Dear Editor,

We would like to thank you and the reviewers for the work you put in the review of this manuscript. We have made corrections in order to respond to comments from Pierre Druilhet and David Makowski, which we detail below.

Comments of reviewers are in *italic*.

1 Response to Pierre Druilhet

1) l.160. It is not clear to the reader which variable is used as the blocking variable: the farm or some spatial or temporal variables or other blocking variables.

The block effect is a within-trial spatial variable. We've tried to be clearer lines 124–127.

2) l. 216 : Is there any point in using a truncated normal distribution for ϵ rather than a usual Gamma or inverse-Gamma distribution?

According to Gelman (2006), the estimation of a variance parameter is more sensitive to an inverse-Gamma(ϵ, ϵ) prior than to a half-Gaussian or half-t prior, in particular when this variance parameter is small.

3) l.216-217 : A Gamma distribution is chosen for ν with the constraint $\nu > 2$. Such a constraint generally reduces the efficiency of the MCMC used for inference. An alternative might be to use the parameter $\nu = 2 + \nu$ where ν is Gamma-distributed.

Thank for suggestion. We have changed this and the calculations seem to be faster (6min instead of 9min).

4) l.267 : I don't get the same variance decomposition since the second and third terms are not independent (given the hyperpriors). They may change the interpretation of the results.

It turns out that the terms θ_j and $\eta_i\theta_j$ are not correlated. Since η_i and θ_j are independent, $\mathrm{E}(\eta_i\theta_j) = \mathrm{E}(\eta_i)\,\mathrm{E}(\theta_j) = 0$. Thus, both θ_j and $\eta_i\theta_j$ are centered. It follows that their covariance is equal to

$$\mathrm{Cov}(\theta_j, \eta_i\theta_j) = \mathrm{E}(\theta_j \times \eta_i\theta_j) = \mathrm{E}\left(\eta_i\theta_j^2\right).$$

Since η_i and θ_j^2 are independent, we obtain

$$\mathrm{Cov}(\theta_j, \eta_i\theta_j) = \mathrm{E}(\eta_i)\,\mathrm{E}\left(\theta_j^2\right) = 0,$$

so that θ_j and $\eta_i\theta_j$ are not correlated. The absence of correlation between θ_j and $\eta_i\theta_j$ has been mentioned in the article line 220.

Since α_i, θ_j, $\eta_i\theta_j$ and ε_{ij} are not correlated, the variance of an observation is equal

$$\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\alpha_i + \theta_j + \eta_i\theta_j + \varepsilon_{ij}) = \mathrm{Var}(\alpha_i) + \mathrm{Var}(\theta_j) + \mathrm{Var}(\eta_i\theta_j) + \mathrm{Var}(\varepsilon_{ij}).$$

Since η_i and θ_j are independent, we obtain

$$\mathrm{Var}(\eta_i\theta_j) = \mathrm{E}\left(\eta_i^2\theta_j^2\right) = \mathrm{E}\left(\eta_i^2\right)\mathrm{E}\left(\theta_j^2\right) = \sigma_\eta^2\sigma_\theta^2.$$

It follows that

$$\mathrm{Var}(Y_{ij}) = \sigma_\alpha^2 + \sigma_\theta^2 + \sigma_\eta^2\sigma_\theta^2 + \mathrm{Var}(\varepsilon_{ij}).$$

Thus, we think that there is no mistake in the expression of $\mathrm{Var}\,(Y_{ij})$.

This variance decomposition for a hierarchical FW model is a result of the article. To our knowledge, it is not present in the literature. So, we added a section on variance decomposition in the discussion, lines 451–463.

2 Response to David Makowski

– The paper addresses an interesting topic concerning participatory plant breeding. It aims at comparing various Bayesian models for analyzing data collected in a highly unbalanced designs with a high share of missing data. I found the paper interesting but I had a lot of difficulties to understand what were the precise objectives of the on-farm trials analyzed by the authors. They didn't explain clearly enough what were the objectives of the genotype selection they aimed to perform with their on-farm trials.

In this PPB programme, selection goals are specific to each farmer and will be strongly based on the results of their trials, it might include both selection among populations to keep the most adapted to its farm and possibly mass selection within the most interesting ones. However, when farmers want to integrate and test new populations in their trials, they critically lack informations on which populations to test. So, one aim of the research is to provide information on varieties across all the trial network and potentially help in varietal choice. We've tried to make it clearer lines 56–65.

The main objective of this study was to develop new statistical methods for analyzing on-farm trials. Therefore, we emphasize statistical models more than the application to the wheat data. We have clarified that the main objective was to develop statistical methods (line 109). Moreover, the statistical methods have been moved before the application in the Materials and Methods section.

– In addition, the description of the data is quite unclear and it is difficult to understand the different types of designs included in the datasets, and how this diversity could be properly handled.

We have simplified the description of the experimental designs (Table 3). We grouped together the designs where the blocks were incomplete and those where no blocks were present. As this article deals with the analysis of series of trials rather than individual trials, we prefered to limit the description of the designs of individual trials.

We have also explained that the data analysis was a two-stage process (first step, germplasm means adjusted for block effects are calculated for each trial, second step, the analysis is carried out at the scale of the trial network) lines 124–127, and have pointed out its limitations lines 444–449.

– Moreover, some aspects of the modelling framework are not fully justified. In particular, I don't understand why your Bayesian models would be more suitable to analyze unbalanced data obtained in non-randomized experiments than other types of models.

Bayesian modeling allows us to better:

- handle data imbalance using hierarchical models that shrink estimates, for example FW regression coefficients were first estimated in hierarchical form by Lian and de los Campos (2016) in a Bayesian framework,
- handle extreme data using Student distributions (no GxE model that uses such distribution of residuals in the literature to our knowledge),

- compare models using LOO cross validation (that is very difficult to estimate quickly with complex frequentist models).

– I don't see why your models would solve the issues related to the high level of heterogeneity of your datasets.

Our models take account of two sources of heterogeneities :

- the heterogeneity of germplasm sensitivities to environments, through the parameters η_i,
- the heterogeneity of within-trial error variances and replications, and more generally extreme data, through the t distribution (see below).

– L17-20: The results presented in the abstract are qualitative. More quantitative results would be useful. In particular, it is unclear whether the results of the proposed model were accurate enough to be used in practice.

We have added quantitative results in the summary. This article provides general knowledge on varieties, that completes the within-trial variety comparisons (average and stability of different traits across environments, FW sensibility) lines 56–65.

– L17-20: Because of the use of an unbalanced design and of the lack of randomization, there is a risk of bias that may lead to errors in the ranking of the cultivars. I would be useful to reflect on this aspect in the abstract.

We have checked that germplasm main effects, environment main effects and germplasm sensitivities were identifiable for the additive and FW models in our application. Thus, the use of a very unbalanced design did not bias estimates through a lack of identifiability or through the confusion of some model parameters. A paragraph on identifiability has been added in the article lines 297–306.

All varieties were randomized within farms, but not randomized between farms. At the outset, the farmers started whith fairly random assignment of populations, but owertime it is true that farmers would tend to keep the ones that work best for their farm. Accurate predictions would accelerate this and create more biais, which might need to be addressed in more mature networks. We have clarified that at lines 270-271.

– L20: « mixtures ». Do you mean « genotype mixtures » ? Risk of confusion with "mixtures of probability distributions ».

We have removed the sentence.

– L30: Unclear (environments always depend on pedoclimatic conditions, weathers etc.). I guess you mean that crops may be impacted by a greater diversity of limiting factors in OA than in conventional agriculture.

We have added "more" line 7. " [Fields managed using organic farming practices] are **more** sensitive to on pedoclimatic conditions, yearly weather, farmers' management " practices and interactions between these factors than conventional management practices.

– L41-42: this is true only if we are able to test a large number of G in a large number of E. Otherwise, it does not allow you to estimate G x E, or for a very limited of G only. In addition, randomized trials are often more difficult to conduct on farms. The benefit is thus uncertain.

Part of what the model is trying to predict is performaance on farms which have not yet tested that population, and that relies on the abilities to predict GxE to some extent, even if the selection that matters to the farmer is G + GxE 56–65 lines.

There is uncertainty about the effectiveness of decentralized selection, just as there is uncertainty about the benefits of centralized selection if the production environments are too different from the selection environments.

– L54, 58, 65: the term « populations » is not clear. Maybe you mean « non-hybrid genotype » but I am not sure. This needs to be clarified in order to allow non-specialists to undertand.

We have removed the term « population » from the introduction and have defined it lines 118–122. A population variety is defined as a set of individuals which may be genetically different but which are derived from the same selection process conducted in a certain environment using certain agronomic practices.

– L66: « as very few populations were present»: this appears to be in contradiction with the sentence « a large number of populations was evaluated » (L58). Moreover, for a non-specialist, it is not fully clear what is the difference between a genotype and a germplasm.

There are few varieties per environment (Farm X Year) but many varieties in the entire trial network (206). We defined a germplasm lines 118–122 as « any biological entity whose individuals are derived from the same breeding process, including varieties registered in the official catalog, landraces, historic varieties, mixtures or populations stemming from crosses » .

– L67-68: « genotype main effect and stability ». Unclear. Need to be explained. In fact, the concrete objective of the network of trials is unclear.

We have replaced genotypic effect by genetic effect lines 72 and 80.The objective of the network of trials and of the research study have been specified lines 56-68 and 109.

– How do you want to use the data to select genotypes?

The data collected in each farm are used by farmers to choose the varieties they want to cross, mix or grow in their farm. However, this work aims at analysing the whole dataset to improve the robustness and the accuracy of some trait estimates and produce additional information on varieties like their sensitivity to environment potential or their stability that could help farmers to select genotypes (lines 56–65).

– What are the criteria relevant for the assessment of the genotypes?

The traits measured were chosen through a collaborative process (see Dawson et al., 2011, for more information). They are commonly observed by farmers when making their choices. Depending on the case, other traits such as yield are sometimes measured, but the quantity of data was not sufficient for a GxE analysis.

– Do you want to select one genotype in each farm or select one cultivar for a group of farms sharing similar E?

Each farm must produce seed of their own for unregistrered varieties, so there is less benefit to selecting a variety across many farms. Farmer is interested in selecting the population variety the most adapted to his farm, but if we could characterize germplasms for their response to environmental conditions, it might help farmers sharing similar environments to access faster to relevant germplasms. For that, FW parameter can help to identify the most suitable germplams according to the potential of the environment.

– L72-77: in this description, it is unclear what are the fixed parameters. Although, only some of the random parameters are listed, making you text hard to follow. In particular, it is not clear how the genotype sensitivities are defined. Do they correspond to genotype-specific random environmental effects (i.e., random GxE interactions). Definitions are provided much later in the paper, but this is not an optimal way to organize your paper.

The presentation of fixed/random effects has been revised. We've tried to be clearer line 77.

– Section 2.2.1: I don't understand what designs were used in the categories Regional, Statellite and unreplicated shown in Table 2.

We have simplified the presentation of experimental designs (Table 3).

– 161: « obvious outliers were excluded ». Define « obvious ».

We have tried to be more precise (line 283): « Outliers with respect to agronomic knowledge of the traits were excluded (for example, a plant taller than 3 meters) ».

– The models are presented in section 2.3. The models look quite standard, with main effects and interactions. It is difficult to understand why these models are able to analyze on-farm trials with unbalanced and missing data.

We proposed various statistical improvements (most of which are not widely available in the literature) and checked their ability to better handle data imbalance and extreme data using two criteria (LOO cross validation and estimation precision). In addition, we tried to take account of the GxE interaction with the parsimonious FW model, which is well suited to data imbalance.

– In 2.3.1, based on the equations, « static stable » corresponds to absence of environmental effect, while « dynamic stable » corresponds to absence of interactions, but this is not how the authors presented these two types of stability in the text.

There is the statistical vision and the plant breeding vision. We prefer the definition of the breeders as it is more pragmatic in our opinion. We have outlined these definitions in the introduction lines 65–68.

– 189: "Finlay and Wilkinson (1963) defined their coefficient as". Which coefficients? The model FWs is not based on a hierarchical structure. I don't understand what assumptions are made on the parameters in this model. Did you assume a specific parameter value for each germplasm and environment? What were the priors? This needs to be clarify.

We have removed the sentence "Finlay and Wilkinson (1963) defined their sensitivity coefficient as $b_i = 1 + \eta_i$" which could indeed create confusion. Compared with the article of Finlay and Wilkinson (1963), there is a difference of 1 between the way we define the regression coefficient on the environments (we used the formalism of Perkins and Jinks, 1968, we made this choice to help with variance decomposition). In both cases this parameter depends on the varieties. The way in which the parameters are estimated and the associated priors are defined in the Section 2.1.2. Weakly informative priors were used in model FWs (please, see lines 188–191).

– 2.3.2. Residual terms. The different types of designs included in your dataset are so different that it is unlikely that they could share the same residual variance. However, this is the assumption made in your models; you used a unique residual variance for all designs.

We have used homoscedastic models for three reasons.

- The within-trial residual variances are difficult to estimate in our application since few germplasm are replicated within the trials. We have clarified that most of the germplasm were not replicated within the trials lines 150–154 and line 270.
- The between-trial residuals are due not only to the within-trial residuals, but also to the genotype-environment interactions that are not explained by the term $\eta_i\theta_j$.
- Student residuals better handle data heteroscedasticity than normal residuals, since they can be written as

$$\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_{ij}^2\right), \quad \sigma_{ij}^{-2} \sim \Gamma(\nu/2, \nu\sigma_\varepsilon^2/2).$$

  We have clarified this point in the article lines 168–172.

We agree that this simplification might reduce the efficiency of the analysis (Welham et al., 2010; Yates and Cochran, 1938), and that it would be interesting to extend the method to better take account of the heterogeneity of the within-trial residual variances. The discussion on this point has been revised lines 444–449.

– 2.3.5. The use of leave-one-out cross validation is not fully justified as the data are not independent, in particular the data collected in the same environments are correlated.

It turns out that the leave-one-out cross validation criterion can be used to compare models whose observations are conditionally independent only. Our models satisfy this assumption. For example, with our models, the data collected in the same environment are correlated through environment effects, but they are independent given germplasm main effects, environment main effects, germplasm sensitivities, σ_ε and ν. Thus, this criterion can be used to compare our models. Please, see Vehtari et al. (2017) for more information.

– 2.3.6. Variance decomposition. I guess this decomposition is not valid for the model FWs. Please clarify.

Indeed, our variance decomposition is only valid for fully hierarchical models such as FWHs (see comment 4 of Pierre Druilhet). We have clarified this point line 219.

– L223: Here as well, because of the heterogeneity of the designs included in your dataset, it is unlikely that the between-germplasm and between-trial variances are the same for all types of trials.

The variances σ_α^2 and σ_θ^2 defined line 223 are the variances of germplasm and environment main effects. These variances are defined at the level of the series of trials rather than at the level of the trials. Thus, we do not make the assumption that these variances are the same for all types of trials in the article.

– 281-283: Here, you define static and dynamic stability. But these definitions should be provided much earlier, the first time you are using these concepts.

We now provide these definitions in the introduction lines 65–68.

– 290: why are they approximations? What would be their exact expressions?

This sentence has been revised and moved to the discussion (line 509). We meant that comparisons between germplasm stability indicators only take account of the part of GxE interactions explained by the FW

model.

– L292-294: It is unconsistent to use frequentist statistical tests while your analysis is based on Bayesian models. You could use your Bayesian inference to derive credibility intervals for all quantities of interest.

We have followed your suggestion and changed our method to use a Bayesian method to compare types of germplasm pairwise.

– 3.1.1. This is interesting but, as mentioned above, the LOO CV does not seem well adapted to the clustering nature of your dataset.

Please see our previous response on this point.

– 3.2. The title is very general and could be applied to any part of the paper.

We have removed this title.

– I found the results on the student distributions interesting, especially because they seem able to handle extreme data, but I am wondering why the authors only considered student and gaussian distributions and not others.

We have used Student distributions for three reasons.

- These distributions have heavier tails than normal distributions and better handle extreme data.
- They better handle data heteroscedasticity than normal distributions (please, see the point on homoscedasticity above).
- They are rather easy to implement, since they are available in programs such as stan. For example, the exponential power family also includes distributions with heavier tails than normal distributions (Box and Tiao, 2011). However, these distributions are not available in stan or jags, so one would have to program a MCMC algorithm from scratch to implement such distributions.

– It was difficult to see how the models could be used in practice and how their use would change the practical recommandations made to farmers.

The data of the French PPB program were analyzed each year using within-trial analyses only, which allowed to provide support to farmers for choosing the best germplasms among those grown on their farms. The objective of these new models would be to complement these analyses with between-trial analyses. Results on the global agronomic values of germplasm (germplasm average performances and sensitivities) could help farmers to choose new populations to integrate in their trials as they critically lack informations on which populations to test. These results might also help researchers to have a better understanding of the biological processes involved in this on-farm breeding program. For example, a farmer worried by economical risks may stop testing a germplasm, if this germplasm turns out to be sensitive to environments. Researchers may be interested to know that mixture populations tend to have lower sensitivities to environments. However, it is difficult to anticipate how useful these new information will be in this particular application.

The main objective of this article is to present new statistical methods for analyzing on-farm trials. They could be applied to other series of on-farm trials than the French PPB program.

References

Box GEP and GC Tiao (2011). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Dawson JC, P Rivière, JF Berthellot, F Mercier, P de Kochko, N Galic, S Pin, E Serpolay, M Thomas, S Giuliano, and I Goldringer (2011). Collaborative Plant Breeding for Organic Agricultural Systems in Developed Countries. *Sustainability* 3, 1206–1223. https://doi.org/10.3390/su3081206.

Finlay KW and GN Wilkinson (1963). The Analysis of Adaptation in a Plant-Breeding Programme. *Australian Journal of Agricultural Research* 14, 742–754. https://doi.org/10.1071/AR9630742.

Gelman A (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1, 515–533.

Lian L and G de los Campos (2016). FW: An R Package for Finlay–Wilkinson Regression That Incorporates Genomic/Pedigree Information and Covariance Structures Between Environments. *G3 Genes|Genomes|Genetics* 6, 589–597. https://doi.org/10.1534/g3.115.026328.

Perkins JM and JL Jinks (1968). Environmental and Genotype-Environmental Components of Variability. *Heredity* 23, 339–356.

Vehtari A, A Gelman, and J Gabry (2017). Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC. *Statistics and Computing* 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Welham SJ, BJ Gogel, AB Smith, R Thompson, and BR Cullis (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Australian & New Zealand Journal of Statistics* 52, 125–149.

Yates F and WG Cochran (1938). The Analysis of Groups of Experiments. *The Journal of Agricultural Science* 28, 556–580. https://doi.org/10.1017/S0021859600050978.