Response to Referee Comments for gc-2020-42

We would like to thank the referees for the opportunity to revise our manuscript. We have revised the manuscript based these suggestions and the changes are shown in the tracked changed version of the manuscript. The line numbers which we refer to are the ones in the tracked change manuscript.

Referee 1

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 on the manuscript. 1. We have moved the text from L242 to L253 to the Appendix section (L510 to L524). 2. 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. 5. The text on L316 to L317 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.

	6. Table 12 has been moved to the Appendix and has been renamed to Table A3. 7. The text on L370 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 L469 to L472: 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 L472 to L479. 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 referee 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 edited the text at L37: 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.	We agree with the referee and we have replaced 'confidence interval' with 'prediction interval' at L11, L46, L153, L158, L159, L279, L283, L378, L405, L428, L434, L436, L438, L439, L441, L489, L491, L508 and L509. We also have made this change on Table 1; Figures 2, 3, 4, 5 and 6 in the manuscript. The change will also be applied to the Figure S1, S10 and S11 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 easily derived from the kriging prediction and kriging variance. They will be more accurate than those obtained using indicator kriging.	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 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 L129 to L130 we have 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 from L95 to L98: 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.	The suggested edit has been made on L89:

	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-transform returns the median, not the mean)? Information about the back-transform should be added.	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 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 have added a paragraph from L118 to L123 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 L139 and L140 because there we are referring to observed kriging errors (L139) 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	In this study we were explicitly considering the kriging variance as a measure of prediction uncertainty, just as one might use

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.	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. We have expanded the text from L140 to 144 and from L149 to L151 to explain this: 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 also have added a comment about this in the discussion section on L429 to L431. 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 from L200 to L202 has been made on the manuscript. Evaluation of communication methods was 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 the sentence from L213 to L214 to: Participants visited each poster in a randomised order 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 L235 to L238 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 responses in cell [*I,j*].
(L225) Two times “between the”.	Suggested edit on L255 has been made 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 L309 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 S9 .

(L324) a there is --> there is a.	Suggested edit on L375 to L377 has been made on the manuscript. Fig.4 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 L386 to L388. The difference maybe because the set of stakeholders 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 from L473 to L480.

Referee 2

Referee Comment	Author Response
The paper includes a lot of statistical terminology and detail of methods. I assume intended audience is those with knowledge of statistical terminology and methods. Possible lost opportunity to appeal to a wider audience given that emphasis on communicating uncertainties.	Thank you for the suggestions, which parallel the first comment from Referee 1. Please see our responses there. In summary, we have removed some of the statistical detail to an Appendix, including text and figures, so that the key arguments should be clearer to a general reader.
Table 1. Would like to see the poster designs. This would add context to the subsequent discussion	In the manuscript we mentioned on L81to L82 that the posters are presented in the supplementary materials. In order to make this clear for the reader we have added the figure number on the following lines in the manuscript: L151to L152 – "To investigate the utility of the kriging variance as a method to communicate uncertainty, one poster showed a map of conditional mean of Se concentration in grain (Section 2.1.1), with a map of kriging variance (see Table 1, Fig S1)" L157 to L159- "One poster showed a map of conditional mean of 135 Se concentration in grain plus the lower and upper bounds of the 95% confidence intervals mapped separately to communicate the uncertainty (see Table 1, Fig S3)." L183 to L185- "Therefore, we presented three posters, each showing a map of conditional mean of Se concentration in grain (Section 2.1.1.), plus probability presented as (1) raw probability scale (see Fig S4), (2) IPCC verbal scale (see Fig

	S2) and (3) raw probability scale plus pictographs (see Fig S5), communicating the uncertainty (see Table 1). .
Might the questions in Table 3 encourage participants to say 'Message clear' to show they understand what they are being shown? Does this introduce bias in the way the question is worded? If author agrees, there is an opportunity here to acknowledge this or show has this has been accounted for in subsequent questions.	We do not think that such a bias was likely in the context of the workshops. All responses were anonymous, and this was made very clear to participants at the start of the meeting. Furthermore (i) in the workshop we emphasized the point that the questions were not tests of the participants' understanding but rather of the efficacy of the methods for communication. (ii) It is clear in the questionnaire (and again, was emphasized in the workshop) that the participant was not being asked to interpret the representations. Rather, the interpretation was stated (e.g., "Our confidence that grain Se concentration exceeds 38 μg kg−1 is greater at x than at z") and the participant was then asked whether this was made clear by the representation. (iii) the fact that the participant was being asked to answer the same question about different methods to convey the same information emphasizes that their responses may differ between methods, even though the fixed interpretation is clear in their minds. This appears to have happened. We noted at L397 that in Malawi a large proportion of respondents selected "Not clear" as a response for the poster which used confidence intervals. In response to comments raised by Referee 1, we added a paragraph the end of the discussion with the reflection on possible limitations of the study. To expand the discussion on the limitations of the study, we have added the following paragraph from L481 to L486:

	"We accept that a possible source of bias in any such study is that a participant feels that they are being tested on their interpretative skills, and so might select a response which suggests, in a general sense, that they understand the input (e.g. "Message clear" for the case in Table 3). However, all participants were aware that their responses were strictly anonymous, and it was emphasized that the task involved their evaluation of several methods for the communication of an interpretation which was provided. In future studies it might be useful to include some final questions which actually are "tests of interpretation" secondary to the main task, to see whether this affects the responses given for different methods."
Figure 2 – Perhaps add a key to explain what the O indicates. This isn't that clear to a non-specialist	We added the key to Fig A1 (renamed from Fig 2) as suggested by the referee.
L21 – Perhaps worth alluding to the ethical issues surrounding the ethics of interventions to improve the dietary intake of Se. Whilst this is not the subject of the paper, worth noting perhaps.	This is an interesting suggestion. We do not think that the general ethics of food-based interventions is within the scope of this study. However, we added the following comment in the Conclusions from L496 to L503: "Because decisions on interventions to address nutrient deficiencies may have positive and negative effects on peoples' health and well-being, the interpretation of information such as that we have used is not value-neutral, and uncertainty in information has ethical implications (given that all spatial information is uncertain, how much uncertainty is ethically acceptable in the decision process?). While these considerations are outside the scope of the study reported here, it would be interesting in future research to examine

	how individual attitudes to the ethics of fortification interventions affect their responses, and whether individuals' perspectives on the ethical implications of basing decisions on uncertain information differs between different methods to communicate that uncertainty.''
L32 – Nugget variance – assumption that readers will know what this is. Author could include glossary/footnote	We have expanded the text from L32 to L34 to explain the nugget variance. "Predictions are subject to uncertainty because of spatial variability resulting from multiple factors operating at different scales (Lark et al., 2014). In addition to environmental factors (geology, climate), there is also uncertainty due to measurement error in the analysis of material, and sampling error in the field where a single crop or soil sample is collected."
L225 – Good to see acknowledgement off possible differences between different groups. Suggest further group work with other participants may increase validity of study. Could this be a suggestion for future work?	The reviewer makes an important point, and we made the following edit to the text on L453 to L456 to emphasize this point. Further work to address this question and examine how stakeholders interpreted each poster will require an elicitation with sufficient numbers of participants with different mathematical background.
L225-232 Good recognition of potential for bias	Thank you for the acknowledgement.
L232 Different learning styles may also affect how people interpret posters	We agree and therefore we expected this to affect their responses. However, due to unbalanced numbers of participants when we categorised them by level of

	mathematical education, it was not possible to do further analysis. We acknowledge this and we highlighted this as a future work from L453 to L456: Further work to address this question and examine how stakeholders interpreted each poster will require an elicitation with sufficient numbers of participants with different mathematical background. This would be useful to understand better how different learning styles influence the interpretation of uncertain information.
L350 – Conclusion about users finding information presented accessible and clear – responses could have been affected by the desire to show understand the representation. I think the leading nature of the question could be seen as significant. Suggest consider acknowledging this possibility	Please see our response to the third point above. We do not agree that the participants were asked a leading question. They were asked to select among responses to a question about whether it was clear from the poster that a certain statement was true, and possible responses included “Not clear” and “More information needed” as well as “Message clear”.
L360-362 – Agree with statement that further work is needed	Thank you for the acknowledgement.
L419-420 – Would like to see how measures of uncertainty are presented – and how These less effective methods of communication (kriging variance and confidence intervals) could be presented in a more effective way	Thank you for acknowledging this point and we strongly believe this is a scope for future research work on methods of communicating uncertainties in spatial predictions.