Natural Hazards and Earth System Sciences Discussions

Nat. Hazards Earth Syst. Sci. Discuss., referee comment RC3
https://doi.org/10.5194/nhess-2021-96-RC3, 2021

Comment on nhess-2021-96

Claire Witham (Referee)

Referee comment on "An ensemble of state-of-the-art ash dispersion models: towards probabilistic forecasts to increase the resilience of air traffic against volcanic eruptions" by Matthieu Plu et al., Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2021-96-RC3, 2021

General Comments

The paper presents a multi-model study of a short period of the Eyjafjallajokull eruption in May 2010. As such it is a very limited study, but it highlights some useful challenges with using and comparing a multi-model ensemble for probabilistic volcanic ash forecasting. Some further details on the source terms used would be helpful for the reader to understand the differences between the model outputs. The authors apply the Fractions Skill Score as their measure of model skill, but I have some serious concerns about the way it has been applied here. The WRF-CHEM a priori simulation is clearly much worse than the other models, but this is really not reflected in the FSS. I believe the application of the pixel filtering is biasing the FSS results to suggest the models are performing better than they are. The use of the 99th percentile is also not mathematically justified. With some revisions to the statistics and data displayed, this would be a more balanced and hence interesting paper for the community.

Specific comments

Line 23: The statement “predefined procedures during such events were missing” is factually incorrect. There were very well established procedures in place laid down by the International Civil Aviation Organisation which can be found in the Handbook on International Airways Volcano Watch (IAVW). This statement needs to be deleted or amended. It is true that these procedures did not cover hazardous concentration levels.

Line 116: Please could the authors provide further details on the size distribution, in particular the size range. This is important in order to understand differences between the models. From the description of it being aerosol the implication is that the sizes are small (<2.5 um) compared to FLEXPART, but this may not be the case.

Line 131-133: It would be much more helpful to the reader if the size bins could be given in their metric equivalent rather than the phi scale. This would allow direct comparison with the other models. I believe this is effectively saying that the size range included is 0.1 – 62.5 um diameter, implying the largest particles are 3 times larger than those in FLEXPART.

Line 168 and Line 171: As there was no umbrella cloud present during this phase of the Eyjafjallajokull eruption, please could the authors provide more details on how the models work. Perhaps this is just a poor use of terminology and the use of "umbrella" needs to be changed to e.g. "maximum plume height".

Throughout this whole section there is no information about the mass erupted. This seems a strange omission. As the mass eruption rate (or total mass) is one of the fundamental source parameters, it needs to be mentioned. For example, is the same mass applied in each case even though different particle size ranges (and hence different fractions of the total ash mass) are being represented? If not, then the implications of using different masses (and particle sizes) need to be discussed.

It seems unfortunate to have used different a priori source terms in the four models. Presumably this is because these are the data that were available. It would still be helpful if the authors could justify why they did not rerun the models with the same a priori. It is hardly surprising that the a posteriori ensemble outperforms the a priori ensemble as it contains less uncertainty.

The WRF_CHEM output is a clear outlier in fig5. I think the authors need to provide some explanation as to why this is. From figure 1, it doesn't appear that the WRF-CHEM simulation is emitting more mass than the other simulations, but it clearly contains more mass in the other figures. The fig1 caption says the "the source terms of fine ash"… Please explain in the text what this means. I also wonder if this is the issue..? i.e. is additional non-fine ash being emitted in the WRF-CHEM run? Also why does this simulation start 6 days before the others? Does this mean it contain extra sources and hence extra ash compared to the others? This is why it's important to provide full details of the mass and particle sizes as mentioned above.

Line 267: Can I check that the value of 0.2 g/m2 is correct here? Based on the plots in Figure 3, this appears to be a large area, rather than the highest contamination area. Perhaps the colour scale in Figure 3 is what is not helpful for the interpretation here, as 0.2 is white, which also appears to be the colour of no ash. A different plotting scale is needed if that is the case.

Following on from this, a more detailed description of what is meant by "For each model output, the G grid points with the highest ash concentration in the domain are kept for further analysis and used to calculate the FSS." is needed. Firstly, I assume that total column load, not "concentration" was used? Second, it implies that a different set of G grid points is derived compared to those determined from the satellite data. This is the approach used by Harvey and Dacre, but this is not at all obvious in this paragraph for those unfamiliar with the FSS.

I have some serious reservations about the current application of the FSS approach to compare the different models. Particularly when the paper is aimed at quantitative model output. As noted above, the WRF-CHEM a priori data is a clear outlier, but using this approach the FSS is only slightly worse. I understand the aim of applying the normalisation, but this does not sit well with me and I think gives a widely spread model a better score than it should. For this work, I would suggest that the application of the FSS would be more scientifically rigorous if a threshold approach was used. The obvious choice would be to apply the same 0.2 mg/m2 threshold to each model. This would be a true measure of the quantitative spatial skill of the model and allow a better comparison both between models and between the a priori and a posteriori results. Currently the statistics are very misleading.

Line 289: FLEXPART appears to perform worse with the a posteriori according to these FSS. Are the numbers correct? If they are, then this aspect should be discussed.

Line 298: It would be helpful to discuss that the reason for these differences in ash load for the a priori runs is because different masses are used in the sources. That they look more similar in the a posteriori in partly due to using the same mass.

Line 307: The Marenco et al measurements were taken along a NW-SE trending line over the UK on the 16 May, so it's unfortunate that the authors have chosen to use a SW-NE trending line for their cross-sections. There is a suggestion in Fig 6 that the FLEXPART and MATCH modelled ash over the UK is indeed at altitude and so the authors may be being overly critical of their results. It is hard to see clearly in Fig 6. Ideally the authors would produce new plots with a NW-SE cross-section. If this is not possible, then it would be helpful if the longitudes of the UK coastline could be marked on the x-axes, as this would allow a better comparison to the Marenco results. Because the y axes are in pressure coordinates it would also be helpful to provide the corresponding pressure values in the text alongside the "4 and 6 km" reference.

Line 335: The 99% percentile has no meaning for a 4 member ensemble (valid values are 0, 25, 50 and 100) and strictly does not apply for a 12 member ensemble either. I recommend the authors use 100% for a meaningful statistic rather than 99%. This is unlikely to affect their results.

Line 428: "To estimate the long-term damage due to high ash dose, we recommend using the median of the ensemble as this gives the best estimate of ash distribution." No scientific justification or proof of this statement is provided, therefore it is just conjecture. I recommend this sentence is deleted.

Technical Corrections

Line 22: Use of English: "forced to cancel" replace by "forced the cancellation of"

Line 25: "type" should be "types"

Line 28-29: These thresholds are incorrect. The contamination levels for which information is provided by the two VAACs are 0.2-2mg/m3, 2-4mg/m3 and >4mg/m3. Please correct.

Line 34: please clarify if "medium ash concentration" referred to here is the same as the medium contamination level specified earlier in the paragraph

Line 35: the use of "shorter intervals" is ambiguous here, it could refer to shorter exposure intervals, it would be helpful to be specific "shorter maintenance intervals"

Line 45: Recommend changing "can" to "could" as this has not yet been demonstrated in practice in a real event. There are other factors that would also need to be considered, such as traffic volume.

Line 88: change "regardless the" to "regardless of the"

Figure 3 and Figure 5: The colour bar captions say "Total ash concentration". This needs to be changed to "Ash column load".

Line 318: change "30-years" to "30-year"

Line 334: change "in in" to "in"