Comments on “Evaluation and Bias Correction of Probabilistic Volcanic Ash Forecasts” by Crawford et al.

Summary
This paper considers ensemble-based volcanic ash forecasts using the HYSPLIT model with 31 GEFS meteorological forecasts for the October 2020 Bezymianny eruption. The cumulative distribution function (CDF) matching is used to reduce ensemble bias, and different metrics (rank histograms, reliability diagrams, fractions skill score, and precision recall curves) are used for further forecast verification. Different sets of runs are considered with various source terms and particle sizes, e.g. setting the source as on the operational source setting (run A) or through inversion (runs B and M) considering also different assimilation periods.

General comments

1. I found interesting the use of the Cumulative Distribution Function (CFD) matching as a way to correct ash forecast bias. This is, to my knowledge, quite novel in the field. I have some questions/comments here (Section 5):

- 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)?
- 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.
- 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.
- 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?

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

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??
- Section 6.5.1 confused me. What is the purpose?

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?
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).
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?
Line 310: "P(..), giving the probability of observing…"
Line 359: (b)?

Line 395: You can also cite Folch et al. (2022) here, where skill scores are generalized to probabilistic contexts.
Line 406: Figures → Figure