frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 June 2025
DOI 10.3389/fpls.2025.1594736

OPEN ACCESS

EDITED BY
Pablo Velasco,
Spanish National Research Council (CSIC), Spain

REVIEWED BY
João Ricardo Bachega Feijó Rosa,
RB Genetics & Statistics Consulting, Brazil
Sikiru Adeniyi Atanda,
North Dakota State University, United States

*CORRESPONDENCE
Maurício S. Araújo
mauricioaraujj@usp.br
José B. Pinheiro
Jbaldin@usp.br

RECEIVED 16 March 2025
ACCEPTED 15 May 2025
PUBLISHED 06 June 2025

CITATION
Araújo MS, Pavan JPS, Stella AA, Fregonezi BF, Lima NF, Leles EP, Santos MF, Goldsmith P, Chigeza G, Diers BW and Pinheiro JB (2025) Optimizing soybean variety selection for the Pan-African Trial network using factor analytic models and envirotyping.
Front. Plant Sci. 16:1594736.
doi: 10.3389/fpls.2025.1594736

Optimizing soybean variety selection for the Pan-African Trial network using factor analytic models and envirotyping

Maurício S. Araújo[1*], João P. S. Pavan[1], André A. Stella[1], Bruno F. Fregonezi[1], Natally F. Lima[2], Erica P. Leles[3], Michelle F. Santos[3], Peter Goldsmith[3], Godfree Chigeza[4], Brian W. Diers[3] and José B. Pinheiro[1*]

[1]Genetics Diversity and Breeding Laboratory, Department of Genetics, University of São Paulo, Piracicaba, São Paulo, Brazil, [2]Allogamous Plant Breeding Laboratory, Department of Genetics, University of São Paulo, Piracicaba, São Paulo, Brazil, [3]Feed the Future Innovation Lab, University of Illinois Urbana-Champaign, United States Agency for International Development (USAID), Washington, DC, United States, [4]International Institute of Tropical Agriculture, Consultative Group on International Agricultural Research (CGIAR), Ibadan, Oyo, Nigeria

Soybean is a global food and industrial crop, however, climate change significantly affects its grain yield. Therefore, the selection of varieties with high adaptation to target population of environments is imperative in Sub-Saharan Africa. This study aimed to identify soybean varieties with high overall performance and stability using multi-environment trial data from the Pan-African Soybean Trial Network. Additionally, we sought to determine the environmental factors influencing yield through envirotyping tools. In two South-Eastern African countries, a total of 169 soybean varieties were evaluated across 83 environments in 19 locations in Malawi (47 trials) and 14 locations in Zambia (36 trials). The trials followed a randomized complete block design with three replications. Data for 37 environmental features were obtained from NASA POWER and SoilGrids. We fitted factor analytic models (FA) to estimate genotype adaptation across environments. Additionally, we applied an environmental kernel approach and the XGBoost method to assess the number of mega-environments. The FA model with four factors provided the best fit, explaining 82.44% and 81.95% of the variance and the average semi-variance ratio (ASVR), respectively. Approximately, 59.6% of the genotype-by-environment interaction were crossover. Varieties V025, V035, and V158 exhibited high yield potential and reliability but displayed moderate stability. Three mega-environments were identified, with growing degree days, mean temperature, and photosynthetically active radiation use efficiency being the most associated features for soybean grain yield. To enhance the identification of variety adaptation in these environments, integrating machine learning models with crop growth modeling is essential to assess associations between environmental features and soybean yield.

KEYWORDS

***Glycine max*, linear mixed models, environmental data, adaptation, stability**

1 Introduction

Soybean (*Glycine max* L.) is a commodity crop of great global importance (Mishra et al., 2024). Its grains are widely utilized in agro-industry, primarily for oil production, high-protein food products, and animal feed formulation (Zhi et al., 2020). Its nutritional composition is determined by proteins, oil, carbohydrates, isoflavones, and minerals. However, population growth and the ever increasing demand for protein sources, both for human consumption and animal feed, highlights the need to expand global soybean production (Messina, 2022). In this context, improving production efficiency in new agricultural frontiers through the development of more adapted varieties becomes essential to ensure food security for future generations. In light of that, genetic improvement programs have focused on developing highyielding varieties with resistance to pests and diseases, as well as broad adaptation to target environmental conditions (Favoretto et al., 2025). These advancements have been driven by the optimization of breeding strategies and the adoption of effective agricultural practices (Carciochi et al., 2019).

Plant breeders rely on multi-environment trials (METs) to evaluate genotype performance across diverse conditions, representing the target population of environments (TPE) and assessing genotype adaptation to specific or broad environments (Poupon et al., 2023; Malosetti et al., 2016; Costa-Neto et al., 2023; Vitale et al., 2024). When crossover interactions occur, genotype rankings vary across environments (Fehr, 1987; Cooper and Delacy, 1994), and neglecting genotype-by-environment (G×E) interaction can introduce some bias and reduce selection efficiency (van Eeuwijk et al., 2016). To quantify G×E interaction, various methods have been explored, each with distinct assumptions and applications. These include analysis of variance (Plaisted and Peterson, 1959; Shukla, 1972), regression models (Finlay and Wilkinson, 1963; Eberhart and Russell, 1966), non-parametric approaches (Lin and Binns, 1998), multiplicative models such as GGE Biplot (Yan et al., 2000) and AMMI (Gauch and Zobel, 1997; Gauch, 2008), linear mixed models (Henderson, 1949, 1950), factor analytic (FA) models — which are extensions of linear mixed models — (Piepho, 1997a, b; Smith et al., 2001b), and Bayesian approaches (Cotes et al., 2006), all widely applied in plant breeding.

Factor analytic (FA) models are a specific class of linear mixed models (LMMs) that are particularly robust in handling diverse data structures, especially unbalanced data. As a parsimonious approximation of the unstructured model, they indirectly construct the full genetic covariance structure, accounting for heterogeneous variances and covariances. This capability allows for the exploration of genetic covariance between environments or traits, making FA models well-suited for METs. Their effectiveness stems from dimensionality reduction through latent variables, known as factors (Smith et al., 2001b; Piepho, 1998). Additionally, as linear mixed models, they facilitate the inclusion of relatedness information, whether genomic (marker-based) or ancestral (pedigree) (Smith et al., 2005). Building on these principles, Smith and Cullis (2018) introduced the Factor Analytic Selection Tools (FAST), which incorporate parameters for assessing overall performance (OP) and stability via Root Mean Square Deviation (RMSD). These metrics enhance breeders' decision-making by providing a statistically sound and comprehensive evaluation framework. Today, FA models are the benchmark for handling unbalanced MET data within the LMM framework (Tolhurst et al., 2022; Araújo et al., 2024), with recent insights by Piepho and Williams (2024) emphasizing their utility in predicting genotype performance in METs.

Beyond selecting the most appropriate statistical methods, modern plant breeding demands additional tools to enhance the predictive ability of models. Over the past decade, environmental features have emerged as valuable resources for improving predictions in METs (Xu, 2016; Resende et al., 2024). Although the integration of environmental data into genetic analyses is not a new concept (Van Eeuwijk and Elgersma, 1993; Wood, 1976), advances in hardware and data processing have enabled the use of large datasets, facilitating the incorporation of environmental features into statistical genetic models. Enviromics, a specialized field at the intersection of environmental data, statistics, and quantitative genetics, leverages plant ecophysiology to better understand how environmental factors influence plant development and the plasticity of key agronomic traits (Costa-Neto and Fritsche-Neto, 2021). In this context, envirotypes represent all sources of environmental variation affecting plant development and can serve as environmental markers in statistical genetic models, aiding in the prediction of genotypic performance in non-evaluated environments (Xu, 2016; Resende et al., 2025).

The addition of information derived from Geographic Information System (GIS) techniques into predictive models has been encouraged to improve the efficiency of breeding programs (Guarino et al., 2002). An initial effort was made by Booth (1990) aiming to indicate climatically suitable regions for the introduction of tree species at a global scale based on the environmental conditions where they were collected. Annicchiarico et al. (2006) assessed how GIS-based methodologies could aid the recommendation of durum wheat genotypes in MET, as compared to traditional methodologies. The integration of machine learning, quantitative genetics, enviromics, and GIS tools enhances the identification of environmental patterns in target environments. These resources enable the exploration of environmental homogeneity and the determination of factors influencing climatic variability, facilitating the incorporation of G×E interaction and the selection of cultivars adapted to specific conditions.

Soybean variety selection is becoming increasingly important due to its high nutritional value and economic significance in the global market. Despite its potential, generally, the adaptation of soybean varieties to Sub-Saharan African environments specifically in the South-Eastern countries of Malawi and Zambia remains largely unexplored, limiting the availability of high-performing cultivars suited to the region's diverse agro-ecological conditions. This gap is particularly concerning given the rapid population growth and the escalating demand for affordable protein based food sources, which underscore the necessity of expanding and

optimizing soybean production. Moreover, climate change exacerbates environmental variability, increasing the urgency for resilient cultivars capable of maintaining stable yields across unpredictable conditions (Sousa et al., 2019). To address this challenge, this study employs advanced selection tools to identify superior varieties with high overall performance and stability within the Pan-African Trials Network. Furthermore, the integration of envirotyping methodologies enables the exploration of associations between environmental variables and G×E interactions, facilitating the identification of specific adaptations critical for sustainable soybean production in Malawi and Zambia.

2 Material and methods

2.1 Phenotypic data and field trials

Soybean variety yield trials are part of the Soybean Innovation Lab (SIL). This program was established to select high-yielding varieties adapted to target population environments (TPE) in Africa, to support cultivation by smallholder farmers. This initiative led to the creation of the Pan-African Soybean Variety Trials (PATs) through partnerships with the African Agricultural Technology Foundation (AATF), the Syngenta Foundation for Sustainable Agriculture (SFSA), and the International Institute of Tropical Agriculture (IITA) (Santos, 2019). The PATs program plays a key role in identifying and disseminating varieties capable of adapting to diverse Agro-ecological conditions, thereby contributing to enhanced food security and economic growth across selected Africa countries. The African continent was divided into 33 Agro-ecological Zones (AEZs), classified according to criteria such as climatic zones (tropical, temperate, etc.), length of the growing season, soil type, and altitude, with a resolution of 5 arc-minutes (≈ 9.2 km × 9.2 km) (Figure 1) (Food and Agriculture Organization of the United Nations, 2025).

A total of 169 soybean varieties were evaluated over the 2017/18 to 2023/24 seasons (Supplementary Figure S1) in trials conducted in two South-Eastern African countries of Malawi and Zambia. In Malawi, 47 trials were conducted across 19 distinct locations, each defined as the interaction between location and season (Figure 1B). In Zambia, 36 environments were carried out across 14 locations (Figure 1C). The trials followed a randomized complete block design (RCBD) with three replications. Each plot consisted of four rows measuring five meters in length (4 × 5 m), spaced 50 cm apart, with 20 plants per row. grain yield (kg ha^{-1}) was measured from the two central rows. Agronomic management practices adhered to the specific technical recommendations for soybean cultivation.

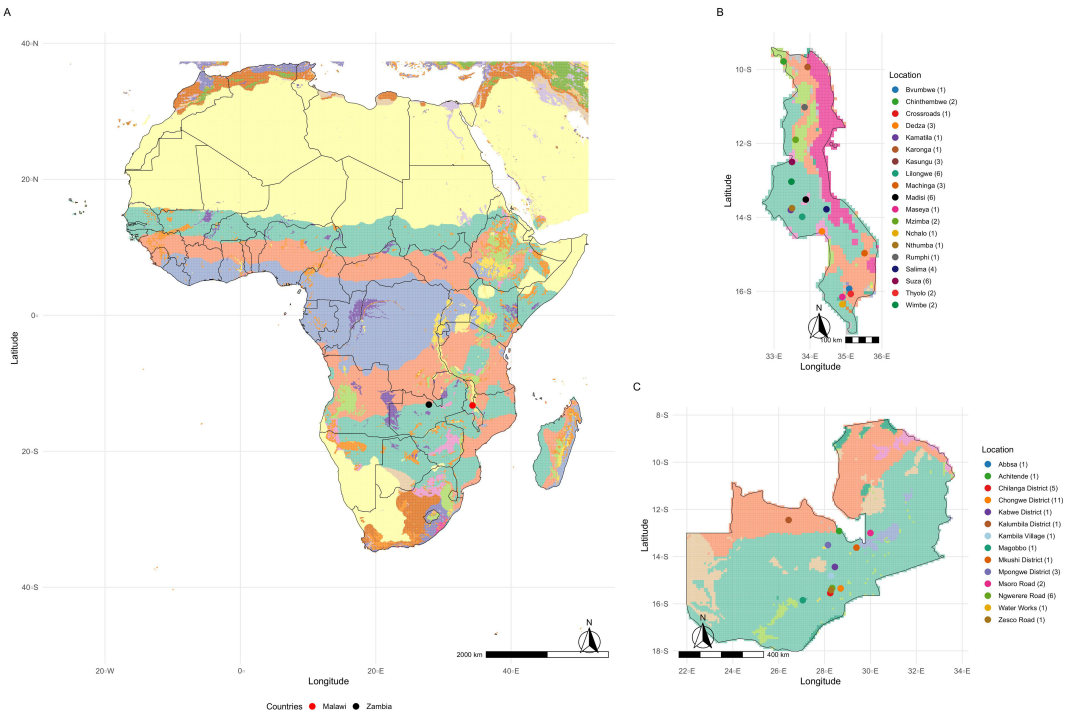

FIGURE 1
(A) displays the map of Africa with Agro-ecological Zones (AEZ) classified into 33 distinct categories based on climatic variables, topography, and the chemical and physical properties of the soil. Each color on the map represents a specific AEZ class. Refer to Food and Agriculture Organization of the United Nations (2025) for detailed identification of each class. The red and black points on the map highlight the countries of Malawi and Zambia, respectively. **(B)** presents the map of Malawi, highlighting its respective AEZs. The colors of the points indicate the locations where the trials were conducted, and the number in parentheses represents the number of trials carried out at each site. **(C)** shows Zambia with the distribution of trial locations, along with the number of experimental trials conducted in each region.

2.2 Envirotyping

Throughout the crop's growing season, we collected data on 37 environmental features (Table 1). Each genotype's sowing and harvesting dates were used to retrieve environment-specific variables, enabling the characterization of trial conditions and the assessment of their similarity. The environmental covariates encompassed geographic, climatic, and soil information. The climatic variables were obtained using the EnvRtype package (Costa-Neto et al., 2021), which accesses the NASA POWER database (https://power.larc.nasa.gov/) (Sparks, 2018; NasaPower, 2022). Soil attributes were retrieved from the SoilGrids database via API using the httr package for web access (Wickham, 2023) and jsonlite for JSON parsing (Ooms, 2014). Static variables such as altitude and soil properties were associated with the trial location coordinates.

Prior to kernel construction, we applied quality control filters to remove missing or inconsistent values and standardized all continuous variables using Z-score normalization to ensure comparability across different measurement scales (Equation 1):

$$Z_{ij} = \frac{x_{ij} - \bar{x}_{\cdot j}}{s_{\cdot j}} \tag{1}$$

where $\bar{x}_{\cdot j}$ and $s_{\cdot j}$ denote the mean and standard deviation, respectively, of the j-th variable across all locations.

To reduce multicollinearity, we examined the Pearson correlation matrix and flagged variable pairs with correlation coefficients. Redundant variables were removed based on domain knowledge and exploratory principal component analysis (PCA), which was implemented using the factoextra version 1.0.7 package (Kassambara and Mundt, 2016).

The final environment-by-variable matrix $\mathbf{W}$ was then used to compute the enviromic similarity kernel *KE* as described in Equation 2.

$$\mathbf{K}_E = \frac{\mathbf{W}\mathbf{W}^\top}{\text{trace}(\mathbf{W}\mathbf{W}^\top)/n} \tag{2}$$

where $\mathbf{W}^\top$ is the transpose of $\mathbf{W}$, and n is the number of environments. This standardization ensures unit trace, allowing comparability across analyses and interpretation of diagonal elements as average similarities. The matrix $\mathbf{W}$ contains standardized environmental covariates (e.g., climatic and soil variables), with rows representing environments (location-by-year combinations) and columns corresponding to environmental descriptors.

2.2.1 Identification of mega-environments

Initially, environments were grouped into mega-environments based on an enviromic similarity matrix, denoted as the enviromic kernel (*KE*). This matrix integrated 37 environmental covariates and grain yield. Hierarchical clustering was applied using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm (Sokal and Michener, 1958). The optimal number of clusters was defined using the Elbow method, and the most influential covariates were explored via principal component analysis (PCA) (Pearson, 1901). To prevent methodological circularity, the dataset was randomly split into training (70%) and test (30%) subsets prior to unsupervised learning. PCA and K-means clustering were applied exclusively to the training subset, and the resulting cluster assignments were used as categorical labels for model training.

Classification was performed using the XGBoost (*Extreme Gradient Boosting*) algorithm (Chen and Guestrin, 2016), implemented via the xgboost package. The model was configured for multi-class classification (multi:softmax) and trained using the first three principal components. The hyperparameters used were: tree depth of 6, learning rate (η) of 0.3, and 100 boosting iterations. The objective function minimized by the algorithm included both the predictive loss and regularization terms, and is expressed in Equation 3:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t), \tag{3}$$

where ℓ denotes the multinomial log-loss function, and the regularization term $\Omega(f_t)$ for each tree f_t is defined in Equation 4:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2, \tag{4}$$

in which T is the number of leaves, w_j is the score on leaf j, γ is the complexity penalty for the number of leaves, and λ controls the L2 regularization on leaf weights. All analyses were performed in R (version 4.3.1) using the following packages: cluster (Maechler et al., 2019), caret (Kuhn et al., 2020), xgboost (Chen et al., 2022), and dendextend (Galili, 2015).

To explore the relationship between environmental variables and grain yield, we fitted a multiple linear regression model using the adjusted mean yield for each environment as the response variable. The model is specified in Equation 5:

$$\mathbf{y} = \mu + \sum_{i=1}^{t} \boldsymbol{\beta}_i \mathbf{X}_i + \mathbf{e} \tag{5}$$

where $\mathbf{y}$ represents the adjusted mean yield in each environment; μ is the intercept of the model, corresponding to the overall mean yield; $\boldsymbol{\beta}_i$ denotes the coefficient associated with the i-th environmental variable; $\mathbf{X}_i$ corresponds to the value of the i-th environmental feature; $\mathbf{e}$ is the random error term, assumed to follow a normal distribution with zero mean and constant variance. Adjusted means used as the response variable were obtained by fitting separate linear mixed models for each environment, in which genotype was included as a fixed effect and replication as a random effect. From these models, empirical best linear unbiased estimates (eBLUEs) of genotype means were extracted. Subsequently, the mean of the eBLUEs within each environment was calculated and used as the environment-level adjusted mean in the subsequent analyses.

2.3 Statistic analysis

We analyzed the phenotypic data using the linear mixed-effects model described by Henderson (1949) and Henderson (1950). Estimation of variance components was performed using the

10.3389/fpls.2025.1594736

TABLE 1 Summary statistics of 37 environmental features grouped into geographical, climatic, and soil-related categories.

Class	Features	ID	Unit	Min	Mean	Max
Geographical	Altitude	alt	meters (m)	70.00	1039.00	1359.00
Climatic	Mean temperature	tmean	°C	17.57	21.90	26.29
	Maximum temperature	tmax	°C	24.23	27.11	31.27
	Minimum temperature	tmin	°C	11.59	17.60	23.68
	Precipitation	prec	mm/day	0.02	5.66	11.63
	Wind speed	wsm	m/s	1.60	2.27	4.07
	Relative humidity	rhm	%	49.96	77.31	88.06
	Dew point temperature	tmdew	°C	9.19	17.13	21.43
	Longwave radiation	lw	$MJ/m^2/day$	28.96	32.41	35.77
	Shortwave radiation	sw	$MJ/m^2/day$	16.79	20.08	22.89
	Growing degree days	gdd	°C d−1	10.33	14.36	18.38
	Radiation use efficiency	fue	–	0.47	0.65	0.84
	Temperature range	tmrange	°C day	4.67	9.52	13.94
	Vapor pressure deficit	vpd	kPa	0.43	0.84	1.81
	Slope of vapor pressure curve	spv	kPa/°C	0.13	0.17	0.20
	Potential evapotranspiration	etp	mm/day	7.63	9.02	10.38
	Precipitation deficit	petp	mm/day	-9.29	-3.36	3.19
	Total precipitation	totprec	mm	2.69	772.73	1451.87
	Average precipitation	aveprec	mm/day	0.02	5.66	11.63
	Evapotranspiration tolerance	etptol	mm	847.00	1300.00	2185.00
	Water balance	watbal	mm	-1802.70	-526.80	376.20
Soil	Bulk density of fine earth	bdod	kg/m^3	120.00	142.60	155.00
	Cation exchange capacity	cec	cmol/kg	60.00	89.28	142.00
	Coarse fragments volume	cfvo	%	2.00	23.62	67.00
	Clay content	clay	%	105.00	205.40	436.00
	Nitrogen content	nit	g/kg	84.00	118.00	170.00
	Organic carbon density	ocd	kg/m^3	165.00	210.90	257.00
	Soil pH (H_2O)	phh2o	–	54.00	61.19	64.00
	Sand content	sand	%	336.00	653.10	811.00
	Silt content	silt	%	63.00	141.50	257.00
	Soil organic carbon	soc	g/kg	115.00	153.70	206.00
	Soil water content at 10 kPa	wv0010	–	249.00	308.30	383.00
	Soil water content at 33 kPa	wv0033	–	184.00	230.90	331.00
	Soil water content at 1500 kPa	wv1500	–	68.00	104.20	199.00
	Soil temperature	tsoil	°C	226.17	253.46	292.83
	Temperature seasonality	sts	°C	86.10	155.20	255.70
	Isothermality	iso	–	-84.60	13.67	30.70
	Mean diurnal range	mdr	–	-2.00	1.17	2.40

Data were collected from soybean varieties evaluated in Malawi and Zambia during the 2017–2024 seasons through the Pan-African Trials Network. Climatic features were obtained from NASA POWER, and soil variables from SoilGrids.

10.3389/fpls.2025.1594736

residual maximum likelihood (REML) method (Patterson and Thompson, 1971). The model was implemented using the ASReml-R package (version 4.1.2) (Butler et al., 2018) within the R software environment (R Core Team, 2022). Prior to model fitting, we assessed the validity of key model assumptions through standard residual diagnostics. The normality of residuals was evaluated using quantile–quantile (Q-Q) plots, as recommended by Kozak and Piepho (2018). Residual independence was assumed, and heteroscedasticity across environments was addressed by specifying a diagonal residual covariance matrix, allowing each environment to have its εown residual variance. The applied model follows Equation 6.

$$\mathbf{y} = \mu\mathbf{1}_n + \mathbf{X}_1\mathbf{s} + \mathbf{X}_2\mathbf{b} + \mathbf{Z}_1\mathbf{g} + \boldsymbol{\epsilon} \tag{6}$$

In which $\mathbf{y}^{(n\times1)}$ is the vector of phenotypic data across t environments, where $n = \sum_{j=1}^{t} n_j$, and n_j is the number of observations in each environment j; μ is the model intercept; $\mathbf{s}^{(t\times1)}$ is the vector of fixed effects for environments; $\mathbf{b}^{(b\times1)}$ is the vector of fixed effects for the blocks, where $b = \sum_{j=1}^{t} b_j$ and b_j is the number of blocks within environment j; $\mathbf{g}^{(v\times1)}$ is the vector of random effects for the v genotypes evaluated across environments, where $\mathbf{g} \sim \text{MVN}(\mathbf{0}, \mathbf{G}\otimes\mathbf{I}_v)$. Although genotypes are conceptually common across environments, the factor analytic (FA) model implicitly nests genotypes within environments by modeling the genotype-by-environment interaction through the $\mathbf{G}$ matrix, which captures the variance–covariance structure among environments. $\boldsymbol{\epsilon}^{(n\times1)}$ is the vector of residual effects, where $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{R}\otimes\mathbf{I}_n)$. Here, $\mathbf{R}$ is a diagonal matrix of order t, allowing for heterogeneous residual variances across environments, i.e., $\mathbf{R} = \text{diag}(\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \ldots, \sigma^2_{\epsilon_t})$. $\mathbf{X}_1^{(n\times t)}$, $\mathbf{X}_2^{(n\times b)}$, and $\mathbf{Z}_1^{(n\times v)}$, represent the incidence matrices of the vectors accompanying them in the model. $\mathbf{1}_n^{(n\times1)}$ is a vector of ones; and $\mathbf{I}_v$ and $\mathbf{I}_n$ areidentity matrices of orders v and n, respectively.

The genotypic effect vector $\mathbf{g}$, for an FA model of order K, is then expressed in Equation 7:

$$\mathbf{g} = (\hat{\boldsymbol{\Lambda}}\otimes\mathbf{I}_v)\hat{f} + \delta \tag{7}$$

where $\hat{\boldsymbol{\Lambda}}^{(t\times K)}$ is the matrix containing the K factor loadings for each of the t environments $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_t)$, $\hat{f}^{(Kv\times1)}$ is the vector containing the v factor scores of genotypes in each environment $[\mathbf{f}_1^{\mathrm{T}}, \mathbf{f}_2^{\mathrm{T}}, \ldots, \mathbf{f}_v^{\mathrm{T}}]^{\mathrm{T}}$, and $\hat{\boldsymbol{\delta}}^{(tv\times1)}$ is the vector representing the model's lack of fit. The joint distribution of $\hat{\mathbf{f}}$ and $\hat{\boldsymbol{\delta}}$ is given in Equation 8:

$$\begin{pmatrix}\hat{f}\\ \hat{\delta}\end{pmatrix} \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbf{I}_K\otimes\mathbf{I}_v & 0\\ 0 & \Psi\otimes\mathbf{I}_v\end{pmatrix}\right] \tag{8}$$

In which $\Psi^{(t\times t)}$ is the diagonal matrix of specific variances $(\Psi_1, \Psi_2, \ldots, \Psi_t)$ for each environment, i.e., what the factors couldn't capture.

The selection of the most parsimonious model was based on the explained variance v_{kt}, which was utilized for all K factors and for each factor per environment (k-th) (Equation 9) (Smith et al., 2015), and the average semi-variance ratio (ASVR) (Equation 10) (Piepho, 2019; Chaves et al., 2023), respectively.

$$v_{k_t} = \frac{\hat{\lambda}^{\star^2}_{k_t} d_k}{\sum_{k=1}^{K}\hat{\lambda}^{\star^2}_{k_t} d_k + \hat{\psi}_t} \times 100 \tag{9}$$

$$ASVR = \frac{\frac{2}{t\times(t-1)}\sum_{t=1}^{t-1}\sum_{t'=t+1}^{t}\frac{1}{2}\times(\sum_{k=1}^{K}\hat{\lambda}^{\star^2}_{k_t} + \sum_{k=1}^{K}\hat{\lambda}^{\star^2}_{k_{t'}}) - \sum_{k=1}^{K}\hat{\lambda}^{\star}_{k_t}\hat{\lambda}^{\star}_{k_{t'}}}{\frac{2}{t\times(t-1)}\sum_{t=1}^{t-1}\sum_{t'=t+1}^{t}\frac{1}{2}\times[(\sum_{k=1}^{K}\lambda^{\star^2}_{k_t} + \psi_t) + (\sum_{k=1}^{K}\hat{\lambda}^{\star^2}_{k_{t'}} + \psi_{t'})] - \sum_{k=1}^{K}\hat{\lambda}^{\star}_{k_t}\hat{\lambda}^{\star}_{k_{t'}}} \times 100 \tag{10}$$

The generalized heritability by Cullis et al. (2006) was obtained through the Equation 11:

$$H^2 = 1 - \left(\frac{\bar{v}_{BLUP}}{2\sigma^2_g}\right) \tag{11}$$

Where $\bar{v}_{BLUP}$ is the average pairwise prediction error variance, and σ^2_g is the genotypic variance.

The coefficient of variation was calculated using Equation 12.

$$CV = \frac{\hat{\sigma}_e}{\hat{\mu}} \tag{12}$$

Where $\hat{\sigma}_e$ is the estimated residual standard deviation, and $\hat{\mu}$ is the overall mean of each environment.

We estimated the genetic correlation between pairs of environments as described by Cullis et al. (2010), given by Equation 13:

$$\rho_{g_{tt'}} = \frac{\sum_{k=1}^{K}\lambda_{tk}\lambda_{t'k}}{\sqrt{\hat{\sigma}^2_{gt}\hat{\sigma}^2_{gt'}}} = \mathbf{DGD} \tag{13}$$

where, $\hat{\sigma}^2_{gt}$ and $\hat{\sigma}^2_{gt'}$ represent the genotypic variance components in environments t and t' respectively, while the matrix $\mathbf{D}$ is a diagonal matrix composed of the reciprocal square roots of the diagonal elements of matrix $\mathbf{G}$.

The crosser interaction was estimated using Equation 14:

$$\sigma^2_{ge_{rk}} = 1 - \frac{\sigma^2\left(\sqrt{\sigma^2_{g_t}}\right)}{\sigma^2_{ge}} \tag{14}$$

The variance component for the genotype-by-environment G ×E interaction, denoted as σ^2_{ge}, was estimated using a compound symmetry (CS) model. In this structure, the variance-covariance matrix of the genetic effects is definedas $\sigma^2_g\mathbf{J} + \sigma^2_{ge}\mathbf{I}_j$, where $\mathbf{J}$ is a matrix of ones. The CS model was adopted following the conceptual framework proposed by Cooper and Delacy (1994), which enables the partitioning of G×E interaction into simple (related to genotypic response consistency) and crossover (due to changes in genotype ranking) components. By assuming equal genetic variances and covariances across environments, the CS structure provides a neutral and interpretable baseline, from which deviations can be attributed to crossover interaction. This approach avoids conflating model-derived correlation structures, such as those in FA models, with the theoretical decomposition of the G×E variance.

10.3389/fpls.2025.1594736

2.4 Factor Analytic Selection Tools

To address identifiability issues and enable biological interpretability in factor analytic (FA) models, we adopted the constraints implemented in ASReml-R (Butler et al., 2018), as described by Smith et al. (2021). Specifically, for models with more than one factor ($K > 1$), the upper triangular elements of the loading matrix $\mathbf{\Lambda}$ were set to zero, and the factor scores were assumed to have a diagonal covariance matrix with decreasing elements. The constrained loading matrix is denoted as Λ^*, and the corresponding factor scores as $\boldsymbol{f}^*$. To recover the original (rotated) parameterization while preserving the variance structure implied by the model, we performed a singular value decomposition (SVD) of $\mathbf{\Lambda}^*$ as follows in Equation 15:

$$\Lambda^{\star} = \mathbf{U}\mathbf{L}^{1/2}\mathbf{V}^{\top}, \tag{15}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices of dimensions $t \times K$ and $K \times K$, respectively, and $\mathbf{L}$ is a diagonal matrix with singular values sorted in decreasing order. The final rotated loading matrix is then obtained as $\Lambda = \Lambda^{\star}\mathbf{V}\mathbf{L}^{-1/2} = \mathbf{U}$, and the diagonal matrix of factor variances is $\mathbf{D} = \mathbf{L}$. Accordingly, the scores f are reconstructed as $(\mathbf{L}^{1/2}\mathbf{V}^{\top} \otimes \mathbf{I}_v)f^{\star}$, ensuring that the variance of the factors satisfies $\text{var}(f) = \mathbf{D} \otimes \mathbf{I}_v$, as required for proper modeling of the random effects in the FA structure. These constraints facilitate identifiability and maintain the interpretability of the latent dimensions while preserving the implied genetic covariance structure across environments.

To support genotype selection within the environments evaluated, we used FA Models and applied the selection tools proposed by Smith and Cullis (2018). Specifically, the overall performance (OP_v) (Stefanova et al., 2009) of the v-th genotype was calculated using Equation 16:

$$OP_v = \frac{1}{t}\sum_{t=1}^{T}\hat{\lambda}^{*}_{1t}\tilde{f}^{*}_{1v} \tag{16}$$

In the provided equations, $\hat{\lambda}^{\star}_{1t}$ represents the rotated factor loading associated with the t-th environment for the first latent factor, and $\tilde{f}^{*}_{1v}$ denotes the rotated score of the v-th genotype for the first latent factor.

The remaining factors evaluate the stability parameter. The overall stability of the v-th genotype can be calculated by the root mean square deviation ($RMSD_v$) using the following Equation 17:

$$RMSD_i = \sqrt{\frac{1}{t}\sum_{t=1}^{T}\epsilon^{*}_{t}} \tag{17}$$

In the given expressions, ϵ^{*}_{vt} represents the deviation of the prediction associated with the first factor, which can be obtained as follows: $\epsilon^{*}_{vt} = \tilde{\boldsymbol{\beta}}_{vt} - \hat{\lambda}^{*}_{1t}\tilde{f}^{\star}_{1v}$, where $\tilde{\boldsymbol{\beta}}_{vt}$ is the linear combination of loadings and factor scores from all factors except the first.

The responsiveness of genotype v to the k-th factor ($\mathbf{RE_{vk}}$) was computed as shown in Equation 18:

$$RE_{vk} = (\bar{\lambda}^{*}_{k} - \bar{\lambda}^{*}_{k-})f^{*}_{vk} \tag{18}$$

where $\bar{\lambda}^{*}_{k+}$ and $\bar{\lambda}^{*}_{k-}$ represent the mean of the positive and negative rotated loadings, respectively, associated with the k-th latent factor.

We evaluated the reliability of each genotype using Equation 19:

$$R_v = 1 - \frac{PEV_v}{\bar{\sigma}^2_g} \tag{19}$$

In which PEV_v is the prediction error variance of the v-th genotype, and $\bar{\sigma}^2_g$ is the mean genotypic variance across environments.

An ideal genotype should present both high overall performance (OP_v) and low root mean square deviation ($RMSD_v$). The ideal genotype is selected based on the construction of an index ($FAST_v$) (Chaves et al., 2023; Cowling et al., 2023) (Equation 20):

$$FAST_v = \left(2 \times \frac{OP_v - \bar{OP}}{\sqrt{\sigma^2_{(OP)}}} - \frac{RMSD_v - \bar{RMSD}}{\sqrt{\sigma^2_{(RMSD)}}}\right) \times R_v \tag{20}$$

3 Results

Environmental kernel-based analyses incorporated climate and soil data from trials between 2017 and 2024. Principal component analysis (PCA) explained 52.7% of the total variance, with 33.1% attributed to the first principal component (PC1) and 19.6% to the second (PC2) (Figure 2A). Ten environmental features contributed most to climate variation among trials, with growing degree days (gdd), mean temperature (tmean), and photosynthetically active radiation use efficiency (fue) showing the strongest loadings in PC1 (Figure 2B). Hierarchical clustering applied to environmental similarities (based on the XGBoost model) suggested three mega-environment groups (Figure 2C). Regarding yield, the variables fue (radiation use efficiency), spv (seasonal precipitation variation), and tmrange (thermal amplitude) were associated with the largest regression coefficients. Additionally, fue, tmdew (mean dew point), wsm (soil moisture), and rhm (mean relative humidity) showed statistically significant associations with yield ($p < 0.05$) (Figure 2D).

The M4 model, with a factor analytic (FA) variance-covariance structure consisting of four factors (Table 2), exhibited the best fit for the dataset (Supplementary Figure S2). This selection was based on a threshold of 82.44% of the explained variance and 81.95 (%) of ASVR for the model with four factors (FA4). This criterion considered not only the explanatory capacity of the data but also the parsimony.

The Pan-Africa Trial Network demonstrated high experimental precision, with values ranging from 0.07 (M18s2E006) to 0.50 (M21s1E051). Broad-sense heritability coefficients (H^2) were also substantial, ranging from 0.46 (M19s2E015) to 0.85 (Z21s2E059) (Figure 3). Based on the distribution, the coefficient of variation (CV) showed a median of 0.229, with first and third quartiles of 0.183 and 0.272, respectively. Similarly, H^2 values had a median of 0.768, with Q1 = 0.710 and Q3 = 0.789 (Supplementary Figure S3). The average yield across the trials was 2,508.54 kg ha^{-1}; however, there was considerable variation among the experiments, ranging from 523.82 kg ha^{-1} (Z19s2E027) to 4,410.92 kg ha^{-1} (M22s2E062) (Supplementary Table S1). Considering the two countries

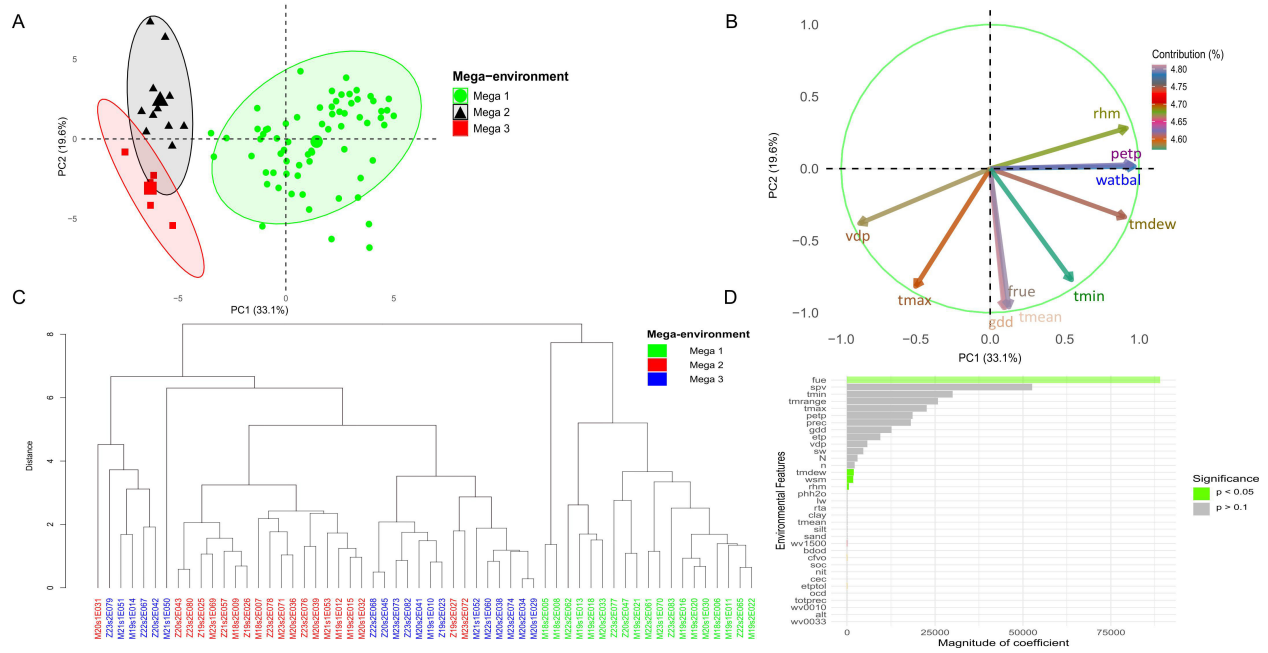

FIGURE 2
The **(A)** displays a principal component analysis (PCA) based on the environmental kernel, where the colors green, black, and red correspond to mega-environments, Mega 1, Mega 2, and Mega 3, respectively. The **(B)** highlights the environmental variables that contribute the most across all evaluation sites. The **(C)** presents a dendrogram based on the XGBoost model, used to test cluster mega-environments in Malawi and Zambia during the 2017 to 2023/24 growing seasons. Meanwhile, the **(D)** represents the variables with the greatest influence on yield performance in the trials.

individually, the average yield in Malawi was 3,171.10 kg ha^{-1}, while in Zambia it was 2,555.94 kg ha^{-1}.

Figure 4 shows a heatmap of pairwise genetic correlations between environments based on the factor analytic (FA) model. The strongest negative correlation was observed between trials Z19s2E028 and Z22s2E067 (r = −0.99), indicating a strong crossover interaction. Environments Z21s2E059, Z21s2E056, and Z21s2E057 showed high variability in correlations with other trials (SD > 0.48), suggesting inconsistent genotype responses. In contrast, Z19s2E024 and Z22s2E065 were among the most stable environments, with the lowest standard deviation in correlations (SD< 0.23). Trials such as Z20s2E046 and M20s2E039 exhibited the highest mean correlations with other environments (mean r > 0.20), highlighting their potential as representative environments for genotype recommendation. These results reflect substantial heterogeneity in genotype-by-environment interactions across trials conducted in Malawi and Zambia from 2017/18 to 2023/24, emphasizing the importance of environment-specific selection.

TABLE 2 Log-likelihood (LogL), deviance, number of parameters (Par.), explained variance (var%), and average semi-variance ratio (ASVR) for the models tested.

Model	LogL	Deviance	Par.	var (%)	ASVR (%)
M1	-57739.31	115478.62	174	37.23	34.00
M2	-57621.79	115243.58	260	65.52	62.58
M3	-57487.28	114974.56	345	77.31	77.17
M4	**-57395.72**	**114791.44**	**429**	**82.44**	**81.95**
M5	-57317.67	114635.34	512	87.90	87.58
M6	-57230.54	114461.08	594	93.11	92.83

The deviance (D) was calculated as $D = -2 \times \log L$. The model in bold is the selected one. The selection threshold was set at 80% for both explained variance (Var%) and ASVR(%), balancing goodness-of-fit and parsimony.

The varieties V020, V075, V137, V158, V035, V025, and V031 exhibited the best performance, as indicated by the highest OP values (Y-axis). Regarding stability, V013 showed the best fit, with the lowest RMSD values (X-axis) according to the FAST index. Varieties V025, V035, and V158 demonstrated high yield and reliability but exhibited medium stability (Figure 5).

Figure 6 presents the response of the variables to the second (Figure 6A), third (Figure 6B), and fourth (Figure 6C) factors. Responsiveness to specific factors facilitates the identification of environmental conditions associated with the environments that contribute to these factors. In this context, varieties V075, V020, and V137 demonstrated high overall performance and stability across factors 2, 3, and 4, respectively. Conversely, genotypes exhibiting low reliability (< 0.4%), such as V029, V110, V100, and V105 (Figure 5), also consistently demonstrated the poorest overall performances across all four evaluated factors, highlighting their limited adaptability and potential. Additionally, the variety V13 maintained the best fit in terms of OP, suggesting a higher stability and suitability under the tested conditions (Figure 6). These findings suggest that the associated factors may reflect meaningful environmental characteristics that can be leveraged for specific adaptation.

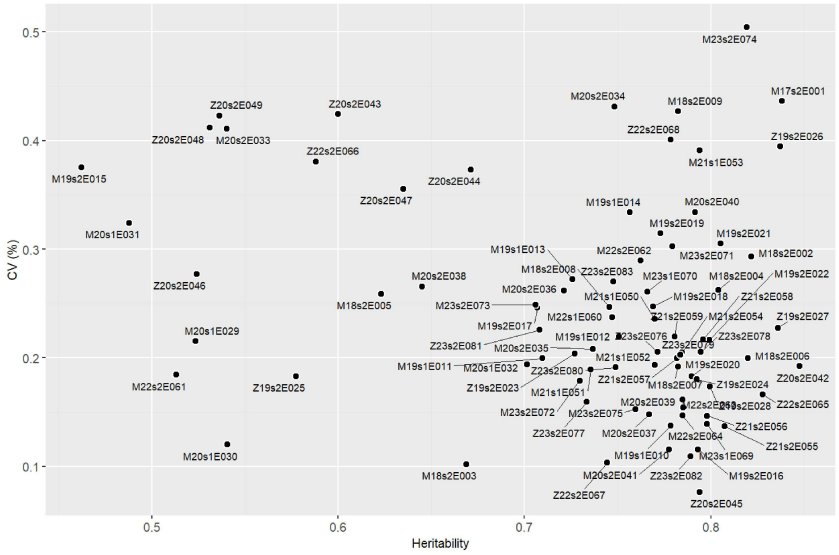

FIGURE 3
Scatterplot showing the relationship between the coefficient of variation (CV) and heritability (H^2) across 83 soybean yield trials conducted in Malawi and Zambia. Each point represents an environment (trial), positioned according to its heritability (X-axis) and CV (Y-axis), with labels indicating the environment codes.

4 Discussion

In this study, we applied FAST tools for selecting soybean varieties with high overall performance and stability in grain yield across METs. Additionally, we utilized GIS and envirotyping tools to explore associations between environmental features and grain yield, and to define mega-environments. Integrating environmental data into genetic-statistical models facilitated the characterization of G×E interaction patterns and their association with yield performance (Tolhurst et al., 2022). Furthermore, identifying environmental similarities between the experimental network and the TPE can enhance genetic gains through selection (Chaves et al., 2024).

The yield components of soybean are strongly influenced by the environmental effect (Araújo et al., 2024), thus being subject to the

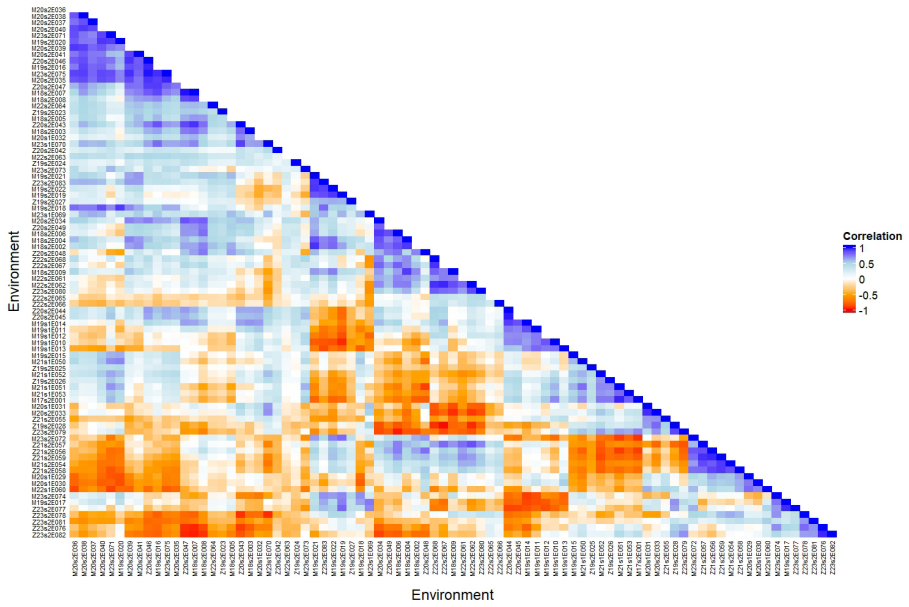

FIGURE 4
Heatmap showing pairwise genotypic correlations between environments based on the factor analytic (FA) model. Each cell represents the genetic correlation between two trials, with a color scale ranging from −1 to 1. Trial names are shown along both axes, and the figure emphasizes patterns of genetic similarity across environments. The evaluations were conducted in Malawi and Zambia from the 2017/18 to 2023/24 seasons, focusing on soybean grain yield.

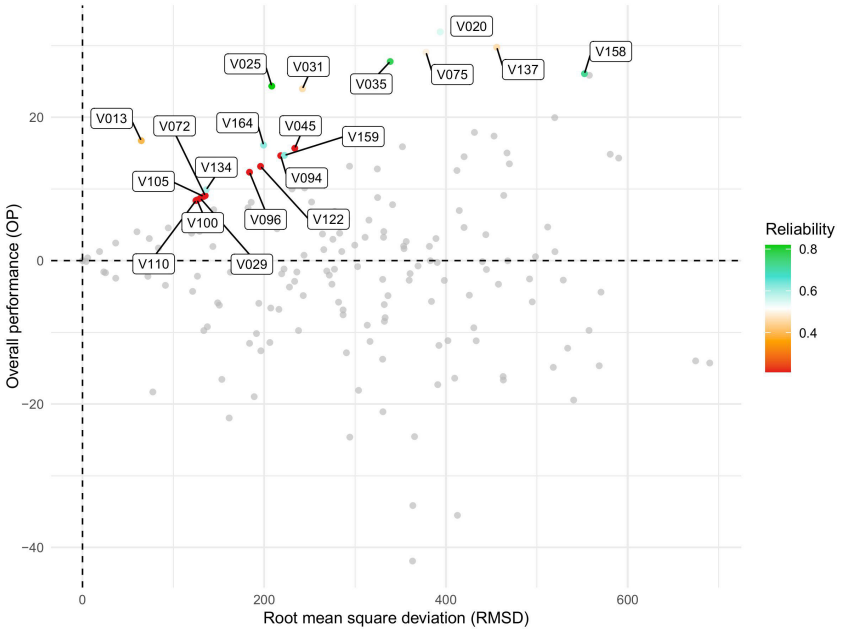

FIGURE 5
Graph showing the relationship between overall performance (OP) and stability, measured as root mean square deviation (RMSD), for soybean varieties evaluated in the Pan-African Trials Network across the 2017–2023/24 seasons. OP represents the mean performance of each genotype across environments, while RMSD quantifies the deviation from the average response, with lower values indicating higher stability. Each point corresponds to a genotype, and colors represent the reliability of the estimated performance–stability values, with the color scale ranging from red (low reliability) to green (high reliability). Axes labels and the legend have been enlarged to improve readability. This visualization summarizes results from the FAST (Factor Analytic Selection Tools) analysis.

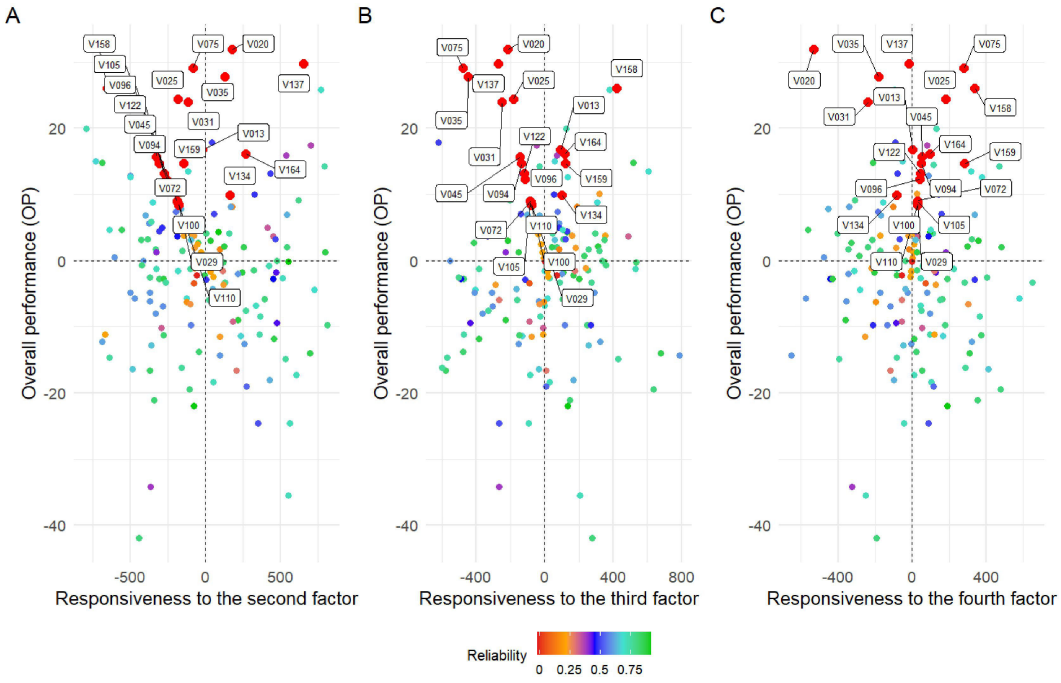

FIGURE 6
Overall performance (OP) vs. stability (RMSD) for all 169 soybean varieties from the Pan-African Trials Network. Biplot **(A)** represents responsiveness to the second factor, **(B)** to the third factor, and **(C)** to the fourth factor. Each point represents a genotype, with color indicating the reliability of its estimated performance–stability score. The color scale ranges from red (low reliability) to green (high reliability), as shown in the accompanying legend. Axes labels and the reliability legend have been enlarged to enhance readability.

G × E interaction (Meyer et al., 2024; Agoyi et al., 2024; Abebe et al., 2024). Over the years, overall performance and stability parameters have been assessed using methods based on analysis of variance (ANOVA) and linear regression. However, several limitations have been identified, such as: (*i*) modeling the genotype effect only as fixed; and (*ii*) the use of balanced data. We fitted a model of the genotype effect as random, employing the factor analytic structure (Piepho, 1997b; Smith et al., 2001a). This approach allows for the estimation of genetic parameters, using the heterogeneous random effect, enabling the evaluation of genetic progress over breeding cycles in various locations, seasons, and different agricultural years (Gogel et al., 2018; Chaves et al., 2023).

The genetic correlation heatmap in Figure 4 reveals high heterogeneity in genetic variances and low genetic correlations among environments, highlighting the crossover nature of the G × E interaction (Cullis et al., 2010). In other words, as the intensity of the interaction increases, the genetic correlation between pairs of environments decreases. This phenomenon is explained by the disparity in genetic variance values in each environment and the covariance between pairs of environments (Cooper and Delacy, 1994). Heinemann et al. (2022) demonstrated, in the context of crossover G×E interaction, the influence of environmental features on yield components. This can be explained by the direct effect of specific environmental variables on the adaptation of genotypes in METs. Therefore, it becomes crucial to identify environmental factors (climate, soil, spatial trends, among others) and genetic factors influencing the G × E interaction. To achieve this, robust methodologies are necessary to dissect this interaction and enable more precise selection (Kang et al., 1989).

The FA model stands out for its efficiency in handling diverse data structures (Piepho, 1998). This approach is commonly employed in MET, particularly during the stages of cultivar selection and recommendation (Kelly et al., 2007). This becomes possible due to the derivation of orthogonal factors from a set of correlated variables (Cullis et al., 2014). These factors represent linear combinations of the factor loadings associated with each environment, along with the corresponding scores for each cultivar. It is worth noting that the structure of the FA model resembles that of an unstructured covariance matrix but distinguishes itself by its greater parsimony. A study conducted by Chaves et al. (2023) demonstrated the effectiveness and flexibility of FAST in selecting tropical maize genotypes, aiming for overall performance and stability across different locations and seasons. The authors suggested incorporating pedigree or genomic data into the statistical model, applying optimization methods, and using environmental features as strategies to enhance selection estimates.

The evaluation of genotypes with high overall performance and stability can be done through latent regression graphs. Although these graphs provide valuable information, selecting the best cultivars using this methodology can be labor-intensive, as it requires evaluating individualized regression for each genotype. In order to overcome these limitations, Smith and Cullis (2018) proposed FA selection tools, aiming to assess the overall performance and stability of each genotype across the entire dataset. Overall performance is achieved when the loadings of the first factor are positive and rotated, corresponding to the main effects of the genotypes. In this scenario, there is no complex G × E interaction, as the ranking of genotypes remains unchanged across different environments. The RMSD is used to estimate stability by measuring the deviation of each genotype from the line drawn by the latent regression. In this study, weights were assigned to both parameters since, for this specific dataset, productive performance was deemed more critical than stability. Consequently, some studies managed to achieve genetic gains using MET data, employing FAST for cultivar recommendation (Smith and Cullis, 2018; Tolhurst et al., 2019; Bakare et al., 2022).

The environmental and altitudinal characteristics of Malawi and Zambia significantly influence local climatic conditions, vegetation distribution, and land use (Supplementary Figure S4). Both countries are situated in high-altitude regions, with Malawi exhibiting altitudes ranging from 500 to 1,500 m, reaching 3,002 m in the Mulanje Mountains (Lancaster, 1980), while Zambia maintains an average altitude between 1,000 and 1,500 m, with Mount Mafinga as its highest peak (2,339 m). These altitudinal variations directly impact temperature regimes, precipitation patterns, and agricultural potential, aligning with previous studies on the influence of topography on African ecosystems. Higher elevations in Malawi are associated with milder temperatures and increased precipitation, which favor diverse vegetation and agricultural systems. In contrast, low-altitude areas, such as regions near Lake Malawi and the Shire Valley, experience warmer and more humid conditions, influencing local biodiversity and crop adaptability. Similarly, Zambia's elevated plateaus contribute to a moderate climate, reducing temperature extremes and promoting stable precipitation levels (Rawlins and Kalaba, 2020).

The analysis of mega-environments aims to identify target regions or environments with consistent patterns of G×E interaction over multiple years (Yan et al., 2023). When these patterns are stable and repeatable, the target region can be subdivided into sub-regions or mega-environments (Cooper and Hammer, 1996). However, when data are limited to a single year, the mega-environment concept may not be appropriate, as these environments should represent repeatable G×E interaction patterns (Basford and Cooper, 1998). In addition to yield data, incorporating environmental variables such as edaphoclimatic characteristics (elevation, temperature, precipitation, and soil type) can enhance the delineation of mega-environments. These variables provide a more comprehensive understanding of environmental influences on genotype performance, facilitating more precise recommendation strategies for different regions.

In this context, we observed that the variables growing degree days (gdd), mean temperature (Tmean), photosynthetically active radiation use efficiency (fue), seasonal precipitation variability (spv), and temperature range (Tmrange) were the most important factors influencing soybean grain yield in these environments. In tropical and subtropical regions such as Malawi and Zambia, adequate GDD accumulation is essential to ensure that soybean reaches maturity at the appropriate time. Mean temperature directly affects soybean metabolic rates, and excessively high temperatures can induce heat

stress, negatively impacting photosynthesis and grain formation. Factors such as light intensity, temperature, and water availability influence fue. In regions with high solar radiation, such as Malawi and Zambia, soybean has the potential for high fue, provided that other factors, such as water and nutrient availability, are not limiting. Irregular precipitation patterns, including severe droughts, can adversely affect soybean development from germination to grain filling. A moderate temperature range is beneficial for soybean, promoting improved carbon assimilation and balanced growth. Understanding the influence of environmental variables on soybean cultivation and modeling the G×E interaction enables the identification of specific adaptations, assisting breeders in decision-making regarding which varieties can have their genetic potential fully exploited (Araújo et al., 2024). Integrating robust statistical models, machine learning techniques (Crossa et al., 2024), and crop growth models (Bustos-Korts et al., 2022) can enhance the accuracy of these recommendations.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

MA: Investigation, Conceptualization, Supervision, Writing – review & editing, Data curation, Methodology, Software, Visualization, Resources, Funding acquisition, Validation, Writing – original draft, Project administration, Formal Analysis. BF: Writing – original draft, Methodology, Writing – review & editing, Formal Analysis. AS: Writing – review & editing, Writing – original draft, Methodology, Formal Analysis. JPP: Formal Analysis, Writing – review & editing, Methodology, Writing – original draft. NL: Methodology, Writing – original draft, Writing – review & editing, Formal Analysis. EL: Resources, Supervision, Conceptualization, Writing – review & editing, Writing – original draft, Data curation. MS: Project administration, Data curation, Conceptualization, Writing – review & editing, Supervision, Writing – original draft. PG: Validation, Writing – review & editing, Project administration, Supervision, Writing – original draft, Funding acquisition, Visualization. GC: Project administration, Supervision, Writing – review & editing, Writing – original draft, Funding acquisition, Resources. BD: Validation, Writing – original draft, Supervision, Writing – review & editing. JBP: Conceptualization, Visualization, Resources, Validation, Data curation, Project administration, Formal Analysis, Methodology, Investigation, Writing – review & editing, Writing – original draft, Software, Supervision, Funding acquisition.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Mauricio dos Santos Araújo was supported by FAPESP (São Paulo Research Foundation, Grant 2024/01868). We are grateful to São Paulo Research Foundation (FAPESP), similarly, we would like to acknowledge our sincere appreciation to the University of São Paulo and the University of Illinois, Urbana-Champaign for their support in this study.

Acknowledgments

We want to thank the coordinators and participants of the United States Agency for International Development (USAID) Feed the Future Soybean Innovation Lab Pan-African Soybean Variety Trials for their valuable contributions in providing the soybean data used in this study. We kindly thank Innocent Vulou Unzimai for the reviews.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2025.1594736/full#supplementary-material

References

Abebe, A. T., Adewumi, A. S., Adebayo, M. A., Shaahu, A., Mushoriwa, H., Alabi, T., et al. (2024). Genotype × environment interaction and yield stability of soybean (*Glycine max* L.) genotypes in multi-environment trials (METs) in Nigeria. *Heliyon* 10, e38097. doi: 10.1016/j.heliyon.2024.e38097

Agoyi, E. E., Ahomondji, S. E., Yemadje, P. L., Ayi, S., Ranaivoson, L., Torres, G. M., et al. (2024). Combining AMMI and BLUP analysis to select high-yielding soybean genotypes in Benin. *Agron. J.* 116, 2109–2128. doi: 10.1002/agj2.21615

Annicchiarico, P., Bellah, F., and Chiari, T. (2006). Repeatable genotype × location interaction and its exploitation by conventional and GIS-based cultivar recommendation for durum wheat in Algeria. *Eur. J. Agron.* 24, 70–81. doi: 10.1016/j.eja.2005.05.003

Araújo, M. S., Chaves, S. F., Dias, L. A., Ferreira, F. M., Pereira, G. R., Bezerra, A. R., et al. (2024). GISFA: an approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting. *Theor. Appl. Genet.* 137, 80. doi: 10.1007/s00122-024-04579-z

Bakare, M. A., Kayondo, S. I., Aghogho, C. I., Wolfe, M. D., Parkes, E. Y., Kulakow, P., et al. (2022). Parsimonious genotype by environment interaction covariance models for cassava *Manihot esculenta*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.978248

Basford, K., and Cooper, M. (1998). Genotype× environment interactions and some considerations of their implications for wheat breeding in Australia this review is one of a series commissioned by the advisory committee of the journal. *Aust. J. Agric. Res.* 49, 153–174. doi: 10.1071/A97035

Booth, T. H. (1990). Mapping regions climatically suitable for particular tree species at the global scale. *For. Ecol. Manage.* 36, 47–60. doi: 10.1016/0378-1127(90)90063-H

Bustos-Korts, D., Boar, M. P., Layton, J., Gehringer, A., Tang, T., Wehrens, R., et al. (2022). Identification of environment types and adaptation zones with self-organizing maps: applications to sunflower multienvironment data in Europe. *Theor. Appl. Genet.* 135, 2059–2082. doi: 10.1007/s00122-022-04098-9

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., and Thompson, R. (2018). *ASReml estimates variance components under a general linear* (VSN International Ltd).

Carciochi, W. D., Rosso, L. H. M., Secchi, M. A., Torres, A. R., Naeve, S., Casteel, S. N., et al. (2019). Soybean yield, biological N2 fixation and seed composition responses to additional inoculation in the United States. *Sci. Rep.* 9, 19908. doi: 10.1038/s41598-019-56465-0

Chaves, S. F., Damacena, M. B., Dias, K. O. G., de Almada Oliveira, C. V., and Bhering, L. L. (2024). Factor analytic selection tools and environmental feature-integration enable holistic decision-making in eucalyptus breeding. *Sci. Rep.* 14, 18429. doi: 10.1038/s41598-024-69299-2

Chaves, S. F. S., Evangelista, J. S. P. C., Trindade, R. S., Dias, L. A. S., Guimarães, P. E., Guimarães, L. J. M., et al. (2023). Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. *Crop Sci.* 63, 1114–1125. doi: 10.1002/csc2.20911

Chen, T., and Guestrin, C. (2016). "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. (New York, NY, USA: ACM) 785–794. doi: 10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., et al. (2022). *Xgboost: Extreme gradient boosting. R package version 1.6. 0.1*. doi: 10.32614/CRAN.package.xgboost

Cooper, M., and Delacy, I. H. (1994). Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. *Theor. Appl. Genet.* 88, 561–572. doi: 10.1007/BF01240919

Cooper, M., and Hammer, G. L. (1996). *Plant adaptation and crop improvement* (IRRI).

Costa-Neto, G., Crespo-Herrera, L., Fradgley, N., Gardner, K., Bentley, A. R., Dreisigacker, S., et al. (2023). Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. *G3 Genes|Genomes|Genetics* 13, jkac313. doi: 10.1093/g3journal/jkac313

Costa-Neto, G., and Fritsche-Neto, R. (2021). Enviromics: bridging different sources of data, building one framework. *Crop Breed. Appl. Biotechnol.* 21, e393521S12. doi: 10.1590/1984-70332021v21Sa25

Costa-Neto, G., Galli, G., Carvalho, H. F., Crossa, J., and Fritsche-Neto, R. (2021). EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3 Genes|Genomes|Genetics* 11, jkab040. doi: 10.1093/g3journal/jkab040

Cotes, J. M., Crossa, J., Sanches, A., and Cornelius, P. L. (2006). A Bayesian approach for assessing the stability of genotypes. *Crop Sci.* 46, 2654–2665. doi: 10.2135/cropsci2006.04.0227

Cowling, W. A., Castro-Urrea, F. A., Stefanova, K. T., Li, L., Banks, R. G., Saradadevi, R., et al. (2023). Optimal contribution selection improves the rate of genetic gain in grain yield and yield stability in spring canola in Australia and Canada. *Plants* 12, 383. doi: 10.3390/plants12020383

Crossa, J., Montesinos-Lopez, O. A., Costa-Neto, G., Vitale, P., Martini, J. W., Runcie, D., et al. (2024). Machine learning algorithms translate big data into predictive breeding accuracy. *Trends Plant Sci* 30, 167–184. doi: 10.1016/j.tplants.2024.09.011

Cullis, B. R., Jefferson, P., Thompson, R., and Smith, A. B. (2014). Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. *Theor. Appl. Genet.* 127, 2193–2210. doi: 10.1007/s00122-014-2373-0

Cullis, B. R., Smith, A. B., Beeck, C. P., and Cowling, W. A. (2010). Analysis of yield and oil from a series of canola breeding trials. part II. exploring variety by environment interaction using factor analysis. *Genome* 53, 1002–1016.

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *J. Agricultural Biological Environ. Stat* 11, 381–393. doi: 10.1198/108571106X154443

Eberhart, S. A., and Russell, W. A. (1966). Stability parameters for comparing varieties. *Crop Sci.* 6, 36–40. doi: 10.2135/cropsci1966.0011183X000600010011x

Favoretto, V. R., Murithi, H. M., Leles, E. P., da Santos, F. M., Chigeza, G., Goldsmith, P., et al. (2025). Soybean rust-resistant and tolerant varieties identified through the pan-african trial network. *Pest Manage. Sci.* 1, 1–7. doi: 10.1002/ps.8639

W. R. Fehr (Ed.) (1987). "Principles of Cultivars Development," in *Contents v. I. Theory and technique; v.2. Crop species (Macmillan). Includes bibliographies and indexes*.

Finlay, K., and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Aust. J. Agric. Res.* 14, 742.

Food and Agriculture Organization of the United Nations (2025). *Global Agro-Ecological Zones (GAEZ) Data Portal*.

Galili, T. (2015). Dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31, 3718–3720. doi: 10.32614/CRAN.package.dendextend

Gauch, H. G.Jr (2008). Statistical analysis of yield trials by AMMI and GGE. *Crop Sci.* 46, 1488–1500. doi: 10.2135/cropsci2007.09.0513

Gauch, H. G. Jr, and Zobel, R. (1997). Identifying mega-environments and targeting genotypes. *Crop Sci.* 37, 311–326. doi: 10.2135/cropsci1997.0011183X003700020002x

Gogel, B., Smith, A., and Cullis, B. (2018). Comparison of a one- and two-stage mixed model analysis of Australia's national variety trial southern region wheat data. *Euphytica* 214, 44. doi: 10.1007/s10681-018-2116-4

Guarino, L., Jarvis, A., Hijmans, R. J., and Maxted, N. (2002). "Geographic information systems (GIS) and the conservation and use of plant genetic resources," in *Managing plant genetic diversity. Proceedings of an international conference, Kuala Lumpur, Malaysia, 12–16 June 2000* (CABI publishing Wallingford UK), 387–404.

Heinemann, A. B., Costa-Neto, G., Fritsche-Neto, R., Matta, D. H., and Fernandes, I. K. (2022). Enviromic prediction is useful to define the limits of climate adaptation: a case study of common bean in Brazil. *Field Crops Res.* 286, 108628. doi: 10.1016/j.fcr.2022.108628

Henderson, C. R. (1949). Estimates of changes in herd environment. *J. Dairy Sci.* 61, 294–300.

Henderson, C. R. (1950). Estimation of genetic parameters. *Ann. Math. Stat* 21, 309–310.

Kang, M. S., Harville, B. G., and Gorman, D. P. (1989). Contribution of weather variables to genotype × environment interaction in soybean. *Field Crops Res.* 21, 297–300. doi: 10.1016/0378-4290(89)90011-7

Kassambara, A., and Mundt, F. (2016). *Factoextra: extract and visualize the results of multivariate data analyses* (CRAN: Contributed Packages).

Kelly, A. M., Smith, A. B., Eccleston, J. A., and Cullis, B. R. (2007). The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. *Crop Sci.* 47, 1063–1070. doi: 10.2135/cropsci2006.08.0540

Kozak, M., and Piepho, H. P. (2018). What's normal anyway? residual plots are more telling than significance tests when checking ANOVA assumptions. *J. Agron. Crop Sci.* 204, 86–98. doi: 10.1111/jac.12220

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., et al. (2020). *Caret: Classification and regression training. R package version 6.0-86*. doi: 10.32614/CRAN.package.caret

Lancaster, I. N. (1980). Relationships between altitude and temperature in Malawi. *South Afr. Geographical J.* 62, 89–97. doi: 10.1080/03736245.1980.10559624

Lin, C. S., and Binns, M. R. (1998). A superiority measure of cultivar performance for cultivar × location data. *Can. J. Plant Sci.* 68, 193–198. doi: 10.4141/cjps88-018

Mächler, M., Rousseeuw, P. J., Struyf, A., Hubert, M., Hornik, K., Studer, M., et al. (2019). Finding groups in data: Cluster analysis extended rousseeuw. R Package version 2.0.6. doi: 10.32614/CRAN.package.cluster

Malosetti, M., Bustos-Korts, D., Boer, M. P., and Eeuwijk, F. A. V. (2016). Predicting responses in multiple environments: issues in relation to genotype × environment interactions. *Crop Science.* 56, 2210–2222. doi: 10.2135/cropsci2015.05.0311

Messina, M. (2022). Perspective: Soybeans can help address the caloric and protein needs of a growing global population. *Front. Nutr.* 9. doi: 10.3389/fnut.2022.909464

Meyer, E., Prenger, E., Mahmood, A., da Fonseca Santos, M., Chigeza, G., Song, Q., et al. (2024). Evaluating genetic diversity and seed composition stability within Pan-African soybean variety trials. *Crop Sci.* 64, 3272–3292. doi: 10.1002/csc2.21356

Mishra, R., Tripathi, M., Sikarwar, R., Singh, Y., and Tripathi, N. (2024). Soybean (*Glycine max* L. Merrill): A multipurpose legume shaping our world. *Plant Cell Biotechnol. Mol. Biol.* 25, 17–37. doi: 10.56557/pcbmb/2024/v25i3-48643

NasaPower (2022). *Prediction of worldwide energy resource*.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and R objects. *arXiv* 1. doi: 10.48550/arXiv.1403.2805

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545–554. doi: 10.1093/biomet/58.3.545

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Magazine J. Sci.* 2, 559–572.

Piepho, H. P. (1997a). Analysis of a randomized block design with unequal subclass numbers. *Agron. J.* 89, 718–723. doi: 10.2134/agronj1997.00021962008900050002x

Piepho, H. P. (1997b). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 53, 761–766. doi: 10.2307/2533976

Piepho, H. P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. *Theor. Appl. Genet.* 97, 195–201. doi: 10.1007/s001220050885

Piepho, H. P. (2019). A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical J.* 61, 860–872. doi: 10.1002/bimj.201800270

Piepho, H. P., and Williams, E. (2024). Factor-analytic variance–covariance structures for prediction into a target population of environments. *Biometrical J.* 66, e202400008. doi: 10.1002/bimj.202400008

Plaisted, R. L., and Peterson, L. C. (1959). A technique for evaluating the ability of selections to yield consistently in different locations or seasons. *Am. Potato J.* 36, 381–385.

Poupon, V., Gezan, S. A., Schueler, S., and Lstibůrek, M. (2023). Genotype × environment interaction and climate sensitivity in growth and wood density of European larch. *For. Ecol. Manage.* 545, 121259. doi: 10.1016/j.foreco.2023.121259

Rawlins, J., and Kalaba, F. K. (2020). "Adaptation to climate change: Opportunities and challenges from Zambia." in *African Handbook of Climate Change Adaptation*, eds. W. Leal Filho (Editor-in-Chief), N. Oguge, D. Ayal, L. Adeleke, and I. da Silva (Cham: Springer), 1–20. doi: 10.1007/978-3-030-45106-6_30

R Core Team (2022). *R: A Language and environment for statistical computing* (R Foundation for Statistical Computing).

Resende, R. T., Hickey, L., Amaral, C. H., Peixoto, L. L., Marcatti, G. E., and Xu, Y. (2024). Satelliteenabled enviromics to enhance crop improvement. *Mol. Plant* 17, 848–866.

Resende, R. T., Xavier, A., Silva, P. I. T., Resende, M. P., Jarquin, D., and Marcatti, G. E. (2025). GISbased G×E modeling of maize hybrids through enviromic markers engineering. *New Phytol.* 245, 102–116. doi: 10.1111/nph.19951

Santos, M. (2019). Soybean varieties in Sub-Saharan africa. *Afr. J. Food Agriculture Nutr. Dev.* 19, 15136–15139. doi: 10.18697/ajfand.88.SILFarmDoc06

Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity* 29, 237–245. doi: 10.1038/hdy.1972.87

Smith, A. B., and Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214, 1–19. doi: 10.1007/s10681-018-2220-5

Smith, A., Cullis, B., and Thompson, R. (2001a). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006-341X.2001.01138.x

Smith, A. B., Cullis, B., and Thompson, R. (2001b). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57, 1138–1147. doi: 10.1111/j.0006-341X.2001.01138.x

Smith, A. B., Cullis, B. R., and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *J. Agric. Sci.* 143, 449–462. doi: 10.1017/S0021859605005587

Smith, A. B., Ganesalingam, A., Kuchel, H., and Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theor. Appl. Genet.* 128, 55–72. doi: 10.1007/s00122-014-2412-x

Smith, A., Norman, A., Kuchel, H., and Cullis, B. (2021). Plant variety selection using interaction classes derived from factor analytic linear mixed models: Models with independent variety effects. *Front. Plant Sci.* 12, 1–17.

Sokal, R. R., and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* 38, 1409–1438.

Sousa, T., Rocha, M., Damasceno-Silva, K. J., Bertini, C., M., Silveira, L., et al. (2019). Simultaneous selection for yield, adaptability, and genotypic stability in immature cowpea using REML/BLUP. *Pesquisa Agropecuária Bras.* 54, e01234. doi: 10.1590/S1678-3921.pab2019.v54.01234

Sparks, A. H. (2018). Nasapower: a nasa power global meteorology, surface solar energy and climatology data client for R. *J. Open Source Software* 3, 1035. doi: 10.21105/joss.01035

Stefanova, K. T., Smith, A. B., and Cullis, B. R. (2009). Enhanced diagnostics for the spatial analysis of field trials. *J. Agricultural Biological Environ. Stat* 14, 392–410. doi: 10.1198/jabes.2009.07098

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

Tolhurst, D. J., Mathews, K. Y. L., Smith, A. B., and Cullis, B. R. (2019). Genomic selection in multienvironment plant breeding trials using a factor analytic linear mixed model. *J. Anim. Breed. Genet.* 136, 279–300. doi: 10.1111/jbg.12404

van Eeuwijk, F. A., Bustos-Korts, D. V., and Malosetti, M. (2016). What should students in plant breeding know about the statistical aspects of genotype × environment interactions? *Crop Sci.* 56, 2119–2140. doi: 10.2135/cropsci2015.06.0375

Van Eeuwijk, F. A., and Elgersma, A. (1993). Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. *Heredity* 70, 447–457. doi: 10.1038/hdy.1993.66

Vitale, P., Crossa, J., Vaccino, P., and De Vita, P. (2024). Defining the target population of environments for wheat (*Triticum aestivum* L.) breeding in Italy based on historical data. *Plant Breed.* 143, 518–533. doi: 10.1111/pbr.13192

Wickham, H. (2023). *Package 'httr': Tools for Working with URLs and HTTP. Version 1.4.7*.

Wood, J. (1976). The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity* 37, 1–7.

Xu, Y. (2016). Envirotyping for deciphering environmental impacts on crop plants. *Theor. Appl. Genet.* 129, 653–673. doi: 10.1007/s00122-016-2691-5

Yan, W., Hunt, L. A., Sheng, Q., and Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.* 40, 597–605. doi: 10.2135/cropsci2000.403597x

Yan, W., Nilsen, K. T., and Beattie, A. (2023). Mega-environment analysis and breeding for specific adaptation. *Crop Sci.* 63, 480–494. doi: 10.1002/csc2.20895

Zhi, Y., Sun, T., Zhou, Q., and Leng, X. (2020). Screening of safe soybean cultivars for cadmium contaminated fields. *Sci. Rep.* 10, 12965. doi: 10.1038/s41598-020-69803-4