Fugitive road dust alters annual plant physiology but perennial grass growth appears resistant

Response to Review

We appreciate the opportunity to continue the revision process for our submission to *Plant Ecology*. Please find below responses addressed to each reviewer in turn.

Reviewer 1

We appreciate the comments on the additions to the ecological context.

Each of the line comments have been corrected as the reviewer suggested.

We appreciate the reviewer's interest in properly reporting results. Indeed, we share the critique of p-values and null hypothesis testing, which in fact motivated our use of confidence intervals. In light of the reviewer's confusion about how CIs are presented and discussed, we've edited for clarity.

Reviewer 2

We've structured our response by the themes addressed in the review:

Random effect terms The reviewer has provided a thorough tutorial on building robust random effects terms. In our paper, we state that we developed our random effect terms to encompass known sources of potential random variation. The reviewer appears to at least agree with us that the relevant sources have been identified. In conducting these analyses, I have found that different formats of random effect construction do change how variation is allocated among the included terms, but these variations do not affect the amount of variation

In my experience with mixed-effect models, I am most interested in ensuring that random variation is accounted for, but not parsing the error structure within the random effect. I believe this applies here—we are not interested in which of the terms contribute more or less to the overall variation, just that the overall variation is accounted for. If we were in fact interested in parsing the error structure, we would need to give much greater scrutiny to how the random error is structured, not just that it has been accounted for. This explains our use of the random effect term given in our script: It appropriately accounts for random variation contributed by the modeled sources, but does not provide the unnecessary information about which term contributed what proportion of that variation.

Reviewer #2: Review: VEGE-D-20-00268 (first revision)

Thank you very much for taking the time to reanalyse the data and revise the ms.

I read the responses by the authors to reviewers' comments first and then went over the entire ms once. Then I spent several days considering different options for my approach to the revised ms.

I reiterate that I appreciate very much for the responses by the authors. When I read the previous version of the ms, I had a difficult time understanding the exact design of the experiment, and at the time, I decided to focus my comments on the m & m section. Thanks to the authors, with the improved m & m section, I can now review and comment on the entire ms.

In this review, I will use "AIC" to mean "AICc".

The authors insisted that they prefer the model selection approach using AIC rather than testing for terms in the full model using likelihood ratio tests. I respect their opinion.

The authors explained that the statistical data analysis followed a principle developed by Cheng et al. (2010). I have not read Cheng et al (2010); however, I have some questions about the statistical data analysis described in the ms.

An example of the model used in the analysis is shown below. conc0 ¡- lme4::lmer(diff 0 + (1—block:round:pot)) In this model, the random effect term is specified as: pots are nested within rounds, and rounds are nested within blocks. For the sake of explanation, let's assume that the specification of the random effect term is correct for the time being. To me the model above is strange. Normally, diff 0 + (1—block) + (1—block:round) + (1—block:round:pot) is the starting full model. First, one would test to see if the last term is needed. So, one would compare the full model and the reduced model below. diff 0 + (1—block) + (1—block:round) If the reduced model was much better (say, having much smaller AIC) than the full model (meaning that variation due to pots within each round is not contributing significantly to the model), then one would go on and test for significance of the second term in the random effect. The above reduced model would be compared against the reduced model below. diff 0 + (1—block) The process of model selection described here is the STANDARD METHODOLOGY. Existence of models like diff 0 + (1—block:round:pot) is very strange to me indeed. Did Cheng et al. (2010) explain how one can get to a model like this? What is their theoretical justification? If there was a theoretically valid explanation, I am very keen to learn about it. If there is no theoretically valid explanation, then I think the authors would be better off following the standard process of model selection.

When I read the earlier draft, I did notice the diff 0 + (1—block:round) being strange. I did make a suggestion to deal with spatial autocorrelation and temporal autocorrelation separately. By making the suggestion, I was hoping that the authors will arrive at a proper model. However, the authors came back with the model: diff 0 + (1—block:round:pot)

Onto my second point about models: Before getting onto my point, please accept my sincere apology for not being precise enough myself. In my previous set of comments, I suggested to build models incorporating repeated measurements. I was not precise enough because "repeated measurements" can mean different things these days. In a sense, authors have been correct because pots are included in the model as (1—block:round:pot). This is how some books and websites explain how to specify the random effect with repeated measurements. What I have meant, however, is the repeated measurements over time, and these days, these are called longitudinal data. I did suggest the correct model for analysing the longitudinal data: For example, for a given species, concentration 0 + date + (date—pot) My lame excuse for being imprecise is that when I have learned about the analysis of "longitudinal data" more than 30 years ago, it has been called the analysis of "repeated measurements". The term "nested analysis" has been used for what is now called "analysis of repeated measurements". So, I still have a bad habit of using the "longitudinal data" and "repeated measurements" interchangeably. I am very sorry about that. If one looks up the term "analysis of repeated measurements" on the internet, one would find that the "longitudinal data" and "repeated measurements" are still used interchangeably by some.

Anyhow, UP TO THIS POINT, the model for the analysis in this study should be as follows: diff 0 + round + (1—block) + (round—pot)

My third point about models: In my previous set of comments, I suggested options for dealing with spatial autocorrelation. However, the authors' response was that such approach was unnecessary. I still think that it is worthwhile to look at relevance of spatial autocorrelation at some scales smaller than the block. I suggested different options. The simplest of my options is as follows: Relevance of spatial autocorrelation at some scales smaller than the block can be achieved easily by comparing the two models below. diff 0 + round + (1—block) + (1—block:plot) + (round—pot) full model diff 0 + round + (1—block) + (round—pot) reduced model Here, the "plot" is the group of six pots that are the unit of dusting treatment (being in the same tent). If the reduced model was better (say, having a smaller AIC) than the full model, then the next step is to compare the following to models: diff 0 + round + (1—block) + (round—pot) and diff 0 + round + (round—pot) Whichever model with a smaller AIC should be chosen as the model for data analysis.

In their responses, the authors stated that "we see no evidence that some sort of spatial autocorrelation occurred on the greenhouse benches and have never heard of such a spatial analysis of pots on benches". I am not suggesting that there is spatial variation at the scale smaller than the benches. However, in field and greenhouse experiments, the traditional approach is to use one of the standard experimental designs to account for spatial variation. If the experiment did not use a standard design (as in this study), I am suggesting that one can look to see if spatial variation is present. One may never know that spatial variation existed until one looks for it. I wonder why the authors have decided that there was "no evidence that some sort of spatial autocorrelation occurred on the greenhouse benches" without even looking for it. I am also wondering why the authors have included the block as a random factor without even testing for its significance. As for "have never heard of such a spatial analysis of pots on benches", there may or may not be published experiments using such a spatial analysis of pots on benches; however, there are field experiments with the analysis incorporating spatial variation at the scale smaller than blocks. These are called analyses with incomplete blocks. I happen to have published a paper describing one such study (Floyd et al. 2002 Agricultural and Forest Entomology 4:109-115). I have uploaded a PDF for the authors. When I did the analysis back in the 20th Century, linear mixed models were still in development, and we have not been able do what we can today. I consulted a couple of statisticians with wealth of experience using linear mixed models and settled on the linear mixed model being used in the paper.

I am puzzled by the approach to data analysis. Figures 2 and 3 show responses that are species-specific (the top part in both figures) and response of all species combined (the bottom part). Did Cheng et al. (2010) suggest this type of analysis? I believe that the standard procedure in statistical data analysis is that if a main effect is significant (in the ms, species are showing species-specific responses), then DO NOT proceed any further (meaning that do not do the analysis combining all species). If species are showing species-specific responses, there simply is no statistically valid reason for examining the combined response. Also, there is no biological meaning in showing the mean response of all species combined. Species are showing different responses. That is that. Looking at the top part of figures 2 and 3, it seems that whether there is a species-specific response depends upon variables. On the one hand, with stomatal conductance, the dust effect is not significantly different from zero across all seven species. On the other hand, with photosynthetic yield, dust effect is significantly positive in pinto bean and sunflower, significantly negative in maize and lentil, and not different from zero in barley, wheat, and sorghum in figure 2. So, for photosynthetic yield, there is no point in showing the result of combined analysis. The same for leaf temperature and chlorophyll concentration in both figures 2 and 3. In a strictly ideal world, one should show the combined

results for stomatal conductance and species-species results for photosynthetic yield, leaf temperature, and chlorophyll concentration. However, in this case, it would probably be simpler to present only the species-specific results for all four variables.

I am not sure what the author intend to show in Tables 2 and 3. Do these tables show the process of model selection? Or, do they show contributions of different fixed-effect terms in the best model? Either way, please read following paragraphs carefully.

I am also puzzled by the results in Table 2. Did Cheng et al. (2010) suggest this type of analysis? As explained earlier, the standard approach to model selection is to start from the full model. (I took all R scripts below from the supplemental information. For the sake of simplicity, I am not repeating my explanation about the random-effect term here, but my explanations earlier about the random-effect term also apply here, of course.) The full model is: scale(lconc) 0 + spp * t_c + (1—block:date:pot) If we fully spell out this model, it becomes as below: scale(lconc) 0 + spp + t_c + spp:t_c + (1—block:date:pot) The spp:t_c term is the species x dust interaction. So, the first comparison should be between the following two models (and, as far as I can tell, the authors have done this): scale(lconc) 0 + spp + t_c + spp:t_c + (1—block:date:pot) scale(lconc) 0 + spp + t_c + (1—block:date:pot) If the full model is much better (meaning having much smaller AIC value) than the reduced model, then one really does not need to look at other reduced models such as: scale(lconc) 0 + spp + (1—block:date:pot) or scale(lconc) 0 + t_c + (1—block:date:pot) Alternatively, if one wants to look at significance of each term in the full model, simply do this: summary(fullmod) to look at random-effect terms and anova(fullmod) to look at the fixed-effect terms: the main effects and the interaction. In this series of analysis, I am also puzzled by some of the models that the authors have chosen. Specifically, why do the following models include the intercept, while the full model and the reduced model without the interaction do not include the intercept? scale(lconc) spp + (1—block:date:pot) scale(lconc) t_c + (1—block:date:pot) To me, the correct comparisons would be as follows: scale(lconc) 0 + spp + t_c + (1—block:date:pot) and scale(lconc) 0 + spp + (1—block:date:pot) for looking at the dust effect. Also, scale(lconc) 0 + spp + t_c + (1—block:date:pot) and scale(lconc) 0 + t_c + (1—block:date:pot) for looking at the species effect.

Also, why did the authors choose the model with just the intercept as the null model? scale(lconc) 1 + (1—block:date:pot) What is the justification? Once the best model (the one with the smallest AIC), then all other models were compared against the best model (see below). So, there is no statistically valid meaning in using this model as the null model. Moreover, there is no biological meaning in this null model, either. Isn't this model simply unnecessary and meaningless?

Looking at Table 2, authors are calculating the change in AIC by subtracting the AIC of each reduced model from the best model. For example, in chlorophyll concentration, the change in AIC between the species only model (AIC = 4140.6) and the best model (i.e., species x dust model: AIC = 4023.2): change in AIC = 117.4 = 4140.6 - 4023.2. This suggests that the comparisons made are between the following two models. scale(lconc) 0 + spp + t_c + spp:t_c + (1—block:date:pot) and scale(lconc) spp + (1—block:date:pot) However, the correct comparison should have been as follows: scale(lconc) 0 + spp + t_c + (1—block:date:pot) and scale(lconc) 0 + spp + (1—block:date:pot) And, the result of this comparison reveals the effect of dust (NOT species).

Theoretically speaking, comparisons of models using AIC is best performed between nested models, although there is no law against using AIC in comparisons of non-nested models (and apparently many people have been comparing non-nested models without realising it). It is said that when comparing non-nested models, results are not as reliable as in comparing nested models. For example, the following is a comparison of two nested models: scale(lconc) 0 +

spp + t_c + (1—block:date:pot) and scale(lconc) 0 + spp + (1—block:date:pot) However, the following is a comparison of two non-nested models (an example taken from table 3): recovery 0 + species*trt + (1—block:event:pot) recovery 0 + photo*trt + (1—block:event:pot)

The theoretical reason for this is as follows: AIC has two parts: the -2*log-likelihood (LL) and penalty (p): AIC = LL + p. However, there in fact is another term (c), so AIC is actually AIC = LL + p + c. The reason why this c is omitted in the formula for AIC is that c is the same for all nested models. So, when one is comparing two nested models, the cs cancel out. Thus, there is no need to include the c in the formula. However, when one is comparing two non-nested models, the two cs may not be the same. Please look up a mathematical derivation of the AIC formula, if you do not believe me.

The above explanation about model comparisons also applies to Table 3.

I am very sorry, but I am not convinced that the approach of comparing models in tables 2 and 3 is useful. We choose a model because we want to see the statistical results from the model. Comparing models should not really be a part of the result section. What I (and probably most readers of this paper) want to know is whether the dust application has effects on physiological parameters and whether these effects vary among the species in the study. One can easily achieve this by simply doing the "anova(fullmod)" as I have suggested above. This will tell everyone whether the species x dust interaction is significant, whether the species effect is significant, and whether the dust effect is significant. Simple and clear-cut. Moreover, this procedure is theoretically valid. May I also add an explanation here that there are two schools of thought about choosing the model for the analysis using summary() and anova(). One school of thought is modelling-driven and choose the best model (say, one with the smallest AIC) and do summary(bestmod) and anova(bestmod). The other school of thought is design-based and choose the full model (even if the full model did not have the smallest AIC) and do summary(fullmod) and anova(fullmod). Statisticians have been arguing about which school of thought is better for decades. As far as I can see, arguments by both schools are equally valid. As long as one is consistently choosing the school, it should be okay. As far as I can tell, the authors of this ms seem to belong to the modelling-driven school. That is fine by me. My point is, however, please do not do the model comparison as have been done in tables 2 and 3. Just choose the best model, do anova(bestmod), and show us whether species, dust application, and the interaction are significant. This is slightly off the topic, but when comparing models and looking at significance of effects, one would use effects models (meaning the models with the intercept) rather than means models (meaning the models without the intercept). Means models are useful for calculating treatment means (for example, the mean values of a physiological parameter for the species used in the experiment).

I do not see the need for Table 1 in the supplemental information. As explained earlier, it is clear that the responses of species to dust application vary among the physiological parameters. There is no need to put a p-value, I don't think. Also, I am puzzled by the fact that the authors included "species" in the random effect (and the fact that species were nested within blocks, too). What is the justification? Why would the authors want to make an inference about species that have not been used in this study? This is a technical point, but if the authors just wanted to see the significance of the main effect ("response" in the model), then just do diff1 ¡- lme4::lmer(diff 0 + response + (1—block:round:spp:pot), data= diff_dat) anova(diff1) Please note that I would use REML here. Please see my explanation below.

When using likelihood ratio tests, ML should be used to compare models with different fixed-effect terms. However, the authors are not using likelihood ratio tests to compare models. Why do the authors not use REML? Did Cheng et al. (2010) suggest this? Please explain.

In my previous set of comments, I have asked the authors to clarify the bottom part of figure 2 in the supplemental information. The authors came back and explained that, in all figures, "different colors indicate the mean and standard error of undusted (orange) and dusted plants (blue)". Thank you very much for the added explanation. However, the bottom part of the figure 2 is still confusing to me. The reason for my confusion stems from the fact that the labels on the horizontal axis say "no-dust" and "dust". Now that the authors have clarified what different colors meant, I can guess that the labels on the horizontal axis might mean "pre-dusting-treatment" and "post-dusting-treatment", respectively. If I am correct, then would you please change the labels accordingly?

Thank you very much for checking the normality of residuals (This process is called model checking). Would you please include "plot()" after each model in the R script in the supplemental information? I want more people to understand importance of checking residuals whenever running data analysis using statistical models. I was surprised to find that not a single outlier was removed from the data after model checking. Many datasets in ecological studies include at least some outliers that are needed to be removed.

Lastly, I would ask the authors to consider my comments carefully. Please understand that I have no intention of being antagonistic. I think that results in this ms have much to offer. I just want the data to be analysed properly. There may be research scientists who read this paper and repeat the method of data analysis used in it without consulting statisticians or study statistics textbooks. This is why I provide critical and careful review on statistical data analysis in all manuscripts that I review.

With development of R and increasing popularity of it, more and more research scientists are discovering the statistical data analysis using models such as liner mixed models, generalized liner models, generalized liner mixed models, and generalized additive models. However, mathematical background of these models is well above and beyond most research scientists (except those who are in numerically oriented fields such as physics and engineering). Fortunately (or unfortunately), there are many books on how to analyse data using R (with R scripts), and it is easy for research scientists without sufficient background in statistical theories to use these methods of data analysis these days. Consequently, there is ever increasing number of papers with incorrectly or inappropriately analysed data being published. Therefore, "previously-published framework of data analysis" means little to me. As far as I can see, a large proportion of published papers in ecology in recent years include incorrect or inappropriate data analyses, especially papers that use lme4 or MASS to model data using LMM, GLM, or GLMM.

Please do not just assume that data analyses in published papers are all correct. Most reviewers and editors are simply unable to evaluate the data analyses these days. Best thing a research scientist can do is to consult statisticians when designing the next study.

I have two comments on issues other than statistical data analysis. In my previous set of comments, I mentioned that the plants used in the experiment might have been stressed (judging from Fv/Fm ¡ 0.8). The authors appear to have decided to ignore this comment. However, I think that it is rather important. I would like to see some discussion on what might have been the cause of the stress and whether the effects (or no effect) of dust on physiological parameters seen in this study can be reproduced if plants were not stressed and why. Also, would the species used in this study experience similar levels of stress without the dust in the field?

I think that the discussion section has improved a lot. Thank you very much for making the effort. May I suggest a few points to consider and include in the section? Firstly, I think that effects of fugitive dust on plants are likely to be variable among different environment. I

wonder if it would be possible to include succinct discussion about the environmental characteristics of the district in focus and how these might have affected physiological traits of perennial grasses and their responses to fugitive dusts. Also, would you please have a few words about how the crop species used in this study might respond to fugitive dusts in different environments. Can people in different region expect the same crop species to respond in similar ways to fugitive dust in this study? Why or why not? I hear that, due possibly to the climate change, the dust in crop fields in the US is again becoming a serious problem in recent years. These are all big issues, but I think that it would be good to put the results of this study into context. Secondly, in the last paragraph, would it be possible to add a sentence about dusts on plants might reduce damage by herbivores (large and small)? There are papers on kaolin clay (that have been sprayed onto plants to reduce photoinhibition) reducing herbivory by insects. Also, even large herbivorous mammals do not like gritty plants. We appreciate your close attention to our work and your helpful comments and suggestions.