Theor Appl Genet (2014) 127:463–480
DOI 10.1007/s00122-013-2231-5

ORIGINAL PAPER

Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions

Nicolas Heslot · Deniz Akdemir · Mark E. Sorrells · Jean-Luc Jannink

Received: 21 March 2013 / Accepted: 31 October 2013 / Published online: 22 November 2013

Abstract

Key message **Development of models to predict genotype by environment interactions, in unobserved environments, using environmental covariates, a crop model and genomic selection. Application to a large winter wheat dataset.**

Abstract Genotype by environment interaction (G*E) is one of the key issues when analyzing phenotypes. The use of environment data to model G*E has long been a subject of interest but is limited by the same problems as those addressed by genomic selection methods: a large number of correlated predictors each explaining a small amount of the total variance. In addition, non-linear responses of genotypes to stresses are expected to further complicate the analysis. Using a crop model to derive stress covariates from daily weather data for predicted crop development stages, we propose an extension of the factorial regression model to genomic selection. This model is further extended to the marker level, enabling the modeling of quantitative trait loci (QTL) by environment interaction (Q*E), on a genome-wide scale. A newly developed ensemble method, soft rule fit, was used to improve this model and capture non-linear responses of QTL to stresses. The method is tested using a large winter wheat dataset, representative of the type of data available in a large-scale commercial breeding program. Accuracy in predicting genotype performance in unobserved environments for which weather data were available increased by 11.1 % on average and the variability in prediction accuracy decreased by 10.8 %. By leveraging agronomic knowledge and the large historical datasets generated by breeding programs, this new model provides insight into the genetic architecture of genotype by environment interactions and could predict genotype performance based on past and future weather scenarios.

Communicated by A. E. Melchinger.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2231-5) contains supplementary material, which is available to authorized users.

N. Heslot · D. Akdemir · M. E. Sorrells · J.-L. Jannink (✉)
Department of Plant Breeding and Genetics, Cornell University, 240 Emerson Hall, Ithaca, NY 14853, USA
e-mail: jeanluc.jannink@ars.usda.gov

N. Heslot
e-mail: nh269@cornell.edu

N. Heslot
Limagrain Europe, CS3911, 63720 Chappes, France

J.-L. Jannink
USDA-ARS, R.W. Holley Center for Agriculture and Health, Cornell University, Ithaca, NY 14853, USA

Abbreviations

BLUP	Best linear unbiased predictor
GBLUP	Genomic estimated best linear unbiased predictor
GEBV	Genomic estimated breeding value
G*E	Genotype by environment interactions
GS	Genomic selection
MET	Multi-environment trials
QTL	Quantitative trait locus
Q*E	QTL by environment interaction
SGL	Sparse group lasso
SNP	Single nucleotide polymorphism
TPE	Target population of environments

Introduction

Genotype by environment interactions (G*E) are one of the most important issues in plant breeding. It is a frequent

observation in multi-environment trials (MET) that genotype performance varies across environments leading to variance differences and rank changes among genotypes (Cooper and DeLacy 1994). These forms of G*E are called non-cross-over (variance differences) and cross-over G*E (rank changes). Rank changes complicate selection for broad adaptation as there might not be one best performing genotype everywhere. In a context of climate change and reduced usage of fertilizers and pesticides, crop environments will likely be more variable, increasing the importance of G*E.

Genomic selection (GS) was first proposed by Meuwissen et al. (2001) to better estimate breeding values based on the simultaneous use of whole-genome markers. A number of empirical and theoretical studies suggest that it could increase genetic gain per unit of time beyond what is possible with phenotypic selection (Heffner et al. 2010; Lorenz et al. 2011).

The GS concept provides an important breakthrough toward a better genotype to phenotype mapping. It solves problems encountered in the application of quantitative trait loci (QTL) study results to breeding, such as overestimation of the identified QTL effects. GS also potentially enables the use of historical breeding data for current breeding efforts. However, an important part of the mapping, the differential response of genotypes to the environment, which causes G*E (Van Eeuwijk et al. 2005) has yet to be included directly into GS approaches. Genomic predictions have so far focused on the computation of breeding values that are single point estimates of genotype performance presumed to be useful across all environments.

In the context of classical plant breeding the G*E issue has been tackled in several ways (DeLacy et al. 1996), the most common is to ignore G*E in the analysis by considering it to be noise. Another approach is to identify repeatable G*E patterns in the data by dividing the environment targeted by breeding into mega-environments that minimize G*E within mega-environments. This allows genotype targeting and increases the trait heritability within the mega-environments, provided that sufficient breeding resources are allocated to each mega-environment (Windhausen et al. 2012). Numerous approaches have been developed to group environments, such as AMMI (Gauch 2006), and clustering (Cooper and DeLacy 1994). Repeatable G*E patterns can also be identified using external data (Löffler et al. 2005; Chenu et al. 2013). For example, if drought stress is known to be the main driver of G*E, environments can be clustered based on drought patterns (Chapman et al. 2000b).

The most powerful integration of G*E within quantitative genetics theory is to consider G*E as a lack of genetic correlation between environments (Falconer and Mackay 1996). Genetic correlations between environments are obtained by considering performances in different environments as different correlated traits.

When G*E is considered as a lack of correlation, it can be taken into account using multiplicative mixed models such as the factor analytic structure to model the covariance between environments responsible for G*E (Piepho 1998; Burgueño et al. 2008; Kelly et al. 2009; Cullis et al. 2010). Those models can be used for GS prediction, by using a relationship matrix based on markers in the mixed model (Burgueño et al. 2012). In practice, those approaches have numerical limitations due to the highly unbalanced nature of most multi-environment plant breeding datasets. In addition, because they are based on observed covariance among environments, they are explanatory a posteriori rather than predictive. They do not allow prediction for a new climatic scenario, a given level of a weather-related stress, or a new environment directly.

A way to gain predictive capability of G*E is to investigate the genetic basis of G*E by identifying the environment parameters responsible for G*E and determining genotype sensitivity. The class of models implementing this approach is termed factorial regression (Denis 1988; Piepho et al. 1998). Genotype performances are the sum of the main genotype effect and of the genotype sensitivity to the stress covariate. Factorial regression has been extended to the differential response of QTL in the biparental mapping case for a few detected QTL (Crossa et al. 1999; Malosetti et al. 2004; Boer et al. 2007).

Including stress covariates in the analysis present some of the same issues encountered using genomic selection (GS) methods for estimating breeding value. A very high number of covariates can potentially be obtained, each explaining a small amount of the total variance, while being highly correlated with each other (Brancourt-Hulmel et al. 2000). Leo Tolstoy's Anna Karenina provides a metaphor for the difficulty of modeling G*E: "Happy families are all alike; every unhappy family is unhappy in its own way." In the context of crops, when genotypes perform poorly in a given environment, it can be due to many different stresses, and deriving general results is a daunting task. On an operational level, the Anna Karenina effect occurs when most of the G*E cannot be explained by a few major stresses or a simple geographic partition of the data.

To be successful for prediction, an approach focusing on the genetic basis of G*E would have to be genome-wide and include numerous stress covariates at the same time. Because response and development curves are often exponential or 'S' shaped (Van Eeuwijk et al. 2005), the framework should accommodate non-linear responses of QTL to the environment variables. Such a framework should enable the modeling of G*E at the allele level focusing on QTL by environment interaction (Q*E).

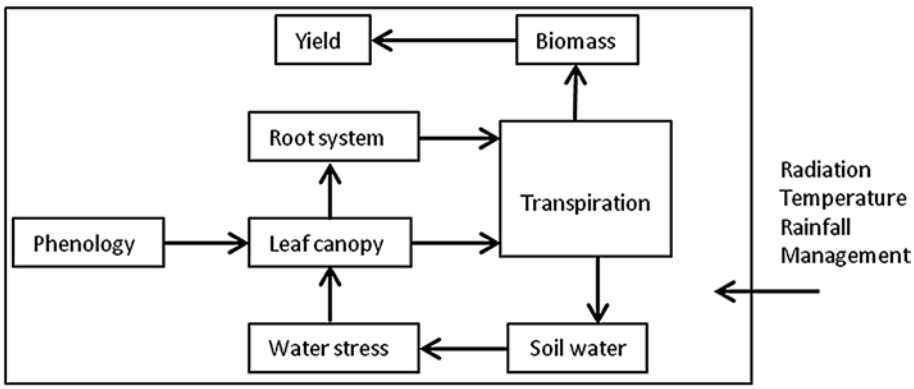

Fig. 1 Typical structure of a crop model

Considering those challenges, some groups have focused on crop modeling to better understand G*E and incorporate external information about the crop (Chenu et al. 2008; Messina et al. 2009). Crop models are sets of equations developed by extensively studying the behavior of a few genotypes under a range of growing conditions. Their main purpose is to predict the development of a crop and the genesis of the different yield components. Figure 1 presents the schematic structure of a crop model. Crop models were initially developed to assist in crop management decisions, strategic planning, yield forecasting, and definition of research needs.

By design, a crop model integrates environment inputs such as weather and soil data in a non-linear way. Crop models, however, have been criticized by crop geneticists for not taking enough genetic variation into account (Hammer et al. 2002). To use crop models for trait prediction across genotypes, research focuses on variation in the parameters of the crop model (Quilot et al. 2004; Jullien et al. 2011). Those parameters are expected to be more heritable than the trait predicted by the crop model, less prone to G*E and to have a less complex genetic architecture. It is a very appealing strategy, but it also requires the measurements of specific phenotypes to recover the model parameters. This is not trivial, as large numbers of environments have to be sampled to study G*E.

Another possibility is to use crop modeling as a tool to perform a physiological integration of environmental data in order to derive stress covariates (Landau et al. 1998, 2000; Boer et al. 2007). These covariates are then used as independent variables in quantitative and statistical genetic models for effect estimation and prediction. This has the advantages of using a genetic model to predict the main genotype effect, whose optimality properties are well known.

Using a crop model to parameterize, the environment data reduce data dimensions from daily weather variables to a few covariates per crop growth stage. The daily weather data are composed of numerous correlated variables most of which have little or no impact on the crop. Moreover, there is a wealth of agronomic and physiological knowledge about the sensitivity of specific growth stages to specific abiotic stresses (Meynard and Sebillotte 1994) (Fig. 2). In addition, the use of a crop model to predict development stages may capture part of the non-linear response of genotypes to the environment by modeling non-linear development processes such as vernalization. It also eliminates the need for specific experiments to measure crop model parameters, and thus enables the use of large commercial breeding datasets that often contain no more than measurements on the final trait of interest, yield. This makes the assumption that the stress response genetic architecture is the same among genotypes at a given developmental stage. Furthermore, interpretability of the model is improved because it uses stress covariates defined by growth stage.

The broad objective of our research is to propose new solutions to integrate environmental data and crop modeling into the genomic selection framework to predict G*E. To this end, we explicitly model whole-genome markers and their differential response to the environment in the GS context to better understand the genetic architecture of G*E. In this study, we extended factorial regression to the GS context and developed a new machine learning approach to capture the response of QTL to stresses non-parametrically. This approach was used along with a crop model to enable the use of daily weather data in prediction models. Those G*E predictions could be used to make breeding decisions for specific adaptation. However, as presented in the discussion, we believe that they are more useful as a tool to understand the interaction and the structure of the target population of environments (TPE). The TPE is the mixture of environments expected for the intended target region (Comstock 1977). This information could then be used to optimize phenotyping.

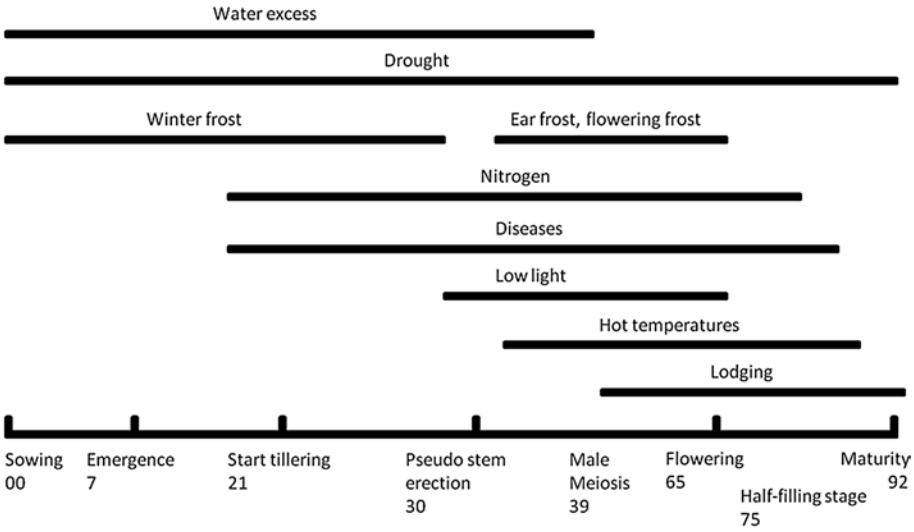

Fig. 2 Major stresses by development stage for winter wheat from Meynard and Sebillotte (1994). The numbers below the development stage names correspond to the Zadoks scale (Zadoks et al. 1974)

Materials and methods

Phenotypic and environment data

Limagrain Europe (Chappes, France) provided a large commercial winter wheat breeding dataset to assess the predictive power of the new models. It consisted of 2,437 genotypes tested for grain yield in 12 locations across France from 2006 to 2011 for a total of 44 environments (year-location combinations). The total number of yield plots was 23,265 generating 9,024 within environment-adjusted genotype means. Those environment means accounted for experimental design which varied from complete block to alpha-lattice and were corrected for spatial variation. The numbers of genotypes observed in each environment ranged from 52 to 974. The data were generated by a single breeding program and corresponded to all trials for the three first years of yield testing of genotypes. As the genotypes were advanced in the breeding program, the level of replication and the number of locations increased, while some genotypes were discarded based on their past performance. The dataset was unbalanced with non-random missingness of the genotypes due to the historical nature of the dataset and because France can be divided in two target environments for winter wheat: the South favoring early maturing genotypes to escape drought and high temperature, and the North favoring late maturing genotypes that yield more. Genotypes were dropped from testing based on their previous performance and new ones added over time. Our data are similar to the winter wheat example (Piepho and Möhring 2006): because an important focus of the breeding is quality, which tends to be negatively correlated with yield, the genetic variance for yield should not decrease dramatically over time. They did not find strong evidence of downward bias for variances of genotype main effects and genotype by year interactions. On average, each genotype was observed in 3.7 environments with 760 genotypes observed in only one environment. The genotypes and the environments corresponded to only one breeding program. This means that any G*E present in this dataset is considered to be small enough that it is manageable by having only one breeding program. Trials were conducted using standard agronomic practices including appropriate use of fertilizers and fungicides.

The lines were genotyped with 1,287 SNP and haplotype-based markers covering the whole genome. Some markers provided perfect linkage with major adaptation loci (dwarfing genes: *Rht-B1, Rht-D1*; vernalization genes: *Vrn-A1, Vrn-A2, Vrn-D5, Vrn-B3*; photoperiod sensitivity genes: *Ppd-D1, Ppd-B1, Ppd-A1*).

In addition, latitude and longitude of each trial location were available along with the sowing date. Daily weather data were obtained from the AGRI4CAST action of the Joint Research Center of the European Commission. (http://mars.jrc.ec.europa.eu/mars/About-us/AGRI4CAST/Introduction). Those data are used to generate crop yield forecasts for European policy makers. The database contains daily meteorological data from 1975 to the last calendar year completed, covering the European Union and neighboring countries. The meteorological parameters are interpolated to a 25 × 25 km grid from a network of meteorological stations. Details about the interpolation procedure and calculations can be found in Van der Goot and Orlandi (2003). The variables available were the mean, minimum, and maximum daily temperature, daily precipitation, daily global radiation, and the ETP [Penman potential evapotranspiration from a crop canopy (mm/day)]. The quality of the data was verified using independent temperature and rainfall records obtained from two trial sites over several years. The interpolated data were well correlated

with the observed data for temperature (correlation above 0.9), less so for rainfall with a correlation of 0.6 going up to 0.8 when considering a weekly scale.

Derivation of stress covariates from the weather data

An intuitive approach to include weather data in an interpretable way, to reduce the number of variables, and to accommodate some non-linearity of response was to define stress covariates by development stage (Landau et al. 1998, 2000; Boer et al. 2007). Brancourt-Hulmel et al. (1999, 2000) and Lecomte (2005) developed a set of stress indices for winter wheat. In their studies, they determined stress covariates by analysis of yield components such as thousand kernel weight and number of kernels per surface area, to identify yield-limiting factors per stage. In addition, they compiled winter wheat sensitivity to stresses at specific development stages from previously published work. Figure 2 shows which stresses are expected to occur for each development stage.

Some of the original stress covariates were excluded because of the lack of necessary information to compute them. They included stress covariates accounting for winter frost damage, disease pressure and nitrogen availability.

Use of a crop model to predict development stages

To derive the stress covariates from the daily weather data, the development stage timing has to be known. This information is often difficult to obtain or not available. To alleviate that need, a crop model, SiriusQuality, was used (Martre et al. 2006). This model is process-based and used a modified version of the phenology model proposed by (Jamieson et al. 1998).

The daily weather data were retrieved from the database using the longitude and latitude of the trial locations. Those data were used as input parameters, to obtain development stages for the crop model SiriusQuality, with default parameters for non-limiting water and nitrogen. The stages predicted by the model used the Zadoks scale code (Zadoks et al. 1974) and stages were 30 (pseudo stem erection), 39 (flag leaf ligule just visible, male meiosis), 65 (anthesis), 75 (half-filling stage), and 92 (maturity). The calendar date of stage 55 (heading date) was derived from the daily sum of temperatures taking stage 39 (male meiosis) as a reference (Gate 1995): first stepi, the sum of daily mean temperature in base 0 °C from planting to heading was calculated using the daily sum of temperature in base 0 °C from planting to meiosis stmei obtained from the crop model asstepi $=$ (stmei $-$ 74)/0.864. Then stepi was converted into a calendar date.

Ideally, the growth stages of each genotype in each environment should be obtained or computed. As this was not feasible and the data covered a wide range of maturities, three sets of development stages were obtained for three elite genotypes, Soissons, Thésee, and Renan, which are early, mid, and late maturing, respectively (He et al. 2012). Those three genotypes were all commercially successful at some point in the last 30 years.

Once the development stages are known, the daily weather data can be used to compute the stress covariates described in Table 1. For example, for the stress covariate stmpmf, the average daily temperature when it is above 0 °C is summed between the predicted meiosis and flowering date, and this is for the three sets of development stages. The covariates capturing frost stress in the spring (ndefr, sti4, ndt0f, st0f) were removed because they had no variance. The three sets of stress covariates plus latitude and longitude of environments were used together (for a total of 101 covariates) and were standardized to zero mean and unit variance before further use. The whole set of covariates (101) was used for all the genotypes regardless of their maturity. This means that for each stress covariate and each observation, there are three values corresponding to the three genotypes used to parameterize the crop model.

Mixed model formulation of G*E

The simplest model to analyze multi-environment trials, in the balanced case, with one observation per genotype and environment combination, for m genotypes and t environments is (Piepho et al. 2008)

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + \varepsilon_1 \qquad \text{(Model 1)}$$

where $\otimes$ indicates the Kronecker product, y, observed phenotypes, β, effects due to the design such as environment in the case above (although block effects could also be represented) and generally treated as fixed, and u is random line effects. In the classic one-step GS approach, the covariance of u is $A\sigma_u^2$ with A the realized relationship matrix computed using molecular markers as $A = VV^t$ with V the marker score matrix. V has dimensions $m \times n$, n numbers of markers. Markers are coded $\{aa, Aa, AA\} = \{-1, 0, 1\}$. σ_u^2 is the genetic variance to be estimated, for example with restricted maximum likelihood. This approach is often referred to as GBLUP for Genomic BLUP. To prevent singularity issues in A, the relationship matrix was bended by adding a small scalar to the diagonal elements (Piepho et al. 2012). ε_1 is assumed to be normally i.i.d. Homoscedasticity is probably an incorrect assumption here. However, when the number of replicates per adjusted mean was used to weight the observations, no gain in accuracy was observed and accuracy even slightly decreased. This is attributable to the variety of experimental design used for the trials forming this dataset. (RCB, alpha-lattice, unreplicated designs). It suggests both that there is no easy way to adjust for heteroscedasticity

Table 1 Stress covariates used and references, modified from Lecomte (2005) page 87

Abbreviation	Description
stmpw, stmpem, stmpmf	Sum of the daily average temperatures (°C) above 0 by development periods
sradw, sradem, sradmf, sradfh, sradhm	Sum of the daily radiation (J/cm^2) by development periods (Gallagher and Biscoe 1978; Monteith 1972)
rdtmpw, rdtmpem, rdtmpmf, rdtmpf30	Ratio srad/stmp by development periods (Fischer 1985)
watxw	Sum of the daily differences *P*-ETP (mm) > 0 from sowing to pseudo stem erection
spetpw, spetpem, spetpmf, spetpfh, spetphm	Sum of the daily differences *P*-ETP (mm) < 0 by development stages
spetpe1	Sum of the daily *P*-ETP (mm) from pseudo stem erection −150 dd to pseudo stem erection +350 dd
nsddr	Number of successive dry days ($P \leq$ ETP in mm) from pseudo stem erection −150 dd to pseudo stem erection +350 dd
ntddr	Number of total dry days ($P \leq$ ETP in mm) from pseudo stem erection stage −150 dd to pseudo stem erection +350 dd
ndefr	Number of days of ear frost (minimal temperature ≤−4 °C) from pseudo stem erection to flowering (Gate 1995)
sti4	Sum of the daily minimal temperatures <−4 °C from pseudo stem erection stage to flowering
ndt0f	Number of days when the daily minimal temperature is ≤0 °C from heading to heading +300 dd
st0f	Sum of the daily minimal temperatures <0° from heading to heading +300 dd
sradmg, sradmm	Sum of the daily radiation (J/cm^2) from meiosis-100 dd to heading, or from meiosis-5 d to meiosis+5 d (Demotes-Mainard et al. 1996)
ndi10m	Number of days when the radiation is ≤1,045 J/cm^2 from meiosis minus 5 days to meiosis plus 5 days
sri10m	Sum of the daily radiation <1,045 J/cm^2 from meiosis minus 5 days to meiosis plus 5 days
nd25m	Number of days when the maximal temperature is ≥25 °C meiosis minus 5 days to meiosis plus 5 days
st25m	Sum of the daily maximal temperatures >25 °C from meiosis minus 5 days to meiosis plus 5 days
nd25ef	Number of days when the maximal temperature is ≥25 °C from heading to flowering (Tashiro and Wardlaw 1990)
st25ef	Sum of the daily maximal temperatures >25 °C from heading to flowering
nd25fh	Number of days when the maximal temperature is ≥25 °C from flowering to half-filling stage (Hunt 1991; Sofield et al. 1977; Stone and Nicolas 1998)
st25fh	Sum of the daily maximal temperatures >25 °C from flowering to half-filling stage
nd25hm	Number of days when the maximal temperature is ≥25 °C from half-filling stage to maturity
st25hm	Sum of the temperatures maximal daily >25 °C from half-filling stage to maturity

w, Winter period—sowing to 1 cm-ear stage; em, 1 cm-ear stage to meiosis; mf, meiosis to flowering; f30, flowering–30 days to flowering; fh, flowering to half-filling stage; hm, half-filling stage to maturity; *P*, rainfall in mm; ETP, potential evapotranspiration in mm; dd, degrees days

here and that we pay no penalty for the simplifying homoscedasticity assumption. In the following models, we assume that this assumption holds. Here the G*E variance is absorbed by the residual ε_1. Habier et al. (2007) have shown the equivalence of GBLUP and ridge regression. That model can then be equivalently written:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \varepsilon_2 \qquad \text{(Model 2)}$$

where α is a marker effect vector: u from Model 1 is equal to $V\alpha$ here.

G*E can be accounted for explicitly by the following model:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + (I_t \otimes I_m)ge + \varepsilon_3 \qquad \text{(Model 3)}$$

With ge, a vector of G*E deviations with covariance proportional to $I_t \otimes A$. However, this model does not provide for a detailed analysis of G*E. To integrate G*E in the mixed model analysis, it is most straightforward to consider G*E as a lack of genetic correlation between environments (Falconer and Mackay 1996). Then, performances in different environments are different traits. This model can be written as

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (I_t \otimes I_m)\eta + \varepsilon_4 \qquad \text{(Model 4)}$$

where η is the vector of environment-specific effects for each genotype with covariance $G \otimes A$. G is the covariance matrix of genotype effects in environments with dimensions $t \times t$. G can be estimated using a factor analytic model, for example (Piepho 1998; Burgueño et al. 2008; Kelly et al. 2009; Cullis et al. 2010). However, Model 3

and Model 4 do not allow prediction of G*E in unobserved environments.

Introducing in the model differential genotype sensitivity to specific stress covariates can capture part of the G*E interaction and allow prediction of performance in unobserved environments. This class of model is termed factorial regression model (Denis 1988; Van Eeuwijk et al. 1996).

For the balanced case, the model can be written in matrix notation:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + (S \otimes I_m)\gamma + \varepsilon_5 \quad \text{(Model 5)}$$

where S is a matrix of dimensions $t \times q$ which contains the centered and scaled observed scores of the q stress covariates in each of the t environment, and γ is the vector, mq long, of stress-specific sensitivities for each genotype with covariance $H \otimes A$, where H is a $q \times q$ covariance matrix of the stress covariates. The definition of an adequate covariance structure for γ is a problem noted in the literature (Smith et al. 2005). Here, the GBLUP framework was extended to factorial regression. This gave additive genotype sensitivity to any given covariate, which is of interest for breeding as well as providing a way to cope with unbalanced data.

Model 5 can be seen as a random regression model, with u random intercept for each genotype and γ genotype-specific random slope. Then, the assumption that the average of the regression slopes across genotypes is zero is not realistic. The assumption could be relaxed by including a fixed effect for each stress covariate. However, here, the stress covariates are confounded with the environments such that the environment effect β captures the mean regression of genotypes on the stress covariates.

For the Model 5 to be scale invariant to linear transformations of the stress covariates, it would be necessary to include a covariance term between u and γ, which would further increase its complexity. The lack of scale invariance means that the model is limited when making inferences about the variance components or testing for fixed effects. The limitation noted, for example, in Smith et al. (2005) remained, as often any given stress covariate explains only a small proportion of the G*E interaction, and a large number of variance components has to be estimated. Considering an equivalent model at the marker level instead of the genotype was investigated as a way to simplify Model 5.

Factorial regression at the QTL level

The factorial regression framework was extended at the QTL level, by modeling each QTL effect as a combination of a main effect and a function of the stress covariates (Crossa et al. 1999; Malosetti et al. 2004), with application in Boer et al. (2007). The design matrix for the linear sensitivity of the markers to each of the stress covariates is in the balanced case: $S \otimes V$, with V the design matrix for all n markers. Combining Model 2 and Model 5, we obtain:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + (S \otimes V)\varphi + \varepsilon_6 \quad \text{(Model 6)}$$

where α is a marker effect vector, and φ is a vector of linear sensitivities of the markers to the stress covariates. Note that φ contains $q \times n$ parameters, so that this model has high dimension. Model 6 can be interpreted as a large penalized regression. However, different levels of shrinkage are desirable for each group of variable (α, β, φ). It is expected that the φ (marker sensitivities to the environment) would have to be shrunk more than the α (main marker effects). One way to solve this problem in a single-step analysis would be to use the sparse group lasso (SGL) (Friedman et al. 2010b) and defining groups of variables for α, β, and φ to provide differential shrinkage of the different groups of variables and enforce model sparseness.

Two-step approach

Algorithms for the SGL are not as computationally efficient as lasso methods, and this became a major hurdle when dealing with thousands of predictors and thousands of observations. Initial testing with the R package *scoop* (Chiquet et al. 2012) showed that the SGL was too slow to be a practical method with a dataset as large as ours. We, therefore, adopted a two-step approach where the main marker and environment effects were first computed using the Bayesian Lasso (Park and Casella 2008) implemented in the R package BLR (Pérez et al. 2010). Residuals from the main effect model (Model 2), and the corresponding β and α were extracted. The residuals of Model 2 ε_2 are deviations from an additive model. They contain G*E deviations in addition to random error. This residual extraction corresponded to step 3 of Fig. 3. At this point, there is only one group of effects, φ, to be estimated and this can be done using a simple penalized regression method. To gain full equivalence with a single-step analysis, the residuals ε_2 should be regressed not on the predictors for φ but on predictors corrected for the main marker effects. However, Model 6 is not a simple multiple regression model, because the factorial regression model involves only a regression of the residual from additivity on the environmental covariates. As a consequence, it does not seem strictly necessary to correct the predictors for φ. For step 4 of Fig. 3, the predictors for the $(S \otimes V)\varphi$ term in (Model 6) were regularized on the residuals of the main effect model with an elastic net (Zou and Hastie 2005). It relies on a combination of both the L_1 (lasso) and L_2 (ridge) norm penalties. For the regression problem $\varepsilon_2 = \mu + (S \otimes V)\varphi + \varepsilon$, The estimator of φ is

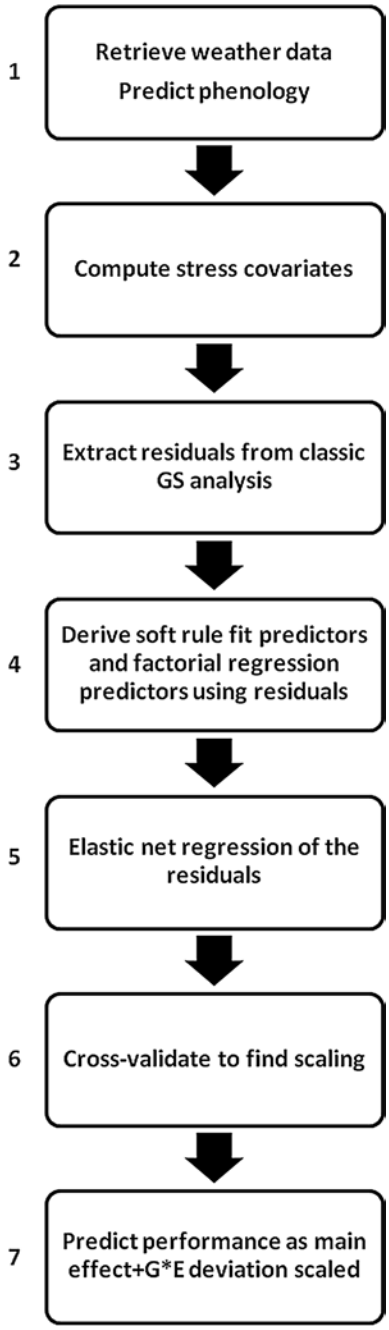

Fig. 3 Modeling flow diagram

$$(1+\lambda_2)\arg\min_{\varphi}\Big([\varepsilon_2-(S\otimes V)\varphi]'[\varepsilon_2-(S\otimes V)\varphi] + \lambda_2\|\varphi\|_2^2+\lambda_1\|\varphi\|_1\Big),$$

With λ_1 and λ_2 shrinkage parameters. This was implemented using the R package glmnet (Friedman et al. 2010a). The shrinkage coefficients were estimated from the data using cross-validation. Fitting these residuals corresponded to step 5 in Fig. 3. Using elastic net had the advantage of avoiding strong assumptions about the optimal model sparseness between a lasso and a ridge regression. For the elastic net properties to hold, all predictors $S \otimes V$ were centered and scaled prior to the analysis.

Because the prediction for the markers main effect α and the predicted G*E deviation $V\varphi$ were both generated by penalized regression methods, they had to be rescaled before combining them to predict the phenotype. Model 6 can then be rewritten as:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta(S \otimes V)\varphi + \varepsilon_7 \quad \text{(Model 7)}$$

where θ is a scaling parameter which was determined by cross-validation within the training set to maximize phenotypic prediction accuracy within environments. This optimal θ was further used with the full model fitted on the whole training set to predict genotype performance in unobserved environments. Thus, the phenotypic performance in environment i was predicted as $V\alpha + \theta\left(b_i^t S\right) \otimes V\varphi$. b_i is a column vector with the ith element equal to one and zero elsewhere. This two-step approach is very similar to the back-fitting algorithm for generalized additive models proposed by Breiman and Friedman (1985). It was also suggested by (Gianola et al. 2006) for combining a classic additive model with a non-parametric component in a mixed model to predict, for example, milk production in cows. It involves several iterations of the process of fitting the model terms sequentially on the residuals of the previous terms. This is in essence the procedure we use here.

Selection of a marker subset to use for factorial regression

Fitting linear marker sensitivity to each covariate would require as many predictors per marker, as there are covariates. To reduce the dimensionality of the problem, a subset of markers was selected as follows. In each environment, marker effects were computed separately as in Heslot et al. (2013). Using the Bayesian Lasso (Park and Casella 2008) implemented in the R package BLR (Pérez et al. 2010), the model was run for 60,000 iterations, and the first 20,000 were discarded as burn-in, and the chains were not thinned. Model convergence was visually assessed based on the trace of parameter samples over iterations.

The variance of marker effects across environments was computed based on the table of marker effects in each environment. Markers were ranked accordingly and sets of different size S of the most variable markers across environments were used to build predictors for the factorial regression at the marker level. This is an important difference from previous approaches taken, for example, in Boer et al. (2007), as they constructed factorial regression predictors at the marker level only for the QTL detected in at

least one environment. The optimal number of markers s to be included in the model was determined by cross-validation. Model 7 then became:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta(S \otimes V_s)\varphi + \varepsilon_8$$

(Model 8)

V_s is a subset of $V$the marker design matrix, containing a subset of S markers selected as described above.

An ensemble method to model complex responses of QTLs to stresses

An important limitation of the factorial regression method is the difficulty to properly model non-linear responses of QTL to stress covariates (Van Eeuwijk et al. 2005). To our knowledge, this modeling has only been attempted in biparental QTL mapping and with one covariate (Ma et al. 2002).

From physiology knowledge, Model 7 is expected to be inadequate because it is linear, but the relationship between the response and the predictors is expected to be non-linear. The response could be approximated using polynomials or splines, but this would require a very large number of predictors and is impractical. In a more general case, Model 8 can be rewritten as:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta f(S, V)\xi + \varepsilon_9$$

(Model 9)

Here the G*E response is determined by a function $f(.)$ which depends on the genotype and the stress covariates. ξ corresponds to effects of the predictors from $f(S, V)$. Model 9 reduces to Model 8 by setting $f(S, V)\xi = (S \otimes V_s)\varphi$. This function is expected to be complex but could be approximated using machine learning techniques suited for non-linear problems. It is also expected that linear response of the markers to the environment would be able to capture a large part of the response of the underlying QTL to stresses. $f(.)$ can then be rewritten as $f(S, V)\xi = (S \otimes V_s)\varphi + g(S, V)\phi$, with φ effects of the factorial regression predictors and ϕ effects of the predictors from $g(S, V)$. As previously, a two-step approach was used. Estimation of β and α was performed using (Model 2) as for the two-step factorial regression. This restricted the machine learning task to the most complex part of the problem, the estimation of $g(S, V)$ from Model 9. Minimizing machine learning enables the use of classic predictors for β and α with well-known optimality and properties. To approximate $g(S, V)$, a non-linear function of unknown form with sparsity, we used soft rule fit, a modified version of the ensemble method RuleFit (Friedman and Popescu 2008; Akdemir and Heslot 2012). Soft rule fit has demonstrated good predictive ability for a number of machine learning tasks. It can capture non-linearity in the data as well as interactions between predictors (Friedman and Popescu 2008; Akdemir and Heslot 2012) in a sparse model without specifying a priori the predictors for all the interactions or the shape of the response.

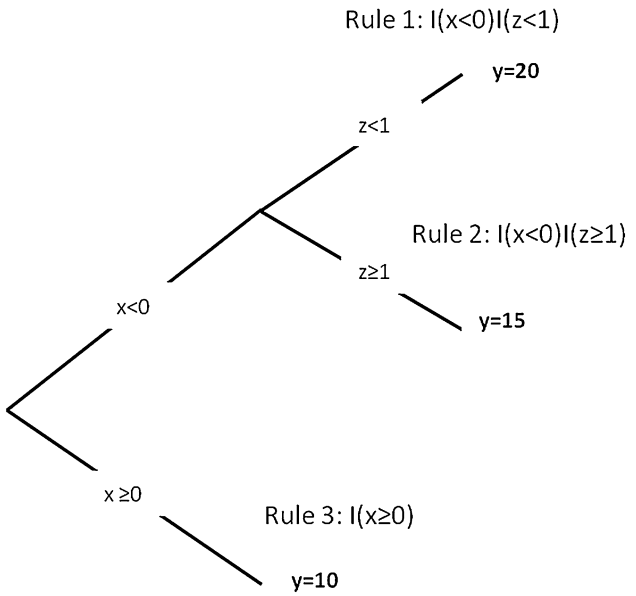

Fig. 4 A simple regression tree built on a bootstrap sample of observations and variables. Each leaf node defines a rule which can be expressed as a product of indicator functions of half spaces. An indicator function takes the value 1 if the condition it takes as input is true else it takes the value 0. Each rule specifies a "simple" rectangular region in the input

Ensemble learning is a relatively new approach to modeling, providing solutions to complex problems by combining simultaneously a number of models (Friedman and Popescu 2003). One of these methods, random forest (Breiman 2001) has already seen some applications in genetics (Bureau et al. 2005; Ogutu et al. 2011). Instead of identifying a single best performing model, the idea is to generate a very large number of predictors built on bootstrap samples of observations and variables. Those predictors are combined together using averaging (random forest) or penalized regression methods (Friedman and Popescu 2008) on the complete dataset.

This approach requires the definition of a family of models, used to generate predictors on bootstrap samples of observations and variables. Soft rule fit uses regression trees to generate predictors. Regression trees are a classic data-mining method that partitions the data into sets, each of which are simply modeled using regression methods. The key aspect for the application here is that it groups observations based on the response variable and the predictors in a non-linear way.

Figure 4 gives a graphical representation of a regression tree built on a bootstrap sample of observations and

variables. For the response variable y and the predictors x and z, the regression tree algorithm identifies "splitting rules," defining nodes that partition the data. Figure 4 shows that the algorithm determined that the greatest variance reduction on the sample was obtained by dividing the data into two subsets based on whether x was positive or not. It was further identified that for x negative, the data were best modeled depending on whether z was superior or not to 1. This defined a complete partition of the observations based on response variable and predictors. This partition can be summarized by three binary rules that indicate to which group an observation belongs.

Soft rule fit uses those rules to derive a probabilistic assignment of each observation to a group using logistic regression. Each rule then provides a predictor taking values between 0 and 1 (Step 4 on Fig. 3). The complexity of the rules is measured by the number of variables involved in a given rule. It is simply controlled by the number of observation groups allowed in the regression tree algorithm. This provides a simple way to control the level of complexity captured by the model as it fixes an upper bound on the number of variables involved in a given predictor.

By repeated sampling of the observations and variables, a matrix W of the soft rule predictors was obtained. W has dimension $mt \times h$ with h number of derived rules. W was further used as predictor of the residuals of the main effect model alone or with the factorial regression predictors described above and regressed on the residual of the main effect model using elastic net (Fig. 3 step 5).

Model 9 can then be rewritten as follows:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta[(S \otimes V_s)\varphi + W\phi] + \varepsilon_{10} \quad \text{(Model 10)}$$

With ϕ effects of the soft rule fit predictors. Model 10 can be fitted without the soft rule fit predictors (in which cases it reduces to Model 8) or without the factorial regression predictors W.

Model 10 was fitted in a two-step procedure as described for models 6 and 7. Briefly, $(S \otimes V_s)$ and W were simultaneously regressed with an elastic net on the residuals in model 2 to obtain φ andϕ. θ, the scaling parameter, was then obtained by cross-validation on the training data. Higher prediction accuracy with the soft rules predictors included in the model (Model 10) compared to (Model 8) would provide evidence of non-linear effects.

About 5,000 initial rules were derived from the data combining markers and stress covariates using the RuleFit algorithm (http://www-stat.stanford.edu/~jhf/r-rulefit/rulefit3/R_RuleFit3.html). The maximum number of groups allowed in the regression trees was three or four. Rules were derived from bootstrap samples of the data using as response variable the residuals from Model 2. As a consequence stress covariates or markers alone might be associated with an apparent main effect on the bootstrap sample, even if overall the data are corrected for the main environment and genotype effect. Of the rules generated, a large number combined only markers or only stress covariates. Such rules were removed to keep only rules combining one or several covariates with one or several markers, thus ensuring that both genotype and environment affected the rule.

Evaluation of model performance

To evaluate the performance of the models presented here, the accuracy was defined as the correlation between the predicted performance and the observed performance in a given environment. The main interest was in the capacity to discriminate between genotypes in unobserved environments. This is a simple way to assess the capacity of the model to capture G*E. To perform a valid statistical test, the dataset was split randomly into two sets of 22 environments, balanced across years and locations with 4,184 observations in the training set and 4,840 observations in the validation set. There were 2,195 genotypes in the validation set of which 544 were absent from the training set. The 22 validation environments were then predicted, and the accuracy computed for each environment. The predictive ability of the model can then be assessed by the mean cross-validated accuracy across environments. Given this set-up, pairs of accuracies in the validation set are independent conditional on the training set and can be used to assess statistical significance of accuracy differences between models. For the pairs of correlations to be strictly (unconditionally) independent, separate sets of training environments would be required for each pair of correlations. Instead, each pair of correlations derives from the same set of training environments. Therefore, the pairs of correlations are only independent conditional on the chosen set of 22 training environments. Thus, the statistical inference we can make is limited to the chosen set of 22 training environments. We cannot extend inferences to freshly chosen sets of training environments. This inference is quite limited and is justified by the data we show. What is tested is whether the model with weather data would do better than the baseline model on a new environment. A paired Wilcoxon rank sum test was used to assess the significance of the difference of the mean accuracies between models. The variance of the accuracy in the validation environments is an unbiased estimator of the variance of prediction accuracy and was reported as the coefficient of variation.

Inference about the genetic architecture of G*E

If the modeling of Q*E with stress covariates captures a significant part of the G*E variance and increases the model predictive power, it can be used to infer the genetic architecture of G*E. An importance measure can be derived for each rule and factorial regression predictor (Friedman and Popescu 2008) from model 10. All predictors were standardized to unit variance and were continuous such that the importance measure was simply the absolute value of the coefficient of each G*E predictor (columns of $S \otimes V$ or W). From the importance of each of the predictors, the importance of the input variables (markers and stress covariates) can be derived as proposed by Friedman and Popescu (2008). It is computed as the sum of the importance of the predictors (soft rules and factorial regression) in which a given input variable appears, divided by the number of input variables involved in each predictor, such that, input variables involved in a predictor equally shared in its importance. The importance of the kth stress covariate is then written:

$I_k = \sum_{i=1}^{s} \varphi_{101(k-1)+i}/2 + \sum_{j=1}^{p} \phi_j / r_j$ with p numbers of rules including the kth stress covariate and r_j the number of variables included in the jth predictor including the kth stress covariate. $\varphi_{101(k-1)+i}$ indicates the $101(k-1)+i$ elements of φ. A coefficient of one half is used for the factorial regression predictors, because each of them has two input variables. Additionally, because each stress covariate was present in the model with three maturity-level parameterizations, importances were summed per stress covariate across those levels. The significance of the importance of the stress covariates was tested by permutations of the stress covariates between environments 100 times to generate a null distribution of the importance measure for each covariate. This permutation was done using the rules discovered on the non-permuted data to limit the computational load required. Using these rules produce a more stringent test than generating new rules using permuted data, because some information from the real data is retained in these rules. Stress covariates were permuted in blocks between environments such that the covariance structure between stress covariates was preserved and each stress covariate had a single value in each environment. Then, soft rules and factorial regression predictors were generated using the permuted stress covariates and regressed on the residuals (Fig. 3, step 5). A given stress covariate importance was considered significant, if it was larger than the greatest importance obtained for that covariate with the permutations.

Model 10 allowed prediction of performance of any genotype in any environment based on the stress covariates and the markers Then, the G*E prediction term $(S \otimes V_s)\varphi + W\phi$ is an estimate of η, the environment-specific effect for each genotype from Model 4, even for unobserved environments. This estimate of η can be further used to estimate G the covariance matrix of genotype effects in environments from Model 3. For the best predictive model, the G*E term was predicted for all 2,437 genotypes in all 44 environments. The derived table of predicted G*E response was used to estimate G. This corresponds to the predicted levels of genetic correlation between environments. Environments were clustered based on this covariance matrix using unweighted pair-group average agglomerative hierarchical clustering. Because of the machine learning predictors, there is no closed form estimate of G based on model coefficients as in random regression.

Results

Mixed model results

Using ASreml-R (Gilmour et al. 2009), Model 3 was fitted to generate simple variance component estimates. The additive genetic variance was estimated to be 6.5, the environment variance 172.2, the G*E variance 11.2, and the error variance 67.7. Heritability was computed using a generalized heritability measure suitable for unbalanced data using the predicted error variance from the mixed model (Piepho and Möhring 2007) (formula 20) and was equal to 0.54.

Using ASReml-R, the size of the factorial regression problem (Model 5) quickly became intractable. It was not possible to fit genotype-specific sensitivity on the full dataset, even for one stress covariate, as the model required more than 250 GB of RAM (Model 5) when covariance structures were included for u and γ. Fit was nevertheless possible when no covariance was included for u and γ. Focusing instead on marker sensitivities to the covariates seemed a suitable approach to decrease the dimensionality of the problem and it took advantage of powerful penalized regression methods (Model 6).

Marker variability

Focusing on marker sensitivity created a dimensionality problem. In the case of linear sensitivity, 1,29,987 additional predictors (1,287 markers × 101 stress covariates) would have to be fitted in the model. To overcome that limitation, subsets of markers with particularly variable effects across environments were selected. The histogram plotted on a log scale in Fig. 5 suggested that a few markers were extremely variable, the most variable being the marker for *Ppd-D1*, the main photoperiod sensitivity locus, with a variance of 0.043. The other markers with the largest variance were not associated with any of the known major adaptation loci, despite the inclusion of diagnostic markers for those loci in the analysis.

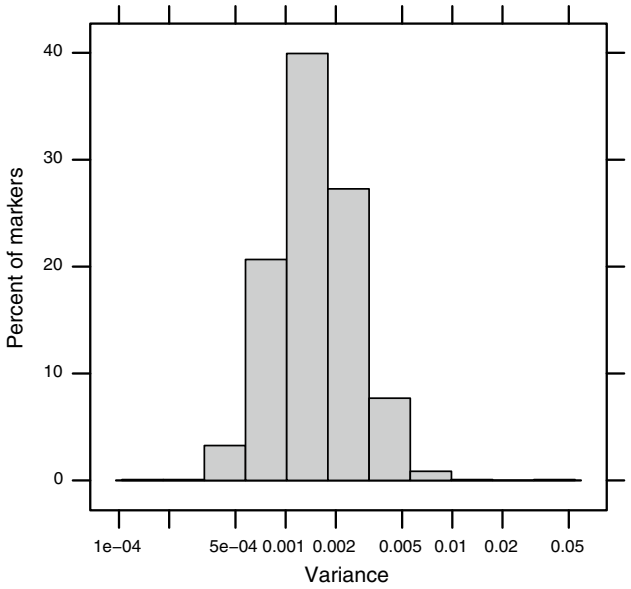

Fig. 5 Distribution of the variance of marker effects, computed in each environment, across environments, plotted on a log scale

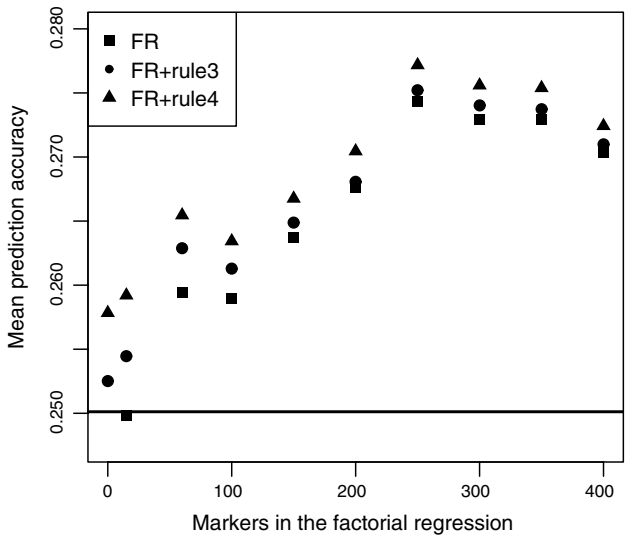

Fig. 6 Predictive performances of the models as a function of the number of markers for which a linear sensitivity to each covariate is fitted from the cross-validation. Three scenarios are plotted, with no inclusion (*square*) or inclusion of soft rule fit predictors of order three (*circle*) or order four (*triangle*). Predictive performance is measured as the mean prediction accuracy for the 22 environments removed from the training set. The first points with 0 abscissa correspond to a model with rules only

Two-step approach results

Despite an appealing simplicity, using the SGL to perform, a one-step analysis was too computationally intensive to be feasible. Thus, all the results presented here are from the two-step approach (Model 10).

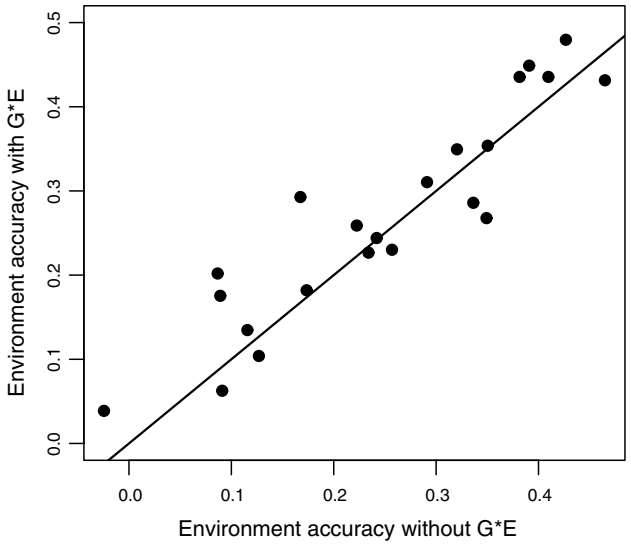

Fig. 7 Comparison of the prediction accuracy in each predicted environment from the cross-validation between the main effect model, and the model with rules of order four and 250 markers for the factorial regression added to the main effect prediction. The line indicates the identity

Figure 6 presents the mean prediction accuracies for different models capturing G*E compared to the base model with no modeling of G*E included in cross-validation. Results indicated a gain in mean prediction accuracy (0.25–0.277) compared to the base model accuracy when predictors of G*E were included. Most of the gain in accuracy came from the linear response of markers to the stress covariates. With all three models considered, accuracy reached a maximum when 250 markers were included for factorial regression. The best model in cross-validation comprised 1,584 soft rules, each containing at least one stress covariate. The best model (with soft rules of order four and 250 factorial regression predictors) provided an 11.1 % increase in accuracy and a 10.8 % decrease in accuracy coefficient of variation over Model 2. On the same cross-validation settings, the G*E model captured on average 3.7 % of the variance of residuals of the base model in each validation environment. Inclusion of the soft rule fit predictor improved accuracy slightly, for any number of markers included. The best model (with soft rules of order four and 250 factorial regression predictors) was significantly better than a model with 250 factorial regression predictors, and no rules included (P value $= 0.093$). This indicated that part of the G*E response was due to a non-linear response of the QTL to the environment. When stress covariates were permuted between environments, including a G*E term decreased the predictive ability of the model. This result rules out overfitting by the model.

Figure 7 presents the detailed cross-validation results for the model with rules of order four and 250 markers in

the factorial regression. The line corresponds to the identity such that if a validation environment is over the line, there is a prediction gain by modeling G*E for that environment. Residuals of a few environments were poorly predicted by the model, as expected under the Anna Karenina effect.

Inference of the genetic architecture of G*E

All the following results are based on the model with rules of order four and 250 markers for the factorial regression, which was the most predictive model in cross-validation.

Six hundred eighty-nine markers (53.5 % of all markers) had an importance different from 0, and thus were included in the model to predict G*E. However, inclusion does not mean significance.

The most important marker was the marker for *ppd-D1*, the photoperiod sensitivity locus with an importance twice the importance of the second most important marker. Ninety-nine percent of this importance was due to the factorial regression predictors and not to the rules, indicating that the effect of the *ppd-D1* locus on yield changes linearly with stress. This marker was also the most variable marker across environments. However, apart from *ppd-D1*, there was no correlation between variability across environment and importance based on the factorial regression. This could be explained by stresses not taken into account by the covariates, or this can point out the inefficiency of the marker selection procedure. The *ppd-D1* photoperiod insensitive allele had a mean frequency of 51.8 % in the different environments with a minimum of 8.6 % and a maximum of 76.8 %. When *ppd-D1* was fitted alone in a factorial regression model, the cross-validated accuracy was equal to the accuracy obtained when fitting Model 1 alone.

Other perfect markers for vernalization, photoperiod sensitivity, or dwarf status had little or no importance. The other most important markers did not correspond to known loci affecting phenology. There was no correlation between the marker importance and their main effect on yield. Similarly the marker importance was not correlated to their main effect on heading date (data not shown).

All stress covariates were included in the soft rule terms. This is evidence of the complexity of the G*E response, as it involves all the stress covariates in the model covering abiotic stresses over the whole plant cycle. Despite this overall complexity, only ten stress covariates had an importance larger than the importance observed in 100 permutations and are presented in Table 2.

From those most important stress covariates, a clear picture emerges about the stress creating the most G*E on winter wheat in France. Almost all the stress covariates presented in Table 2 related to stresses before flowering such as drought stress in early spring (ntddr, spetpe1) and heat stress before flowering (stmpmf, st25ef, nd25ef). Heat stress at the beginning of grain filling was also important (st25fh, nd25fh). Latitude captured a North/South gradient in weather patterns. This is evidence that the most critical stage for G*E and abiotic stress sensitivity unexplained by geography are stresses before flowering and not the late stage stresses.

Table 2 Importance of the eight stress covariates with a significant importance, for the model with rules of order 4 and 250 markers for the factorial regression. Importance is rescaled to give a score of 100 to the largest one

Stress covariate	Stress type	Importance
stmpmf	Sum temperature meiosis to flowering	100.00
ntddr	Drought early spring	88.97
spetpe1	Drought early spring	84.99
st25ef	Heat stress before flowering	79.69
nd25ef	Heat stress before flowering	77.53
st25fh	Heat stress early grain filling	76.25
nd25fh	Heat stress early grain filling	71.65
Latitude	North/South trend	29.52

As the model enabled the prediction of part of the G*E response in each environment based on markers and stress covariates, fitted values were calculated for all genotypes all environments in the dataset. However, this could also be done for any environment with daily weather data. Those predicted values could be used to study the stability of genotypes in a set of environments. Consequently, for each of the 2,437 genotypes, the variance of the predicted G*E response in a set of environments was compared to the main genotype effects, and there was no correlation. Those fitted values were also used to compute a predicted G*E correlation matrix between environments. This correlation matrix corresponds to the genetic correlation between environments as captured by the model (Fig. 8).

The correlation matrix indicated divergent G*E response between environments with a wide range of correlations between environments. Several clusters of environments corresponding mostly to year were identifiable on the dendrogram (Fig. 8). Those clusters were also confirmed by looking at the heatmap of the correlation matrix (Supplementary Figure S1). They corresponded mostly to a clustering by year. Clear clusters can be identified for 2007, 2008, and 2011. The cluster of environments on the left of the dendrogram spanned several years and corresponded mostly to locations in the south of France. A similar cluster analysis was performed directly using the stress covariates produced a pattern that was less differentiated by year and some outlier environments that disappeared in the clustering approach based on predicted G*E (data not shown).

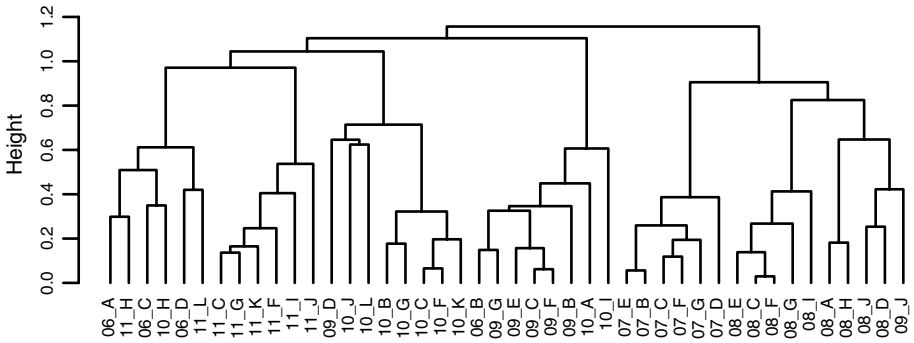

Fig. 8 Hierarchical agglomerative clustering of the environments based on the predicted G*E response of all genotypes. Environments were named using the last two digits of the year followed by a location code

Discussion

A disappointing gain in accuracy?

Our model was able to predict part of the G*E response of genotypes in unobserved environments. Though the gain was statistically significant, it was small. Figure 7, however, indicates that gains were much larger in some environments, especially environments where prediction accuracy was low with the baseline model. In a survey of the G*E literature, we found only two papers (Burgueño et al. 2011, 2012) concerned with predicting G*E to report cross-validation results. Papers we identified using factorial regression such as Crossa et al. 1999 usually only reported a model fit to the whole dataset and did not assess prediction accuracy.

The average gain we observed (11.1 %) was higher than the one reported in Burgueño et al. (2011). They reported a gain of 6 % on average across six datasets when using a factor analytic model. Analysis of results presented in Burgueno et al. (2012) suggested an average accuracy gain of 7.2 % (0.44–0.471) when predicting genotypes absent from the training set and a gain of 19.6 % (0.474–0.55) when predicting genotypes present in the training set in a different environment using a factor analytic model and the realized relationship matrix. For their model to be predictive, some genotypes needed to be observed in the environment to be predicted.

Here we are predicting G*E deviation for unobserved environments which is clearly a more difficult task. The cross-validation setting is rather stringent here with the training dataset effectively reduced to half the size of the total dataset. With more data, in particular more environments, we would expect better performance. We also show that the gain in accuracy is statistically significant and that when using permuted covariates, the model has no predictive power. These results together indicate that the model is picking up real G*E signal in the data.

It is also important to remember that the reported accuracies are correlations between predicted values and phenotypes measured in a single environment with a low number of replicates and consequently a low repeatability. Based on the variance components estimated from Model 3, the within environment repeatability should be approximately $\left(\sigma_u^2 + \sigma_{ge}^2\right) \Big/ \left(\sigma_u^2 + \sigma_{ge}^2 + \sigma_e^2\right) = 0.207$ with σ_u^2, the additive genetic variance, σ_{ge}^2 the G*E variance, and σ_e^2 the error variance estimated with Model 3. So the maximum accuracy would be about 0.455 if all genetic and G*E variance was predicted.

Strategies integrating statistical and crop growth models for phenotype prediction

Levins (1966) states that in building models, there is a tension among the goals of realism, generality, and accuracy making it impossible to create a model that fulfills all goals simultaneously. Purely statistical models (e.g., regression and machine learning) sacrifice realism and generality in favor of accuracy. These models minimize prediction error, but usually cannot be extrapolated to conditions outside those previously observed, and their parameters have little interpretive value relative to the underlying biology of the problem. They are pure black box models. Crop growth models, in contrast, sacrifice accuracy in favor of a certain level of realism and generality. These models do not completely lack predictive power. They can indicate which conditions will increase or decrease performance and serve best to provide qualitative rather than quantitative, error-minimizing, predictions. For the problem of G*E prediction in breeding, we require accuracy because we will seek to select on the basis of model predictions. Generality is also required because the environments that interest us are those of the future and are therefore, by definition, unobserved. Finally, while model realism is probably relegated to the lowest priority, we do not want to discard it entirely because interpretation of the model can guide future experimental and selection efforts.

In broad terms, we place crop growth and statistical models in series with the outputs of the former providing the inputs of the latter. Thus, we aim for the crop growth model to provide some level of generality and realism, first

by condensing massive quantities of weather variables in a limited number of covariates and second by tying those covariates to phenologically relevant and interpretable plant stresses. The statistical model must then provide predictive accuracy. We note that placing the models in the opposite order (statistical then crop growth) is also a possibility. In that case, the statistical model would predict crop growth model parameters, and the crop growth model would then combine those with weather data to produce a prediction. For that strategy, the crop growth model parameters are used as traits (Reymond et al. 2004; Reymond et al. 2003), which would require specific phenotyping experiments. The strategy that we use requires only phenotypic data generated by a breeding program for the purpose of selection. In addition, as discussed above, there are concerns that crop models are not sensitive enough to capture the subtle performance differences between elite genotypes (White et al. 2008).

In this paper, we integrated environment data in the analysis using a crop model, as a tool to generate metadata about the trial that included phenology. This is a key point because phenology data are not usually collected in plant breeding trials, with the possible exception of heading date. Furthermore, the determination of most of the developmental stages is difficult and labor intensive. Once development stages are known, agronomy and plant physiology knowledge can be leveraged to define stress covariates by stage. This has multiple advantages because it reduces the dimensionality of the data to a few dozen covariates. It also enables use of the large datasets generated by plant breeding activities to study G*E. While a major data reduction was achieved, there were still more predictors (environmental covariates) than observations (environments). Here, only weather data were used, but the framework developed could accommodate other kinds of environment variables such as soil quality types and disease pressure. This type of data reduction strategy with further use in QTL mapping was first proposed by Boer et al. (2007). Here, we extended it to a large number of stress covariates and genome-wide markers while capturing non-linearity of responses.

Inference about the genetic architecture of G*E

The best performing model included predictors for linear responses to the stress covariates as well as soft rule fit predictors capturing non-linearity. This suggests that part of the G*E is not amenable to modeling by a linear response. In addition, the use of a crop model and the definition of stress covariates by growth stage also captured some non-linearity. For example, heat stress is expected to be critical at flowering and less so earlier in the cycle. This creates major non-linearity issues in using weather data for modeling directly. The use of stress covariates provides a simplification of the problem which is difficult to quantify.

Assuming that the model captured enough G*E variance to provide useful insight in the genetic architecture of G*E, the use of biologically meaningful stress covariates facilitated the interpretation of the model. The significantly important stress covariates were related to radiation and water stress before flowering rather than to terminal stresses. This has important implications because it suggests that breeding efforts for stress tolerance should focus more on those specific stresses. Alternatively, as terminal stress was expected a priori to be important, it could suggest that the breeding program from 2006 to 2011 did not sample environments with terminal stress. Despite an expected large North–South G*E pattern, results indicated that annual rather than latitudinal variation was more important.

One of the most important loci for G*E was *Ppd-D1*, a photoperiod sensitivity locus, indicating that a major determinant of the G*E response in this dataset is phenology. However, *Ppd-D1* alone did not capture a significant part of the G*E variance. Other important loci for G*E did not correspond to any of the other known major adaptation genes, and these loci warrant further investigation. Results suggest that Q*E is pervasive and characterized by small interactions among large numbers of regions of the genome and a large number of stresses as expected under the Anna Karenina effect. The importance of the markers for G*E prediction was not related to their main effect, and the genotype main effect was not related to the variability of G*E response. These results also suggest that genotypes can be selected for stability without penalizing performance. However, the model does not explain a large part of the G*E variance. If the alternative hypothesis of a correlation between main and interaction effects holds, we do not know what power we would have to detect it. Stability was defined here as the variance of the predicted G*E for each genotype. Other definitions are possible. Stability is not necessarily a desirable trait if it means consistently poor performance. Capacity of genotypes to take advantage of good growing conditions is a favorable trait which is potentially associated with performance instability according to the definition we used.

Our results indicated that the QTLs causing the most Q*E have small main effects. This suggests that focusing on markers with large overall main effects when trying to identify Q*E is inefficient.

We hypothesize that those QTLs with a large Q*E effect on yield are not likely to be detected in mapping experiments across environments because their measured main effect will not be consistent across environments. If they are detected within environment, they are unlikely to be validated in separate mapping experiments in different

environments. If identifying consistent QTL main effects is the goal, those QTLs might only be detected when focusing directly on an underlying physiology or plant architecture trait. For example, in this dataset, *ppd-D1* did not have a consistent main effect on yield across environments, but would have been detected if the mapping focused on the main effect of photoperiod sensitivity.

A new tool to deal with G*E in breeding programs

Historical weather data or predicted weather data from climate change models could be used in simulations to investigate the target population of environments (TPE). The TPE is the mixture of environments expected for the intended region of production (Comstock 1977). In most cases, the composition of the TPE is unknown. Simulation studies show that in the case of cross-over G*E, it is beneficial to weight the trials by their expected frequency of occurrence in the TPE (Podlich et al. 1999). Our approach provided an accessible way to determine those frequencies. By predicting genotype performance using historical weather records, the frequency of occurrence of the clusters identified in Fig. 8 can be calculated. This could be used to optimize the phenotypic testing strategy. Most of the environment clustering we observed was by year suggesting that our sample of environments, while large, was not large enough to cover all expected environment types. Using this data only provided a partial glimpse of the TPE. This interpretation supposes that the locations are a spatially representative sample of the TPE.

Using the predicted G*E response instead of the stress covariates to cluster environments allowed clustering on the predicted level of genetic correlation between environments, which is the parameter of interest for breeding purposes. This was possible even for environments with no phenotypic data. Multiple stresses were considered, and none of them was suspected to be the main cause of G*E. Consequently, it was not meaningful to directly use the stress covariates to group environments as in previous studies (Chapman et al. 2000a, 2000b, 2000c). In their studies, the main cause of G*E was drought stress, such that environments could be clustered based on the pattern of drought stress. Using the predicted G*E response also captured non-linear responses of genotypes to stresses and threshold effects.

By leveraging agronomic knowledge and the large historical datasets generated by breeding programs, this new model provides insight into the genetic architecture of genotype by environment interactions and predicts genotype performance based on past and future weather scenarios. The model can therefore provide a better knowledge of the current and future TPE. This knowledge should translate to an improved design of phenotypic testing strategies.

Acknowledgments We thank Pierre Martre for providing the crop model. The reviewers provided excellent comments that significantly improved the paper. JRC-MARS—Meteorological Data Base—EC–JRC provided access to the interpolated meteorological data. This research was supported in part by USDA-NIFA-AFRI grants, award numbers 2009-65300-05661, 2011-68002-30029, and 2005-05130 and by Hatch project 149-449. Limagrain Europe provided financial support for N. Heslot.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The experiments comply with the current laws of the countries in which they were performed.

References

Akdemir D, Heslot N (2012) Soft rule ensembles for statistical learning. Arxiv Prepr Arxiv 1205:4476

Boer MP, Wright D, Feng L et al (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. doi:10.1534/genetics.107.071068

Brancourt-Hulmel M, Lecomte C, Meynard JM (1999) A diagnosis of yield-limiting factors on probe genotypes for characterizing environments in winter wheat trials. Crop Sci. doi:10.2135/cropsci1999.3961798x

Brancourt-Hulmel M, Denis JB, Lecomte C (2000) Determining environmental covariates which explain genotype environment interaction in winter wheat through probe genotypes and biadditive factorial regression. Theor Appl Genet. doi:10.1007/s001220050038

Breiman L (2001) Random forests. Mach Learn. doi:10.1023/A:1010933404324

Breiman L, Friedman J (1985) Estimating optimal transformations for multiple regression and correlation. J Am Stat Assoc 80:580–598

Bureau A, Dupuis J, Falls K et al (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol. doi:10.1002/gepi.20041

Burgueño J, Crossa J, Cornelius PL, Yang RC (2008) Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. Crop Sci. doi:10.2135/cropsci2007.11.0632

Burgueño J, Crossa J, Cotes JM et al (2011) Prediction assessment of linear mixed models for multienvironment trials. Crop Sci. doi:10.2135/cropsci2010.07.0403

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 Sci. doi:10.2135/cropsci2011.06.0299

Chapman SC, Cooper M, Butler D, Henzell R (2000a) Genotype by environment interactions affecting grain sorghum I. Characteristics that confound interpretation of hybrid yield. Aust J Agric Res. doi:10.1071/AR99020

Chapman SC, Cooper M, Hammer G, Butler D (2000b) Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust J Agric Res 51:209–221

Chapman SC, Hammer G, Butler D, Cooper M (2000c) Genotype by environment interactions affecting grain sorghum III. Temporal sequences and spatial patterns in the target population of environments. Aust J Agric Res. doi:10.1071/AR99022

Chenu K, Chapman SC, Hammer G et al (2008) Short-term responses of leaf growth rate to water deficit scale up to whole-plant and crop levels: an integrated modelling approach in maize. Plant Cell Environ. doi:10.1111/j.1365-3040.2007.01772.x

Chenu K, Deihimfard R, Chapman SC (2013) Large-scale characterization of drought pattern: a continent-wide modelling approach applied to the Australian wheatbelt—spatial and temporal trends. New Phytol. doi:10.1111/nph.12192

Chiquet J, Grandvalet Y, Charbonnier C (2012) Sparsity with sign-coherent groups of variables via the cooperative-lasso. Ann Appl Stat. doi:10.1214/11-AOAS520

Comstock RE (1977) Quantitative genetics and the design of breeding programs. In: Pollak E, Kempthorne O, Bailey TB (eds) Proceedings of the international conference on quantitative genetics. Iowa State University Press, Ames, pp 705–718

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

Crossa J, Vargas M, Van Eeuwijk FA et al (1999) Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. Theor Appl Genet. doi:10.1007/s001220051276

Cullis BR, Smith AB, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome. doi:10.1139/G10-080

DeLacy IH, Basford KE, Cooper M et al (1996) Analysis of multi-environment trials—an historical perspective. In: Cooper M, Hammer G (eds) Plant adaptation and crop improvement. CAB International, Wallingford, pp 39–124

Demotes-Mainard S, Doussinault G, Meynard JM (1996) Abnormalities in the male developmental programme of winter wheat induced by climatic stress at meiosis. Agronomie. doi:10.1051/agro:19960804

Denis JB (1988) Two-way analysis using covariates. Statistics 19:123–132

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Pearson Prentice Hall, Harlow

Fischer RA (1985) Number of kernels in wheat crops and the influence of solar radiation and temperature. J Agri Sci. doi:10.1017/S0021859600056495

Friedman J, Popescu BE (2003) Importance sampled learning ensembles. J Mach Learn Res 94305:1–32

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann Appl Stat 2:916–954

Friedman JH, Hastie T, Tibshirani R (2010a) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1

Friedman JH, Hastie T,Tibshirani R (2010b) A note on the group lasso and a sparse group lasso. Arxiv Prepr Arxiv:10010736

Gallagher JN, Biscoe PV (1978) Radiation absorption, growth and yield of cereals. J Agri Sci. doi:10.1017/S0021859600056616

Gate P (1995) Ecophysiologie du blé. De la plante à la culture. Tec & Doc, Paris, p 430

Gauch HG (2006) Statistical analysis of yield trials by AMMI and GGE. Crop Sci. doi:10.2135/cropsci2005.07-0193

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedure. Genetics. doi:10.1534/genetics.105.049510

Gilmour AR, Gogel B, Cullis BR, et al (2009) ASREML user guide release 3.0. VSN International Ltd.

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics. doi:10.1534/genetics.107.081190

Hammer G, Kropff MJ, Sinclair TR, Porter JR (2002) Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur J Agron. doi:10.1016/S1161-0301(02)00093-X

He J, Le Gouis J, Stratonovitch P et al (2012) Simulation of environmental and genotypic variations of final leaf number and anthesis date for wheat. Eur J Agron. doi:10.1016/j.eja.2011.11.002

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci. doi:10.2135/cropsci2009.11.0662

Heslot N, Jannink J-L, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci. doi:10.2135/cropsci2012.07.0420

Hunt LA (1991) Post anthesis temperature effects on duration and rate of grain filling in some winter and spring wheats. Can J Plant 617:609–617

Jamieson PD, Semenov MA, Brooking IR, Francis GS (1998) Sirius: a mechanistic model of wheat response to environmental variation. Eur J Agron. doi:10.1016/S1161-0301(98)00020-3

Jullien A, Mathieu A, Allirand JM et al (2011) Characterization of the interactions between architecture and source-sink relationships in winter oilseed rape (*Brassica napus*) using the GreenLab model. Ann Bot-Lond. doi:10.1093/aob/mcq205

Kelly AM, Cullis BR, Gilmour AR et al (2009) Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet Sel Evol. doi:10.1186/1297-9686-41-33

Landau S, Mitchell RA, Barnett V et al (1998) Testing winter wheat simulation models' predictions against observed UK grain yields. Agric Forest Meteorol. doi:10.1016/S0168-1923(97)00069-5

Landau S, Mitchell RA, Barnett V et al (2000) A parsimonious, multiple-regression model of wheat yield response to environment. Agric Forest Meteorol. doi:10.1016/S0168-1923(99)00166-5

Lecomte C (2005) Experimental evaluation of varietal innovations. Proposition of genotype—environment analysis tools adapted to the diversity of needs and constraints of the professionals of the seeds industry. Diss AgroParisTech p 262

Levins R (1966) The strategy of model building in population biology. Am Sci 54:421–431

Löffler CM, Wei J, Fast T et al (2005) Classification of maize environments using crop simulation and geographic information systems. Crop Sci. doi:10.2135/cropsci2004.0370

Lorenz AJ, Chao S, Asoro FG et al (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron. doi:10.1016/B978-0-12-385531-2.00002-5

Ma CX, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161:1751–1762

Malosetti M, Voltas J, Romagosa I et al (2004) Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica. doi:10.1023/B:EUPH.0000040511.46388.ef

Martre P, Jamieson PD, Semenov MA et al (2006) Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. Eur J Agron. doi:10.1016/j.eja.2006.04.007

Messina C, Hammer G, Dong Z et al (2009) Modelling crop improvement in a GXEXM framework via gene-trail-phenotype relationships. In: Sadras VO, Calderini D (eds) Crop physiology: applications for genetic improvement and agronomy. Elsevier, Netherlands, pp 235–265

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

Meynard JM, Sebillotte M (1994) L'élaboration du rendement du blé, base pour l'étude des autres céréales à paille. In: Picard D,

Combe L (eds) Elaboration du rendement des principales cultures annuelles. INRA, Paris, pp 31–51
Monteith J (1972) Solar radiation and productivity in tropical ecosystems. J Appl Ecol 9:747–766
Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. doi:10.1186/1753-6561-5-S3-S11
Park T, Casella G (2008) The bayesian lasso. Am Stat Assoc. doi:10.1198/016214508000000337
Pérez P, De los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R. Plant Gen. doi:10.3835/plantgenome2010.04.0005
Piepho HP (1998) Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor Appl Genet. doi:10.1007/s001220050885
Piepho HP, Möhring J (2006) Selection in cultivar trials—is it ignorable? Crop Sci. doi:10.2135/cropsci2005.04-0038
Piepho HP, Möhring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. Genetics. doi:10.1534/genetics.107.074229
Piepho HP, Denis JB, Van Eeuwijk FA (1998) Predicting cultivar differences using covariates. J Agric Biol Environ Stat. doi:10.2307/1400648
Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica. doi:10.1007/s10681-007-9449-8
Piepho HP, Ogutu JO, Schulz-Streeck T et al (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci. doi:10.2135/cropsci2011.11.0592
Podlich DW, Cooper M, Basford KE (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. Plant Breed. doi:10.1046/j.1439-0523.1999.118001017.x
Quilot B, Génard M, Kervella J, Lescourret F (2004) Analysis of genotypic variation in fruit flesh total sugar content via an ecophysiological model applied to peach. Theor Appl Genet. doi:10.1007/s00122-004-1651-7
Reymond M, Muller B, Leonardi A et al (2003) Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol. doi:10.1104/pp.013839.soil
Reymond M, Muller B, Tardieu F (2004) Dealing with the genotype x environment interaction via a modelling approach: a comparison of QTLs of maize leaf length or width with QTLs of model parameters. J Exp Bot. doi:10.1093/jxb/erh200
Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agric Sci. doi:10.1017/S0021859605005587
Sofield I, Evans L, Cook M, Wardlaw I (1977) Factors influencing the rate and duration of grain filling in wheat. Aust J Plant Physiol. doi:10.1071/PP9770785
Stone P, Nicolas M (1998) The effect of duration of heat stress during grain filling on two wheat varieties differing in heat tolerance: grain growth and fractional protein accumulation. Aust J Plant Physiol. doi:10.1071/PP96114
Tashiro T, Wardlaw I (1990) The response to high temperature shock and humidity changes prior to and during the early stages of grain development in wheat. Aust J Plant Physiol. doi:10.1071/PP9900551
Van der Goot E, Orlandi S (2003) Technical description of interpolation and processing of meteorological data in CGMS. Joint Research Centre of the European Commission, Ispra, Italy, p 23
Van Eeuwijk FA, Denis J-B, Kang MS (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environments tables. In: Kang MS, Gauch HG (eds) Genotype-by-environment interaction. CRC Press, Boca Raton, pp 15–50
Van Eeuwijk FA, Malosetti M, Yin X et al (2005) Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. Aust J Agric Res. doi:10.1071/AR05153
White JW, Herndl M, Hunt LA et al (2008) Simulation-based analysis of effects of loci on flowering in wheat. Crop Sci. doi:10.2135/cropsci2007.06.0318
Windhausen VS, Wagener S, Magorokosho C et al (2012) Strategies to subdivide a target population of environments: results from the CIMMYT-led maize hybrid testing programs in Africa. Crop Sci. doi:10.2135/cropsci2012.02.0125
Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growth stages of cereals. Weed Res. doi:10.1111/j.1365-3180.1974.tb01084.x
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. doi:10.1111/j.1467-9868.2005.00503.x