Referee Comments for Geoscience Communication gc-2020-42

Thank you for the opportunity to revise our manuscript. We would like to thank the referee for their time and for the constructive comments they have provided. We have revised the manuscript based on your suggestions. We reply to each of the comments below. Our suggested edits in the paper are in blue below, with line numbers indicating where we wish to make the changes.

Referee Comment	Author Response
My main concern is that Sections 2.3 and 3 go in very much detail on the statistical analysis. This part is difficult for an audience that has a not so strong background in statistics and it distracts the reader from the main topic of the paper: how can we communicate uncertainty about spatial predictions effectively? I strongly recommend to move large parts of these sections, including quite a few of the tables, to the Supplementary Information. Instead, more attention could be paid to what we learn from the experiment conducted on communicating uncertainty (i.e., Table 12, Figures 7 and 8). Further, a more thorough comparison should be made with findings on spatial uncertainty communication and visualisation from the cartography and geo-information literature (I added a few entry citations at the end of this review).	Thank you for the suggestions. To address the concerns our referee on sections 2.3 and 3 in the manuscript, we have expanded the Supplementary Materials section and added an Appendix section on the manuscript. We made the following changes. 1. On L197 we changed the text to The 'full table' illustrated in Fig. A1 is an example of this. 2. We have moved the text from L211 to L222 to the Appendix section. We also moved Fig. 2 to the Appendix and renamed it Fig. A1. 3. Fig. 3 has been moved to supplementary information of the manuscript and is now Fig. S9. 4. Tables 4 & 5 have been moved to Appendix and renamed to Tables A1 & A2. The text on L284 has been edited to The full tables for responses for responses both locations and all posters to question Q1 are shown in Table A1 in the Appendix. The responses pooled for both meeting locations are shown in Table A2.

	5. Table 12 has been moved to the Appendix and has been renamed to Table A3. The text on L319 has been edited to The responses for Q5 are shown in Table A3. We acknowledge, the importance of the topic raised by the referee of comparing findings on spatial uncertainty communication and visualisation from cartography. We have added a paragraph in the discussion from L404: The findings of this study complement work that has been done on cartography and visualization for spatial information (Kunz et al., 2011; Beven et al., 2015). Our findings show the importance of finding cartographic solutions to represent probability information, and to develop interactive methods for interpretation in a GIS environment (e.g., to produce pictographs, like those we have used, for sites of interest, or to find more effective ways to represent the 95% prediction interval).
I also think the experiment could have been conducted in a better way and that some basic mistakes were made in preparing the posters. These and some other points are worked out in the detailed comments below. I do not require that the experiments are redone but recommendations how to do better in future could be included in the Discussion and/or Conclusion	Thank for citing this, we wish to address this comment by adding a paragraph, in the discussion section, focusing on the limitations of the study. We added a paragraph in the discussion to address the concerns of the referee from L404: It is good practice to use a consistent colour scale for the three legends showing lower and upper 95% prediction interval and the conditional median. However, in our study we could not use one colour legend for the three maps for Fig. S1 (Poster 1) because of the marked differences in the predicted values on back-transformation. This made it difficult to find a working

	colour scale from the minimum value in the lower bound to the maximum in the upper bound on which one would see the variation in all three maps. We opted to use a continuous legend on the map of the mean and discrete ones for the lower and upper limits. This might have hindered interpretation. However, we suspect that there is a need for fundamentally different ways to visualize confidence intervals, perhaps using interactive methods to display them in a GIS environment.
(L36) Not in all kriging algorithms is the prediction a linear combination of the data.	The reviewer is correct here in that, in some circumstances, the kriging prediction may be a linear combination of some non-linear function of the data (see, for example, Webster and Oliver, 2007). It remains, however, a linear model in the parameters, hence the term “Best Linear Unbiased Predictor” for the prediction from a Linear Mixed Model. We edit the text at L36: The prediction is a linear combination of the data, sometimes after a non-linear transformation, which is optimal.
(L39, L41, L130, etc.) Authors use the term ‘confidence interval’, but technically this should be ‘prediction interval’. There is a principal difference between the two, for example see https://en.wikipedia.org/wiki/Confidence_and_prediction_bands.	Thank you for raising this suggestion, we wish to address this comment by replacing ‘confidence interval’ with ‘prediction interval’ at L11, L39, L41, L130, L134, L135, Table 1, L245, L 252, L296, L327, L345, L367, L371, L373, L374, L375, L377, and L408. We wish to make this change also on Table 1; Figures 4, 5, 6, 7 and 8. The change will also be applied to the Figure S1, S9 and S10 in the supplementary material.
(L46, L110-114, L138) There is no need to use indicator kriging to compute exceedance probabilities. By invoking the normal distribution assumption for kriging prediction errors (which authors do, se L38), these exceedance probabilities can be	While it certainly is possible to compute probabilities on the assumption that ordinary kriging errors are normally distributed, this does introduce an additional potential source of error. This is why methods such as indicator and

easily derived from the kriging prediction and kriging variance. They will be more accurate than those obtained using indicator kriging.	disjunctive kriging have been developed. We therefore do not accept the reviewer's view that indicator kriging would necessarily produce less accurate results than the assumption of normal errors. At line L111 we inserted the following text. While exceedance probabilities could be computed on the assumption of normally distributed errors, we chose to use the widely-applied non-parametric method, indicator kriging, which requires no such assumption.
(L61-65) I may be opening a box of Pandora, but authors will know that the uncertainty in the mapped concentrations of micronutrients in grain are heavily influenced by the support of the observations and predictions (i.e., the area or volume over which observations and predictions are made). Authors do not apply a change of support so the predictions and associated uncertainty refer to the support of the observations. Is this appropriate? What was it? This is not explained in L82-96: there is a lot of attention for the spatial sampling design but we learn nothing about how the field sampling was done. Were these point samples or bulk/composite samples? This is of key importance when addressing uncertainty.	This is a fair point. The sampling method is described in detail elsewhere (Gashu et al., 2020). We have added further information about sampling in the following text we have added below L65. The sample support for these data consisted of a bulk grain sample formed from aliquots collected from grain samples within a single field, as described by Gashu et al. (2020). The predictions, and quantifications of uncertainty, therefore, relate to grain nutrient concentrations at individual field scale. This is appropriate when considering possible health implications for smallholder and subsistence producers.
(L89) was --> were.	Suggested edit on L89 has been made to the manuscript. In total, 455 sampling points were obtained, including 136 and 113 locations where teff and wheat were sampled, respectively
(L103) This implies that predictions need to be back-transformed. How was this done (note that a naive back-	This is also an important point. The back-transformation, to be unbiased, requires a term in the kriging variance. However, this introduces a potential source of uncertainty. For this

transform returns the median, not the mean)? Information about the back-transform should be added.	reason it is commonly advocated (e.g. Pawlowsky-Glahn and Olea (2004). Geostatistical analysis of compositional data, Oxford University Press) that the simple back-transformation by exponentiation is used. This is median-unbiased (i.e. estimates the conditional median). Pawlowsky-Glahn and Olea (2004) note that this is a more useful predictor than the conditional mean for a strongly skewed variable. We propose to expand the text at L103 to explain this, and to use the term "conditional median" rather than "conditional mean". Note, however, that the prediction interval retains its usual interpretation on back-transformation.
(Eq. 1, L123) Here it should be upper case Z instead of lower case z, while in L132 and L134 it should be lower case z instead of upper case Z.	An upper-case Z is used to refer to the random variable, and a lower-case z to refer to a realization. We follow sources such as Webster and Oliver (2007). We do not think that it makes a difference whether an upper or lower-case z is used for the first term in the bracket in Equation 1. We are willing to make that change at the reviewer's suggestion. However, the cases should remain unchanged at lines L132 and L134 because there we are referring to observed kriging errors (132) and are retaining the same notation for the kriging prediction (upper case) as in Equation 1.
(L129, L340, Figure S3) Poster 3 should have shown the kriging standard deviation instead of the kriging variance. The kriging variance has different measurement units (the square of microgram per kilogram) and one cannot expect decision makers to account for this. Poster 3 also does not list the measurement units of the kriging variance. Moreover, the numbers are extremely small (around 1) and are almost certainly incorrect.	In this study we were explicitly considering the kriging variance as a measure of prediction uncertainty, just as one might use the variance as a measure of variability. In this case we cannot back-transform the variance (or by extension the standard error) to the original units of measurement, so the kriging variance is simply presented as a relative measure of uncertainty across the mapped area. This may well be one of its disadvantages. We are not sure why the reviewer thinks the kriging variances are incorrect, we did check them by cross-

	validation. Perhaps they did not realize that these are on the log scale. At L129 we add the following text. The kriging variance is on the transformed (log) scale, as a back-transformation of this quantity is not possible. The variations in kriging variance therefore give the interpreter an impression of the variations in prediction uncertainty across the mapped area, but not in interpretable units. We added a comment about this in the discussion section on L368. The difficulty of interpreting the kriging variance is compounded when a transformation is necessary, and that, in other circumstances, the kriging standard error, on the original units of measurement, may be more interpretable.
(Section 2.1.4, Figures S2 and S4) I doubt that computing the probability that the true value exceeds or lies below a threshold quantifies the uncertainty of predictions. For example, if the threshold is 38, the kriging prediction is 55 and the kriging standard deviation 8 then the probability of exceeding the threshold is extremely large (suggesting very small uncertainty, category “virtually certain”), while a kriging prediction of 36 with standard deviation 3 leads to large uncertainty (we end up in the category “about as likely as not”). But 8 is larger than 3, so can we maintain that the uncertainty of the predictions is quantified? These complications should have been addressed.	The reviewer makes an important point, but we do not agree that probabilities are not communicating uncertainty in these circumstances. If the prediction distribution has a large variance, but the mean is well above the threshold, then, from the perspective of a data user making a decision about nutritional interventions, the uncertainty about the contribution from staple crops is indeed small, and smaller than for a second case where the prediction variance is smaller, but the mean is near or on the threshold.
(L174) were --> where; where --> were.	Suggested edit on L174 has been made on the manuscript.

	Evaluation of communication methods were done through a questionnaire, as shown in Table 3, but without putting the participants in a situation where they felt they were being tested on their mathematical skills and understanding.
(L186) Visiting posters in randomised order does not avoid carry-over effects, it only makes sure that the effects cancel out over a larger group. Perhaps rephrase this sentence to make this clear. Note also that instead of randomising it would have been better to have a deterministically determined sequence that guarantees that all posters occur in a completely balanced order.	In view of this, we rephrased L186 to say To avoid any bias resulting from carry-over effects from one poster to another when the individual responses were pooled for analysis. Regarding the second point, this would still be done by randomization (e.g., a behavioural Latin square), but was not done for logistical reasons (i.e., to reduce the overall numbers of groups of participants that had to be managed in the exercise).
(L207) Symbol $o_{i,j}$ not defined in the main text.	The symbol $o_{i,j}$ was defined on L205 in the following way The evidence for the saturated model, as a better model for the data than the additive model, is provided by the likelihood ratio statistic or deviance for the two models, L, where $$L = \sum_{i=1} \sum_{j=1} o_{i,j} \log \frac{o_{i,j}}{e_{i,j}}$$ and $o_{i,j}$ are the number of observed response in cell [*I,j*].
(L225) Two times “between the”.	Suggested edit on L225 was made on the manuscript.

	However, it was first necessary to consider whether there was evidence for differences in the responses between the two sets of respondents at different locations.
(L229, L230, L235, L296, L301) “was conducted”, “participants are drawn”, “This gives us”, “there was”, “There is”. Please check entire manuscript on correct use of present and past tense.	Thank you for this suggestion. We have checked the entire manuscript to correct on the use of present and past tense.
(L277) less --> fewer.	Suggested edit on L277 was made on the manuscript. In the Ethiopia meeting, we had fewer participants (64%) who had studied mathematics and statistics up to degree level and above, than in the Malawi meeting (88%), see Fig. 3.
(L324) a there is --> there is a.	Suggested edit on L277 was made on the manuscript. Fig.6 shows the responses to Q5 for the separate posters for pooled counts graphically. We can see that there is a greater proportion of respondents selecting the response `Message clear' on threshold-based methods, Posters 2 (IPCC verbal scale), 4a (raw probability) and 4b (raw probability plus pictograph), than on general based.
(L334-335) Can and do you explain why the p-values were so different between Ethiopia and Malawi?	The difference could be as result of differences in compositions of the groups in Ethiopia and Malawi. We added this text to explain the difference in the manuscript on L335. The difference maybe because the stakeholder in the Malawi meeting was more homogeneous in terms of professional group (a less even distribution among them) and level of

	mathematical education than the stakeholders in the Ethiopia meeting.
(Figure S1) Poster 1 has some important deficiencies. First, the mean has a continuous legend while the lower and upper limits have discrete units. This affects the map (discrete colour jumps in the limit maps). Second, all three maps should have had the same colour legend. For an example, see Figure 7 in https://onlinelibrary.wiley.com/doi/full/10.1111/ejss.12998.	The reviewer makes an important point, and we must acknowledge it was difficult to find a working colour scale in which one could see the variation in all three maps, given the marked difference in the ranges. Hence, we decided to use different colours and discrete units. However, as guided by our referee, we have added paragraph explaining the limitations of the study which we wish to add from L404.