We thank the reviewers for the detailed comments on the manuscript. Please find below our response (in bold and italic) to all the remarks requiring revisions or elucidations, and how and where the manuscript has been modified according to the comments.

Comment on acp-2021-858

Anonymous Referee #1

Referee comment on "Refining an ensemble of volcanic ash forecasts using satellite

retrievals: Raikoke 2019" by Antonio Capponi et al., Atmos. Chem. Phys. Discuss.,

https://doi.org/10.5194/acp-2021-858-RC1, 2021

This paper presents a filtering method for volcanic ash simulations and it is applied in a case of study for the Raikoke 2019 eruption. They use a a large ensemble of simulations and column integrated ash load retrievals from a geostationary satellite for filtering the ensemble. They show the methodology and their results include improvement in the forecasts of ash, also showing how the filtered forecast could affect the planning a set of flight routes over the Pacific Ocean, compared with the raw ensemble forecast. The manuscript explain in details the context and the method (including most of their assumptions and limitations). It show the results in a clear and concise way and it presents the conclusions accordingly. The article is well structured, well written, and its length is according to the content. The figures are clear and provide interesting information. Therefore, I recommend for publication in ACP after minor corrections.

I would like to ask to the authors some details on the averaging and collocation procedures of the Himawari mean ash column loadings retrievals and their uncertainties ; and, if possible, clarification on the correlated LHS sampling algorithm, and a few clarifications on the general overview of the algorithm (see the minor comments below). I am also wondering which of the decisions were made for avoiding the filter divergence, giving the large number of perturbed parameters of the ensemble and their implications of this known issue in particle filters, which is particularly relevant for high-dimensional problems.

The satellite hourly averages are an average of data available for XX00 and YY30 (where XX is the hour of the data we are looking at and YY = XX-1) files. The averaging process from those two times is the only one done before comparing observations to the model data. We tested to see how much it would change by comparing the instantaneous loading values output every 6h by NAME with observations averaged over a period 3 to 6 hours centered on the verification time. However, not much changed, and on a few occasions the averaging reduced the number of grid boxes available for comparison. Therefore, not averaging seemed the best way to avoid discrepancies between hourly averages and instantaneous values, despite the fact that it meant not making use of 5/6 of the observations. We have added most of these details in the revised manuscript to make it clearer and to specify that we do not time-average any further the observations before comparison in LL209-215.

We added more information about the regridding in the manuscript, LL346-350, and also a reference to the algorithm used (Iris). Both satellite retrievals and their uncertainties are regridded before comparison with the model output. For the regridding, we used Iris in order to make the comparison possible. In the final algorithm, we use the NAME dataset as target grid, and regrid the observations and uncertainty information based on that one. At the beginning of the development, we did a comprehensive testing to see how much using the different schemes available for regridding would affect the results. We also did the same tests by regridding each ensemble based on the satellite grid. Differences were minimal except that regridding the entire ensemble takes a long time. Therefore, we decided to use NAME as target and area-weighted regridding as the most appropriate scheme.

Regarding the resampling strategies, we added a description of each resampling step in the text, L440-455, and a figure (the new figure 5).

Regarding the filter divergence, for this application we decided to base our method purely on limits of acceptability. Simulations outside our limits are rejected, and the ones within are used for

resampling. The system is general, so that can be easily applied to all parameters and type of observations and as also pointed out in a later comment, we implemented dynamic thresholds that are automatically adjusted during the verification to avoid the potential that all simulations are rejected or too few members are retained for moving to the next step. Despite its simplicity, we believe that for this first application it is a suitable method, which allows rejecting models failing to simulate the observations. We added a note on this in lines 389-391 to make it clearer. One of the next steps should be defining the posterior by weighting each simulation output according to some measure of its fit to the observations, in a way that takes proper account of the epistemic uncertainties in the satellite retrievals or any other available information, and added this possible next step in LL. 596-598)

Giving the length of this review, I would be happy if the authors can address the general comments above and the related minor comments below, but I understand that addressing all of them could take more time than expected for a minor review.

Minor comment

L14: The filtering technique was developed in Beven and Binley (1992), not here. The rejection metrics, design of the ensemble, use of the retrievals, etc. are unique for this work (not the technique itself). I would propose to change the word "technique" by "system" or similar, to reflect the large amount of work done in this paper that it not related to the method itself.

We changed "technique" with "method" (L14)

L18:L20 I suggest (but not strictly needed) to rephrase this sentence: "The ensemble members are filtered, based on their level of agreement with the ash column loading and their uncertainty of the Himawari satellite retrievals, to produce a constrained posterior ensemble" .

Following the suggestion, we rephrased the sentence (L18-20)

L31: "at a given time" is not needed

We removed "at a given time"

L32: which dynamics? Atmospheric? Ash? Wouldn't be clearer : "...VATDMs solves numerical representations of equations related to ash processes in the atmosphere to evolve…" ?

As suggested, we rephrased the sentence (LL32-33)

L35: Distributions over which dimension? Are you meaning spatial distributions (as inL40)? Size distributions? Temporal distributions?

We removed 'distribution' and specified what kind of observations are usually obtained from remote sensing instruments (concentrations, size distributions, mass loadings) (L36).

L39: I suggest to remove "forward". It could be the inverse model, depending on the retrieval system.

We removed "Forward" and changed to "inverse" (L40). The word forward model is generally a more accurate term for the modelling done in the retrieval. I.e the retrieval works by forward modelling the radiances you would expect given an ash plume of given properties, this is then compared to the observed radiances to update the ash plume properties with the intent of better matching the forward modelled radiances to the observed radiances. This is achieved by minimising the cost function. 'Inverse' may be misleading to how the retrieval works, but is more appropriate in the context of the sentence, within the introduction and illustrating retrievals in general.

L40: I suggest to avoid the word "distribution" here. "volcanic ash column loading" (instead of volcanic ash distributions (known as column loadings)) is clear enough.

We changed to "volcanic ash column loadings" (L41).

L41: Would you add retrievals errors too?

We are not sure what the reviewer refers to when saying 'retrieval errors'. If it refers to errors that come from the minimisation of the cost function, then these errors can be significant, but so can the others already mentioned in the manuscript. We reworded the sentence to make it clearer (L42).

L42: wouldn't be clearer: "This error information … "?

We added "error" in the sentence (L43)

L50: Indeed, this is the only the "Jo" or the observational part of the cost function (yH(x)), where y are the satellite retrievals and H(x) the simulations of the retrievals by the VATDM. The VATDM errors shall be included in this observational error covariance matrix (usually denoted by R or Po). The other term in the simplest form of the variational formulation refers to the prior errors (x-xb)^T B^-1 (x-xb), which is the difference of the control with the prior.

We agree with the above statement. We do not think the text in the introduction needs changing following this comment.

L53: I see a small change of the colour of the text after "plume" (but it could be only an artefact of my pdf viewer).

We did not have the same issue, but we re-checked the font colour throughout the manuscript.

L50, L57-L60 Please reconsider to use the word "sequential". Variational data assimilation can also provide estimates of the system states sequentially (for example the 3DVar or 3DVar-FGAT), and Kalman filtering ("sequential" in the manuscript) can also use a fixed time window (but usually it is shorter than the variational method), for example in the NOAA's GFS.

We agree with the comment, however we do not think the terminology is inappropriate here. In the introduction the terms are useful for providing a general overview of the two main approaches for data assimilation (variational and sequential). We have added the word 'typically uses observations from a fixed time window' to allow for the exceptions to this rule (L58).

L64: Shouldn't be better to write "... this estimates a probability density …", or "empirical pdf' ? Can the the full (usually continuous) pdf be computed by perturbing the VATDM parameters and meteorology?

As suggested, we changed the sentence (L65).

L65:66 Just a suggestion: it could be clearer that, instead of using the "filtering pdf" wording, the authors use "prior pdf" in L65 and "posterior pdf" in L66.

We agree, and changed to "prior" and "posterior" (LL66-67)

L71: Please add "unbiased" before Gaussian.

We added 'unbiased' in the sentence (L72)

L72: It is not clear for me why they are more sensitive to the tails of the prior distribution.

Because no assumptions are made on the prior pdf, it is assumed uniform and, therefore, samples in the tails are more likely to represent the true state of the system than in Gaussian priors. We have clarified the text by removing the confusing element of this sentence which refer directly to tail sensitivity (LL72-74).

L74: I suggest some rephrasing like :"Bayesian inference is used in particle filtering to constrain simulation parameters with observations".

Following the suggestion, we rephrased the sentence (L75)

L75: I suggest to replace "derived" by "computed".

We changed to "computed" (L76) .

L87:103: Wouldn't be better to move this to Sect. 4?

We prefer to leave the series of bullet points in the Introduction, as it helps to illustrate and differentiate better the GLUE methodology of Beven and Binley (1992) from the work done in this paper.

L124: Advanced Himawari Imager (AHI).

We added the full name within the sentence (L125).

L125. Please indicate if this is the retrieval used in this work, or at least link the Met Office retrieval with Francis et al. (2012) in the text.

We modified the sentence, specifying that the Met Office algorithm by Francis et al (2012) is used here (L124), and added a second reference to Francis et al (2012) in the following paragraph (L127).

L147: This implies that it is needed that all the 10-minute (within 1 hour) retrievals are also flagged as "clear", to have a "clear" regrided pixel, right? Isn't this too restrictive?

As detailed in our response to the general comment, and in LL209-215, the average is only data available for XX00 and YY30, so it doesn't require all of the pixels in 6 slots (i.e every 10 minutes). The 'all' in L147 refers only to the classified pixels. So, what that actually means is that you require all the pixels to not contain ash and 90% of them to be defined as clear sky (i.e no significant ash or meteorological cloud).

L149: What about the regridding of the uncertainties?

As explained in more details in our response to the general comment, we do regrid the uncertainties as well. To make this clearer in the text, we added more details on this later in the text, in LL346-350

Table 1: Any particular reason for not using the control MOGREPS-G for your control run?

We performed the control run using parameters, including the driving meteorology, which are default for the operation configuration of NAME. As detailed in Sect. 2.1 of Witham et al. (2019), operationally the Global Met Office weather forecast model) provides the meteorological fields use by NAME.

L188:189: There is no information on the tropical cyclone on Figure 1.

We do not reference a tropical cyclone. If the reviewer meant the extratropical cyclone, we removed the reference to Figure 1.

L189: Please choose Himawari-8 or Himawari in the manuscript

We changed all to Himawari-8 and homogenized the terminology throughout the text

Figure 1: Please add latitude and longitude to the Figure. The space between the panes could be decreased and you could save space by including only one colorbar for all the panels.

Figure 1 was modified following the comment: each panel now shares the same colorbar, and latitude and longitude information were added to the figure.

L203: Following L167, the NAME model outputs 6h averaged values, and you are comparing with 1-hour averaged Himawari retrievals. This difference in temporal collocation can be important. Why this mismatch? Can you compare them with a consistent time collocation, by setting the model outputs to 1h averages, for example? Would you expect changes in your results?

We replied in detail on the averaging and temporal collocation in the general comment and added more details on the averaging in LL209-215.

L231: I am not sure if this sensitivity is used later in this manuscript. If it is not used, this sentence can be removed.

We did not use the sensitivity in the manuscript, but only to check the filter behaviour at each time verification with past observations. We removed this information from the text.

L233: It was later in the text that I understand that the resampling mentioned in L226 was, in fact, to create a full new 1000 member ensemble sampled (and not resampled) from the posterior pdf, and it was not meaning the usual resampling technique (that basically produces copies of existing members, weighted by the posterior empirical pdf). Is there any way to clarify this here, also indicating (and thus repeating L162) that the new ensemble is run from T0 up to Tn? Is this right or I misunderstood?

Yes, that is correct. Following the resampling, a new 1000-member is created, and each member run a 96h forecast using parameters from the new posterior pdf. We reworded the Step 5 (LL239-242) to make this clearer and added a reference to Table 2 in Step 6 (L244), to point to the verification times for each posterior ensemble.

Figure 2: This figure could be clearer. Shouldn't be "posterior pdf" in box 5 instead of "parameters resampling"?. The resampling is done in the "LHS" box, or not?. Are the "ensemble creation" and "NAME runs" the same step (i.e., do you consider than an ensemble is the set of perturbed variables/parameters, or an ensemble is the set of NAME outputs?)?

We modified the figure, removing the "resampling box". LHS now is the main step for resampling the parameters, and the "Ensemble creation" and "NAME runs" share the same box.

L243 : Please define MOGREPS the first time it appears in the text. Also, is MOGREPS the same as MOGREPS-G? (please check the text for consistency).

We defined MOGREPS-G now in Table 1 and L256. It's called MOGREPS-G to differentiate it from MOGREPS-UK which is a high-resolution ensemble. MOGREPS-UK is a more recent creation which is why some early literature simply refers to MOGREPS. We homogenized the terminology throughout the text.

L296: Wouldn't have more sense to, instead of assuming a constant release duration and constant value of the parameters, to assume a parametrised temporal variation of them, following the qualitative information wrote in this paragraph?. How much this assumption could affects the filtering of the ensemble? Can you envisage how to improve this issue taking advantage of the high temporal resolution of Himawari retirievals? (see my very last comment)

Keeping the parameters constant may represent a limitation of this method. However, the main parameters that could be modified based on observations (for the Raikoke case) with some confidence are only the plume height H and duration. The range of duration was already chosen based on the information provided by KVERT. As for H, it is possible to configure a multi-phase source and perform a control run with ash released at different heights at different times. However, perturbing H variations over a range of time windows for a 1000-member ensemble is not trivial and it would increase considerably the already large number of perturbed parameters. Instead, we decided to keep the release constant along a vertical line, to reflect the typical operational set-up for NAME. To tackle this issue, we should first focus on code optimization to make the ensemble size more manageable and efficient (as also pointed out in the conclusions). With a reduced number of ensemble members, it may be possible to introduce within each simulation the temporal variation of H as perturbed parameters. We have added a note in the conclusion regarding this, LL594-595.

L336: Is the filtering of the "clear" pixels (ie., a subset of the no-matching pixels) biasing your HR metric? With this matching pixels procedure, you are removing from your dataset those pixels where the model simulates ash but the satellite indicates "clear". Could this play a role in the possible overestimate of the ash horizontal extension shown in Figure 6?

Generally, the overestimation of simulated ash clouds compared to the observations is expected, and the NAME simulated cloud is always more extensive than the satellite one. With the matching pixels procedure (and MPD calculation), we identify those members in agreement with the observations for pixels where ash ≥0.2 g/m^2. However, when creating the probability maps in Fig 6, the entire dataset

from those members model output is used, without removing any of pixels where ash ≥0.2 g/m^2, despite not being detected by the satellite.

L353: Please see comment on L149. I guess that you are comparing with the regridded uncertainties from the Himawari retrieval. The information on the regridding of these uncertainties is missing in the paper.

We do regrid the uncertainties and we added more details on this in the text, in LL346-350

L360: As the thresholds are defined later in the text, I suggest to add "Sect. 4.2.4" just before "Fig. 3d"

L368: Same as L360

In both cases, we added the reference to Sect. 4.2.4 (L375 and L384)

L366 : Do you mean the absolute value of the difference between simulated and observed values, or you are allowing negatives values of PDs?

We reworded the sentence, to make it clear that we mean absolute values (L381).

L371:372 Both HR and MPD are normalised by the number of matching pixels (Eq. 2 for HR and the averaging of PD for MDP), and I do not see any obvious argument for this statement as the main reason. Aren't these dynamically adjusted thresholds implemented in an attempt to avoid filter divergence rather than justified by the number of retrievals?

As explained in our response to the general comment, we use the dynamic thresholding to avoid filter divergence. However, the system is also in place as there might be a case where all the simulations are rejected – for the reasons explained in those lines. This was never the case for the 11 ensembles, and the thresholds did not vary substantially as Table 2 shows. However, this system was in place to avoid the situation of having insufficient samples for resampling at the next step. We rephrased LL389-391 to make this clearer.

L397: I understand that you cannot show all the ENS, but why did you skip ENS03, while it is the ENS with most retrievals in Table2? (and you also provide ENS03 in the supplementary dataset).

The figure was initially designed to show the difference between two subsequent Ensembles (01 and 02) and between two ensembles after 12 hours (Ens 02 and 04). We now modified the figure 4 by removing the correlation matrix comparison (shown now in the new Figure 5), and adding ENS03 as well, showing the evolution for parameter distributions for EN01, 02, 03 and 04 (L416, Fig. 4).

Figure 4 : This figure is very interesting. "Each parameter in the box plots is normalized by dividing each individual value from the ensemble members by the mean of that entire parameter range from the selected ensemble": Is this meaning that the normalisation is different for the blue and orange boxplots? How can they be compared side by side? Wouldn't be better to normalise by the fixed sampling range of Table 1, such that a zero value means the lower bound of sampling range, and an unity value the upper bound of the sampling range?

The normalization is not different for the blue and orange boxplots, and it is the same within the same plot for an ensemble. For each ensemble, we calculate the mean of each parameter over all the 1000 members. Then both the values for the full ensemble (blue) and for the members WLoA (orange) are normalized using the same mean value calculated for the relevant parameter. This was specified in the original figure caption, LL432-434.

Figure 4: As panel (d) shows relatively small differences of both boxplot colours in comparison with those of (a) , I am wondering how Figure 4 looks like in the following iterations of the filter. Are they similar to panel d? Can you identify signs of filter divergence? It is worth to add this in the Appendix?

We added in the Appendix a new figure (Fig. A1) showing the evolution of the parameter distributions for the ensembles not shown in the main manuscript. Depending on the verification time, the distributions of most of the perturbed parameters for ENS05 to ENS11 is similar to ENS01-

04. We did not notice parameters deviating noticeably from the distributions observed in ENS01-04, although the DFAF and MER F show a higher variability. Both DFAF and MER F are used to perturb the MER, calculated using H. With H being better constrained by the comparison with the observations, their higher variability is not surprising, as they try to perturb MER within the constrained ranges of acceptable plume heights.

L425: The LHS sampling with correlated variables is not trivial, and the Cholesky reference do not add useful information, unless all the process is described. Since this step is fundamental in the method, could you add a reference or explain how the correlated LHS is done?

We added the details for each different step in the main text (LL440-455). In addition, we added a new Figure, Fig. 5, showing an example of distribution fitting for one of the ESPs (Fig. 5a) and moved the comparison between correlation matrixes originally in Fig. 4b to Fig. 5b.

L428: I do not understand why do you use the posterior pdf for some parameters and keep using the uniform pdfs for the internal parameters. Could you provide a justification for this?

We did not observe the same constraining behaviour for the internal parameters, and this seems to apply to all 11 ensembles (Fig. 4 and new figure A1) (We specified that we observed this behaviour over all ensembles in LL425-426). Therefore, we decided to leave the internal model parameters unconstrained and sampled from the same uniform range, and evaluate the performance of each ensemble based primarily on the evolution of ESPs.

L432: This is a qualitative statement and it should be presented as such. Unless you perform the formal proof of it, I would recommend to rephrase this sentence. In addition, could you state explicitly which EPS shows distributions similar to a normal distribution (particle density, duration)?

We both rephrased the sentence and added details on which parameter seems to approximate a normal distribution (L466).

Figure 5: Very nice figure. Similar to a comment above, since the ENS06 shows small peaks in the H, DFAF and MER panels, I am wondering what would be the equivalent of these figures for ENS07 to ENS11.

Figure 5: Just a comment: it is interesting to see that in the H panel for ENS01 the distribution is very skewed but while the algorithm iterates, the right tail of the distribution start to be heavier and the peak smaller. Is this continuing in later iterations? Do this have implications?

Figure 5 If the ENS01 perturbations have very well defined range limits and the filtering is done by rejection criteria, how it is possible to draw samples from values outside these range limits in the next iterations?. For example, the particle density, duration, MERF and DAF panels have tails of the distribution that are away from the ENS01 shading. Are you smoothing the pdfs in the ABC procedure?

For the above comments on figure 5 (now Figure 6): we added in the Appendix the evolution of the distributions for ENS07 to ENS11 (new Fig. B1). The trends in these later iterations for the various parameters is similar to ENS02-06. However, a noticeable variation occurs in the fourth day of the eruption, i.e. ENS10 and 11, where the H increase noticeably. This variation agrees with variations in plume heights observed in Muesel et al (2020) and Bruckert et al (2021), where the authors reported an increase of 6 km or more of the maximum plume top height in the days following the eruption. This increase was related to aerosol–radiation interaction leading to warming of the ash and a subsequent increase of plume height, reaching ~20 km of height in the fourth day after the eruption.

Regarding the limits of the parameters, we did not smooth the pdfs, and those extended ranges result from the resampling strategy described in the general comment. We noticed the increase in range during the design of the ensembles, but decided not to enforce any manual limit in the parameters, due to the small number of members with values outside the ones defined in Table 1 (e.g., for ENS03, 24 members with density values <1350 kg/m^3 and 30 members > 2500 kg/m^3)

Figure 6: I suggest to plot the probability of 0-30 percent of ENS03 in orange colour, to avoid confusion with ENS01. Also, please add x and y axis labels and values.

Figure 6: Why do you compare mean values of satellite burden with these probabilities, and not with the mean value (or median, etc) of your ensemble? Could you add a panel with this information? I think that the probability maps are good, but the mean/median value information could be missing.

For the above comments on figure 6, now Figure 7: we modified the figure, removing the 0-30% contour to make the plot clearer, and leaving the comparison only for the 30-100% contours among ENS01, 03 and Control Run. We also added coordinates labels. A new figure in the Appendix C, same as Figure 7, keeps the 0-30% contours for reference.

The figure was not intended as direct comparison between observed and modelled mean ash value, but to show how the posterior manages to capture more accurately the presence of ash compared to the prior and the control run, in regions where the satellite observations detect ash. The probabilities were chosen instead of the mean, as it is more informative for a hazard mitigation point of view.

L435: Why do you have such large mismatch in the area covered by ash loading? It is because limitation in the Himawari retrievals (and clouds)? Or it is because you assumed a constant flux of ash emission during the period? Other reasons?

Please see comments for L336 regarding the NAME overestimation.

L470: Why ENS08 at T10 and not ENS10? It is because you are interested in 12-hour forecasts?

"T10" was a typo and we changed it to "T08" (L506). ENS08 was originally compared with observations at T08, as shown also in Table 2.

Figure 7: I cannot easily see the colour shading of ENS08 for ash > 2 mg/m2 in this figure. I would suggest to only keep the 0.2 threshold (cyan) for ENS08.

We modified the figure (now Figure 8), and we added coordinates labels as well.

L495: Can you be more specific in the NCAR dataset used here? Why using NCAR reanalysis data if you can use the same meteorological (control) simulation of your ensemble?

The flight tracks were created using NCAR data as they were created for another project that did not have access to the Met office meteorological wind fields. The NCAR data is freely available and so is a commonly used dataset. Comparison of analysis data from a variety of global weather forecasts show that it performs well. We do not expect our results to depend on the choice of global wind fields used to create the flight tracks.

L479: Is there any independent set of observations to validate your results?

We did not have any additional set of observations available at the time of writing. We suggested in the conclusion that a useful follow up would be to use different and independent datasets to validate these results.

Figure 8: Please add latitude and longitude axis and labels.

We modified the figure (now Figure 9).

L537: As in the abstract, I would replace "methodology" by "system".

We replaced 'methodology' with 'system' (L573)

L543:544: Please note that this is not true for two of the six panels of Figure 5.

We specified in L580 that density and duration are not showing the same behaviour as the other parameters.

L546: This is a design choice, not a result of the filtering procedure. Please see my comment above on the justification of leaving these internal parameters unconstrained versus the others.

Please see our reply to the previous comment for L438 regarding our design choice.

L557: After this work, is there any plan to revise the set of parameters to perturb? For example, to add more flexibility on the emission timing and duration pulses for other eruptions; to add some degree of freedom to the temporal variability of the parameters and variables; or improved assumptions on the vertical distribution of the emitted ash?

As originally mentioned in the Conclusions (L591), a first follow up should be the use of different datasets for validating the results, and code optimization. As part of this, especially if in view of a potential operational application, the number of perturbed parameters and members should be revised. Any reduction in perturbed parameters and ensemble size that could lead to similar or improved results may then facilitate the addition of parameters such as those suggested by the referee. We added a note on this in LL594-595.

Anonymous Referee #2
Referee comment on "Refining an ensemble of volcanic ash forecasts using satellite retrievals: Raikoke 2019" by Antonio Capponi et al., Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2021-858-RC2, 2022

The paper "Refining an ensemble of volcanic ash forecasts using satellite retrievals: Raikoke 2019" by Antonio Capponi et al. presents a methodology to improve numerical forecasts of volcanic ash by assimilating satellite retrievals of the investigated ash cloud. The authors apply this methodology to the volcanic cloud produced by the recent Raikoke eruption, and show interesting results with implications for volcanic hazard assessment. I found the paper timely, interesting and clear. The methodology is well described, and the application to the Raikoke eruption is clear. The discussions are in line with the results presented and the conclusions are very interesting. For these reasons, I suggest publication after minor revisions. Here below my comments and questions.

L68 The recent paper "Data assimilation of volcanic aerosol observations using FALL3D+PDAF" by Mingari et al., 2022 could be cited as well.

We added the suggested reference in L70

L140 Just for curiosity, with this retrieval algorithm can you also provide estimates of plume height? For the future, I think it would be very interesting to include in the data assimilation methodology the height of the volcanic cloud. Would it be feasible?

Yes, the algorithm can provide plume height estimates within the produced files for use within InTEM. In the operational form of InTEM at the Met Office this information is not used. The main reason for this is that it would likely not be useful. In general, the accuracy of the NAME modelled height of the ash should be better than the satellite retrieved height, when considering a location where both the satellite and NAME model has ash. This is because there is usually sufficient vertical shear in the atmospheric winds that for NAME to get the ash at the correct horizontal location it needs to be at close to the correct vertical height. As a result, it would likely be a waste of effort to use this data in a sensible way.

L162 "All simulations" indicate the simulations forming all the ensembles, not only the first one, right? Please add information.

Yes, it is all simulations forming each ensemble. We specified this at the beginning of the sentence (L167)

L212 Could you provide information on the computational time necessary to run 1000 simulations? Do you run them in parallel or in serial mode?

We added the information on computational times necessary to run both a single simulation (1 member of an ensemble) and an entire ensemble in L154-157

L268 Please describe all the parameters of eq. 1. not only H, but also g,h and r.

g hr^{-1} specifies the measurement unit for the Mass eruption rate. We moved the unit in L281 to avoid confusion.

L310 Is MOGREPS-G the same as MOGREPS?

It's called MOGREPS-G to differentiate it from MOGREPS-UK which is a high-resolution ensemble. MOGREPS-UK is a more recent creation which is why some early literature simply refers to MOGREPS. We changed all to MOGREPS-G throughout the text.

L397 Why did you exclude ENS03?

We wanted to show the difference between two subsequent Ensembles (01 and 02) and between two ensembles after 12 hours (Ens 02 and 04). We now modified the figure by removing the correlation matrix comparison (shown now in the new Figure 5), and adding ENS03, showing the evolution of distributions for EN01, 02, 03 and 04.

L423-428 I think that more details should be given on the resampling strategy for the posterior pdfs. The correlation matrix (also in Fig.4) and the Cholesky decomposition should be better described.

We have added the details of our resampling strategy in the new LL440-455. In addition, we have added a new Figure, Fig. 5, showing an example of distribution fitting for one of the ESPs (Fig. 5a) and moved the comparison between correlation matrices originally in Fig. 4b to Fig. 5b.

L434 Is the trend of the distributions confirmed also for the ensembles not shown in Fig. 5? I think that this figure should be described giving more details. In particular, the fact that most of the ESPs are skewed towards the lower end of the initial range is interesting and I am curious to understand if you could validate these findings on the ESPs with independent observations of the same quantities. Could you compare the height of the column that emerges from this methodology with independent observations? Particle density seems to be conserved. Could you provide an explanation for that?

We have added in the Appendix the evolution of the distributions for ENS07 to ENS11 (Fig. B1). The trends in these later iterations are similar to ENS02-06. However, a noticeable variation occurs for the plume height in ENS10 and 11, with H increasing noticeably. This variation agrees with variations in plume heights observed by Muesel et al (2020) and Bruckert et al (2021). Here, the authors reported an increase of more than 6 km or more of the maximum plume top height in the days following the eruption starting from around 12 km (+/- 1.5 km) after the onset of the eruption and reaching ~20 km of height in the fourth day after the eruption. This increase was related to aerosol–radiation interaction leading to warming of the ash and a subsequent increase of plume height. This top height agrees with the outcome of the ENS10 and 11 and increase in H that can be also observed for ENS08 and 09 (Fig. B1). Generally, as also stated in the conclusions, a useful follow up would be to use different and independent datasets to validate these results.

Regarding the density: despite it seems to be conserved compared to other parameters, it slowly seems to peak as well. The fact that it is not showing clear variations compared to other ESPs, may be related to the initial range of density: it is based on literature values and probably it was a good approximation for Raikoke from the start.

L461 The figure is well done, but it is really difficult to see the contour lines. Is there a way to improve the readability of the panels?

We modified the figure (now Figure 7), removing the 0-30% contour to make the panels clearer (but we kept the original as Appendix C1), and leaving the comparison only for the 30-100% contours among ENS01, 03 and Control Run. We also added coordinates labels.

L532 It is not immediately evident that panels (a) and (b) indicate the concentration risk of the prior ensemble and ens08, respectively. Maybe subtitles could be added in order to make the comparison between the prior ensemble and ens08 more easy and immediate. The same for the ash dose risk of panels (c) and (d)

We modified the figure (now Figure 9) and added the labels to all panels.