Do Spatial Designs Outperform Classic Experimental Designs?

Raegan HOEFLER, Pablo GONZÁLEZ-BARRIOS, Madhav BHATTA,
Jose A. R. NUNES, Ines BERRO, Rafael S. NALIN, Alejandra BORGES,
Eduardo COVARRUBIAS, Luis DIAZ-GARCIA, Martin QUINCKE, and
Lucia GUTIERREZ

Controlling spatial variation in agricultural field trials is the most important step to compare treatments efficiently and accurately. Spatial variability can be controlled at the experimental design level with the assignment of treatments to experimental units and at the modeling level with the use of spatial corrections and other modeling strategies. The goal of this study was to compare the efficiency of methods used to control spatial variation in a wide range of scenarios using a simulation approach based on real wheat data. Specifically, classic and spatial experimental designs with and without a two-dimensional autoregressive spatial correction were evaluated in scenarios that include differing experimental unit sizes, experiment sizes, relationships among genotypes, genotype by environment interaction levels, and trait heritabilities. Fully replicated designs outperformed partially and unreplicated designs in terms of accuracy; the alpha-lattice incomplete block design was best in all scenarios of the medium-sized experiments. However, in terms of response to selection, partially replicated experiments that evaluate large population sizes were superior in most scenarios. The AR1 × AR1 spatial correction had little benefit in most scenarios except for the medium-sized experiments with the largest experimental unit size and low GE. Overall, the results from this study provide a guide to researchers designing and analyzing large field experiments.
Supplementary materials accompanying this paper appear online.

Key Words: Experimental design; Autoregressive process; Prediction accuracy; Response to selection; Spatial correction; Randomization-based experimental designs.

Raegan Hoefler and Pablo Gonzalez-Barrios have contributed equally to this work.

Raegan Hoefler, Pablo González-Barrios, Madhav Bhatta, Jose A. R. Nunes, Ines Berro, and Lucia Gutierrez (✉) Department of Agronomy, University of Wisconsin–Madison, 1575 Linden Dr., Madison, WI 53706, USA (E-mail: *gutierrezcha@wisc.edu*). Pablo Gonzalez-Barrios, Ines Berro, Alejandra Borges, and Lucia Gutierrez Statistics Department, Facultad de Agronomía, Univesidad de la República, Garzón 780, Montevideo, Uruguay. José A. R. Nunes Department of Biology, Federal University of Lavras, Lavras, Minas Gerais State, Brazil. Rafael S. Nalin Department of Genetics, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, São Paulo 131418-900, Brazil. Eduardo Covarrubias CGIAR Excellence in Breeding Platform (EiB), El Batan, Mexico and International Maize and Wheat Improvement Center (CIMMYT), El Batan, Mexico. Luis Diaz-Garcia Instituto Nacional de Investigaciones Forestales, Agricolas y Pecuarias, 20676 Aguascalientes, Mexico. Martin Quincke Programa Nacional de Investigación Cultivos de Secano, Instituto Nacional de Investigación Agropecuaria, Est. Exp. La Estanzuela, 70000 Colonia, Uruguay.

Abbreviations

ALPHA	Alpha-lattice incomplete block design
A-PREP	Augmented partially replicated experimental design
$AR1 \times AR1$	Two-dimensional autoregressive process
AYT	Advanced yield trial
COR	Pearson correlation coefficient between predicted and true genotypic values
CRD	Completely randomized design
EYT	Elite yield trial
GE	Genotype by environment interaction
IBD	Incomplete block design
LV	Linear variance
MET	Multi-environment trial
NSC	No spatial correction
PREP	Partially replicated experimental design
PYT	Preliminary yield trial
RBE	Randomization-based experimental design
RCBD	Randomized complete block design
R-CD	Row–column alpha-lattice design
SP	Spatial experimental design
UNREP	Unreplicated experimental design

1. INTRODUCTION

The importance of controlling spatial heterogeneity in agricultural field trials to efficiently and accurately estimate treatment effects has been widely understood for decades (Brownie et al. 1993; Casler 2015; Fisher 1935; Smith et al. 2005). Spatial heterogeneity in a field due to fertility, moisture, slope, shade, or management practices can bias the estimation of treatment effects (Grondona et al. 1996), making it difficult to accurately differentiate between treatments (Zystro et al. 2019). Yet, it is often difficult to predict patterns of spatial variation, even with years of experimentation in a field (Casler 2015). Fisher's (1926; 1935) experimental design principles provide some measure of protection against spatial variation, but many agree that additional levels of spatial control are beneficial (Borges et al. 2019; Cullis et al. 2006; John and Eccleston 1986; Papadakis 1937; Piepho et al. 2013; Piepho and Williams 2010; Stefanova et al. 2009; Wilkinson et al. 1983; Zimmerman and Harville 1991).

Spatial control can be approached in both the design and the analysis phases of experimentation. Randomization-based experimental (RBE) designs with some level of spatial control include among others, randomized complete block (RCBD, Fisher 1926), incomplete blocks (Yates 1936) including alpha-lattice (ALPHA, Cochran and Cox 1957; Patterson and Williams 1976), row–column (R-CD, Fisher 1926), augmented (Federer 1956; Federer and Raghavarao 1975), and partially replicated (PREP, Cullis et al. 2006; Moehring et al. 2014; Williams et al. 2011) experimental designs. A class of experimental designs similar to these

are spatial (SP) designs, which have additional restrictions on their randomization based on a dependence correlation structure for the field that is either known or assumed a priori (Coombes 2002; Eccleston and Chan 1998; Martin and Eccleston 1991; Williams and Piepho 2013). The SP designs generally follow a R-CD or an ALPHA design and are often constructed using linear variance (LV, Piepho and Williams 2010; Williams 1986; Williams et al. 2006; Williams and Piepho 2013) or autoregressive (AR1) models (Coombes 2002; Eccleston and Chan 1998; Martin and Eccleston 1991).

The first and most basic attempt at controlling spatial variability in field trials was through the use of completely randomized designs (CRD) which rely strictly on randomization to provide a valid estimate of the experimental error variance (Fisher 1926) and unbiased estimates of treatment effects (Casler 2015) and comparisons (Piepho et al. 2003). Therefore, a CRD assumes that all extraneous variables affect all experimental units equally. As an improvement, the RCBD restricts the randomization of a complete set of treatments to within a block to control for extraneous variables like global spatial variation and management practices that may affect blocks differently (Brownie et al. 1993). This further increases the precision of treatment effects and reduces experimental error (Mead 1997). Because of the effectiveness and simplicity of this blocking scheme, the RCBD is the most commonly used experimental design in agricultural experimentation (Casler 2015; Van Es et al. 2007). However, some researchers argue that other ways of controlling this variability are needed because within block variation is common in large experiments (Grondona et al. 1996). The class of incomplete block designs (IBDs) was created to allow for smaller block sizes to mitigate this issue (Yates 1936). One specific IBD is the resolvable alpha-lattice design (ALPHA, Patterson and Williams 1976; Williams et al. 2002) which is characterized by having its treatments cycled through the incomplete blocks so that pairs of treatments within an incomplete block occur a set number of times. Common schemes include having all pairs occur the same number of times or having pairs occur either never or once (ALPHA$_{(0,1)}$). Many studies illustrate the effectiveness of the ALPHA design (Gonçalves et al. 2010; Masood et al. 2008; Patterson and Hunter 1983; White et al. 1996; Williams and John (1999)). Furthermore, Borges et al. (2019) showed that the ALPHA design was superior to CRDs, RCBDs, and partially replicated designs at controlling spatial heterogeneity within a single testing site for three different sized experiments with both high and low spatial variation. Row–column designs, including the row and column incomplete blocks or alpha designs (John and Eccleston 1986; Williams 1986; Williams and John 1989), utilize blocking in two dimensions to control spatial variation present in both directions and to lessen the effects of possibly blocking in the wrong one-dimensional direction (Zystro et al. 2019). Williams and Piepho (2013) found that the R-CDs were more efficient than the RCBD for different trials and experiment sizes.

Federer (1956) proposed augmented experimental designs to evaluate more treatments using similar resources by having checks replicated in a particular experimental design (i.e., RCBD or ALPHA) and then *augmenting* or filling in the experiment with unreplicated treatments (Federer and Crossa 2012). The experimental error is estimated from the replicated checks and used for statistical inference, and therefore, the error variance of the checks should be similar to those from the unreplicated entries to avoid biases (Kempton 1984). Cullis et al. (2006) proposed to use entries from the population to substitute replicated checks

in grid plot designs to better avoid this bias and called this a partially replicated experimental design (PREP). Later, Williams et al. (2011) proposed the augmented PREP (A-PREP) design to extend Cullis' idea (Cullis et al. 2006) of the PREP for use in multi-environment trials (METs) where all entries are evaluated with replications in one environment and are unreplicated in all other environments. Furthermore, replicated entries for the whole MET would be randomized according to an ALPHA or other experimental design. The main difference between the PREP and A-PREP designs is that the replicated entries are the same in all locations in the PREP (Cullis et al. 2006) and different for each environment in the A-PREP (Moehring et al. 2014; Williams et al. 2011).

Designing experiments according to an assumed spatial structure and a planned method of spatial analysis was first suggested by Wilkinson et al. (1983) who proposed that trials should be designed specifically with their method of nearest neighbor analysis in mind. A group of similar models recommended by Besag and Kempton (1986) and Williams (1986) were developed, so the neighbor relationship between plots only existed for plots within the same incomplete block. The one-dimensional LV plus incomplete block model from Williams (1986) was extended to two dimensions in an additive form by Williams et al. (2006) and a separable form by Piepho and Williams (2010). Eccleston and Chan (1998) characterized designs resulting from a separable AR1 model, a LV model, and a separable moving average model and showed that the AR1 × AR1 and LV designs are robust when the assumed dependence structure is incorrect. PREP designs created specifically for analysis with the Gilmour et al. (1997) model were developed by Cullis et al. (2006). Williams et al. (2006) discussed the construction of resolvable spatial R-CD using a two-dimensional form of the LV model and found that designs generated with an AR1 model were more efficient when the autocorrelations in rows and columns were low. Williams and Piepho (2013) also found comparable results in their study on the efficiency of designs generated using separable AR1 models and separable and additive LV models for a range of dependence structure parameters.

Spatial control can also be achieved at the analysis level of an experiment, regardless of the randomization procedure, by using a plot's position to correct for spatial patterns through trend analysis (i.e., polynomial regression, Federer and Schlottfeldt 1954; Tamura et al. 1988), neighbor analysis (Bartlett 1978; Kempton and Howes 1981; Williams 1986), or modeling correlated errors in either rows or columns (Papadakis 1937; Zimmerman and Harville 1991).

The inclusion of non-traditional spatial components in the analysis of RBE designs has also been used to better control field spatial variation. Zimmerman and Harville (1991) proposed a correlated error plus trend model to directly control for both large-scale and small-scale spatial heterogeneity and showed that this model was more efficient than previously developed models. Gilmour et al. (1997) recommended the use of a two-dimensional separable autoregressive process to model the spatial variance component of a row–column model (AR1). Recent work has been done to compare RBE designs with the addition of spatial components similar to those in the Zimmerman and Harville (1991) model. Borges et al. (2019) and González-Barrios et al. (2019) compared several spatial error covariance structures including the AR1 process and the two-dimensional exponential process with a two-dimensional spline model using novel simulation approaches. They reported higher

accuracy when spatial corrections were included and found that none of the spatial corrections outperformed the use of good experimental designs.

The setting of RBE designs can also influence the effectiveness of the design and its corresponding analysis through factors like the heritability of the trait (Borges et al. 2019), the genotype by environment interaction (GE) structure of the MET (Moehring et al. 2014), experimental unit and experiment size (Casler 2015; Lin and Binns 1986), and modeling the relationship among genotypes (Moehring et al. 2014).

Broadly summarizing, there is a need to evaluate strategies to control spatial heterogeneity through the design and analysis of field experiments, to improve RBE designs and analysis models with spatial components, to improve non-traditional spatial analysis models, to develop designs specifically to accommodate non-traditional spatial analysis models, and to compare the various methodologies within each group. The goal of this study was to use a simulation approach with real wheat performance data to compare classic and spatial randomization-based experimental designs in their ability to efficiently control spatial variability as extensions to Moehring et al. (2014), Borges et al. (2019), and González-Barrios et al. (2019). Specifically, the objectives were to evaluate the performance of SP designs compared to classic experimental designs and to evaluate the performance of all the RBEs with and without spatial corrections in a large number of scenarios of differing experimental unit size, experiment size, GE level, relationship information among genotypes, and trait heritability.

2. MATERIALS AND METHODS

2.1. General Procedure

The basic procedure for the simulation follows those of Borges et al. (2019) and González-Barrios et al. (2019) and is outlined in Fig. 1 and Supplemental Figure 1. First, phenotypic data were obtained for 1314 advanced wheat lines evaluated in 60 location-year environments of a real field performance trial. The performance for five environments was simulated for two levels of genotype by environment interaction (GE, GE:G = 0.2 and 2.5) to create a vector of true genotypic values for yield per environment. Simultaneously, spatial variability of a real wheat uniformity trial (64 ha) was obtained to create a grid of plots for each of two experimental unit sizes (2×3 and $4 \times 6\,\text{m}^2$). At each of 100 independently sampled sites within the uniformity trial, a random sample of true genotypic effects for yield were randomly assigned to an experimental unit according to one of ten experimental designs to create a new vector of simulated yields. Genotypic effects were randomized over the experimental units of field variation 100 times for each scenario of the small-sized experiments (50–83 genotypes) and 10 times for each scenario of the medium-sized experiments (200–333 genotypes). Each vector of simulated yield was analyzed according to the respective experimental design, two methods of spatial correction (no spatial correction, NSC; and a two-dimensional autoregressive process, $AR1 \times AR1$), and two assumptions about the relationship among genotypes ($\mathbf{I}$ and $\mathbf{K}$) to retrieve a vector of estimated genotypic effects for yield in each environment. The correlation between the true and the predicted genotypic

values and the response to selection were used to compare experimental designs and spatial correction performance in all scenarios.

2.2. Real Data from Advanced Wheat Lines

2.2.1. Real Phenotypic Data

A total of 1314 advanced inbred wheat lines from the Wheat Breeding Program of the National Agricultural Research Institute (INIA) of Uruguay (IWBP) were assessed for grain yield (kg ha^{-1}) in five locations in Uruguay (Dolores, 33° 50′ S, 58° 14′ W; Durazno, 33° 33′ S, 56° 31′ W; La Estanzuela, 34° 20′ S, 57° 42′ W; Young, 32° 76′ S, 131 57° 57′ W; and Ruta2, 33° 45′ S, 57° 90′ W) from 2010 to 2017. For each of the 60 location-year environments, genotypes were evaluated in some or all of the following trials as ALPHA designs: elite yield trials (EYT, F_9) with four replications, advanced yield trials (AYT, F_8) with three replications, and preliminary yield trials (PYT, F_7) with three replications (Fig. 1a). Trials were further classified based on short-, intermediate-, and long-cycle maturity. A subset of this population was described in detail in Lado et al. (2016).

2.2.2. Molecular Marker Data

The IWBP population was genotyped using genotyping-by-sequencing (GBS) (Elshire et al. 2011) using a modification for wheat (Poland and Rife 2012) as described in Lado et al. (2016). A total of 81,999 markers were used to estimate the realized additive relationship matrix (**K**, Fig. 1a). The **K** matrix was estimated as the cross-product of the centered and standardized marker states divided by the number of markers and was estimated using the *rrBLUP* package (Endelman 2011).

2.2.3. Statistical Analysis of the Real Phenotypic Data

The real phenotypic data were analyzed using a two-step approach. First, the empirical best linear unbiased estimates (E-BLUEs) were estimated for all genotypes in each maturity group in each trial in each location-year environment following Lado et al. (2016). The term E-BLUE was used here to indicate that variance components were unknown and estimated from the data. The E-BLUEs of the genotypes in all environments (y_{ij}) were then analyzed in a second step according to the following statistical model:

$$y_{ij} = \mu + g_i + e_j + ge_{ij} + e_{ij}, \tag{1}$$

where μ is the overall mean or intercept, g_i is the random effect of the ith genotype, e_j is the effect of the jth environment, ge_{ij} is the random effect of the interaction of the ith genotype evaluated in the jth environment, and e_{ij} is the residual error. The vector of genotypic effects (**g**) is $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{K}\sigma_g^2\right)$ where **K** is the realized additive relationship matrix estimated with the markers and σ_g^2 is the estimated genotypic variance. The vector of genotype by environment interaction effects (**ge**) is $\mathbf{ge} \sim N\left(\mathbf{0}, \mathbf{K} \otimes \Sigma\sigma_{ge}^2\right)$ where Σ is the variance–covariance matrix among environments modeled as a factor analytic of order 1 (FA1) structure, $\otimes$ is the
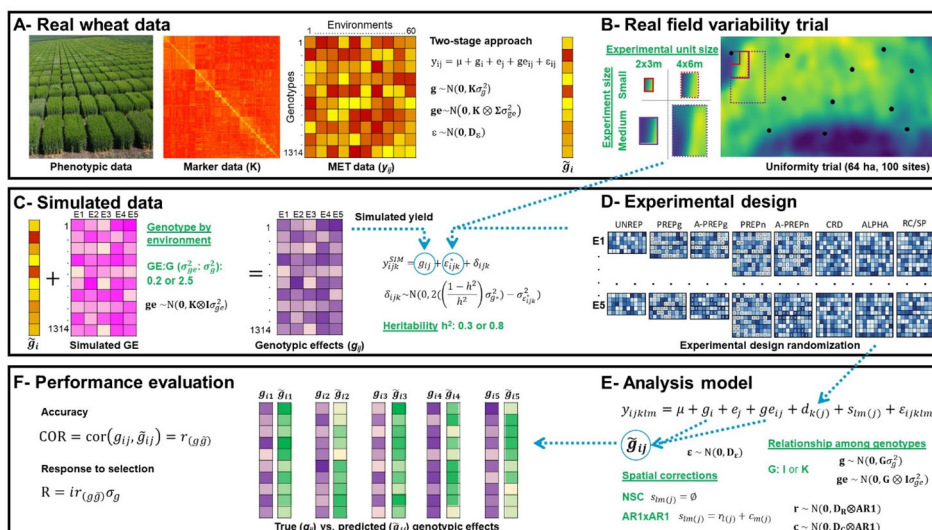
Figure 1. Diagram of the simulation strategy used to compare experimental designs and spatial correction methods for different scenarios of experiment size, experimental unit size, genotype by environment interaction (GE), heritability, and use of molecular marker information. Real data (**a**); a real wheat multi-environment trial was conducted where 1314 genotypes, all with marker information, were evaluated in 60 location-year environments in Uruguay. A two-stage approach was taken to first analyze each trial separately and then together to account for GE. The result was a vector of empirical best linear unbiased predictors (E-BLUPs) for yield for each genotype ($\tilde{g}_i$). Real field variability information (**b**); yield data were collected for a large (64 ha) wheat uniformity trial, and kriging was performed to produce grids of experimental units across the uniformity trial for all combinations of experimental unit size (2×3 and 4×6 m^2) so that each experimental unit contained a single value of yield (kg ha^{-1}) to be used as a measure of spatial variation (ε_{ijk}^{*}). One-hundred independent sites were then chosen within the uniformity trial so a multi-environment trial with five environments could be simulated at each site. Simulated phenotypic data (**c**); to simulate GE for five environments, effects were sampled from a multivariate normal distribution which had a covariance matrix as a function of the marker-based relationship matrix, the genetic variance, and predetermined values of $\sigma_{ge}^2 : \sigma_g^2$ (0.2 and 2.5). These effects were added to the vector of E-BLUPs from A to produce a vector of true genotypic effects for each of five environments for each GE level. Experimental design (**d**); at each site in the uniformity trial, an independent sample of genotypic effects from **c** was added to experimental units of spatial variation from **b** according to one of ten experimental designs. Additional field noise (δ_{ijk}) as a function of two predetermined levels of yield heritability (0.3 and 0.8) was also added to produce the final simulated phenotypic data (y_{ijk}^{SIM}). For small experiments, a re-randomization of the sample of genotypes at each of the 100 sites occurred 100 times. For medium experiments, this occurred 10 times at each of the 50 sites. Model analysis (**e**); for each scenario, the vector of simulated yield from part D was analyzed according to a model with genotypic effects, environmental effects, GE, the corresponding design terms, and additional spatial correction terms. Genotypic effects and GE were modeled using the **g** matrix either specifying no relationship among genotypes (**i**) or a marker-based relationship among genotypes (**k**). Either no additional spatial correction terms were included in the model (NSC) or terms for rows and columns according to a first-order autoregressive structure (AR1$\times$AR1) were included. Performance evaluation (**f**); two criteria were used to evaluate the performance of the experimental designs and methods of spatial corrections in all scenarios: (1) the correlation between the true genotypic values from part C and the predicted genotypic values from part **e** (COR), and (2) the response to selection when selecting both locally and globally estimated using the breeder's equation as a function of COR, three standardized selection intensities, and the true genotypic variances from part **c**.

Kronecker product, and σ_{ge}^2 is the estimated GE variance. The following correlation structures among environments were evaluated, and the best model (i.e., FA1) was selected based on both AIC and BIC statistics (BIC values are shown here): diagonal (57,334), compound symmetry (55,830), heterogeneous compound symmetry (55,940), and factor analytic of order 1 (54,912). The factor analytic of order 2 and the unstructured models failed to converge. The vector of residual errors (**e**) is $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D_e})$ where $\mathbf{D_e}$ is a block diagonal matrix with the error variances within environments estimated in step one. Both the first and second steps of this analysis were performed using the *ASReml-R* package (Butler et al. 2009) of the R software (R Core Team 2019). Empirical best linear unbiased predictors (E-BLUP) of the genotypes ($\mathbf{\tilde{g}}$) from the second step of the analysis were used for the simulation study (Fig. 1a).

2.3. REAL DATA FROM THE UNIFORMITY TRIAL AND CREATION OF THE SPATIAL VARIABILITY MAPS

A uniformity trial of 64 ha was sown in Dolores (33° 50′ S, 58° 14′ W), Uruguay, with the wheat cultivar 'Nogal' (USDA-ARS 1992) at a density of 120 kg ha^{-1} (Fig. 1b). The field was harvested on June 20, 2008, in 1445 rectangular plots of size 15×5 m^2, and grain yield (kg ha^{-1}) was recorded via a yield monitor. From yield monitor data, an empiric variogram was computed by Matheron's method of moments considering a linear trend surface according to x–y coordinates, and a Matern variogram model (kappa = 1) was fitted. Universal kriging was performed using the best fitted model and two different cell sizes of the prediction grid in order to obtain two yield maps, each with a different experimental unit size: 2×3 and 4×6 m^2 (Fig. 1b). Kriging allows for interpolation at a smaller scale than the original continuously sampled field. Yield maps were created using the *sp* package (Pebesma and Bivand 2005), and experimental variograms and kriging predictions were obtained using the *gstat* package (Pebesma 2004) of the R software (R Core Team 2019). One-hundred sites were randomly chosen within the uniformity trial yield maps for the small experiments, while 50 were chosen for the medium experiments, and the real spatial variation obtained from these sites for each of the two prediction grids was used in the simulation procedure (Figs. 1b, 2).

2.4. SIMULATED SCENARIOS

A vector of simulated yield was created for each simulated scenario. Each scenario included one level of each of the following factors: GE, experimental unit size, experiment size, trait heritability, and experimental design.

2.4.1. Genotype by Environment Interaction

Genotype by environment interaction (GE) effects were simulated for all genotypes (1314 wheat lines) for two GE levels according to a compound symmetry structure with $\sigma_{GE}^2 : \sigma_G^2$ ratios of 0.2 and 2.5 (Fig. 1c) to create yield data for each genotype in a MET with five environments. GE effects were sampled from a multivariate normal distribution,
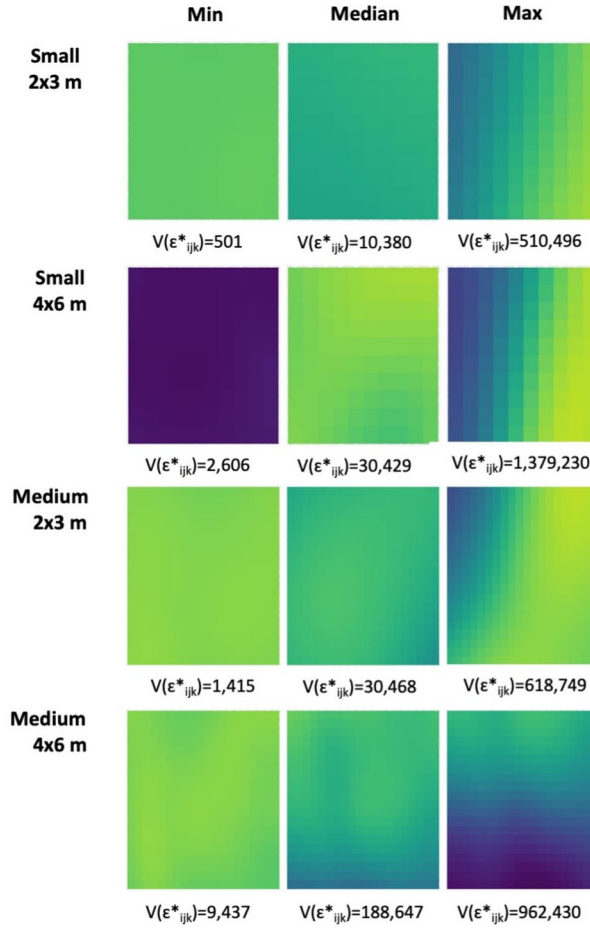
Figure 2.   Characterization of the spatial variability of 100 sites in the uniformity trial after kriging raw yield (kg ha^{-1}) data for both experimental unit sizes (2×3 and $4 \times 6\,\text{m}^2$). All 100 sites were used for the small experiments ($n_g = 50-83$), but only 50 sites were used for the medium experiments ($n_g = 200-333$). A yield variability map for the site with the minimum variation (Min), the median variation (Median), and the maximum variation (Max) for each combination is shown. The number of plots in each figure corresponds to designs having 100 or 400 experimental units for small- and medium-sized experiments, respectively (i.e., PREP$_n$, CRD, ALPHA, R-CD, and SP designs).

$\textbf{ge} \sim N\left(\mathbf{0}, \mathbf{K} \otimes \mathbf{I}_E \sigma_{ge}^2\right)$, where $\mathbf{I}_E$ is the identity matrix of order of the number of environments (i.e., five), σ_{ge}^2 is the GE variance calculated based on the GE to G ($\sigma_{ge}^2 : \sigma_g^2$) variance components ratio, and σ_g^2 was estimated from model (1). Sampled values from this distribution were added to the E-BLUPs ($\tilde{\mathbf{g}}$) from model (1) to simulate the true genotypic values within each of the five environments ($\tilde{g}_i + ge_{ij} = g_{ij}$) for each GE level.

2.4.2. Experimental Unit Size

Two experimental unit sizes were evaluated as part of the simulation scenarios as described before: 2×3 and 4×6 m^2 (Fig. 1b). This was accomplished by using different grid sizes for the kriging of the uniformity trial. Consideration of the experimental

unit size here was an evaluation of the field size and reflects the level of spatial variability due to a larger occupied area. It is not a true evaluation of experimental unit size because we used two different scales of the same kriged data and did not include any effect of intra-plot competition or changes in experimental error variance due to the use of smaller experimental units (Fig. 1b).

2.4.3. Experiment Size

Two sizes of experiments were defined in terms of the number of genotypes (n_g) and experimental units (n) evaluated. The CRD provides the base number of n_g and n for both small ($n_g = 50$, $n = 100$) and medium ($n_g = 200$, $n = 400$) experiment sizes. The other designs were constructed appropriately using these base numbers with $n_g = 50$–83 and $n = 50$–100 for the small experiments and $n_g = 200$–333 and $n = 200$–400 for the medium experiments depending on the design (Table 1). These experiment sizes are common in plant breeding and are similar to those reported in other studies (Borges et al. 2019; González-Barrios et al. 2019).

2.4.4. Heritability

Two heritability values of yield on a genotype mean basis within environments were evaluated as part of the simulation scenarios: low (0.3) and high (0.8). This was accomplished by increasing the noise level at each site (δ_{ij} see model (3) in the simulation procedure described below; Fig. 1c).

Table 1. Number of genotypes (n_g) and the number of experimental units (n) evaluated for small and medium sized experiments for each experimental design

Experimental design[a]	Replications within each environment	Small experiment size		Medium experiment size	
		Number of genotypes (n_g)	Number of experimental units (n)	Number of genotypes (n_g)	Number of experimental units (n)
UNREP	1	50	50	200	200
PREP$_g$	1/2	50	60	200	240
PREP$_n$	1/2	83	100	333	400
CRD	2	50	100	200	400
ALPHA	2	50	100	200	400
R-CD	2	50	100	200	400
SP	2	50	100	200	400

[a] Unreplicated (UNREP); partially replicated (PREP) including classic PREP (Cullis et al. 2006) and augmented PREP (A-PREP, Williams et al. (2011)) as well as PREP keeping the total number of genotypes constant (PREP$_g$) and PREP keeping the total number of experimental units constant (PREP$_n$); completely randomized experimental design (CRD); alpha incomplete block design (ALPHA); resolvable row–column design (R-CD); and spatial designs (SP) including two autocorrelation parameters: the site autocorrelation parameters (SP$_{site}$) and using 0.8 as the autocorrelation parameters (SP$_{0.8}$) at every site

2.4.5. Experimental Design

Genotypes were assigned to experimental units based on one of the following ten experimental designs: an unreplicated design (UNREP), four partially replicated designs (PREP$_n$, A-PREP$_n$, PREP$_g$, A-PREP$_g$), a completely randomized design (CRD), an alpha-lattice incomplete block design (ALPHA), a resolvable row–column design (R-CD), and two spatial designs (SP$_{site}$ and SP$_{0.8}$, Supplemental Figure 2, Fig. 1d). All experimental designs were assessed in small- and medium-sized experiments except for the PREP$_g$ and PREP$_n$ designs which were not evaluated in medium-sized experiments. Each experimental design was independently randomized over five environments at each site considering the GE structure to create a multi-environment trial.

For the UNREP designs, all genotypes were unreplicated within each environment. The CRD was randomized with two replications within each environment (Table 1, Fig. 1d, Supplemental Figure 2).

For the PREP designs, 20% of the genotypes were replicated twice in a randomized complete block design, and unreplicated genotypes were assigned at random to the remaining plots. PREP designs with the same number of genotypes as the CRD were called PREP$_g$, while PREP designs with the same number of experimental units as the CRD were called PREP$_n$ (Table 1). PREP designs following Cullis et al. (2006) with the same replicated genotypes in all locations were called PREP, while A-PREP designs following Williams et al. (2011) with different genotypes replicated in each environment were called A-PREP. Replicated genotypes were chosen at random other than with these restrictions (Table 1, Fig. 1d, Supplemental Figure 2).

ALPHA designs were randomized with two complete replications (Table 1, Fig. 1d, Supplemental Figure 2). The small ($n_g = 50$) experiment size was randomized following a 5×10 array (i.e., ten incomplete blocks of size five per replication), while the medium ($n_g = 200$) size was randomized following a 10×20 array (i.e., twenty incomplete blocks of size ten per replication). For both experiment sizes, pairs of genotypes occurred either never or once in all incomplete blocks (ALPHA$_{(0,1)}$).

For the R-CD, a resolvable incomplete block design in both rows and columns with two complete replications in each environment was used (Table 1, Fig. 1d, Supplemental Figure 2). Similarly to the randomization of the ALPHA design, an ALPHA$_{(0,1)}$ design was optimized so that each pair of genotypes was compared either never or once within a row and within a column (Table 1, Fig. 1d, Supplemental Figure 2).

Two SP designs, SP$_{site}$ and SP$_{0.8}$, were constructed for a separable autoregressive (AR) covariance structure where the variance parameters are known in advance as in Williams and Piepho (2013). A resolvable R-CD was used, but the randomization of genotypes was further optimized according to the spatial variation for each site (Table 1, Fig. 1d, Supplemental Figure 2). Optimization was defined according to the A-optimality criteria of minimizing the $trace(\mathbf{X}'\mathbf{V}^{*+}\mathbf{X})^{-1}$ where $\mathbf{X}$ is the design matrix and $\mathbf{V}^{*+}$ is the variance of y (Eccleston and Chan 1998; Williams and Piepho 2013) defined as:

$$\mathbf{V} = \gamma_E \mathbf{I_r} \otimes \mathbf{I_s} \otimes \mathbf{I_k} + \gamma_C \mathbf{I_r} \otimes \mathbf{I_s} \otimes \mathbf{J_k} + \gamma_R \mathbf{I_r} \otimes \mathbf{J_s} \otimes \mathbf{I_k} + \gamma_S \mathbf{I_r} \otimes \sum\nolimits_{\mathbf{R}} \otimes \sum\nolimits_{\mathbf{C}} \qquad (2)$$

where γ_E, γ_C, γ_R, and γ_S are the variance parameters for plot error (E), column (C), row (R), and the spatial component (S), respectively. $\mathbf{J_k}$ and $\mathbf{J_s}$ are kxk and sxs matrices of ones, and $\mathbf{I_r}$, $\mathbf{I_s}$, and $\mathbf{I_k}$ are the identity matrices of sizes rxr, sxs, and kxk, respectively, where r is the number of replications, s is the number of columns, and k is the number of rows. Additionally, $\sum_\mathbf{R}$ and $\sum_\mathbf{C}$ each have an AR(1) structure such that $\sum_\mathbf{R} = \{\rho_R^{|j_1 - j_2|}\}$ and $\sum_\mathbf{C} = \{\rho_C^{|i_1 - i_2|}\}$ where i is the index of s columns, j is the index of k rows, and ρ_R and ρ_C are the correlation parameters for rows and columns, respectively (Williams and Piepho 2013). For SP_{site}, autocorrelation parameters were calculated at each one of the 100 sites in the uniformity trial and used as input to optimize the randomization. This represents the best possible, although unlikely, situation where the actual spatial autocorrelation is known before conducting the experiment. The ranges for ρ_R and ρ_c, were 0.93–0.99 and 0.94–0.99, respectively, for 2×3 m^2 experimental units and 0.80–0.99 and 0.92–0.99 for 4×6 m^2 experimental units. For the $SP_{0.8}$ design, an autocorrelation parameter of $\rho_R = \rho_C = 0.8$ was chosen to enhance the breadth of the study while modeling comparable autocorrelation in the rows and columns following Williams et al. (2006).

The CRD and PREP designs were randomized using custom-built codes. The ALPHA design was randomized with the *agricolae* package (de Mendiburu 2019) of the R software (R Core Team 2019). The R-CD and both SP designs were randomized in DiGGer (Coombes 2002) with a maximum of 50,000 phase 1 interchanges when optimizing within row and columns under no spatial correlation and a maximum of 500,000 phase 2 interchanges when optimizing according to spatial correlation (Williams and Piepho 2013).

2.5. SIMULATION PROCEDURE

Simulated yield for each iteration of each simulation scenario at each site in the uniformity trial was calculated as follows (Fig. 1c):

$$y_{ijk}^{\text{SIM}} = g_{ij} + \varepsilon_{ijk}^* + \delta_{ijk}, \tag{3}$$

where y_{ijk}^{SIM} is the simulated yield corresponding to the ith genotype assigned to the kth experimental unit in the jth environment, and g_{ij} is the true genotypic value of the ith genotype in the jth environment as described earlier. ε_{ijk}^* is the field experimental error obtained from the given plot in the uniformity trial, and δ_{ijk} is a repeatability error following Borges et al. (2019). We assumed $\delta_{ijk} \sim N(0, \sigma_\delta^2)$ where σ_δ^2 is a random noise variance computed as a function of the predefined heritability of yield ($\sigma_\delta^2 = 2\left(\frac{1-h^2}{h^2}\right)\sigma_{g*}^2 - \sigma_{\varepsilon*}^2$) where $h^2 = 0.3$ or 0.8 as described earlier, σ_{g*}^2 is the genotypic variance calculated from the true genotypic values of sampled genotypes at each site, and $\sigma_{\varepsilon*}^2$ is the field error variance within the site. At each site in the small experiments, 100 iterations of model (3) were conducted for each plot in each simulation scenario, and ten iterations were used for the medium experiments. This resulted in 10,000 simulations for each scenario of the small experiments and 500 for each scenario of the medium experiments. Random samples of genotypes were independent across sites.

2.6. STATISTICAL ANALYSIS OF THE SIMULATED DATA

The vectors of simulated yield from model (3) for each of ten experimental designs in all simulation scenarios were analyzed according to a general statistical model using two methods of spatial corrections and two types of relationship matrices in order to retrieve the estimated genotypic effects (Fig. 1e):

$$y_{ijklm} = \mu + g_i + e_j + ge_{ij} + d_{k(j)} + s_{lm(j)} + \varepsilon_{ijklm}, \tag{4}$$

where y_{ijklm} is the simulated yield for the ith genotype, the jth environment, kth replication (when applicable), lth row, and mth column; μ is the overall mean, g_i is a random effect of the ith genotype, e_j is the effect of the jth environment, ge_{ij} is the random effect of the interaction between the ith genotype and the jth environment, $d_{k(j)}$ is the effect of the design structure factors nested within the jth environment, $s_{lm(j)}$ is the effect of the terms corresponding to spatial corrections within the jth environment, and ε_{ijklm} is the residual error. Correspondingly, $\mathbf{g} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_g^2\right)$ and $\mathbf{ge} \sim N\left(\mathbf{0}, \mathbf{G} \otimes \mathbf{I}_E \sigma_{ge}^2\right)$ where $\mathbf{G}$ is the relationship matrix among genotypes, σ_g^2 is the genetic variance, $\mathbf{I_E}$ is the identity matrix of order of the number of environments, σ_{ge}^2 is the genotype by environment interaction variance, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{D}_\varepsilon)$, and $\mathbf{D}_\varepsilon$ is the block diagonal matrix of error variances for each environment. Two variance–covariance structures were used to model the relationship among genotypes ($\mathbf{G}$): The identity matrix ($\mathbf{G} = \mathbf{I}_{ng}$) assuming unrelated genotypes and the realized additive relationship matrix based on marker information ($\mathbf{G} = \mathbf{K}$) assuming the correlation among genotypes are given by the genetic relationship among genotypes. The covariance among all random effects was assumed zero (Fig. 1e). Design structure factors were defined based on each experimental design. For the UNREP and CRD, $d_{k(j)}$ is null. For the PREP designs, $d_{k(j)} = \beta_{t(j)}$, where $\beta_{t(j)}$ is a fixed effect of the tth complete block within the jth environment. For the ALPHA design, $d_{k(j)} = \beta_{t(j)} + b_{q(jt)}$, where $b_{q(jt)}$ is a random effect of the qth incomplete block within the tth replication within the jth environment, $\mathbf{b} \sim N\left(\mathbf{0}, \mathbf{D}_b\right)$, and $\mathbf{D_b}$ is a block diagonal matrix of the incomplete block variances for each environment. For the R-CD and SP designs, $d_{k(j)} = \beta_{t(j)} + r_{l(j)} + c_{m(j)}$ where $r_{l(j)}$ is a random effect of the lth row within the jth environment, $\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R)$, D_R is a block diagonal matrix of the row variances for each environment, $c_{m(j)}$ is a random effect of the mth column within the jth environment, $\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C)$, and $\mathbf{D_C}$ is a block diagonal matrix of the column variances for each environment. Two spatial correction methods were considered by including additional terms in the model ($s_{lm(j)}$). A no spatial correction (NSC) model was used with $s_{lm(j)}$ as null. In this case, it was assumed that the experimental design structure (e.g., blocking factors when applicable) is efficient at controlling the entire systematic environmental variation. In the second method, the random row and column effects nested within environments were added to the model (when these terms are not already included), and a first-order autoregressive covariance structure (AR1) was used in both rows and columns ($s_{lm(j)} = r_{l(j)} + c_{m(j)}$, where $\mathbf{r} \sim N(\mathbf{0}, \mathbf{D}_R \otimes \mathbf{AR1})$ and $\mathbf{c} \sim N(\mathbf{0}, \mathbf{D}_C \otimes \mathbf{AR1})$). This was used instead of modeling the AR1 × AR1 correction in the R matrix due to software constraints. All analyses were performed using the *Sommer* package (Covarrubias-Pazaran 2019) of the

R software (R Core Team 2019) using the Newton–Raphson restricted maximum likelihood method.

2.7. Performance Evaluation Criteria

The performance of the model in each scenario was evaluated using prediction accuracy and response to selection. Prediction accuracy was estimated by the Pearson correlation coefficient (COR) between the true (g_{ij}) and the predicted $(\tilde{g}_{ij})$ genotypic values. The response to selection (R) was estimated from the breeder's equation following Lorenz (2013) as:

$$R = i\, r_{(g\tilde{g})} \sigma_g \tag{5}$$

where i is the standardized selection intensity, $r_{(g\tilde{g})}$ is the accuracy estimated as the Pearson correlation (COR) between true and predicted genotypic values, and σ_g is the true genetic standard deviation. Three values of standardized selection intensity equivalent to 5, 10, and 15% of individuals selected in the CRD (i.e., $n_g = 3, 5$, and 8 in all small experiments, and $n_g = 10, 20$, and 30 in all medium experiments) were compared for local (i.e., selection within environments) and global (i.e., selection across environments) response to selection. This means that eight out of 50 genotypes are selected for each design in the small-sized experiments besides the PREP$_n$ design in which eight out of 83 genotypes are selected. To additionally evaluate the performance of simulations, the mean-based heritability was estimated following Cullis et al. (2006):

$$H^2_{\text{Cullis}} = 1 - \frac{\bar{v}^{\text{BLUP}}_{\Delta_{..}}}{2\hat{\sigma}^2_g} \tag{6}$$

where $\bar{v}^{\text{BLUP}}_{\Delta_{..}}$ is the mean variance of a difference of two BLUPs for the genotypic effect, $\hat{\sigma}^2_g$ is the genotypic variance, and both were estimated from model (4). Similar to the response to selection, a global estimate (i.e., across locations) of the heritability was estimated from model (4), while a local estimate (i.e., for each environment) was estimated from a modification of model (4) for each location.

Each criterion was calculated for all iterations at all sites. Box plots were created to summarize the evaluation criteria estimates using the *ggplot2* package (Wickham 2016) of the R software (R Core Team 2019). Finally, model convergence was estimated as the proportion of times each scenario (i.e., combinations of experimental unit size, experimental design, GE level, heritability level, type of relationship among genotypes, and spatial corrections) converged out of 10,000 iterations for the small experiments and 500 iterations for the medium experiments.

Table 2. Correlation (COR) between true and predicted genotypic values for yield (kg ha^{-1}) for ten experimental designs for each scenario of GE, experimental unit size, trait heritability, relationship among genotypes, and method of spatial correction for the small experiments (i.e., 50–83 genotypes)

Scenario	GE:G=0.2										GE:G=2.5									
	UNREP	PREPg	A-PREPg	PREPn	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8	UNREP	PREPg	A-PREPg	PREPn	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8
Experimental unit size: 2 × 3 m²																				
Heritability = 0.3																				
G=I, NSC	0.560	0.589	0.587	0.601	0.597	0.676	0.679	0.674	0.673	0.674	0.297	0.435	0.432	0.446	0.447	0.542	0.540	0.539	0.539	0.539
AR1 × AR1	0.562	0.593	0.590	0.605	0.600	0.679	0.682	0.678	0.678	0.679	0.297	0.434	0.432	0.449	0.450	0.544	0.542	0.543	0.545	0.543
G=K, NSC	0.585	0.634	0.631	0.663	0.661	0.713	0.716	0.711	0.711	0.711	0.342	0.533	0.535	0.580	0.580	0.624	0.625	0.622	0.623	0.624
AR1 × AR1	0.587	0.636	0.633	0.666	0.664	0.717	0.719	0.715	0.715	0.715	0.343	0.535	0.536	0.582	0.581	0.626	0.627	0.626	0.628	0.628
Heritability = 0.8																				
G=I, NSC	0.795	0.877	0.877	0.881	0.882	0.914	0.921	0.912	0.912	0.912	0.496	0.840	0.840	0.843	0.843	0.894	0.895	0.892	0.892	0.892
AR1 × AR1	0.795	0.880	0.879	0.887	0.888	0.919	0.922	0.913	0.913	0.913	0.495	0.838	0.839	0.843	0.843	0.895	0.895	0.894	0.894	0.894
G=K, NSC	0.795	0.898	0.898	0.906	0.906	0.927	0.932	0.925	0.924	0.924	0.486	0.863	0.862	0.872	0.872	0.905	0.906	0.903	0.904	0.903
AR1 × AR1	0.796	0.900	0.900	0.911	0.911	0.930	0.933	0.925	0.925	0.925	0.486	0.862	0.861	0.872	0.872	0.906	0.906	0.905	0.905	0.905
Experimental unit size: 4 × 6 m²																				
Heritability = 0.3																				
G=I, NSC	0.558	0.588	0.585	0.598	0.595	0.676	0.679	0.670	0.673	0.671	0.300	0.432	0.433	0.449	0.446	0.542	0.541	0.538	0.538	0.538
AR1 × AR1	0.560	0.591	0.589	0.604	0.601	0.681	0.682	0.675	0.678	0.675	0.301	0.433	0.433	0.451	0.448	0.544	0.543	0.543	0.545	0.543
G=K, NSC	0.583	0.633	0.630	0.664	0.661	0.714	0.717	0.710	0.712	0.710	0.347	0.535	0.533	0.579	0.579	0.625	0.623	0.622	0.622	0.622
AR1 × AR1	0.586	0.637	0.634	0.669	0.666	0.719	0.720	0.714	0.716	0.714	0.348	0.537	0.535	0.581	0.581	0.627	0.625	0.627	0.627	0.626
Heritability = 0.8																				
G=I, NSC	0.794	0.883	0.883	0.884	0.884	0.914	0.924	0.917	0.919	0.914	0.496	0.842	0.842	0.845	0.845	0.894	0.898	0.893	0.894	0.893
AR1 × AR1	0.796	0.889	0.889	0.891	0.891	0.919	0.926	0.918	0.920	0.915	0.495	0.843	0.842	0.848	0.848	0.897	0.899	0.895	0.896	0.895
G=K, NSC	0.794	0.903	0.903	0.908	0.908	0.927	0.935	0.928	0.930	0.925	0.488	0.865	0.865	0.873	0.873	0.905	0.909	0.905	0.905	0.905
AR1 × AR1	0.796	0.908	0.908	0.914	0.914	0.931	0.936	0.929	0.931	0.926	0.489	0.865	0.865	0.876	0.876	0.908	0.910	0.907	0.907	0.906

Mean prediction accuracy (COR) between true and predicted genotypic values for ten experimental designs and two methods of spatial correction (no spatial correction, NSC; and two-dimensional autoregressive model of order 1, AR1 × AR1) for the small-sized experiments under all combinations of experimental unit size, heritability, relationship among genotypes, and GE. Experimental designs include: unreplicated (UNREP); partially replicated (PREP) including classic PREP (Cullis et al. 2006) and augmented PREP (A-PREP, Williams et al. 2011) as well as PREP keeping the total number of genotypes constant (PREPg) and PREP keeping the total number of experimental units constant (PREPn); completely randomized experimental design (CRD); resolvable alpha incomplete block design (ALPHA); resolvable row-column design (R-CD); and spatial designs (SP) including two parameter comparisons (i.e., using the site autocorrelation parameters (SP$_{site}$) and using 0.8 (SP$_{0.8}$) as autocorrelation parameters for every site). Each scenario was evaluated 100 times in each of 100 uniformity trial sites

Table 3. Correlation (COR) between true and predicted genotypic values for yield (kg ha^{-1}) for eight experimental designs for each scenario of GE, experimental unit size, trait heritability, relationship among genotypes, and method of spatial correction for the medium experiments (i.e., 200–333 genotypes)

Scenario	GE:G = 0.2								GE:G = 2.5							
	UNREP	A-PREPg	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8	UNREP	A-PREPg	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8
Experimental unit size: 2 × 3 m²																
Heritability = 0.3																
G = I																
NSC	0.563	0.612	0.614	0.687	0.698	0.684	0.685	0.684	0.286	0.465	0.469	0.553	0.556	0.552	0.553	0.553
AR1 × AR1	0.565	0.616	0.626	0.696	0.700	0.685	0.686	0.685	0.287	0.466	0.471	0.554	0.557	0.553	0.554	0.554
G = K																
NSC	0.639	0.721	0.753	0.778	0.786	0.776	0.777	0.775	0.380	0.653	0.690	0.713	0.714	0.712	0.712	0.714
AR1 × AR1	0.641	0.724	0.760	0.786	0.788	0.777	0.778	0.776	0.381	0.654	0.691	0.714	0.715	0.713	0.714	0.715
Heritability = 0.8																
G = I																
NSC	0.799	0.889	0.889	0.915	0.931	0.920	0.919	0.918	0.491	0.849	0.848	0.894	0.908	0.896	0.896	0.896
AR1 × AR1	0.803	0.900	0.901	0.923	0.932	0.920	0.919	0.918	0.492	0.852	0.862	0.905	0.909	0.896	0.896	0.896
G = K																
NSC	0.805	0.921	0.927	0.936	0.945	0.938	0.938	0.936	0.489	0.890	0.899	0.919	0.926	0.919	0.919	0.919
AR1 × AR1	0.808	0.926	0.932	0.941	0.946	0.938	0.938	0.936	0.489	0.892	0.907	0.926	0.927	0.920	0.919	0.919
Experimental unit size: 4 × 6 m²																
Heritability = 0.3																
G = I																
NSC	0.561	0.610	0.611	0.686	0.727	0.647	0.644	0.646	0.286	0.464	0.471	0.552	0.559	0.548	0.549	0.548
AR1 × AR1	0.568	0.624	0.658	0.722	0.729	0.642	0.640	0.641	0.287	0.466	0.480	0.560	0.561	0.549	0.550	0.549

Table 3. *(Continued)*

Scenario	GE:G = 0.2								GE:G = 2.5							
	UNREP	A-PREPg	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8	UNREP	A-PREPg	A-PREPn	CRD	ALPHA	RC	SPsite	SP0.8
G = K																
NSC	0.640	0.716	0.748	0.777	0.804	0.757	0.756	0.756	0.378	0.651	0.688	0.710	0.715	0.707	0.708	0.708
AR1 × AR1	0.645	0.727	0.777	0.803	0.806	0.752	0.753	0.751	0.380	0.653	0.693	0.716	0.717	0.708	0.708	0.709
Heritability = 0.8																
G = I																
NSC	0.798	0.887	0.887	0.914	0.929	0.937	0.934	0.937	0.492	0.852	0.847	0.894	0.907	0.900	0.899	0.900
AR1 × AR1	0.800	0.892	0.899	0.923	0.930	0.873	0.868	0.864	0.495	0.865	0.859	0.903	0.908	0.897	0.897	0.897
G = K																
NSC	0.804	0.919	0.922	0.936	0.944	0.940	0.939	0.941	0.489	0.891	0.898	0.918	0.927	0.920	0.920	0.920
AR1 × AR1	0.805	0.922	0.926	0.941	0.945	0.899	0.897	0.899	0.491	0.901	0.905	0.924	0.927	0.918	0.918	0.917

Mean prediction accuracy (COR) between true and predicted genotypic values for eight experimental designs and two methods of spatial correction (no spatial correction (no spatial correction, NSC; and two-dimensional autoregressive model of order 1, AR1 × AR1) for the medium-sized experiments under all combinations of experimental unit size, heritability, relationship among genotypes, and GE. Experimental designs include: unreplicated (UNREP); augmented PREP (A-PREP, Williams et al. 2011) keeping the total number of genotypes constant (A-PREPg) and keeping the total number of experimental units constant (A-PREPn); completely randomized experimental design (CRD); resolvable alpha incomplete block design (ALPHA); resolvable row–column design (R-CD); and spatial designs (SP) including two parameter comparisons (i.e., using the site autocorrelation parameters (SPsite) and using 0.8 (SP0.8) as autocorrelation parameters for every site). Each scenario was evaluated 10 times in each of 50 uniformity trial sites
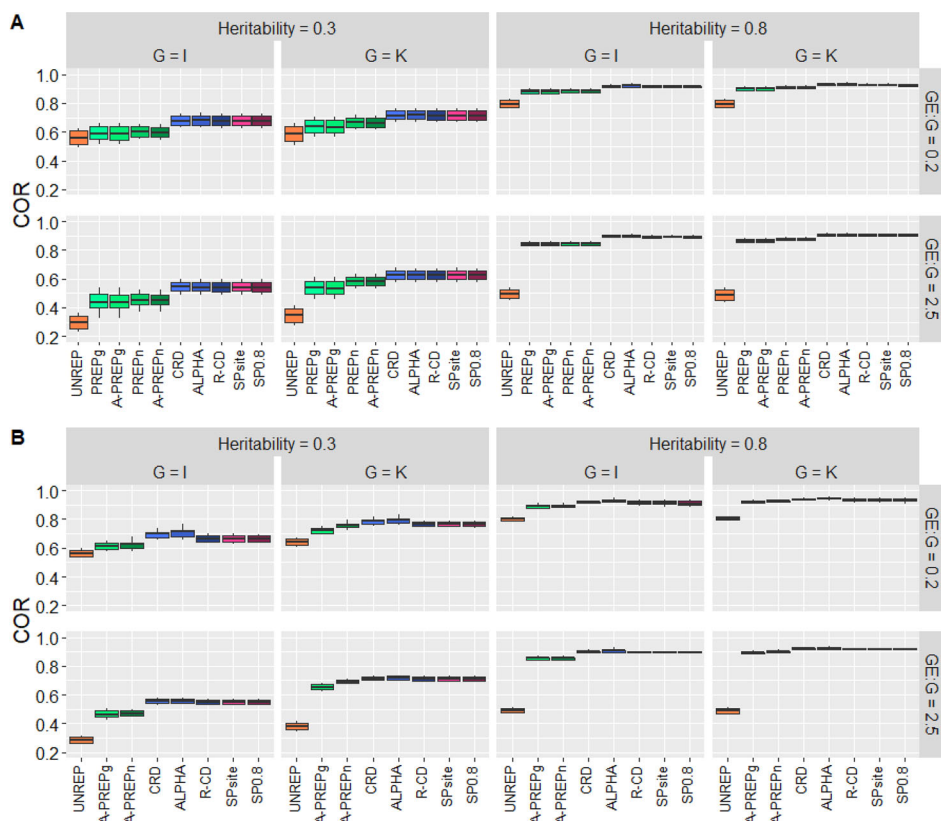
Figure 3. Correlation (COR) between true and predicted genotypic values for yield (kg ha^{-1}) for ten experimental designs, two heritability levels (i.e., low=0.3, high=0.8), two relationships among genotypes, and two GE structures for the **a** small (i.e., 50–83 genotypes), and **b** medium (i.e., 200–333 genotypes) experiment sizes. Experimental designs include: unreplicated (UNREP); partially replicated (PREP) including classic PREP (Cullis et al. 2006) and augmented PREP (A-PREP, Williams et al. 2011) as well as PREP keeping the total number of genotypes constant (PREP$_g$) and PREP keeping the total number of experimental units constant (PREP$_n$); completely randomized experimental design (CRD); resolvable alpha incomplete block design (ALPHA); resolvable row–column design (R-CD); and spatial designs (SP) including two parameter comparisons (i.e., using the site autocorrelation parameters (SP$_{site}$) and using 0.8 (SP$_{0.8}$) as autocorrelation parameters for every site). PREP$_g$ and PREP$_n$ designs were not evaluated for the medium-sized experiments. Each scenario of the small-sized experiments was evaluated 100 times in each of 100 uniformity trial sites, and each scenario of the medium-sized experiments was evaluated 10 times in each of 50 uniformity trial sites.

3. RESULTS

3.1. ACCURACY OF EXPERIMENTAL DESIGNS

The highest COR was achieved by the fully replicated experimental designs such as the ALPHA design, SP designs, R-CD, and CRD (Tables 2, 3, Fig. 3). The experimental design with the highest overall COR was the ALPHA (Tables 2, 3, Fig. 3). Furthermore, the superiority of the ALPHA experimental design was more noticeable in the medium-sized experiments with low heritability, low GE, and assuming independent genotypes (i.e., one of the hardest situations tested for any model, Fig. 3). The SP designs performed similarly
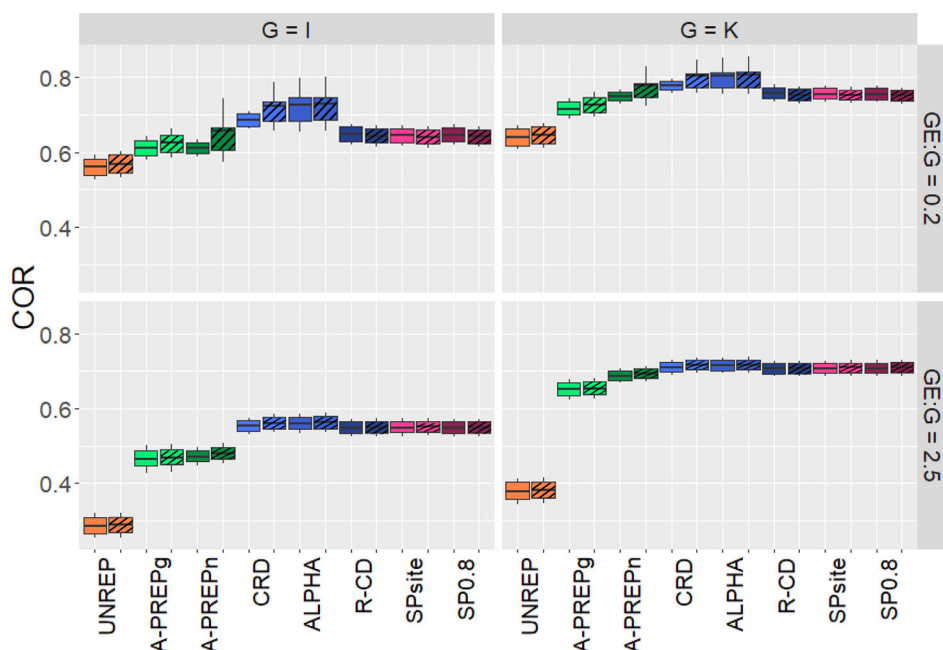
Figure 4. Correlation (COR) between true and predicted genotypic values for yield (kg ha^{-1}) for eight experimental designs with and without additional spatial corrections for the medium-sized experiments with the 4×6 m^2 experimental unit and low heritability across all levels of GE (i.e., GE:G=0.2 or 2.5) and relationship among genotypes (i.e., $\mathbf{G} = \mathbf{K}$ or $\mathbf{I}$). Each experimental design is shown (colored boxes) with no spatial correction (NSC, no pattern) and the AR1 $\times$ AR1 correction (diagonal pattern). Experimental designs include: unreplicated (UNREP); augmented partially replicated (A-PREP, Williams et al. 2011) as well as PREP keeping the total number of genotypes constant (PREP$_g$) and PREP keeping the total number of experimental units constant (PREP$_n$); completely randomized experimental design (CRD); resolvable alpha incomplete block design (ALPHA); resolvable row–column design (R-CD); and spatial designs (SP) including two parameter comparisons (i.e., using the site autocorrelation parameters (SP$_{site}$) and using 0.8 (SP$_{0.8}$) as autocorrelation parameters for every site). Medium-sized experiments were evaluated with 10 iterations at each of 50 sites.

to the R-CD in all scenarios. The UNREP was the experimental design with the lowest COR in all scenarios followed by the PREP designs. When the additive relationship matrix was used, the PREP$_n$ experimental designs had higher COR than the PREP$_g$ experimental designs (Fig. 3).

3.2. Accuracy Spatial Corrections

Models with spatial corrections (i.e., AR1 $\times$ AR1) had higher COR than models without spatial corrections (i.e., NSC) in medium-sized experiments with the large experimental unit size (i.e., 4×6 m^2) and at low heritability (i.e., $h^2 = 0.3$) for the ALPHA, CRD, PREP$_g$ and PREP$_n$ experimental designs (Fig. 4). However, there was little effect of spatial corrections for the other experimental designs or simulation scenarios in the medium-sized experiments or for any scenarios of the small-sized experiments (Tables 2, 3), and the ALPHA had only a small improvement.
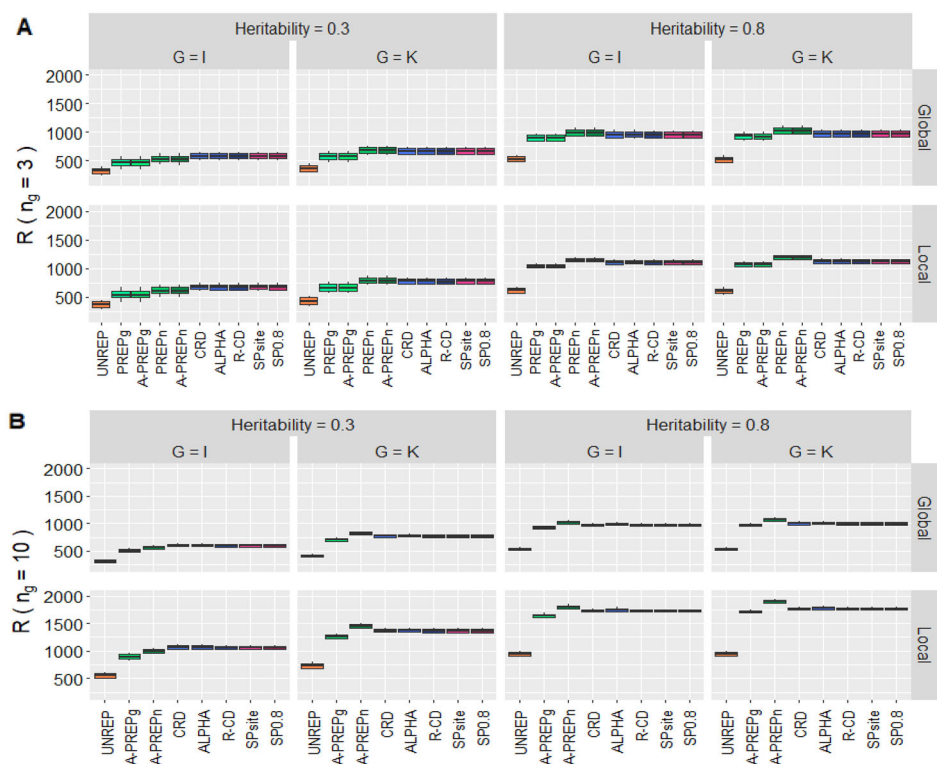
Figure 5.   Response to selection for yield (kg ha^{-1}) when selecting the top 5% of all genotypes for all experimental designs in the **a** small-sized experiments (i.e., selecting 3 genotypes out of 50) and **b** medium-sized experiments (i.e., selecting 10 individuals out of 200), respectively, across two heritability levels (i.e., low = 0.3, high = 0.8), two relationships among genotypes (modeling correlation among genotypes using the realized additive relationship matrix, **G = K**, or assuming uncorrelated genotypes, **G = I**), and for the higher level of genotype by environment interaction (GE:G = 0.2 and 2.5). Experimental designs include: unreplicated (UNREP); partially replicated (PREP) including classic PREP (Cullis et al. 2006) and augmented PREP (A-PREP, Williams et al. 2011) as well as PREP keeping the total number of genotypes constant (PREPg) and PREP keeping the total number of experimental units constant (PREPn); completely randomized experimental design (CRD); resolvable alpha incomplete block design (ALPHA); resolvable row–column design (R-CD); and spatial designs (SP) including two parameter comparisons (i.e., using the site autocorrelation parameters (SP$_{site}$) and using 0.8 (SP$_{0.8}$) as autocorrelation parameters for every site). Small-sized experiments were evaluated with 100 iterations in each of 100 sites and medium-sized experiments were evaluated with 10 iterations in each of 50 sites.

3.3. Accuracy Depending on Heritability, Relationship Among Genotypes, and GE Level

Overall, experimental design performance based on COR increased when markers were used and with high heritability (Fig. 3). However, markers had less of an effect at high heritability and high GE for all experimental designs (Fig. 3). The effect of the use of markers on COR is more noticeable in the PREP$_n$ than in the other experimental designs (Fig. 3).

Overall performance of the experimental designs was lower with higher GE, but this difference was larger for the UNREP experimental design at high heritability (Fig. 3).

Markers had no effect on COR for the UNREP experimental design at high heritability (Fig. 3).

3.4. Response to Selection

Higher response to selection for yield (kg ha^{-1}) was observed under higher selection intensity, higher heritability, with the use of markers (Fig. 5), and for low GE (Supplemental Tables 1 and 2) although more differences between the levels of each factor were observed for the medium-sized experiments compared to the small-sized experiments (Fig. 5). Local selection resulted in higher response to selection than global selection (Fig. 5) with larger differences between the two methods found at high GE compared to low GE (Supplemental Figures 1 and 2). However, a similar ranking of experimental designs in their response to selection was observed for both local and global selection strategies (Fig. 5). The PREP$_n$ designs had the largest response to selection for all scenarios and selection intensities at high heritability and at low heritability when markers were used. This advantage was less evident when selection intensity was low, for lower heritability, and for the small-sized experiments (Fig. 5). The UNREP design showed the worst response to selection in all scenarios (Fig. 5). Designs with similar levels of replication including the CRD, ALPHA, R-CD, and SP designs were similar in their response to selection for all scenarios and have the largest response to selection at low heritability when markers are not used.

3.5. Model Performance Based on Heritability

The local heritability for the CRD showed similar values as those simulated with an average of 0.27 for low heritability scenarios (i.e., $h^2 = 0.3$) and 0.77 for high heritability scenarios (i.e., $h^2 = 0.8$) (Supplemental Figure 3). Higher heritabilities were obtained for the other replicated experimental designs. Global heritability was similar for low GE, but lower heritabilities were obtained for all experimental designs with high GE. Overall, simulated data recovered heritability scenarios as expected.

3.6. Model Convergence

Model convergence was high (>80%) for all scenarios of the small-sized experiments except for the UNREP at low heritability and high GE (Supplemental Table 3). For the medium-sized experiments, model convergence was lower for the UNREP, A-PREPg, A-PREPn, CRD, and ALPHA at high heritability and low GE (Supplemental Table 4).

4. DISCUSSION

A large amount of spatial variability in agriculture field experiments occurs as a result of spatial, temporal, and spatiotemporal variations in fertility, moisture, slope, shade, management practices, disease and pests incidence, and microclimatic variations (Borges et al. 2019; González-Barrios et al. 2019; Grondona et al. 1996; Stefanova et al. 2009). Effectively controlling this spatial variation within a field is necessary to produce accurate and unbiased

estimates of treatment effects (Grondona et al. 1996). Several studies have evaluated methodologies proposed to solve this complex problem (Borges et al. 2019; González-Barrios et al. 2019; Moehring et al. 2014), but evaluations under real field variability for a broad set of situations are still lacking. This study evaluated ten experimental designs with and without spatial corrections in 100 sites of a uniformity trial with differing amounts of spatial variation to determine their relative abilities at controlling the spatial variation. The designs were tested in a range of scenarios including differing experiment sizes, experimental unit sizes, trait heritabilities, levels of GE, and the relationship among genotypes. The results from this study provide a guide to researchers for designing and analyzing routinely conducted experiments.

4.1. Experimental Design

The setting of both spatial and classic experimental designs can influence the effectiveness of the experimental design and its corresponding analysis through factors like the heritability of the trait (Borges et al. 2019; de S. Bueno and Gilmour 2003; González-Barrios et al. 2019; Mramba et al. 2018; Piepho and Möehring 2007), the GE structure of the MET (Borges et al. 2019; Casler 2015; González-Barrios et al. 2019; Lado et al. 2016; Moehring et al. 2014; Robbins et al. 2012), the experimental unit size (Borges et al. 2019; Casler 2015; González-Barrios et al. 2019; Mramba et al. 2018; Velazco et al. 2017), and specifying the relationship among genotypes (González-Barrios et al. 2019; Lado et al. 2016; Mramba et al. 2018). In this study, completely replicated designs (i.e., ALPHA, SP designs, R-CD, and CRD) outperformed PREP and UNREP designs in terms of COR in all tested scenarios of trait heritability, relationship among genotypes, and GE levels regardless of the size of the experiment (i.e., small or medium). This result was expected as replicated designs enable more accurate estimates of genotypic effects, variance components, and especially the error variance (Fisher 1926). Among all experimental designs, the ALPHA was the overall best for most scenarios. More prominent differences in the relative performance of experimental designs were observed in medium experiments than in small experiments. For instance, the ALPHA design showed higher COR at low heritability and at low GE than other designs in the medium experiments. However, the CRD, R-CD, and spatial designs (i.e., SP_{site} and $SP_{0.8}$) performed similarly to each other and had only slightly lower COR than the ALPHA design in many scenarios of the small experiments. The similar performance of the CRD compared to designs that utilize blocking indicates that blocking may not have been effective at controlling block variation in our study. Blocking was found to be less effective in specific sites with low field variability in Borges et al. (2019), and therefore, this might indicate low levels of spatial variability in our study. The fields evaluated in Borges et al. (2019) had an area 1.2 to 20 times larger than those evaluated in our study.

The performance of the PREP designs in terms of COR was intermediate to the performance of the UNREP and completely replicated designs, which was expected (Moehring et al. 2014) because fewer replications were used in comparison with the fully replicated designs (Cullis et al. 2006; Moehring et al. 2014; Williams et al. 2011). Other studies found that for a fixed trial size, PREP designs are efficient and can outperform replicated designs in METs (Cullis et al. 2006; Moehring et al. 2014). Comparing among PREP designs, $PREP_n$

designs outperformed PREP$_g$ designs in terms of COR when markers were used, and this was more evident at low heritability. One of the most critical factors affecting design performance is the number of error degrees of freedom (Clarke and Stefanova 2011), and this number is determined by the experiment size and the number of replicated plots. Therefore, PREP$_n$ designs with more error degrees of freedom have more power than PREP$_g$ designs, especially at low heritability when the increase in degrees of freedom is more impactful than at high heritability. The PREP$_n$ design also appears to partially compensate for the lack of replications when the relationship among genotypes from markers is included.

The two SP designs (i.e., SP$_{site}$ and SP$_{0.8}$) evaluated in this study had similar COR to the R-CD for all scenarios in both small- and medium-sized experiments. SP designs are not efficient when the autocorrelation between plots is high (Williams et al. 2006) as in our study. Williams and Piepho (2013) found very similar relative efficiencies for SP designs and R-CDs for a range of spatial variation parameter sets, but the SP designs and the R-CD were always better than RCBDs in their study. Therefore, it is not always necessary nor beneficial to know the underlying spatial structure of a field a priori. Additionally, Piepho et al. (2013) emphasized that the analysis of SP designs does not have the same randomization protection compared to randomized-based designs. Therefore, when designing an experiment, it is important to consider the balance between the benefit of protection due to randomization and better prediction and precision for treatment mean comparisons. We constructed our SP designs using a row–column design with an AR1 correlation structure, and the same spatial structure was used as an additional spatial correction.

The PREP$_n$ design had the highest response to selection for all scenarios of both the small and medium experiments except when the heritability was low and markers were not included in the model. Because the accuracy of the PREP$_n$ is smaller than that of the ALPHA, the response to selection is mainly driven by the larger selection intensity that can be achieved when selecting the same number of individuals from a larger population (i.e., $n_g = 50$ for the small ALPHA versus $n_g = 83$ for the small PREP$_n$). Moehring et al. (2014) showed similar results for METs. González-Barrios et al. (2019) found a similar result for the use of larger population sizes, although they found a trade-off between the number of replications and the number of genotypes evaluated. As for the other designs, the UNREP always had the lowest response to selection, and the CRD, ALPHA, R-CD, and SP designs were similar in their response to selection and usually intermediate to the PREP$_n$ and UNREP designs. Additionally, the response to selection was higher for local selection compared to global selection which was attributed to GE being exploited at a local scale.

4.2. Spatial Correction

Spatial variability can be controlled using different R matrix correlation structures in one or two dimensions (Cullis and Gleeson 1991; Gilmour et al. 1997; Qiao et al. 2000), trend analysis, or a combinations of both (Brownie et al. 1993; Casler and Undersander 2000; Zimmerman and Harville 1991). Spatial variation was modeled in this study by adding random row and column effects with first-order autoregressive covariance structures to the systematic part of the model instead of modeling the R matrix due to software constrains. In general, modeling the spatial correlation structure with the AR1 × AR1 terms had a slight

to no improvement for any design in any scenario except for the medium-sized PREP_g, PREP_n, CRD, and ALPHA designs with the larger experimental unit size (i.e., $4 \times 6 \text{ m}^2$) under low heritability. More prominent differences due to spatial correction found in our largest field tested (i.e., medium experiments with $4 \times 6 \text{ m}^2$ experimental units) indicate that experiments that cover larger fields would probably benefit more from spatial corrections in their model.

Experimental design randomization may have played an important role in addressing some patterns of spatial variation (Piepho et al. 2013) which could be one of the reasons for the marginal to no response of spatial corrections. Contrary to the results here, other studies found that spatial corrections greatly improve the performance of experimental designs (Borges et al. 2019; Casler 2010; Federer 1998; González-Barrios et al. 2019). The AR1 (Borges et al. 2019), AR1 × AR1 (Cullis and Gleeson 1991; Moehring et al. 2014; Piepho and Williams 2010), and spline (González-Barrios et al. 2019) models have been identified as superior spatial correction models, but it has also been shown that the best spatial model is case specific and depends on the experimental design and field heterogeneity (Beeck et al. 2010; Borges et al. 2019; Cullis and Gleeson 1991; Müller et al. 2010; Moehring et al. 2014; Richter and Kroschewski 2012; Stefanova et al. 2009; Williams 1986; Wu et al. 1998). In our study, the spatial corrections were not modeled in the R matrix as in other studies due to software restrictions. This is probably the reason why our spatial corrections were not as effective as in other studies.

4.3. Experimental Unit and Experiment Size

Each experimental design was evaluated in two sizes of experiments (based on the number of genotypes evaluated) with two sizes of experimental units (i.e., plots) resulting in different overall field sizes. Small experiments (i.e., $n_g = 50$–83) with the small experimental unit (i.e., $2 \times 3 \text{ m}^2$) occupied 300–600 m^2 depending on the experimental design, while small experiments with the larger experimental unit (i.e., $4 \times 6 \text{ m}^2$) occupied 1200–2400 m^2. Medium experiments (i.e., $n_g = 200$–333) with the smaller experimental unit ranged from 1200 to 2400 m^2, and the medium experiments with the large experimental unit ranged from 4800 to 9600 m^2.

No differences in COR were observed with the change in experimental unit sizes in our study, and the COR was higher for medium experiment sizes than small experiments for most of the experimental designs. However, smaller COR in larger experiments was associated with large field variability in other large field experiment studies (Borges et al. 2019). According to Smith's law (Smith 1938), there is a negative asymptotic relationship between variance on a single-plot basis and plot size, but this relationship is affected by several factors and their interactions (such as species or environment). Therefore, in terms of reducing the variance, the optimum between an increase in plot size resulting in a larger experimental area and a reduction in plot size to minimize this area depends on the shape of this relationship and therefore on the spatial variability of the whole field (Casler 2015). In our study, we modified the field size by evaluating different experimental unit and experiment sizes, but we did not evaluate different relationships for the intra-plot variance.

4.4. Heritability, Relationship Among Genotypes, and Genotype by Environment Interaction

Trait heritability is one of the most important factors affecting the performance of prediction models (Bhatta et al. 2020; Zhang et al. 2017) and in particular, experimental designs (Borges et al. 2019; de S. Bueno and Gilmour 2003; González-Barrios et al. 2019; Mramba et al. 2018; Piepho and Möehring 2007) because for traits with high heritability, less noise is present and identifying genotypic signals becomes easier no matter the experimental design. In the current study, COR was higher for high trait heritability compared to low trait heritability for all designs. Modeling the relationship among genotypes increased the COR for all experimental designs, but the effect was larger under low heritability. Incorporating the relationships among genotypes into the models was beneficial because molecular markers can provide additional information for predicting genotypic performance by borrowing information from relatives, essentially acting as additional replications (Piepho et al. 2008). The performance of genotypes for traits with high heritability can easily be predicted with phenotypic information (i.e., $\mathbf{G}=\mathbf{I}$) evaluated in proper experimental designs; therefore, the correlation among genotypes does not add additional information, and its effect is not noticeable under high trait heritability. As expected, the COR was reduced with the increase in GE, but the reduction in COR was less evident at high trait heritability for all designs except the UNREP.

4.5. Simulation Performance

The lowest convergence occurred for the small-sized UNREP design at low heritability and high GE. Moderate convergence was observed for the medium-sized PREP, CRD, and ALPHA designs at high heritability and low GE. Higher proportions of convergence (>80%) were observed for most other scenarios. The main reason for the convergence issue with UNREP may be associated with the heterogeneous covariance error assumption. Other simulation studies have also recognized convergence problems in the analysis of experimental designs associated with autoregressive variance models with several fixed and random effects or with autocorrelation near unity while fitting spatial error structures within blocks (Moehring et al. 2014; Robbins et al. 2012; Stefanova et al. 2009).

Simulation performance was also evaluated with the Cullis et al. (2006) estimate of mean-based heritability. According to these authors, the heritability is best defined in terms of pairwise comparisons among genotypes, particularly in the presence of unbalanced data (Cullis et al. 2006). In general, data were simulated properly in terms of their heritability; the local heritability (i.e., per location) realized from the CRD experiments was similar to that simulated for the low heritability (i.e., 0.27 vs. 0.3 and 0.77 vs. 0.8). Global heritability (i.e., for all locations combined) was lower when GE was high, which was also expected as the GE variance is larger. Lower heritabilities were found for all experimental designs and GE levels in global and local selection when the genotypic relationship matrix was used (i.e., $\mathbf{G}=\mathbf{K}$, data not shown), which was attributed to smaller estimates of the genotypic variance. Underestimation of mean-based heritability was found for other studies modeling the genotypic relationship matrix (Kruijer et al. 2015) and specifically for marker-based

estimates with the Cullis et al. (2006) heritability (Ould Estaghvirou et al. 2013). Larger standard errors and sometimes unrealistic biological values were also obtained in these cases (Kruijer et al. 2015). Schmidt et al. (2019) recommends the Cullis et al. (2006) estimates for unrelated genotypes estimations ($\mathbf{G=I}$).

The genotype by environment ratios were higher than the ones simulated for all scenarios (i.e., 0.5 for GE:G=0.2 and larger than 10 for GE:G=2.5). This may be a consequence of the simulation strategy or the modeling component where GE includes part of the variation in the main effects.

5. CONCLUSION

Our study evaluated the comparisons of classic experimental designs and spatial designs with and without spatial corrections in a range of scenarios. The ALPHA design was, in general, more accurate at predicting true genotypic values than other experimental designs for any combination of heritability, relationship among genotypes, and GE level, whereas the UNREP design was the worst in terms of both predicting the true genotypic values and response to selection. The ALPHA design had similar values of response to selection to those of the CRD, R-CD, and SP designs, but the $PREP_n$ design outperformed all other designs. Spatial corrections improved the performance of some designs such as $PREP_g$, $PREP_n$, and CRD designs, especially for medium-sized experiments and at low trait heritability. This study provides results covering a broad range of scenarios that could be applicable to many plant breeding efforts and worked to unify some of the methodologies regarding experimental design and analysis.

ACKNOWLEDGEMENTS

DATA AVAILIBILITY

The datasets generated during and/or analyzed during the current study are available in the Figshare repository, which is available at https://figshare.com/s/52c998883a43be3b74d4.

FUNDING

REFERENCES

Bartlett, M. S. (1978). "Nearest Neighbour Models in the Analysis of Field Experiments." *Journal of the Royal Statistical Society: Series B (Methodological)* 40(2):147–58.

Beeck, C.P., Cowling, W.A., Smith, A.B. and Cullis, B.R. (2010). "Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic mixed models with pedigree information." *Genome*, *53*(11):992-1001.

Besag, J. and Kempton, R. (1986). "Statistical Analysis of Field Experiments Using Neighbouring Plots." *Biometrics* 42(2):231-251.

Bhatta, M.,L., Gutierrez., L,. Cammarota, F., Cardozo, S. Germán, B. Gómez-Guerrero, M.F. Pardo, V. Lanaro, M. Sayas, and Castro A.J. (2020). "Multi-trait Genomic Prediction Model Increased the Predictive Ability for Agronomic and Malting Quality Traits in Barley (Hordeum vulgare L.)." *G3: Genes, Genomes, Genetics*, 10(3): 1113-1124.

Borges, A., González-Reymundez, A., Ernst, O., M. Cadenazzi, O., Terra, J., and Gutiérrez, L. (2019). "Can Spatial Modeling Substitute for Experimental Design in Agricultural Experiments?" *Crop Science* 59(1):44–53.

Brownie, C., Bowman, D. T. and Burton, J. W. (1993). "Estimating Spatial Variation in Analysis of Data from Yield Trials: A Comparison of Methods." *Agronomy Journal* 85(6):1244–53.

Butler, D., Cullis, B.R., Gilmour, A.R., and Gogel, B.J. (2009). "Analysis of Mixed Models for S Language Environments ASReml-R Reference Manual ASReml-R Estimates Variance Components under a General Linear Mixed Model by Residual Maximum Likelihood (REML)."

Casler, M.D. (2010). "Changes in mean and genetic variance during two cycles of within-family selection in switchgrass." *BioEnergy research* 3(1): 47-54.

Casler, M.D. (2015). "Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments." *Agronomy Journal* 107(2):692–705.

Casler, M.D. and Undersander, D.J. (2000). "Forage yield precision, experimental design, and cultivar mean separation for alfalfa cultivar trials." *Agronomy Journal 92*(6):1064-1071.

Clarke, G.P.Y. and Stefanova, K.T. (2011). "Optimal Design for Early-Generation Plant-Breeding Trials with Unreplicated or Partially Replicated Test Lines." *Australian and New Zealand Journal of Statistics* 53(4):461–80.

Cochran, W.G. and Cox, G.M. (1957). "Experimental Designs", 2nd Ed. Oxford, England: John Wiley & Sons.

Coombes, N. (2002). "The Reactive TABU Search for Efficient Correlated Experimental Designs." Liverpool John Moores University.

Covarrubias-Pazaran, G. 2019. "Quick Start for the Sommer Package." 1–12.

Cullis, B. R. and Gleeson, A.C. (1991). "Spatial Analysis of Field Experiments-An Extension to Two Dimensions." *Biometrics* 47(4):1449-1460.

Cullis, B.R., Smith, A. B. and Coombes, N. E. (2006). "On the Design of Early Generation Variety Trials with Correlated Data." *Journal of Agricultural, Biological, and Environmental Statistics* 11(4):381–93.

de S. Bueno Filho, J.S. and Gilmour, S.G. (2003). "Planning incomplete block experiments when treatments are genetically related." *Biometrics*, 59(2):375-381.

de Mendiburu, F. (2019). "Package Agricolae: Statistical Procedures for Agricultural Research." 156.

Eccleston, J.A. and Chan, B. (1998). *Design Algorithms for Correlated Data: In Compstat*. Physica, Heidelberg.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species." *PLoS ONE* 6(5).

Endelman, J.B. (2011). "Ridge Regression and Other Kernels for Genomic Selection with R Package RrBLUP." *The Plant Genome Journal* 4(3):250-255.

Federer, Walter T. (1998). "Recovery of interblock, intergradient, and intervariety information in incomplete block and lattice rectangle designed experiments." *Biometrics* 54(2): 471-481.

Federer, W.T. and Raghavarao, D. (1975). "On Augmented Designs." *Biometrics* 31(1):29-35.

Federer, W.T. and Schlottfeldt, C.S. (1954). "The Use of Covariance to Control Gradients in Experiments." *Biometrics* 10(2):282-290.

Federer, W.T. and Crossa, J. (2012). "I.4 Screening Experimental Designs for Quantitative Trait Loci, Association Mapping, Genotype-by Environment Interaction, and Other Investigations." *Frontiers in Physiology* 3 JUN.

Federer, W.T. (1956). "Augmented (or Hoonuiaku) Designs. Hawaii Plant." *Records* 55:191–208.

Fisher, R.A. (1926). "The Arrangement of Field Experiments."

Fisher, R.A. (1935). *The Design of Experiments*. 1st ed. Available at: https://trove.nla.gov.au/work/7546506?q&sort=holdings+desc&_=1571073338279&versionId=45128278# (Accessed: 14 October 2019).

Gilmour, A. R., Cullis, B. R. and 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):269–93.

Gonçalves, E., St.aubyn, A. and Martins, A. (2010). "Experimental Designs for Evaluation of Genetic Variability and Selection of Ancient Grapevine Varieties: A Simulation Study." *Heredity* 104(6):552–62.

González-Barrios, P., Díaz-García, L. and Gutiérrez, L. (2019). "Mega-Environmental Design: Using Genotype × Environment Interaction to Optimize Resources for Cultivar Testing." *Crop Science* 59(5):1899-1915.

Grondona, M.O., Crossa, J., Fox, P.N., and Pfeiffer, W.H. (1996). "Analysis of Variety Yield Trials Using Two-Dimensional Separable ARIMA Processes." *Biometrics* 52(2):763.

John, J. A. and Eccleston, J.A. (1986). "Row-Column a-Designs". *Biometrika* 73:301-306

Kempton, R. A. (1984). "The Use of Biplots in Interpreting Variety by Environment Interactions." *The Journal of Agricultural Science* 103(1):123–35.

Kempton, R.A. and Howes, C.W. (1981). "The Use of Neighbouring Plot Values in the Analysis of Variety Trials." *Applied Statistics* 30(1):59-70.

Kruijer, W., M. P. Boer, M. Malosetti, P. J. Flood, B. Engel et al., 2015. "Marker-based estimation of heritability in immortal populations". Genetics 199: 379–398.

Lado, B., González Barrios, P., Quincke, M., Silva, P., and Gutiérrez, L. (2016). "Modeling Genotype × Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program." *Crop Science* 56(5):2165–79.

Lin, C.S. and Binns, M.R. (1986). "Relative Efficiency of Two Randomized Block Designs Having Different Plot Sizes and Numbers of Replications and of Plots per Block1." *Agronomy Journal* 78(3):531-534.

Lorenz, A.J. (2013). "Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment." *G3: Genes, Genomes, Genetics* 3(3):481–91.

Martin, R.J. and Eccleston, J.A. (1991). "Optimal Incomplete Block Designs for General Dependence Structures." *Journal of Statistical Planning and Inference* 28(1):67–81.

Masood, M.A., Farooq, K., Mujahid, Y. and Anwar, M.Z. (2008). "Improvement in Precision of Agricultural Field Experiments through Design and Analysis." *Pakistan Journal of Life and Social Science* 6(2):89–91.

Mead, R. (1997). "Design of Plant Breeding Trials." Pp. 40–67 in *Statistical Methods for Plant Variety Evaluation*. Springer Netherlands.

Moehring, J., Williams, E. R. and Piepho, H.P. (2014). "Efficiency of Augmented P-Rep Designs in Multi-Environmental Trials." *Theoretical and Applied Genetics* 127(5):1049–60.

Mramba, L.K., Peter, G.F., Whitaker, V.M. and Gezan, S.A. (2018). "Generating improved experimental designs with spatially and genetically correlated observations using mixed models." *Agronomy*, 8(4):40.

Müller, B.U., Kleinknecht, K., Möhring, J. and Piepho, H.P. (2010). "Comparison of Spatial Models for Sugar Beet and Barley Trials." *Crop Science* 50(3):794–802.

Ould Estaghvirou, S. B., J. O. Ogutu, T. Schulz-Streeck, C. Knaak, M. Ouzunova et al., 2013 Evaluation of approaches for esti- mating the accuracy of genomic prediction in plant breeding. BMC Genomics 14: 860.

Papadakis, J. S. (1937). "Méthode Statistique Pour Des Expériences Sur Champ." *Thessalonike: Institut d'Amélioration Des Plantes à Salonique* 23:1–30.

Patterson, H.D. and Hunter, E.A. (1983). "The Efficiency of Incomplete Block Designs in National List and Recommended List Cereal Variety Trials." *The Journal of Agricultural Science* 101(2):427–33.

Patterson, H.D. and Williams, E.R. (1976). "A New Class of Resolvable Incomplete Block Designs." *Biometrika* 63(1):83–92.

Pebesma, E.J. and Bivand, R.S. (2005). *Classes and Methods for Spatial Data: The Sp Package*.

Pebesma, E.J. 2004. "Multivariable Geostatistics in S: The Gstat Package." *Computers and Geosciences* 30(7):683–91.

Piepho, H.P., Büchse, A. and Emrich, K. (2003). "A Hitchhiker's Guide to Mixed Models for Randomized Experiments." *Journal of Agronomy and Crop Science* 189(5):310–22.

Piepho, H.P., and J. Möehring. 2007. Computing heritability and selection response from unbalanced plant breeding trials. *Genetics* 177:1881–1888.

Piepho, H.P., J. Möhring, A. E. Melchinger, and Büchse, A. (2008) "BLUP for phenotypic selection in plant breeding and variety testing." *Euphytica* 161(1-2): 209-228.

Piepho, H.P., Moehring, J., Williams, E. R. (2013). "Why randomize Agriculture experiment?" *Journal of agronomy and plant science* 199 (2013):374-383

Piepho, H.P. and Williams, E.R. (2010). "Linear Variance Models for Plant Breeding Trials." *Plant Breeding* 129(1):1–8. Piepho, H.P., , Williams, E.R., and Michel, V. (2015). "Beyond Latin Squares: A Brief Tour of Row-Column Designs." *Agronomy Journal* 107(6):2263–70.

Poland, J.A. and Rife, T. W. (2012). "Genotyping-by-Sequencing for Plant Breeding and Genetics." *The Plant Genome Journal* 5(3):92-102.

Qiao C.G., Basford, K.E., DeLacy, I.H., Cooper, M. (2000). "Evaluation of experimental designs and spatial analyses in wheat breeding trials." *Theoritical Applied Genetics* 100:9–16.

R Core Development Team. 2019. "A Language and Environment for Statistical Computing."

Richter, C., and Kroschewski. B. (2012) "Geostatistical models in agricultural field experiments: Investigations based on uniformity trials." *Agronomy journal* 104(1): 91-105.

Robbins, K. R., Backlund, J. E. and Schnelle, K. D. (2012). "Spatial Corrections of Unreplicated Trials Using a Two-Dimensional Spline." *Crop Science* 52(3):1138–44.

Schmidt, P., Hartung, J., Rath, J., and Piepho, H.P. (2019). "Estimating Broad-Sensse Heritability with Unblanced Data from Agricultural Cultivar Trials." *Crop Science* 59(2):525-536.

Smith, H.F. (1938). "An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops." *The Journal of Agricultural Science* 28(1):1–23.

Smith, A.B., Cullis, B.R and Thompson, R. (2005). "The Analysis of Crop Cultivar Breeding and Evaluation Trials: an Overview of Curretn Mixed Model Approaches." *Journal of Agricultural Science* 143:449-462.

Stefanova, K.T., Smith, A. B. and Cullis, B. R. (2009). "Enhanced Diagnostics for the Spatial Analysis of Field Trials." *Journal of Agricultural, Biological, and Environmental Statistics* 14(4):392–410.

Tamura, R. N., Nelson, L. A. and Naderman, G. C. (1988). "An Investigation of the Validity and Usefulness of Trend Analysis for Field Plot Data." *Agronomy Journal* 80(5):712-718.

USDA-ARS. (1992). Germplasm Resources Information Network (GRIN) database. Triticum aestivum L. subsp. aestivum. POACEAE 'Nogal.' Retrieved from Cloud U.S. National Plant Germplasm System website: https://npgsweb.ars-grin.gov/gringlobal/accessiondetail.aspx?id=1457618

Van Es, H., Gomes, C., Sellmann, M., and Van Es, C. (2007). Spatially-balanced complete block desings for field experiments. *Geoderma* 140:346-352.

Velazco, J.G., Rodríguez-Álvarez, M.X., Boer, M.P., Jordan, D.R., Eilers, P.H., Malosetti, M. and van Eeuwijk, F.A. (2017). "Modelling Spatial Trends in Sorghum Breeding Field Trials Using a Two-Dimensional P-Spline Mixed Model." *Theoretical and Applied Genetics* 130(7):1375–92.

White, L., John, J. A. and Williams, E. R. (1996). "Cyclic and Computer Generated Designs." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159(3):625.

Wickham, H. 2016. "ggplot2: Elegant Graphics for Data Analysis." (2d ed) Springer-Verlag New York

Wilkinson, G.N., Eckert, S.R., Hancock, T.W. and Mayo, O. (1983). "Nearest Neighbour (Nn) Analysis of Field Experiments." *Journal of the Royal Statistical Society: Series B (Methodological)* 45(2):151–78.

Williams, E.R. (1986). "A Neighbour Model for Field Experiments." *Biometrika* 73(2):279–87.

Williams, E.R. and John, J.A. (1989). "Construction of Row and Column Designs with Contiguous Replicates." *Applied Statistics* 38(1):149-154.

Williams, E.R. and John, J.A. (1999). "Construction of Resolvable Designs with Nested Treatment Structure." *Biometrical Journal* 41(3):341–49.

Williams, E.R., John, J.A., and Whitaker, D. (2006). "Construction of Resolvable Spatial Row-Column Designs." *Biometrics* 62(1):103–8.

Williams, E.R., Matheson, A. C. and Harwood, C. E. (2002) "Experimental Design and Analysis for Tree Improvement." *Experimental Design and Analysis for Tree Improvement.* (Ed. 2).

Williams, E.R. and Piepho, H.P. (2013). "A Comparison of Spatial Designs for Field Variety Trials." *Australian and New Zealand Journal of Statistics* 55(3):253–58.

Williams, E.R., Piepho, H. P. and Whitaker, D. (2011). "Augmented P-Rep Designs." *Biometrical Journal* 53(1):19–27.

Wu, T.X., Mather, D.E., and Dutilleul, P. (1998). "An application of geostatistical and neighbor analsyses to data from plant breeding trials." *Crop Science* 38:1545-1553.

Yates, F. (1936). "Incomplete Randomized Blocks." *Annals of Eugenics* 7(2):121–40.

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., Olsen, M., Prasanna, B.M., Crossa, J., Yu, H., and Zhang, X. (2017). "Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 Bi-Parental Tropical Maize Populations." *Frontiers in Plant Science* 8.

Zimmerman, D. L. and Harville, D. A. (1991). "A Random Field Approach to the Analysis of Field-Plot Experiments and Other Spatial Experiments." *Biometrics* 47(1):223.

Zystro, J., Colley, M. and Dawson, J. (2019). "Alternative Experimental Designs for Plant Breeding." *Plant Breeding Reviews* 42:87–117.