Response to Reviewer 2's Comments

I would like to thank the reviewer for their comprehensive review of the paper. They have highlighted some interesting issues which I have attempted to answer both within the paper and in the response below. The reviewer's comments are shown in italics and my response to their comments is shown in plain text.

General comments

Line 120, 178 and 420-422. The discussion of configuring ensembles to perform better for certain variables and certain parts of the atmosphere is interesting and would benefit from a lengthier description. In this study, simulations are performed using the MOGREPS-G meteorological ensemble. Has this ensemble been optimised to produce a maximum growth rate of the ensemble spread at a certain forecast lead time? Would differently configured ensembles be more suitable for dispersion applications?

The configuration of meteorological ensembles and their suitability for dispersion ensembles is a very interesting topic. In the past meteorological ensembles were optimised to produce a maximum growth rate of the ensemble error at a certain forecast lead times but recent work in this field has focussed on ensuring that the ensemble is optimised for all forecast lead times. As far as the authors are aware dispersion studies using ensemble meteorology have focussed on single case studies and single ensemble meteorological data sets (or multi-model ensembles) so have not considered whether differently considered ensembles would be more suitable for dispersion applications. We have added a sentence noting this below line 178.

Line 331. The authors correctly state that the BSS provides a comparison of the performance of the ensemble relative to the deterministic forecast and does not provide information about the individual performance of the ensemble. Therefore, if the deterministic forecast is accurate the BSS can be negative even if the ensemble forecasts are also representative of the analysis. I would like to see this argument in the introduction section if possible as it's an important point for interpreting these relative skill scores. This is particularly exemplified in figures 12 and 13. By eye the ensemble forecast appears to perform in a very similar manner to the deterministic forecast, but the BSS shows that relatively, this ensemble is worse.

We have expanded the text mentioning this point in the location where the Brier skill score is first mentioned towards the end of section 2.0.3.

Line 204, 282, 291 and elsewhere. The Brier Score is calculated for a single output grid square. Does the size of the grid matter? For example, the authors state that the ensemble runs perform better than the deterministic runs at later time steps and hypothesise that this is due to increased ensemble spread at later times. Another reason could be that the plume has spread out more at later times reducing the potential for a double penalty issue. This issue also highlighted in figures 5 and 6, do the negative BSS occur when the plume is narrow, i.e. at the start of the simulations? When calculating BSS at the grid scale small displacements in the plume location can result in large differences compared to the analysis. This occurs particularly when the size of the eddies causing dispersion are large compared to the width of the plume. Would it be possible to show the BSS vs area covered by plume, in an analogous way to fig 7.

Investigating the impact of the grid size was out of the scope of this project. However, I have plotted the Brier skill score against area of the plume (below). This shows that the spread of Brier skill scores

is greater when the area exceeding the threshold is smaller but there is no bias towards negative or positive skill scores for large or small areas.

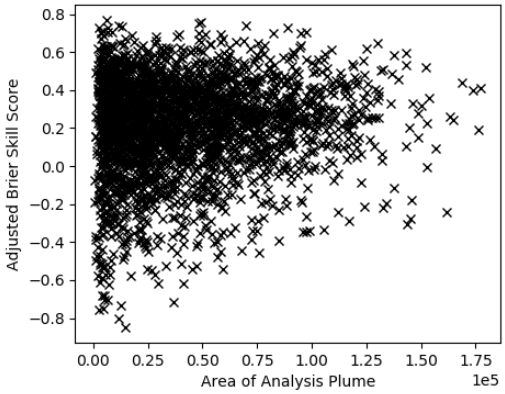

Figure 1: Brier skill score plotted against area of forecast exceeding the threshold for total integrated air activity of Cs-137 above 50kBqs/m3. The area exceeding the threshold is determined from the analysis plume.

In the text, to help clarify this point we have replaced:

“There are fewer negative scoring runs at later forecast time steps implying that the ensemble is more likely to perform better than the deterministic at later time steps. This is possibly due to the increase in ensemble spread at later time steps.”

with “At later forecast time steps there are fewer negative scoring runs and the range of Brier skill scores is narrower. The reduction in negative scoring runs implies that the ensemble is more likely to perform better than the deterministic at later time steps. This is possibly due to the increase in ensemble spread at later time steps. The reduction in the range of the Brier skill scores is likely to be due to the increase in area exceeding the threshold. At early time steps when the plume is narrow the Brier skill score is dominated by a few grid cells and the ensemble tends to be less spread resulting in either a high Brier skill score or a very low Brier skill score. At later time steps the plume is more spread out and the ensemble is more spread so there is a greater range of Brier scores for the different grid cells and the Brier skill score tends to be closer to zero.”

Finally on line 282; do the authors know why there is a difference in the rate at which the Brier skill score increases with forecast time for different flight levels? One explanation could be that the plume spreads more rapidly at the lower levels due to increased turbulence and remains tightly constrained at upper levels?

This is an interesting question but unfortunately one which the authors were unable to investigate within the project.

Although not the aim of this paper, it would be of value in the conclusions to discuss how forecasters/decision makers might make use of ensemble dispersion forecast output.

The aim of this paper was to examine the value of using meteorological ensembles to provide meteorological uncertainty information to dispersion modelling. An equally important component of the forecasting process is how forecasters and/or decision makers make use of ensemble dispersion forecast information. The authors believe that this part of the process is worthy of a paper (or many

papers) and cannot be covered in a short statement. However, a sentence acknowledging this has been added to the conclusions.

Specific comments

Why is the title posed as a question? The answer is clearly yes, but ensembles add value is what is being addressed here.

Title modified to: Assessing the value meteorological ensembles add to dispersion modelling using hypothetical releases

Line 55. The authors refer to the computational expense of running a statistical emulator. In my experience statistical emulators are built precisely because they can mimic the response of a dynamical model but much faster because they only rely on statistical relationships. Perhaps I have misunderstood the meaning of this sentence?

The reviewer is correct, running a statistical emulator is not computationally expensive. However, a new statistical emulator needs to be constructed for each different dispersion event and therefore the construction of emulators for multiple events is expensive. In addition, to construct emulators for multiple events would require a significant amount of effort. We have corrected this sentence to highlight where the computational expense is.

Line 75. Not all ensemble systems perturb both the initial model state and the model physics. Therefore 'and' should be 'and/or' in this sentence.

Sentence modified as suggested

Line 96. In this section there is reference to the Brier skill score and use of lagged ensembles. These terms should be explained. For example, does the 'most recent ensemble' refer to the lagged ensemble with the shortest lead time?

I've replaced "They used the Brier skill score to show that a 24-member ensemble performed better than the regional model for both eruptions. However, although the lagged ensemble outperformed the most recent ensemble for Kelut the most recent ensemble performed better for Rinjani."

with "Performance was assessed using the Brier skill score, a skill score that measures the accuracy of probabilistic predictions, to show that a 24-member ensemble performed better than the regional model for both eruptions. \citet{dare:2016} and \citet{zidikheri:2018} also compared the performance of a forecast generated using meteorological data initialized 24 hours earlier than the latest forecast at the start of the eruption and showed that although the ensemble using the older forecast outperformed the forecast using the most recent ensemble for Kelut the forecast using the most recent ensemble performed better for Rinjani."

Line 99. The term 'dispersion ensembles' is somewhat ambiguous. All Lagrangian model dispersion simulations are ensembles in the sense that they release an ensemble of particles and track their motion. I guess the authors are referring to dispersion simulations run using ensemble of meteorological fields. This is a bit wordy but should be explained in full the first time to avoid ambiguity.

Replaced: "These studies suggest that for those events that have been examined dispersion ensembles outperform dispersion models run using a single meteorological model."

with: “These studies suggest that for those events that have been examined dispersion models run using ensemble meteorology (hereafter dispersion ensembles) outperform dispersion models run using a single meteorological model.”

Line 151. Here the authors use a 20kmx20km horizontal grid spacing, but earlier (line 136) they use a 10km x 10km grid spacing. Why as a different grid spacing used for the two scenarios?

The scenarios were set up based on typical grid spacings used within services delivered by the Met Office for volcanic ash forecasting and radiological dispersion forecasting. An increase in resolution was applied to both scenarios to reflect likely future increases in resolution of both services.

Line 213. Do the authors have a reason or hypothesis for why the highest threshold is exceeded around 100km from the release location for all the release locations?

This work was carried out following the Horizon2020 CONFIDENCE project (https://portal.iket.kit.edu/CONFIDENCE/index.php?action=confidence&title=objectives) so the setup of the radiological scenario reflects choices made within that project. So, the threshold exceedance at around 100km was chosen based on the distance thresholds were exceeded in the modelling carried out in that project. However, re-reading the CONFIDENCE reports we realised that the distances were based on typical distances at which deposition thresholds were exceeded rather than air concentration thresholds. The text has been modified to reflect that and to link to the final CONFIDENCE paper.

Line 239. The analysis and deterministic met have the same grid spacing while the ensemble met has coarser grid spacing. Does this impact the results? If so, why not coarse grain the analysis and deterministic met to the same grid spacing as the ensemble met?

It is possible that the different resolutions of the ensemble meteorology and the deterministic and analysis meteorology has an impact on the results, but we don’t believe it is a dominant impact. Most meteorological centres sacrifice resolution for ensemble size, so most ensemble meteorological data are at a lower resolution than their deterministic counterparts. In the event of a real atmospheric dispersion incident, we would use the ensemble and/or deterministic meteorology at their native resolutions and therefore we wished to assess their performance at their native resolutions in this study. We have added a note explaining this below line 239 (original submission).

Lines 358-400. These two paragraphs are a repetition of the methodology and are not conclusions. Therefore, it is not appropriate for them to be in the conclusions section.

This section has been re-titled as summary and conclusions.

Table 1. What time period are the accumulations over?

The volcanic ash is accumulated from the start of the eruption up to the forecast time so for the 3-hour forecast time it is the total accumulated over 3 hours, for the 12-hour forecast time it is the total accumulated over 12 hours.

Figure 2, 3 and 4. Are the averages over all releases?

Yes

Figure 3. Why is there a larger spread in the average maximum distances for the different concentration thresholds in the lowest layer (FL000-200) compared to the higher layers? Is this due to deposition of particles to the surface? If the particles did not deposit to the surface does this difference in spread decrease?

Investigating the spread in average maximum distances for the different concentration thresholds was out of the scope of this project.

Typographical errors

Both errors below have been corrected.

Line 146. There is an extra space before 800m.

Line 308. ‘at and’ should be ‘and’