Theoretical and Applied Genetics (2024) 137:80
https://doi.org/10.1007/s00122-024-04579-z

ORIGINAL ARTICLE

# GIS-FA: an approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting

Maurício S. Araújo[1] · Saulo F. S. Chaves[1] · Luiz A. S. Dias[1] · Filipe M. Ferreira[2] · Guilherme R. Pereira[1] · André R. G. Bezerra[3] · Rodrigo S. Alves[4] · Alexandre B. Heinemann[6] · Flávio Breseghello[6] · Pedro C. S. Carneiro[4] · Matheus D. Krause[7] · Germano Costa-Neto[5] · Kaio O. G. Dias[4]

Received: 15 July 2023 / Accepted: 6 February 2024 / Published online: 12 March 2024

## Abstract

***Key message*** **We propose an "enviromics" prediction model for recommending cultivars based on thematic maps aimed at decision-makers.**

**Abstract** Parsimonious methods that capture genotype-by-environment interaction (GEI) in multi-environment trials (MET) are important in breeding programs. Understanding the causes and factors of GEI allows the utilization of genotype adaptations in the target population of environments through environmental features and factor-analytic (FA) models. Here, we present a novel predictive breeding approach called GIS-FA, which integrates geographic information systems (GIS) techniques, FA models, partial least squares (PLS) regression, and enviromics to predict phenotypic performance in untested environments. The GIS-FA approach enables: (i) the prediction of the phenotypic performance of tested genotypes in untested environments, (ii) the selection of the best-ranking genotypes based on their overall performance and stability using the FA selection tools, and (iii) the creation of thematic maps showing overall or pairwise performance and stability for decision-making. We exemplify the usage of the GIS-FA approach using two datasets of rice [*Oryza sativa* (L.)] and soybean [*Glycine max* (L.) Merr.] in MET spread over tropical areas. In summary, our novel predictive method allows the identification of new breeding scenarios by pinpointing groups of environments where genotypes demonstrate superior predicted performance. It also facilitates and optimizes cultivar recommendations by utilizing thematic maps.

## Introduction

Crossover interaction refers to changes in the ranking of genotypes caused by the lack of genotypic correlation and negative correlations between environments, which is the most critical source of genotype-by-environment interaction (GEI) for plant breeders (Cooper and Delacy 1994; Crossa et al. 2004). Cultivar development programs for crops evaluate experimental genotypes (i.e., prior to release) in multi-environmental trials (MET) to (i) depict GEI patterns for future cultivar placement and (ii) increase the accuracy of selection. Therefore, analytical methods that fully explore the GEI patterns from MET are needed for decision-making (Malosetti et al. 2013; van Eeuwijk et al. 2016; Dias et al. 2022; Tolhurst et al. 2022).

The first attempt to consider the GEI in plant breeding was proposed by Yates and Cochran (1938), who decomposed the part due to the interaction from the total phenotypic variation. Later, Finlay and Wilkinson (1963) used marginal environmental means as independent variables in

Communicated by Hiroyoshi Iwata.

Maurício S. Araújo and Saulo F. S. Chaves contributed equally to this work.

✉ Kaio O. G. Dias
kaio.o.dias@ufv.br

[1] Department of Agronomy, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[2] Department of Crop Science - College of Agricultural Sciences, São Paulo State University, Botucatu, São Paulo, Brazil

[3] Limagrain Brazil S.A., Jataí, Goiás, Brazil

[4] Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[5] Institute for Genomics Diversity, Cornell University, Ithaca, NY, USA

[6] Brazilian Agricultural Research Corporation (Embrapa Rice and Beans), Santo Antônio de Goiás, Goiás, Brazil

[7] Department of Agronomy, Iowa State University, Ames, IA, USA

the regression analysis to depict GEI, and several approaches were developed within that framework (Eberhart and Russell 1966; Li et al. 2018). Multivariate techniques such as additive main effects and multiplicative interaction (AMMI) (Gauch Jr and Zobel 1997) and the genotype plus GEI (GGE) biplot (Yan et al. 2000) have also been extensively used (Yan et al. 2007; Balestre et al. 2009; Silva et al. 2021). Further model expansions were made possible by the development of the linear mixed model equations (Henderson 1949, 1950), which allowed for the incorporation of covariance between relatives and environments and the relaxation of assumptions such as homogeneous residual variances (Piepho et al. 2008). Factor-analytic (FA) mixed models (Piepho 1997; Smith et al. 2001) can be employed to explore the covariance between environments. These models offer the flexibility to account for heterogeneous genotypic (or genetic) covariances between environments using a few latent variables known as factors (*K*). In addition to the overall (i.e., across environments) and conditional (i.e., within environments) performance, metrics such as stability and sensitivity can also be computed from FA models to facilitate the decision-making process (Stefanova and Buirchell 2010; Cullis et al. 2014; Dias et al. 2018; Smith and Cullis 2018; Smith et al. 2021).

An extension to statistical models that address GEI involves incorporating environmental information, such as physical and chemical soil properties, as well as environmental features like temperature and rainfall precipitation (Tolhurst et al. 2022). The advantages of integrating environmental features into a prediction model include (i) the capability to untangle environmental determinants and the crossover GEI main drivers and (ii) the ability to predict phenotypic performance in yet-to-be-seen environments (Sae-Lim et al. 2014; Oliveira et al. 2020; Tolhurst et al. 2022). Furthermore, categorizing similar environments into homogeneous groups facilitates resource optimization and the identification of mega-environments (Wood 1976; Denis 1988; Van Eeuwijk and Elgersma 1993; Millet et al. 2019; Costa-Neto et al. 2021c; Krause et al. 2022). Therefore, advances in computational resources, along with the development of geographic information systems (GIS) techniques, are essential for designing novel prediction strategies in MET (Cooper and Messina 2021; Rogers et al. 2021; Cooper et al. 2022; Diepenbrock et al. 2022).

GIS techniques have been defined as computer-based systems used for analyzing and interpreting spatially referenced information and are powerful tools in the integration of genetics and environmental information (Beebe et al. 1997; Guarino et al. 2002; Jarquín et al. 2014; Hernández et al. 2019; Costa-Neto and Fritsche-Neto 2021). For example, Annicchiarico et al. (2006) identified consistent genotype-by-location interactions using GIS-based models to recommend cultivars for durum wheat in Algeria. Costa-Neto et al. (2020) applied a GIS-based tool with factorial regression to analyze spatial trends and create thematic maps of yield performance for upland rice in Brazil. In addition, Costa-Neto et al. (2021b) integrated GIS techniques with nonlinear kernels to model additive, dominance, and GEI effects. All the mentioned techniques fall under the umbrella of "envirotypic-assisted selection," which integrates genomic and environmental data to improve the accuracy of selection in plant breeding programs (Resende et al. 2021).

The combination of statistics, quantitative genetics, and GIS techniques enabled the introduction of the field of enviromics in the plant breeding community (Cooper et al. 2014; Xu 2016; Costa-Neto and Fritsche-Neto 2021). Coupled with knowledge from plant ecophysiology, this field aims to describe how the environment impacts plant development and the phenotypic plasticity of important agronomic traits (Costa-Neto and Fritsche-Neto 2021). Accordingly, envirotypes are all sources of environmental variations related to plant development that can act as environmental markers in statistical genetics models to predict genotypic effects in non-evaluated environments (Xu 2016; Resende et al. 2021). However, integrating phenotypic and genomic data with environmental features can generate two statistical problems: high correlation among predictors resulting in multicollinearity and the curse of dimensionality when the number of observations is smaller than the predictors. In these situations, methods such as partial least squares (PLS), which combine features from principal components analysis and multiple regression (Wold et al. 2001), and Bayesian factor analytic models (Nuvunga et al. 2019), can be applied to identify linear combinations of predictors that capture the underlying structure of the data (Montesinos-López et al. 2022a,b).

Here, we present a novel predictive breeding approach called GIS-FA that combines FA, PLS, and enviromics to predict the phenotypic performance of experimental genotypes in untested environments. The GIS-FA uses environmental information collected from GIS tools to predict the factor loadings of untested environments via PLS, where the estimated factor loadings from the observed environments are used as the training set. The empirical best linear unbiased predicted values (eBLUPs) of genotypic means in untested environments are then calculated as the linear combination of the predicted loadings via PLS and genotypic scores from the FA models. We hypothesize that the GIS-FA model has higher prediction accuracy compared to a PLS model trained with eBLUPs within observed environments (henceforth called GIS-GGE). We tested this hypothesis using two MET datasets from Brazil: rice trials located in the Brazilian Savanna (Cerrado) and the Amazon rainforest, as well as soybean trials located in the state of Mato Grosso do Sul. Thus, this study aims to: (i) propose the GIS-FA methodology for predicting genotypes' performance in untested environments and compare its predictive ability with the GIS-GGE methodology; (ii) apply GIS-FA to select the best-ranking genotypes based on their overall performance (OP) and stability using the FA selection tools; and (iii) create thematic maps that illustrate the genotypes' performance across environments in the breeding zone.

## Material and methods

### Phenotypic data

We exemplify the GIS-FA model using two datasets from MET covering tropical areas in Brazil. These trials have been used to make decisions regarding the release of cultivars by both public and proprietary breeding organizations. The soybean dataset contains three years of field trials conducted in the state of Mato Grosso do Sul (represented by triangles in Fig. 1), whereas the rice dataset includes two years of field trials conducted across eight states (represented by circles in Fig. 1). It is important to note that the variation in elevation varies across the studied area (Fig. 1b). This factor, along with latitude and longitude, influences changes in both weather and soil conditions, as indicated by the Köppen–Geiger classification (Alvares et al. 2013) in Fig. 1c and the Brazilian Soil Classification System (Santos 2018) in Fig. 1d. Both datasets include field trials planted in the same location and year but during different planting seasons. Thus, henceforth, the term "environment" refers to the combination of location, year, and planting season. Another common characteristic shared by both datasets is that not all genotypes were evaluated in all environments (Supplementary Figure 1). This has three main reasons: (i) seed availability, (ii) discarding low-performing lines at the end of each agricultural year, and (iii) including cultivars/genotypes from partner breeding programs for evaluation in the target population of environments (TPE). It is expected that the inclusion/

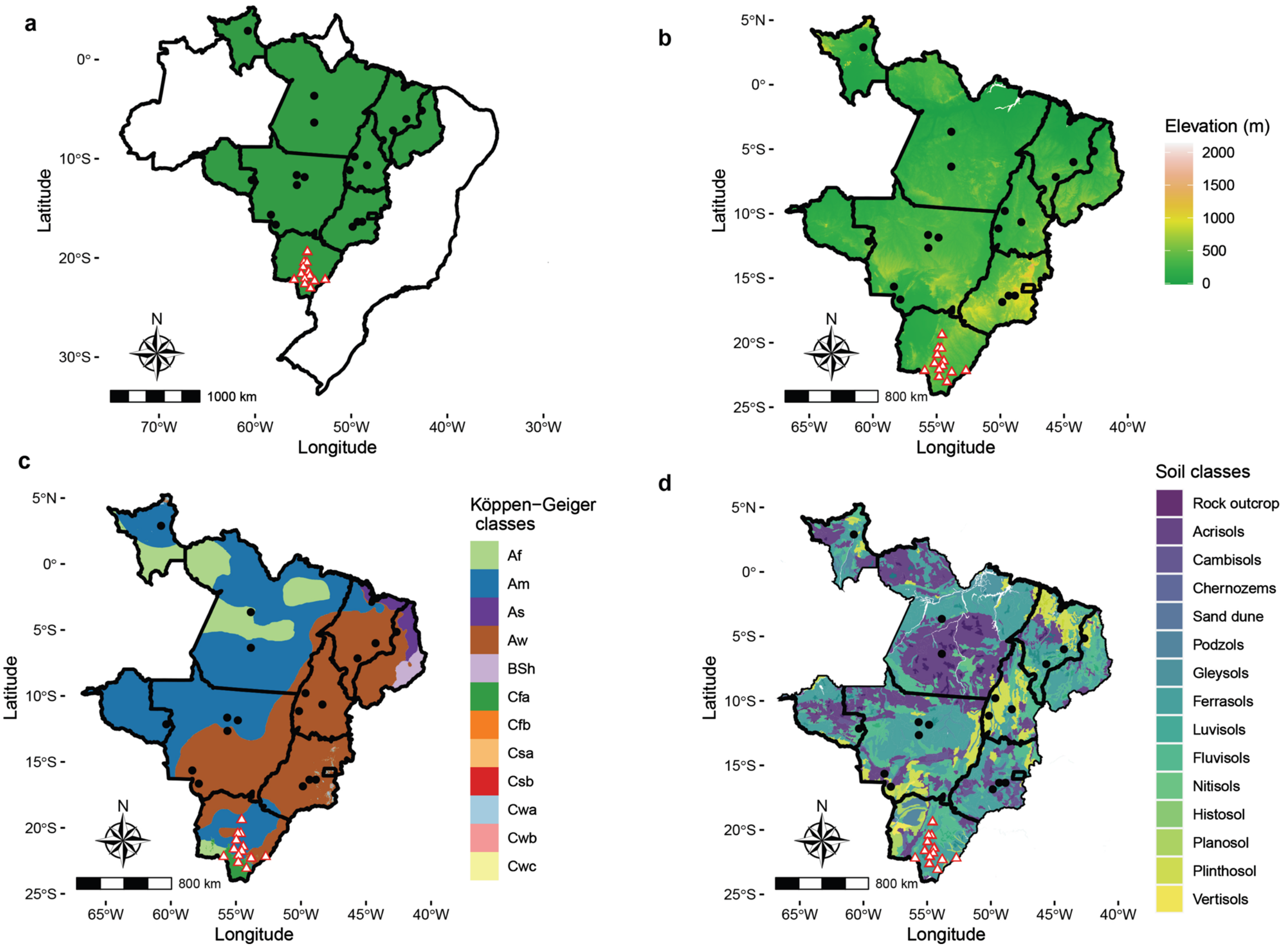

**Fig. 1** Maps of the studied area. **a** Shows the map of Brazil, highlighting the states where the rice (circles) and soybean (triangles) trials were conducted. We subset these states in **b–d**. **b** Depicts the elevation in meters, **c** displays the Köppen–Geiger classification (Alvares et al. 2013), and **d** highlights the Brazilian soil classification adapted to FAO classification (Santos 2018; FAO 2014)

exclusion of selected candidates in the MET does not yield relevant bias in the variance component estimates (Piepho and Möhring 2006; Hartung and Piepho 2021).

### Rice dataset

The rice dataset is composed of 80 pure lines developed by the Brazilian Agricultural Research Corporation (Embrapa Rice and Beans). These pure lines plus three commercial cultivars were evaluated for their value of cultivation and use (VCU) in 21 environments during the cropping seasons of 2009/2010 and 2010/2011. Candidate cultivars that demonstrate high yield and agronomic stability in the TPE will be registered for commercial use. The TPE of the Upland Rice Breeding Program is located within the geographical coordinates of 1 ° North to 17° South and 42° West to 70° West. It includes eight states from the Mid-West (Mato Grosso and Goiás), the Northeast (Maranhão and Piauí), and the North (Pará, Rondônia, Roraima, and Tocantins). Further details are presented in Supplementary Table 1. Eighteen locations were sampled in the TPE (Fig. 1), where trials were arranged in randomized complete blocks with four replications. Experimental plots consisted of four 5 m rows spaced 0.3 m apart, totaling an area of 6 m$^2$, with 60 seeds sown per meter. Seed yield (kg ha$^{-1}$) was measured in the two central rows. Management practices in these regions followed the technical recommendations adopted for upland rice.

### Soybean dataset

The soybean dataset comprises 195 pure lines that were evaluated over three cropping seasons (2019/2020, 2020/2021, and 2021/2022) at 13 locations in the state of Mato Grosso do Sul and the Central-West region of Brazil (Fig. 1). Trials were conducted under rainfed conditions and overseen by the Mato Grosso do Sul Foundation (Fundação MS) in 49 different environments. The experimental design involved randomized, complete blocks with three replications. The plots consisted of five 12 m-long rows spaced 0.5 m apart, with a total area of 30 m$^2$. Seed yield (kg ha$^{-1}$) was measured in the three central rows and corrected for 13% moisture. Weed and pest control were carried out following the recommendations for the region.

## GIS-FA workflow

Here, we will summarize the procedures for applying the GIS-FA methodology. The method was created to evaluate the OP and stability of genotypes in untested environments and to plot the spatial prediction on thematic maps. This enables breeders to define strategies for recommending adaptable cultivars, prospect new target environments that maximize genetic gain through selection, and define breeding zones based on the pattern of environmental features. The procedures to apply the GIS-FA are:

- *Step 1—Geographic data collection from tested and untested environments*: To implement the GIS-FA method, it is imperative to acquire geographic information. This includes, but is not limited to, latitude and longitude. For the tested environments, such data can be obtained in situ in the experimental area or via GIS tools. For the untested environments, one can sample pixels (coordinates) of the breeding region (or the area under consideration for prediction). These pixels must be representative of the different environmental conditions found in the breeding region. We detail the sampling process adopted in this study in section "Environmental information."
- *Step 2—Environmental data collection*: This step requires information on the sowing and harvest times for each trial. More detailed results can be achieved by using genotype-specific harvest dates. The process of envirotyping (data collection and processing) is crucial for understanding the environmental factors that drive the G × E interaction and shape the development of the plant (Cooper et al. 2014; Xu 2016; Costa-Neto et al. 2021a). Environmental features can be obtained in the form of in situ data (e.g., from sensors attached to drones or high-throughput phenotyping stations) or in raster format (e.g., historical series for a given geographic point stored on online platforms as rasters). Other methods of obtaining these data include meteorological stations, the National Centers for Environmental Information (NCEI) (NOAA 2023), the Climate Forecast System Reanalysis (CFSR) (CFSR 2018), the European Centre for Medium-Range Weather Forecasts (ECMWF) (ECMWF 2023), the Global Historical Climatology Network (GHCN) (GHCNd 2023), the NASA Earth Observing System Data and Information System (EOSDIS) (EOSDIS 2023), WorldClim (Fick and Hijmans 2017), Climatologies at High Resolution for the Earth's Land Surface Areas (CHELSA) (CHELSA 2023), and the Climate Research Unit Time-Series (CRU TS). Soil data can be collected through analysis conducted in the experiment itself or obtained from databases such as SoilGrids (SoilGrids 2022). We detail the collection of environmental features in both datasets analyzed in section Environmental information. The incorporation of environmental features in statistical-genetic models is based on Shelford's Law (Shelford 1911), which states that the growth of a species is regulated by environmental factors (within a range of maximum and minimum values). The environmental features can serve as environmental markers, enabling a deeper understanding of phenotypic expression. This concept was introduced in

the context of G × E analysis for plant breeding by Costa-Neto et al. (2021a), in which more details of its theoretical application are provided in the text. In this case, there is an association between the environmental marker and the evaluated genotype. Environmental features can also be used to characterize both tested and untested environments, allowing for the determination of the similarity of the sampled points to the TPE (see section Environmental similarity and interpolation grid for details).

- *Step 3—Phenotypic data analysis*: In this step, we fit FA models with different numbers of factors and choose one based on parsimony and/or explanatory ability (as detailed in section FA model selection). After choosing the model, we use the FA selection tools (Stefanova and Buirchell 2010; Smith and Cullis 2018) to build a selection index and select the best-ranking genotypes across different environments (further details in section Selection tools for overall performance and stability).
- *Step 4—Prediction for the untested environments*: The matrix of rotated loadings of the chosen FA model is used to train a PLS regression model with the gathered environmental features. The goal is to predict the factor loading of untested environments only by providing the model with environmental information about these locations. Once the loadings are predicted, they are used in linear combinations with the experimental genotypes' factor scores to predict the eBLUPs in untested environments. This process is thoroughly detailed in section Spatial predictions in the breeding zone.
- *Step 5—Map-based recommendation*: The prediction phase provides the performance of each genotype in the new locations that were sampled in the first step. To extrapolate to the whole breeding region, an interpolation process is required (detailed in section Environmental similarity and interpolation grid). We proposed three types of thematic maps, considering interpolation: (i) adaptation zones, which allow for the identification of adaptation areas for each genotype, i.e., areas where genotypes are expected to have better responses to the local environmental effects; (ii) pairwise comparisons, which compare the performance of two genotypes (or a genotype and a commercial check) in untested environments; and (iii) which-won-where, used to identify the most promising experimental genotypes in the breeding region. At (i) and (ii), one can make a pre-selection of which genotypes to evaluate using the FA selection tools and perform a detailed study about these selection candidates' adaptation throughout the breeding region.

## Environmental information

We used 32 environmental features in this study, including three geographical coordinates (altitude, latitude, and longitude), 16 related to weather conditions, and 13 soil traits (Table 1). The weather variables for each environment were obtained as daily averages for the growing season (i.e., between sowing and harvest dates) and processed using the R (version 4.2.3, R Core Team 2023) package EnvRtype (Costa-Neto et al. 2021c), which retrieves raw data from the NASA database (Sparks 2018; NasaPower 2022). Most of the soil variables for each location (i.e., latitude/longitude combination) were acquired using the geodata package (Hijmans et al. 2023), which downloads rasters from the SoilGrids platform (SoilGrids 2022). Only the raster data for soil temperature, isothermality, temperature seasonality, and mean diurnal range were manually downloaded from the platform of Lembrechts et al. (2022). Soil rasters were downloaded for a depth interval of 5–15 cm with a resolution of 30 arcseconds. Each pixel represents an area of approximately 1 km$^2$ and was processed using the raster package (Hijmans 2020).

In this study, we aimed to perform spatial predictions using environmental information in a three-step procedure as follows: (i) defining the scope of the prediction area based on the political borders of the Brazilian states where trials were conducted; (ii) implementing a sampling approach to generate a cloud of geographical points (latitude/longitude) for collecting environmental data. Fifty points were sampled from each municipality within states, ensuring an unbiased sampling of possible environmental conditions in the states; and (iii) using the data from (ii), performed a spatial interpolation to cover the entire area of the state(s) and computed the spatial predictions. In (ii), the soil-related environmental features were obtained as previously described for the tested environments. Monthly averages for the weather-related environmental features were obtained from 2000 to 2021. Further details will be provided in the following sections.

## Environmental similarity and interpolation grid

The package pdist (Wong 2022) was used to quantify the environmental similarity by calculating the Euclidean distances between the observed and unobserved (i.e., sampled points) environments. Let $\mathbf{W}$ be a $J \times P$ matrix of scaled values representing $P$ environmental features in $J$ observed environments, and let $\boldsymbol{\Omega}$ be a matrix containing the same information but for $U$ unobserved environments. The environmental features were scaled to variance 1. Then, the Euclidean distance between an observed environment $j$ and an unobserved environment $u$ ($D_{ju}$) is given by the distances between the rows of $\mathbf{W}$ and $\boldsymbol{\Omega}$ that correspond to $j$ and $u$, respectively:

**Table 1** Summary statistics of the 32 environmental features classified into three groups: geographical, climatic, and soil-related

| Group | Environmental information | ID | Unit | Rice data | | | Soybean data | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Min | Mean | Max | Min | Mean | Max |
| Geographical | Altitude | alt | meters (m) | 70.00 | 410.19 | 1033.00 | 234.00 | 410.35 | 661.00 |
| | Latitude | lat | degrees (°) | – 16.85 | – 11.68 | 2.90 | – 23.10 | – 21.66 | – 19.38 |
| | Longitude | lon | degrees (°) | – 60.75 | – 52.05 | – 42.65 | – 56.55 | – 54.65 | – 52.72 |
| Climatic | All sky insolation incidents on a horizontal surface | sw | $\mathrm{MJ/m^2/day}$ | 16.43 | 18.90 | 20.83 | 21.38 | 22.65 | 24.69 |
| | Clear sky insolation incident on a horizontal surface | lw | $\mathrm{MJ/m^2/day}$ | 380.72 | 405.20 | 424.89 | 401.17 | 408.72 | 417.74 |
| | Total precipitation | prec | mm/day | 4.58 | 7.23 | 9.77 | 1.87 | 3.65 | 6.32 |
| | Relative humidity | rh | % | 79.15 | 85.47 | 91.92 | 57.30 | 66.27 | 77.09 |
| | Slope of saturation vapor pressure curve | spv | kPa°C | 0.16 | 0.19 | 0.22 | 0.19 | 0.22 | 0.24 |
| | Potential evapotranspiration | etp | mm.day | 7.53 | 8.64 | 9.36 | 9.88 | 10.51 | 11.47 |
| | Deficit by precipitation | pept | mm.day | – 4.49 | – 1.40 | 2.21 | – 9.60 | – 6.86 | – 4.01 |
| | Vapor pressure deficit | vpd | kPa | 0.36 | 0.60 | 0.96 | 0.94 | 1.70 | 2.22 |
| | Wind speed at 2 m above ground | ws | m/s | 0.06 | 1.04 | 1.82 | 0.62 | 1.69 | 2.14 |
| | Dew/Frost Point Temperature | tdew | °C | 18.50 | 22.03 | 24.07 | 17.90 | 19.50 | 21.02 |
| | Daily temperature range | trange | °C day | 5.65 | 7.22 | 8.87 | 9.36 | 12.16 | 14.00 |
| | Temperature at 2 m above ground | tmean | °C | 21.99 | 24.84 | 27.40 | 25.06 | 27.48 | 28.96 |
| | Maximum temperature at 2 m above ground | tmax | °C | 26.69 | 28.69 | 31.71 | 29.94 | 33.91 | 36.08 |
| | Minimum temperature at 2 m above ground | tmin | °C | 17.82 | 21.47 | 23.58 | 20.53 | 21.75 | 22.80 |
| | Growing degree days | gdd | $^\circ\mathrm{C\ d^{-1}}$ | 14.26 | 17.08 | 19.65 | 17.26 | 19.83 | 21.19 |
| | Effect of temperature on radiation use efficiency | frue | – | 0.65 | 0.78 | 0.89 | 0.78 | 0.89 | 0.94 |
| Soil | Bulk density of the fine earth fraction | bdod | $\mathrm{kg\ dm^{-3}}$ | 1.10 | 1.27 | 1.40 | 1.20 | 1.28 | 1.40 |
| | Clay (< 0.002 mm) in fine earth | clay | % | 18.00 | 28.58 | 42.00 | 16.00 | 36.22 | 52.00 |
| | Silt (0.002– 0.05 mm) in fine earth | silt | % | 11.00 | 19.88 | 32.00 | 10.00 | 17.76 | 23.00 |
| | Sand (> 0.05 mm) in fine earth | sand | % | 39.00 | 51.65 | 69.00 | 25.00 | 45.94 | 66.00 |
| | Volume fraction of coarse fragments (> 2 mm) | cfvo | % | 1.00 | 4.38 | 11.00 | 2.00 | 3.59 | 5.00 |
| | Nitrogen content | nit | $\mathrm{g\ kg^{-1}}$ | 0.80 | 1.48 | 2.40 | 1.30 | 1.69 | 2.10 |
| | Organic carbon density | ocd | $\mathrm{kg\ m^{-3}}$ | 1.90 | 2.46 | 3.20 | 2.00 | 2.45 | 3.00 |
| | pH ($H_2O$) | phh2o | – | 4.30 | 5.27 | 5.80 | 5.20 | 5.39 | 5.80 |
| | Soil organic carbon in fine earth | soc | $\mathrm{g\ kg^{-1}}$ | 8.80 | 18.63 | 35.40 | 14.10 | 17.85 | 24.20 |
| | Soil temperature | tsoil | – | 226.17 | 253.46 | 292.83 | 237.17 | 257.94 | 270.00 |
| | Temperature seasonality | sts | – | 86.10 | 155.20 | 255.70 | 242.70 | 349.49 | 401.90 |
| | Isothermality | iso | – | – 84.60 | 13.67 | 30.70 | 15.30 | 19.74 | 23.30 |
| | Mean diurnal range | mdr | – | – 2.00 | 1.17 | 2.40 | 1.90 | 2.47 | 3.00 |

Climatic features were obtained from 2000 to 2021

$$D_{ju} = \sqrt{\sum_{p=1}^{P}(w_{jp} - \omega_{up})^2} \tag{1}$$

where $w_{jp}$ and $\omega_{up}$ are entries of $\mathbf{W}$ and $\mathbf{\Omega}$ that represent the value of the $p$th environmental feature for the $j$th tested environment and the $u$th untested environment, respectively.

After calculating the distances between all $J$ and $U$ environments, we expanded these results to include all possible environments within the delimited prediction area using the inverse distance weighting (IDW) interpolation method. The IDW was performed using the `Spatstat` package (Baddeley et al. 2015). Let $u^\star$ represent an untested and unsampled environment ($u^\star = 1, 2, \ldots, U^\star$, with $U^\star \gg U$). The Euclidean distance between a given $j$ and $u^\star$ is defined as:

$$D_{u^\star j} = \frac{\sum_{u=1}^{U} \frac{1}{||u^\star - x_u||^\tau} D_{uj}}{\sum_{u=1}^{U} \frac{1}{||u^\star - x_u||^\tau}} \tag{2}$$

where $||u^\star - x_u||$ represents the Euclidean distance between $u^\star$ and a given sampled point $x_u$ within the observation window, and $\tau$ is a power of the multiplication determined through cross-validation (CV). Values of $\tau$ ranging from 0.1 to 5.0, with an increment of 0.1, were tested in the CV. The value that yielded the lowest mean squared error between the predicted and observed values at the sampled points was selected.

Once we have performed the interpolation and obtained the Euclidean distances between all tested and untested environments, we consider the environmental similarity between the $u$th (or $u^*$th) untested environment and the observed environments of the TPE to be the minimum distance of $u$ (or $u^\star$) to any $j$:

$$S_u = \min(D_{uj}) \quad \& \quad S^\star_u = \min(D_{u^\star j}) \tag{3}$$

## Phenotypic analysis

The phenotypic analyses across environments for both data sets were performed using the following linear mixed model (Henderson 1949, 1950) in the `ASReml-R` package (version 4.1.2, The VSNi Team 2023). Variance components were estimated using residual maximum likelihood (Patterson and Thompson 1971).

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}_1\mathbf{s} + \mathbf{X}_2\mathbf{r} + \mathbf{Z}_1\mathbf{g} + \boldsymbol{\epsilon} \tag{4}$$

where $\mathbf{y}$ is the vector of phenotypic records, $\mu\mathbf{1}$ is the intercept, $\mathbf{s}$ is the vector of fixed effects of environments with design matrix $\mathbf{X}_1$, $\mathbf{r}$ is the fixed vector of within-environment block effects with design matrix $\mathbf{X}_2$, $\mathbf{g}$ is the vector of random genotypic effects nested within environments with incidence matrix $\mathbf{Z}_1$, and $\boldsymbol{\epsilon}$ is the residual term. The distributional assumptions for $\mathbf{g}$ and $\boldsymbol{\epsilon}$ are detailed below.

Using the available information on the coordinates (row and column) of each plot in the soybean dataset, we implemented a strategy to control the spatial trends in a single step, following the approach proposed by Gogel et al. (2018). In summary, we conducted model testing in each environment, considering spatial analysis. These adjustments included incorporating autoregressive processes in the error term as well as linear and nonlinear effects as fixed or random terms, as previously demonstrated by Gilmour et al. (1997). We identified the best-fitting model for each specific environment. Once we determined the optimal model for each environment, we incorporated the factors from these models into Eq. (4). Each additional factor followed a block diagonal covariance structure, with non-nil effects only for environments where these factors were present in the best within-environment model. Detailed information about this procedure can be found in Supplementary Table 2. For spatially adjusted trials, the residual effects are distributed as $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}[\boldsymbol{\Gamma}_{C_j} \otimes \boldsymbol{\Gamma}_{R_j}])$, where $\boldsymbol{\Gamma}_{C_j}$ and $\boldsymbol{\Gamma}_{R_j}$ are autocorrelation matrices of dimensions $C_j \times C_j$ and $R_j \times R_j$, respectively. Here, $C_j$ represents the number of columns, and $R_j$ represents the number of rows in the $j$th trial. These matrices have a value of 1 on the diagonal, and the off-diagonal elements represent the autocorrelation coefficients that quantify the spatial trends in the column or row directions. For environments where no spatial adjustment was necessary, $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}\mathbf{I}_{N_j})$, where $\mathbf{I}_{N_j}$ is an identity matrix of order $N_j$, which corresponds to the number of phenotypic records per environment. $\oplus$ represents the direct sum, which generates a block diagonal matrix, and $\otimes$ denotes the Kronecker product. For the rice dataset, since we did not have access to spatial information, $\epsilon \sim MVN(\mathbf{0}, \oplus_{j=1}^{J}\sigma^2_{\epsilon_j}\mathbf{I}_{N_j})$.

Genotypic effects were modeled using the FA covariance structure (Piepho 1997; Smith et al. 2001):

$$\mathbf{g} = (\hat{\boldsymbol{\Lambda}} \otimes \mathbf{I}_V)\tilde{\mathbf{f}} + \tilde{\boldsymbol{\delta}} \tag{5}$$

where $\hat{\boldsymbol{\Lambda}}$ is the $J \times K$ matrix of $K$ loadings for the $J$ environments ($\hat{\boldsymbol{\Lambda}} = \{\hat{\lambda}_{k_j}\}$), $\tilde{\mathbf{f}}$ is a vector of $K$ scores for the $V$ genotypes ($\tilde{\mathbf{f}} = \{f_{k_v}\}$), and $\tilde{\boldsymbol{\delta}}$ is the vector of the $VJ$ lack of fit effects ($\tilde{\boldsymbol{\delta}} = \{\hat{\delta}_{v_j}\}$). $\mathbf{I}_V$ is an identity matrix of order $V$. $\tilde{\mathbf{f}}$ and $\tilde{\boldsymbol{\delta}}$ are independent and distributed as multivariate Gaussian with zero means and variances given by $\mathbf{D} \otimes \mathbf{I}_V$ and $\boldsymbol{\Psi} \otimes \mathbf{I}_V$, respectively. $\mathbf{D}$ is a $K \times K$ symmetric positive (semi)-definite factor score variance matrix, and $\boldsymbol{\Psi}$ is a $J \times J$ diagonal matrix of environment-wise variances that were not captured by any factor ($\hat{\boldsymbol{\Psi}} = \{\hat{\psi}_j\}$). For more information about the estimation process of $\hat{\boldsymbol{\Lambda}}$, $\tilde{\mathbf{f}}$, and $\tilde{\boldsymbol{\delta}}$, refer to Smith et al. (2001), Thompson et al. (2003), and Tolhurst et al. (2022).

### Rotation

We followed the rotation process recommended by Smith et al. (2021), where two constraints are imposed for the sake of interpretability: $\mathbf{D}$ is a diagonal matrix with elements arranged in decreasing order, and $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ is an identity matrix, i.e., $\boldsymbol{\Lambda}$ is composed of orthonormal columns. To address these conditions, we performed the singular value decomposition of $\hat{\boldsymbol{\Lambda}}$:

$$\hat{\boldsymbol{\Lambda}} = \mathbf{U}\mathbf{L}^{\frac{1}{2}}\mathbf{V}' \tag{6}$$

where $\mathbf{U}$ is an $J \times K$ orthonormal matrix whose columns are the eigenvectors of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$, $\mathbf{L}$ is a $K \times K$ diagonal matrix with elements given by the eigenvalues of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$ in decreasing order, and $\mathbf{V}$ is a $K \times K$ orthonormal matrix whose columns are the eigenvectors of $\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Lambda}}$. Note that $\mathbf{U}$ meets the conditions of the second constraint, so $\hat{\boldsymbol{\Lambda}}^\star = \mathbf{U}$, in which $\hat{\boldsymbol{\Lambda}}^\star$ is the matrix of rotated loadings. By considering $\mathbf{D} = \mathbf{L}$, we fulfill the condition of the first constraint. The rotated scores were obtained as $\tilde{\mathbf{f}}^\star = (\mathbf{D}\mathbf{V}' \otimes \mathbf{I}_V)\mathbf{f}$, where $\tilde{\mathbf{f}}^\star$ is the vector of rotated scores. After rotation, the conditional distribution of the genotypic effects is $\mathbf{g} \sim MVN[\mathbf{0}, (\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}}) \otimes \mathbf{I}_V]$.

## FA model selection

FA models with different numbers of factors were fitted and compared in terms of their explanatory ability. We used the average semivariance ratio (ASR, Piepho 2019; Chaves et al. 2023a) as a selection criterion. By calculating the ratio between the average semivariance of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star\prime}$ and the average semivariance of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star\prime} + \mathbf{\Psi}$, it is possible to investigate the amount of total covariance that is being captured by the factors of the FA model. The ASR is given as follows:

$$\text{ASR} = \frac{\frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \frac{1}{2}\left(\sum_{k=1}^{K} \hat{\lambda}_{k_j}^{\star 2} d_k + \sum_{k=1}^{K} \hat{\lambda}_{k_{j'}}^{\star 2} d_k\right) - \sum_{k=1}^{K} \hat{\lambda}_{k_j}^{\star} \hat{\lambda}_{k_{j'}}^{\star} d_k}{\frac{2}{J(J-1)} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \frac{1}{2}\left[\left(\sum_{k=1}^{K} \lambda_{k_j}^{\star 2} d_k + \hat{\psi}_j\right) + \left(\sum_{k=1}^{K} \hat{\lambda}_{k_{j'}}^{\star 2} d_k + \hat{\psi}_{j'}\right)\right] - \sum_{k=1}^{K} \hat{\lambda}_{k_j}^{\star} \hat{\lambda}_{k_{j'}}^{\star} d_k} \times 100 \tag{7}$$

where $d_k$ is the $k$th element of the diagonal of $\mathbf{D}$.

We defined an ad hoc threshold of 75% for explanatory ability. As complementary information, we also estimated the proportion of genetic variance explained by the $k$th factor in the $j$th environment ($v_{k_j}$, Smith et al. 2015):

$$v_{k_j} = \frac{\hat{\lambda}_{k_j}^{\star 2} d_k}{\sum_{k=1}^{K} \hat{\lambda}_{k_j}^{\star 2} d_k + \hat{\psi}_j} \times 100 \tag{8}$$

From the best-fit model, we estimated some useful parameters to investigate the experimental precision, such as the environment-wise generalized heritabilities (Cullis et al. 2006) and coefficients of experimental variation (CV), which are given by the following equations, respectively:

$$H_j^2 = 1 - \left(\frac{\bar{v}_{\Delta}^{\text{BLUP}}}{2\sigma_{g_j}^2}\right) \tag{9}$$

$$\text{CV}_j = \frac{\sigma_{e_j}}{\mu_j} \tag{10}$$

where $\bar{v}_{\Delta}^{\text{BLUP}}$ is the average pairwise prediction error variance, $\sigma_{g_j}^2$ is the genotypic variance for the $j$th environment, taken from the diagonal elements of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}}$; $\sigma_{e_j}$ is the estimated residual standard deviation for the $j$th environment, and $\mu_j$ is the mean of the trait for the $j$th environment.

## Genotype-by-environment interaction investigation tools

We investigated the GEI dynamics in the datasets by examining the pairwise genetic correlations between environments and the partitioning of GEI variance into crossover and non-crossover patterns. The pairwise genetic correlation between environments ($\rho_{jj'}$) is given as follows (Cullis et al. 2010):

$$\mathbf{\Upsilon} = \mathbf{\Delta}(\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}})\mathbf{\Delta} \tag{11}$$

where $\mathbf{\Upsilon}$ is a $J \times J$ matrix of genetic correlations, and $\mathbf{\Delta}$ is a diagonal matrix whose elements are the inverse of the square roots of the diagonal values of $\hat{\mathbf{\Lambda}}^{\star}\mathbf{D}\hat{\mathbf{\Lambda}}^{\star'} + \hat{\mathbf{\Psi}}$.

The decomposition of the GEI variance was performed using the following equation, adapted from Cooper and Delacy (1994):

$$\sigma^2_{\text{ge}_{\text{rank}}} = 1 - \frac{\text{Var}\left(\sqrt{\sigma^2_{g_j}}\right)}{\sigma^2_{ge}} \tag{12}$$

where $\sigma^2_{\text{ge}}$ is the variance attributed to the GEI, which is determined by fitting a compound symmetry model. This model has the same structure as Eq. (4), but the variance–covariance matrix of genetic effects has the form $\sigma^2_g\mathbf{J} + \sigma^2_{\text{ge}}\mathbf{I}_J$, where $\mathbf{J}$ is a $J \times J$ matrix of ones.

## Selection tools for overall performance and stability

The target features of most breeding programs are to achieve high performance and stability across the TPE. Using the best-fit FA model, we estimated metrics to assess the performance and stability of genotypes. The performance was measured using the OP metric ($\text{OP}_v$), which was obtained as follows (Stefanova and Buirchell 2010; Smith and Cullis 2018):

$$\text{OP}_v = \frac{1}{J}\sum_{j=1}^{J} \hat{\lambda}_{1_j}^{\star} f_{1_v}^{\star} \tag{13}$$

Note that only the first factor is used to compute the $\text{OP}_v$. This factor captures the largest portion of the total variance. Thus, it provides a generalized measure of the genetic main effects (Supplementary Figure 2; Stefanova and Buirchell 2010). According to empirical observations by Smith and Cullis (2018), this is valid when the majority of loadings in the first factor are positive, indicating the absence (or insignificance) of crossover GEI in the first factor. Using this principle, the other factors are used to represent stability. Considering that the genetic effect of a given genotype $v$ at the $j$th environment, disregarding the lack of fit effect, is $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \hat{\lambda}^{\star}_{2_j} f^{\star}_{2_v} + \cdots + \hat{\lambda}^{\star}_{K_j} f^{\star}_{K_v}$, which is equivalent to $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \epsilon_{vj}$, the stability of $v$ is given by:

$$\mathrm{RMSD}_v = \sqrt{\frac{1}{J}\sum_{j=1}^{J} \epsilon^2_{vj}} \tag{14}$$

in which $\mathrm{RMSD}_v$ is the root-mean-square deviation of $v$, representing the distance between the point and the slope in a latent regression given by $g_{vj} = \hat{\lambda}^{\star}_{1_j} f^{\star}_{1_v} + \epsilon_{vj}$ (Smith and Cullis 2018).

A desirable genotype $i$ has a high $\mathrm{OP}_i$ and a low $\mathrm{RMSD}_i$. Following these principles, we applied a selection index ($\mathrm{SI}_v$) with these metrics (Chaves et al. 2023b; Cowling et al. 2023), given as follows:

$$\mathrm{SI}_v = 2 \times \frac{\mathrm{OP}_v - \overline{\mathrm{OP}}}{\sqrt{V(\mathrm{OP})}} - \frac{\mathrm{RMSD}_v - \overline{\mathrm{RMSD}}}{\sqrt{V(\mathrm{RMSD})}} \tag{15}$$

In addition to the selection index, the reliability of the $v$th genotype (Mrode 2014) was calculated as follows:

$$r_v = 1 - \frac{\mathrm{PEV}_v}{\overline{\sigma^2_g}} \tag{16}$$

where $\mathrm{PEV}_v$ represents the prediction error variance of the $v$th genotype, and $\overline{\sigma^2_g}$ is the average genotypic variance across environments. The reliability metric associated with the selection index is useful for improving the accuracy of selection, especially when dealing with unbalanced data sets. We adopted a selection intensity of 15% for both datasets.

## Spatial predictions in the breeding zone

In this study, GIS tools were used to: (1) collect georeferenced data from the evaluated trials, (2) build environmental markers, and (3) perform spatial predictions for a larger area. Here, we used PLS regression (Wold 1966; Aastveit and Martens 1986) to make the predictions. This method is useful when the number of predictors is much larger than the number of observations and when these predictors are correlated. When PLS is used to predict genotypic performances in untested environments, the response variable is the genotypic performance in the testing set. In this situation, the response variable is a $J \times 1$ vector ($\mathbf{y}$) of phenotypic records if a genotype-wise PLS model is fitted or a $J \times V$ matrix ($\mathbf{Y}$) when a multivariate PLS model is fitted considering all genotypes at once (Monteverde et al. 2019; Costa-Neto et al. 2022). We refer to the multivariate model as GIS-GGE.

We modified GIS-GGE by using the rotated loadings of the tested environments ($\hat{\lambda}^{\star}_{k_j}$) as response variables instead of the within-environment phenotypic records of the genotypes. This procedure we called GIS-FA. We obtained these loadings from the previously chosen FA model (section FA model selection). With the predicted loadings and the previously estimated scores for each genotype from the FA model, we can predict the empirical BLUPs of the genotypes in untested environments. The PLS regression model was trained using the rotated loadings and environmental features of the tested environments:

$$\hat{\mathbf{\Lambda}}^{\star} = \mathbf{W}\mathbf{B}^{\star} + \mathbf{E} \tag{17}$$

where $\mathbf{B}^{\star}$ is a $P \times K$ vector of coefficients, $\mathbf{E}$ is a $J \times K$ matrix of lack-of-fit effects, and $\hat{\mathbf{\Lambda}}^{\star}$ and $\mathbf{W}$ are previously described in sections Phenotypic analysis and Environmental similarity and interpolation grid, respectively. We obtained $\mathbf{B}^{\star}$ using a kernel PLS algorithm (Lindgren et al. 1993; Dayal and MacGregor 1997) implemented in the `pls` package (Liland et al. 2022). This algorithm is detailed in Appendix A.

After training the model, we substituted $\mathbf{W}$ with $\mathbf{\Omega}$ to predict the $K$ loadings of the $U$ untested environments:

$$\hat{\mathbf{\Lambda}}^{\star}_U = \mathbf{\Omega}\mathbf{B}^{\star} + \mathbf{E} \tag{18}$$

Recall from section Environmental information that $\mathbf{\Omega}$ was built using historical weather data from 2000 to 2021, as well as soil environmental features. Once we predicted the loadings of untested environments, we used them in linear combinations with the previously predicted scores of each genotype (see section Phenotypic analysis) to estimate their eBLUPs within untested environments:

$$\mathbf{g}_U = (\hat{\mathbf{\Lambda}}^{\star}_U \otimes \mathbf{I}_V)\tilde{\mathbf{f}}^{\star} \tag{19}$$

Note that we use the same scores to predict the eBLUPs of both tested and untested environments. Nevertheless, the scores are predicted based solely on the data collected from the tested environments. In other words, the environments in the data set must accurately reflect the TPE so that the loadings of the untested environments closely match the loadings of the tested ones.

A CV process is required to obtain $\mathbf{B}^{\star}$. We employed a leave-one-out scheme, where data from a single environment were removed (the testing set), and predictions were made using the information provided by the remaining environments (the training set). The predicted eBLUPs were then correlated with the actual eBLUPs and eBLUEs of each environment to determine the predictive ability of the PLS regression model. The model with the highest number of components demonstrating predictive ability was chosen. We leveraged the same CV scheme to compare the predictive ability of GIS-FA and GIS-GGE. In this study, the PLS regression of GIS-GGE was trained with the within-environment empirical eBLUPs of each genotype as response variables.

### Thematic maps

Thematic maps combine cartographic principles and GIS tools to represent and analyze spatial and geographic phenomena. The incorporation of spatial interpolation methods enables the estimation of values in untested locations, resulting in a seamless representation of the phenomenon. This facilitates the identification of patterns and trends, aiding decision-making across various fields of study (Costa-Neto et al. 2020).

Recall that $\mathbf{\Omega}$ has $U$ rows, and the predictions must be extrapolated to all $U^{\star}$ untested environments within the targeted area. For this purpose, we used an interpolation process similar to the one described in section Environmental similarity and interpolation grid. The difference is that for the environmental similarity maps, we interpolated Euclidean distances, while for the thematic maps described in this section, we interpolated eBLUPs. Once the spatial prediction was interpolated across the whole breeding region, we built thematic maps to aid in the visualization and interpretation of the results. We created maps with three themes:

- Adaptation zones: These maps depict the expected spatial prediction of each selection candidate across the breeding zone. The adaptation of a genotype to an environment is assessed by the expected response of that genotype when it is planted in that environment. Thus, in this context, "adaptation" is used as a synonym for specific performance. For improved visualization, we divided the predicted eBLUPs into eight categories (from expected yield lower than 2500 kg ha$^{-1}$ to expected yield higher than 4000 kg ha$^{-1}$), and each category was then assigned a specific color.
- Pairwise comparisons: These maps allow for a direct comparison of the expected responses of different genotypes in specific environments. Two distinct colors, one for each candidate, were used to indicate that the superior selection candidate was superior in each location on the map. This visual representation helps to quickly identify which selection candidate outperforms the other in each pixel, facilitating the interpretation of competitive advantages among genotypes in specific environments.
- Which-won-where: The genotype that achieved the best performance in each location on the map is highlighted. This map provides a clear depiction of the winning genotype for each specific location, enabling a comprehensive understanding of the distribution of high-performing genotypes across the breeding zone.

These maps, like all the other plots, were built using the `ggplot2` package (Wickham 2016), with the addition of the `ggspatial` (Dunnington 2023) and `sf` (Pebesma and Bivand 2023) packages. The shapefiles we used are freely available at the Brazilian Institute of Geography and Statistics (IBGE in the Portuguese acronym) website (https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html), or they can be downloaded using the `geodata` package. The Supplementary Material has the commented `R` scripts used to perform GIS-FA, and users can reproduce it using the soybean dataset, freely available at https://github.com/Kaio-Olimpio/GIS-FA/tree/main.

## Results

### Experimental accuracy

In the rice dataset, CV$_j$ ranged from 0.11 (E20) to 0.34 (E13), and $H_j^2$ ranged from 0.31 (E08) to 0.78 (E18) (Fig. 2a). In the soybean dataset, CV$_j$ ranged from 0.04 (E31) to 0.17 (E42), and $H_j^2$ ranged from 0.31 (E18) to 0.77 (E31) (Fig. 2b). Spatial trends were modeled in 37 out of 49 soybean trials (Supplementary Table 2).

### Genotype recommendations for tested environments

The FA model with four factors (FA4) met our criteria for both datasets. It explained more than 75% of the variance (Table 2). This model captured most of the within-environment variance in both datasets (Supplementary Figure 3).

The genotypic correlations ranged from − 0.0031 (E07 *vs.* E19) to 0.8936 (E13 *vs.* E19) for the rice dataset (Fig. 3a) and from − 0.0010 (E07 *vs.* E41) to 0.9753 (E031 *vs.* E32) for the soybean dataset (Fig. 3b). In the rice dataset, environments E17 and E18 exhibited the most contrasting patterns compared to the other environments. Their correlations with the remaining environments were predominantly negative or close to zero. Similarly, in the soybean dataset, negative or negligible correlations were observed for contrasts involving environments E18, E33, E34, E43, E46, and E47. These findings indicate substantial differences between these specific environments and the rest of the dataset. The wide range of correlation magnitudes is reflected in the percentage of crossover GEI in the datasets: 76 and 81% of the total GEI were due to crossover interactions in the rice and soybean datasets, respectively.

The selected candidates based on the selection index are highlighted in Fig. 4. Despite the low reliability of the rice

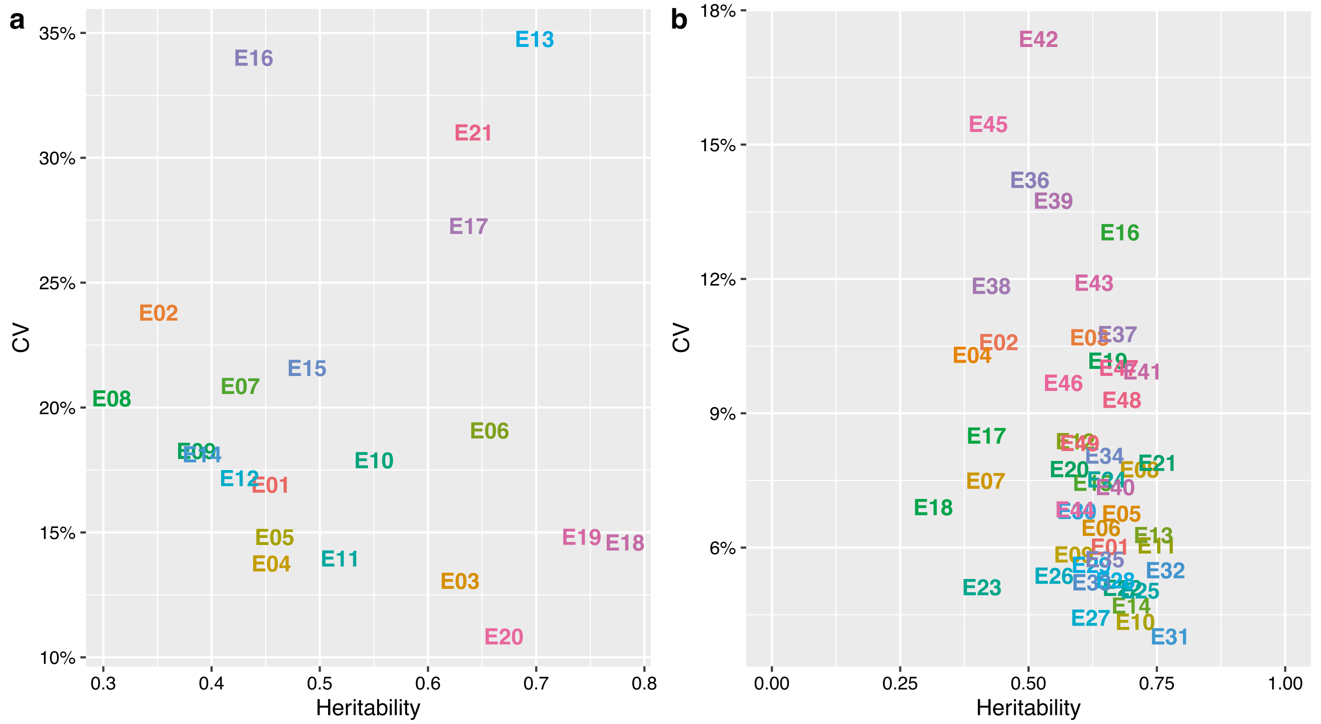

**Fig. 2** Scatter plot representing the experimental coefficient of variation (CV, on a decimal scale) in the *y*-axis and the generalized heritability in the *x*-axis for grain yield (kg ha$^{-1}$) of rice (**a**) and seed yield (kg ha$^{-1}$) of soybean (**b**) trials

**Table 2** Fitted factor-analytic mixed models for each dataset (rice and soybean) and their respective logarithm of the likelihood function (LogL), number of parameters (no. par.), and average semivariance ratio (ASR)

| Model | LogL | no. par | ASR |
|---|---|---|---|
| *Rice data* | | | |
| FA1 | - 10,482.19 | 61 | 21.38 |
| FA2 | - 10,470.01 | 76 | 51.60 |
| FA3 | - 10,456.57 | 92 | 66.85 |
| FA4 | **- 10,444.36** | **108** | **78.70** |
| *Soybean data* | | | |
| FA1 | - 37,722.05 | 227 | 22.58 |
| FA2 | - 37,627.68 | 276 | 54.08 |
| FA3 | - 37,578.47 | 332 | 69.08 |
| FA4 | **- 37,528.83** | **336** | **76.63** |
| FA5 | - 37,462.72 | 400 | 83.52 |
| FA6 | - 37,418.20 | 433 | 91.40 |
| FA7 | - 37,374.42 | 468 | 93.92 |

In the rice dataset, models with five factors onward had singularity issues. The selected models are in bold

dataset, genotypes G23, G18, G29, G31, and G26 stand out for their high stability. Genotypes G10, G09, G03, and G01 presented high OP and reliability. The check treatment (C83) had the highest OP, but it exhibited low stability and reliability compared to the other selected genotypes (Fig. 4a). Among the soybean genotypes, G178, G031, G101, G052, and G035 exhibited the highest stability. On the other hand, G177, G100, G144, G088, and G016 were notable for their high OP. Genotype G16 showed high OP, stability, and reliability (Fig. 4b). The reliability of the selected candidates was higher in the soybean dataset.

## Predictions using environmental markers in untested environments

### Environmental similarity

The rice trials are spread throughout the breeding region and effectively capture the environmental conditions of the area being studied (Fig. 5a). On the other hand, the trials in the soybean dataset are concentrated in the central part of the state, while there is a region to the west that exhibits low similarity. This area corresponds exactly to the Pantanal biome, which is a protected area with legal restrictions on soybean planting (Fig. 5b). This is probably the reason why there is no trial in this region.

### GIS-FA validation

In comparison with GIS-GGE, our proposal yields a higher prediction accuracy (as measured by the simple correlation between predicted and observed values) for both datasets.

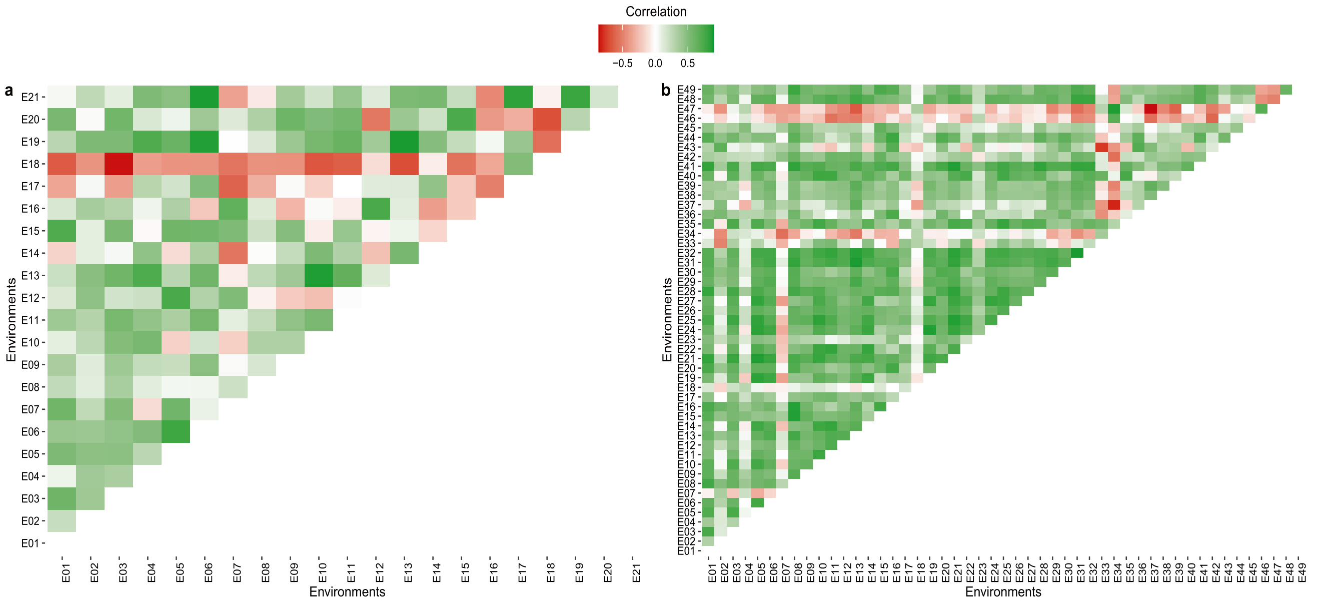

**Fig. 3** Heatmap representing the genetic correlation between pairs of environments in the rice (**a**) and soybean (**b**) datasets. The color gradient depicts the direction of the correlation: Red designates a negative correlation, whereas green represents a positive correlation

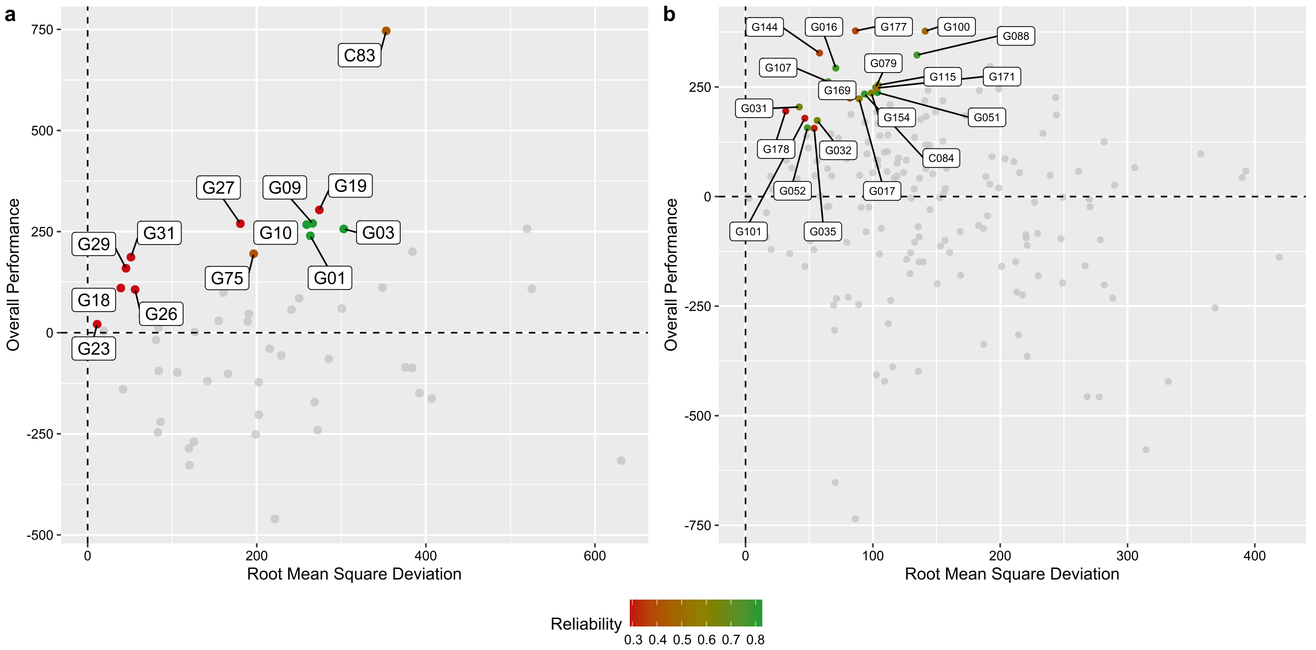

**Fig. 4** Overall performance (*y*-axis) and root-mean-square deviation (*x*-axis) of the experimental genotypes in the rice (**a**) and soybean (**b**) datasets. The most productive genotypes are oriented toward the upper part on the *y*-axis, and the most stable ones are toward the left in the *x*-axis

For predicting eBLUEs, GIS-FA is 10 and 1% better than GIS-GGE in the rice and soybean datasets, respectively. For predicting eBLUPs, GIS-FA is 9 and 5% more effective than GIS-GGE. A second way to assess the predictive ability of the methods is to check the coincidence between the top 10% of observed and predicted values (Fig. 6). GIS-FA provides more assertive results (Fig. 6a, b) than GIS-GGE (Fig. 6c, d). In other words, when recommending elite candidates based on predicted values, it is more probable that the true top performers will be recommended using GIS-FA than using GIS-GGE. In the rice dataset (Fig. 6a, c), GIS-FA has an accuracy that is 13.15 percentage points higher than

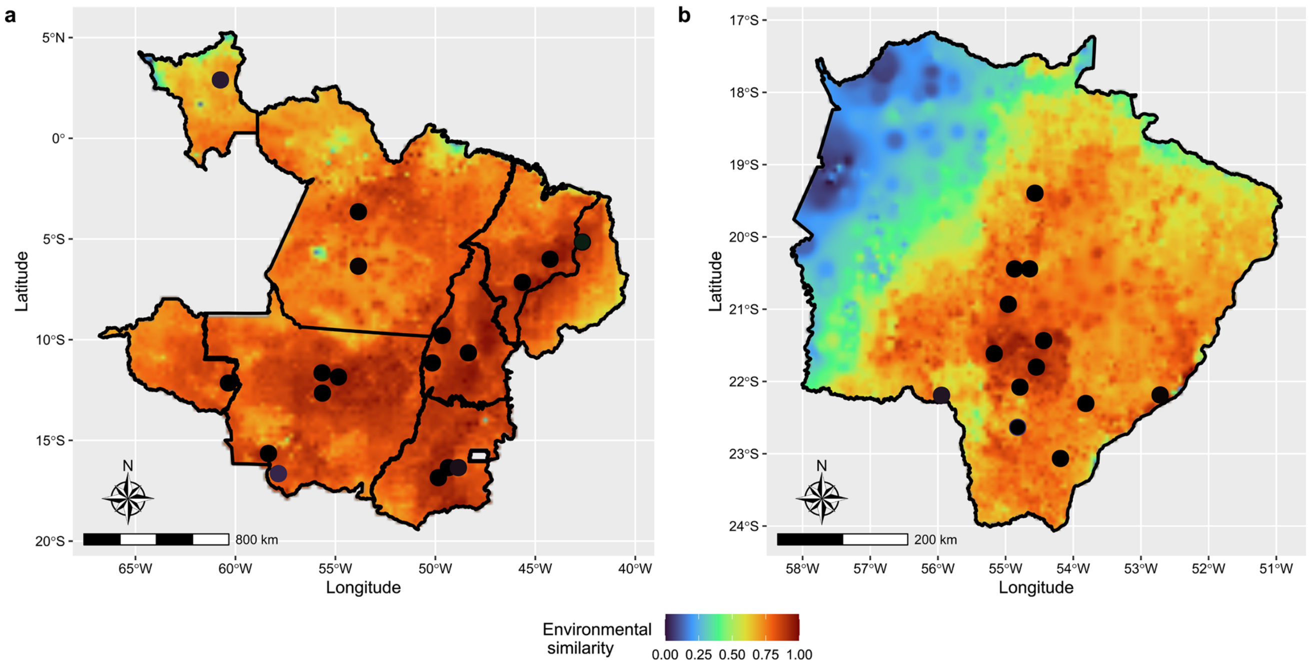

**Fig. 5** Environmental similarity between tested and untested environments in the target population of environments in the rice (**a**) dataset and in the soybean (**b**) dataset. The warmer the color, the higher the similarity, and consequently, the higher the prediction reliability. Colored circles represent the trials' locations

GIS-GGE. In the soybean dataset (Fig. 6b, d), GIS-FA is 21.19 percentage points more advantageous than GIS-GGE.

### Thematic maps of adaptation zones

The spatial prediction done by GIS-FA was useful in assessing the expected performance of the experimental genotypes in untested environments. This helps to define adaptation zones for each genotype, which are the theme of the maps in Fig. 7. For example, G16 of the rice dataset, shown in Fig. 7a, seems to be well adapted only in a small portion of Goiás State (green region), and it responds poorly to the environmental effects of other locations within the breeding region. Conversely, G27 of the rice dataset, shown in Fig. 7b, exhibits a broader spectrum in terms of adaptation in the breeding region. The same interpretation applies to the genotypes in the soybean dataset. G064 (Fig. 7c) is an unstable candidate, with a very restricted area where it is better adapted (in the northern part of the breeding region). On the other hand, G088 (Fig. 7d) is a stable genotype, meaning it possesses alleles that respond favorably to the environmental effects of different locations across the state. In each map, we provide the OP and RMSD of the corresponding genotype. We have deliberately chosen two promising candidates (Rice's G27 and soybean's G088, which are among those selected in Fig. 4), as well as two low-yielding genotypes (Rice's G16 and soybean's G064), to be included in Fig. 7. Nevertheless, we recommend using OP and RMSD as criteria to choose the genotype for which an adaptation map should be created.

### Thematic maps of pairwise comparison

To support the decision-making process, we developed a second thematic map: the pairwise comparison maps (Fig. 8), which facilitate the comparison of two candidates. Take, for example, G10 and G19 in Fig. 4a and G100 and G177 in Fig. 4b. These candidates have somewhat similar performances, according to their OP and RMSD. However, they are clearly adapted to different zones within the breeding region. G10 shows better responses at lower latitudes, while G19 is more suitable for higher latitudes (Fig. 8b). G100 is better adapted to the central region of the soybean's breeding region, and G177 is more compatible with the environmental conditions at the breeding region's horizontal extremes (Fig. 8d).

### Thematic maps of which-won-where

The which-won-where map (Fig. 9) shows the experimental genotype that is most suitable for a specific environment within the breeding zone. In the rice dataset (Fig. 9a), G10 emerges as the most promising experimental genotype in almost all environments in the central and northern portions of the breeding zone, while G19 prevails in the southern and eastern regions. G09, G16, G17, and G20 are the most suitable for specific environments. The breeding region of the soybean dataset is more diverse, with G177, G100, G170, and G088 being the most important experimental genotypes, as they have emerged as the winners in the widest area. The other selection candidates, including a cultivar check (C054), are the top performers in only a few restricted environments (Fig. 9b).

## Discussion

The GIS-FA method represents the integration of modern statistical genetics with GIS principles. We showed how GIS-FA can aid plant breeders in making decisions by considering the observed performance in tested environments and spatial predictions in untested environments. For observed environments, GIS-FA leverages the resources of FA models to provide useful inferences about the dynamics of the GEI and to select candidates with high performance and stability using customized selection tools (Stefanova and Buirchell 2010; Smith and Cullis 2018). In untested environments, GIS-FA allows the recommendation of cultivars based on spatial predictions derived from soil characteristics, climatic conditions, and empirical data parameters (i.e., factor loadings for genotypes). The GIS-FA method allows for data-driven decision-making with the aid of graphical tools such as thematic maps. These maps include (i) adaptation zone maps, which depict the expected spatial prediction of each genotype within the entire breeding zone; (ii) pairwise comparison maps, which facilitate the comparison of performance between two selection candidates (or a candidate and a commercial check); and (iii) which-won-where maps, which show the most promising experimental genotype (the winner) in each location within the breeding zone.

### Genotype-by-environment interaction and selection in tested environments

Increasing crop yield and adapting to different growing conditions are important goals in plant breeding. These traits are the outcomes of a plethora of small quantitative trait loci (QTLs) effects that are highly influenced by the environment (Lynch and Walsh 1998; Crossa 2012). In terms of cultivar recommendation in the TPE, the most concerning source of the GEI is the lack of genotypic correlation between environments (Cooper and Delacy 1994), as observed in both data sets (Fig. 3). As a consequence, it is unlikely that the same set of experimental genotypes will exhibit similar performance across uncorrelated environments. In this case, if a global (i.e., across environments) recommendation is needed, metrics such as the selection index, which combines performance and stability, might be employed. The weight of each metric in the selection index is determined by the breeder (Chaves et al. 2023b). Here, we prioritized performance over stability.

In the GIS-FA method, we leverage the resources of FA mixed models (Piepho 1997; Smith et al. 2001) that explore the complexity of the GEI while handling highly unbalanced data sets. Furthermore, FA models allow for a parsimonious estimation of environment-wise genotypic variances and pairwise covariances. These covariances can be used to investigate the dynamics of GEI, as fully described in this study. The efficiency of the GIS-FA method depends on the choice of the number of factor loadings in the FA model, i.e., a poor choice will provide erroneous results. In GIS-FA, it is important to note that the factor loadings of observed environments are used as the training set. This allows for the prediction of the loadings of untested environments in the testing set. Thus, when selecting the best-fit FA model, selection criteria such as the ASR should always be considered. Naturally, using more factors will provide greater explanatory ability. Nevertheless, it will hinder parsimony and computational efficiency, especially in large data sets.

Assuming that the observed environments accurately represent the expected environmental conditions throughout the breeding zone, the most promising genotypes in the tested environments are probably the best ones in the untested environments. Thereby, the idea is to prioritize selected experimental genotypes when drawing the thematic maps "genotype-wise adaptation" and "pairwise comparisons."

### Spatial interpolations in untested environments

Like molecular markers, environmental feature similarity can be used for both inference and prediction purposes. Inference models aim to determine the effect of each environmental feature on phenotypic expression and the GEI, which is analogous to QTL mapping models (Denis 1988; Van Eeuwijk and Elgersma 1993; Crossa et al. 1999; Costa-Neto et al. 2021c; Heinemann et al. 2022). In this work, we focused on environmental-wide predictions, regardless of the particular effect of each EF on phenotypic expression and GEI. As polygenic models are used to perform whole-genome regressions (Meuwissen et al. 2001), we assumed that the core of ecophysiological effects captured by the environmental feature could be sufficient to generate genotype-wise predictions across the spatial grid. The benefits of incorporating environmental features into predictive breeding are advantageous in most cases, whether integrated with genomic information or not (de los Campos et al. 2020; Buntaran et al. 2021; Jarquín et al. 2021; Costa-Neto et al. 2022). However, recent work from Crossa et al. (2023) demonstrated that the inclusion of environmental covariates could either increase or decrease prediction accuracy, depending on the specific case. Techniques such as feature selection (Crossa et al. 2023) and exhaustive search

(Li et al. 2018) can be considered when selecting environmental features.

### Environmental similarity

Environmental similarity maps revealed a need to perform an adequate sampling of the different environmental types within a given target breeding region (Fig. 5). This entails including samples from various climatic conditions and soil traits that may be encountered in future predictive environments. Essentially, these maps illustrate a metric of reliability for spatial predictions by benchmarking the similarity between observed and unobserved environments. They demonstrate the environmental similarity between tested and untested environments. In other words, the more similar an untested environment is to a tested environment, the higher the chances of making an assertive prediction. The results depicted in the maps of Fig. 5 can be attributed to the geographical distribution of trials in relation to the Brazilian biomes [refer to Figure 1 of Chaves et al. 2023b for a map with the Brazilian biomes]. The soybean breeding region comprises two biomes, namely the Pantanal (wet lowlands) and the Cerrado (highland savanna conditions). All trials were conducted in the Cerrado, which explains the lack of similarity between the TPE and the environments in the Pantanal biome. Consequently, the prediction for this particular region is likely to be compromised. The rice breeding region also includes two biomes: Amazonia (a wet tropical rainforest) and Cerrado. Unlike the soybean dataset, there are representative trials from both biomes, providing comprehensive coverage of the relevant environmental conditions.

### Predicting using partial least squares regression

The association between PLS regression, GEI, and environmental features was introduced by Aastveit and Martens (1986) for inference purposes. Their aim was to address challenges related to the curse of dimensionality and multicollinearity in explaining the dynamics of GEI using two datasets. Their model was later expanded to include information on molecular markers to investigate QTL-by-environment interactions (Crossa et al. 1999; Vargas et al. 2006). Nevertheless, employing environmental features in statistical models to explain and predict GEI has not gained significant popularity among plant breeders (Vargas et al. 2001; Ortiz et al. 2007; Ramburan et al. 2012; Porker et al. 2020). With the advancement of computational technology and the democratization of "enviromics" resources, PLS has emerged as a suitable method for exploring big data and performing spatial predictions of experimental genotypes in new environments (Monteverde et al. 2019; Rincent et al. 2019; Guo et al. 2021; Costa-Neto et al. 2022). In fact, PLS has emerged as a relevant alternative for prediction purposes, even when breeders do not specifically incorporate environmental data into the model (Ortiz et al. 2023).

In most studies that employed PLS regression for prediction purposes, the training set typically consisted of the performance per se of genotypes and environmental features from the tested environments (Monteverde et al. 2019; Costa-Neto et al. 2022). Our study demonstrated that associating environmental features with the rotated factor loadings of the tested environment yields superior results. Through GIS-FA, we achieved higher prediction accuracy (Table 3) and enhanced the ability to distinguish high-performance experimental genotypes when relying solely on predicted values (Fig. 6). By predicting the factor loadings for untested environments, we establish a connection between the observed environmental feature values and the underlying causes of GEI, as well as the genetic covariance that exists between environments. A prior study by Rincent et al. (2019) also utilized PLS models to predict latent factors of the AMMI components for untested environments. This approach enabled them to construct an appropriate covariance structure that improved the accuracy of their predictions. The findings of Rincent et al. (2019) and the results of this work provide evidence of the potential of using PLS models to indirectly perform spatial predictions by initially predicting the latent elements that contribute to a particular performance. A similar strategy was proposed in a single-step model by Tolhurst et al. (2022), who demonstrated the efficiency of combining known and latent environmental features to predict both tested and untested environments.

### Thematic maps

An important feature of GIS-FA is the illustration of the spatial predictions from selection candidates using thematic maps (Figs. 7, 8, and 9). Figure 7 offers information on the areas within the breeding zone where the experimental genotypes are expected to thrive. Figure 7 allows the evaluation of the merit of a certain candidate cultivar based on its ability to outperform a commercial cultivar used as a reference or another promising experimental genotype.

**Table 3** Prediction accuracy of eBLUEs and eBLUPs using the proposed method GIS-FA and the conventional method GIS-GGE

| Model | Prediction | Prediction accuracy | |
|---|---|---|---|
| | | Rice data | Soybean data |
| GIS-GGE | BLUE | 0.40 | 0.53 |
| GIS-GGE | BLUP | 0.55 | 0.71 |
| GIS-FA | BLUE | 0.44 | 0.55 |
| GIS-FA | BLUP | 0.60 | 0.74 |

For more information about these methods, see the Material and Methods section

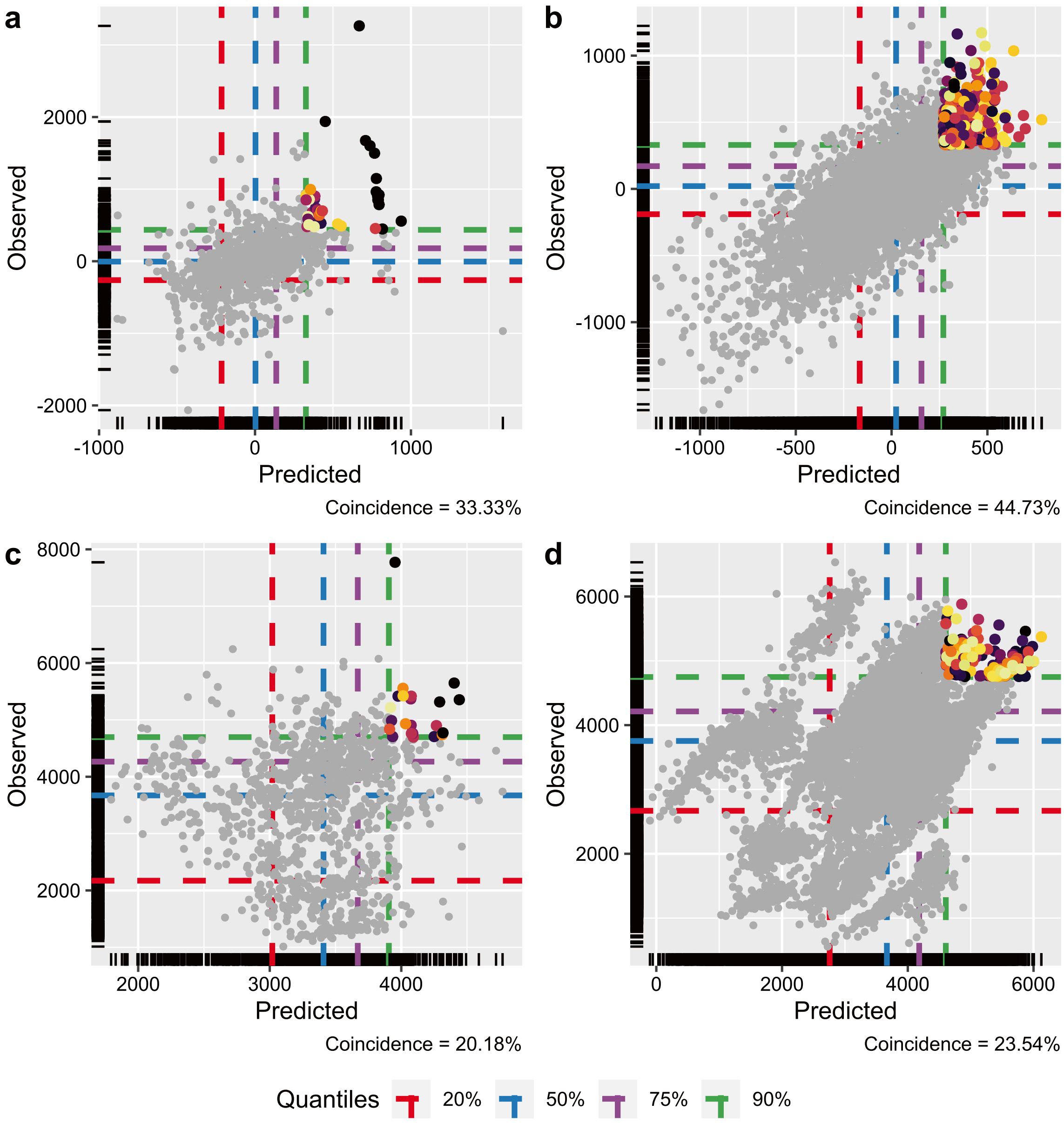

**Fig. 6** Scatter plot of all predicted values ($x$-axis) in the leave-one-out cross-validation scheme against observed values ($y$-axis). The dashed lines represent the empirical percentiles (20, 50, 75, and 90%) associated with the trait value. The colored dots represent the coincident selection candidates when selecting the top 10% performers using observed and predicted values. Each color represents a different genotype. "Coincidence" in the lower left corner of each plot depicts the accuracy of selecting the top 10% using the predicted values. **a, b** Illustrate the results for the GIS-FA method in the rice and soybean datasets, respectively. **c, d** Represent the results for the GIS-GGE method in the rice and soybean datasets, respectively

Figure 9 provides a straightforward solution for genotype recommendations across the breeding region, indicating which candidate is more suitable for a specific environment within the breeding zone. Thematic maps serve as valuable tools in decision-making, assisting in the allocation of genotypes in the breeding region (Costa-Neto et al. 2020; Bustos-Korts et al. 2022). In addition, the thematic maps provide information on the genotypes' stability and adaptation from a geographic perspective. Costa-Neto et al. (2020) suggested that, in a GIS context, "stability" means lower variability in spatial patterns, while "adaptation" refers to the expected performance in a specific environment in the breeding region.

One advantage of this approach is the possibility of integrating high-quality satellite images from diverse platforms. Here, we used freely available geographic databases on online platforms to achieve an efficient prediction method without incurring any additional costs.

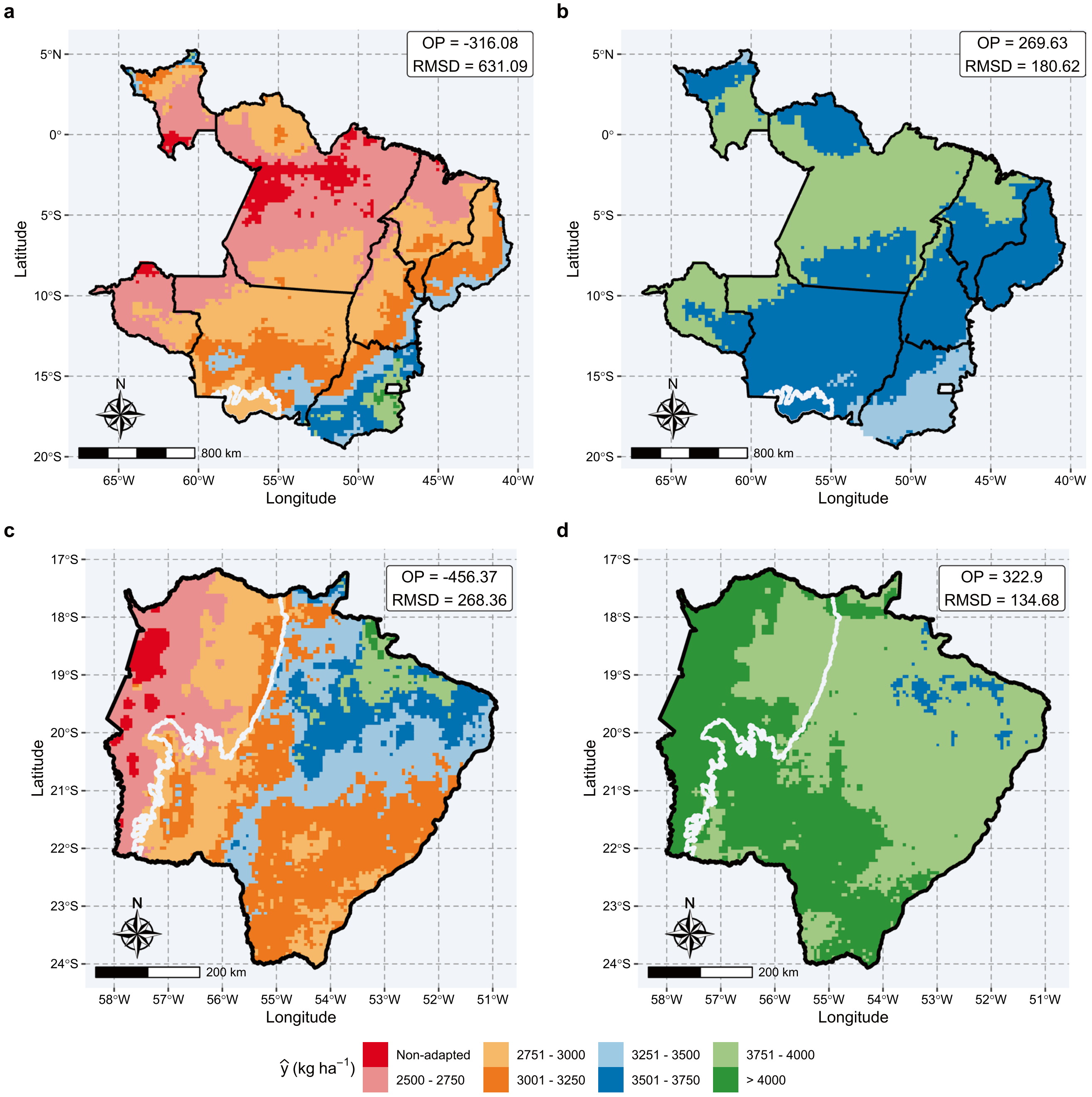

**Fig. 7** Genotype-wise adaptation map showing the adaptation zones of the genotypes G16 (rice dataset, **a**), G27 (rice dataset, **b**), G064 (soybean dataset, **c**), and G088 (soybean dataset, **d**). The color scale represents the expected yield classes, from non-adapted (intense red) to more than 4000 kg ha$^{-1}$ (intense green). The white contour delimits the Pantanal biome. On the upper right of each map, we provide the overall performance (OP) and root-mean-square deviation (RMSD) of each genotype

Furthermore, implementing partial geographic visualizations can optimize resource allocation when defining the experimental network of trials. The higher resolution of the satellite-based data could enable the delivery of spatial predictions at the farmer's level. This could benefit the product development and placement stages by extending this methodology to accommodate satellite-based enviromics while also accounting for historical agronomic records.

## Future directions

The statistical models of GIS-FA can be improved by integrating molecular information to leverage the covariance

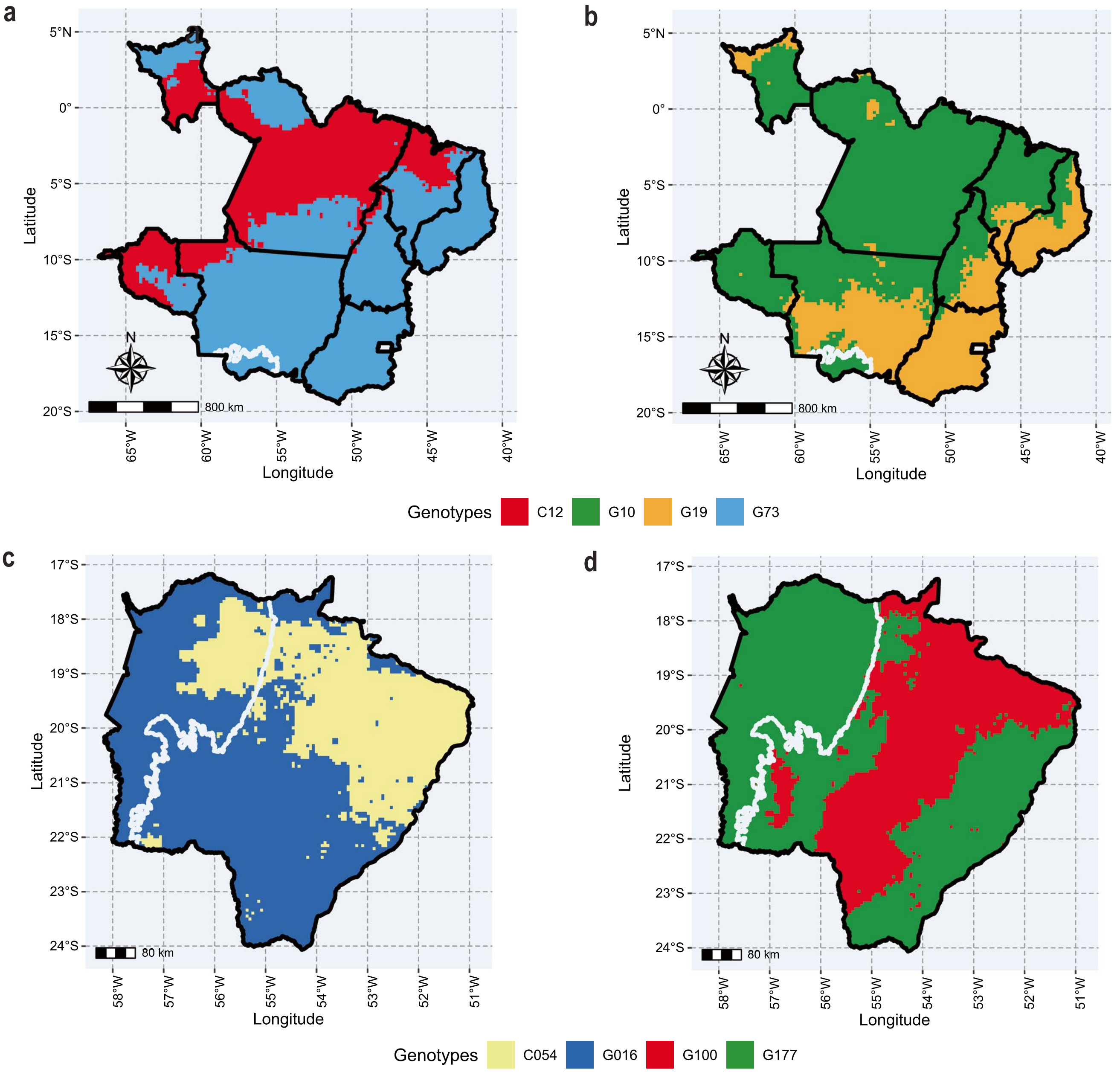

**Fig. 8** Pairwise comparison map showing the regions within the rice (**a, b**) and soybean (**c, d**) target populations of environments where a selection candidate outperforms a given peer. The colors across the map represent the winning genotype. **a, c** Are examples of pairwise comparisons between an experimental genotype and a commercial check, while **b, d** contrast the performance of two promising experimental genotypes along the breeding region. The white contour in **c** and **d** delimits the Pantanal biome

between relatives and employing more informative environmental features in the PLS model (Dias et al. 2018; Monteverde et al. 2019; Crossa et al. 2023). The utilization of ecophysiological environmental features in crop growth models could enhance our understanding of the link between phenotypic expression and environmental factors (Rincent et al. 2019; Costa-Neto et al. 2021a). GIS-FA can also be benchmarked with other enviromic-based approaches fit for predicting genotypes in untested environments (Jarquín et al. 2014; Tolhurst et al. 2022; Costa-Neto et al. 2020). Other statistical resources and even artificial intelligence methods can replace the PLS in the prediction step (Guo et al. 2021; Heinemann et al. 2022). Finally, future research can explore the potential risks associated with assigning genotypes to specific environments using GIS-FA. This can be done through the application of probabilistic methods (Dias et al. 2022).

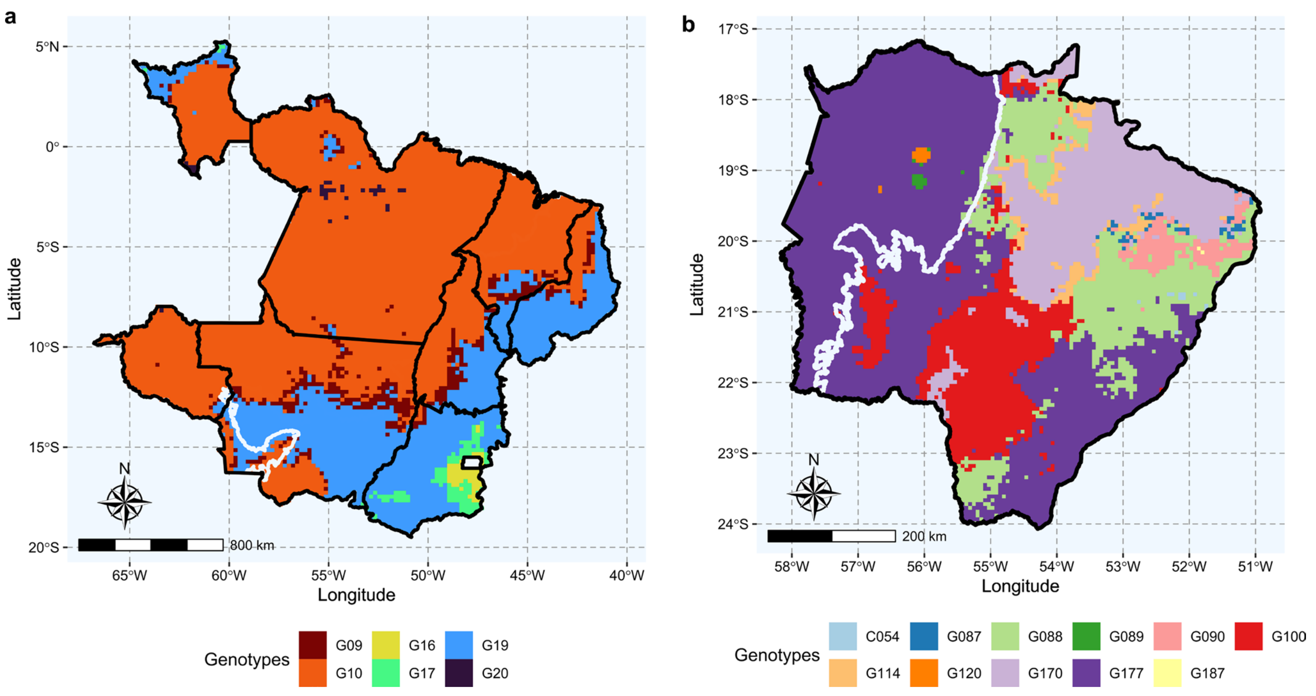

**Fig. 9** Which-won-where map depicting the most promising genotype at each location across the target population of environments of the rice dataset (**a**) and the soybean dataset (**b**) Each color represents the experimental genotype that wins in a specific environment within the breeding region. The white contour in **b** delimits the Pantanal biome

## Appendix A: Partial least squares regression

Here, we employed the kernel PLS algorithm (Lindgren et al. 1993; Dayal and MacGregor 1997) to predict the factor loadings of untested environments. Details about this algorithm are presented below:

Take the following multiple regressions as a starting point:

$$\hat{\boldsymbol{\Lambda}}^{\star} = \mathbf{WB} + \mathbf{E} \tag{A1}$$

where $\hat{\boldsymbol{\Lambda}}^{\star}$ is the $J \times K$ matrix of $K$ rotated loadings for the $J$ observed environments, $\mathbf{W}$ is a $J \times P$ matrix of scaled values for $P$ environmental features in the $J$ observed environments, $\mathbf{B}$ is a $P \times K$ vector of coefficients, and $\mathbf{E}$ is a $J \times K$ matrix of lack of fit effects. Note that most of the environmental features are correlated (Supplementary Figure 4), so $\mathbf{W}$ has multicollinearity problems, and $\mathbf{B} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\hat{\boldsymbol{\Lambda}}^{\star}$ does not yield a proper solution. To overcome this issue, we employed kernel PLS regression to transform $\mathbf{B}$ into $\mathbf{B}^{*}$, using the following equation:

$$\mathbf{B}^{\star} = \boldsymbol{\Phi}(\boldsymbol{\Theta}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Xi}' \tag{A2}$$

where $\boldsymbol{\Phi}$ is a $P \times C$ matrix of weights for $\mathbf{W}$ ($\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1\, \boldsymbol{\phi}_2 \dots \boldsymbol{\phi}_C\}$), with $C$ being the number of PLS components; $\boldsymbol{\Theta}$ is a matrix of loadings for $\mathbf{W}$ ($\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1\, \boldsymbol{\theta}_2 \dots \boldsymbol{\theta}_C\}$) and has the same dimension as $\boldsymbol{\Phi}$, and $\boldsymbol{\Xi}$ is a $K \times C$ matrix of weights for $\boldsymbol{\Lambda}$ ($\boldsymbol{\Xi} = \{\xi_1 \xi_2 \dots \xi_C\}$). We describe the CV procedure that defined the number of components ($c = 1, 2, \dots, C$) in section Spatial predictions in the breeding zone. $\boldsymbol{\Phi}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Xi}$ were defined using an iterative process that leveraged the kernel functions of $\mathbf{W}$ and $\boldsymbol{\Lambda}$. First, $\boldsymbol{\phi}_c$ is estimated as the eigenvector that is equivalent to the largest eigenvalue of the kernel $\mathbf{W}'\hat{\boldsymbol{\Lambda}}^{\star}\hat{\boldsymbol{\Lambda}}^{\star\prime}\mathbf{W}$. We used this vector to initialize an iterative process whose number of repetitions is equivalent to $C$. Let $\mathbf{R} = \boldsymbol{\Phi}(\boldsymbol{\Theta}'\boldsymbol{\Phi})^{-1}$, with $\mathbf{R} = \{\mathbf{r}_1\ \mathbf{r}_2\ \dots\ \mathbf{r}_C\}$. In the first iteration, $\mathbf{r}_1 = \boldsymbol{\phi}_1$. Subsequently, $\mathbf{r}_c = \boldsymbol{\phi}_c - \theta'_{c-1}\boldsymbol{\phi}_c\xi'_{c-1}$. On each iteration, $\boldsymbol{\theta}_c$ and $\xi_c$ are estimated as follows:

$$\theta_c = \frac{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})}{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})\mathbf{r}_c} \qquad \xi_c = \frac{\mathbf{r}'_c(\mathbf{W}'\hat{\boldsymbol{\Lambda}}^{\star})}{\mathbf{r}'_c(\mathbf{W}'\mathbf{W})\mathbf{r}_c} \tag{A3}$$

The solutions of these equations are stored in $\boldsymbol{\Theta}$ and $\boldsymbol{\Xi}$, respectively, and are used to update the covariance matrix for the next iteration as follows:

$$(\mathbf{W}'\hat{\boldsymbol{\Lambda}}^{\star})_{c+1} = (\mathbf{W}'\hat{\boldsymbol{\Lambda}}^{\star})_c - \theta_c\xi'_c[(\mathbf{W}\mathbf{r}_c)'\mathbf{W}\mathbf{r}_c] \tag{A4}$$

When the iteration process is finished, $\mathbf{B}^{*}$ provides a proper solution to Eq. (A1) and can be used for prediction purposes. We used $\mathbf{B}^{*}$ in Eq. (17) to train the PLS model and in Eq. (18) to make predictions.

**Acknowledgements** This work was supported by the Minas Gerais State Agency for Research and Development (FAPEMIG), the Brazilian National Council for Scientific and Technological Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), the Mato Grosso do Sul Foundation (Fundação MS), the Brazilian Agricultural Research Corporation (Embrapa Rice and Beans), and the Federal University of Viçosa (UFV).

**Author contribution statement** M.S.A., S.F.S.C., and K.O.G.D. conceived the research. M.S.A. and S.F.S.C. executed the statistical analyses and drafted the initial manuscript. M.D.K. and G.C.N. provided insights into the methodology. L.A.S.D., F.M.F., G.R.P., R.S.A., P.C.S.C., M.D.K., and G.C.N. provided critical revisions of the paper drafts. A.R.G.B. provided knowledge on the structure of the soybean dataset, while A.B.H. and F.B. provided information about the rice dataset. M.S.A., S.F.S.C., and M.D.K. built the tutorial available in the Supplementary Material. All authors approved the final version of the manuscript.

**Funding** This research was supported by the Minas Gerais State Agency for Research and Development (FAPEMIG), the Coordination for the Improvement of Higher Education Personnel (CAPES), and the Brazilian National Council for Scientific and Technological Development (CNPq). Fundação de Amparo à Pesquisa do Estado de Minas Gerais. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Conselho Nacional de Desenvolvimento Científico e Tecnológico.

**Data availability** The R codes and both datasets used in this study are freely available: https://github.com/Kaio-Olimpio/GIS-FA. Supplementary Material contains a detailed tutorial with a commented script describing the steps for performing GIS-FA analysis with the soybean dataset.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

## References

Aastveit AH, Martens H (1986) ANOVA interactions interpreted by partial least squares regression. Biometrics 42(4):829–844. https://doi.org/10.2307/2530697

Alvares CA, Stape JL, Sentelhas PC et al (2013) Köppen's climate classification map for Brazil. Meteorol Zeitschrift 22:711–728. https://doi.org/10.1127/0941-2948/2013/0507

Annicchiarico P, Bellah F, 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. https://doi.org/10.1016/j.eja.2005.05.003

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. J Stat Softw 75:1–6. https://doi.org/10.18637/jss.v075.b02

Balestre M, Von Pinho RG, Souza JC et al (2009) Genotypic stability and adaptability in tropical maize based on AMMI and GGE biplot analysis. Genet Mol Res 8(4):1311–1322. https://doi.org/10.4238/vol8-4gmr658

Beebe S, Lynch J, Galwey N et al (1997) A geographical approach to identify phosphorus-efficient genotypes among landraces and wild ancestors of common bean. Euphytica 95:325–338. https://doi.org/10.1023/A:1003008617829

Buntaran H, Forkman J, Piepho HP (2021) Projecting results of zoned multi-environment trials to new locations using environmental covariates with random coefficient models: accuracy and precision. Theor Appl Genet 134:1513–1530. https://doi.org/10.1007/s00122-021-03786-2

Bustos-Korts D, Boer MP, Layton J et al (2022) Identification of environment types and adaptation zones with self-organizing maps: applications to sunflower multi-environment data in europe. Theor Appl Genet 135:2059–2082. https://doi.org/10.1007/s00122-022-04098-9

CFSR (2018) Climate forecast system reanalysis (CFSR), for 1979 to 2011. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00765/

Chaves SFS, Alves RS, Dias LAS et al (2023) Analysis of repeated measures data through mixed models: an application in *Theobroma grandiflorum* breeding. Crop Sci 63(4):2131–2144. https://doi.org/10.1002/csc2.20995

Chaves SFS, Evangelista JSPC, Trindade RS et al (2023) Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. Crop Sci 63(3):1114–1125. https://doi.org/10.1002/csc2.20911

CHELSA (2023) Climatologies at high resolution for the earth's land surface areas. https://chelsa-climate.org/

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

Cooper M, Messina CD (2021) Can we harness "enviromics'' to accelerate crop improvement by integrating breeding and agronomy? Front Plant Sci 12(735):143. https://doi.org/10.3389/fpls.2021.735143

Cooper M, Messina CD, Podlich D et al (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci 65:311. https://doi.org/10.1071/CP14007

Cooper M, Messina CD, Tang T et al (2022) Predicting genotype × environment × management (G×E×M) interactions for the design of crop improvement strategies, pp 467–585. https://doi.org/10.1002/9781119874157.ch8

Costa-Neto G, Fritsche-Neto R (2021) Enviromics: bridging different sources of data, building one framework. Crop Breed Appl Biotechnol 21:e393,521S12. https://doi.org/10.1590/1984-70332021v21Sa25

Costa-Neto G, Morais Júnior OP, Heinemann AB et al (2020) A novel GIS-based tool to reveal spatial trends in reaction norm: upland rice case study. Euphytica 216:37. https://doi.org/10.1007/s10681-020-2573-4

Costa-Neto G, Crossa J, Fritsche-Neto R (2021a) Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. Front Plant Sci 12(717):552. https://doi.org/10.3389/fpls.2021.717552

Costa-Neto G, Fritsche-Neto R, Crossa J (2021b) Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. Heredity 126(1):92–106. https://doi.org/10.1038/s41437-020-00353-1

Costa-Neto G, Galli G, Carvalho HF et al (2021c) EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. G3 Genes|Genomes|Genetics 11(4):jkab040. https://doi.org/10.1093/g3journal/jkab040

Costa-Neto G, Crespo-Herrera L, Fradgley N et al (2022) Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data. G3: Genes|Genomes|Genetics 13(2):jkac313. https://doi.org/10.1093/g3journal/jkac313

Cowling WA, Castro-Urrea FA, Stefanova KT 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. https://doi.org/10.3390/plants12020383

Crossa J (2012) From genotype × environment interaction to gene × environment interaction. Curr Genom. 13(3):225–244. https://doi.org/10.2174/138920212800543066

Crossa J, Vargas M, Van Eeuwijk FA et al (1999) Interpreting genotype× environment interaction in tropical maize using linked molecular markers and environmental covariables. Theor Appl Genet 99:611–625. https://doi.org/10.1007/s001220051276

Crossa J, Yang RC, Cornelius PL (2004) Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. J Agric Biol Environ Stat 9(3):362–380. https://doi.org/10.1198/108571104x4423

Crossa J, Montesinos-López OA, Crespo Herrera LA et al (2023) Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy? Front Genet 14:7016. https://doi.org/10.3389/fgene.2023.1209275

Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11:381. https://doi.org/10.1198/108571106X154443

Cullis B, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome 53:1002–1016. https://doi.org/10.1139/G10-080

Cullis BR, Jefferson P, Thompson R et al (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. https://doi.org/10.1007/s00122-014-2373-0

Dayal BS, MacGregor JF (1997) Improved PLS algorithms. J Chemom 11(1):73–85

de los Campos G, Pérez-Rodréguez P, Bogard M et al (2020) A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nat Commun 11:4876. https://doi.org/10.1038/s41467-020-18480-y

Denis BJ (1988) Two way analysis using covariates. Statistics 19(1):123–132. https://doi.org/10.1080/02331888808802080

Dias KOG, Gezan SA, Guimarães CT et al (2018) Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity 121:24–37. https://doi.org/10.1038/s41437-018-0053-6

Dias KOG, Santos JPR, Krause MD et al (2022) Leveraging probability concepts for cultivar recommendation in multi-environment trials. Theor Appl Genet 135:1385–1399. https://doi.org/10.1007/s00122-022-04041-y

Diepenbrock CH, Tang T, Jines M et al (2022) Can we harness digital technologies and physiology to hasten genetic gain in us maize breeding? Plant Physiol 188(2):1141–1157. https://doi.org/10.1093/plphys/kiab527

Dunnington D (2023) ggspatial: spatial data framework for ggplot2. https://CRAN.R-project.org/package=ggspatial, r package version 1.1.8

Eberhart SA, Russell WA (1966) Stability parameters for comparing varieties. Crop Sci 6:36–40. https://doi.org/10.2135/cropsci1966.0011183X000600010011x

ECMWF (2023) European centre for medium-range weather forecasts. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ncdc:C00765/

EOSDIS (2023) Nasa earth observing system data and information system. https://worldview.earthdata.nasa.gov

FAO (2014) World reference base for soil resources 2014. www.fao.org/3/i3794en/I3794en.pdf

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 32:4302–4315. https://doi.org/10.1002/joc.5086

Finlay K, Wilkinson G (1963) The analysis of adaptation in a plant-breeding programme. Aust J Agric Res 14:742. https://pdf.usaid.gov/pdf_docs/PNAAS139.pdf

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

GHCNd (2023) Global historical climatology network daily. https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-daily/

Gilmour AR, Cullis B, Verbyla Ap (1997) Accounting for natural and extraneous variation in the analysis of field experiment. J Agric Biol Environ Stat 2:269–293. https://doi.org/10.2307/1400446

Gogel B, Smith A, 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. https://doi.org/10.1007/s10681-018-2116-4

Guarino L, Jarvis A, Hijmans RJ et al (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, pp 387–404

Guo Y, Xiang H, Li Z et al (2021) Prediction of rice yield in East China based on climate and agronomic traits data using artificial neural networks and partial least squares regression. Agronomy 11(2):282. https://doi.org/10.3390/agronomy11020282

Hartung J, Piepho HP (2021) Effect of missing values in multi-environmental trials on variance component estimates. Crop Sci 61(6):4087–4097. https://doi.org/10.1002/csc2.20621

Heinemann AB, Costa-Neto G, Fritsche-Neto R et al (2022) Enviromic prediction is useful to define the limits of climate adaptation: a case study of common bean in Brazil. Field Crop Res 286(108):628. https://doi.org/10.1016/j.fcr.2022.108628

Henderson CR (1949) Estimates of changes in herd environment. J Dairy Sci 61:294–300

Henderson CR (1950) Estimation of genetic parameters. Ann Math Stat 21:309–310

Hernández MV, Ortiz-Monasterio I, Pérez-Rodríguez P et al (2019) Modeling genotype × environment interaction using a factor analytic model of on-farm wheat trials in the Yaqui Valley of Mexico. Agron J 111(6):2647–2657. https://doi.org/10.2134/agronj2018.06.0361

Hijmans R (2020) raster: Geographic data analysis and modeling. R package version 3.6-3. https://CRAN.R-project.org/package=raster

Hijmans RJ, Barbosa M, Ghosh A et al (2023) geodata: Download geographic data. https://CRAN.R-project.org/package=geodata, r package version 0.5-8

Jarquún D, Crossa J, Lacaze X et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127(3):595–607. https://doi.org/10.1007/s00122-013-2243-1

Jarquún D, de Leon N, Romay C et al (2021) Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. Front Genet 11(592):769. https://doi.org/10.3389/fgene.2020.592769

Krause MD, Dias KOG, Singh AK et al (2022) Using large soybean historical data to study genotype by environment variation and identify mega-environments with the integration of genetic and non-genetic factors. bioRxiv 4:487885. https://doi.org/10.1101/2022.04.11.487885

Lembrechts JJ, van den Hoogen J, Aalto J et al (2022) Global maps of soil temperature. Glob Chang Biol 28(9):3110–3144. https://doi.org/10.1111/gcb.16060

Li X, Guo T, Mu Q et al (2018) Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc

Natl Acad Sci 115(26):6679–6684. https://doi.org/10.1073/pnas.1718326115

Liland KH, Mevik BH, Wehrens R (2022) PLS: partial least squares and principal component regression. https://CRAN.R-project.org/package=pls, r package version 2.8-1

Lindgren F, Geladi P, Wold S (1993) The kernel algorithm for PLS. J Chemom 7(1):45–59. https://doi.org/10.1002/cem.1180070104

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits, 1st edn. Sinauer Associates, Sunderland

Malosetti M, Ribaut JM, Eeuwijk FAV (2013) The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Genet Sel Evol 4:44. https://doi.org/10.3389/fphys.2013.00044

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(1819–1829):11290733

Millet EJ, Kruijer W, Coupel-Ledru A et al (2019) Genomic prediction of maize yield across European environmental conditions. Nat Genet 51(6):952–956. https://doi.org/10.1038/s41588-019-0414-y

Monteverde E, Gutierrez L, Blanco P et al (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

Montesinos-López OA, Montesinos-López A, Kismiantini, Roman-Gallardo A et al (2022a) Partial least squares enhances genomic prediction of new environments. Front Genet 13:920689.https://doi.org/10.3389/fgene.2022.920689848

Montesinos-López OA, Montesinos-López A, Sandoval DAB et al (2022b) Multi-trait genome prediction of new environments with partial least squares. Front Genet 13:966775. https://doi.org/10.3389/fgene.2022.966775851

Mrode RA (2014) Linear models for the prediction of animal breeding values, 3rd edn. CABI

NasaPower (2022) Prediction of worldwide energy resource. https://power.larc.nasa.gov/data-access-viewer

NOAA (2023) Climate data online. https://www.ncei.noaa.gov/cdo-web

Nuvunga JJ, Silva CP, Oliveira LA et al (2019) Bayesian factor analytic model: an approach in multiple environment trials. PLoS ONE 14(8):e0220290. https://doi.org/10.1371/journal.pone.0220290

Oliveira IC, Guilhen JHS, Ribeiro PCO et al (2020) Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. Field Crop Res 257(107):929. https://doi.org/10.1016/j.fcr.2020.107929

Ortiz R, Crossa J, Vargas M et al (2007) Studying the effect of environmental variables on the genotype × environment interaction of tomato. Euphytica 153:119–134. https://doi.org/10.1007/s10681-006-9248-7

Ortiz R, Reslow F, Montesinos-López A et al (2023) Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. Sci Rep 13(1):9947. https://doi.org/10.1038/s41598-023-37169-y

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58:545–554. https://doi.org/10.2307/2334389

Pebesma E, Bivand R (2023) Spatial data science: with applications in R. https://r-spatial.org/book/

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

Piepho HP (2019) A coefficient of determination ($r^2$) for generalized linear mixed models. Biom J 61(4):860–872. https://doi.org/10.1002/bimj.201800270

Piepho H, Möhring J (2006) Selection in cultivar trials–is it ignorable? Crop Sci 46(1):192–201. https://doi.org/10.2135/cropsci2005.04-0038

Piepho HP, Möhring J, Melchinger AE et al (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228. https://doi.org/10.1007/s10681-007-9449-8

Porker K, Coventry S, Fettell N et al (2020) Using a novel PLS approach for envirotyping of barley phenology and adaptation. Field Crop Res 246(107):697. https://doi.org/10.1016/j.fcr.2019.107697

R Core Team (2023) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramburan S, Zhou M, Labuschagne M (2012) Integrating empirical and analytical approaches to investigate genotype-environment interactions in sugarcane. Crop Sci 52(5):2153–2165. https://doi.org/10.2135/cropsci2012.02.0128

Resende RT, Piepho HP, Rosa GJM et al (2021) Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor Appl Genet 134:95–121. https://doi.org/10.1007/s00122-020-03684-z

Rincent R, Malosetti M, Ababaei B et al (2019) Using crop growth model stress covariates and AMMI decomposition to better predict genotype-by-environment interactions. Theor Appl Genet 132(12):3399–3411. https://doi.org/10.1007/s00122-019-03432-y

Rogers AR, Dunne JC, Romay C et al (2021) The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. G3: Genes|Genomes|Genetics 11(2):jkaa050. https://doi.org/10.1093/g3journal/jkaa050

Sae-Lim P, Komen H, Kause A et al (2014) Identifying environmental variables explaining genotype-by-environment interaction for body weight of rainbow trout (Onchorynchus mykiss): reaction norm and factor analytic models. Genet Sel Evol 46(16):1–11. https://doi.org/10.1186/1297-9686-46-16

Santos HG (2018) Sistema brasileiro de classificação de solos (in Portuguese), 5th edn. Embrapa, Brasília, DF. https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1094003/sistema-brasileiro-de-classificacao-de-solos

Shelford VE (1911) Animal communities in temperate America as illustrated in the Chicago region. Biol Bull 21:95–167. https://doi.org/10.5962/bhl.title.34437

Silva KJ, Teodoro PE, da Silva MJ et al (2021) Identification of mega-environments for grain sorghum in Brazil using GGE biplot methodology. Agron J 113:1–12. https://doi.org/10.1002/agj2.20707

Smith AB, Cullis BR (2018) Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. Euphytica 214:143. https://doi.org/10.1007/s10681-018-2220-5

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

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

Smith A, Norman A, Kuchel H et al (2021) Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. Front Plant Sci 12(978):248. https://doi.org/10.3389/fpls.2021.737462

SoilGrids (2022) Soilgrids—global gridded soil information. https://www.isric.org/explore/soilgrids/

Sparks AH (2018) NasaPower: a NASA power global meteorology, surface solar energy and climatology data client for R. J Open Source Softw 3(30):1035. https://doi.org/10.21105/joss.01035

Stefanova KT, Buirchell B (2010) Multiplicative mixed models for genetic gain assessment in lupin breeding. Crop Sci 50(3):880–891. https://doi.org/10.2135/cropsci2009.07.0402

The VSNi Team (2023) asreml: Fits linear mixed models using REML. www.vsni.co.uk, r package version 4.2.0.267

Thompson R, Cullis B, Smith A et al (2003) A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. Aust N Z J Stat 45(4):445–459. https://doi.org/10.1111/1467-842X.00297

Tolhurst DJ, Gaynor RC, Gardunia B et al (2022) Genomic selection using random regressions on known and latent environmental covariates. Theor Appl Genet 135:3393–3415. https://doi.org/10.1007/s00122-022-04186-w

Van Eeuwijk FA, Elgersma A (1993) Incorporating environmental information in an analysis of genotype by environment interaction for seed yield in perennial ryegrass. Heredity 70(5):447–457. https://doi.org/10.1038/hdy.1993.66

van Eeuwijk FA, Bustos-Korts DV, Malosetti M (2016) What should students in plant breeding know about the statistical aspects of genotype × environment interactions? Crop Sci 56(5):2119–2140. https://doi.org/10.2135/cropsci2015.06.0375

Vargas M, Crossa J, Van Eeuwijk F et al (2001) Interpreting treatment-environment interaction in agronomy trials. Agron J 93(4):949–960. https://doi.org/10.2134/agronj2001.934949x

Vargas M, van Eeuwijk FA, Crossa J et al (2006) Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. Theor Appl Genet 112(6):1009–1023. https://doi.org/10.1007/s00122-005-0204-z

Wickham H (2016) ggplot2: elegant graphics for data analysis, 2nd edn. Springer, Cham

Wold HOA (1966) Estimation of principal components and related models by iterative least squares. Academic Press, New York

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58:109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wong J (2022) Pdist: partitioned distance function. https://CRAN.R-project.org/package=pdist, r package version 1.2.1

Wood J (1976) The use of environmental variables in the interpretation of genotype–environment interaction. Heredity 37(1):1–7. www.nature.com/articles/hdy197661

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

Yan W, Hunt LA, Sheng Q et al (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. Crop Sci 40:597–605. https://doi.org/10.2135/cropsci2000.403597x

Yan W, Kang MS, Ma B et al (2007) GGE biplot vs. AMMI analysis of genotype-by-environment data. Crop Sci 47:643–653. https://doi.org/10.2135/cropsci2006.06.0374

Yates F, Cochran WG (1938) The analysis of groups of experiments. J Agric Sci 28:556–580. https://doi.org/10.1017/S0021859600050978

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.