scientific reports

Check for updates

OPEN

Enhancing enviromics based predictions in common bean multi-environment trials

Gabriel M. Blasques[1], Luiz A. S. Dias[1], Mauricio S. Araújo[2], Alisson F. Chiorato[3], Sergio A. M. Carbonell[3], Lucas P. Corrêdo[1], Saulo Chaves[2]✉ & Kaio O. G. Dias[4,5]✉

Enviromic approaches enhance predictive models by incorporating environmental data into selection frameworks. By integrating factor analytic (FA) models, enviromics, and Geographic Information Systems (GIS), the GIS-FA method was proposed to improve genotype prediction in untested environments. This study aims to refine GIS-FA by implementing Random Forest Spatial Interpolation for enhanced environmental data interpolation and optimizing spatial sampling to exclude non-agricultural areas, thereby improving environmental characterization. We applied the improved GIS-FA framework to common bean trials conducted across 23 environments in São Paulo, Brazil, evaluating 59 genotypes from the "Carioca" and "Black" market classes. The enhanced method increased empirical Best Linear Unbiased Predictions (eBLUPs) accuracy from 0.46 to 0.53 in leave-one-out cross-validation, representing a 15.2% improvement and enabling more reliable genotype performance predictions. Additionally, integrating GIS-FA with Factor Analytic Selection Tools improved the interpretation of stability and adaptability metrics by allowing predictions at untested locations. This provided a comprehensive spatial view of genotype performance across the entire Target Population of Environments (TPE). High-resolution thematic maps generated through GIS-FA facilitated genotype recommendation across São Paulo. These findings demonstrate the value of incorporating machine learning-based interpolation and spatial optimization into GIS-FA, reinforcing its potential to support selection strategies and advance environment-informed prediction in modern plant breeding.

Keywords Multi-environment trials, Genotype-by-environment interaction, Factor analytic, GIS-FA method, Variety recommendation

Common bean (*Phaseolus vulgaris* L.) plays a fundamental role in both the Brazilian diet and agriculture, with the "Carioca" and "Black" types accounting for approximately 90% of total production and 78% of national consumption[1]. Its cultivation spans the entire Brazilian territory, demonstrating remarkable adaptability to diverse edaphoclimatic conditions and seasonal variations. This adaptation has led to the establishment of three main growing seasons: (i) the rainy season, (ii) the dry season, and (iii) the winter season[2]. Such a broad geographic and seasonal distribution ensures a continuous and diversified supply, which is crucial for meeting the country's dietary demands.

Given the wide range of environmental conditions to which the crop is exposed, multi-environment trials (MET) are implemented to evaluate genotypes across a series of environments representative of the recommendation regions. Assessing different genotypes in MET reveals a key factor for statistical analysis: the genotype by environment (G×E) interaction[3]. This phenomenon is crucial for understanding how genotypes adapt to varying environmental conditions, playing a pivotal role in effective selection and recommendation processes[4].

For analysis of MET data, mixed models offer significant advantages due to their robustness in handling unbalanced data, their flexibility in accommodating both fixed and random effects, and their capacity for more coherent co-variance modeling[5]. Given the frequent high crossover G×E interaction in these trials, the unstructured model would be the most appropriate choice for modeling the variance-covariance matrix. However,

[1]Department of Agronomy, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil. [2]Department of Genetics, Luiz de Queiroz College of Agriculture, University of São Paulo, Piracicaba, São Paulo, Brazil. [3]Agronomic Institute (IAC), Campinas, São Paulo, Brazil. [4]Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil. [5]Institute of Artificial and Computational Intelligence (IDATA), Federal University of Viçosa, Viçosa, Minas Gerais, Brazil. ✉email: saulochaves@usp.br; kaio.o.dias@ufv.br

this approach substantially increases the number of parameters to be estimated, potentially compromising model parsimony and reducing the precision of parameter estimation[6].

Recognized as the gold standard for modeling G×E interactions, factor analytic (FA) mixed models[7,8] represent genotypic effects as random regressions on latent environmental covariates (loadings), with each genotype associated with a corresponding slope (score)[9]. Their prominent role in MET analysis arises from their ability to reduce the dimensionality of data by extracting latent factors, allowing for an approximation of unstructured variance-covariance modeling in a parsimonious manner[10,11]. Although FA models were initially underutilized due to limitations in the interpretability of their results, their adoption increased with the contributions of Smith and Cullis[12], who proposed the Factor Analytic Selection Tools (FAST). These tools provide breeders with performance and stability metrics, making FA more suitable for the selection and recommendation of varieties.

Enviromics has emerged as a transformative approach in G×E data analyses, enhancing MET studies by addressing the environmental component with greater precision and depth. Unlike traditional methods that treat the environment as a "black box", enviromics integrates detailed environmental data from tools such as remote sensing, high throughput platforms, and big data analytics[13–17]. This enables the characterization of environmental factors as structured, measurable variables in predictive models. Through the combination of environmental data with genomic and phenotypic information, enviromics improves the modeling of G×E interactions, resulting in more accurate genotype performance predictions across different environments[18–20].

By integrating FA with enviromics through partial least squares (PLS) regression, Araújo et al.[16] proposed the GIS-FA methodology. This approach leverages meteorological, soil, and topographic data obtained via Geographic Information System (GIS) to predict the factor loadings of FA models for each environment using PLS. From these predicted loadings, the empirical best linear unbiased predictions (eBLUPs) of genotypes can be estimated in untested environments by linearly combining the predicted loadings with the genotypic scores of FA models. Furthermore, this methodology makes it possible to create thematic maps, providing a deeper understanding of variety performance across environments within the target population of environments (TPE), therefore improving variety recommendation.

In predictive modeling using environmental features, the characterization, representativeness, sampling, and spatio-temporal resolution of environmental data are critical to achieve accurate and reliable results[21–23]. Within this framework, machine learning methodologies, such as Random Forest Spatial Interpolation (RFSI), have demonstrated superior performance over traditional interpolation techniques, including inverse distance weighting (IDW) and kriging (with or without external drift), by integrating random forest predictions with information from neighboring points and their distances, this approach enhances spatial accuracy and improves the reliability of spatialized environmental data[24–26]. Originally, the GIS-FA methodology employs random sampling of points within the TPE to extract environmental information and utilizes the IDW method for interpolation at a low spatial resolution, highlighting key areas for improvement.

To improve GIS-FA predictions, this study aims to refine the sampling by ensuring an evenly spaced distribution of points while excluding rivers, roads, and urban areas, improve interpolation using RFSI, and utilize the highest available resolution from environmental databases. Furthermore, we seek to optimize variety recommendation by integrating recommendation maps with FAST for common bean trials conducted by the Common Bean Breeding Program of the Agronomic Institute (PMGF-IAC) in the state of São Paulo, Brazil.

Results

From the single trial analyses of the experiments, the coefficients of variation (CV_a) ranged from 6.31% (Moc18R) to 24.9% (Cam21D) and the heritability (H_a^2), from 0.36 (Vot20W) to 0.97 (Moc20D) (Fig. 1). This heritability metric reflects the accuracy of genotype ranking based on BLUPs, with values approaching 1 indicating high precision in distinguishing genotypic differences and lower values reflecting limited discriminatory power due to stronger residual influence. All evaluated environments exhibited significant genotypic effects according to the likelihood ratio test. In the MET analysis, the FA4 model showed the best fit and was used for subsequent predictions, as indicated by the adopted selection criteria (Table 1).

Genetic correlations between environments ranged from −0.59 (Tat20R vs Mon19R) to 0.995 (Vot19W vs Cam18R), with an average correlation of 0.35. This mean value indicates a general trend toward positive correlations, which accounted for 87.35% of the total estimates. The environments Moc19D, Cam18W, Vot19W, and Cam18R exhibited the highest similarity when compared to all others, with mean correlations ranging from 0.57 to 0.58. Conversely, we identified Tat20R, Moc20D, Cap19R, and Cap19D as the most divergent, with mean correlations ranging from −0.04 to 0.14. High positive genetic correlations between environments indicate minimal crossover G×E interactions, whereas correlations close to zero or highly negative reflect, respectively, the presence and increasing intensity of these interactions (Fig. 2).

We identified the top 15% most productive and stable genotypes using the FAST methodology in the MET analysis for "Carioca" and "Black" bean types (Fig. 3), along with their respective reliabilities. To ensure an unbiased evaluation of new genotypes, we excluded the check varieties (IAC 2051 and IAC Veloz) from the selection percentage. For the "Carioca" type, we selected G16, G12, G39, G32, G45, and G31, while for the "Black" type, we selected G19, G30, and G28. Among them, G39, G45, G32, G31, and G19 exhibited greater stability compared to the check varieties. However, none of the selected genotypes outperformed the checks in terms of absolute performance.

In this study, the improved GIS-FA method reached an accuracy of 0.53 and RMSE of 698.57 under LOOCV. These results represent an improvement over both the original GIS-FA implementation and its performance on the current dataset (Table 2). The eBLUPs predicted from this model were subsequently used to generate thematic maps, offering a spatial representation of genotype performance across the TPE.

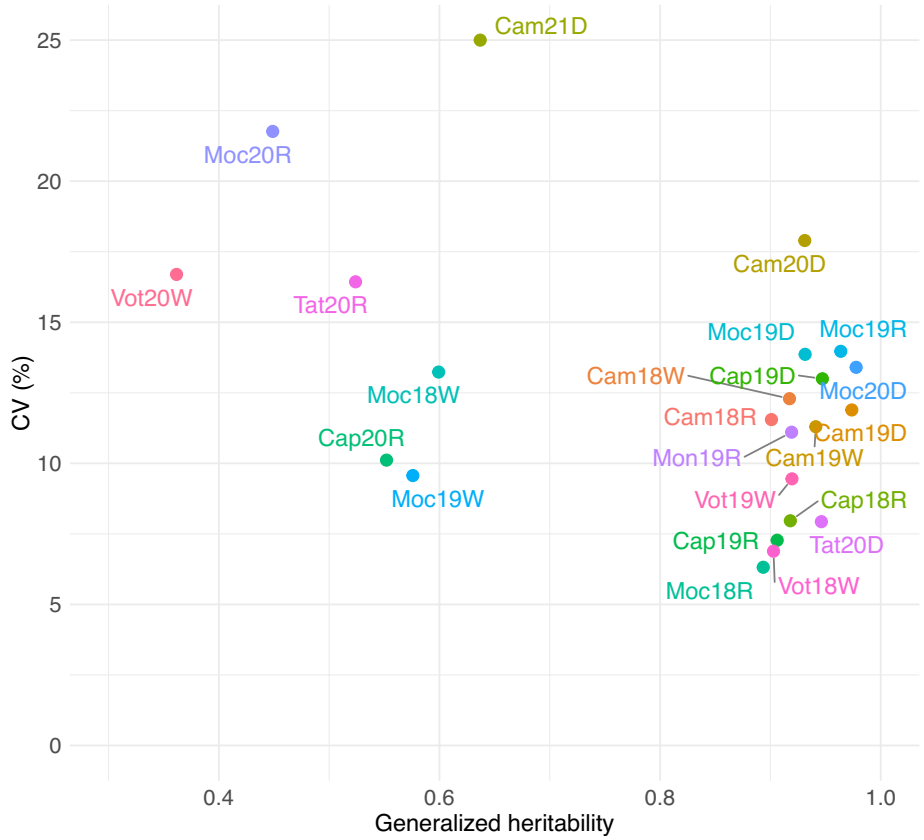


Fig. 1. Scatter plot of the experimental coefficient of variation (CV_a) and generalized heritability (H_a^2) for each of the 23 environments. Environment labels indicate the location (first three letters), year (last two digits), and season ("R" for rainy, "D" for dry, "W" for winter).

Model	LogL	AIC	ASR
FA1	−13644.89	27417.79	28.03
FA2	−13623.26	27416.52	40.23
FA3	−13603.39	27408.78	51.26
FA4	**−13590.56**	**27417.12**	**71.20**
FA5	−13579.82	27429.63	74.00

Table 1. Values related to the log-likelihood (LogL), Akaike Information Criterion (AIC), and average semi-variance ratio (ASR) for the adjusted models. Selected model is highlighted in **bold**. Models with more than five factors did not converged.

Six "Carioca" genotypes were the most suitable for specific environments in the TPE. However, only four (IAC 2051, G16, G42, and G02) covered large geographic regions, while G22 and G12, despite excelling in some environments, had limited spatial representation (Fig. 4). For "Black" genotypes, we observed less variability among the winning genotypes, with IAC Veloz and G28 dominating most of the TPE, while G46 and G30 being the best for limited areas (Fig. 5).

Through the adaptation zone analysis, we categorized the eBLUPs of genotypes and generated individual maps to assess their specific performance across the TPE. We constructed these thematic maps only for the most promising genotypes identified by FAST and the Which-won-where map. They showed that the "Carioca" genotypes IAC 2051 and G16 exhibited high productivity across large areas of the TPE, while the less stable G02 and G42 achieved high productivity only for specific regions, namely northwest and southeast, respectively (Fig. 6). For the "Black" genotypes, only G28 demonstrated high productivity across a significant portion of the TPE, while IAC Veloz and G30 performed well in more restricted areas. Additionally, G19 stood out for its high stability (Fig. 7).

We mapped the geographic winning distribution between two genotypes by comparing their predicted eBLUPs across the TPE using the pairwise comparison thematic map (Fig. 8). For the "Carioca" type, we compared the most promising genotypes with the commercial variety IAC 2051. G02 and G16 outperformed

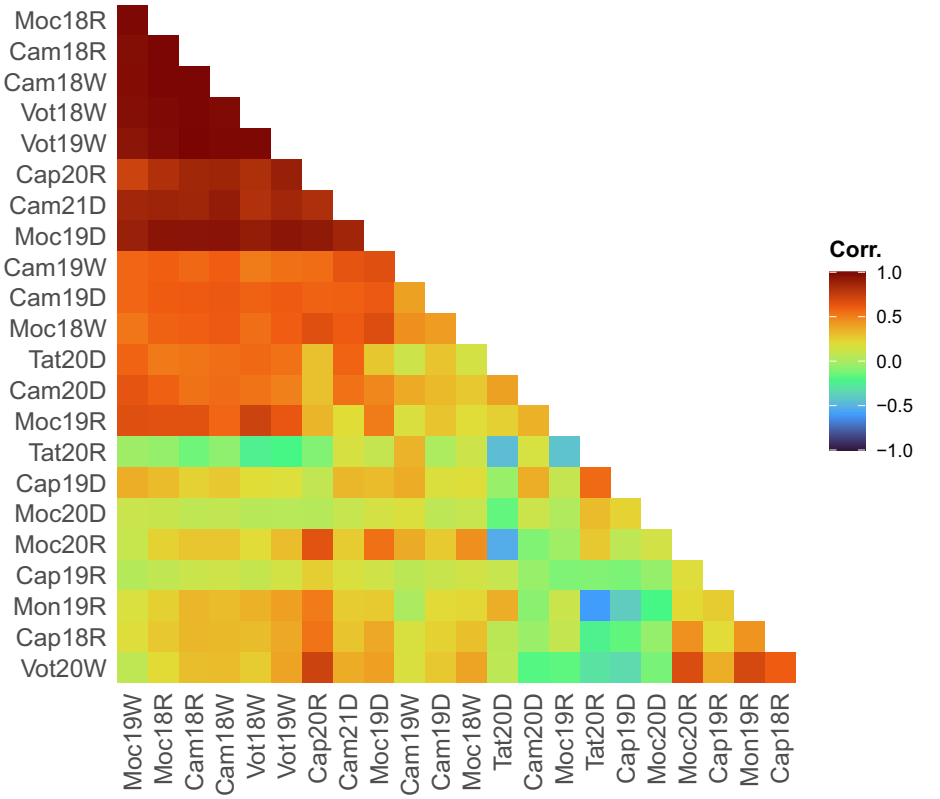


Fig. 2. Heatmap representation of genetic correlations among the 23 environments evaluated in the Multi Environment Trials (MET). Darker red and blue colors indicate strong positive or negative correlations, respectively. Conversely, darker green color represent weaker correlations.

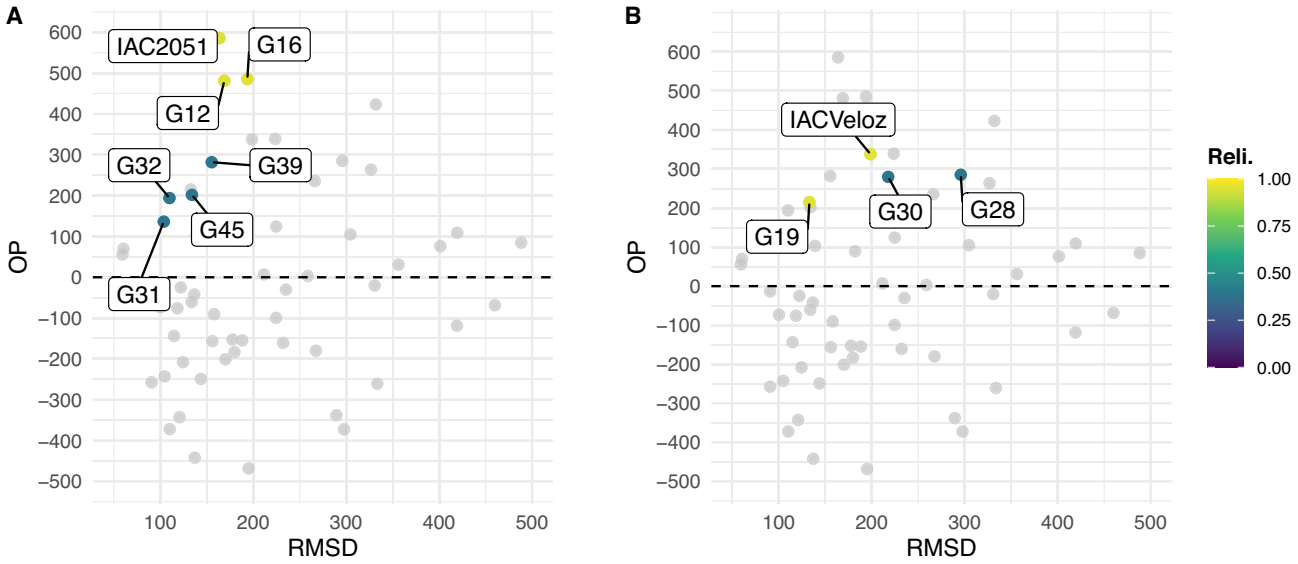


Fig. 3. Scatter plot of overall performance (OP) and stability (RMSD), illustrating the relationship between these two selection criteria across the evaluated lines. The most productive and stable genotypes were identified based on the selection index described in the Material and Methods section and are highlighted. Panel A highlights the selection of "Carioca" lines, while Panel B highlights the selection of "Black" lines. Genotypes positioned in the upper left region of the plot are characterized by high OP and low RMSD, indicating superior adaptability and stability.

Model	Prediction	Prediction accuracy	
		Common Beans	Soybean
GIS-FA	Loadings	0.38	0.34
GIS-FA*	Loadings	0.43	0.35
GIS-FA	eBLUPs	0.46	0.60
GIS-FA*	eBLUPs	0.53	0.63

Table 2. Prediction accuracy of Factor Loadings and eBLUPs using the original method (GIS-FA) and the improved method (GIS-FA*).

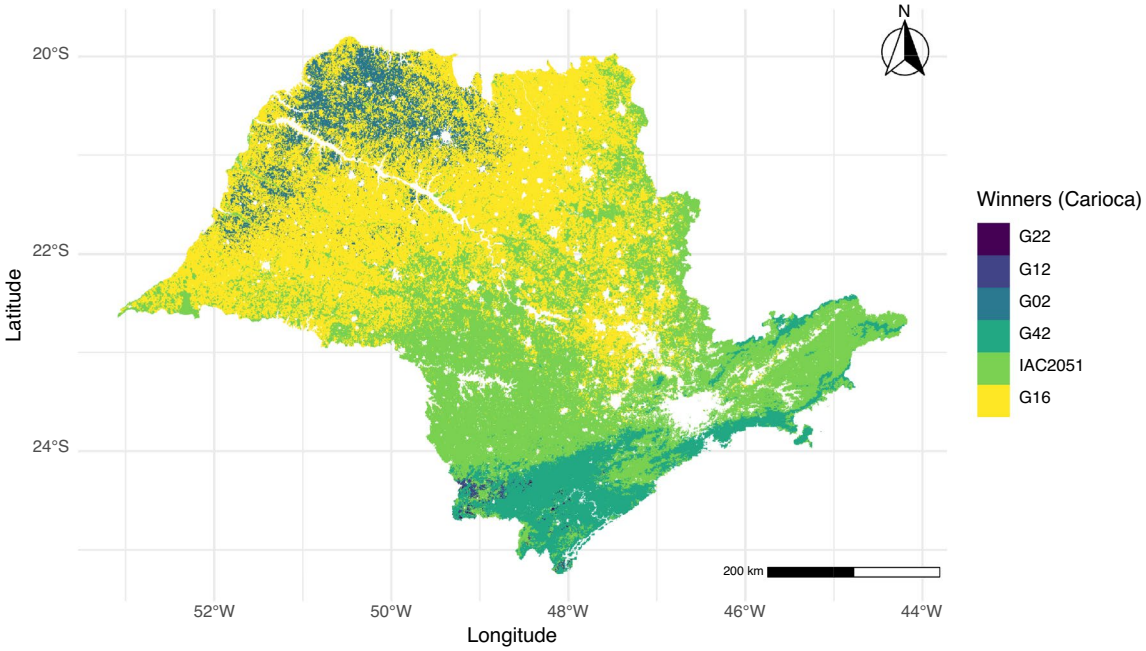


Fig. 4. Which-won-where map representing, through color coding, the best "Carioca" type genotype at each pixel (environment) across the target population of environments (São Paulo state).

the check in the northwest region of the TPE, while G42 surpassed it in the southeast region (Fig. 8a). For the "Black" genotypes, we made comparisons against the commercial variety IAC Veloz, which consistently outperformed all competitors in the southern and southeastern regions of the TPE. G30 and G28 had superior performance in the northern, northwestern, and western regions, while G19 showed superior performance in the northwest (Fig. 8b).

The environmental dissimilarity map (Supplementary Fig. 2) revealed a clear spatial gradient in representativeness of the MET conditions across São Paulo. Central and northwestern regions showed low dissimilarity values, indicating high similarity with the experimental environments and, consequently, greater confidence in extrapolated predictions. In contrast, southeastern and coastal areas displayed higher dissimilarity, reflecting environmental conditions poorly represented in the MET dataset. These regions are therefore associated with lower reliability of genotype predictions. While this map provides a useful spatial perspective for contextualizing the reliability of eBLUP-based recommendations, it should be interpreted as an exploratory tool rather than a definitive guide, highlighting areas where further testing or data collection would be particularly valuable.

Discussion

This study leveraged FA models and FAST tools in MET evaluations to assess common bean genotypes under diverse climatic conditions across the three traditional growing seasons. Additionally, we applied and improved the GIS-FA methodology by enhancing spatial sampling to ensure an evenly distributed set of points while excluding rivers, roads, and urban areas. We also improved interpolation step by implementing Random Forest Spatial Interpolation and utilized the highest available resolution from environmental databases. Furthermore, we built thematic maps for adaptation zones, pairwise comparisons, and which-won-where analyses, facilitating a spatially-explicit understanding of genotypic performance and stability. By integrating these approaches, we provided critical insights for enhancing selection strategies and improving varietal recommendations in response to environmental complexity, reinforcing the role of data-driven selection in modern plant breeding.

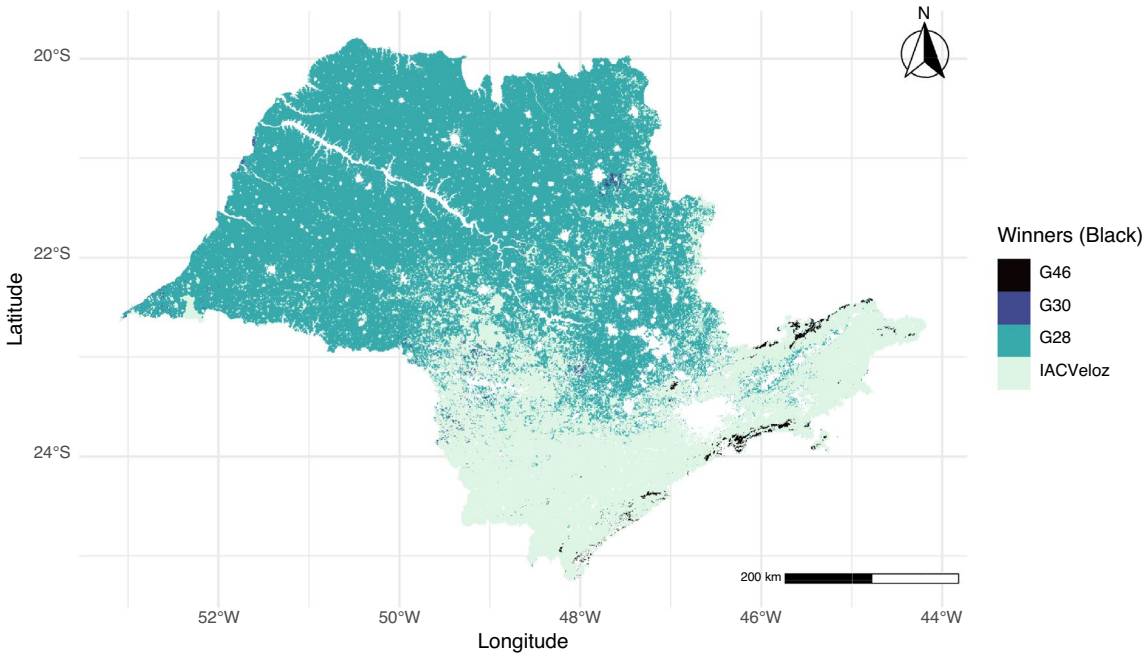


Fig. 5. Which-won-where map representing, through color coding, the best "Black" type genotype at each pixel (environment) across the target population of environments (São Paulo state).

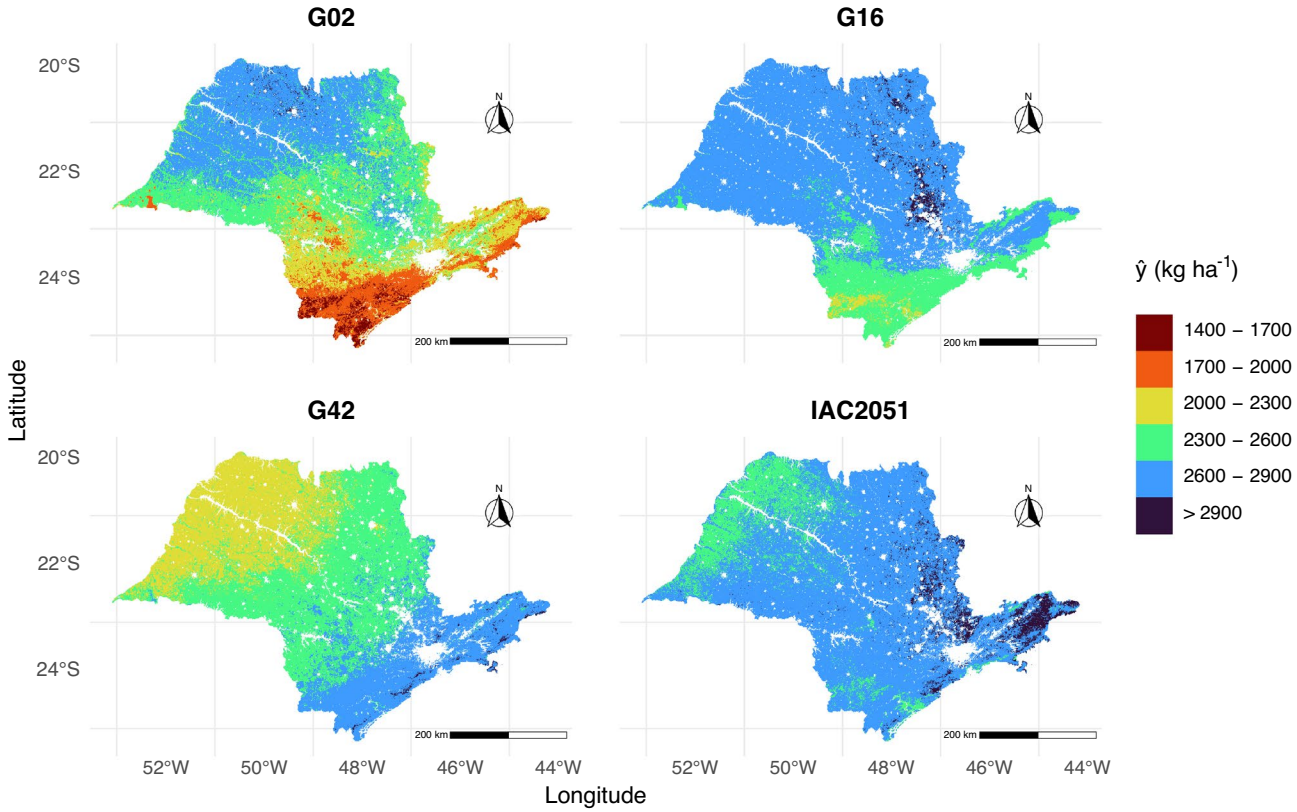


Fig. 6. Adaptation zones map illustrating the expected performance of the "Carioca" genotypes G02, G42, and G16, along with the commercial variety IAC 2051, based on yield classes. The color scale represents predicted yield categories, where darker red shades indicate lower yielding areas, while darker blue shades correspond to high yielding regions.

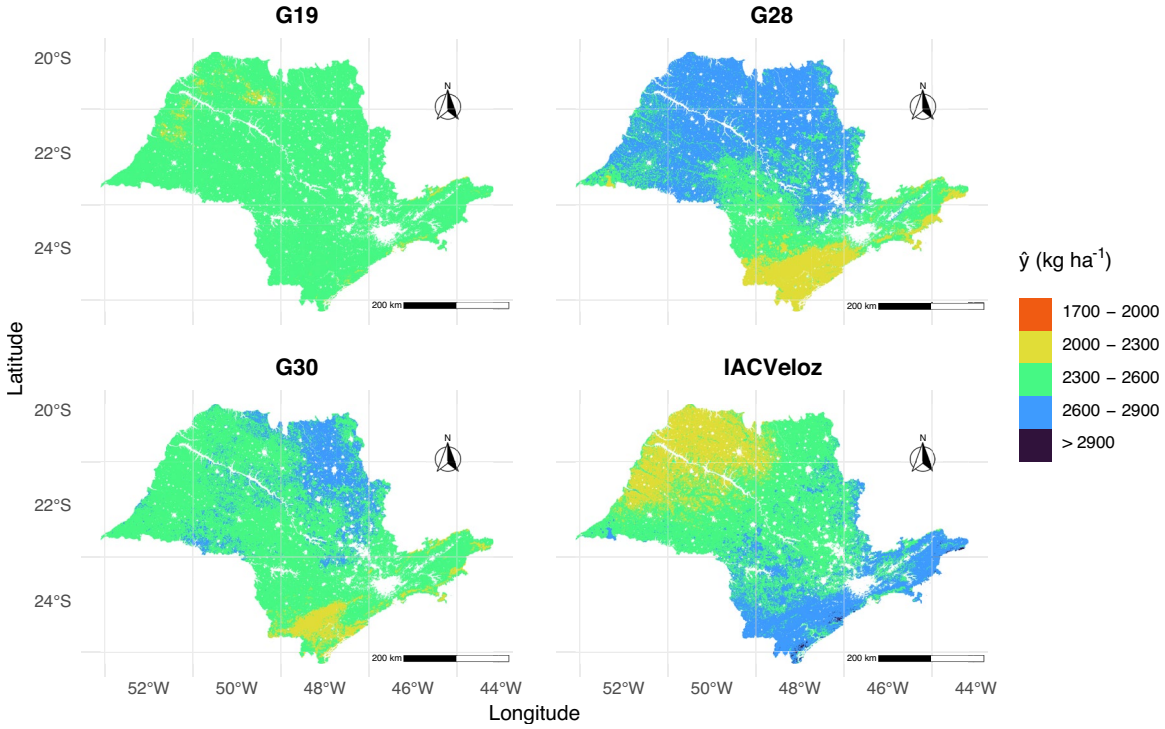


Fig. 7. Adaptation zones map illustrating the expected performance of the "Black" genotypes G19, G28, and G30, along with the commercial variety IAC Veloz, based on yield classes. The color scale represents predicted yield categories, where darker red shades indicate lower yielding areas, while darker blue shades correspond to high yielding regions.

These methodological advancements build upon the long-standing efforts of the Common Bean Breeding Program of the Agronomic Institute (PMGF-IAC), the oldest common bean breeding program in Brazil, established in 1932. Over nearly a century, PMGF-IAC has played a pivotal role in developing and releasing high-yielding, stable, and disease-resistant varieties, significantly contributing to common bean production in Brazil. Initially focused on mass selection of landraces, the program has evolved through the adoption of genealogical and backcrossing selection methods. Today, PMGF-IAC integrates cutting-edge breeding methodologies, including genomic selection and enviromics, to enhance performance and resilience. To date, the program has released over 50 varieties, each tailored to diverse edaphoclimatic conditions and market demands[27].

The FA model has proven to be a robust and parsimonious approach for modeling G×E in MET, effectively capturing the underlying covariance structures and enabling more precise selection decisions in a routine breeding program[4,28–31]. Its integration into modern plant breeding pipelines and the incorporation of environmental covariates allows for a deeper understanding of the environmental factors influencing genotype performance, ultimately refining selection strategies[32]. By integrating FA modeling with environmental covariates via PLS regression, the GIS-FA methodology demonstrated its potential to improve genotype predictions across untested environments, outperforming similar predictive approaches[16]. However, the effectiveness of enviromic predictions relies on the representativeness of sampled environments within METs concerning the TPE, as well as the accuracy of extracted environmental data[13,33].

The integration of GIS-FA methodology with FAST notably enhanced the interpretation of stability and performance metrics by enabling predictions in untested locations. Despite both methodologies identifying superior genotypes, distinct differences emerged regarding the geographic performance and stability of selected genotypes. FAST highlighted the most productive and stable genotypes independently of geographic distribution and environmental information. In contrast, the GIS-FA thematic maps illustrated spatial variations, revealing genotype specific adaptability across distinct environments. For instance, while FAST selected G16, G12, G32, G39, G45, G31, G30, G19, and G28, GIS-FA analyses identified alternative genotypes (G22, G42, and G02 for "Carioca" and G46 for "Black") better adapted to particular geographic regions. Notably, genotypes such as G16, G12, G28, and G30 were consistently selected across both FAST and GIS-FA analyses, indicating their robustness and broad adaptability. Thus, combining FAST with GIS-FA provided complementary insights, balancing the MET genotype performance indicators with spatially explicit adaptability assessments essential for optimized genotype recommendations.

The reliability of each genotype (Supplementary Table 1) provides further insight into the consistency of genotype recommendations. Genotypes evaluated more frequently conditions show higher reliability, indicating greater confidence in their predicted values. However, there is marked variability in reliability across genotypes, reflecting differences in their testing frequency and environmental representativeness. This variability implies that while some genotypes can be recommended with greater confidence, others require more cautious

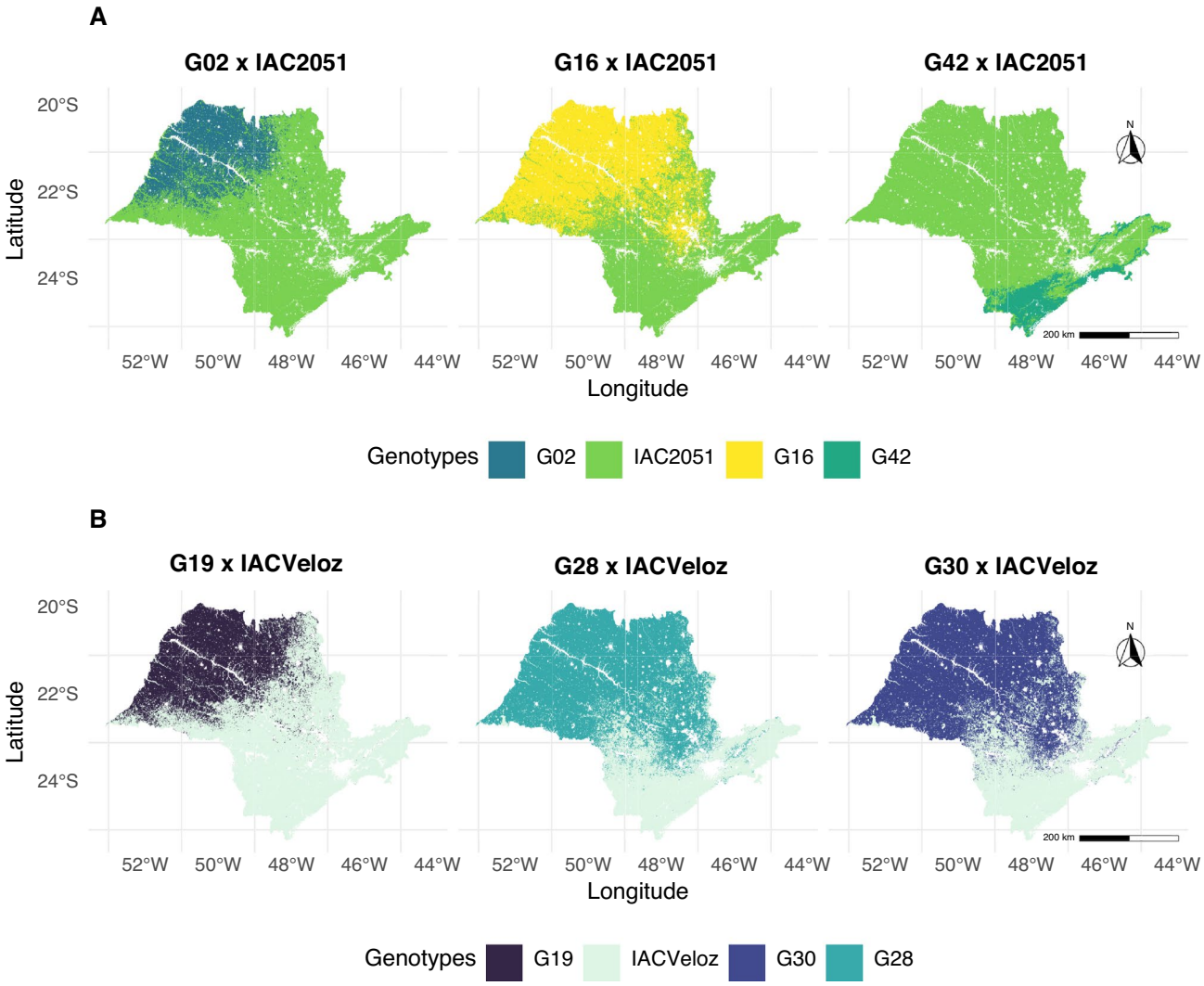


Fig. 8. Pairwise comparison map illustrating the regions within the target population of environments where a selected candidate outperforms a given competitor for "Carioca" (**A**), and "Black" (**B**) type genotypes.

interpretation, especially when evaluated in fewer or unique environments. This pattern reinforces the importance of representativeness in METs and suggests that, in practice, genotype recommendations should account not only for predicted values but also for their associated reliability[32]. While reliability does not capture prediction accuracy, it is a useful indicator of confidence, particularly for genotypes tested in limited or environmentally unique conditions. These considerations were added to help guide interpretation of the predictions and inform decisions about where extrapolation should be made with caution.

Although the geographic distribution of MET sites in this study was limited, the temporal variation introduced by evaluating genotypes across multiple years and three distinct cropping seasons provided a good representation of genotype responses under diverse environmental conditions. Furthermore, improvements in environmental variable extraction and the implementation of RFSI enhanced the characterization of environmental conditions, addressing challenges related to spatial resolution disparities in raw data (Supplementary Table 5). By ensuring a more reliable environmental dataset, the eBLUP predictions became more accurate, yielding an average accuracy gain of 15.2% over the previous methodology (from 0.46 to 0.53).

Despite these advancements, challenges remain in fully capturing the complexity of G×E, particularly under highly variable climatic conditions. This is evident when comparing the mean genetic correlations of an environment with the accuracies obtained for within-environment predictions in the LOOCV. The results highlight that highly divergent or underrepresented environments within the experimental network lead to lower predictive accuracy (Fig. 1, Supplementary Table 2). Although temporal diversity is partially accounted for by considering multiple growing seasons, the limited geographical distribution restricts broad generalizations across the entire TPE. This constraint highlights the exploratory nature of the thematic maps generated in this work, emphasizing the need for cautious interpretation when extrapolating results to environmentally dissimilar regions. Given that poorly represented environmental conditions tend to yield lower prediction accuracy, the inclusion of a spatial map of environmental dissimilarity across the TPE (Supplementary Figure 2) offers

a useful resource for contextualizing predictions. This visualization allows the identification of regions more environmentally similar to the MET conditions, where predictions are expected to be more reliable.

While the GIS-FA approach achieved an accuracy of 0.53 in LOOCV, further refinements could enhance its predictive performance. Recent studies suggest that modeling environmental features can improve both the parsimony and predictive accuracy of the prediction models[34,35]. Moreover, despite computationally demanding, incorporating high resolution spatiotemporal environmental data could refine genotypic performance predictions by enabling the estimation of environmental features at the plot or even plant level[17]. Additionally, integrating genomics, such as high-density molecular markers or genomic relationship matrices, has the potential to significantly boost predictive accuracy by capturing genomic signals associated with G× E and environmental adaptability[36]. Nonetheless, given the limitations of LOOCV in terms of generalizability, particularly when spatially correlated environments are present, future studies should consider more robust cross-validation strategies, such as spatial-block or leave-one-location-out schemes, to provide more reliable assessments of predictive performance[37]. Although the use of LOOCV at the environment level provides a rigorous internal validation strategy, an independent external validation dataset was not available in this study. This remains a limitation, as it prevents direct evaluation of the model's generalizability beyond the experimental network.

Future research should focus on developing genomic assisted crossing strategies based on the spatial suitability of different regions within the TPE, potentially accelerating the development of specific adapted varieties. In parallel, seasonal prediction models could enable the development of genotype recommendations tailored to specific cropping seasons. Extending these predictive models to assess genotypic responses under future climate scenarios would further provide a valuable tool for climate resilient breeding programs, addressing the increasing challenges posed by climate change. Lastly, the absence of explicit uncertainty measures or confidence intervals for spatial predictions represents a significant methodological gap. Addressing this limitation in future studies would enhance the reliability of genotype recommendations and better support informed decision-making and risk management in plant breeding programs.

Methods

Plant material

The phenotypic data were collected from Value for Cultivation and Use (VCU) trials, which represent the final stage of variety evaluation prior to official recommendation for commercial release. The trials comprised 59 common bean genotypes, including "Black" (15 lines) and "Carioca" (41 lines) types, evaluated over four agricultural years (2018–2021) across six municipalities in the state of São Paulo, covering three growing seasons: rainy, dry, and winter. This resulted in a total of 23 environments, defined by the combination of season, location, and year. Each environment was labeled using the first three letters of the location, the last two digits of the year, and the first letter of the season (e.g., Campinas in 2018 during the rainy season is designated as "Cam18R") (Fig. 9). The evaluated dataset exhibited a 44% imbalance in genotype presence across environments (Supplementary Fig. 1), indicating incomplete G×E combinations. The experiments were conducted by the Agronomic Institute (IAC) in accordance with the guidelines established by the Ministry of Agriculture, Livestock, and Supply/ National Cultivar Registry (MAPA/RNC) for VCU trials.

The experiments were laid out in randomized complete block design (RCBD) with three replications. Each plot consisted of four four-meter long rows, spaced 0.50 m apart. The two central rows were designated as the effective plot area (4 m^2). Initial fertilization was applied based on detailed soil analysis and tailored to the specific nutritional requirements of the crop at each location. Following harvest, grain yield was recorded for each plot and subsequently standardized to kilograms per hectare (kg ha^{-1}).

Statistical analyses of phenotypic data

We used the residual maximum likelihood (REML) method to estimate variance components[38], as a necessary step to predict genotypic values using best linear unbiased prediction (BLUP)[39]. We fitted the following model for the single-environment data analysis:

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

where $\boldsymbol{y}_{n\times 1}$ represents the vector of phenotypic observations, μ is the intercept, $\boldsymbol{r}_{b\times 1}$ is the vector of fixed replication effects, $\boldsymbol{g}_{t\times 1}$ is the vector of random genotypic effects [assumed as $\boldsymbol{g} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}\sigma_g^2)$], and $\boldsymbol{\epsilon}_{n\times 1}$ is the vector of random residual effects [with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma_\epsilon^2)$]. Here, $\mathbf{1}_{n\times 1}$ is a vector of ones, and $\boldsymbol{X}_{n\times b}$ and $\boldsymbol{Z}_{n\times t}$ are the incidence matrices associated with the replication and genotype effects, respectively. n is the number of observations (plots), b is the number of replications, and t is the number of genotypes.

We used single-environment data analyses to assess the significance of the genotypic effect employing the likelihood ratio test (LRT)[40]:

$$LRT = -2\left(-\log L + \log L_R\right) \tag{2}$$

where $\log L$ is the log-likelihood of the full model, and $\log L_R$ is the log-likelihood of the reduced model (excluding the tested effect). The LR statistic was compared to tabulated values of the chi-square (χ^2) distribution with one degree of freedom to determine the statistical significance of the genotypic variance component.

For the MET analysis, we included data from all environments where the genotypic effect was statistically significant, following the model:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}_1\boldsymbol{a} + \boldsymbol{X}_2\boldsymbol{r} + \boldsymbol{Z}\boldsymbol{g} + \boldsymbol{\epsilon} \tag{3}$$

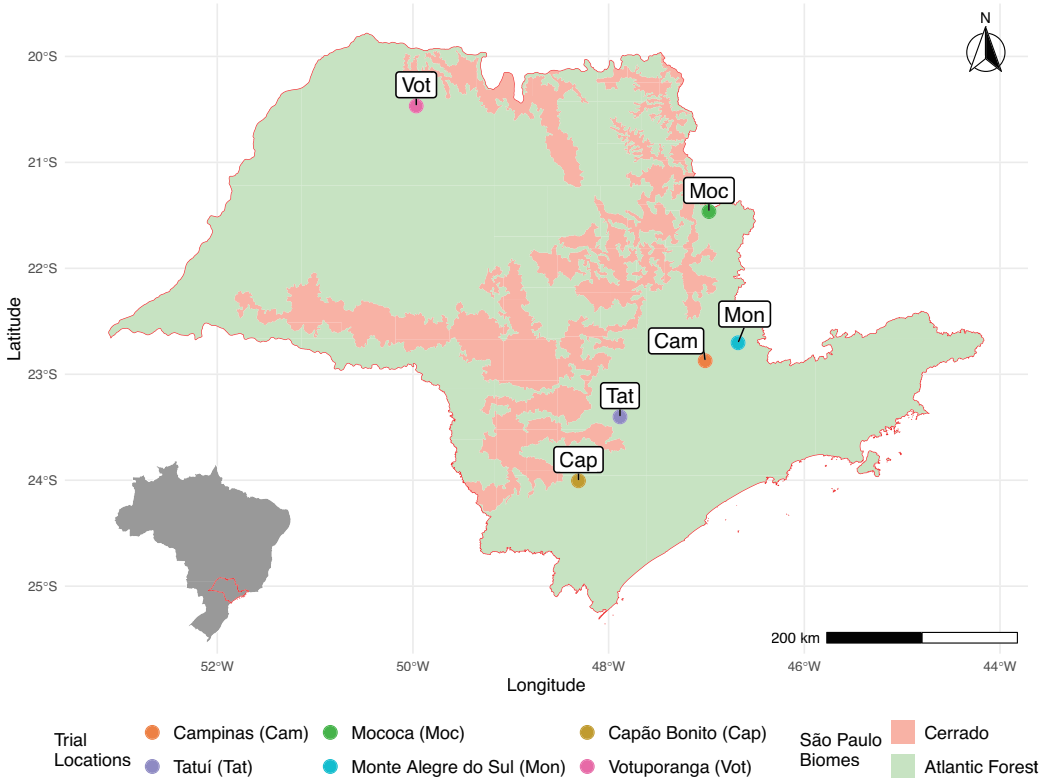


Fig. 9. Geographic distribution of field trials conducted from 2018 to 2021 in the state of São Paulo, Brazil. The map delineates São Paulo's borders in red and marks each trial location with distinct colored points, corresponding to the legend. The shaded areas represent the state's biomes: Cerrado (pink) and Atlantic Forest (green). The inset map of Brazil provides spatial context, highlighting São Paulo's position within the country.

where $\boldsymbol{y}_{v\times 1}$ represents the vector of phenotypic observations, μ is the intercept, $\boldsymbol{a}_{a\times 1}$ is the vector of fixed environmental effects, and $\boldsymbol{r}_{ba\times 1}$ is the vector of fixed replicate effects within environments. The vector $\boldsymbol{g}_{ta\times 1}$ represents the random genotypic effects within environments, assumed to follow a multivariate normal distribution $\boldsymbol{g} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}_g \otimes \boldsymbol{I}_t)$, where $\boldsymbol{\Sigma}_{g_{t\times t}}$ is the genetic covariance matrix, and $\boldsymbol{I}_t$ is an identity matrix of order t. The residual effects are represented by the vector $\boldsymbol{\epsilon}_{v\times 1}$, assumed to follow $\boldsymbol{\epsilon} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}_e \otimes \boldsymbol{I}_v)$, where $\boldsymbol{\Sigma}_{e_{a\times a}}$ is the residual covariance matrix, and $\boldsymbol{I}_v$ is an identity matrix of order v. Here, $\boldsymbol{1}_{v\times 1}$ is a vector of ones, while $\boldsymbol{X}_{1_{v\times a}}$, $\boldsymbol{X}_{2_{v\times b}}$, and $\boldsymbol{Z}_{v\times t}$ are the incidence matrices associated with the environmental, replicate, and genotypic effects, respectively. The quantities are defined as follows: v is the total number of observations, b is the number of replications, t is the number of genotypes, and a is the number of environments. The symbol $\otimes$ denotes the Kronecker product.

We modeled genotypic effects using the FA structure[8], where the genotypic effects ($\boldsymbol{g}$) and its covariance matrix($\boldsymbol{\Sigma}_g$), for an FA model of order K, are given by:

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{4}$$

$$\boldsymbol{g} = (\boldsymbol{\Lambda} \otimes \boldsymbol{I}_t)\boldsymbol{f} + \boldsymbol{\delta} \tag{5}$$

where $\boldsymbol{\Lambda}_{a\times K}$ is the factor loadings matrix, and $\boldsymbol{\Psi}_{a\times a}$ is a diagonal matrix of specific variances. The vector $\boldsymbol{g}_{t\times 1}$ represents the random genotypic effects, modeled as a function of the factor scores ($\boldsymbol{f}_{K\times t}$, where K is the number of factors) and the lack of fit effects ($\boldsymbol{\delta}_{t\times a}$).

In the FA framework, different numbers of factors (1 to $a-1$) can be considered for the analysis of experimental trials. To determine the most appropriate model, we used the Akaike Information Criterion (AIC)[41] and the Average Semi-Variance Ratio (ASR)[4]. We selected the model with the lowest AIC value among those with an ASR exceeding 70%.

When the number of factors exceeds one ($k > 1$), the factor loading matrix is not unique, requiring constraints to ensure identifiability and interpretability. To maintain consistency within the FA model, factor scores were assumed to be independent, with a diagonal (non identity) covariance matrix whose elements were ordered in decreasing magnitude. The factor loadings were constrained such that their inner product resulted in an identity matrix, ensuring orthonormal columns. Additionally, to impose uniqueness, all upper triangular elements of the loading matrix were set to zero. These constraints preserve the expected variance structure of random effects in mixed models while ensuring a biologically meaningful representation of the factors[30].

To address the identifiability issues and enhance the biological interpretability of the latent factors, the factor loading matrix $\mathbf{\Lambda}^*$, constrained to be lower triangular with orthonormal columns, was subjected to a rotation. Specifically, we applied a rotation based on the singular value decomposition (SVD) of $\mathbf{\Lambda}^*$, expressed as:

$$\mathbf{\Lambda}^* = \mathrm{UL}^{1/2}\mathrm{V}^\top$$

where U and V are orthogonal matrices and L is a diagonal matrix containing the singular values. The rotation ensures that the factors remain orthogonal and that the scale and structure of genetic variances across environments are maintained[8]. Accordingly, the corresponding factor scores were rotated as $\mathrm{f}^\star = (\mathrm{LV}^\top \otimes \mathrm{I}_V)\mathrm{f}$, where is the original vector of scores.

After model evaluation, we estimated the generalized heritabilities (H_q^2)[42], and the experimental coefficient of variation (CV_q) for each environment:

$$H_a^2 = 1 - \frac{V(\Delta)}{2\sigma_{g_q}^2} \tag{6}$$

where $\mathbb{V}(\Delta)$ represents the average pairwise prediction error variance, and $\sigma_{g_a}^2$ is the genotypic variance of the a-th environment.

$$CV_a = \frac{\sqrt{\sigma_{e_a}^2}}{\mu_a} \tag{7}$$

where $\sigma_{e_a}^2$ is the residual variance for environment a, and μ_a is the mean phenotypic value for environment a.

We estimated Genetic correlations between environments using the following equation[43]:

$$\boldsymbol{C} = \boldsymbol{D}(\mathbf{\Lambda}^*\mathbf{\Lambda}^{*'} + \mathbf{\Psi})\boldsymbol{D} \tag{8}$$

where $\boldsymbol{C}_{a\times a}$ represents the matrix of genetic correlations between environments, $\boldsymbol{D}_{a\times a}$ is a diagonal matrix whose elements are the inverse of the square roots of the diagonal elements of the genetic covariance matrix $\mathbf{\Sigma}_{\boldsymbol{g}}$. The matrix $\mathbf{\Lambda}^*_{a\times K}$ represents the rotated factor loadings, and $\mathbf{\Psi}_{a\times a}$ is a diagonal matrix of specific variances.

Factor analytic selection tools

Based on the selected model, we applied Factor Analytic Selection Tools (FAST) to assess both overall performance (OP), and genotype stability (RMSD)[12]. The OP is determined by the first factor, which should predominantly consist of positive loadings, providing a generalized measure of the main genotypic effects. The RMSD is derived from the subsequent factors, capturing environment specific variability and the remaining G×E interaction after accounting for overall performance[12,44]. We computed the OP and RMSD using the following equations:

$$OP_j = \frac{1}{a}\sum_{a=1}^{a}\lambda_{1_a}^* f_{1_{ja}}^*, \quad RMSD_j = \sqrt{\frac{1}{a}\sum_{a=1}^{a} e_{ja}^2} \tag{9}$$

where λ_a^* is the rotated loading of the first factor, and f_{ja}^* is the rotated score of genotype j in environment a, and $e_{ja} = \lambda_{2_a}^* f_{j2}^* + \lambda_{3_a}^* f_{j3}^* + \cdots + \lambda_{K_a}^* f_{jK}^*$.

After evaluating the selection tools, we adopted a weighting criterion of 2:1 between OP and RMSD to select the top 15% of lines, according to the following equation[4]:

$$I_j = 2 \times \left(\frac{OP_j - \overline{OP}}{\sqrt{V(OP)}} - \frac{RMSD_j - \overline{RMSD}}{\sqrt{V(RMSD)}} \right) \tag{10}$$

where $\overline{OP}$ represents the mean overall performance, $\overline{RMSD}$ is the mean stability value, $\sqrt{V(OP)}$ and $\sqrt{V(RMSD)}$ are their respective standard deviations.

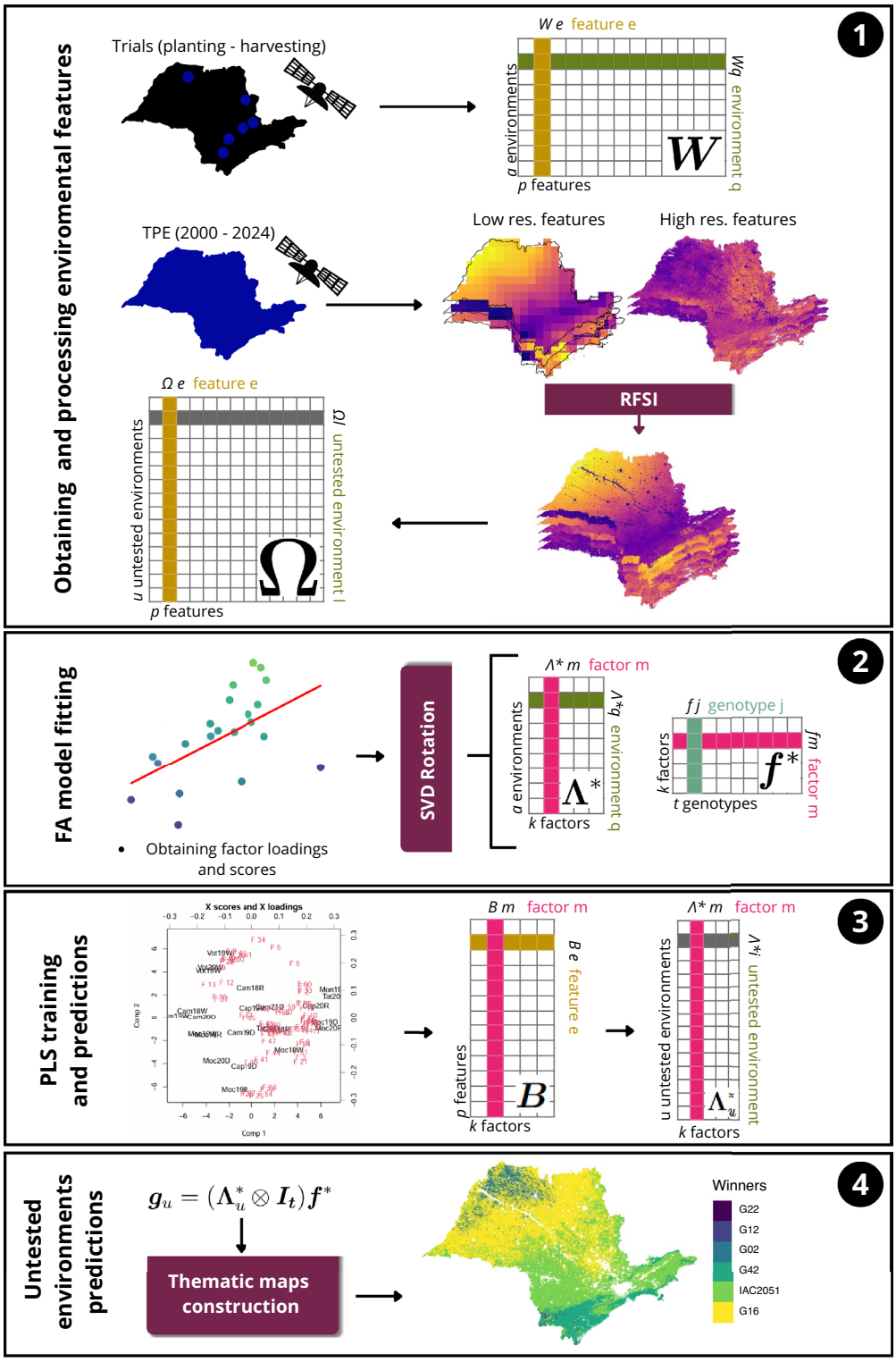
Additionally, we calculated the individual reliabilities of lines following Mrode [45], using the equation:

$$R_j = 1 - \frac{PEV_j}{\sigma_g^2} \tag{11}$$

where PEV_j represents the prediction error variance of the j-th genotype, and σ_g^2 is the average genotypic variance across environments.

Environmental features

A total of 69 environmental features were collected for the predictive analyses in this study. We classified these features into four distinct groups: agroclimatic features (37), soil properties (23), spectral bands (7), and vegetation indices (2) (Supplementary Table 3). We acquired data for the experimental sites using their respective geographic coordinates and planting and harvesting dates, and for the target population of environments (TPE), defined as the entire state of São Paulo, based on a historical series from 2000 to 2024.


1
Obtaining and processing enviromental features
Trials (planting - harvesting)
W e feature e
a environments
Wq environment q
p features
W
TPE (2000 - 2024)
Low res. features
High res. features
RFSI
Ω e feature e
u untested environments
Ωl untested environment l
p features
Ω
2
FA model fitting
Obtaining factor loadings and scores
SVD Rotation
Λ* m factor m
a environments
Λ*k environment q
k factors
Λ*
f j genotype j
fm factor m
k factors
t genotypes
f*
3
PLS training and predictions
X scores and X loadings
B m factor m
B e feature e
p features
k factors
B
Λ* m factor m
u untested environments
Λ*u untested environment
k factors
Λ*u
4
Untested environments predictions
$g_u = (\Lambda_u^* \otimes I_t) f^*$
Thematic maps construction
Winners
G22
G12
G02
G42
IAC2051
G16

◀**Fig. 10**. Enhanced GIS-FA workflow for prediction of genotypic performance in untested environments. (1) Initially, environmental features are obtained for experimental trials based on planting and harvesting dates, as well as for the target population of environments (TPE) across São Paulo state, considering a historical time series (2000–2024). Due to varying original resolutions (low and high), features undergo Random Forest Spatial Interpolation (RFSI) to standardize all variables to high spatial resolution. This results in the feature matrices for observed trial environments ($\boldsymbol{W}$) and untested environments in the TPE ($\boldsymbol{\Omega}$). (2) The FA model is fitted using phenotypic data, obtaining factor loadings and factor scores, which are subsequently rotated via singular value decomposition (SVD), resulting in rotated factor loadings ($\boldsymbol{\Lambda}^*$) and rotated genotype scores ($\boldsymbol{f}^*$). (3) Rotated factor loadings ($\boldsymbol{\Lambda}^*$) and environmental feature matrix for observed trials ($\boldsymbol{W}$) are employed to train a partial least squares (PLS) regression model, deriving the regression coefficients ($\boldsymbol{B}$). These coefficients, combined with the environmental features of the untested environments ($\boldsymbol{\Omega}$), allow for the prediction of rotated factor loadings in unobserved environments ($\boldsymbol{\Lambda}_u^*$). (4) Finally, predicted rotated factor loadings ($\boldsymbol{\Lambda}_u^*$) and rotated genotype scores ($\boldsymbol{f}^*$) are combined to estimate genotypic values for each genotype across all untested environments within the TPE, enabling the generation of thematic prediction maps.

We collected environmental features at a daily temporal resolution and computed average values to characterize each environment. In both experimental sites and the TPE, covariates were extracted at point-based geographic coordinates (latitude and longitude) and averaged between planting and harvesting dates. For the TPE, points followed a regular grid uniformly spaced across the state. To ensure relevance for agricultural prediction and avoid biases from non-cultivable areas, we implemented a systematic sampling strategy with spatial constraints. Specifically, we excluded pixels that overlapped with rivers, roads, or urban areas, using land cover and infrastructure shapefiles from the Brazilian Institute of Geography and Statistics (IBGE). This strategy ensured even spatial coverage, reproducibility, and environmental representativeness of agriculturally viable areas.

We sourced environmental datasets from NASA's Prediction of Worldwide Energy Resource (POWER)[46], SoilGrids[47], and the Moderate Resolution Imaging Spectroradiometer (MODIS)[48]. We conducted data acquisition and processing using the native Application Programming Interfaces (APIs) of these databases, ensuring standardized extraction and seamless integration of the environmental features. The spatial resolution adopted was 50 km^2 for agroclimatic variables, reflecting the highest available resolution from NASA POWER. All other variable classes, including soil properties and spectral data, were obtained at a finer resolution of 0.25 km^2. Accordingly, agroclimatic features were sampled from an evenly spaced grid of 85 points, while soil, spectral, and vegetation indices were extracted from a denser grid of 794,631 points.

Interpolation

To address differences in spatial resolution among the environmental datasets, we applied Random Forest Spatial Interpolation[24] (RFSI). This method leverages high-resolution covariates along with distances between observations to enhance spatial predictions. By incorporating the values of the nearest observations and their respective distances as additional covariates, RFSI effectively captures spatial autocorrelation that standard Random Forest approaches do not explicitly model. In our dataset, agroclimatic variables were originally available at 50 km^2 resolution, while soil properties, spectral bands and vegetation indices were obtained at 0.25 km^2. To standardize the input data for spatial prediction across the entire TPE, we used RFSI to interpolate all variables to a final resolution of 0.25 km^2 (500 m $\times$ 500 m). As a result, after interpolation, all environmental features were standardized to the same spatial resolution across the TPE, producing a total of 794,631 pixels. We adopted five nearest neighbors for each prediction location. The interpolation process follows the model:

$$\hat{z}(s_0) = f\left[x_1(s_0), ..., x_m(s_0), z(s_1), d_1, z(s_2), d_2, z(s_3), d_3, ..., z(s_n), d_n\right] \tag{12}$$

where s_i $(i = 1, ..., f)$ represents the i-th nearest observation location from s_0, and $d_i = |s_i - s_0|$ denotes the distance from each neighboring observation to the prediction location.

Predictions for untested environments

We used the environmental features obtained to predict genotypic eBLUPs in untested environments based on the GIS-FA method[16]. This approach integrates information from the factor analytic model with regression-based prediction using PLS models. We adjusted the modeling based on the factor loadings estimated from the factor analytic model and the extracted environmental features, following the formulation:

$$\boldsymbol{\Lambda}^* = \boldsymbol{W}\boldsymbol{B} + \boldsymbol{E} \tag{13}$$

where $\boldsymbol{W}_{a\times p}$ is the environmental features matrix for the evaluated environments, $\boldsymbol{B}_{p\times k}$ is the regression coefficient matrix, and $\boldsymbol{E}_{a\times k}$ is the residual matrix of the model. Here, p represents the number of environmental features.

Using the estimated regression coefficients ($\boldsymbol{B}$), we predicted the factor loadings for unobserved environments as follows:

$$\boldsymbol{\Lambda}_u^* = \boldsymbol{\Omega}\boldsymbol{B} \tag{14}$$

where $\mathbf{\Lambda}^*_{u\times K}$ is the matrix of rotated factor loadings for unobserved environments, $\mathbf{\Omega}_{u\times p}$ is the matrix of environmental features for unobserved environments, and $\boldsymbol{B}_{p\times k}$ is the previously estimated regression coefficient matrix. Here, u represents the number of untested environments.

Once we estimated the factor loadings for unobserved environments, we predicted the genotypic values using the multiple regression underlying the factor analytic model equation:

$$\boldsymbol{g}_u = (\mathbf{\Lambda}^*_u \otimes \boldsymbol{I}_t)\boldsymbol{f}^* \tag{15}$$

where $\mathbf{\Lambda}^*_u$ is the matrix of rotated factor loadings for unobserved environments, $\boldsymbol{I}_t$ is the identity matrix of order t, and $\boldsymbol{f}^*$ is the vector of rotated factor scores for the genotypes.

Validation and comparative analysis

We employed cross-validation to assess the reliability of the PLS model. We adopted a leave-one-out cross-validation (LOOCV) scheme, where we excluded one environment at a time and refitted both the FA and PLS models using the remaining data. Subsequently, we predicted eBLUPs for the excluded environment. For each LOOCV iteration, prediction accuracy was evaluated using Spearman's rank correlation coefficient between the predicted and observed eBLUPs, as well as the root mean square error (RMSE). Spearman correlation was selected to assess the consistency of genotype rankings across environments, which is critical for selection and recommendation purposes. The overall performance of each method was summarized as the mean Spearman correlation across all LOOCV iterations. This approach differs from that originally proposed by Araújo et al.[16] because we excluded the environment from both the PLS training phase and the FA model fitting (including factor loading estimation), ensuring a more coherent cross-validation process.

In addition to the predictive metrics, we incorporated a spatial map representing environmental similarity between the experimental sites and the broader TPE. Based on the methodology proposed by Araújo et al.[16], euclidean distances were calculated in the multivariate space of scaled environmental covariates between each unobserved location and all trial environments. The minimum distance from each location to any experimental site was retained as an index of similarity.

To assess how the novelties introduced to the method impacts its predictive ability, we also employed the original GIS-FA pipeline and cross-validated as described above. Additionally, we evaluated the performance of these approaches not only on the common bean dataset used in our study, but also on the soybean dataset from the original GIS-FA paper[16]. This dataset is publicly available and can be accessed via the Git repository. It comprises 195 genotypes evaluated under rainfed conditions across 49 distinct environments, encompassing 13 locations distributed throughout the state of Mato Grosso do Sul and the broader Central-West region of Brazil over three cropping seasons. Field trials were conducted by the Mato Grosso do Sul Foundation using a randomized complete block design with three replicates. Experimental units consisted of five rows, each 12 meters long and spaced 0.5 meters apart, totaling 30 m^2 per plot. We performed all analyses using `R`[49], along with the `ASReml-R` [50] and `pls` packages[51].

Changes to the GIS-FA workflow

Given that one of the objectives of this study is to propose enhancements to the GIS-FA methodology, some clarifications regarding modifications in specific stages of the pipeline are necessary to ensure clarity of the innovations proposed here. Originally, the GIS-FA method obtained environmental variables for untested environments by sampling 50 points per municipality within the TPE. Subsequently, following the prediction of genotypic eBLUPs PLS, IDW was employed to spatially interpolate these predicted values across the TPE.

The modified approach presented in this study extracts environmental variables at uniformly spaced points while excluding rivers, roads, and urban areas throughout the entire TPE, leveraging the highest available resolution from each data platform. Following extraction, RFSI is applied to variables initially acquired at lower resolutions. Consequently, this ensures all environmental features are consistently available at each individual location across the TPE. This methodological shift enables direct prediction using the trained PLS model for every pixel. The revised workflow is illustrated in Fig. 10.

Data availability

A step-by-step documented R script along with the environmental datasets used in this study are available at: https://github.com/Kaio-Olimpio/GIS_FA_Common_Bean. Phenotypic data from the multi-environment trials were generated by the Agronomic Institute (IAC) and are not publicly available due to institutional policies, but may be provided upon reasonable request. Correspondence and requests for materials should be addressed to K.O.G.D. (kaio.o.dias@ufv.br).

Received: 21 May 2025; Accepted: 24 September 2025
Published online: 30 October 2025

References

1. Heinemann, A. B., Costa-Neto, G., Fritsche-Neto, R., Matta, D. H. & Fernandes, I. K. Enviromic prediction is useful to define the limits of climate adaptation: A case study of common bean in brazil. *Field Crop Res.* **286**, 108628. https://doi.org/10.1016/j.fcr.2022.108628 (2022).
2. Heinemann, A. B. et al. Drought impact on rainfed common bean production areas in brazil. *Agric. For. Meteorol.* **225**, 57–74. https://doi.org/10.1016/j.agrformet.2016.05.010 (2016).
3. Malosetti, M., Ribaut, J.-M. & Eeuwijk, F. A. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Front. Physiol.* https://doi.org/10.3389/fphys.2013.00044 (2013).

4. Chaves, S. F. S. et al. Employing factor analytic tools for selecting high-performance and stable tropical maize hybrids. *Crop Sci.* **63**, 1114–1125. https://doi.org/10.1002/csc2.20911 (2023).
5. Smith, A. B., Cullis, B. R. & Thompson, R. The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *J. Agric. Sci.* **143**, 449–462. https://doi.org/10.1017/S0021859605005587 (2005).
6. Eeuwijk, F. V., Cooper, M., DeLacy, I., Ceccarelli, S. & Grando, S. Some vocabulary and grammar for the analysis of multi-environment trials, as applied to the analysis of fpb and ppb trials. *Euphytica* **122**, 477–490. https://doi.org/10.1023/A:1017591407285 (2001).
7. Piepho, H. Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* **53**, 761–766 (1997).
8. Smith, A., Cullis, B. & Thompson, R. 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 (2001).
9. Cullis, B. R., Jefferson, P., Thompson, R. & Smith, A. B. 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 (2014).
10. Smith, A. B., Ganesalingam, A., Kuchel, H. & Cullis, B. R. 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 (2015).
11. Tolhurst, D. J., Mathews, K. L., Smith, A. B. & Cullis, B. R. Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *J. Anim. Breed. Genet.* **136**, 279–300. https://doi.org/10.1111/jbg.12404 (2019).
12. Smith, A. B. & Cullis, B. R. 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 (2018).
13. Costa-Neto, G., Crossa, J. & Fritsche-Neto, R. Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. *Front. Plant Sci.* https://doi.org/10.3389/fpls.2021.717552 (2021).
14. Resende, R. T. et al. Enviromics in breeding: Applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* **134**, 95–112. https://doi.org/10.1007/s00122-020-03684-z (2021).
15. Tolhurst, D. J., Gaynor, R. C., Gardunia, B., Hickey, J. M. & Gorjanc, G. 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 (2022).
16. Araújo, M. S. et al. GIS-FA: An approach to integrating thematic maps, factor-analytic, and envirotyping for cultivar targeting. *Theor. Appl. Genet.* **137**, 80. https://doi.org/10.1007/s00122-024-04579-z (2024).
17. Resende, R. T. et al. Satellite-enabled enviromics to enhance crop improvement. *Mol. Plant* **17**, 848–866. https://doi.org/10.1016/j.molp.2024.04.005 (2024).
18. Xu, Y. Envirotyping for deciphering environmental impacts on crop plants. *Theor. Appl. Genet.* **129**, 653–673. https://doi.org/10.1007/s00122-016-2691-5 (2016).
19. Resende, R. T. et al. Satellite-enabled enviromics to enhance crop improvement. *Mol. Plant* **17**, 848–866. https://doi.org/10.1016/j.molp.2024.04.005 (2024).
20. Fernandes, I. K., Vieira, C. C., Dias, K. O. G. & Fernandes, S. B. Using machine learning to combine genetic and environmental data for maize grain yield predictions across multi-environment trials. *Theor. Appl. Genet.* **137**, 189. https://doi.org/10.1007/s00122-024-04687-w (2024).
21. Bhowmik, A. K. & Costa, A. C. Representativeness impacts on accuracy and precision of climate spatial interpolation in data-scarce regions. *Meteorol. Appl.* **22**, 368–377. https://doi.org/10.1002/met.1463 (2015).
22. Li, Y., Wu, H., Chen, H. & Zhu, X. A robust framework for resolution enhancement of land surface temperature by combining spatial downscaling and spatiotemporal fusion methods. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–14. https://doi.org/10.1109/TGRS.2023.3283614 (2023).
23. Klinges, D. H. et al. Proximal microclimate: Moving beyond spatiotemporal resolution improves ecological predictions. *Glob. Ecol. Biogeogr.* **33**, e13884. https://doi.org/10.1111/geb.13884 (2024).
24. Sekulic, A., Kilibarda, M., Heuvelink, G. B., Nikolic, M. & Bajat, B. Random forest spatial interpolation. *Remote Sens.* https://doi.org/10.3390/rs12101687 (2020).
25. Jiao, S. et al. Spatial prediction using random forest spatial interpolation with sample augmentation: A case study for precipitation mapping. *Earth Sci. Inf.* **16**, 863–875. https://doi.org/10.1007/s12145-023-00936-6 (2023).
26. Sekulić, A., Kilibarda, M., Protić, D. & Bajat, B. A high-resolution daily gridded meteorological dataset for serbia made by random forest spatial interpolation. *Sci. Data* **8**, 123. https://doi.org/10.1038/s41597-021-00901-2 (2021).
27. Bezerra, L. M. C., Fredo, C. E., Chiorato, A. F. & Carbonell, S. A. M. The research, development, and innovation trajectory of the IAC common bean breeding program. *Crop Breed. Appl. Biotechnol.* **21**, e36872124. https://doi.org/10.1590/1984-70332021v21n2a33 (2021).
28. Dias, K. O. D. G. et al. Estimating genotype × environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. *Crop Sci.* **58**, 72–83. https://doi.org/10.2135/cropsci2016.07.0566 (2018).
29. Ferrante, A., Cullis, B. R., Smith, A. B. & Able, J. A. A multi-environment trial analysis of frost susceptibility in wheat and barley under australian frost-prone field conditions. *Front. Plant Sci.* https://doi.org/10.3389/fpls.2021.722637 (2021).
30. Smith, A., Norman, A., Kuchel, H. & Cullis, B. Plant variety selection using interaction classes derived from factor analytic linear mixed models: Models with independent variety effects. *Front. Plant Sci.* **12**, 737462. https://doi.org/10.3389/fpls.2021.737462 (2021).
31. Bakare, M. A. et al. Parsimonious genotype by environment interaction covariance models for cassava (*Manihot esculenta*). *Frontiers in Plant Science* https://doi.org/10.3389/fpls.2022.978248 (2022).
32. Chaves, S. F. S., Damacena, M. B., Dias, K. O. G., de Almada Oliveira, C. V. & Bhering, L. L. Factor analytic selection tools and environmental feature-integration enable holistic decision-making in eucalyptus breeding. *Sci. Rep.* **14**, 18429. https://doi.org/10.1038/s41598-024-69299-2 (2024).
33. Rogers, A. R. & Holland, J. B. Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3: Genes, Genomes, Genetics* https://doi.org/10.1093/g3journal/jkab440 (2022).
34. Piepho, H.-P. & Blancon, J. Extending finlay–wilkinson regression with environmental covariates. *Plant Breed.* **142**, 621–631. https://doi.org/10.1111/pbr.13130 (2023).
35. Resende, R. T. et al. GIS-based G×E modeling of maize hybrids through enviromic markers engineering. *New Phytol.* **245**, 102–116. https://doi.org/10.1111/nph.19951 (2025).
36. Crossa, J. et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends Plant Sci.* **22**, 961–975. https://doi.org/10.1016/j.tplants.2017.08.011 (2017).
37. Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929. https://doi.org/10.1111/ecog.02881 (2017).
38. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554. https://doi.org/10.1093/biomet/58.3.545 (1971).
39. Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447 (1975).
40. Wilks, S. S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**, 60–62 (1938).
41. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723. https://doi.org/10.1109/TAC.1974.1100705 (1974).
42. Cullis, B. R., Smith, A. B. & Coombes, N. E. On the design of early generation variety trials with correlated data. *J. Agric. Biol. Environ. Stat.* **11**, 381–393. https://doi.org/10.1198/108571106X154443 (2006).

43. Cullis, B. R., Smith, A. B., Beeck, C. P. & Cowling, W. A. 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 (2010).
44. Stefanova, K. T. & Buirchell, B. Multiplicative mixed models for genetic gain assessment in lupin breeding. *Crop Sci.* **50**, 880–891. https://doi.org/10.2135/cropsci2009.07.0402 (2010).
45. Mrode, R. A. *Linear models for the prediction of animal breeding values* (CABI, 2014), 3rd edn.
46. NASAPOWER. Nasa power data access viewer (2025). Accessed: 2025-01-04.
47. Poggio, L. et al. Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *Soil* **6**, 359–381. https://doi.org/10.5194/soil-6-359-2020 (2020).
48. MODIS. Modis data products (2025). Accessed: 2025-01-04.
49. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (2024).
50. The VSNi Team. *ASReml: Fits linear mixed models using REML* (2023). R package version 4.2.0.332.
51. Liland, K. H., Mevik, B.-H. & Wehrens, R. *PLS: Partial least squares and principal component regression* (2024). R package version 2.8-5.

Acknowledgements
The authors express their sincere gratitude to the researchers and field assistants who conducted the field experiments at the Common Bean Breeding Program of the Instituto Agronômico (PMGF-IAC).

Author contributions
G.M.B. conceived the study, developed the methodology, performed the formal analysis, wrote the original draft, and contributed to the review and editing of the manuscript. M.S.A., L.A.S.D., S.C., and K.O.G.D. contributed to the conceptualization, methodology, and review and editing of the manuscript. A.F.C. and S.A.M.C. provided resources and participated in the validation of results. L.P.C. contributed to the methodology and validation. All authors reviewed the manuscript.

Funding
This work was supported by the Minas Gerais State Agency for Research and Development (FAPEMIG), including the project APQ-02529-22, the São Paulo Research Foundation (FAPESP, Grant 2024/01868), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES).

Declarations

Competing interests
The authors declare that they have no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-21882-x.

Correspondence and requests for materials should be addressed to S.C. or K.O.G.D.