of soil organic matter in sustaining soil fertility. Nature 371(6500):783–785

Tiezzi F, de Los Campos G, Gaddis KP, Maltecca C (2017) Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. Journal of dairy science 100(3):2042–2056

Tobler WR (1970) A computer movie simulating urban growth in the detroit region. Economic geography 46(sup1):234–240

Tomczak M (1998) Spatial interpolation and its uncertainty using automated anisotropic inverse distance weighting (idw)-cross-validation/jackknife approach. Journal of Geographic Information and Decision Analysis 2(2):18–30

Trnka M, Rötter RP, Ruiz-Ramos M, Kersebaum KC, Olesen JE, Žalud Z, Semenov MA (2014) Adverse weather conditions for european wheat production will become more frequent with climate change. Nature Climate Change 4(7):637–643

Washburn JD, Cimen E, Ramstein G, Reeves T, O'Briant P, McLean G, Cooper M, Hammer G, Buckler ES (2021) Predicting phenotypes from genetic, environment, management, and historical data using CNNs. TAG Theoretical and applied genetics Theoretische und angewandte Genetik 134(12):3997–4011, DOI 10.1007/s00122-021-03943-7, place: Germany

Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A, et al. (2017) Omics-based hybrid prediction in maize. Theoretical and applied genetics 130(9):1927–1939

Widener S, Graef G, Lipka AE, Jarquin D (2021) An assessment of the factors influencing the prediction accuracy of genomic prediction models across multiple environments. Frontiers in Genetics 12

Yang W, Feng H, Zhang X, Zhang J, Doonan JH, Batchelor WD, Xiong L, Yan J (2020) Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Molecular Plant 13(2):187–214

Zingaretti LM, Gezan SA, Ferrão LFV, Osorio LF, Monfort A, Muñoz PR, Whitaker VM, Pérez-Enciso M (2020) Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. Frontiers in plant science 11:25

Zubaidi SL, Kot P, Hashim K, Alkhaddar R, Abdellatif M, Muhsin YR (2019) Using lars–wg model for prediction of temperature in columbia city, usa. In: IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol 584, p 012026

Summary

In plant breeding, genotype-by-environment (G × E) interactions represent a substantial source of variation underlying complex phenotypic traits. A better understanding of G × E interactions can be beneficial to design varieties specifically adapted to certain types of environmental conditions, as well as to optimize the set of environments to include in the training set used in genomic selection. A caveat of modeling G × E interactions with only year-location labels is the impracticality of making predictions for new environments that have never been tested, such as potential future weather conditions. Nowadays, a large wealth of information, such as large volumes of high-throughput environmental, genomic and phenotypic data, can be jointly analyzed to estimate the sensitivity of the phenotypic response of selection candidates to a set of weather and soil conditions. Numerous statistical methods, mostly based on mixed models, have been proposed for integrating these large datasets and to disentangle G × E interactions. However, they rely on strong statistical assumptions, that cannot identify nonlinear responses of genes to environmental conditions. Machine learning approaches are of utmost interest to harness these observational datasets, in particular because they can handle mixed data formats, capture nonlinear and linear interactions and cope intrinsically with irrelevant input variables.

The main objectives of this dissertation were therefore to examine machine learning methods for (i) enhancement of prediction performance using genomic, environmental, and management data, and for (ii) better understanding of the environmental factors impacting predictive abilities of complex agronomic traits. In addition, specific pitfalls and challenges associated with machine learning methods, such as optimization of hyperparameters and utilization of regularization methods, were studied.

In **chapter 1**, we give a general introduction to the topic of G × E interactions and of the collection and processing of environmental data. We also present some general characteristics of machine learning models for genomic prediction in the context of multi-environment trials (METs).

In **chapter 2**, we examine the predictive ability of gradient boosted tree algorithms, a relatively recent machine learning framework, against reaction norm models, for environment-specific predictions. The data we analyzed consist of multi-environment trials from 32 locations across the

United States and Canada from 2014 to 2017, in which maize hybrids were phenotyped for various traits like plant height and grain yield. Soil, management (irrigation information) and weather data were used in prediction models in addition to molecular genome-wide marker data. In particular, genotype-specific environmental covariates were used to summarize daily weather data, to take into account variability in earliness. Results demonstrate an improvement of predictive ability using nonlinear gradient boosting frameworks harnessing environmental data, for the trait grain yield, in a challenging cross-validation scheme aiming at predicting new genotypes in a new year. Modeling explicitly G $\times$ E interactions yielded a gain in predictive ability for the class of random effects models. The effect of environmental factors on grain yield was also investigated, and those related to heat stress, precipitation and soil fertility were ranking among the most important variables.

In **chapter 3**, we describe an R package (*learnMET*) that provides a user-friendly pipeline to evaluate machine learning algorithms for prediction of genotype performance in different multi-environment prediction scenarios. Weather data can be retrieved from a public satellite-based platform (NASA POWER) or derived from field weather stations data. Well-known relationships in ecophysiology (vapour pressure deficit, reference evapotranspiration) and abiotic stress covariates are computed based on the available climate data. Additionally, various methods are proposed to summarize the daily climate data into temporal window sizes, some of which attempting to predict the timing of important developmental stages based on accumulated thermal time. Different evaluation metrics are provided as output when a cross-validation scheme is evaluated, to allow users to decide on the best model to use with their own data. We assessed some of the proposed prediction tools against a parametric benchmark method. Further, the fitted model can be used to gain insights into the relative contribution of different environmental or genetic factors, as we implemented gateways to other expert R packages for machine learning model interpretation.

In **chapter 4**, a new method is introduced to build environmental similarity matrices using a nonlinear distance measure, named Dynamic Time Warping (DTW), calculated between weather time series characterizing the crop growing season in each environment. This metric was used to cluster crop growing events and applied to two MET datasets (the Genomes to Fields dataset from **chapter 2**, and a wheat dataset from CIMMYT). Reaction norm models defined in a similar manner to the models implemented in **chapter 2** were tested (i) with a similarity matrix based on environmental covariates, and (ii) with a similarity matrix derived from DTW distance. According to our results, the latter explained a larger part of the environmental variance than the former, and better captured additive-by-environment and dominance-by-environment interaction effects. Therefore, we encourage further exploring DTW distance as an effective and simple approach to quantify similarity between time series, which could be applied to other types of datasets, such as time series measurements with high-throughput phenotyping data.

Finally, in **chapter 5**, we discuss limitations and possibilities related to these studies. In particular, special attention should be given to environmental data quality, to the design of the training set

data in order to avoid extrapolation, and to preprocessing techniques for improving predictive performance.

Zusammenfassung

In der Pflanzenzüchtung repräsentieren Genotyp-Umwelt-Interaktionen (G × E) eine bedeutende Quelle von Variation in komplexen phänotypischen Merkmalen. Ein besseres Verständnis von G × E Interaktionen kann einen Mehrwert für die Selektion von Sorten bieten, die für bestimmte Umweltbedingungen angepasst sind. Zusätzlich lassen sich diese Erkenntnisse nutzen, um die Auswahl der Umwelten, die das Training Set für genomische Selektion bilden, zu optimieren. Ein Nachteil der Modellierung von G × E Interaktionen, die lediglich Informationen zur Klassifizierung von Jahr und Ort nutzt, ist, dass es nahezu unmöglich ist Vorhersagen für bisher ungeteste Umwelten, wie beispielweise zukünftige Wetterverhältnisse, zu treffen. Heutzutage lassen sich große Mengen an Umwelt-, genomischen und phänotypischen Daten gemeinsam auswerten, um die Sensitivität des Phänotyps von Selektionskandidaten gegenüber Wetter- und Bodenverhältnissen zu schätzen. Zahlreiche statistische Methoden, die häufig auf gemischten Modellen basieren, wurden bisher auf ihre Eignung überprüft große Datensätze gemeinsam auszuwerten und G × E Interaktionen aufzuschlüsseln. Diese Ansätze basieren allerdings auf statistischen Annahmen, die es verbieten, nicht-lineare Zusammenhänge zwischen Genen und Umweltbedingungen zu identifizieren. Ansätze aus der Domäne des maschinellen Lernens sind hier von großem Interesse, da sie es ermöglichen Daten mit unterschiedlicher Kodierung auszuwerten, und dabei sowohl lineare als auch nicht-lineare Interaktionen zu berücksichtigen und automatisch mit uninformativen Variablen umzugehen.

Die Hauptziele dieser Arbeit bestehen daher darin Ansätze des maschinellen Lernens dahingehend zu überprüfen, ob sie (i) die Vorhersageleistung mittels genomischer, Umwelt- und Managementdaten verbessern können und (ii) es ermöglichen ein verbessertes Verständnis darüber zu erlangen welche Umweltfaktoren die Vorhersagegenauigkeit komplexer, agronomischer Merkmale beeinflussen. Zusätzlich wollten wir untersuchen, wie sich beispielsweise die Optimierung von Hyperparametern sowie die Nutzung von Regulierungsmethoden, welche bei der Anwendung von Algorithmen im Bereich maschinelles Lernen berücksichtigt werden müssen, auf die Vorhersage der Merkmale auswirkt.

Das erste Kapitel liefert eine allgemeine Einleitung der G x E Thematik sowie der Erfassung und Verarbeitung von Umweldaten. Zusätzlich präsentieren wir allgemeine Eigenchaften von Modellen der Domäne maschinellen Lernens im Zusammenhang mit genomischen Vorhersagen vor dem

Hintergrund von mehrortigen Versuchen (METs).

Im zweiten Kapitel untersuchen wir die Fähigkeit von "gradient-boosted tree"-Algorithmen, im Vergleich mit Reaktionsnormmodellen, umweltspezifische Vorhersagen zu treffen. Die untersuchten Daten umfassen phänotypische Messungen diverser Phänotypen von Maishybriden, darunter Pflanzenhöhe und Kornertrag, welche in mehrortigen Versuchen erfasst wurden. Dieser Datensatz beinhaltet Messungen an 32 Orten, die über die gesamten USA und Kanada verteilt sind, und in den Jahren 2014 bis 2017 erhoben wurden. Zusätzlich wurden Bodenparameter, Managementinformationen (beispielsweise Bewässerungsdaten) Wetterdaten und genomische Markerdaten für die Vorhersagemodelle verwendet. Im Besonderen wurden genotypspezifische Umwelt-Kovariablen eingesetzt, um tägliche Wetterdaten zusammenzufassen und damit Unterschiede in der Reife von Genotypen berücksichtigen zu können. Unsere Ergebnisse zeigen einen Mehrwert nicht-linearer "gradient-boosting"-Algorithmen bei der Nutzung von Umweltdaten zur Vorhersage des Merkmals Kornertrag für Genotyp-Umwelt-Kombinationen, die bisher nicht experimentell untersucht wurden. Die explizite Modellierung von G x E Interaktionen erzielte einen Zugewinn an Vorhersagegenauigkeit für die Klasse von Modellen mit zufälligen Effekten. Der Einfluss von Umweltfaktoren auf das Merkmal Kornertrag wurde ebenfalls untersucht. Insbesondere Hitzestress, Niederschlag und Bodenfruchtbarkeit konnten in diesem Zusammenhang als Faktoren mit starkem Einfluss auf den Kornertrag identifiziert werden.

Im dritten Kapitel beschreiben wir ein R-Paket (*learnMET*) welches eine nutzerfreundliche Zusammenstellung von Algorithmen des maschinellen Lernens bietet, um unterschiedliche Konstellationen von Genotypen und mehrortigen Versuchen, hinsichtlich der Vorhersagefähigkeit genotypischer Leistung, zu untersuchen. Zu diesem Zweck können Wetterdaten automatisch von einer öffentlichen, satellitenbasierten Plattform (NASA POWER) bezogen werden oder, alternativ, extern ermittelt und integriert werden, sofern Daten von Wetterstationen vorliegen. Allgemein bekannte ökophysiologische Beziehungen (z.B. Dampfdruckdefizite und Evapotranspiration) sowie abiotische Stresskovariablen werden, basierend auf den vorliegenden Klimadaten, berechnet. Zusätzlich schlagen wir unterschiedliche Methoden zur Aggregierung der täglichen Klimadaten in größere, zeitliche Fenster vor; darunter Ansätze zur Vorhersage des Zeitpunkts wichtiger Entwicklungsstadien auf der Grundlage akkumulierter Wärmeeinheiten. Im Zuge einer Kreuzvalidierung liefert die Software den Nutzern unterschiedliche Metriken zur Bewertung der Eignung unterschiedlicher Modelle für ihre Daten. Zur Validierung unserer Ansätze haben wir diese mit weitverbreiteten, parametrischen Modellen verglichen. Neben unseren eigenen Ansätzen haben wir explizit eine Integration anderer Softwarepakete verfolgt, um die Erfassung des relativen Beitrags unterschiedlicher genetischer oder Umweltfaktoren zu ermöglichen ohne sich mit der Syntax diese Pakete vertraut machen zu müssen.

In Kapitel Vier stellen wir eine neue Methode zur Erstellung von Ähnlichkeitsmatrizen auf der Grundlage von Umweltdaten vor. Diese basieren auf einer nicht-linearen Distanzmetrik, welche als Dynamic Time Warping (DTW) bezeichnet wird, und mittels einer Zeitreihenanalyse berechnet

wird, welche charakteristisch für die Wachtumsperiode einer Kulturart in einer gegebenen Umwelt ist. Diese Metrik wurde verwendet um bestimmte Entwicklungsstadien zu gruppieren und auf zwei mehrortige Datensätze angewandt (Genomes to Fields Datensatz aus dem zweiten Kapitel und Weizendaten des CIMMYT). In einem ersten Schritt wurden Reaktionsnormmodelle mit einer Ähnlichkeitsmatrix, auf der Basis von Umweltkovariablen, geprüft. In einem zweiten Schritt wurden Rekationsnormmodelle mit einer Ähnlichkeitsmatrix, die mittels DTW-Distanz erstellt wurde, verwendet. Unsere Ergebnisse zeigen, dass DTW-basierte Ähnlichkeitsmatrizen einen größeren Anteil der Umweltvarianz erklären konnten als der alternative Ansatz und zusätzlich vorteilhaft sind um additive und dominante Interaktionseffekte zwischen Genotyp und Umwelt zu modellieren. Daher regen wir an DTW-Distanz, als effektiven und simplen Ansatz zur Quantifizierung der Ähnlichkeit zwischen Zeitreihen, auf andere Datensätze, wie beispielsweise Zeitreihenanalysen von phänotypischen Daten aus Hochdurchsatzverfahren, anzuwenden.

Abschließend diskutieren wir in der General Discussion die Möglichkeiten und Grenzen unserer bisherigen Studien. Ein Hauptaugenmerk sollte dabei auf die Qualität von Umweltvariablen, auf das Design des Traingsdatensatzes zur Vermeidung von Extrapolation sowie auf Datenaufbereitungsmethoden zur Verbesserung der Vorhersageleistung gelegt werden.

Acknowledgments

I first want to thank my two PhD supervisors, Prof. Dr. Timothy Beissinger and Prof. Dr. Henner Simianer, who have guided me in the last three and a half years. Tim, thank you for your continuous support, the insightful discussions on statistical and mathematical topics that certainly helped me to improve this thesis. I really appreciated your sincere curiosity about machine learning applications throughout my PhD studies, and your enthusiasm about the topic of environmental data in genomic prediction. You also offered me an amazing flexibility in terms of projects and collaborations, for which I feel grateful. To my second supervisor, Prof. Dr. Simianer, whom I first met when I was completing my internship at KWS in 2018: thank you for having accepted to supervise my work and for your high-standard scientific direction. Your pertinent suggestions and careful review of my work have very positively challenged me during my PhD studies. Learning to master the different aspects of my project was a real challenge, and I feel lucky that both Tim and Prof. Simianer strongly encouraged me. I also had the chance to attend many high-quality academic courses and conferences and to defend my work at different workshops and conferences.

I would like to thank Prof. Dr. Alexander Lipka for becoming a member of my examination board, and for all the strong support in statistics and ideas he shared with me during our collaborative meetings. Many thanks also to Prof. Dr. Diego Jarquin for accepting to review my thesis on a short notice.

I also thank KWS SAAT SE for financing the PhD thesis. In particular, I acknowledge the interest for my research and the comments and advise from Dr. Gregory Mahone, who is also a member of my thesis committee, and from Dr. Sofia Pereira da Silva. Many thanks also to Dr. Milena Ouzunova for showing her interest in my work early, as I completed two internships at KWS during my engineering studies. I also thank co-authors of my first paper Dr. Patrick Thorwarth, Dr. Malthe Schmidt and Jan-Christoph Richter.

I also want to express my gratitude for having had the amazing opportunity to work with the Genomes to Fields dataset, that has been extremely valuable for my entire PhD thesis. I feel indebted to Prof. Dr. Natalia de Leon for having offered me the chance to present my research at the 2020 Genomes to Fields workshop, and to discuss with other collaborators about their

experience with environmental data in breeding. Overall, I am deeply grateful to all collaborators and people having contributed to the generation of this rich dataset, that will undoubtedly be very useful in the future to better understand interactions between genomics and environment.

I would like to thank Dr. Johannes Martini for the ideas he shared with me and for the collaboration on the fourth chapter of this thesis.

I am extremely grateful to Mrs Döring and Mrs Belaed for their great and reliable help with administrative matters.

I thank all my colleagues from the Plant Breeding Methodology Group for the support and the very pleasant working atmosphere, even in a context of remote work. Special thanks to Prof. Dr. Wolfgang Link for his true interest in my work and the applications it may have in applied plant breeding: it has been a great source of motivation for me. I would like to thank Baris Alaca, for sharing his knowledge when working together in the field, and Mila Tost, for her positive mindset and the great exchange about tricks in R or for using the HPC server. It was also a great pleasure to meet and work with Julia Hagenguth, Alex Windhorst, Rebecca Tacke, Dr. Antje Schierholt, Dr. Stefanie Griebel, Dr. Medhat Mahmoud and all unmentioned colleagues. Thanks also to my colleagues from the Animal Breeding and Genetics group, in which I first started my PhD journey. Thanks also to Matthew Murphy, for the interesting discussions on the topic of epistasis.

Je souhaite remercier l'une de mes professeures de mathématiques de classe préparatoire au lycée Saint-Louis, Mme Bernet, pour m'avoir encouragée à rester tenace dans cette matière et pour les solides fondations qu'elle a contribuées à bâtir dans ma réflexion mathématique, et scientifique en général. Mon intérêt pour les mathématiques s'est surtout révélé quelques années plus tard, mais il est certain que j'ai largement bénéficié de la qualité de l'enseignement reçu lors de mes années en classe préparatoire.

Parce qu'un doctorat est aussi associé à des moments de doute et de découragement, je remercie du fond du coeur mes amies de l'Agro d'avoir été tellement encourageantes et compréhensives, malgré le Rhin qui nous sépare. Anna, Chloé, Maud, Marie-Gabrielle, Laurène, Blandine et Morgane: merci pour tous ces fous rires et marques de soutien au cours de ces dernières années!

Vielen Dank an Lisa, für die schöne Abendessen und viele Lachen in Einbeck. Mit deiner Unterstützung hat sich oft meine Stimmung aufgehellt.

Vielen Dank an meine deutsche Familie, die mich mit große Sorgfalt und Wärme angenommen haben, und auch viele Geduld hatten, als ich oft am Wochenende an irgendwelche Präsentation noch arbeiten musste. Merci à mes parents Christian et Cécile, pour leur encouragement, lorsque je faisais face à de grands moments d'incertitude, et de n'avoir jamais cessé de croire en moi.

Évidemment, tout cela n'aurait pas été possible sans celui qui est devenu mon mari. Matthias, du hast unglaublich viel dafür getan, dass diese Arbeit gelungen ist. Du hast immer sehr viel

Interesse an dem Thema gezeigt, und du hast immer an mich geglaubt. Die Promotion hätte mich ohne dich wahrscheinlich verrückt gemacht! Ich bin froh, dass wir zusammen so viele schöne Orte in Niedersachsen oder im Harz beim Wandern entdeckt haben; jetzt kommt für uns ein neues Abenteuer in der malerischen Schwäbischen Alb!

Curriculum Vitae

Name:	Cathy Colette Westhues, born Jubin
Date and place of birth:	04.06.1994, Beaupréau, France
Nationality:	French

School Education

09/2010 – 06/2012	Degree: Baccalauréat scientifique, Lycée Saint-Sauveur, Redon, France

University Education

09/2012 – 06/2014	Classe préparatoire BCPST, Lycée Saint-Louis, Paris, France
09/2014 – 09/2018	Dipl.-Ing. agr. (equivalent to M.Sc.Eng.) Agronomy and Plant Breeding, AgroParisTech, France Thesis: Comparison of two haplotyping block approaches for genetic diversity analyses and genome-wide association studies. Supervisor: Dr. Julie Fiévet
11/2018 - 04/2022	Ph.D. student in the Division of Plant Breeding Methodology, University of Göttingen, Germany Supervisors: Prof. Dr. Timothy M. Beissinger and Prof. Dr. Henner Simianer

Professional Experience

03/2018 - 08/2018	Internship and master thesis in maize biostatistics group, KWS SAAT SE, Einbeck, Germany
04/2017 - 08/2017	Molecular phytopathologist intern, Selecta Klemm GmbH & Co. KG, Stuttgart, Germany
09/2016 - 02/2017	Internship in maize biostatistics group, KWS SAAT SE, Einbeck, Germany

Publications

Jubin, C., Griebel, S., and Beissinger, T. (2021). Improving genomic tools for outcrossing crops. Molecular Plant, 14(4), 538-540. doi: 10.1016/j.molp.2021.03.013

Westhues, C. C., Mahone, G. S., da Silva, S., Thorwarth, P., Schmidt, M., Richter, J. C., Simianer, H. and Beissinger, T. M. (2021). Prediction of maize phenotypic traits with genomic and environmental predictors using gradient boosting frameworks. Frontiers in plant science, 12. doi: 10.3389/fpls.2021.699589

Westhues, C.C. , Simianer, H., Beissinger, T.M. (2022) learnMET: an R package to apply machine learning methods for genomic prediction using multi-environment trial data. G3, 12(11), jkac226. doi: https://doi.org/10.1093/g3journal/jkac226