We thank Arnau Folch for the careful review and thoughtful comments which will improve the clarity of our manuscript. Some responses are posted in blue below the comment they correspond to. Comments which are not addressed here will be addressed in the final review period.

- It is unclear to me how the bias correction is effectively applied in each assimilation cycle. Is it a cell-wise correction? How mass load corrections are converted to concentrations (or to "particle" masses in your Lagrangian framework)?
- Since the bias correction is applied to the gridded mass loadings (or concentrations) the procedure would be the same for an Eulerian model. The procedure isn't applied to the computational particle masses but to the gridded mass loading field that was estimated form the mass distribution of the computational particles.
- It is a cell-wise correction. s' = (1-m)s - b where s' is the corrected mass loading for the cell, s is the original mass loading. m is the slope and b is the intercept.
- As for applying the correction to concentrations - Line 171 states "There are several practical considerations in adding or subtracting a constant value to the simulated mass loading or concentration values. Propagating the correction to ash concentrations would involve some assumptions such as dividing the additive correction evenly among the number of ash layers present". This can be seen because the mass loading at each cell, is the sum of the concentrations,c, over each ith level in the column, multiplied by the level thicknesses, L.
- $s = \sum_i LiCi$
- $s' = (1 - m) \sum_i LiCi$ – b
- Ci' = (1-m)Ci – bi
- $\sum_i bi = b$

  The corrected concentration Ci' can be calculated by multiplying the uncorrected concentration by (1-m) and then subtracting a fraction of the intercept, bi. The only constraint is that the sum of the bi equal to b. Different strategies could be devised for estimating the bi values. We could go into some ideas in more depth but it seems outside the scope of the paper.

- The CDF linear fit considers the difference between the model/observation pairs (i.e. the absolute error) as a function of the forecast value, with line intercept shifting mass loading values in the direction opposite the sign of the intercept. My impression is that using absolute errors does not correct evenly the concentrated and the diluted (distal) parts of the cloud. Instead, would make sense to consider relative errors for the linear fit? I am curious so see if this would affect the resulting bias corrections.
- Is the impression from Figure 6 (a) and (c) showing the line fit deviating from the actual differences at higher forecast values? This is not always the case and the fit can sometimes deviate more for lower values. To make the correction more even, a higher order fit may be used. It would also be possible to use a weighted linear regression to make the fit better for higher mass loadings at the cost of possibly making it worse for lower mass loadings. This could be desirable if certain ranges were more important for end users. Currently we use

y-x = mx + b. Where y is observation and x is model value. Utilizing (y-x)/x = mx +b would be equivalent to y-x = mx2 + bx. Which is equivalent to using a second order fit with the y intercept forced to zero. It is probably preferable to just utilize a second order fit.

- It is obvious that the CDF does not inform about cloud location; two clouds may have exactly the same CDFs without any overlap. You mention (line 166) that model points with no observation pair are discarded. This is unclear. Does it mean that bias correction (i.e. "data assimilation") is applied only in the overlap regions (i.e. where both model and observations are non-zero)? Or do you simply mean that the observations domain can be smaller than the model domain but that you actually assimilate zero observation values? Please clarify.
- The following re-wording of lines 165-166 may be more clear. "The modeled and observed values are sorted from greatest to least and then paired so the greatest observed value is paired with the greatest modeled value and so forth. Pairs in which either the observed or modeled value are 0 are discarded. Usually there are more modeled values above 0 than observed because modeled values can cover a larger range."
- As opposed to other DA methodologies, your DA strategy essentially brings model to observations without considering any observation error. The authors know well that in many cases large portions of clouds can be obscured. Could you comment on this?
- We will add some discussion this to the conclusions.
- First we note that CDF matching is not affected by errors which do not change the CDF. For instance, errors with Gaussian or uniform distribution and zero mean. Clearly it will be affected by errors that result in the observed CDF being different than the actual CDF such as bias that occurs when the retrieval fails for a portion of the cloud. The CDF matching will then result in model output with a bias close to that of the observations. We note that some other DA methodologies are bias-blind as well.
- If the model bias is very large, as it was for RunA, then it may still be useful to correct it using the incomplete observations as the observation bias might be significantly smaller than the model bias. However, for a run which has already assimilated observations in some way such as RunM where the model bias is expected to be fairly small, then using CDF matching with incomplete observations would probably not be useful.
- It may be possible to add in some accounting for incomplete observations by trying to correct the bias in the observations first. For instance identifying areas which may be covered by cloud as well as identifying what sort of shape the observed CDF/PDF should take. If the observed CDF can be shown to take a certain form that can be parameterized, then this could be possible.

2. I understand that, for bias correction (DA), you bring model to the observations space, which in the case of himawari-8 implies ~2km pixel size. I missed some discussion about whether discretization issues may exist in Lagrangian models like HYSPLIT. In other words, is 20.000 particles a number large enough to guarantee convergence of model loads, particularly in the distal and/or at the end of the simulation? I strongly suggest running a simulation (e.g. run A or M) with 40k particle and see how this affects results, similar to what authors did with particle size or source width. I refer not only to CDF but also to the metrics in 6.7

We actually bring observations to the model space by regridding to a 0.1x0.1 degree grid and time averaging over an hour. This is discussed in section 3, line 70. "For comparison to model output, the satellite data is parallax corrected using estimated cloud top heights and then composited by first regridding to a regular 0.1 degree latitude-longitude grid and subsequently taking the average of all retrievals within an hour time frame"

We agree that it is worthwhile to discuss the choice of particle number in more depth and will add an appendix with the following discussion. The lowest concentration or mass loading that a Lagrangian model can resolve with the method of estimation used here is determined by the grid size, time step, and amount of mass on the computational particle.

We suppose that the quantity of 0.1 g m^{-2} should be represented by at least 10 computational particles. Then with a horizontal resolution of 0.1o ×0.1o at about 54o latitude, the amount of mass on each computational particle should be no larger than 7.24×10^5 g. Therefore the total mass of each emission chunk which is represented by 2×104 particles should be no larger than about 0.0145 Tg. This condition is generally satisfied for this case. To test we created runs identical to runB but with 1×10^5 particles per emission chunk as well as 2×10^3 particles per emission chunk. Both of these runs produced almost identical emissions estimates, that is, Figure 4 is almost the same for these runs.

For RunA, 2×10^4 particles were released over a 2 hour time period. With a mass eruption rate of 3.75×10^3 kg s^{-1}, a model time step of 5 minutes, and an averaging time of 1 h, the lowest mass loading that the the model can produce (from one particle spending one time step in a grid cell) is 0.0016 g m^{-2}. The mass loadings of interest are about 100 times this, so we conclude that the particle number is sufficient. The situation for RunM is somewhat more complicated. For the individual runs for the inversion algorithm, 2×10^4 particles were used as described above for RunB. Then a run with emissions that vary in time and space was created from the emissions estimates. Currently HYSPLIT evenly distributes the number of particles in time and space so when the emissions are varying, the amount of mass on the computational particles also varies. This makes a simple calculation such as done above difficult. To be on the safe side and because we did not have time constraints on the runs we ran with more than 2×10^6 particles total. The exact number varied for different ensemble members because of the way we handled emission chunks with essentially 0 emissions.

3. Section 6.5 is hard to follow; I lost the thread even after reading two or three consecutive times. In particular:

- Did not understand how the reliability plots are computed and what the vertical axes in Figures 11 and 12 (b,c) show. Please explain better.
- Related to the previous point, what do you mean by "probability of observing the event (line 310)? Do you have an ensemble of observations??
- The addition of the following may answer the previous two points. "The modeled probability of the event on the x axis indicates the fraction of ensemble members which indicate the event occurred. If the modeled probability is 50% then we would

expect that if we look at all the times the modeled probability was 50%, half the time the event would be observed and half the time it would not be observed. The y axis gives the actual fraction of times the event was observed. Ideally, the calibration function lies along the 1:1 line. When the function lies below the 1:1 line, the modeled probabilities are overconfident. For instance a point at (0.80,0.50) means that out of all the times the model predicted there was an 80% chance of occurrence, the actual event was observed only 50% of the time."

- Section 6.5.1 confused me. What is the purpose?
- This may be more clear if the content is moved after or within 6.5.2. The purpose is to provide a more intuitive understanding of the temporal evolution of the rank histogram and reliability diagrams. Also to convey that simply having a flat rank histogram does not necessarily mean a great forecast. The section is referred to in lines 383-385. "As forecast time increases, the case with bias correction approaches, but does not reach, the simple case discussed earlier. The ensemble members overlap less and less with each other which is indicated in the refinement distribution. The calibration function becomes flat with all simulated probabilities corresponding to a low actual probability. The rank histogram becomes quite flat on the lower end indicating a large number of points with below threshold value for the observations and more than half the ensemble members. This is due to increasing difficulty in predicting the location of the ash."

4. A Table summarizing all the metrics you use (and the range of their possible values) would help.

Minor comments and typos

Line 24: "has developed" ➔ "have developed"?
Line 38: 9km a.s.l.
Line 68: In addition to these physical mechanisms, could dilution below detection threshold explain also part of this decrease?
Yes, this is what is meant by "dissipation due to dispersion." Wording will be updated to be more clear.
Line 100:
Line 103: "This is expected to produce a better forecast than for instance using only one cycle"…this is to be checked. For example, by mixing several forecast cycles you may introduce inconsistency in the wind fields, something undesirable in Eulerian frameworks (may be not that bad for HYPSPLIT).
This is a good point. Wording will be changed to reflect. Inconsistency in wind fields manifests different problems in a Lagrangian model vs. Eulerian. For instance,violation of mass conservation is not an issue in Lagrangian model. However violation of the well-mixed condition could occur which could result in spurious higher concentrations in some areas. This is usually a rather subtle effect and the benefits of more accurate wind speeds and directions would be expected to dominate. However this has not been investigated.
Line 111: a mass fraction of ash of 0.1?
Line 187: "right" ➔ "left"?
Line 194: 0:00Z ➔ 0:00 UTC
Line 197: "This indicates a possible issue with turbulence parameterization in the model which control the rate at which the plume disperses". This is true, but actually could be more complex as diffusion effects can actually mix with wind shear advecting

differently as particles settle down. With passive (non vertically resolved) satellite observations it is impossible to distinguish the single contributions from these two effects.

Line 266, eq (1). Please make evident that C depends also on time.

Line 275: “information about the relationship between concentrations in adjacent grid cells for each member is not preserved”. I do not understand this sentence. Do you mean that HYSPLIT does not output concentration at height levels at different periods?

Lines 275-285. Argument difficult to follow, please explain better. Why dosage cannot be computed individually for each ensemble member and then do your percentiles?

We do state this in Line 268 “For probability of exceeding a critical dosage D_C, this equation should be applied to each ensemble member and ensemble relative frequency of exceeding the dosage computed from the resulting ensemble of dosages” The conversation here was meant to convey that if an end user does not have access to the full ensemble data, and only has APL or ATL data, then information on dosages cannot be accurately inferred from just that information. The following changes may convey this better (underlined is added).

“However~~, combining these~~ utilizing the probability of exceeding a concentration or ATL to get probability of exceedance of dosage along a route containing multiple grid cells is not possible because information about the relationship between concentrations in adjacent grid cells for each member is not preserved by the ATL field.