Received: 23 August 2022 | Accepted: 31 March 2023 | Published online: 25 May 2023

DOI: 10.1002/csc2.20995

Crop Science

ORIGINAL ARTICLE

Crop Breeding & Genetics

Analysis of repeated measures data through mixed models: An application in *Theobroma grandiflorum* breeding

Saulo F. S. Chaves[1] | **Rodrigo S. Alves**[2] | **Luiz A. S. Dias**[1] | **Rafael M. Alves**[3] | **Kaio O. G. Dias**[2] | **Jeniffer S. P. C. Evangelista**[2]

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

[2]Department of General Biology, Federal University of Viçosa, Viçosa, Brazil

[3]Brazilian Agricultural Research Corporation (Embrapa), Eastern Amazon unit, Belém, Brazil

Correspondence
Saulo F. S. Chaves, Federal University of Viçosa, Department of Agronomy, Viçosa, Brazil.
Email: saulo.chaves@ufv.br

Assigned to Associate Editor Philippe Seguin.

Funding information
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Fundação de Amparo à Pesquisa do Estado de Minas Gerais; Conselho Nacional de Desenvolvimento Científico e Tecnológico

Abstract

Theobroma grandiflorum is a perennial fruit tree native to the Amazon region. As a perennial species with continuous production throughout the years, breeders should seek well-conducted trials, accurate phenotyping and adequate statistical models for genetic evaluation and selection that can leverage the information provided by the repeated measures. We evaluated 13 models with different covariance structures for genetic and residual effects for *T. grandiflorum* evaluation, using an unbalanced dataset with 34 hybrids from the triple-crossing of nine parents, planted in a randomized complete block design. For nine consecutive years, the fruit yield of these hybrids was evaluated. Each model had its goodness-of-fit tested by the Akaike information criterion. The most adequate model for estimating the variance components and the breeding values were modelled with the first-order heterogeneous autoregressive for residual effects and third-order factor analytic for genetic effects. From this model, we used the factor analytic selection tools for selecting the top 10 families, providing a genetic gain of 10.42%. These results are important not only for *T. grandiflorum* breeding but also to show that in any repeated measures' data from fruit-bearing perennial species the modelling of genetic and residual effects should not be neglected.

1 | INTRODUCTION

Theobroma grandiflorum Willd. (Ex. Spreng) Schum., Malvaceae, is a fruit-bearing perennial species native to the Brazilian Amazon. The pulp and seeds of its fruit, the cupuassu, are important commercial products (Pereira et al., 2018). The species has great potential due to the agro-industrial value of its fruit. Indeed, its production has grown in northern Brazil, where there are 16,000 hectares cultivated, which generates employment and revenues for the local population. The first orchards were established with seeds from wild individuals, i.e., without any breeding. For this reason, most of the Amazonian orchards have low yield and high occurrence of the witches' broom disease [*Moniliophthora perniciosa* (Stahel) Aime & Phillips-Mora] (Alves & Chaves, 2020). The great intra- and inter-population divergence, conditioned by the species' allogamy and potentiated by its self-incompatibility, also contributes to this fact (Alves et al., 2007).

Embrapa Amazônia Oriental (Brazilian Agricultural Research Corporation, Eastern Amazon unit) started a breeding programme to overcome these issues more than 30 years ago. The programme started with plants being collected in the mid-1980s, from the species' putative centre of origin

and along the Amazon River channel (Alves et al., 2007). The genetic materials collected were conserved in an ex situ on-farm germplasm bank (GB). In the following years, genotypes from other locations were collected to enrich the GB. This population was the base for developing superior genotypes, i.e., holders of both disease resistance and yield alleles. To this day, the programme has released ten commercial cultivars: one improved population and nine clonal cultivars (Alves & Chaves, 2020).

Nevertheless, the perennial nature of the species demands differentiated experimental and statistical methods. Note that *T. grandiflorum* can keep producing for several decades, and, consequently, data are often collected from the same plant several times. In fact, *T. grandiflorum* has a production window of 6 months, so data can be collected repeatedly within a year. Such a pattern characterizes repeated measures data, in which the most important peculiarity is the presence of genetic and environmental covariance between measures (Faveri et al., 2017; Piepho & Eckl, 2014; Wolfinger, 1996). This cannot be ignored if breeders intend to follow the basic principles of plant breeding, namely (i) appropriate experimental design, (ii) accurate data collection and (iii) adequate statistical models for genetic evaluation (Stringer et al., 2017). Note that a "measure" can mean a different harvest (data collection within a year), or even a different year, as it is in this study's case (see next section for more details).

A suitable method for analysing repeated measures data is the use of linear mixed models, which allows for the modelling of covariance structures with different complexity levels. The simplest and most parsimonious models usually assume that the variances are the same throughout measures. This condition, however, is rarely met, so these models may not reflect the plants' real genetic value. This is because environmental variations observed in each measure may lower the correlation between them and make the variances heterogeneous (Faveri et al., 2017). On the other hand, the multivariate model is more complete and informative. It considers each measure as a different trait, particularizing not only variances but also pairwise covariances (Mrode, 2014). Nonetheless, this model has a higher statistical and computational complexity (Meyer, 2009).

This motivated us to test several models with different covariance structures to analyse repeated measures data in *T. grandiflorum* breeding. We hypothesize that by modelling the genetic and non-genetic random effects, we will reach higher genetic gains and more accurate results than the traditional models that are commonly used. Thus, our goals were (i) to compare models with different covariance structures and (ii) to select high-performance hybrids and parents of *T. grandiflorum* across multiple years.

Core Ideas

- Modelling the covariance structure of random effects leads to higher accuracy and genetic gains.
- Factor Analytic Selection Tools are suitable for performing selection in repeated measures context.
- The modelling of genetic and residual effects should not be neglected for fruit-bearing perennial species
- *Theobroma grandiflorum* breeding may also benefit from early selection and precocious multi-environment trials

2 | MATERIAL AND METHODS

2.1 | Site description

The trial was established at Tomé-Açu municipality (latitude 02° 24' 05.6'' S; longitude 48° 00' 08.8'' W; altitude of 45 m above sea level), in the Brazilian Amazon, in February 2007. The climate of the region is the type B4rA'a', according to the Thornthwaite classification (Moraes et al., 2020). The soil at the experimental site was an Oxisol with medium texture, low natural fertility, and deep with good drainage. During the years of evaluation, the average temperature, humidity and rainfall were 27°C, 85.5%, and 2,729 mm, respectively. Weather data were obtained in situ, from a local weather station.

Theobroma grandiflorum needs partial shading in the juvenile period (i.e., first 3 years). To fulfil this requirement, we selected an area that was already being used for black pepper (*Piper nigrum* L.) cultivation. Thus, *T. grandiflorum* plants were intercropped with 3-year-old, full-production black pepper plants. The black peppers were planted in a double row scheme, with 2 m spacing between inner rows and 4 m spacing between double rows. Each *T. grandiflorum* was planted between black peppers' double rows, at a 6 × 4 m spacing. From the third year after trial establishment onwards, black peppers' mortality increased due to the natural occurrence of fusariosis, caused by *Fusarium solani f. sp. piperis*. At the end of the fourth year, black peppers were all dead, and *T. grandiflorum* plants were alone in the field. Figure 1 illustrates the trial arrangement with and without the black pepper plants.

The fertilization was done using cattle manure and an NPK fertilizer with the formulation 10:28:20. The quantity varied with the stage of maturity of the plants. In the dry season (usually from July to November in the locality where the trial was conducted), plants were irrigated with 60 L of water per day. In the rainy season, irrigation was not

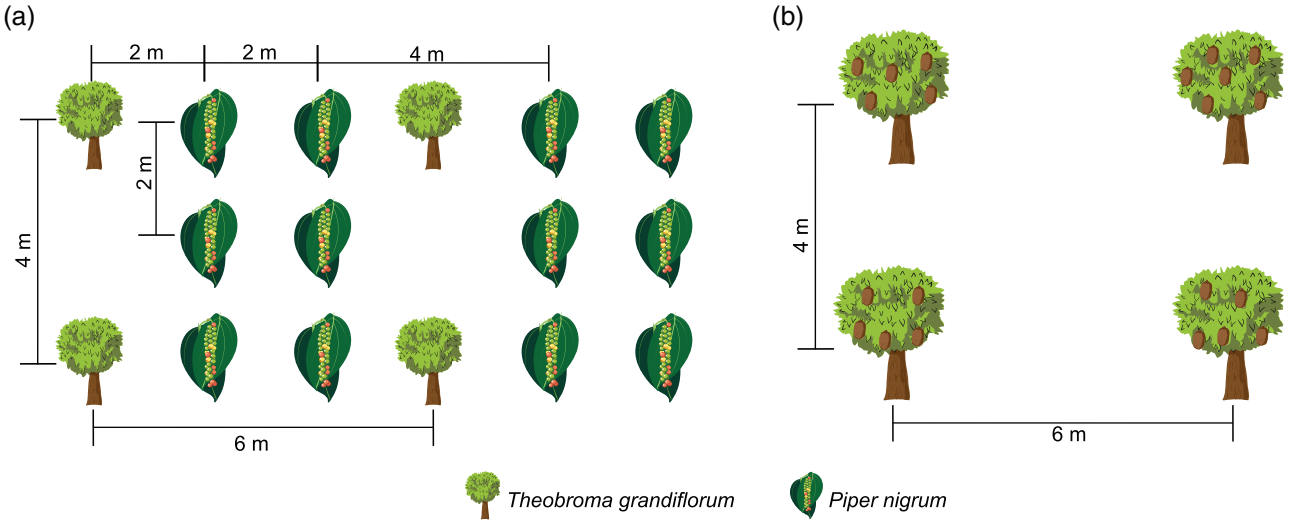

FIGURE 1 Illustration of the field trial arrangement in the first 4 years (a) and after the fifth year onwards (b).

necessary. There was no usage of agrochemicals to control pests or diseases, although herbicides were used to control weeds under the plants' crown, to avoid interspecific competition. Witches' broom was controlled by pruning any symptomatic branch.

2.2 | Plant material and experimental design

The evaluated hybrids were acquired by triple crossing following the scheme (C1 × C2) × C3, i.e., the genic contribution of three different clones (see Supplemental Table S1 for more information about the hybrids and their parents). In total, we evaluated 34 hybrids assigned to a randomized complete block design with five to ten replicates and three plants per plot. To leverage the genetic covariance for predicting the breeding values of both hybrids and parents, the $t \times t$ numerator kinship matrix (Figure 2) was estimated, where t is the number of genotypes (hybrids and parents). Henceforth, this matrix will be called **A**. **A** is composed of the additive relationship between genotypes ($\alpha_{ii'}$), given by $\alpha_{ii'} = 2 \times \Theta_{ii'}$, where $\Theta_{ii'}$ is the coancestry coefficient between the ith and the i'th genotypes. We used the `nadiv` package (Wolak, 2012) for estimating **A**.

From the third year after establishment onwards (from 2010 to 2019), we evaluated on a plot basis the fruit yield per plant (kg), the product of the number of fruits per plant by the average fruit weight of the same plant. *T. grandiflorum* production window coincides with the rainy season in the Amazon region (Venturieri, 2011). In the locality where the trial was conducted, this happens from mid-December to the end of May. During this period, we performed four harvests, with about 45 days of difference between them. The objective was to determine the yield for the whole production window, allowing a fair comparison between genotypes with early and late production. The value of 1 year is the sum of the harvests performed in a production window. To simplify the analyses and interpretations, we allowed production windows to match with calendar years.

2.3 | Statistical analyses

We performed all the analyses using the linear mixed models' methodology [REML/BLUP, Patterson and Thompson (1971), Henderson (1975)]. Following Smith et al. (2007) and Faveri et al. (2015), we used a base model as a starting point for modelling the genetic and residual effects, sequentially. In the mathematical notations below, n is the total number of observations ($n = n_1 + n_2 + \cdots + n_j$, where n_j is the number of observations per year), m is the number of years ($j = 1, 2, 3, \ldots, 9$), r is the number of replicates ($f = 1, 2, 3, \ldots, 10$), q is the number of plots ($s = 1, 2, 3, \ldots, 250$), and t is the number of genotypes, both hybrids and parents ($i = 1, 2, 3, \ldots, 51$). The base model was:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}_1\mathbf{a} + \mathbf{X}_2\mathbf{b} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the $n \times 1$ vector of phenotypes, μ is the intercept associated with an $n \times 1$ vector of ones ($\mathbf{1}$), $\mathbf{a}$ is the $m \times 1$ vector of the fixed effects of years with incidence matrix $\mathbf{X}_1^{(n \times m)}$, $\mathbf{b}$ is the $mr \times 1$ vector of the fixed effects of replicates within years, with incidence matrix $\mathbf{X}_2^{(n \times mr)}$; $\mathbf{g}$ is the $mt \times 1$ vector of random additive genetic effects for individual years [$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G} = \boldsymbol{\Sigma}_g \otimes \mathbf{A})$, where $\boldsymbol{\Sigma}_g$ is the $m \times m$ matrix of genetic covariances between years, and $\mathbf{A}$ is the $t \times t$ numerator kinship matrix] with associated design matrix $\mathbf{Z}_1^{(n \times mt)}$, $\mathbf{p}$ is the

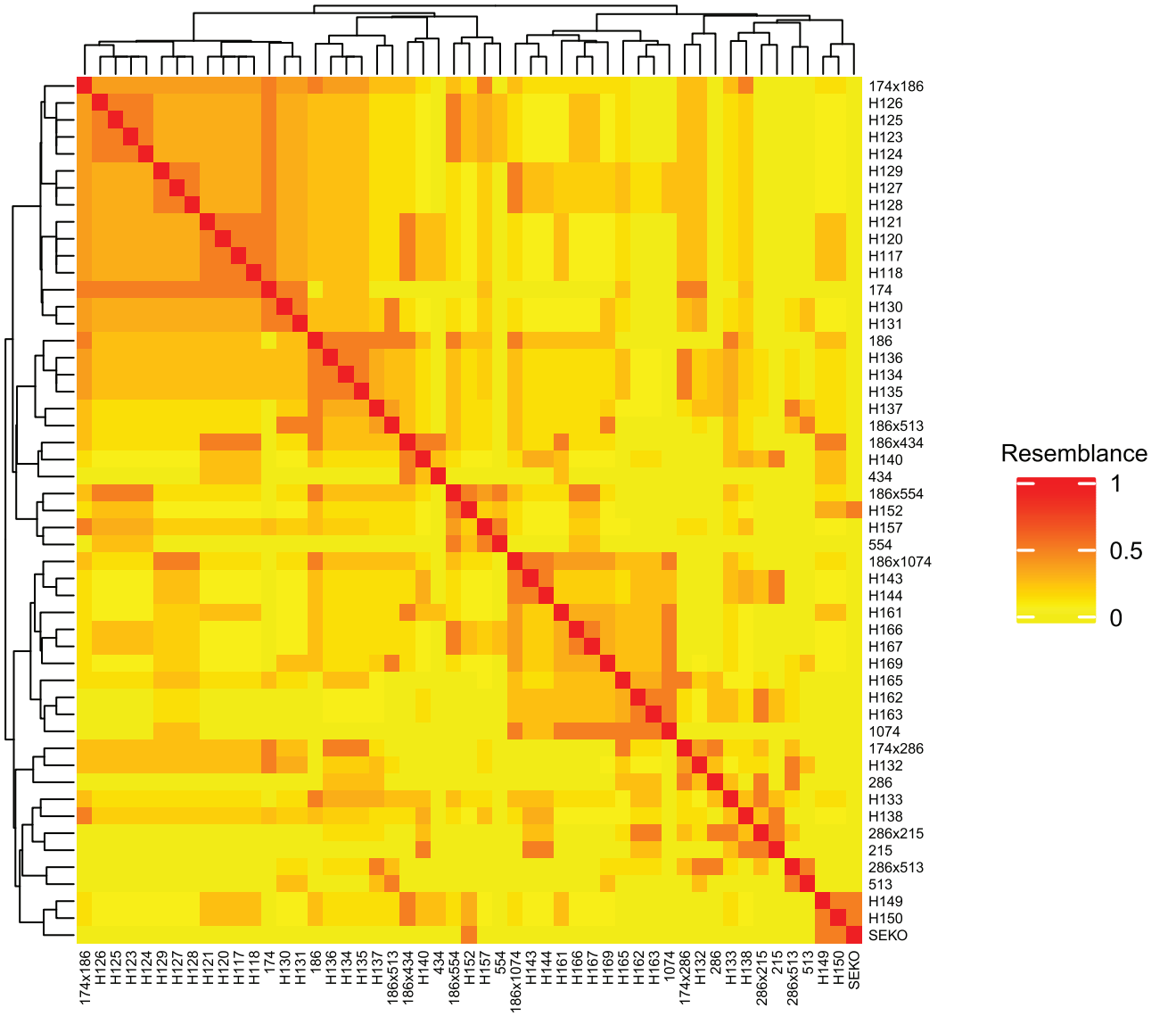

FIGURE 2 Graphical representation of the numerator kinship matrix, illustrating the resemblance between *Theobroma grandiflorum* hybrids (H) and their parents.

$q \times 1$ vector of permanent plot effects [$\mathbf{p} \sim N(\mathbf{0}, \sigma_p^2 \mathbf{I}_q)$] with associated incidence matrix $\mathbf{Z}_2^{(n \times q)}$ and $\mathbf{e}$ is the $n \times 1$ vector of residual effects [$\mathbf{e} \sim N(\mathbf{0}, \mathbf{R} = \mathbf{\Sigma}_e \otimes \mathbf{I}_{n_j}$), where $\mathbf{\Sigma}_e$ is the $m \times m$ matrix of residual covariances between years, and $\mathbf{I}_{n_j}$ is an identity matrix of order n_j]. Within repeated measures data, we can divide the environmental effects into permanent and temporary (Mrode, 2014, pp. 4–5). In this study's context, the permanent environmental effect refers to multiple observations per plot, comprising the permanent plot effect. By considering this effect in the linear mixed model, the residue will comprise the temporary environmental effects. Otherwise, both permanent and temporary environmental effects will be part of the error effects. Here, we chose to account for the permanent plot effect ($\mathbf{p}$) in the model, assuming its homogeneity ($\sigma_p^2 \mathbf{I}_q$). In fact, our modelling will be restricted to $\mathbf{\Sigma}_g$ and $\mathbf{\Sigma}_e$ covariance matrices between years, respectively.

2.3.1 | Modelling the genetic effects

First, we modelled $\mathbf{\Sigma}_g$ with eight different covariance structures, fixing the homoscedasticity (variances identity: IDV) for the residual effect. In the matrix representations of $\mathbf{\Sigma}_g$, described in the following paragraphs, each row/column represents a year. The diagonal contains the covariance between a year and itself, i.e., the variance of that particular year. Off-diagonal values represent the covariance between the years j and j'. Since the dataset comprises 9 years, $\mathbf{\Sigma}_g$ is a 9×9 matrix. The first structure was the homogeneous compound symmetry, which divides the genetic variance into the main

effect and the effect due to the genotypes-by-years interaction (GYI):

$$\begin{bmatrix} \sigma_g^2+\sigma_{gy}^2 & \sigma_g^2 & \sigma_g^2 & \dots & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2+\sigma_{gy}^2 & \sigma_g^2 & \dots & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2+\sigma_{gy}^2 & \dots & \sigma_g^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_g^2 & \sigma_g^2 & \sigma_g^2 & \dots & \sigma_g^2+\sigma_{gy}^2 \end{bmatrix}$$

$$\otimes\mathbf{A} = (\sigma_g^2\mathbf{J}_m + \sigma_{gy}^2\mathbf{I}_m) \otimes \mathbf{A}, \quad (2)$$

where σ_g^2 is the variance due to genetic main effects, σ_{gy}^2 is the variance due to the GYI, $\mathbf{I}_m$ is an identity matrix of order m, $\mathbf{J}_m$ is an $m \times m$ matrix of ones and $\otimes$ is the direct product.

The second covariance structure was the diagonal (DIAG), which considers the heterogeneity of genetic variances between years and the nullity of pairwise covariances:

$$\begin{bmatrix} \sigma_{g_1}^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_{g_2}^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_{g_3}^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_{g_9}^2 \end{bmatrix} \otimes \mathbf{A} = \mathbf{D} \otimes \mathbf{A}, \quad (3)$$

where $\mathbf{D}$ is an $m \times m$ diagonal matrix, in which each element represents the genetic variance of the jth year ($\mathbf{D} = \{\sigma_{g_j}^2\}$).

Sequentially, we used the first-order homogeneous autoregressive (AR1) to model $\boldsymbol{\Sigma}_g$. In this structure, the covariances are products of the genetic variance with autocorrelation coefficients (ρ^t), which have an exponent as high as the distance between years:

$$\sigma_g^2 \times \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^8 \\ \rho & 1 & \rho & \dots & \rho^7 \\ \rho^2 & \rho & 1 & \dots & \rho^6 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^8 & \rho^7 & \rho^6 & \dots & 1 \end{bmatrix} \otimes \mathbf{A} = \sigma_g^2\boldsymbol{\Gamma} \otimes \mathbf{A}, \quad (4)$$

where $\boldsymbol{\Gamma}$ is a $m \times m$ matrix with ones on the diagonal and the autocorrelations on the off-diagonal.

We also structured $\boldsymbol{\Sigma}_g$ with the AR1 heterogeneous counterpart (AR1H). This covariance structure considers the heterogeneity of variances and the covariances are obtained by the product of the genetic standard deviations of two years and the ρ^t:

$$\begin{bmatrix} \sigma_{g_1}^2 & \rho\sigma_{g_1}\sigma_{g_2} & \rho^2\sigma_{g_1}\sigma_{g_3} & \dots & \rho^8\sigma_{g_1}\sigma_{g_9} \\ \rho\sigma_{g_2}\sigma_{g_1} & \sigma_{g_2}^2 & \rho\sigma_{g_2}\sigma_{g_3} & \dots & \rho^7\sigma_{g_2}\sigma_{g_9} \\ \rho^2\sigma_{g_3}\sigma_{g_1} & \rho\sigma_{g_3}\sigma_{g_2} & \sigma_{g_3}^2 & \dots & \rho^6\sigma_{g_3}\sigma_{g_9} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^8\sigma_{g_9}\sigma_{g_1} & \rho^7\sigma_{g_9}\sigma_{g_2} & \rho^6\sigma_{g_9}\sigma_{g_8} & \dots & \sigma_{g_9}^2 \end{bmatrix}$$

$$\otimes\mathbf{A} = \sqrt{\mathbf{D}}\boldsymbol{\Gamma}\sqrt{\mathbf{D}} \otimes \mathbf{A}. \quad (5)$$

The next covariance structure was the heterogeneous compound symmetry (CSH). Different from AR1H, there is no autocorrelation, i.e., the correlation coefficient is not penalized by the distance between years:

$$\begin{bmatrix} \sigma_{g_1}^2 & \rho\sigma_{g_1}\sigma_{g_2} & \rho\sigma_{g_1}\sigma_{g_3} & \dots & \rho\sigma_{g_1}\sigma_{g_9} \\ \rho\sigma_{g_2}\sigma_{g_1} & \sigma_{g_2}^2 & \rho\sigma_{g_2}\sigma_{g_3} & \dots & \rho\sigma_{g_2}\sigma_{g_9} \\ \rho\sigma_{g_3}\sigma_{g_1} & \rho\sigma_{g_3}\sigma_{g_2} & \sigma_{g_3}^2 & \dots & \rho\sigma_{g_3}\sigma_{g_9} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_{g_9}\sigma_{g_1} & \rho\sigma_{g_9}\sigma_{g_2} & \rho\sigma_{g_9}\sigma_{g_3} & \dots & \sigma_{g_9}^2 \end{bmatrix}$$

$$\otimes\mathbf{A} = \{\sqrt{\mathbf{D}}[\mathbf{I}_m + \rho(\mathbf{J}_m - \mathbf{I}_m)]\sqrt{\mathbf{D}}\} \otimes \mathbf{A}. \quad (6)$$

After testing the aforementioned structures, we modelled $\boldsymbol{\Sigma}_g$ using multivariate structures. The first was the unstructured (US), which considers each year as a different trait:

$$\begin{bmatrix} \sigma_{g_1}^2 & \sigma_{g_{12}} & \sigma_{g_{13}} & \dots & \sigma_{g_{19}} \\ \sigma_{g_{21}} & \sigma_{g_2}^2 & \sigma_{g_{23}} & \dots & \sigma_{g_{29}} \\ \sigma_{g_{31}} & \sigma_{g_{32}} & \sigma_{g_3}^2 & \dots & \sigma_{g_{39}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{g_{91}} & \sigma_{g_{92}} & \sigma_{g_{93}} & \dots & \sigma_{g_9}^2 \end{bmatrix} \otimes \mathbf{A}, \quad (7)$$

where $\sigma_{jj'}$ is the covariance between the jth and the j'th years.

Then, we used the factor analytic covariance structure (FA) (Piepho, 1997; Smith et al., 2001). The core idea behind FA is to explore the covariance between years to reduce the dimensionality into latent variables, namely factors. Thus, it is an alternative to the often over-parameterized unstructured model (Meyer, 2009; Resende & Thompson, 2004). Despite being proposed to the multi-environment context, the application for repeated measures data is straightforward, even though the inferences are slightly different. We tested FA models of first-, second-, third- and fourth-order, i.e. FA1, FA2, FA3 and FA4. The general structure is

$$\begin{bmatrix} \sum_{k=1}^{K}\lambda_{1_k}^2+\psi_1 & \sum_{k=1}^{K}\lambda_{1_k}\lambda_{2_k} & \sum_{k=1}^{K}\lambda_{1_k}\lambda_{3_k} & \dots & \sum_{k=1}^{K}\lambda_{1_k}\lambda_{9_k} \\ \sum_{k=1}^{K}\lambda_{2_k}\lambda_{1_k} & \sum_{k=1}^{K}\lambda_{2_k}^2+\psi_2 & \sum_{k=1}^{K}\lambda_{2_k}\lambda_{3_k} & \dots & \sum_{k=1}^{K}\lambda_{2_k}\lambda_{9_k} \\ \sum_{k=1}^{K}\lambda_{3_k}\lambda_{1_k} & \sum_{k=1}^{K}\lambda_{3_k}\lambda_{2_k} & \sum_{k=1}^{K}\lambda_{3_k}^2+\psi_3 & \dots & \sum_{k=1}^{K}\lambda_{3_k}\lambda_{9_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^{K}\lambda_{9_k}\lambda_{1_k} & \sum_{k=1}^{K}\lambda_{9_k}\lambda_{2_k} & \sum_{k=1}^{K}\lambda_{9_k}\lambda_{3_k} & \dots & \sum_{k=1}^{K}\lambda_{9_k}^2+\psi_9 \end{bmatrix}$$

$$\otimes\mathbf{A} = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \otimes \mathbf{A}, \quad (8)$$

where $\boldsymbol{\Lambda}$ is the $m \times K$ matrix of loadings ($\boldsymbol{\Lambda} = \{\lambda_{j_k}\}$), where K is the number of factors and k represents each factor; and

$\boldsymbol{\Psi}$ is the $m \times m$ diagonal matrix of specific variances ($\boldsymbol{\Psi} = \{\psi_j\}$).

When using a factor analytic model with orders higher than one, plausible inferences are only possible after rotating the loadings and scores. This is because mathematical constraints imposed on $\boldsymbol{\Lambda}$ when $K > 1$ hinder its biological sense. More details about the rotation process are available in the specialized literature (Cullis et al., 2010, 2014; Smith & Cullis, 2018).

2.3.2 | Modelling the residual effects

We used three structures for modelling the residual effects, DIAG, AR1, and AR1H, which were described in Equations (3), (4), and (5), respectively. The only differences are the matrix after the direct product, which will be $\mathbf{I}_{n_j}$ instead of $\mathbf{A}$; and the variance component, which will be σ^2_e, not σ^2_g.

2.3.3 | Model selection

We used the Akaike information criterion [AIC; Akaike (1974)] to evaluate the goodness-of-fit of the tested models:

$$AIC = -2 \log L + 2t, \tag{9}$$

where L is the maximum point of the residual likelihood function and t is the number of estimated parameters. The lower the AIC, the more adequate the model.

In the FA models, we also used the average semivariances ratio (ASR) as an auxiliary parameter. Average semivariances were used by Piepho (2019) when proposing a generalized estimator for the coefficient of determination (R^2). We leveraged this metric to quantify the portion of the total covariance that is being explained by the FA model:

$$ASR = \frac{\frac{2}{m\times(m-1)} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} \frac{1}{2} \times (\sum_{k=1}^{K} \lambda^2_{j_k} + \sum_{k=1}^{K} \lambda^2_{j'_k}) - \sum_{k=1}^{K} \lambda_{j_k} \lambda_{j'_k}}{\frac{2}{m\times(m-1)} \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} \frac{1}{2} \times [(\sum_{k=1}^{K} \lambda^2_{j_k} + \psi_j) + (\sum_{k=1}^{K} \lambda^2_{j'_k} + \psi_{j'})] - \sum_{k=1}^{K} \lambda_{j_k} \lambda_{j'_k}} \times 100. \tag{10}$$

2.3.4 | Estimation of genetic and non-genetic parameters

We estimated the generalized heritability using the formula given by Cullis et al. (2006):

$$H^2 = 1 - \frac{\overline{V}(\Delta_g)}{2\sigma^2_g}, \tag{11}$$

where $\overline{V}(\Delta_g)$ is the average pairwise prediction error variance.

We also estimated the accuracy (Mrode, 2014):

$$r = \sqrt{1 - \frac{PEV}{\sigma^2_g}}, \tag{12}$$

where PEV is the prediction error variance.

We estimated these parameters in the first model and in the best-fitted model for comparison. Note that if the best-fitted model was heterogeneous, H^2 and r are year wise.

Using the US or FA structures, we estimated the pairwise genetic correlations (ρ_g), given by, respectively, Piepho (2018) and Cullis et al. (2014):

$$\rho_g = \frac{\sigma_{jj'}}{\sqrt{\sigma^2_j \sigma^2_{j'}}}, \tag{13}$$

$$\rho_g = \frac{\sum_{k=1}^{K} \lambda^2_{j_k} \lambda^2_{j'_k}}{\sqrt{\sigma^2_{g_j} \sigma^2_{g_{j'}}}}. \tag{14}$$

Besides H^2 and r, we compared the ranking of genotypes provided by each model using the Spearman correlation, given by

$$\rho_{sp} = 1 - \frac{6 \sum D}{n(n^2 - 1)}, \quad (15)$$

where D is the difference between ranks and n is the number of pairs of data.

We also estimated and compared the genetic gains (GG) of each model, given by:

$$GG = \frac{BV_s}{\mu}, \quad (16)$$

where BV_s is the mean of the selected genotypes' breeding value and μ is the phenotypic mean.

We performed all the analyses within the `R` software environment, version 4.2.1 (R Core Team, 2022), using `ASReml-R` (version 4.1) (Butler et al., 2018). We built the *heatmaps* using the `ComplexHeatmap` package, version 2.12.0 (Gu et al., 2016); and the other plots using the `ggplot2` package (Wickham, 2016). The Supporting Information section contains a file with the R scripts followed by the detailed results of each model, with different covariance structures for the genetic and residual effects. Phenotypic data, full `R` scripts and weather data throughout the years of evaluation (2010 to 2019) are available at https://github.com/saulo-chaves/Vcov-RM-grandiflorum.

3 | RESULTS

Considering the genealogy in the analyses through **A** improved models' fitness (e.g., the first model had an AIC of 12,312 without **A** and 12,290 with **A**). For that reason, we showed only results of models with **A** in Table 1. The AR1H structure was the most appropriate for explaining residual effects. The number of parameters associated with these effects is related to residual variances of the 9 years and the autocorrelation coefficient (Table 1). The FA3 was the best covariance structure for the genetic covariance matrix when accounting for both parsimony (AIC = 11,990) and explicative power (83.29%). Thus, there are 36 parameters related to the genetic effects: nine factorial loadings for the first, second and third factors and nine specific variances (Table 1). Note that Table 1 reports only parameters related to residual and genetic effects, but all models have one additional parameter related to the plot effects. The complete table containing models with and without **A** is in the Supporting Information (Table S2).

In the 13th model (M13), particularizing genetic and residual variance components for each year allows for visualizing the scale of these effects' influence on the phenotypic expression (Table 2). Consequently, parameters that are directly related to both genetic and residual effects, such as heritability and accuracy, are also particularized, which permits a more detailed study of the population behaviour over the years. Heritability ranged from 0.63 to 0.76, whereas accuracy varied from 0.78 to 0.87. These parameters had a moderate fluctuation throughout years, given the low difference between their mean (0.72 for heritability and 0.84 for accuracy) and their minimum and maximum values. Note from Tables 1 and 2 that the accuracy of the best-fitted model (M13) was higher than the accuracy of the first model (M1). The average fruit yield per year was increasing until the fifth year, the eighth year after planting, a pre-climax period. At this stage, *T. grandiflorum* reaches productive maturity, a behaviour that is usually observed at commercial orchards and was also perceived by Alves and Chaves (2023) and Alves et al. (2021).

The genetic correlation between years is useful to interpret the GYI, substituting the general interpretation provided by the CS structure, in M1 (σ^2_{gy}). In the correlation *heatmap* (Figure 3), if the GYI was low, one would expect correlations closer to 1. This did not happen in this study, as lower associations were detected. Thus, genotypes have different performances according to the environmental conditions to which they are subjected, which change on a yearly basis.

In the last instance, modelling provokes changes in the selection of genotypes. Besides the natural modifications in the breeding values from one model to another, more complex models provide extra resources that can aid in the selection. It is the case of the FA3 structure in the M13. In this model, the empirical breeding values are estimated using $BV_{ij} = \lambda_{j_1} f_{i_1} + \lambda_{j_2} f_{i_2} + \lambda_{j_3} f_{i_3} + \delta_{ij}$, in which δ_{ij} is the lack of fit effect. Cullis et al. (2010) recommend disregarding δ_{ij}, so the breeding values comprise only the common effects between years, i.e., the marginal breeding values. Following Cullis et al. (2014), one can make a second division, detaching the first factor ($\lambda_{j_1} f_{i_1}$) from the remainder. According to those authors, the first factor reflects genotypes' performance, whilst the others indicate stability. Based on these principles, Smith and Cullis (2018) proposed the Factor Analytic Selection Tools (FAST): overall performance (OP_i) and root mean square deviation ($RMSD_i$, representing stability), which we used herein to perform our selection. In our study, these parameters were calculated by the following formulae:

$$OP_i = \frac{1}{m} \sum_{j=1}^{m} \lambda_{j_1} f_{i_1}, \quad (17)$$

$$RMSD_i = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (\lambda_{j_2} f_{i_2} + \lambda_{j_3} f_{i_3})^2}, \quad (18)$$

the higher the OP_i and the lower the $RMSD_i$, the better the genotype. Bear in mind that we could only calculate these parameters after rotating the loadings and scores.

TABLE 1 Covariance structures for the residual (Σ_e) and genetic effects (Σ_g), logarithm of the likelihood function (log$L$), number of parameters associated with Σ_e (Σ_e par) and Σ_g (Σ_g par) matrices, Akaike information criterion (AIC) values, average semivariances ratio (ASR) (FA models only) and accuracy of the thirteen tested models in *Theobroma grandiflorum* repeated measures data analysis.

Model	$\Sigma_e^{\ddagger}$ structure	$\Sigma_g^{\ddagger}$ structure	AIC	ASR	Σ_e par	Σ_g par	log L	Accuracy
M1	IDV	CS	12,290	-	1	2	−6141	0.80
M2	IDV	DIAG	12,256	-	1	9	−6117	0.67
M3	IDV	AR1	12,297	-	1	2	−6145	0.92
M4	IDV	AR1H	12,248	-	1	10	−6112	0.71
M5	IDV	CSH	12,238	-	1	10	−6107	0.76
M6‡	IDV	US	-	-	-	-	-	-
M7	IDV	FA1	12,219	36.28	1	18	−6090	0.82
M8	IDV	FA2	12,213	64.64	1	27	−6084	0.83
M9	IDV	FA3	12,218	91.17	1	36	−6080	0.84
M10	IDV	FA4	12,220	93.44	1	45	−6077	0.79
M11	DIAG	FA3	12,006	83.43	9	36	−5967	0.84
M12	AR1	FA3	12,211	91.15	2	36	−6078	0.84
M13	**AR1H**	**FA3**	**119,90**	**83.29**	**10**	**36**	**−5957**	**0.84**

Note: Values in bold present the best-fitted model.
Abbreviations: IDV: variances identity; DIAG: diagonal; CS: compound symmetry; CSH: heterogeneous compound symmetry; AR1: first-order autoregressive; AR1H: heterogeneous first-order autoregressive; FA1: first-order factor analytic; FA2: second-order factor analytic; FA3: third-order factor analytic; FA4: fourth-order factor analytic; US: unstructured. ‡ Did not converge.

TABLE 2 Estimates of variance components and genetic and non-genetic parameters for fruit yield (kg plant^{-1}) estimated by the first (simplest and most parsimonious, M1 in Table 1) and 13th (best-fitted, M13 in Table 1) models in *Theobroma grandiflorum* repeated measures data analysis. Each measure is represented by a year ("Y").

	M1	M13								
Component/parameter[a]	Y01 to Y09	Y01	Y02	Y03	Y04	Y05	Y06	Y07	Y08	Y09
σ_g^2	19.39	7.28	16.79	12.80	34.10	37.91	88.59	34.18	84.80	11.49
σ_{gy}^2	14.33	-	-	-	-	-	-	-	-	-
σ_p^2	35.42	25.03								
σ_e^2	70.90	29.72	28.21	42.96	51.32	74.25	95.27	161.83	89.58	103.16
ρ	-	0.12								
H^2	0.78	0.66	0.69	0.74	0.74	0.76	0.75	0.63	0.76	0.74
r	0.80	0.79	0.83	0.85	0.85	0.86	0.86	0.78	0.87	0.83
μ	28.36	16.58	18.36	21.07	27.12	30.89	29.51	46.35	31.1	34.33

[a]σ_g^2 is the additive genetic variance, σ_{gy}^2 is the genotypes-by-years interaction, σ_p^2 is the variance due to the permanent plot effects, σ_e^2 is the residual variance, ρ is the autocorrelation, H^2 is the heritability, r is the accuracy and μ is the phenotypic mean.

To compare the rankings between the first model and the best-fitted model, we standardized OP_i and $RMSD_i$ and constructed a selection index, as performed by Chaves et al. (2023), giving equal weights for both. Figure 4 shows the differences between the models' ranks. Only half of the genotypes kept the same position from one model to another (ranking correlation = 0.528). Note that the selection is based solely on breeding values in M1, whereas the "OP-RMSD" selection index in M13 penalizes the genotypes' breeding value due to its instability.

We plotted OP_i and $RMSD_i$ in a scatter plot and based the selection following the criteria of $OP > 0$ and $RMSD_i < 1.8$. The selected genotypes are highlighted in Figure 5. These criteria can be adjusted according to breeders' objective. Only three genotypes were selected by both models: the parental clone 215, and the hybrids H117 and H143. Note that amongst the selected genotypes, there are both hybrids and parents. For instance, 215 is the parent of 286×215 and H143, and 186×434 is the parent of H117 and H120. This proves that these genotypes have the merits to serve as

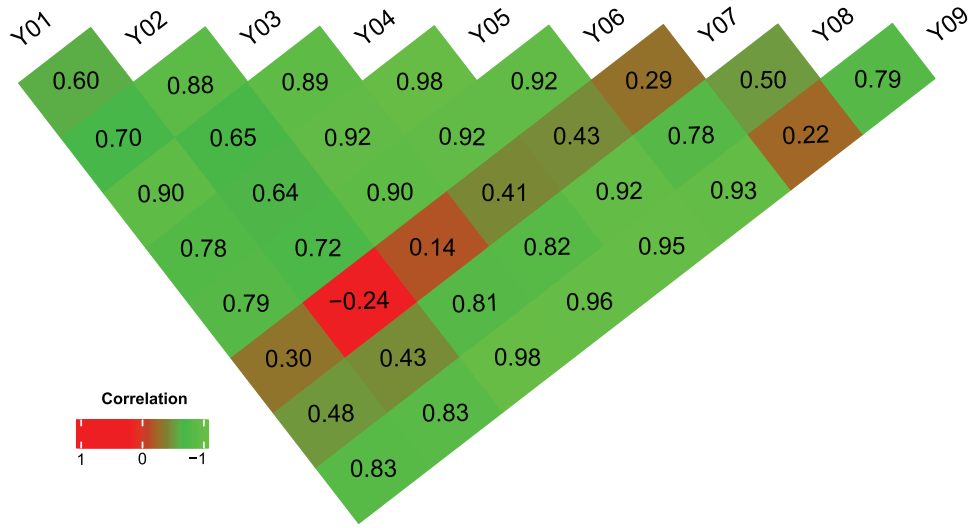

FIGURE 3 *Heatmap* representing the genetic pairwise correlations between years ("Y") in *Theobroma grandiflorum* repeated measures data analysis.

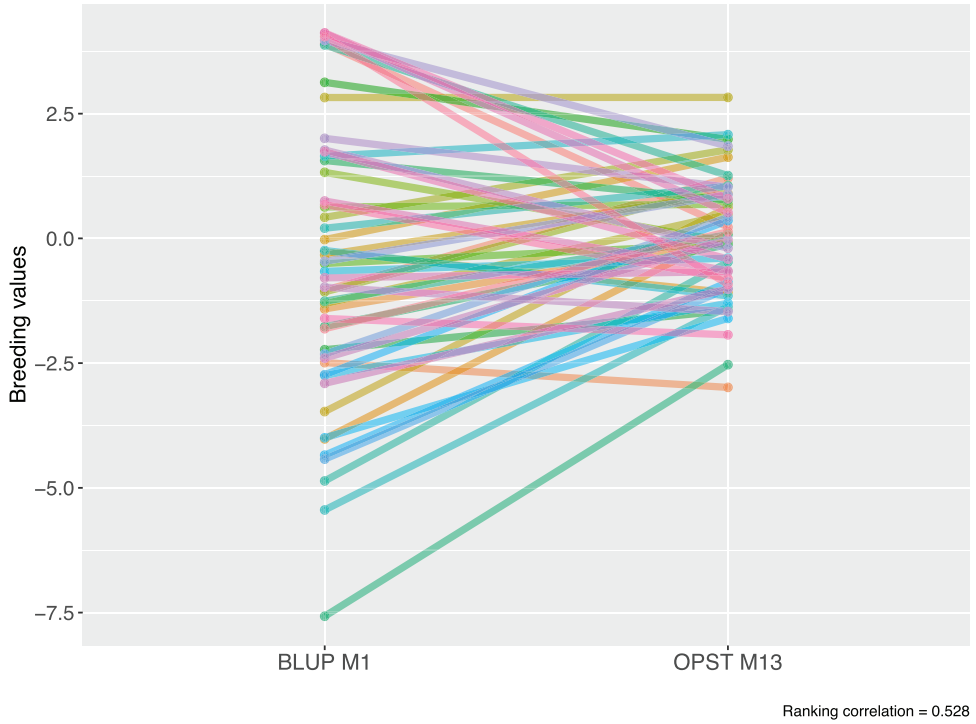

FIGURE 4 Changes between rankings of the first model (BLUP - M1; see Table 1) and 13th model, using the ''overall performance - root mean square deviation" index (OP-RMSD - M13; see Table 1), in *Theobroma grandiflorum* repeated measures data analysis.

parents for future crossings, offering alleles for both yield and stability.

4 | DISCUSSION

Fruit-bearing perennial plant trials have two common characteristics: large experimental area and repeated measures on the same plant throughout time. The effects caused by temporal environmental variations affect the genotypes differently. That depends on genetic factors, such as the genotype's genetic constitution; eco-physiological factors, e.g., the plant maturity stage; and macro- (e.g., climate) or micro-environmental (related to the plant's position in the trial) factors (Lahive et al., 2019). Bear in mind that these factors are all related to each other. Hence, the covariance

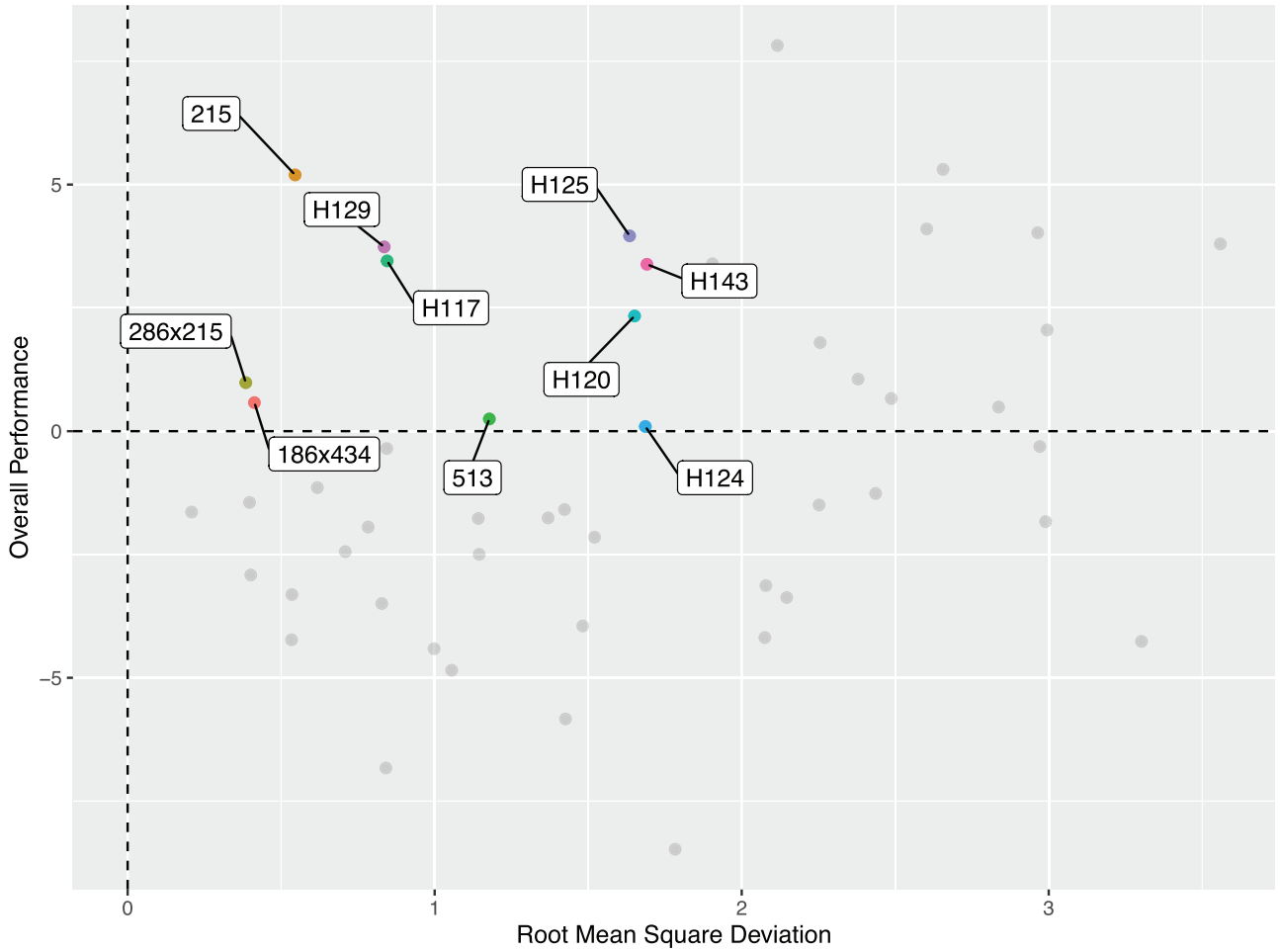

FIGURE 5 Overall performance and stability (root mean square deviation) of the *Theobroma grandiflorum* hybrids and their parents. The selected genotypes are highlighted. The lower the root mean square deviation and the higher the overall performance the better.

heterogeneity is the rule for these species and must be properly accounted for in the statistical model used to analyse the data. This study demonstrates this fact by testing several covariance structures for residual and genetic effects with different levels of complexity. The model with the best fit (lowest AIC) had the AR1H structure for the residual effects and FA3 for genetic effects. Using this model allowed for a thorough analysis of the phenotypic expression over the years. This study also shows that selection tools proposed by Smith and Cullis (2018) are suitable for selecting in the repeated measures' context. We selected the top ten genotypes using this method (Figure 5).

An information criterion such as AIC (Akaike, 1974) is crucial for model selection. In this study, there was an oscillation in the accuracy when different structures were tested in the models' effects. Nevertheless, once we identified the best structure for both genetic and residual effects, the second-highest accuracy among the tested models was reached. This stresses that the accuracy may be inflated in models with higher AIC, and so, it cannot serve as a criterion for selection (Resende & Alves, 2020). In the FA context, AIC may not reflect the most suitable model (Smith et al., 2015; Zhang et al., 2020). Thus, an auxiliary parameter such as the average semivariances ratio can aid the model selection. The idea is to select a model with highly explained covariance, yet parsimonious. Here, the chosen model had an explicative power of 83.29%.

AR1H was the most suitable covariance structure for residual effects. The AR1H covariance structure is often used in spatial analyses, considering that as the distance between two plots increases, the correlation between them decreases (Chavarría-Perez et al., 2020; Gilmour et al., 1997). This concept can be applied in repeated measures context, i.e., the correlation between two measures decreases as the time between them increases (Piepho & Eckl, 2014; Verbyla et al., 2021; Wolfinger, 1996). For *T. grandiflorum* and other fruit-bearing species, this occurs not only because of environmental factors but also because of the plants' maturity stage. The efficacy of the FA structure for modelling the genotypic effects has already been proved for multi-environment trials (Burgueño et al., 2012; Krause et al., 2020; Dias et al., 2018). On a smaller scale, it is also being employed in the context of

perennial plants' repeated measures data (Faveri et al., 2022; Verbyla et al., 2021). FA considers each measure as a distinct trait, allowing for particularized estimations. Nonetheless, this structure is more parsimonious than a non-structured model, since it considers the covariance between measures (or traits) to reduce the dimensionality to few latent variables named factors. Note that when the unstructured covariance matrix was used for modelling genetic effects, the model did not reach convergence (M6 at Table 1). This is expected given the high dimensionality of the data.

Only by structuring the genotypic covariance matrix in the best-fitted model, we could visualize the variation in the magnitude of the genetic variance throughout years. The different maturity stages may be linked to this behaviour. This is related to differential gene expression, which is also connected to environmental factors (Tahi et al., 2019). The fluctuation of residual effects is linked to the temporary environmental effects, such as weather (Cilas & Bastide, 2020; Farrell et al., 2018; Lahive et al., 2019). Note by the low autocorrelation value ($\rho = 0.12$, Table 2) that the years have little association. In other words, there are significant changes in the temporary environmental effects on a yearly basis, which justifies the heterogeneity. Since the genetic and residual effects were particularized, so were heritabilities and accuracies. These parameters had high values if compared to other *T. grandiflorum* studies (Alves et al., 2021; Chaves et al., 2021), which denote the experimental precision and guarantee the reliability of the selection.

T. grandiflorum has a high intra-population divergence, promoted by self-incompatibility (Alves et al., 2007). Furthermore, it has a long juvenile period ($\approx$ 3 years), something common in fruit-bearing perennial species (Rai & Shekhawat, 2014). After the juvenile period, the plant enters a pre-climax productive stage, which lingers until the eighth year. During this period, the plant adapts to the soil and climate conditions of the orchard while keeping the maturing process, characterizing an eco-physiological phenomenon. In other words, there is a differential gene expression throughout the plant's life, which is a source of genetic heterogeneity (Friedman, 2020). It is noteworthy that this heterogeneity can be greater in a multi-environment context, due to the influence exerted by the genotype-by-environment interaction (GEI). In this scenario, both GYI and GEI play an important role in phenotypic expression. In *T. grandiflorum* breeding, selection candidates are evaluated in multi-environment trials (METs) only in the last step. This study, for example, refers to a trial in an intermediate phase of the breeding programme, in which the genotypes were evaluated in a single environment, giving a major focus on the multi-years scenario. The intensity of the GYI observed in this study may justify the usage of METs in earlier phases, so genotypes with both temporal and geographical stability can be prioritized. Still, the high heritabilities and accuracies observed herein (Table 2) provide the required reliability for the selection and conclusions of this study.

The genetic correlations between years indicate the existence of GYI, i.e., genotypes have different performances throughout years. This information is fundamental for prioritizing the selection of resilient genotypes, capable of enduring the increasing climatic changes, as well as variations in management (Ceccarelli et al., 2021; Gateau-Rey et al., 2018). In the *T. grandiflorum* context, breeders may also keep in mind the selection of genotypes that can bear the intra- and interspecific competition, as most orchards in the Brazilian Amazon are agroforestry systems (Alves et al., 2020, 2021).

The existence of GYI provides two reasons why not to use homogeneous models like M1: (i) the variance homogeneity denotes a constant and stable effect over years, which is hardly true, given the reasons stated previously in this study; and (ii) to base the selection only on the performance per se may not be the best choice. Recall that *T. grandiflorum* can keep producing for several decades, so stability is as important as performance. In this context, selection tools proposed by Smith and Cullis (2018) for the FA scenario provide an overview of both performance (OP_i) and stability ($RMSD_i$). In our study, we showed two ways of using these parameters for aiding the selection: building selection indexes or scatter plots. None of these methods are static, i.e., breeders may change the weights in the index or the criteria for selection in the scatter plot according to their objective (Smith & Cullis, 2018). To the best of our knowledge, no studies have employed this method in the repeated measures' context. Following the same criteria for both M1 and M13, i.e., considering only performance, there is only one difference between the top 10: the presence of H129 substituting H143. In this scenario, the best-fitted model provides a gain 14% higher (17.3%) than the first model (14.9%). This increase is usually observed when accounting for the genetic and residual heterogeneity via modelling (Souza et al., 2021). Considering both performance and stability for selection ($OP_i > 0$ and $RMSD_i < 1.8$), a genetic gain of 10.4% was reached. This lower value may not mislead the reader. See from Equation 16 that gains are based solely on breeding values, disregarding the stability parameter. In other words, the selected genotypes in Figure 5 have more evidence of holding alleles for both stability and fruit yield. Analysing this fact concomitantly with the best-fitted model's higher accuracy and higher gains when we adopt the same selection criteria, it is plausible to assume that M13 provides a more reliable selection.

The results of this study offer new possibilities for *T. grandiflorum* breeding. The first major recommendation is to essay genotypes in METs at earlier phases, allowing the selection of candidates with broader stability. In fact, employing METs precociously would diminish the chances of discarding geographically stable genotypes. A second recommendation refers to the number of years to evaluate before selecting. Here, we observed that the fourth year has a high genetic correlation with the following years, opening the possibility of an earlier selection, as suggested by a previous study

(Chaves et al., 2022). Thus, the selection would be performed in the first year after the pre-climax period has passed (Dias & Kageyama, 1998). This would significantly raise the selection gains by decreasing the time between cycles. The final recommendation is valid for any perennial fruit-bearing species: it is important to hand over modelling strategies when dealing with repeated measures data. The testing phase is crucial since the best structures for the dataset studied herein may not be the best for a distinct dataset, in a completely different context. Allying sophisticated statistical genetic methods with classical breeding strategies, and employing the new molecular tools that are being developed for the species (Falcão et al., 2022; Mournet et al., 2020; Niu et al., 2019; Santos et al., 2022), will guide *T. grandiflorum* breeding programme towards the future.

5 | CONCLUSION

Modelling the covariance structures and identifying the best-fitted model proved to be fundamental for the evaluation and selection of *T. grandiflorum* genotypes. The best-fitted model generated more reliable results, both in the estimation of variance components and breeding values, enabling the improvement of one of the most important stages of the breeding programme: the genetic evaluation. In Model 13, matrices of residual and genetic effects were modelled by the first-order heterogeneous autoregressive structures (AR1H) and third-order factor analytic (FA3), respectively. Using the FAST in this model, the genotypes 215, 286×215, 186×434, 513, H129, H117, H125, H143, H120, and H124 were selected, providing a genetic gain of 10.42%.

AUTHOR CONTRIBUTIONS

All authors contributed to the study and conception and design: **Saulo Chaves**: Formal analysis; Writing – Original draft; Visualization. **Rodrigo Alves**: Conceptualization; Methodology; Writing – Review & Editing. **Luiz Dias**: Supervision; Resources; Writing – Review & Editing. **Rafael Alves**: Data curation; Validation; Writing – Review & Editing. **Kaio Dias**: Validation; Methodology; Writing – Review. **Jeniffer Evangelista**: Conceptualization; Writing – Review & Editing.

ACKNOWLEDGEMENTS

We express our gratitude to all field workers who work in the *Theobroma grandifloum* breeding programme of Embrapa Amazônia Oriental. Without them, nothing would be possible. We acknowledge the financial support from Minas Gerais State Agency for Research and Development (FAPEMIG), Brazilian National Council for Scientific and Technological Development (CNPq) and Coordination for the Improvement of Higher Educational Personnel (CAPES) - Finance Code 001.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest

ORCID

Saulo F. S. Chaves https://orcid.org/0000-0002-0694-1798
Rodrigo S. Alves https://orcid.org/0000-0002-3038-6210
Luiz A. S. Dias https://orcid.org/0000-0002-1828-6938
Rafael M. Alves https://orcid.org/0000-0002-9826-4690
Kaio O. G. Dias https://orcid.org/0000-0002-9171-1021
Jeniffer S. P. C. Evangelista https://orcid.org/0000-0001-8820-0434

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alves, R. M., & Chaves, S. F. S (2020). BRS careca, BRS fartura, BRS duquesa, BRS curinga, and BRS golias: New cupuassu tree cultivars. *Crop Breeding and Applied Biotechnology*, *20*(4), e342920413. https://doi.org/10.1590/1984-70332020v20n4c66

Alves, R. M., & Chaves, S. F. S (2023). Selection of *Theobroma grandiflorum* clones adapted to agroforestry systems using an additive index. *Acta Scientiarum. Agronomy*, *45*(1), e57519. https://doi.org/10.4025/actasciagron.v45i1.57519

Alves, R. M., Chaves, S. F. S., Alves, R. S., Santos, T. G., Araújo, D. G., & Resende, M. D. V. (2021). Cupuaçu tree genotype selection for an agroforestry system environment in the Amazon. *Pesquisa Agropecuária Brasileira*, *56*, e02139. https://doi.org/10.1590/S1678-3921.pab2021.v56.02139

Alves, R. M., Chaves, S. F. S, Gama, M. A. P, Pedroza Neto, J. L., & Santos, T. G. (2020). Simultaneous selection of cupuassu tree and Brazilian mahogany genotypes in an agroforestry system in Pará state, Brazil. *Acta Amazonica*, *50*(3), 183–191. https://doi.org/10.1590/1809-4392202000711

Alves, R. M., Sebbenn, A. M., Artero, A. S., Clement, C., & Figueira, A. (2007). High levels of genetic divergence and inbreeding in populations of cupuassu (*Theobroma grandiflorum*). *Tree Genetics & Genomes*, *3*(4), 289–298. https://doi.org/10.1007/s11295-006-0066-9

Burgueño, J., Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). *ASReml-R reference manual version 4*. VSN International.

Ceccarelli, V., Fremout, T., Zavaleta, D., Lastra, S., Imán Correa, S., Arévalo-Gardini, E., Rodriguez, C. A., Hilacondo, W. C., & Thomas, E. (2021). Climate change impact on cultivated and wild cacao in Peru and the search of climate change-tolerant genotypes. *Diversity and Distributions*, *27*(8), 1462–1476. https://doi.org/10.1111/ddi.13294

Chavarría-Perez, L. M., Giordani, W., Dias, K. O. G, Costa, Z. P., Ribeiro, C. A. M., Benedetti, A. R., Cauz-Santos, L. A., Pereira, G. S., Rosa, J. R. B. F., Garcia, A. A. F, & Vieira, M. L. C. (2020). Improving yield and fruit quality traits in sweet passion fruit: Evidence for genotype by environment interaction and selection of promising genotypes. *PLoS ONE*, *15*(5), e0232818. https://doi.org/10.1371/journal.pone.0232818

Chaves, S. F. S, Alves, R. M., Alves, R. S., Sebbenn, A. M., Resende, M. D. V, & Dias, L. A. S (2021). *Theobroma grandiflorum* breeding optimization based on repeatability, stability and adaptability information. *Euphytica*, *217*(12), 211. https://doi.org/10.1007/s10681-021-02944-3

Chaves, S. F. S, Dias, L. A. S, Alves, R. S., Alves, R. M., Evangelista, J. S. P. C., & Dias, K. O. G (2022). Leveraging multi-harvest data for increasing genetic gains per unit of time for fruit yield and resistance to witches' broom in *Theobroma grandiflorum*. *Euphytica*, *218*(12), 171. https://doi.org/10.1007/s10681-022-03126-5

Chaves, S. F. S., Evangelista, J. S. P. C., Trindade, R. S., Dias, L. A. S., Guimarães, P. E., Guimarães, L. J. M., Alves, R. S., Bhering, L. L., & Dias, K. O. G. (2023). Employing factor analytic tools for selecting high– performance and stable tropical maize hybrids. *Crop Science*, *63*(3), 1114–1125. https://doi.org/10.1002/csc2.20911

Cilas, C., & Bastide, P. (2020). Challenges to cocoa production in the face of climate change and the spread of pests and diseases. *Agronomy*, *10*(9), 1232. https://doi.org/10.3390/agronomy10091232

Cullis, B. R., Jefferson, P., Thompson, R., & Smith, A. B. (2014). Factor analytic and reduced animal models for the investigation of additive genotype-by-environment interaction in outcrossing plant species with application to a *Pinus radiata* breeding programme. *Theoretical and Applied Genetics*, *127*(10), 2193–2210. https://doi.org/10.1007/s00122-014-2373-0

Cullis, B. R., Smith, A. B., Beeck, C. P., & Cowling, W. A. (2010). Analysis of yield and oil from a series of canola breeding trials. Part II. exploring variety by environment interaction using factor analysis. *Genome*, *53*(11), 1002–1016. https://doi.org/10.1139/G10-080

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(4), 381–393. https://doi.org/10.1198/108571106X154443

Dias, K. O. G., Gezan, S. A., Guimarães, C. T., Parentoni, S. N., Guimarães, P. E. O., Carneiro, N. P., Portugal, A. F., Bastos, E. A., Cardoso, M. J., Anoni, C. O., Magalhães, J. V., Souza, J. C., Guimarães, L. J. M., & Pastina, M. M. (2018). Estimating genotype ×environment interaction for and genetic correlations among drought tolerance traits in maize via factor analytic multiplicative mixed models. *Crop Science*, *58*(1), 72–83. https://doi.org/10.2135/cropsci2016.07.0566

Dias, L. A. S., & Kageyama, P. Y. (1998). Repeatability and minimum harvest period of cacao (*Theobroma cacao* l.) in southern Bahia. *Euphytica*, *102*, 29–35. https://doi.org/10.1023/A:1018373211196

Falcão, L. L., Silva-Werneck, J. O., Albuquerque, P. S. B., Alves, R. M., Grynberg, P., Togawa, R. C., Costa, M. M. d. C., Brigido, M. M., & Marcellino, L. H. (2022). Comparative transcriptomics of cupuassu (*Theobroma grandiflorum*) offers insights into the early defense mechanism to *Moniliophthora perniciosa*, the causal agent of witches' broom disease. *Journal of Plant Interactions*, *17*(1), 991–1005. https://doi.org/10.1080/17429145.2022.2144650

Farrell, A. D., Rhiney, K., Eitzinger, A., & Umaharan, P. (2018). Climate adaptation in a minor crop species: Is the cocoa breeding network prepared for climate change? *Agroecology and Sustainable Food Systems*, *42*(7), 812–833. https://doi.org/10.1080/21683565.2018.1448924

Faveri, J., Verbyla, A. P., Cullis, B. R., Pitchford, W. S., & Thompson, R. (2017). Residual variance–covariance modelling in analysis of multivariate data from variety selection trials. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(1), 1–22. https://doi.org/10.1007/s13253-016-0267-0

Faveri, J., Verbyla, A. P., Pitchford, W. S., Venkatanagappa, S., & Cullis, B. R. (2015). Statistical methods for analysis of multi-harvest data from perennial pasture variety selection trials. *Crop and Pasture Science*, *66*(9), 947. https://doi.org/10.1071/CP14312

Faveri, J., Verbyla, A. P., & Rebetzke, G. (2022). Random regression models for multi-environment, multi-time data from crop breeding selection trials. *Crop and Pasture Science*. Advanced online publication. https://doi.org/10.1071/CP21732

Friedman, J. (2020). The evolution of annual and perennial plant life histories: Ecological correlates and genetic mechanisms. *Annual Review of Ecology, Evolution, and Systematics*, *51*(1), 461–461. https://doi.org/10.1146/annurev-ecolsys-110218-024638

Gateau-Rey, L., Tanner, E. V. J., Rapidel, B., Marelli, J.-P., & Royaert, S. (2018). Climate change could threaten cocoa production: Effects of 2015-16 el niño-related drought on cocoa agroforests in Bahia, Brazil. *PLoS ONE*, *13*(7), e0200454. https://doi.org/10.1371/journal.pone.0200454

Gilmour, A. R., Cullis, B. R., & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*(3), 269–293. https://doi.org/10.2307/1400446

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics (Oxford, England)*, *32*(18), 2847–2849. https://doi.org/10.1093/bioinformatics/btw313

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics. Journal of the International Biometric Society*, *31*(2), 423. https://doi.org/10.2307/2529430

Krause, M. D., Dias, K. O. G., Santos, J., Oliveira, A. A., Guimarães, L. J. M., Pastina, M. M., Margarido, G. R. A., & Garcia, A. A. F. (2020). Boosting predictive ability of tropical maize hybrids via genotype-by-environment interaction under multivariate GBLUP models. *Crop Science*, *60*(6), 3049–3065. https://doi.org/10.1002/csc2.20253

Lahive, F., Hadley, P., & Daymond, A. J. (2019). The physiological responses of cacao to the environment and the implications for climate change resilience. A review. *Agronomy for Sustainable Development*, *39*(1), 5. https://doi.org/10.1007/s13593-018-0552-0

Meyer, K. (2009). Factor-analytic models for genotype × environment type problems and structured covariance matrices. *Genetics Selection Evolution*, *41*(1), 21. https://doi.org/10.1186/1297-9686-41-21

Moraes, J. R. S C., Rolim, G. S., Martorano, L. G., Aparecido, L. E. O., Oliveira, M. S. P., & Farias Neto, J. T. (2020). Agrometeorological models to forecast açaí (*Euterpe oleracea* Mart.) yield in the eastern Amazon. *Journal of the Science of Food and Agriculture*, *100*(4), 1558–1558. https://doi.org/10.1002/jsfa.10164

Mournet, P., Albuquerque, P. S. B., Alves, R. M., Silva-Werneck, J. O., Rivallan, R., Marcellino, L. H., & Clément, D. (2020). A reference high-density genetic map of *Theobroma grandiflorum* (Willd. ex Spreng) and QTL detection for resistance to witches' broom disease (*Moniliophthora perniciosa*). *Tree Genetics & Genomes*, *16*(6), 89. https://doi.org/10.1007/s11295-020-01479-3

Mrode, R. A. (2014). *Linear models for the prediction of animal breeding values* (3rd ed.). Wallingford, UK: CABI.

Niu, Y.-F., Ni, S.-B., & Liu, J. (2019). The complete chloroplast genome of *Theobroma grandiflorum*, an important tropical crop.

Mitochondrial DNA Part B, *4*(2), 4157–4158. https://doi.org/10.1080/23802359.2019.1693291

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545

Pereira, A. L. F., Abreu, V. K. G., & Rodrigues, S. (2018). Cupuassu'*Theobroma grandiflorum*. In S. Rodrigues, E. O. Silva, & E. S. Brito (Eds.), *Exotic fruits* (pp. 159–162). Academic Press. https://doi.org/10.1016/B978-0-12-803138-4.00021-6

Piepho, H.-P. (1997). Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics. Journal of the International Biometric Society*, *53*(2), 761–761. https://doi.org/10.2307/2533976

Piepho, H.-P. (2018). Allowing for the structure of a designed experiment when estimating and testing trait correlations. *The Journal of Agricultural Science*, *156*(1), 59–70. https://doi.org/10.1017/S0021859618000059

Piepho, H.-P. (2019). A coefficient of determination (R^2) for generalized linear mixed models. *Biometrical Journal*, *61*(4), 860–872. https://doi.org/10.1002/bimj.201800270

Piepho, H.-P., & Eckl, T. (2014). Analysis of series of variety trials with perennial crops. *Grass and Forage Science*, *69*(3), 431–440. https://doi.org/10.1111/gfs.12054

Rai, M. K., & Shekhawat, N. S. (2014). Recent advances in genetic engineering for improvement of fruit crops. *Plant Cell, Tissue and Organ Culture*, *116*(1), 1–15. https://doi.org/10.1007/s11240-013-0389-9

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

Resende, M. D. V., & Alves, R. S. (2020). Linear, generalized, hierarchical, bayesian and random regression mixed models in genetic/genomics in plant breeding. *Functional Plant Breeding Journal*, *2*(2), 1–31. https://doi.org/10.35418/2526-4117/v2n2a1

Resende, M. D. V., & Thompson, R. (2004). Factor analytic multiplicative mixed models in the analysis of multiple experiments. *Revista de Matemática e Estatística*, *22*(2), 31–52.

Santos, L. F., Silva, R. J. S., Falcão, L. L., Alves, R. M., Marcellino, L. H., & Micheli, F. (2022). Cupuassu (*Theobroma grandiflorum* [Willd. ex Sprengel] Schumann) fruit development: Key genes involved in primary metabolism and stress response. *Agronomy*, *12*(4), 763. https://doi.org/10.3390/agronomy12040763

Smith, A. B., & Cullis, B. R. (2018). Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica*, *214*(8), 143. https://doi.org/10.1007/s10681-018-2220-5

Smith, A. B., Cullis, B. R., & Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, *57*(4), 1138–1138. https://doi.org/10.1111/j.0006-341X.2001.01138.x

Smith, A. B., Ganesalingam, A., Kuchel, H., & Cullis, B. R. (2015). Factor analytic mixed models for the provision of grower information from national crop variety testing programmes. *Theoretical and Applied Genetics*, *128*(1), 55–72. https://doi.org/10.1007/s00122-014-2412-x

Smith, A. B., Stringer, J. K., Wei, X., & Cullis, B. R. (2007). Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. *Euphytica*, *157*(1), 253–266. https://doi.org/10.1007/s10681-007-9418-2

Souza, V. F., Ribeiro, P. C. O., Vieira Júnior, I. C., Oliveira, I. C. M., Damasceno, C. M. B., Schaffert, R. E., Parrella, R. A. C., Dias, K. O. G., & Pastina, M. M. (2021). Exploring genotype × environment interaction in sweet sorghum under tropical environments. *Agronomy Journal*, *113*(4), 3005–3018. https://doi.org/10.1002/agj2.20696

Stringer, J. K., Atkin, F. C., & Gezan, S. A. (2017). Statistical approaches in plant breeding: Maximising the use of the genetic information. In *Genetic improvement of tropical crops* pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-319-59819-2_1

Tahi, M., Trebissou, C., Ribeyre, F., Guiraud, B. S., Pokou, D. N., & Cilas, C. (2019). Variation in yield over time in a cacao factorial mating design: Changes in heritability and longitudinal data analyses over 13 consecutive years. *Euphytica*, *215*(6), 106. https://doi.org/10.1007/s10681-019-2429-y

Venturieri, G. A. (2011). Flowering levels, harvest season and yields of cupuassu *Theobroma grandiflorum*. *Acta Amazonica*, *41*, 143–152. https://doi.org/10.1590/S0044-59672011000100017

Verbyla, A. P., Faveri, J., Deery, D. M., & Rebetzke, G. J. (2021). Modelling temporal genetic and spatio-temporal residual effects for high-throughput phenotyping data. *Australian & New Zealand Journal of Statistics*, *63*(2), 284–308. https://doi.org/10.1111/anzs.12336

Wickham, H. (2016). `ggplot2`: *Elegant graphics for data analysis*. Springer, 2 edition.

Wolak, M. E. (2012). `nadiv`: An `R` package to create relatedness matrices for estimating non-additive genetic variances in animal models. *Methods in Ecology and Evolution*, *3*(5), 792–796.

Wolfinger, R. D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*(2), 205–230. https://doi.org/10.2307/1400366

Zhang, R., Han, D., & Hu, X. (2020). Analyzing the performance of corn in China using a factor-analytic variance-covariance structure with multiple factors. *Crop Science*, *60*(1), 190–201. https://doi.org/10.1002/csc2.20090

SUPPORTING INFORMATION

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

How to cite this article: Chaves, S. F. S., Alves, R. S., Dias, L. A. S., Alves, R. M., Dias, K. O. G., & Evangelista, J. S. P. C. (2023). Analysis of repeated measures data through mixed models: An application in *Theobroma grandiflorum* breeding. *Crop Science*, *63,* 2131–2144. https://doi.org/10.1002/csc2.20995