Response to referee comments

The authors thank our reviewers for their thoughtful and constructive comments, and we will aim to address the points that have been raised in our revised manuscript where it is practicable to do so. Our detailed response to each review is given below.

Response to referee #1

Thank you Slawomir for your helpful review comments. We have revised our manuscript to take account of points that have been raised.

Addressing the general comments section, we have included several further general references on dispersion ensembles in an effort to put our work into a broader context. As the present paper is essentially a follow on to (Leadbetter et al., 2022) we did not wish to repeat too much background information in the introductory sections, but we agree that a little more context would help the reader here. We have also added a further sentence to emphasise the value of this extensive dataset for exploring various aspects of the ensemble behaviour (as suggested, e.g., to be able to compare results for different locations or meteorological conditions). However, we have caveated our statement as we do not have a sufficiently large modelling period or domain coverage to allow comprehensive analysis to be performed (ideally, one would wish to run such simulations sampling a period of several years and with global coverage!).

Specific comments

1. One of the basic questions related to the presented methodology is whether 18 members is enough to produce sufficient statistics to cover interested range of possible results. It seems that there are situations when this is not the case, and the authors are aware that either more ensemble members would be needed or other models can be applied. ECMWF produces large forecast ensembling that can be used to drive atmospheric dispersion calculations, however it'd be very time consuming. The other possibility is to produce multi-model ensemble, which usually has bigger spread than the ensemble based on one dispersion model. In fact there are many articles already published dealing with these issues.

Yes, a single cycle of the MOGREPS-G forecast is limited to an ensemble size of 18 members. The Met Office does use a time-lagged approach (2 x 18 members) for other forecasting applications, which we plan to explore for dispersion ensembles too in the future – but this is beyond the scope of the current paper. We agree that ECMWF would also be an option – providing more members plus the benefit of an independent data assimilation system and NWP model. The multi-model approach is another possibility, of course. However, many of the multi-model ensemble papers have looked at ensembles where both the dispersion model and the meteorological model vary and our aim in this paper is to focus just on meteorological uncertainty and ensembles within the framework of a single dispersion model. Lots of work remains to be done around aspects such as optimal ensemble size, source term ensembles, multi-model or multi-parameter dispersion ensembles!

*2. Table 1 contains thresholds used for both scenarios. Obviously, in case of operational system, the best would be, when these thresholds reflect some criteria used operationally. For radiological scenario mostly doses are applied in various criteria, however in some countries, like Austria also time integrated concentration and deposition are used. For example some agriculture countermeasures can be implemented, if time integrated concentration of Cs-137 exceeds 350 Bq*s/m3 or depostion is higher than 650 Bq/m2 (for iodine I-131 this is respectively 170 Bq*s/m3 and 700 Bq/m2). Thresholds shown in Table 1 are much higher, but this is obviously arbitrary choice of the modellers.*

The selection of thresholds for the radiological case had been kept simple, with thresholds chosen to capture typical downwind distances for monitoring and detection in the event of a moderate radiological accident (i.e., up to several hundreds of kms) rather than using actual thresholds for, e.g., evacuation and sheltering decisions. Our source term has similarly been chosen in a largely arbitrary manner. While our choices are therefore somewhat arbitrary, they are pragmatic and useful. The volcanic ash case uses real thresholds as these are more clearly and uniquely defined in an international context. Our paper is not aiming to assess specific features like sheltering or food bans (with their specific requirements on quantities, averaging periods, thresholds, etc.) but to instead assess the dispersion behaviour in a general manner.

3. The authors use quite simple indicators (rank histogram, attribute diagram, spread-error relation), but it seems they are mostly sufficient. On the other hand it would be convenient to see the values in the form of table (ensemble spread vs error in ensemble mean) to see how the results are changing in time. Some additional indicators can be also considered: like factor of 2 for spread-error diagram.

We have used metrics and diagrams that are commonly used when evaluating ensemble predictions and these should provide a good general overview of ensemble performance especially with respect to the spread and calibration characteristics. We feel that presenting a table of results in this instance could be over-simplistic and might be misleading on its own, whereas the graphics in our figures help to illustrate the level of variation that is seen with dispersion ensembles. We have added the FA2 (factor-2) shaded region on the (binned) spread-error plots to aid interpretation, though we have not evaluated the FA2 metric itself on the raw data points for the reasons outlined above.

4. The way of rank maps presentation with two colour sections is appreciated. However, the reader should be warned against too simple interpretation of these maps. The fact that the ensemble system predicts areas where "real plume" (i.e. from analysis) are not present does not mean that the ensemble gave bad prognosis. If the ensemble shows low probability for such areas it is fine, otherwise you can say that prognosis was not very accurate. The role of ensemble is to predict areas when plume can, but not necessarily, must appear.

We have revised the sentence starting on L385 to take account of the above. The previous text did cover some of these points (e.g., in the discussion around the ‘dark’ and ‘light’ green regions) but we agree that the discussion could be a little clearer here. Also, even in an instance when the forecast probabilities are high and the ‘real plume’ is absent, the ensemble forecast is not necessarily ‘wrong’ (as probabilistic forecasts should be validated over many forecast instances). As an aside, there is a broader challenge when evaluating metrics that use ‘correct rejections’ of how we define the area of plausibility for the comparison between forecasts and observed plumes (that is, how we differentiate between locations where a plume could plausibly reach within the time period of interest and the remote areas where it is impossible for the plume to reach and there are trivial null forecasts). This question of how to handle the ‘zeroes’ problem is mentioned briefly elsewhere in the manuscript.

Technical corrections

The main comment is related to the request of including mathematical formulas for quantities used in the article, firstly, in order to avoid any ambiguity, and secondly simply for the reader's convenience. This concerns also the way how the figures have been constructed.

Formulae have been introduced to define spread and error quantities more clearly. We feel the existing descriptions on the construction of diagrams (in words) are sufficient and that introducing additional notation might not be helpful, but we have included a further reference to Wilks here.

Response to referee #2

We thank the reviewer for their positive comments on our work.

The reviewer has proposed various ideas for extending our analysis to enhance the content of the paper, and while we agree that these are all very good suggestions for future work, they would involve significant additional effort at this time and we would not wish to delay publication of the present manuscript. Addressing the suggestions individually:

a) *verification against observed ash for a real eruption event*
We recognise that the current study is limited to hypothetical events and does not consider any observations of real-world events, which as the reviewer has noted tend to be very limited in practice. However, we would regard objective verification against observations as being a separate study and outside the scope of the current paper. We are pursuing separate research examining the behaviour of our ensemble forecasting system against recent eruption events although the work is not yet mature enough for publication.

b) *comparison of RMSE for ensemble-mean and control, and examination of RPS/CRPS scores*
We agree that these are both very good ideas to examine ensemble performance in greater detail. However, our existing analysis toolkit does not currently provide all of these outputs and would need some further development. We would then need to re-run processing on the entire dataset, which requires significant computational effort and time. It is therefore not straightforward to address this point in the context of the current manuscript. However, our earlier paper (Leadbetter et al., 2022) based on this ensemble dataset has examined the relative benefits of the ensemble approach over a deterministic one through use of Brier score metrics which partly addresses the wider point being raised here.

Specific comments

Our revised manuscript has addressed all the specific comments raised in the review. We would like to thank the reviewer for highlighting these points and for their helpful suggestions to improve the manuscript.

Lines 6/61: we have clarified the wording around “analysed” meteorological fields as per the reviewer’s suggested text.

Figure 5: the graphic has been updated to resolve the earlier issue of the overlapping bin size information. This has been achieved by adjusting the bin widths to make them slightly broader (and they are now more consistent with the other spread-error figures in the manuscript). The use of wider bins gives slightly smoother lines but does not change results in any material way. The text colour for the bin size values has also been changed from green to grey, and the colour for the T+48 data points changed from grey to olive green to improve visibility of the 1-1 line. A ‘within-factor-of two’ shaded region has also been included to aid interpretation. Additional text added to caption to address the other points raised.

Note that Figures 8 and 11 have also been updated with minor formatting changes for consistency with the revised Figure 5. We have also increased the line weighting in Figure 11 b) to improve their visibility.

Line 688: wording corrected.