Modelling $G\times E$ with historical weather information improves genomic prediction in new environments

Jussi Gillberg[1*], Pekka Marttinen[1], Hiroshi Mamitsuka[1,2] & Samuel Kaski[1*]

1. Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, PO Box 15400, 00076 Aalto, Finland.

2. Institute for Chemical Research, Kyoto University, Gokasho, Uji 6110011, Japan

Interaction between the genotype and the environment ($G\times E$) has a strong impact on the yield of major crop plants. Although influential, taking $G\times E$ explictily into account in plant breeding has remained difficult. Recently $G\times E$ has been predicted from environmental and genomic covariates, but existing works have not shown that generalization to new environments and years without access to in-season data is possible and practical applicability remains unclear. Using data from a Barley breeding program in Finland, we construct an in-silico experiment to study the viability of $G\times E$ prediction under practical constraints. We show that the response to the environment of a new generation of untested Barley cultivars can be predicted in new locations and years using genomic data, machine learning and historical weather observations for the new locations. Our results highlight the need for models of $G\times E$: non-linear effects clearly dominate linear ones and the interaction between the soil type and daily rain is identified as the main driver for $G\times E$ for Barley in Finland. Our study implies that genomic selection can be used to capture the yield potential in $G\times E$ effects for future growth seasons, providing a possible means to achieve yield improvements, needed for

feeding the growing population.

Global yield improvements are needed to feed the growing population [1]. One possibility is to breed varieties for higher environmental adaptability, known as *targeted breeding* [2]. By improving the genetic fit of varieties in their growth environments, yield potential in the interaction between the genotype and environment could be realised. While the importance of $G \times E$ for agronomic performance is widely accepted, utilisation calls for methods that predict yields in new environments, because actual experimental data, consisting of yields of plant variety candidates from yield trials, will in practice be available from only a very limited number of environments. Importantly, prediction of a plant's response to a new environment cannot be based on weather data from the growth season, as those will never be available at the time of prediction.

Methods for "cold start" prediction problems [3], where predictions are needed for completely novel instances, have been developed within the machine learning community. Example applications include design of novel drugs for previously unseen cancers [4], and recommendations in on-line shopping for new customers and/or products [3]. These methods are based on using external covariate data that describe properties of the novel instances. We develop a new method, an extension of the Kernelized Bayesian Matrix Factorization [5], to account for the uncertainty in the covariates, which allows the use of historical records to predict weather conditions for future growth seasons, and eventually makes future $G \times E$ prediction for yield possible. Therefore, our new method, unlike the existing alternatives [6–8], does not rely on accurate weather information from the growth season from the new location (Figure 1).

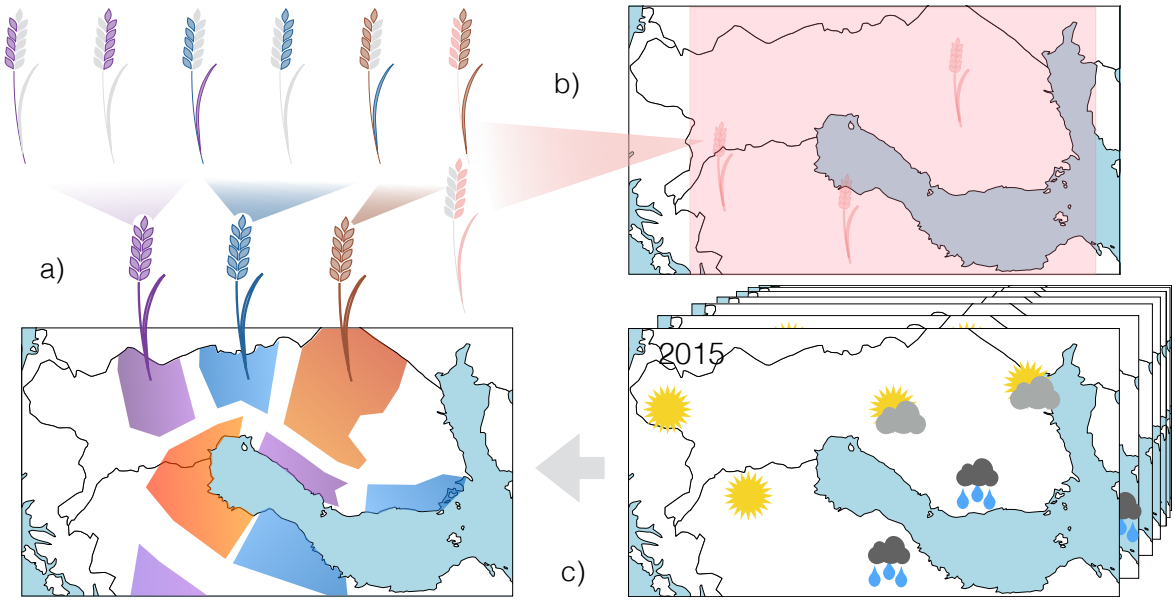


Figure 1: Outline of our approach. a) Precision breeding aims at producing varieties that are optimal for a specific environment. As compared to traditional breeding (b), targeted breeding aims at higher environmental adaptation, i.e., smaller target environments. Weather (microclimate) is a crucial driver for agronomic performance, but as it is unknown for future growth seasons, we use historical weather records (c) to predict the environmental stresses. The growth locations differ with respect to their estimated probabilities of extreme conditions and our method can be used to manage risks by trading-off yield potential for stress tolerance, when the risk in a particular environment is elevated.

In genomic selection (GS) [9], field trials are replaced with genomic predictions to speed-up plant breeding. We formulate an *in silico* experimental setup for GS in targeted breeding that, unlike existing works [7,8,10–12], strictly satisfies all realistic constraints: test locations, years, and genotypes are all genuinely new (not part of the training set) and yields are predicted for the offspring of the training set. In this setup, we demonstrate the feasibility of targeted breeding by investigating the accuracy of $G\times E$ prediction using environmental data including historical weather information but without in-season data (Model M^{hist}_{G+E+GE}). We compare this with multiple competing settings, including the non-realistic ideal situation having in-season data (M_{G+E+GE}), a

model without the $G\times E$ interaction (M_{G+E}), a previous implementation with $G\times E$ interactions using in-season data (`GE-BLUP`) [8], and the industry-standard that does not include $G\times E$ (best linear unbiased prediction using genomic data [13]; `GBLUP`). Data from a barley breeding program in Finland from Boreal Plant Breeding Ltd, including historical weather information for the target environments, are divided into training, validation, and test sets, and the prediction accuracy is measured as the average correlation between predicted and observed yields in the test sets [8] (Figure 2c). A sensitivity analysis is done to explore the impact of model assumptions. A description of the model and the setup can be found in Materials and Methods, and further details are given in the Supporting Information.

Modelling $G\times E$ with historical weather data, M^{hist}_{G+E+GE}, improves predictive accuracy as compared to the industry-standard, `GBLUP` (Figure 2a; p=0.011, a two-sided paired Wilcoxon signed rank test, df=17). The improvement is comparable to using in-season data (M_{G+E+GE}, p=0.023). The Bayesian methods in general show higher accuracy whereas `GE-BLUP` performs poorly with the data available. Overall, the absolute prediction accuracy of all methods was relatively low in this challenging setup, with M^{hist}_{G+E+GE} having the highest correlation of 0.105. Nevertheless, the improvement is considerable over the industry-standard with correlation 0.077, with the proposed new method explaining 85% more of the variation of the phenotype on average.

The sensitivity analysis demonstrates considerable variability between test environments (Figure 2c). Indeed, including $G\times E$ interaction terms into the model decreased accuracy in 1/18 environments, had little effect in 11/18 environments, but improved the accuracy substantially in

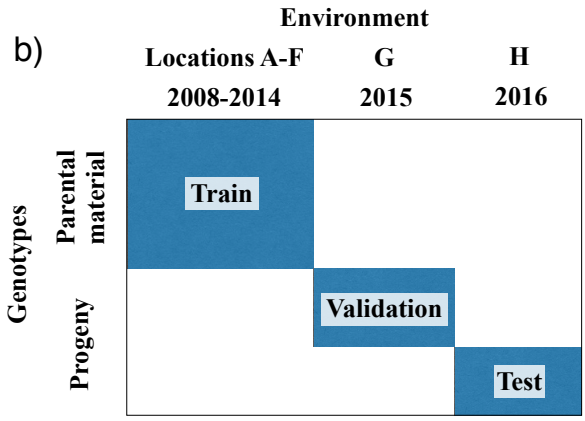
6/18 environments. In the last group, increasing model complexity by adding more $G\times E$ components consistently improved performance, which highlights the potential to increase accuracy through complex modelling of $G\times E$. Importance of different data sources to the predictions can be further analysed by investigating the weights of the different kernels, used to summarise the data sources (Figure S 3). We see that the two most influential kernels were the ones that represented 1) the non-linear interaction between soil type and daily rainfall, and 2) the non-linear effect of rain, matching well the biological understanding of the problem.

Our experiments confirmed that prediction in new environments is a challenging task, as reported earlier [8], our method reaching the highest correlation of 0.105 between predictions and observations. Nevertheless, the usefulness of including multiple $G\times E$ interaction terms and non-linear interactions between environmental covariates became evident from our results. We expect that gains from modelling $G\times E$ will increase in the future as more data, representing further locations and years, will allow more accurately distinguishing the interactions from the main effects. Other ways to improve the predictions inlcude using more detailed genomic modeling, e.g. using Gaussian and other kernels for summarizing the SNP data.

Besides targeted breeding, there are several other needs for $G\times E$ prediction models. They could mitigate the problems of conventional breeding: accounting for historical weather in the actual target population of environments can help prevent overfitting to the conditions in the few field trials performed, as discussed in detail in SI Gains from modelling $G\times E$ for current target population of environments. The assumption of the match between field trials and actual growing

a)

Model	mean	SE_{mean}	P
M_{G+E+GE}	0.104	0.036	0.011
$M^{\text{historical}}_{G+E+GE}$	0.105	0.034	0.023
GE−BLUP	0.016	0.026	0.132
GBLUP	0.077	0.034	N/A
M_{G+E}	0.097	0.035	0.109



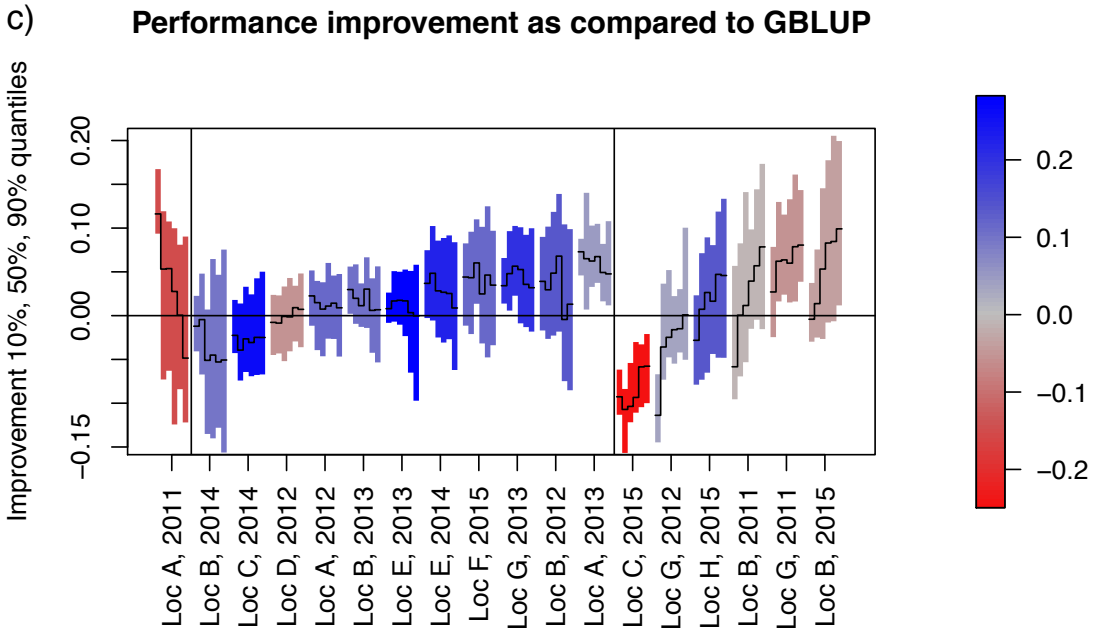


Figure 2: Predicting $G\times E$ with historical weather information improves genomic prediction accuracy in strictly new environments. **a)** Comparison of prediction accuracies; *mean*: correlation between predicted and observed yields, averaged across test environments; SE_{mean}: standard error of the mean; P: p-value compared to the industry standard (GBLUP). **b)** Outline of the *in silico* setup for comparing methods. **c)** Sensitivity analysis: the difference in prediction accuracies (y-axis) between $G\times E$ prediction with historical data (M^{hist}_{G+E+GE}) and the industry standard (GBLUP) is shown in 18 different environments (x-axis); values above the horizontal line mean that M^{hist}_{G+E+GE} is more accurate. Six vertical bars are shown for each environment, representing variability in results (median and 90 % confidence intervals). Starting from the left, they correspond to models with 0, 1, 2, 3, 4 or 5 $G\times E$ interaction terms (0 corresponds to the M_{G+E} model). The color indicates the performance of GBLUP in the environment, red meaning GBLUP performed poorly (Loc C, 2015 were omitted from the comparison as all methods performed poorly there). Vertical lines divide the environments into three groups: $G\times E$ *decreased*: including $G\times E$ terms to the model decreased performance; $G\times E$ *neutral*: 10 environments where $G\times E$ terms had neutral effect; $G\times E$ *increased*: 6 environments where performance increased by adding more $G\times E$ terms.

89 locations is equally crucial for the official variety trials for value of cultivation and use (VCU),
90 required in most countries to evaluate new varieties. $G\times E$ models are also needed in assessing the
91 effects of climate change and to select for varieties that react favourably to the altering conditions
92 [1]. For this purpose, the historical weather observations in M^{hist}_{G+E+GE} can be replaced with climate
93 simulations to assess the performance of varieties under various climate scenarios. To summarise,
94 we showed that $G\times E$ prediction in the setup required by targeted breeding, where the environ-
95 ments are strictly new and predictions are based on historical weather data available at the time of
96 prediction, improves prediction accuracy significantly compared to the industry standard, which is
97 needed to accelerate the implementation of targeted breeding.

98 **Methods**

99 **Data** All data used in the experiment come from a barley breeding program in Finland, which is
100 a part of a larger population of target environments for barley as varieties used in Finland are also
101 used in other Nordic countries. The phenotype consists of (z-transformed) yield measurements
102 (kg/ha) for 2,244 lines observed in trials at 11 locations across the 4 southernmost growth zones
103 in Finland from 2008 to 2015. The total number of observed $location \times year$ combinations is
104 35. In some locations, trials have been performed on several years and several fields with varying
105 soil properties, and a total of 12,277 yield observations have been recorded. The number of ob-
106 servations per genetic line ranges from 1 to 118 (median 4). The lines were genotyped with the
107 Illumina 9k iSelect SNP Chip, SNPs with minor allele frequence (MAF) < 0.05 or with $> 5\%$
108 values missing were omitted. Also all genotypes with $> 5\%$ of SNPs missing were omitted. The

final proportion of missing genotype data is 0.002.

The soil characteristics for each field block are measured in terms of the proportions of sand, silt and clay (*soil classification triangle* [14]) and the proportion of organic content. Meteorological information consists of daily averages of temperature and rainfall, and the distances to the closest meteorological station range from 1 to 40 km (average 13.5 km). The baseline approach `GE-BLUP` [8] requires summarising the weather information per crop stage: vegetative (from sowing to visible awns), heading time (from visible awns to the end of anthesis), and grain filling (from the end of anthesis to maturity). The times of the crop stages are estimated using temperature sum accumulation; the details are given in Section Comparison methods. In the weather observations, the proportion of missing values in daily average temperature and rainfall measurements is < 0.0015 (max 3 missing values/environment) and < 0.0032 (max 2 missing values/environment), respectively.

Experimental setup. To study prediction accuracy, we use a setup that strictly imposes the realistic constraints related to modelling $G \times E$ in targeted breeding for new locations. Predictions are required for new locations (not part of the experimental grid) and for years for which no phenotype data are available (to mimic future growth seasons). Additionally, predictions are needed for the offspring of the lines in the training set, which have no phenotype data observations. More details with a summary of the differences between our setup and earlier works are given in SI Details of experimental setup. We measure prediction accuracy using cross-validation, where the training, validation and test sets are selected to enforce the realistic constraints (Figure 2c). As the

performance measure for prediction accuracy, we follow the conventional approach, i.e., the Pearson correlation between the predicted and observed yields in the test set [8,10–12]. This correlation is computed for each cross-validation fold in turn, and averaged over the test cases. Similarly to Malosetti et al. [8], the test case -specific correlations are transformed into Fisher's z-scores before averaging and back-transformed to obtain the final results. We regress the $G\times E$ interactions on the average characteristics of the growing season: for each yield trial, we use the weather observations from the typical growing season (1$^{\text{st}}$ of May until end of August) regardless of the sowing date. This indirect approach allows predicting with historical weather data. When predicting with historical data, the prediction for each genotype is made for each year for which historical weather observations are available, and the median of those is used as the final predicted value.

We also carry out a sensitivity analysis that allows studying the impact of modelling assumptions, such as inclusion of $G\times E$ interaction components to the model. In detail, the sensitivity analysis shows variability (median and 90% interval) in the predictive performance in a given environment (location, year combination) when we vary $i)$ the hyperparameter values over their spesified ranges, $ii)$ the genotype sets that we are predicting, and $iii)$ the training set by removing any single training environment.

Model. In the models M_{G+E}, M_{G+E+GE} and M^{hist}_{G+E+GE} we assume that *i*) the yield y_{ij} of genotype i in environment j is affected by the genotype, the environmental conditions throughout the growing season and the interactions between the two. We assume that *ii*) the response to the environmental properties is non-linear and that *iii*) it may involve interactions between differ-

ent environmental properties. For instance, temperature/rainfall either too low or too high reduces yield, and the response to rainfall is also affected by the soil type. We further assume that *iv*) the responses to the environmental conditions are highly polygenic. Assumptions *i*-*iv* are encoded using the *kernel trick* [15], in which covariate data are represented as similarities, or kernels, between different data items. Kernel methods are a computationally effective way to model non-linearities and interactions and they have been applied to breeding data [16]. An additional complication in the data is the low number of observed trials compared to the complexity of the problem. To handle this, we constrain our model to only learn the most prominent combinations of environmental conditions affecting yield, by assuming a low-rank approximation for the model parameters accounting for the $G\times E$ effects. Finally, we follow the Bayesian statistical framework [17], and regularise the model by placing priors on all parameters, which alleviates overfitting to the training data and improves prediction accuracy in the test data.

Mathematically, the model for yield is formulated as

$$y_{ij} = g_i + e_j + \xi_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, N_g, j = 1, \ldots, N_e, \tag{1}$$

where g_i is the genetic main effect, e_j is the environmental effect, ξ_{ij} is the effect that arises from interaction between genotype i and environment j, ϵ_{ij} is noise distributed as $N(0, \sigma_j^2)$, and N_g and N_e are the numbers of genotypes and environments. The genetic main effect g_i is modeled as a linear function of the genomic covariates. In detail, the model for the vector of genetic main effects

$\mathbf{g}^* = (g_1, \ldots, g_{N_g})^T$ is given in terms of a linear genomic kernel K_g by

$$\underset{N_g \times 1}{\mathbf{g}^*} = \underset{N_g \times N_g}{K_g} \cdot \underset{N_g \times 1}{\mathbf{a}_{g0}} + \underset{N_g \times 1}{\mathbf{e}_{g0}}\,, \tag{2}$$

where $\mathbf{a}_{g0}$ are kernel regression weights and $\mathbf{e}_{g0}$ is the noise vector with elements distributed independently as $N(0, \sigma^2_{g0})$. The dimension of each matrix is shown in equation (2) below the corresponding matrix symbol. The genomic kernel K_g is computed by first concatenating the genomic covariates $\mathbf{g}_i$ as the rows of a matrix $\mathbf{G}$ and then using the standard linear kernel, $K_g = \mathbf{G}\mathbf{G}^T$.

The environmental main effect e_j in equation (1) is modeled as a random effect,

$$e_j \sim N(0, \sigma^2_{e0}), \quad j = 1, \ldots, N_e.$$

The $G \times E$ terms ξ_{ij} are modelled as non-linear functions of the genomic and environmental covariates, $\mathbf{g}_i$ and $\mathbf{e}_j$. Each environment and genotype is first represented by R latent variables. The interactions ξ_{ij} are modelled as the inner product of the latent variable vectors corresponding to genotype i and environment j, that is,

$$\xi_{ij} = \sum_{r=1}^{R} h^g_{ir} \cdot h^e_{jr}, i = 1, \ldots, N_g, j = 1, \ldots, N_e. \tag{3}$$

Here, h^g_{ik} is the kth latent variable for the ith genotype, and h^e_{jk} is the kth latent variable for the jth

177 enviroment. Using matrix notation, equation (3) can be written as

$$\underset{N_g \times N_e}{\Xi} = \underset{N_g \times R}{H_g} \cdot \underset{R \times N_e}{H_e^T}, \tag{4}$$

178 where $\Xi = [\xi_{ij}]$ is the matrix of interaction terms, and $H_g = [h^g_{ij}]$ and the $H_e = [h^e_{ij}]$ are matri-
179 ces having as their rows the R-dimensional latent variable representations for each genotype and
180 environment, respectively.

The latent variables H_g and H_e are obtained from genotype and environment kernels K_g and K_e:

$$\underset{N_g \times R}{H_g} = \underset{N_g \times N_g}{K_g} \cdot \underset{N_g \times R}{A_g} + \underset{N_g \times R}{E_{H_g}} \quad \text{and}$$

$$\underset{N_e \times R}{H_e} = \underset{N_e \times N_e}{K_e} \cdot \underset{N_e \times R}{A_e} + \underset{N_e \times R}{E_{H_e}},$$

181 where A_g and A_e are kernel regression weights, and E_{H_g} and E_{H_e} are matrices containing error
182 terms distributed independently as $N(0, \sigma_g^2)$ or $N(0, \sigma_e^2)$, respectively. The environmental ker-
183 nel K_e is obtained by combining multiple kernels $K_e^1, \ldots, K_e^E$, computed from enviromental data
184 $\mathbf{e}_j, j = 1, \ldots, N_e$, each kernel representing a different aspect of the environment (weather, soil,
185 etc). Details about processing the raw data into kernels and about combining multiple environmen-
186 tal kernels into a single kernel are presented in SI Data processing and kernels.

187 For inference we use variational approximation [18], which is a computationally feasible way
188 to approximate posterior distributions of parameters in complex models. The variational updates

required here can be derived similarly to Gönen et al. [5], except that we have extended their model and algorithm by including the genotype and environment main effects, i.e., the terms g_i and e_j in equation (1). Detailed distributions of the model parameters and the guidelines for specifying hyperparameter values are given in Sections SI Detailed model specification and SI Specifying hyperparameter values, respectively. Further details about the inference algorithm in SI Details of the variational inference algorithm.

Comparison methods. The mixed model computations for the comparison methods `GBLUP` and `GE-BLUP` are performed using the R library `rrBLUP` [19]. For both methods, fixed effects were used to account for field block-specific effects, corresponding to the terms e_j in M^{hist}_{G+E+GE}, M_{G+E+GE} and M_{G+E}. For `GBLUP`, the genomic kernel (see Section Model) was used as the covariance matrix Σ. For `GE-BLUP`, the environmental kinship model (GE-KE) [8], is used and the full covariance matrix Σ is generated through the Kronecker product $\Sigma = \Sigma_G \otimes \Sigma_E$, where Σ_G and Σ_E are the genetic and environmental covariance matrices, respectively. The environmental covariance matrix Σ_E is generated from the available environmental data to describe soil properties and the growth conditions during the vegetative, heading time and grain filling developmental stages. All soil data and growth zone information are used as such whereas the daily average temperature and rainfall measurements are summarised as the mean and the standard deviation of the daily observations per crop stage. The growth periods are estimated using the sowing date and temperature sum accumulation-based estimates of heading and ripening times (440.2 °C and 905.9 °C, respectively), which were estimated from external breeding data. The vegetative stage is assumed to last 3 weeks starting from sowing, the time of heading is assumed to start 2 weeks

before and last 1 week after the estimated heading time and grain filling was assumed to start after heading and to last 1 week longer than the estimated time of ripening. Wide estimates for the growth periods were used to account for varying growth speeds. The resulting set of environmental covariates is z-normalized and a linear kernel is used, which is further normalized according to equation (()) in SI Data preprocessing and kernels.

Data availability

The data accompanied by the method code will be made available upon publication in the form of kernels to allow reproducing the results.

Acknowledgements

We acknowledge Boreal Plant Breeding Ltd for the access to the plant breeding data and we gratefully thank Outi Manninen, Mika Isolahti and Esa Teperi for very helpful comments. This work was in part funded by Tekes, the Finnish Funding Agency for Innovation (Dnro 1718/31/2014 to J.G, H.M and S.K) and the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, and grant numbers 294238, 292334 to SK and grant numbers 286607 and 294015 to P.M.).

Author contributions

J.G. processed the data from Boreal Plant Breeding Ltd and performed the *in silico* experiments. J.G. and P.M. implemented the method. All authors were involved in the conception and design of the study, analyzed the results and assisted with drafting and critically revising the manuscript.

Competing interests

The authors declare no competing financial interests.

1. Tester M, Langridge P (2010) Breeding Technologies to Increase Crop Production in a Changing World. *Science* 327(5967):818–822.

2. Braun HJ, Rajaram S, Ginkel M (1996) Cimmyt's approach to breeding for wide adaptation. *Euphytica* 92(1):175–183.

3. Schein AI, Popescul A, Ungar LH, Pennock DM (2002) Methods and Metrics for Cold-start Recommendations in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02. (ACM, New York, NY, USA), pp. 253–260.

4. Costello JC, et al. (2014) A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology* 32(12):1202–1212.

5. Gönen M, Kaski S (2014) Kernelized Bayesian Matrix Factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(10):2047–2060.

6. Jarquín D, 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.

7. 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.

8. 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.

9. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157(4):1819–1829.

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.

11. Albrecht T, et al. (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. *Theoretical and Applied Genetics* 127(6):1375–1386.

12. Saint Pierre C, et al. (2016) Genomic prediction models for grain yield of spring bread wheat in diverse agro-ecological zones. *Scientific Reports* 6:27312.

13. 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.

14. Shepard FP (1954) Nomenclature based on sand-silt-clay ratios. *Journal of Sedimentary Petrology* 24(3):151–158.

15. Shawe-Taylor J, Cristianini N (2004) *Kernel methods for pattern analysis*. (Cambridge University Press).

16. Gianola D, Morota G, Crossa J (2014) Genome-enabled prediction of complex traits with kernel methods: What have we learned? in *Proceedings, 10th World Congress of Genetics Applied to Livestock Production*. p. 6.

17. Gelman A, et al. (2013) Bayesian data analysis, 3rd edition.

18. Beal MJ (2003) *Variational algorithms for approximate Bayesian inference*. (Gatsby Computational Neuroscience Unit, University College London).

19. Endelman JB (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4(3):250–255.

20. Gönen M (2012) Bayesian Efficient Multiple Kernel Learning. *Proc. 29th International Conference on Machine Learning, ICML 2012* pp. 1–8.

Supplementary Information (SI)

Data preprocessing and kernels. A summary of different kernels, including transformations specific to each data source, preprocessing and kernel transformations used, is given in Table S 1. The bandwidth parameter of all the Gaussian kernels is set to the conventional default value equal to the number of covariates used to compute the kernel. All kernels K are normalized to make them unit diagonal:

$$\widetilde{K} = (\mathbf{d}^{-1/2} \times \mathbf{d}^{-1/2}) \cdot K \tag{1}$$

where $\mathbf{d}$ is a vector of the diagonal values of kernel K, $\times$ denotes the outer product, and the $\mathbf{d}^{-1/2}$ denotes a vector with all elements of $\mathbf{d}$ raised to the power of $-1/2$. The interaction kernel between the soil type and rainfall is computed from other kernels as

$$K_{\text{soil x rain}} = \widetilde{K}_{\text{soil, Gaussian}} \odot \widetilde{K}_{\text{rain, Gaussian}}, \tag{2}$$

where $\odot$ denotes the Hadamard (elementwise) product. Finally, all kernels are normalized with respect to their summed total variance by multiplication with a constant c

$$\widetilde{\widetilde{K}} = c \cdot \widetilde{K} \tag{3}$$

where $c = \left[\sum_{i=1}^{N_z} \text{Var}(\widetilde{\mathbf{k}}_i)\right]^{-1/2}$ and $\widetilde{\mathbf{k}}_i$ is the ith column of $\widetilde{K}$. The motivation for this normalisation comes from the expectation that *a priori* each kernel explains the same amount of variance, and, when combining the kernels as described below, this prior expectation is realised by the nor-

malisation.

Combining environmental kernels. The final environmental kernel K_e is obtained as a weighted sum of the different normalized ($\widetilde{\widetilde{K}}$) kernels in Table S 1. The weights are learnt from the training data by fitting BEMKL [20], a multiple kernel learning regression method, using experiment-specific yield means as the target variable. For BEMKL, shape (α) and scale (β) parameter values for the prior Gamma distributions are set to 1 except for the λ parameter, for which the scale is fixed to 10, providing stronger regularization. Regression bias term b is set to 0. For further details of BEMKL, see Gönen et al. [20]. Before combining the kernels, the learnt weights are normalized such that their sum of squares is equal to 1, and the largest weight (in absolute value) is positive. The distributions of the normalised kernel weights from the sensitivity analysis are presented in Figure S 3. The composite kernel K_e is again normalized according to equation (3).

Detailed model specification. The distributional assumptions of the model are

$$y_{ij}|H_g, H_e, g_i, e_j, \sigma_j^2 \sim \mathcal{N}(g_i + e_j + (\mathbf{h}_i^g)^T \mathbf{h}_j^e, \sigma_j^2), \quad \forall(i,j)$$

$$\sigma_j^{-2} \sim \mathcal{G}(\alpha_j, \beta_j), \quad \forall(j)$$

$$a_i^{g0}|\lambda_{g0} \sim \mathcal{N}(0, \lambda_{g0}^{-1}), \quad \forall(i)$$

$$g_i|\mathbf{a}_{g0}, K_g, \sigma_{g0}^2 \sim \mathcal{N}(\mathbf{a}_{g0}^T \mathbf{k}_i^g, \sigma_{g0}^2), \quad \forall(i)$$

$$a_{ij}^g|\lambda_g \sim \mathcal{N}(0, \lambda_g^{-1}), \quad \forall(i,j)$$

Variable (unit)	transformation	preprocessing parameters	missing value imputation	kernel transformation(s)
Soil content (%, N_{covs} =3)	log transformation	z-normalization	(none)	linear and Gaussian
Soil organic content (%, N_{covs} =3)	log transformation	z-normalization	(none)	Gaussian
daily rainfall (mm, N_{covs} =123)	7-day moving average (6 previous days)	z-normalization with 3rd order polynomial smoothing	0-imputation	linear and Gaussian
daily average temperature (C°, N_{covs} =123)		z-normalization with 3rd order polynomial smoothing of daily mean/scale parameters	0-imputation	linear and Gaussian
growth zone (1-4, N_{covs} =1)		z-normalization	(none)	Gaussian
genotype markers (SNPs, N_{covs} =5696)		Minor allele frequency scaling for SNP A: $\frac{\text{A} - 2 \cdot \text{MAF}_\text{A}}{\sqrt{2 \cdot \text{MAF}_\text{A} \cdot (1 - \text{MAF}_\text{A})}}$	mean imputation	linear kernel

Table S 1: Preprocessings and kernel functions applied to covariates.

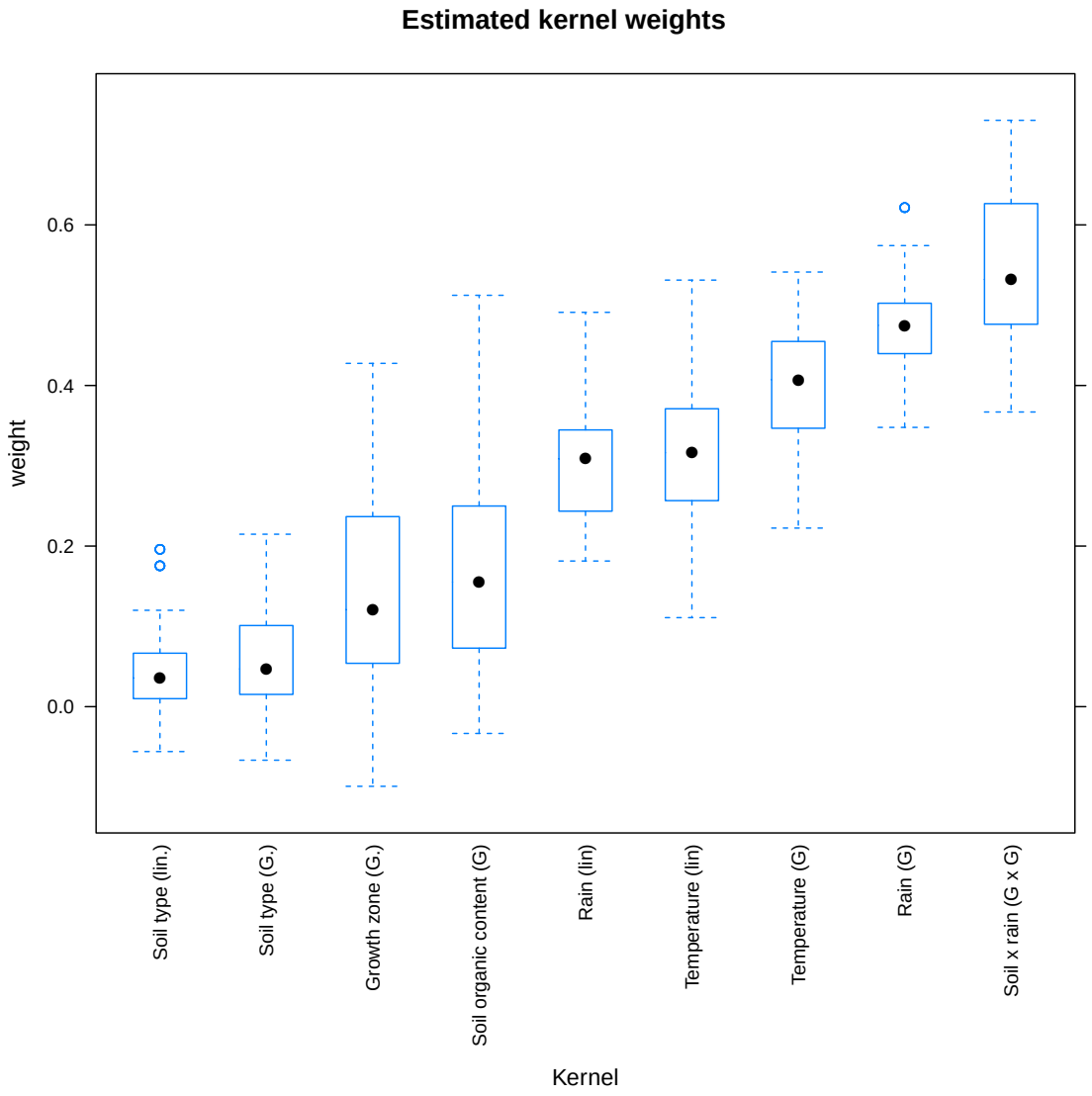


Figure S 3: Estimated normalized kernel weights in the sensitivity analysis.

$$h_{ij}^{g}|A_g, K_g, \sigma_g^2, \sim \mathcal{N}((\mathbf{k}_i^g)^T \mathbf{a}_j^g, \sigma_g^2), \quad \forall(i,j)$$

$$e_j|\sigma_{e0}^2 \sim \mathcal{N}(0, \sigma_{e0}^2), \quad \forall(j)$$

$$a_{ij}^{e}|\lambda_e \sim \mathcal{N}(0, \lambda_e^{-1}), \quad \forall(i,j)$$

$$h_{ij}^{e}|A_e, K_e, \sigma_e^2 \sim \mathcal{N}((\mathbf{k}_i^e)^T \mathbf{a}_j^e, \sigma_e^2), \quad \forall(i,j),$$

where $\mathbf{k}_i^g$, $\mathbf{k}_j^e$, $\mathbf{a}_j^g$, $\mathbf{a}_j^e$, denote columns of matrices K_g, K_e, A_g, A_e, with subscripts i and j specifying the column index; $\mathbf{h}_i^g$ and $\mathbf{h}_j^e$ denote ith and jth rows of H_g and H_e, represented as column vectors; a_i^{g0} is the ith element of vector $\mathbf{a}_{g0}$; a_{ij}^g and a_{ij}^e are the (i,j)th elements in matrices A_g and A_e. $\mathcal{N}$ and $\mathcal{G}$ denote the Gaussian and Gamma distributions, respectively.

Specifying hyperparameter values. Prior knowledge about the approximate weights of different sources of variance, e.g. the relative weight of genetic and environmental main effects, is used to specify hyperparameter values. We determine for each hyperparameter either a single fixed value or a grid of values to be selected from by cross-validation. Parameters (α_j, β_j) of the Gamma distribution for environment-specific residual noise variances σ_j^2 are set to (10, 1), corresponding to an expected value of approximately 0.1 for σ_j^2. The variance of environment mean effects σ_{e0}^2 is fixed to 0.25. To set the parameters λ_{g0} and σ_{g0}^2 that determine the amount of signal and noise in the genetic main effects, we find values for them such that two conditions are satisfied. First, 95% of the variance of the genetic effects $\mathbf{g}^*$ is assumed to be signal, that is,

$$\frac{\text{Var}(K_g \cdot \mathbf{a}_{g0})}{\text{Var}(K_g \cdot \mathbf{a}_{g0}) + \sigma_{g0}^2} = 0.95.$$

The second condition is that the variance of the genetic main effects, $\text{Var}(K_g \cdot \mathbf{a}_{g0}) + \sigma_{g0}^2$, is either 0.2, 0.4, or 0.6. In practice we find these values for σ_{g0}^2 and λ_{g0} by simulating multiple realisations from the model with specific values for the parameters, and select values that on average satisfy the two conditions.

The parameters $\lambda_g, \sigma_g^2, \lambda_e$, and σ_e^2, controlling the proportion of signal and noise in the la-

tent components H_g and H_e that model the $G\times E$ interactions, are selected according to similar principles: by inspecting the proportion of signal of the total variance of the latent factors and the relative contribution of the interaction terms compared to the genetic main effects. In detail, we assume first that

$$\frac{\text{Tr}(\text{Var}(K_g \cdot A_g))}{\text{Tr}(\text{Var}(K_g \cdot A_g)) + R\sigma_g^2} = 0.95, \text{ and}$$
$$\frac{\text{Tr}(\text{Var}(K_e \cdot A_e))}{\text{Tr}(\text{Var}(K_e \cdot A_e)) + R\sigma_e^2} = 0.95,$$

where Tr() denotes the trace of a matrix. Second, we assume that the total variance of the interactions is either the same or half of the total variance from the genetic main effects, i.e.

$$\text{Tr}(\text{Var}(H_g \cdot H_e^T)) = \Phi \times R \times [\text{Var}(K_g \cdot \mathbf{a}_{g0}) + \sigma_{g0}^2],$$

where Φ is either 0.5 or 1, to be selected with cross-validation.

Details of the variational inference algorithm. For short-hand, the hyper-parameters in the model are denoted jointly by

$$\zeta = \{\alpha_j, \beta_j, \sigma_{g0}^2, \sigma_{e0}^2, \sigma_g^2, \sigma_e^2, \lambda_{g0}, \lambda_g, \lambda_e\},$$

and the parameters by

$$\Theta = \{\mathbf{a}_{g0}, A_g, A_e, H_g, H_e, \mathbf{g}^*, \mathbf{e}^*, \sigma_*^2\},$$

where $\sigma_*^2 = (\sigma_1^2, \ldots, \sigma_{N_e}^2)$. In the following the dependence on ζ is omitted for clarity. We assume the factorized variational approximation

$$p(\Theta|K_g, K_e, Y) \approx q(\Theta) = q(\mathbf{a}_{g0})q(A_g)q(A_e)q(H_g)q(H_e)q(\mathbf{g}^*)q(\mathbf{e}^*)q(\sigma_*^2)$$

and define each factor in the ensemble just like its full conditional:

$$\begin{aligned}
q(\mathbf{a}_{g0}) &= \mathcal{N}(\mathbf{a}_{g0}; \mu(\mathbf{a}_{g0}), \Sigma(\mathbf{a}_{g0})) \\
q(A_g) &= \prod_{r=1}^{R} \mathcal{N}(\mathbf{a}_r^g; \mu(\mathbf{a}_r^g), \Sigma(\mathbf{a}_r^g)) \\
q(A_e) &= \prod_{r=1}^{R} \mathcal{N}(\mathbf{a}_r^e; \mu(\mathbf{a}_r^e), \Sigma(\mathbf{a}_r^e)) \\
q(H_g) &= \prod_{i=1}^{N_g} \mathcal{N}(\mathbf{h}_i^g; \mu(\mathbf{h}_i^g), \Sigma(\mathbf{h}_i^g)) \\
q(H_e) &= \prod_{j=1}^{N_e} \mathcal{N}(\mathbf{h}_j^e; \mu(\mathbf{h}_j^e), \Sigma(\mathbf{h}_j^e)) \\
q(\mathbf{g}^*) &= \prod_{i=1}^{N_g} \mathcal{N}(g_i; \mu(g_i), \Sigma(g_i)) \\
q(\mathbf{e}^*) &= \prod_{j=1}^{N_e} \mathcal{N}(e_j; \mu(e_j), \Sigma(e_j)) \\
q(\sigma_*^2) &= \prod_{j=1}^{N_e} \mathcal{G}(\sigma_j^{-2}; \alpha(\sigma_j^{-2}), \beta(\sigma_j^{-2})).
\end{aligned}$$

The parameters in the factor distributions can be derived as by Gönen et al [5], and they are therefore omitted from here.

Initialisation of the variational algorithm. The parameter $\mathbf{g}^*$ was initialised to the main

genetic effects learnt by `GBLUP`, and $\mathbf{e}^*$ was initialised to the average yields in the different environments. Parameters H_g and H_e were initialised by applying the regularized Singular Value Decomposition (SVD) implemented in R library `softImpute` to the yield matrix Y after regressing out the initialised main effects $\mathbf{g}^*$ and $\mathbf{e}^*$. Parameters $\mathbf{a}_{g0}$, A_g and A_e were initialised to 0. Environment-specific residual variance parameters σ^2_* were initialised to environment-specific sample variances.

Details of experimental setup. Different prediction tasks, distinguished by the availability of different data types, are presented in Figure S 4. Setups 1-4 correspond to those studied by Malosetti et al. [8]: in setup 1, phenotype measurements are available for the genotypes and environments to be predicted, and both genotypes and environmental covariates are fully observed. In setups 2 and 3, phenotype measurements are still available but only for the genotypes or the environments to be predicted, but not both, and covariates are fully observed. In setup 4, no phenotype data are available for environments/genotypes to be predicted, but both genetic and environmental covariates are still fully observed.

Two additional setups can be considered. In setups 5 and 6 environmental covariates from the environments of interest are only partially available: location and soil characteristics are known but the in-season weather measurements are not available for the year of interest. However, historical observations for the same locations are available and they are used to estimate the performance of each genotype. Setups 5 and 6 differ depending on whether phenotype measurements are available from some other environment for the genotypes (5) or not at all (6). The results in this paper are for

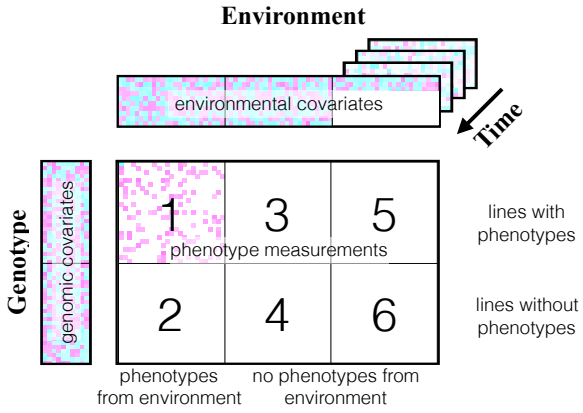


Figure S 4: Comparison of prediction setups with respect to the availability of phenotype data and the genomic and environmental covariates as presented by Malosetti et al [8]. White colour indicates missing value. In setups 1, 3 and 5, "lines with phenotypes", the lines to be predicted have phenotype observations (from some environments). In setups 1 and 2, "phenotypes from environment", phenotypes have been measured from the prediction target environments (for some lines). In setups 1-4 presented by Malosetti et al [8]., environmental covariates are available for all environments, whereas in the new setups 5 and 6, environmental covariates from the trials of interest are missing and they are replaced by using several years of historical data.

setup 6 where no phenotype data are available for any of the lines of interest. We emphasize that a further difference to earlier work [8] is that we strictly require the test environments to be simultaneously both from a location and from a year not included among the training environments and that the genotypes in the test/validation sets are from the progeny of the training set. A summary of the differences between our setup to those presented by earlier works is given in Table S 2

Gains from modelling $G\times E$ for current target population of environments. Our results indicate targeted breeding could improve yields by dividing a single target population of environments (TPE) into several parts, but the same methodology could be used even when developing only 1 variety for a larger population of target environments as in traditional breeding. Traditional breeding makes the implicit assumption that varieties' observed yields $g \in 1, \ldots, G$ in trial experiments in environments (location $\times$ year) $e \in 1, \ldots, E$, are representative of the yield in the TPE,

Publication	New environment	New genotypes
Burgueño et al. 2012 (CV1/CV2) [10]	CV1/CV2: test locations and years are present in the location-year combinations in the training data	new lines in CV1: not restricted to the offspring generation. In CV2 the test lines have phenotype observations
Heslot et al. 2013 [7]	Random split, balanced wrt years and locs $\rightarrow$ year and locations not new	only 544/2195 genotypes have no phenotype observations, test set not restricted to the offspring generation
Albrecht et al. 2014 [11]	the year-location combination is new but the test locations and years are a present in other location-year combinations in the training data	genotypes are new and from the offspring
Malosetti et al. 2016 [8]	time-structured DTD: 2/6 test locations new according to strict criteria; physically structured DTD: none of the environments are strictly new (as the year is not new)	all genotypes within the same family, not from the next generation.
Saint Pierre et al. 2016 [12] (leave-one-side-out)	location new but year part of the training set	test lines have phenotype observations

Table S 2: Comparison of the proposed *in silico* setup to the existing setups.

363 in other words

$$p(\text{yield}_g|\text{TPE}) \approx \frac{1}{E}\sum_e p(\text{yield}_g|\text{environment}_e) \tag{4}$$

However, with geographic field use information and weather data widely available, this strong assumption can be replaced with an estimate for the yield in the TPE given the actual fields and their microclimates:

$$p(\text{yield}_g|\text{TPE}) \approx \sum_f^F P_f \times p(\text{yield}_g|f) \tag{5}$$

$$= \sum_f^F P_f \times \int_{\theta_f} p(\text{yield}_g|\theta_f) \times p(\theta_f) d\theta_f, \tag{6}$$

364 where $f \in 1, \ldots, F$, are fields in the TPE used for cultivation of the new variety, θ_f are parameters
365 (e.g. weather conditions) related to a certain field f, $p(\theta_f)$ is the uncertainty related to these
366 conditions, estimated from historical records, $p(\text{yield}_g|\theta_f)$ is the predictive distribution for the
367 yield under conditions θ_f, obtained from the model, and P_f is the proportion of the total volume
368 cultivated in field f.

	environment	GBLUP	M_{G+E}	M_{G+E+GE}	M^{hist}_{G+E+GE}	GE−BLUP	N_{test}
1	Loc B, 2011	-0.147	-0.146	-0.109	-0.07	0.068	59
2	Loc A, 2011	-0.237	-0.086	-0.206	-0.2	-0.015	59
3	Loc G, 2011	-0.02	-0.028	0.045	0.031	0.183	58
4	Loc B, 2013	-0.016	0.057	0.051	0.046	0.182	182
5	Loc A, 2012	0.167	0.198	0.167	0.163	0.057	106
6	Loc G, 2012	0.023	-0.079	0.017	0.031	-0.166	106
7	Loc D, 2012	0.01	-0.031	-0.01	-0.026	-0.075	105
8	Loc E, 2013	0.124	0.2	0.265	0.245	-0.354	91
9	Loc B, 2012	0.089	0.166	0.17	0.195	-0.047	106
10	Loc G, 2013	0.228	0.21	0.278	0.282	-0.256	91
11	Loc B, 2012	0.171	0.223	0.179	0.192	0.127	260
12	Loc A, 2012	0.046	0.113	0.087	0.043	0.1	243
13	Loc E, 2013	0.467	0.488	0.503	0.518	-0.122	153
14	Loc G, 2013	0.2	0.308	0.353	0.344	-0.107	152
15	Loc C, 2014	0.208	0.197	0.068	0.045	0.054	79
16	Loc B, 2013	0.267	0.309	0.291	0.225	0.147	153
17	Loc A, 2013	0.05	0.128	0.116	0.114	0.089	153
18	Loc E, 2014	-0.02	-0.007	0.027	-0.013	0.354	79
19	Loc B, 2014	-0.037	-0.021	-0.046	-0.044	-0.133	79
20	Loc B, 2014	0.258	0.272	0.012	0.054	0.122	106
21	Loc C, 2014	0.344	0.324	0.375	0.345	0.065	106
22	Loc E, 2014	0.325	0.42	0.371	0.404	0.275	105
23	Loc H, 2015	0.199	0.12	0.166	0.122	0.251	64
24	Loc F, 2015	0.246	0.245	0.197	0.295	0.097	64
25	Loc B, 2015	-0.069	-0.093	0.227	0.257	-0.289	64
26	Loc B, 2013	0.091	0.084	0.078	0.083	0.058	488
27	Loc G, 2013	0.181	0.236	0.151	0.138	-0.015	244
28	Loc E, 2013	0.243	0.222	0.234	0.236	-0.052	244
29	Loc C, 2014	0.292	0.36	0.385	0.382	-0.069	120
30	Loc F, 2015	0.053	0.169	0.186	0.161	0.002	39
31	Loc E, 2014	0.232	0.217	0.25	0.208	-0.062	120
32	Loc B, 2015	-0.056	-0.083	-0.062	-0.071	-0.292	39
33	Loc B, 2014	0.01	-0.016	-0.145	-0.137	0.086	91
34	Loc E, 2014	0.221	0.306	0.382	0.367	-0.157	91
35	Loc C, 2014	0.273	0.25	0.232	0.258	0.089	91
36	Loc H, 2015	0.044	0.006	0.159	0.123	-0.206	42
37	Loc B, 2015	-0.041	0.035	0.145	0.128	0.219	42
38	Loc B, 2015	0.095	-0.039	0.008	0.012	0.039	64
39	Loc F, 2015	-0.037	0	-0.039	0.03	0.078	64
40	Loc H, 2015	0.25	0.259	0.261	0.262	0.253	63
41	Loc C, 2015	-0.204	-0.341	-0.353	-0.359	-0.143	60

Table S 3: Results for individual test folds.