Received: 7 March 2025 | Accepted: 26 July 2025

DOI: 10.1002/tpg2.70105

The Plant Genome

ORIGINAL ARTICLE

# Enhancing the prediction accuracy of groundnut yield by integrating significant markers and modeling genotype × environment interaction

**Nelson Lubanga**[1] | **Velma Okaron**[2] | **Davis M. Gimode**[3] | **Reyna Persa**[4] | **James Mwololo**[5] | **David K. Okello**[6] | **Mildred Ochwo Ssemakula**[2] | **Thomas L. Odong**[2] | **Wilfred Abincha**[7] | **Damaris A. Odeny**[3] | **Diego Jarquin**[4]

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, UK

[2]Department of Agricultural Production, School of Agricultural Sciences, College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda

[3]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)–Nairobi, Nairobi, Kenya

[4]Agronomy Department, University of Florida, Gainesville, Florida, USA

[5]ICRISAT - Lilongwe, Lilongwe, Malawi

[6]National Semi-Arid Resources Research Institute, Serere, Uganda

[7]Kenya Agricultural and Livestock Research Organization (KALRO), City Square, Kenya

**Correspondence**
Damaris A. Odeny, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT)–Nairobi, P.O Box 39063-00623 Nairobi, Kenya.
Email: Damaris.Odeny@icrisat.org

Diego Jarquin, Agronomy Department, University of Florida, 2089 Mccarty Hall B, P.O. Box 110500, Gainesville, FL 32611-7011, USA.
Email: jhernandezjarqui@ufl.edu

**Present address**
Davis M. Gimode, CGIAR Standing Panel on Impact Assessment, Alliance of Bioversity International and CIAT, Palmira

## Abstract

Multi-environment trials are routinely conducted in plant breeding to capture the genotype-by-environment interaction (G × E) effects. Significant G × E could alter the response pattern of genotypes (the change in rankings of genotypes), subsequently complicating the selection process. Four genomic prediction (GP) models were assessed in three groundnut yield-related traits: pod yield (PY), seed weight (SW), and 100 seed weight (SW100), across four environments. The models, M1 (environment + line), M2 (environment + line + genomic), M3 (environment + line + genomic + genomic × environment interaction), and M4 (environment + line + genomic + genomic × environment interaction + significant markers), were tested using four cross-validation (CV) schemes (CV2, CV1, CV0, and CV00), each simulating different practical breeding scenarios. The results revealed that models incorporating marker data (M2, M3, and M4) consistently improved predictive ability

**Abbreviations:** BLUEs, best linear unbiased estimates; CGM, crop growth models; CV, cross-validation; EC, environmental covariables; G × E, genotype-by-environment interaction; G × Y, genotype-by-year; GBLUP, genomic best linear unbiased prediction; GP, genomic prediction; GS, genomic selection; GWAS, genome-wide association study; ICRISAT, nternational Crops Research Institute for the Semi-Arid Tropics; M × E, marker × environment interaction; MET, multi-environment trial; PY, pod yield; SSA, sub-Saharan Africa; SW, seed weight; WGP, whole genome prediction.

Nelson Lubanga and Velma Okaron are equal first authors.

Campus Km 17 vía Cali–Palmira, Palmira, Colombia.

Assigned to Associate Editor Jianming Yu.

**Funding information**
Food and Agriculture Organization (FAO)

in comparison to the phenotypic model (M1). Incorporating G × E (M3 and M4) further improved predictive ability and reduced residual and environmental variances. The inclusion of significant markers and G × E was more advantageous in CV1 and CV00 scenarios, demonstrating that this strategy is especially useful when phenotypic data for the target genotypes is limited or unavailable. Across the CV schemes, predictive ability was higher in CV2, suggesting that including additional information on the performance of genotypes in known environments can increase the accuracy of selecting superior genotypes in breeding programs. Integrating significant markers and modeling G × E in GP models could be an effective approach in groundnut breeding programs to accelerate genetic gains.

**Plain Language Summary**

Genotype-by-environment interaction (G × E) is an important consideration in plant breeding, particularly for crops like groundnut, where environmental variability can significantly influence the yield. Effectively modeling G × E can enhance the ability to predict how different genotypes perform across diverse environments, which is essential for identifying promising genotypes that are high-yielding and stable. Using molecular marker data in genomic prediction models improves the accuracy of selecting the best genotypes. Including significant markers as fixed effects in these models can further increase predictive ability by making use of known genetic information. This approach allows for the direct utilization of known genetic information, thereby refining the selection of superior genotypes in breeding programs. Our findings showed that this approach enhanced the predictive ability and can be implemented in groundnut breeding programs to select superior high-yielding genotypes.

## 1 | INTRODUCTION

Groundnut (*Arachis hypogaea L.*) is a vital crop primarily cultivated by smallholder farmers in tropical and subtropical countries for food security and income. It plays a significant role in sub-Saharan Africa (SSA), where it is a major source of dietary protein (Toomer, 2018), healthy fats (Mora-Escobedo et al., 2015), essential vitamins (Arya et al., 2016), and micronutrients (Kurapati et al., 2021). Despite its significance in SSA, groundnut yields remain remarkably low, averaging ∼1017 kg/ha compared to 4505 kg/ha recorded in the United States (FAO, 2022; USDA-NASS, 2021). Groundnut yield losses in SSA have been further worsened by climate change effects (Yeleliere et al., 2023) that include prolonged dry periods, heavy rainfall, hail damage, increased temperatures, and the emergence of pests and diseases (Batley & Edwards, 2016). Developing climate-resilient groundnut varieties is therefore an important breeding objective for improving yields in SSA (Asibuo et al., 2018; Tabe-O et al., 2023).

Evaluating advanced breeding lines in multi-environment trials (METs) is a crucial step in the varietal development process (Janila et al., 2013). The METs enable the assessment of genotype stability through studying genotype-by-environment interaction (G × E) (Burgueño et al., 2012). Significant G × E occurs when genotypes respond differently to changes in the environment, potentially complicating the selection of high-yielding groundnut genotypes (Asibuo et al., 2018). Similarly, significant G × E can alter the relative performance rankings of genotypes across environments.

Genomic selection (GS) involves the application of genomic prediction (GP) models to estimate breeding values of genotypes based on genome-wide markers (Bernardo, 1994; Meuwissen et al., 2001). The accurate capture of G × E enables the identification of superior and stable genotypes across different environments (Crossa et al., 2006; Crossa et al., 2014). Single-environment GP models ignore G × E, and are therefore unreliable (Crossa et al., 2017) as they fail to capture the heterogeneity of genetic variance across environments, or the imperfect genetic correlation among environments (Lopez-Cruz et al., 2015). Modeling covariance matrices by incorporating G × E allows for the utilization of information from correlated environments, thereby enhancing the accuracy of predicting genotypes (Burgueño et al., 2012).

The accuracy of GP models is evaluated using random cross-validation (CV) methods that partition the dataset into training and testing sets (Pérez-Cabal et al., 2012). The CV2 method, in which known genotypes are evaluated in known environments with unknown combinations of genotype and environments, is particularly well-suited for scenarios involving incomplete field trials (sparse testing) (Burgueño et al., 2012). The primary objective of CV2 is to predict the performance of genotypes in environments where they have not yet been tested. An alternative method, CV1, predicts the performance of untested genotypes in known environments (Burgueño et al., 2012). In the CV1 scenario, the accuracy of prediction is influenced by the genetic relatedness between the training and testing sets. A third alternative is CV0, which is used to predict the performance of tested genotypes in new environments where they have not been evaluated (Jarquín et al., 2017). Another method, CV00, predicts untested genotypes in unknown environments, and is therefore, the most challenging strategy (Jarquín et al., 2017).

The first statistical framework for analyzing G × E was the Finlay–Wilkinson model, also known as the joint regression analysis that was first introduced by Yates and Cochran (1938) and later published by Finlay and Wilkinson (1963). In this method, the performance of each genotype in individual environments is regressed against an environmental index, which represents the overall effect of each environment within the trial. This index is typically defined as the mean performance of all genotypes in a given environment. Burgueño et al. (2012) were the first to use the genomic best linear unbiased prediction (GBLUP) model in a multi-environment context where G × E was modeled with different variance and covariances among environments (factor analytic). However, the model does not incorporate environmental covariables (EC), which limits its ability to predict performance in new environments. Jarquín et al. (2014) introduced the Bayesian reaction-norm model incorporating G × E as functions of markers and ECs. In the Bayesian reaction-norm model, the main effects of markers and EC, as well as their interactions (marker × EC), are introduced using covariance structures that are functions of markers and EC. This model extends the standard GBLUP by fitting all markers, ECs, and marker × EC interaction effects as random effects, modeled jointly through a multiplicative covariance structure. Specifically, the interaction term follows a multivariate normal distribution with covariance defined as the Hadamard (element-wise) product of the genomic and environmental relationship matrices. Including EC in this model allows predicting genotypes in new environments.

Lopez-Cruz et al. (2015) proposed a marker × environment interaction (M × E) model, in which marker effects and genomic values are partitioned into components that are stable across environments (main effects) and others that are environment-specific (interactions). The M × E model estimates the phenotypic correlation between any two environments as the ratio of variance component estimates. The covariance (and thus the correlation) between environments is derived from the variance components associated with the genetic effects across environments. Since variance components are constrained to be nonnegative, this formulation inherently restricts the covariance (and thus correlation) between environments to be positive. Cuevas et al. (2016) extended the M × E model by using a nonlinear (Gaussian) kernel to model G × E. Similarly, Crossa, los Campos, et al. (2016) extended the M × E model using priors that produce shrinkage and variable selection, that is, Bayesian ridge regression and BayesB.

In this study, we implemented the reaction-norm model (Jarquín et al., 2014) by incorporating the interaction between markers and EC. Previous studies have demonstrated that GP models that include significant markers associated with traits of interest as fixed effects, in combination with genome-wide markers, achieve higher accuracy than standard GP models such as multiple linear regression, G-BLUP, Bayesian Lasso, and Bayes C$\pi$ (Rutkoski et al., 2014). Spindel et al. (2016) presented a method where genome-wide association study (GWAS) was conducted on a training set, and markers passing a threshold were modeled as fixed effects in the RR-BLUP model. This approach, which they termed GS + de novo GWAS, improved the prediction accuracy by 10% compared to six other GP and marker-assisted selection (MAS) approaches in rice (*Oryza sativa* L.) and *Capsicum annuum*. Kim et al. (2022) also reported an improved prediction accuracy of 5.3% for capsaicinoid contents in *Capsicum annuum*. These instances provide evidence that including significant markers as fixed effects in GP models has the potential to improve the prediction accuracy. The current study incorporated significant markers from previous GWAS analyses (Okaron et al., 2024) with the following objectives: (1) study the effects of modeling G × E in GP models for groundnut yield and (2) determine whether incorporating significant markers from GWAS as fixed effects in GP models could enhance predictive ability.

**Core Ideas**

- Including marker data in genomic prediction (GP) models significantly increased the predictive ability.
- Incorporating genotype-by-environment interaction (G × E) in GP models improved the predictive ability and reduced residual variances.
- Integrating significant markers and modeling G × E in GP models enhanced the predictive ability.

## 2 | MATERIALS AND METHODS

### 2.1 | Germplasm

A total of 192 advanced breeding lines sourced from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) - Malawi were used in this study (Table S1). ICRISAT utilises extensive germplasm collections to generate high-yielding and climate-resilient varieties in Asia and SSA. The field trial evaluations were conducted in four locations: Serere (1.4994° N, 33.5490° E) and Nakabango (0°19′36″ N, 33°53′36″ E) in Uganda, and Chitala (13°40′59″ S, 34°16′0″ E) and Chitedze (13.9815° S, 33.6372° E) in Malawi. The trials were planted in a 14 × 25 Alpha Lattice design with two replicates. Each plot was 1 × 0.9 m, comprising 3 m-long rows with 45 cm inter-row spacing, 60 cm between plots, and 2 m between replicated blocks. Phenotyping was conducted at Chitedze and Chitala (Malawi) during the 2020/21 cropping season (December–April), and at Serere and Nakabango (Uganda) during the long rainy season of 2021. These sites represent major groundnut-producing agroecological zones in Eastern and Southern Africa.

### 2.2 | Phenotyping, genotyping, and quality control

All genotypes were harvested at the end of the growing season and evaluated for PY and other yield-associated traits, including SW and SW100 in the four environments. Details of the plant material, experimental design, phenotyping and genotyping procedure are described in our recent study by Okaron et al. (2024). In brief, pods from each plot were harvested and sun-dried to a moisture content below 13%, then weighed using a weighing scale, and recorded in grams as PY. For SW, seeds from each plot were shelled and weighed in grams. Additionally, 100 kernels were randomly selected from each plot, weighed, and the data recorded in grams as SW100.

SNP filtering involved removing all markers with >20% missing values and a minor allele frequency below 5%. SNPs with a call rate lower than 80% and heterozygosity larger than 95% were also excluded. The final retained genotypes contained 38,853 SNPs. Filtered SNPs were converted to numeric allele classes (0, 1, and 2) for homozygous minor, heterozygous and homozygous major alleles, respectively using Plink 1.9 (Purcell et al., 2007).

### 2.3 | Phenotypic analysis

The phenotypic analyses were performed in two stages. The first stage involved calculating the best linear unbiased estimates (BLUEs) for each genotype within each environment using Equation (1). The BLUEs for each trait in each environment were extracted using ASReml-R Version 4.2 (Butler et al., 2023).

$$y_{\mathrm{ikl}} = \mu + L_i + R_k + B_{kl} + e_{ikl} \tag{1}$$

Here, $y_{ikl}$ represents the $i$th genotype observed in the the $l$th incomplete block nested within $k$th replicate. $\mu$ is the overall mean (intercept), $L_i$ is the fixed effect of the $i$th genotype, $R_k$ is the random effect of the $k$th replicate, assuming $R_k \overset{\text{iid}}{\sim} N(0, \sigma_R^2)$, $\sigma_R^2$ is the variance of the replicate. $N(.,.)$ denotes a normal density, iid stands for independent and identically distributed observations, $B_{kl}$ is the random effect of the $l$th block nested within the $k$th replicate such that $B_{kl} \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$, $\sigma_B^2$ is the variance of the block effect nested within replication, $e_{ikl}$ is the random error term with $e_{ikl} \overset{\text{iid}}{\sim} N(0, \sigma_e^2), \sigma_e^2$ as the error variance.

To estimate the broad-sense heritability ($H^2$) for each trait at each environment, we extracted the variances of each effect using the following model whereby genotype, replicate, and the block within replicate were considered random effects:

$$y_{ikl} = \mu + L_i + R_k + B_{kl} + e_{ikl} \tag{2}$$

where $y_{ikl}$ represents the observed value, $\mu$ is the intercept, $L_i$ is the random effect of the $i$th genotype, assuming $L_i \overset{\text{iid}}{\sim} N(0, \sigma_L^2)$, $\sigma_L^2$ is the genotype variance, $R_k$ is the random effect of the $k$th replicate, assuming $R_k \overset{\text{iid}}{\sim} N(0, \sigma_R^2)$, $\sigma_R^2$ is the variance of the replicate, $B_{kl}$ is the random effect of the $l$th block nested within the $k$th replicate such that $B_{kl} \overset{\text{iid}}{\sim} N(0, \sigma_B^2)$, $\sigma_B^2$ is the variance of the block effect nested within replication, $e_{ijk}$ is the random error term with $e_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma_e^2), \sigma_e^2$ is the error variance. Broad-sense heritability was then calculated using Equation (3):

$$H^2 = \frac{\sigma_L^2}{\sigma_L^2 + \sigma_{e/r}^2} \tag{3}$$

where $\sigma_L^2$ is genotypic variance, $\sigma_e^2$ is error variance, and $r$ is the number of replicates within each environment.

Broad-sense heritability across the environment was calculated using the following linear model:

$$y_{ijkl} = \mu + L_{\mathrm{i}} + E_j + R_k + B_{kl} + LE_{ij} + e_{ijkl} \tag{4}$$

Here, $y_{ijkl}$ represents the $i$th genotype observed in the $j$th environment at the $l$th incomplete block nested within $k$th replicate. $\mu$ is the overall mean (intercept), $L_i$ is the random effect of the $i$th genotype, assuming $L_i \overset{\text{iid}}{\sim} N(0, \sigma_L^2)$, $\sigma_L^2$ is the genotype variance, $E_j$ is the random effect of the $j$th environment, assuming $E_j \overset{\text{iid}}{\sim} N(0, \sigma_E^2)$, $\sigma_E^2$ is the variance of

the environment, $R_k$ is the random effect of the $k$th replicate, assuming $R_k \overset{iid}{\sim} N(0, \sigma^2_R)$, $\sigma^2_R$ is the variance of the replicate, $B_{kl}$ is the random effect of the $l$th block nested within the $k$th replicate such that $B_{kl} \overset{iid}{\sim} N(0, \sigma^2_B)$, $\sigma^2_B$ is the variance component of the block effect nested within replication, $LE_{ij}$ is the G × E of the $i$th genotype in $j$th environment, $LE_{ij} \overset{iid}{\sim} N(0, \sigma^2_{LE})$, $\sigma^2_{LE}$ is the G × E variance, $e_{ijkl}$ is the random error term with $e_{ijkl} \overset{iid}{\sim} N(0, \sigma^2_e)$, $\sigma^2_e$ is the error variance of the non-explained variability by the previous model terms. Broad-sense heritability was then calculated on an entry-mean basis using the following equation:

$$H^2 = \frac{\sigma^2_L}{\sigma^2_L + \sigma^2_{LE/n} + \sigma^2_{e/n\times r}} \tag{5}$$

where $\sigma^2_L$ is genotypic variance, $\sigma^2_{LE}$ is the G × E variance, $\sigma^2_e$ is error variance, and $r$ is the number of replicates within each environment and n is the number of environments.

## 2.4 | Prediction models

The second stage of the analysis involved using the BLUEs obtained in the first stage (Equation 1) to fit four different prediction models described below in Equations (6–9).

### 2.4.1 | M1: Environment + line (E + L)

In this model, the response variable corresponds to the adjusted phenotypes ($y_{ij}$) (BLUEs) obtained after fitting the corresponding mixed linear model as previously described in Equation (1). $y_{ij}$ represents the response of the $i$th genotype at the $j$th environment. In this case, the environment and line effects are considered random outcomes:

$$y_{ij} = \mu + L_i + E_j + e_{ij} \tag{6}$$

where $L_i$ is the random effect of the $i$th line assuming $L_i \overset{iid}{\sim} N(0, \sigma^2_L)$, where $\sigma^2_L$ is the variance component of the genotype effect, $E_j$ is the random effect of the $j$th environment such that $E_j \overset{iid}{\sim} N(0, \sigma^2_E)$ where $\sigma^2_E$ is the variance component of the environment effect, and $e_{ij}$ is the random error term, such that $e_{ij} \overset{iid}{\sim} N(0, \sigma^2_e)$, where $\sigma^2_e$ is the error variance.

### 2.4.2 | M2: Environment, line, and genomic main effects (E + L + G)

This model is an extension of M1 (Equation 6). It considers the inclusion of the random genomic effect of the $i$th genotype $g_i$, which approximates its additive genetic value using molecular marker information. This model component was defined by the regression on $p$ marker covariates,

$$g_i = \sum_{m=1}^{p} x_{im} b_m \tag{7}$$

where $x_{im}$ is the genomic information (marker matrix) of the $i$th genotype at the $m$th marker, and $b_m$ is the corresponding marker effect. Marker effects are considered as iid draws from normal distributions of the form $b_m \overset{iid}{\sim} N(0, \sigma^2_b)$, with $\sigma^2_b$ as the corresponding variance component, which is homogenous to all markers. The vector of genomic effects $g = \{g_i\}$ follows a multivariate normal density with zero mean and variance-covariance matrix cov $(g) = G\sigma^2_g$ such that $g \sim N(0, G\sigma^2_g)$.

Here, $G\ \alpha\ \frac{XX'}{p}$ is the genomic relationship matrix, $X$ is the centered and standardized (by columns) matrix of molecular markers, and $\sigma^2_g = p\sigma^2_b$ is the genomic variance (additive genetic variance). The resulting model becomes:

$$y_{ij} = \mu + L_i + E_j + g_i + e_{ij} \tag{8}$$

where the genotype effects of the vector of genomic random effects $g$ are correlated such that Model 2 allows the borrowing of information across genotypes. A disadvantage of this model is that the genomic estimated breeding values for a genotype tested in different environments are the same regardless of the environment. To allow a specific genomic effect in each environment, G × E was included as shown in the Model M3.

### 2.4.3 | M3: Environment, line, genomic, and genomic × environment interaction effects (E + L + G + G × E)

This model is an extension of M2 (Equation 8) by adding the interaction between the markers and environments ($gE_{ij}$) via covariance structures as described by Jarquín et al. (2014). This model is an extension of the GBLUP model where the interactions between the markers and each environment are indirectly modeled using the following linear predictor:

$$y_{ij} = \mu + L_i + E_j + g_i + gE_{ij} + e_{ij} \tag{9}$$

where the $gE_{ij}$ term corresponds to the interaction between the genetic value of the $i$th genotype and the $j$th environment. This interaction term is assumed to follow a multivariate normal distribution such that

$$gE = \{gE_{ij}\} \sim N\left(0, Z_g G Z'_g \circ Z_E Z'_E\ \sigma^2_{gE}\right)$$

where $Z_g$ and $Z_E$ are the incidence matrices that connect genotypes and environments with phenotypes, respectively. $\sigma^2_{gE}$ is the variance component associated with $gE$ and "°" represents the Hadamard product (element-by-element product) of the two covariance structures representing the genetic information and the the environmental effects.

#### 2.4.4 | M4: Environment, line, genomic, genomic × environment interaction effects, and significant GWAS SNPs as fixed effects (E + L + G + G × E + S)

GWAS was previously conducted by Okaron et al. (2024) using the same set of SNPs that were employed in the GP models. The significant SNPs ($p < 0.005$) that were associated with PY, SW, and SW100 from the BLINK model were selected to improve GP. Wang and Zhang (2021) demonstrated that the BLINK model offers higher statistical power compared to other GWAS models. Based on this criterion, the number of selected markers as fixed effects for each trait was 111 (PY), 187 (SW), and 96 (SW100).

The M3 model was extended by incorporating the significant markers as fixed effects as follows:

$$y_{ij} = \mu + E_j + L_i + g_i + gE_{ij} + S_i + e_{ij} \quad (10)$$

where $S_i$ is the fixed genomic effect associated with the $i$th genotype.

$$S_i = \sum_{u=1}^{t} x_{iu} b_u$$

where $x_{iu}$ is the genotype at the $u$th ($u = 1, 2,.., t$) selected marker of the $i$th individual, $t$ is the number of associated markers selected for inclusion as fixed-effect covariates and $b_u$ is the fixed additive effect of the $u$th significant marker.

### 2.5 | Model assessment under different CV schemes

To assess the performance of the four GP models, four prediction strategies commonly used by breeders were considered. These prediction scenarios mimic different realistic problems that breeders face at different stages of the breeding pipeline. In all the CVs, predictive ability was calculated as the correlation between observed and predicted values within the same environment. A random five-fold partitioning of the entire population was used for CV1 and CV2, and predictive ability was calculated as the average correlation between predicted and observed values of genotypes within the same environment. In CV0 and CV00, the "leave one environment out" approach was used to predict future (unknown) environments.

In CV2, a fivefold CV was implemented, where 20% of the phenotypic values were randomly assigned to each fold, one fold was used as a testing set, and the remaining four folds (80%) were used as the training set for model calibration. This process was repeated 10 times for all the five folds, and the correlations between the predicted and observed values within the same environment computed for all environments.

Under CV1, a five-fold CV approach was also used, where the dataset was randomly divided into five equal subsets, each containing 20% of the genotypes. One fold was used as the validation set, while the remaining four folds (comprising 80% of the genotypes) were used as the training set. Predictions were generated separately for each fold. Predictive ability was assessed as the correlation between predicted and observed values within the same environment. This process was repeated across 10 random fivefold partitions.

CV0 was done by masking the phenotypic information of all genotypes at each environment (one at a time), then using the remaining environments as the training set. In CV00 scheme, both the phenotypic information of the testing environment and the phenotypic information of the genotype to be predicted were deleted from all environments (one at a time).

All the prediction models were analyzed using the BGLR package (Pérez & de los Campos, 2014).

## 3 | RESULTS

### 3.1 | Phenotypic data analysis

Mean phenotypic values and broad-sense heritability ($H^2$) for PY, SW, and SW100 across the four environments that had been reported in Okaron et al. (2024) are also summarized in Table 1. The mean phenotypic values for PY, SW, and SW100 ranged from 46.97 to 328.85 g, 22.73 to 218.19 g, and 28.08 to 43.01 g, respectively (Figure 1).

The estimated $H^2$ across environments indicated influences of genetic and environmental impacts and their interactions across environments. SW showed the highest proportion of variance explained by G × E at 16%, while SW100 had the highest proportion of variance explained by the residual at 52% (Figure 1). The corresponding narrow sense $h^2$ estimates were 0.38, 0.31, and 0.58 for PY, SW, and SW100, respectively.

### 3.2 | Evaluation of the variance components for each model

The relative percentage of variability explained by the different model terms is shown in Figure 2. Adding G × E effects and significant markers as fixed effects reduced the

**TABLE 1** Mean phenotypic values and heritability of pod yield (PY), seed weight (SW), and 100 seed weight (SW100) in the four environments.

| | PY | | SW | | SW100 | |
|---|---|---|---|---|---|---|
| **Environment** | **Mean (g)** | $H^2$ | **Mean (g)** | $H^2$ | **Mean (g)** | $H^2$ |
| Chitala | 328.85 | 0.63 | 218.19 | 0.67 | 38.69 | 0.76 |
| Chitedze | 117.08 | 0.58 | 62.53 | 0.77 | 28.08 | 0.56 |
| Nakabango | 46.97 | 0.85 | 22.73 | 0.73 | 40.72 | 0.10 |
| Serere | 99.69 | 0.39 | 41.83 | 0.42 | 43.01 | 0.28 |
| Across environments | | 0.42 | | 0.31 | | 0.44 |

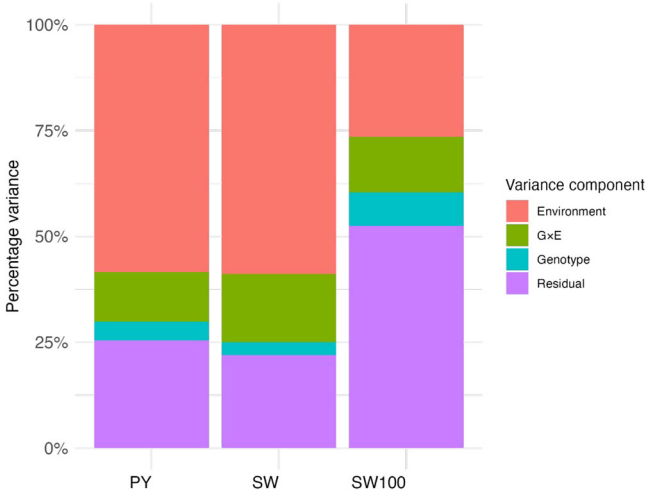


**FIGURE 1** Trait phenotypic variance component estimation and proportion of the total variance explained for the three traits (pod yield [PY], seed weight [SW], and 100 seed weight [SW100]). G × E, genotype-by-environment.

variability captured by the environment and residual terms (Figure 2). As expected, the main effects of the environment consistently explained most of the total variance for all traits. The most complex model, M4, had the lowest variance explained by the environment effect, while M1 had the highest in all the traits. This suggests that M4 may more effectively capture G × E than the other models. Relative to model M1, the inclusion of G × E reduced residual variance by 2.9% for PY, 4.0% for SW, and 4.9% for SW100. A comparable reduction was observed in environmental variance, with decreases of 4.8%, 5.4%, and 4.3% for the respective traits (Figure 2).

## 3.3 | Predictive ability

The mean predictive abilities of all models across the evaluated traits are presented in Figures 3–6. In general, the predictive ability was higher in CV2 compared to CV1, CV0, and CV00, implying that including more information on observed genotypes and known environments in the training set can enhance model performance. Incorporating the G × E interaction in model M3 increased predictive ability compared to the main effect models, that is, M1 and M2, for all traits under CV1 and CV2 schemes (Figures 3 & 4). The inclusion of significant markers as fixed effects, along with G × E, improved the predictive ability in both CV1 and CV00 scenarios (Figures 3 & 5). This finding suggests that accounting for known associated loci can improve the predictive ability by capturing a greater proportion of the underlying genetic variance.

### 3.3.1 | Predictive ability under CV1

All models incorporating marker information exhibited higher predictive ability compared to the phenotype model (M1). Overall, models that incorporated G × E exhibited higher predictive ability than the main-effect models M1 and M2 (Figure 3; Table S2). Additionally, incorporating significant markers as fixed effects in the GP model that accounted for G × E further enhanced predictive ability across all traits and environments (Figure 3). Model M4 consistently demonstrated the highest predictive ability, whereas M1 showed the lowest across all traits. The range in predictive ability was −0.15 to 0.73, −0.15 to 0.67, and −0.11 to 0.77 for PY, SW, and SW100, respectively (Figure 3; Table S2).

### 3.3.2 | Predictive ability under CV2

All traits exhibited moderate to high predictive abilities. Models that incorporated marker information consistently outperformed the phenotypic Model (M1), for all traits except SW100. Overall, the interaction models M3 and M4 did not offer a significant improvement in predictive ability compared to the main-effect models M1 and M2 (Figure 4). The predictive ability ranged from 0.40 to 0.78, 0.35 to 0.71, and 0.69 to 0.90 for PY, SW, and SW100, respectively (Figure 4; Table S2).

### 3.3.3 | Predictive ability under CV00

Models that utilized significant markers as fixed effects consistently outperformed the phenotype model (M1),

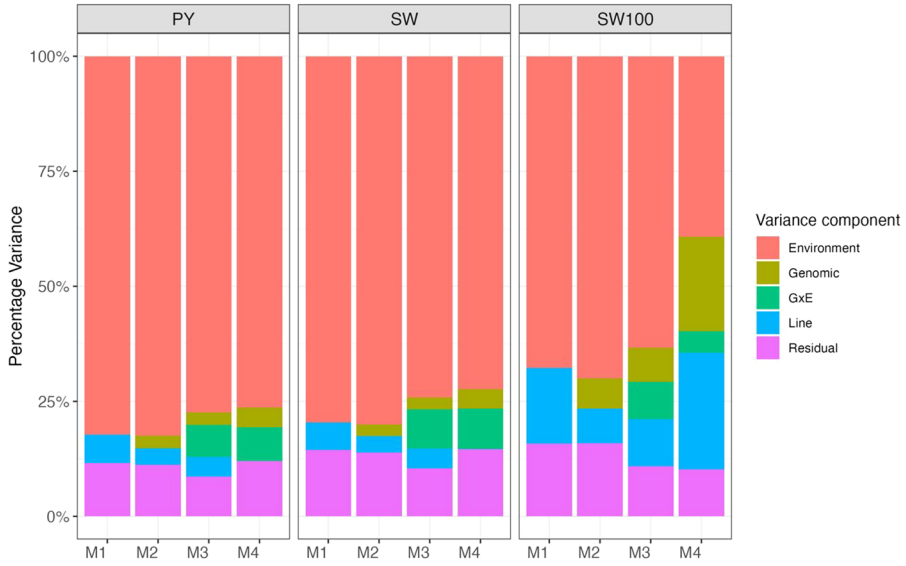


**FIGURE 2** Percentage of explained variance components accounted for pod yield (PY), seed weight (SW), and 100 seed weight (SW100) for the four models (M1, M2, M3, and M4). G × E, genotype-by-environment.

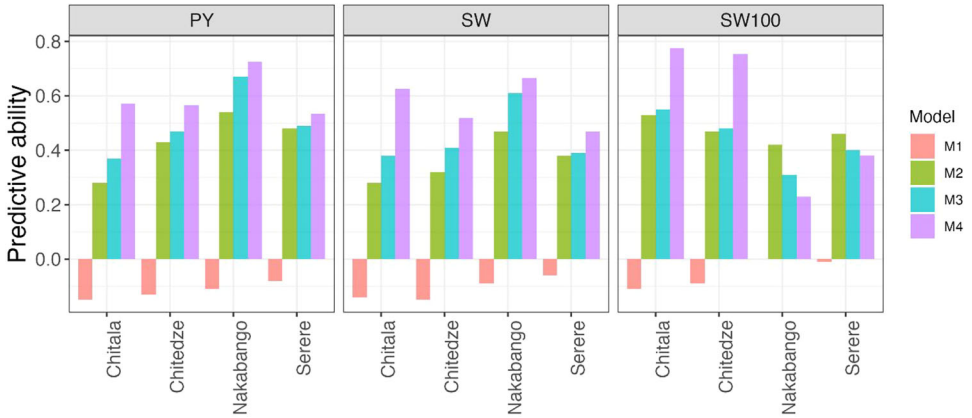


**FIGURE 3** Predictive abilities under CV1 (where CV is cross-validation) scheme of the four models for pod yield (PY), seed weight (SW), and 100 seed weight (SW100) at Chitala, Chitedze, Nakabango, and Serere.

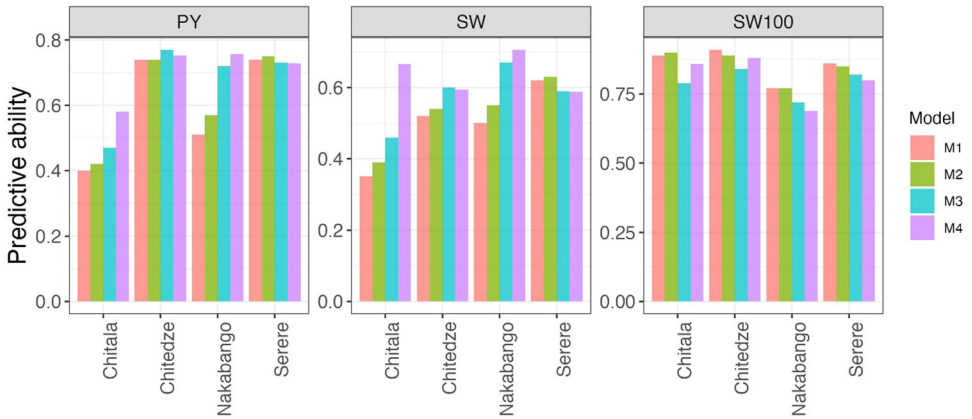


**FIGURE 4** Predictive abilities under CV2 (where CV is cross-validation) scheme for pod yield (PY), seed weight (SW), and 100 seed weight (SW100) at Chitala, Chitedze, Nakabango, and Serere.

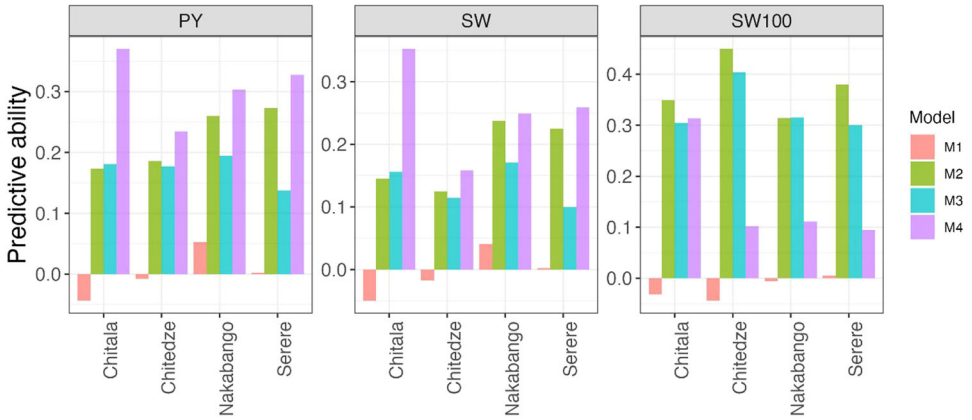


**FIGURE 5** Predictive abilities under CV00 (where CV is cross-validation) scheme for pod yield (PY), seed weight (SW), and 100 seed weight (SW100) at Chitala, Chitedze, Nakabango, and Serere.

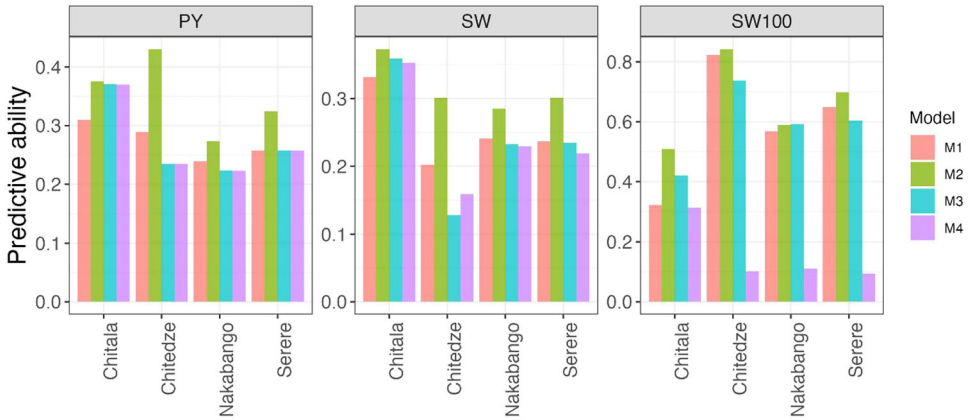


**FIGURE 6** Predictive abilities under CV0 (where CV is cross-validation) scheme for pod yield (PY), seed weight (SW), and 100 seed weight (SW100) at Chitala, Chitedze, Nakabango, and Serere.

highlighting the value of including known genomic information (Figure 5). As observed in CV1, models that accounted for G × E and included significant markers as fixed effects outperformed the main-effect models M1 and M2 across all traits, except SW100. The predictive ability ranged from −0.04 to 0.37, −0.05 to 0.35, and −0.05 to 0.45 for PY, SW, and SW100, respectively (Figure 5; Table S2).

### 3.3.4 | Predictive ability under CV0

Among the models evaluated, the main-effect model M2 consistently demonstrated superior predictive performance compared to M1, M3, and M4 (Figure 6; Table S2). Similar to CV2, the interaction models M3 and M4 did not provide additional predictive benefit compared to the main-effect models M1 and M2 (Figure 6; Table S2). Predictive ability ranged from 0.22 to 0.43 for PY, 0.13 to 0.37 for SW, and 0.10 to 0.84 for SW100 (Figure 6).

## 4 | DISCUSSION

The current study exploited existing phenotypic and genotypic datasets for advanced breeding lines evaluated across four different locations to enhance GP for yield-related traits in groundnut. The results demonstrated the power of modeling G × E to improve GP in groundnut breeding programs. GP models based on METs are inherently complex due to the heterogeneity of genetic variance and imperfect genetic correlations across environments, non-genetic noise, and high-dimensional marker data (Jarquín et al., 2014). Modeling and exploitation of G × E remains one of the major challenges in the analysis of METs in plant breeding and variety testing (Crossa, los Campos, et al., 2016). The use of EC to model the environment component explicitly has been previously shown to increase prediction accuracies in multiple environments (Jarquín et al., 2014), and this also applies to our work. The reaction norm model used in the current study allowed modeling the main and interaction effects of markers

and EC using covariance structures (Jarquín et al., 2014). The M3 model was in all cases superior to or not different from the main effect models (M1 and M2). These results are consistent with other studies that incorporated G × E, resulting in higher predictive accuracy compared to univariate models (Burgueño et al., 2012; Crossa, Burgueño, et al., 2006; Cuevas et al., 2016). The improved predictive ability could be due to the decreased amount of variance captured by both the environment and residual terms in all the traits as compared to the traditional GP model (M2) and the phenotypic-based model M1. In other studies, incorporating G × E was reported to reduce the proportion of variability attributed to the environment and residual terms in soybean (Canella Vieira et al., 2022). Mageto et al. (2020) observed that including G × E led to approximately 30% reduction in the residual variance component in maize. Lubanga et al. (2025) demonstrated that models incorporating G × E effects exhibited the lowest mean square error across various sparse testing designs when compared to models that included only main effects in cassava.

The benefit of using molecular marker information was reflected in the current study, where models M2, M3, and M4 revealed a higher predictive ability than M1 (without marker information). Lubanga et al. (2023) reported higher prediction accuracies for models incorporating markers compared to pedigree and phenotypic-based models. In GBLUP, molecular markers are used to estimate relatedness between individuals and represent realized genomic relationships rather than the expected relationships (Bernardo, 1994; Bernardo & Yu, 2007). This makes it possible to capture distant relationships and variation in sibling relationships due to Mendelian sampling, leading to more accurate estimates of additive genetic variance and breeding values (Goddard et al., 2011).

CV schemes are useful for assessing the predictive ability of GP models in a multi-environment context (Burgueño et al., 2012; Jarquín et al., 2017). The four CV schemes used in this study simulated different realistic problems that breeders face, that is, CV2, CV1, CV0, and CV00 (Jarquín et al., 2017; Persa et al., 2021). Across all models, predictive ability was consistently higher under CV2 compared to CV1, CV0, and CV00, due to the additional benefit of incorporating observed genotypes in known environments, even though the combinations of genotype and environments are unknown. However, CV2 adds time to the generation interval due to the additional field testing requirement for all the selection candidates (Burgueño et al., 2012). Although CV1 enables the selection of lines without field testing, it allows for a reduction in the generation interval, and this generally leads to lower predictive ability compared to CV2. Breeding programs therefore need to prioritize accordingly to avoid compromising the annual rate of genetic progress (Burgueño et al., 2012; Jarquín et al., 2017).

Predicting new environments is a more difficult task (Jarquín et al., 2017). This study used CV0 and CV00 schemes for predicting in untested environments. These scenarios are common for most breeding programs that have trials in multiple locations with the need to predict the performance of superior genotypes in future environments. Predictive performance was consistently low under the CV00 scheme, which represents the most challenging CV scenario.

Among the traits, modeling G × E in SW100 did not improve the results of modeling the main genotype and EC matrices alone. It is possible that the EC used explained only a limited proportion of the across environment interaction for this trait, and for this reason the model incorporating G × E, when fitting covariance matrices for the environment and marker by environment interaction, did not improve predictive ability in comparison to the main effect models.

The findings of the current study align with earlier reports indicating that high heritability of the target trait results in improved predictive ability (Lubanga et al., 2021; Zhang et al., 2017). These findings imply that enhancing heritability of the target trait is a significant component in training populations to improve prediction accuracies in the GS pipeline design. Groundnut breeders will need to consider expanding the number of locations and replications, and testing in multiple years to increase heritability. The current strengthening of breeding networks such as the Groundnut Improvement Network for Africa (http://gnut4africa.org) and the Africa Dryland Crops Improvement Network in Africa can be further leveraged to increase testing locations and make GS practical.

The current results further validate earlier studies that reported improved predictive ability using significant markers from trait association studies as fixed effects in prediction models (Spindel et al., 2016). The implemented model (M4) accounted for G × E by modeling interactions between markers and ECs using covariance functions, and incorporating trait-associated markers from a previous study as fixed effects. The results revealed that including significant markers as fixed effects, along with modeling G × E, improved predictive ability across all traits in CV1 and CV00. However, no clear benefit was observed in CV2 and CV0, perhaps due to the availability of abundant training data leading to the high predictive ability. These results suggest that the advantage of incorporating significant markers and modeling G × E would be more pronounced when predicting unknown genotypes in new environments. The higher influence of the significant markers incorporated as fixed effects in the model could be due to a lower proportion of the total variance associated with the trait explained by the random portion of the model. We, therefore, strongly recommend the addition of trait-associated markers where available to improve GP in groundnut breeding programs. These findings are consistent with those of Li et al. (2019) in maize and Kim et al. (2022) in *Capsicum*

*annuum*, both of which reported improved predictive ability when significant markers were included as fixed effects. Similarly, Rice and Lipka (2019) simulated 216 traits with varying genetic architectures and identified 60 cases in which the inclusion of fixed-effect markers led to improved prediction accuracy. They recommended that the performance of models incorporating significant markers as fixed effects be evaluated on a trait-by-trait basis prior to implementation in breeding programs.

The findings of the present study will require future validation based on some of the limitations. First, the phenotypic data for the three traits were collected over a single season across the four experimental sites located in Uganda and Malawi. Consequently, genotype-by-year (G × Y) interaction effects were not assessed, which may have limited the robustness of the dataset. Consideration of G × E in the context of G × Y remains an important challenge for breeders, especially those from small breeding programs (Monteverde et al., 2019). Second, the current study used the reaction norm model proposed by Jarquín et al. (2014) to analyze data and predict genomic performance of PY, SW, and SW100. The assumption of the model is that the relationships between molecular markers and EC are linear, which is a major limitation since interactions between genes and environmental conditions may take many different forms. The reaction norm model uses the Gaussian prior that does not induce variable selection, and the shrinkage induced by Gaussian prior density may not be particularly appropriate when markers or ECs have large effects. For instance, a study by Monteverde et al. (2019) on G × Y in grain yield and quality of rice that integrated molecular markers and EC demonstrated that partial least squares model (Wold et al., 2001) gave better prediction accuracies than reaction norm model (Jarquín et al., 2014).

Finally, this study did not evaluate varying levels of significance thresholds for molecular marker selection. Future work should focus on fitting models with greater biological realism that can accommodate non-linear and more complex genotype responses across a wide range of environments. Promising directions include (i) the reaction norm model with penalized factorial regression (Avagyan et al., 2025), (ii) integrating crop growth models (CGM) with whole genome prediction (WGP) (CGM–WGP) (Technow et al., 2015), (iii) using models that apply priors that produce variable selection (which cannot be directly implemented under the reaction norm model) (Crossa, los Campos, et al., 2016), and (iv) integrating random regressions on known and latent EC (Tolhurst et al., 2022). Future research should also investigate the effects of different marker threshold levels to determine the optimal balance between model complexity and predictive ability, and to provide guidance on the most effective threshold for marker inclusion.

## 5 | CONCLUSION

This study demonstrates the importance of incorporating G × E and significant markers in GP. G × E remains one of the most critical challenges in accurately estimating breeding values for agronomic traits, as it can obscure the true genetic potential of genotypes across diverse environmental conditions. The higher predictive ability observed in CV2 compared to the other scenarios suggests that incorporating more information from known genotypes and environments could enhance the prediction accuracy. The inclusion of significant markers as fixed effects and G × E was more advantageous in CV1 and CV00, demonstrating that this strategy is especially useful when phenotypic data for the target genotypes is limited or unavailable. Collectively, these results are promising, especially when predicting newly developed genotypes in unknown environments.

## AUTHOR CONTRIBUTIONS

**Nelson Lubanga**: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; writing—original draft; writing—review and editing. **Velma Okaron**: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; writing—original draft; writing—review and editing. **Davis M. Gimode**: Resources. **Reyna Persa**: Data curation; formal analysis; methodology. **James Mwololo**: Resources; writing—review and editing. **David K. Okello**: Resources. **Mildred Ochwo Ssemakula**: Resources. **Thomas L. Odong**: Resources. **Wilfred Abincha**: Formal analysis; resources. **Damaris A. Odeny**: Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; writing—review and editing. **Diego Jarquin**: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; software; validation; visualization; writing—review and editing.

## ACKNOWLEDGMENTS

The first two authors are thankful to the staff of the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in Malawi and the national staff in Uganda who helped in the data collection. This work was supported by the Food and Agriculture Organization (FAO) through the benefit sharing fund of the International Treaty on Plant Genetics for Food and Agriculture. Supplemental funding was provided by the Scholarship Advert for Short Term Academic Mobility (SCIFSA) and the Regional Universities' Forum for Capacity Building in Agriculture (RUFORUM) for Velma Okaron Ph.D fellowship. The testing and cross-validation of GP models was made possible through the analytical resources provided by

the Institute of Biological, Environmental and Rural Sciences (IBERS) High-Performance Computer Cluster.

## DATA AVAILABILITY STATEMENT

The R scripts and the phenotypic and genomic datasets used to perform the analysis are publicly available on the figshare repository: https://figshare.com/articles/dataset/Datasets_and_R_scripts_used_to_perform_Genomic_prediction_in_multi-environment_groundnut_traisl/27324150?file=50058195.

## ORCID

*Nelson Lubanga* https://orcid.org/0000-0002-8975-0793
*David K. Okello* https://orcid.org/0000-0001-5705-6898
*Wilfred Abincha* https://orcid.org/0000-0003-1605-8760
*Damaris A. Odeny* https://orcid.org/0000-0002-3629-3752
*Diego Jarquin* https://orcid.org/0000-0002-5098-2060

## REFERENCES

Arya, S. S., Salve, A. R., & Chauhan, S. (2016). Peanuts as functional food: A review. *Journal of Food Science and Technology*, *53*(1), 31–41. https://doi.org/10.1007/s13197-015-2007-9

Asibuo, J. Y., Forpoh, A. S., & Akromah, R. (2018). Genotype × environment interactions of groundnut (*Arachis hypogaea* L.) for pod yield. *Ecological Genetics and Genomics*, *7–8*, 27–32. https://doi.org/10.1016/j.egg.2018.03.001

Avagyan, V., Boer, M. P., Solin, J., van Dijk, A. D. J., Bustos-Korts, D., van Rossum, B.-J., Ramakers, J. J. C., van Eeuwijk, F., & Kruijer, W. (2025). Penalized factorial regression as a flexible and computationally attractive reaction norm model for prediction in the presence of GxE. *Theoretical and Applied Genetics*, *138*(4), Article 88. https://doi.org/10.1007/s00122-025-04865-4

Batley, J., & Edwards, D. (2016). The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Current Opinion in Plant Biology*, *30*, 78–81. https://doi.org/10.1016/j.pbi.2016.02.002

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, *34*(1), 20–25. https://doi.org/10.2135/cropsci1994.0011183Ĭ003400010003x

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

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. https://doi.org/10.2135/cropsci2011.06.0299

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. G., & Thompson, R. (2023). *ASReml-R reference manual (4.2.)*. VSN International Ltd.

Canella Vieira, C., Persa, R., Chen, P., & Jarquin, D. (2022). Incorporation of soil-derived covariates in progeny testing and line selection to enhance genomic prediction accuracy in soybean breeding. *Frontiers in Genetics*, *13*, 905824. https://doi.org/10.3389/fgene.2022.905824

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., & Krishnamachari, A. (2006). Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science*, *46*(4), 1722–1733. https://doi.org/10.2135/cropsci2005.11-0427

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., & Pérez-Rodríguez, P. (2016). Extending the marker × environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science*, *56*(5), 2193–2209. https://doi.org/10.2135/cropsci2015.04.0260

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y., Bonnett, D., & Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, *112*(1), 48–60. https://doi.org/10.1038/hdy.2013.16

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., Montesinos-López, O. A., & Burgueño, J. (2016). Genomic prediction of genotype × environment interaction kernel regression models. *The Plant Genome*, *9*(3), plantgenome2016.03.0024. https://doi.org/10.3835/plantgenome2016.03.0024

FAO. (2022). *World food and agriculture—Statistical yearbook 2022*. FAO. https://doi.org/10.4060/cc2211en

Finlay, K., & Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742. https://doi.org/10.1071/AR9630742

Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, *128*(6), 409–421. https://doi.org/10.1111/j.1439-0388.2011.00964.x

Janila, P., Nigam, S. N., Pandey, M. K., Nagesh, P., & Varshney, R. K. (2013). Groundnut improvement: use of genetic and genomic tools. *Frontiers in Plant Science*, *4*, 23. https://doi.org/10.3389/fpls.2013.00023

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

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., Battenfield, S., & Crossa, J. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype × environment interactions in Kansas wheat. *The Plant Genome*, *10*(2), plantgenome2016.12.0130. https://doi.org/10.3835/plantgenome2016.12.0130

Kim, G. W., Hong, J.-P., Lee, H.-Y., Kwon, J.-K., Kim, D.-A., & Kang, B.-C. (2022). Genomic selection with fixed-effect markers improves the prediction accuracy for capsaicinoid contents in *Capsicum annuum*. *Horticulture Research*, *9*, uhac204. https://doi.org/10.1093/hr/uhac204

Kurapati, S., Kommineni, R., Variath, M. T., Manohar, S. S., Vemulapalli, P., Vemireddy, L. N. R., & Pasupuleti, J. (2021). Localization and gene action studies for kernel iron and zinc concentration

in groundnut (*Arachis hypogaea* L.). *Euphytica*, *217*(7), Article 143. https://doi.org/10.1007/s10681-021-02872-2

Li, D., Xu, Z., Gu, R., Wang, P., Lyle, D., Xu, J., Zhang, H., & Wang, G. (2019). Correction: Enhancing genomic selection by fitting large-effect SNPs as fixed effects and a genotype-by-environment effect using a maize BC1F3:4population. *PLoS One*, *14*(12), e0226592. https://doi.org/10.1371/journal.pone.0226592

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., Singh, R. P., Autrique, E., & de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *G3 Genes|Genomes|Genetics*, *5*(4), 569–582. https://doi.org/10.1534/g3.114.016097

Lubanga, N., Ifie, B. E., Persa, R., Dieng, I., Rabbi, I. Y., & Jarquin, D. (2025). Sparse testing designs for optimizing resource allocation in multi-environment cassava breeding trials. *The Plant Genome*, *18*(1), e20558. https://doi.org/10.1002/tpg2.20558

Lubanga, N., Massawe, F., & Mayes, S. (2021). Genomic and pedigree-based predictive ability for quality traits in tea (*Camellia sinensis* (L.) O. Kuntze). *Euphytica*, *217*(3), Article 32. https://doi.org/10.1007/s10681-021-02774-3

Lubanga, N., Massawe, F., Mayes, S., Gorjanc, G., & Bančič, J. (2023). Genomic selection strategies to increase genetic gain in tea breeding programs. *The Plant Genome*, *16*(1), e20282. https://doi.org/10.1002/tpg2.20282

Mageto, E. K., Crossa, J., Pérez-Rodríguez, P., Dhliwayo, T., Palacios-Rojas, N., Lee, M., Guo, R., San Vicente, F., Zhang, X., & Hindu, V. (2020). Genomic prediction with genotype by environment interaction analysis for kernel zinc concentration in tropical maize germplasm. *G3 Genes|Genomes|Genetics*, *10*(8), 2629–2639. https://doi.org/10.1534/g3.120.401172

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Monteverde, E., Gutierrez, L., Blanco, P., Pérez de Vida, F., Rosas, J. E., Bonnecarrère, V., Quero, G., & McCouch, S. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) grown in subtropical areas. *G3 Genes|Genomes|Genetics*, *9*(5), 1519–1531. https://doi.org/10.1534/g3.119.400064

Mora-Escobedo, R., Hernández-Luna, P., Joaquín-Torres, I. C., Ortiz-Moreno, A., & Robles-Ramírez, M., & del, C. (2015). Physicochemical properties and fatty acid profile of eight peanut varieties grown in Mexico. *CyTA—Journal of Food*, *13*(2), 300–304. https://doi.org/10.1080/19476337.2014.971345

Okaron, V., Mwololo, J., Gimode, D. M., Okello, D. K., Avosa, M., Clevenger, J., Korani, W., Ssemakula, M. O., Odong, T. L., & Odeny, D. A. (2024). Using cross-country datasets for association mapping in *Arachis hypogaea* L. *The Plant Genome*, *17*(4), e20515. https://doi.org/10.1002/tpg2.20515

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Pérez-Cabal, M. A., Vazquez, A. I., Gianola, D., Rosa, G. J. M., & Weigel, K. A. (2012). Accuracy of genome-enabled prediction in a dairy cattle population using different cross-validation layouts. *Frontiers in Genetics*, *3*, Article 27. https://doi.org/10.3389/fgene.2012.00027

Persa, R., Grondona, M., & Jarquin, D. (2021). Development of a genomic prediction pipeline for maintaining comparable sample sizes in training and testing sets across prediction schemes accounting for the genotype-by-environment interaction. *Agriculture*, *11*(10), 932. https://doi.org/10.3390/agriculture11100932

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Rice, B., & Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *The Plant Genome*, *12*(1), 180052. https://doi.org/10.3835/plantgenome2018.07.0052

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M. N., Jannink, J., & Sorrells, M. E. (2014). Genomic selection for quantitative adult plant stem rust resistance in wheat. *The Plant Genome*, *7*(3), plantgenome2014.02.0006. https://doi.org/10.3835/plantgenome2014.02.0006

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., & McCouch, S. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, *116*(4), 395–408. https://doi.org/10.1038/hdy.2015.113

Tabe-Ojong, M. P. J., Lokossou, J. C., Gebrekidan, B., & Affognon, H. D. (2023). Adoption of climate-resilient groundnut varieties increases agricultural production, consumption, and smallholder commercialization in West Africa. *Nature Communications*, *14*(1), 5175. https://doi.org/10.1038/s41467-023-40781-1

Technow, F., Messina, C. D., Totir, L. R., & Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate bayesian computation. *PLoS One*, *10*(6), e0130855. https://doi.org/10.1371/journal.pone.0130855

Tolhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M., & Gorjanc, G. (2022). Genomic selection using random regressions on known and latent environmental covariates. *Theoretical and Applied Genetics*, *135*(10), 3393–3415. https://doi.org/10.1007/s00122-022-04186-w

Toomer, O. T. (2018). Nutritional chemistry of the peanut (*Arachis hypogaea*). *Critical Reviews in Food Science and Nutrition*, *58*(17), 3042–3053. https://doi.org/10.1080/10408398.2017.1339015

USDA-NASS. (2021). *U.S. Department of Agriculture. 2020*. https://www.nass.usda.gov/Newsroom/archive/2021/index.php

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting power and accuracy for genomic association and prediction. *Genomics, Proteomics & Bioinformatics*, *19*(4), 629–640. https://doi.org/10.1016/j.gpb.2021.08.005

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Yates, F., & Cochran, W. G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, *28*(4), 556–580. https://doi.org/10.1017/S0021859600050978

Yeleliere, E., Antwi-Agyei, P., & Baffour-Ata, F. (2023). Impacts of climate change on the yields of leguminous crops in the Guinea Savanna agroecological zone of Ghana. *Regional Sustainability*, *4*(2), 139–149. https://doi.org/10.1016/j.regsus.2023.04.002

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., Olsen, M., Prasanna, B. M., Crossa, J., Yu, H., & Zhang, X. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Frontiers in Plant Science*, *8*, 1916. https://doi.org/10.3389/fpls.2017.01916

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Lubanga, N., Okaron, V., Gimode, D. M., Persa, R., Mwololo, J., Okello, D. K., Ssemakula, M. O., Odong, T. L., Abincha, W., Odeny, D. A., & Jarquin, D. (2025). Enhancing the prediction accuracy of groundnut yield by integrating significant markers and modeling genotype × environment interaction. *The Plant Genome*, *18*, e70105. https://doi.org/10.1002/tpg2.70105