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.
The sentence has been deleted. In the scope of the new introduction as suggested by RC1, this part has been re-written. More specifically, the new introduction states: "In order to mitigate the consequences of such types of volcanic eruptions on aviation, operational centres continuously watch and issue warnings about ash dispersion in the atmosphere, that support the decisions in the frame of predefined procedures (Bolic et al, 2011). This watching and warning role worldwide has been the duty of Volcanic Ash Advisory Centres (VAACs)." The new introduction is enclosed as a response to RC1.

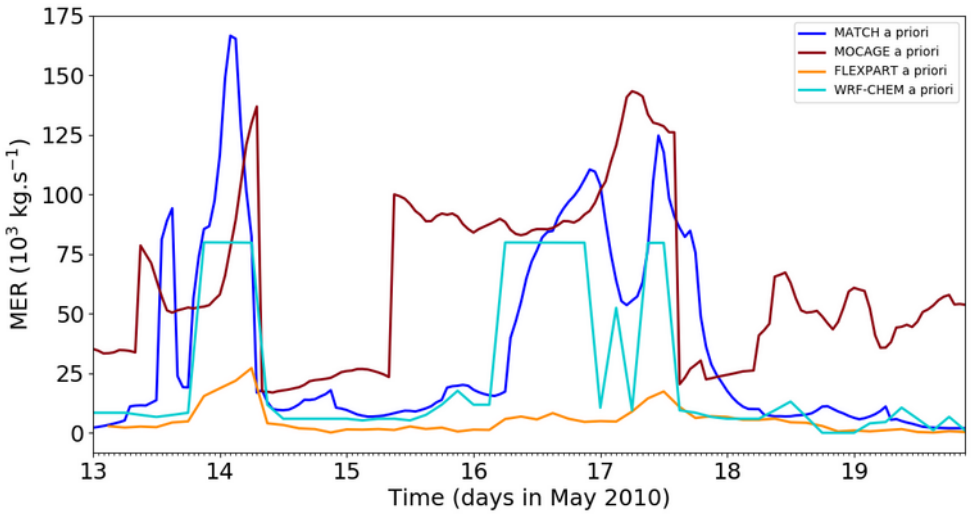
Line 116:
The size distribution of ash in the four models is now described in a table, that is described in the response to RC2. In particular, it is stated that the aerosols in MATCH are given in one single bin regarded as coarse fraction (2.5-10 μm). As stated in the manuscript, the deposition and optical properties of aerosols assume a prescribed size distribution.

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.
In the table that summarizes the ash bin size and mass distribution for the different models (see answer to RC2), the values of bin size are given μm. Reference to phi scale has been removed.

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".
Two models assume an umbrella-shape plume (MATCH and WRF-Chem) even if the plume was not observed as an umbrella during this phase of the eruption. An umbrella-shape is a common assumption that means that most ash mass is released at highest plume levels, as explained in detail in Stuefer et al. (2013) and Hirtl et al. (2019).
Since this "umbrella" property is stated in the manuscript as a model assumption only (no matter what was the shape of the plume in reality), we believe we can use the terminology "umbrella" as it is in the manuscript.

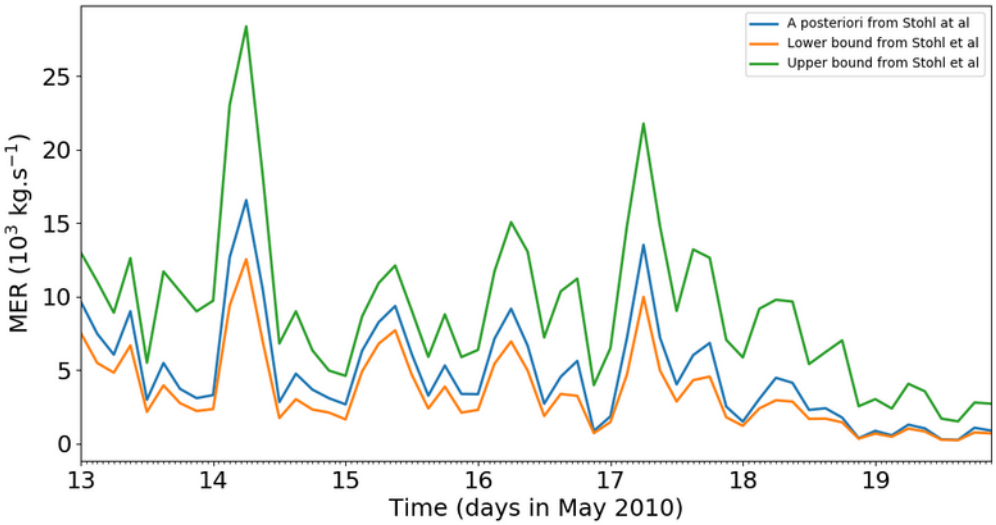
Throughout this whole section there is no information about the mass erupted. This seems a strange omission. As the mass eruption rate (or total mass) is one of the fundamental source parameters, it needs to be mentioned. For example, is the same mass applied in each case even though different particle 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.
In the study that was reported in the submitted version of the manuscript, the mass eruption rates (MER) of the different source terms has not been compared. Following the reviewer's suggestion, we have computed the MER from the different source terms and found some inconsistencies for two a priori model runs (MATCH and WRF-Chem) based on the Mastins et al. (2009) relationship. These source terms were corrected.

After these corrections, the time evolution of the MER of the four a priori source terms are:



The MER of the MATCH and WRF-Chem a priori simulations are in good agreement and rather close to the Mastins et al. (2009) relationship – this has been checked aside. However, some differences remain, which can be explained by different assumptions used for the computation of the source terms: the fraction of fine ash that is kept from the total mass erupted, the eruption height that is assumed as entry of the models, and some parameters such as the density of ash for instance. These are part of the possible uncertainty of an a priori source term. The MOCAGE MER (based on FPLUME source term) generally follows as similar order of magnitude with these two source terms based on Mastins et al. (2009).
However, the FLEXPART a priori simulation, based on the PLUMERIA model, has a lower MER than the other models. Overall, there is now a rather good agreement between the MER and the ash column load, which was, to our understanding, an important concern for the reviewer, as raised in the next comment.

Besides, we have computed the MER for the 3 a posteriori source terms:



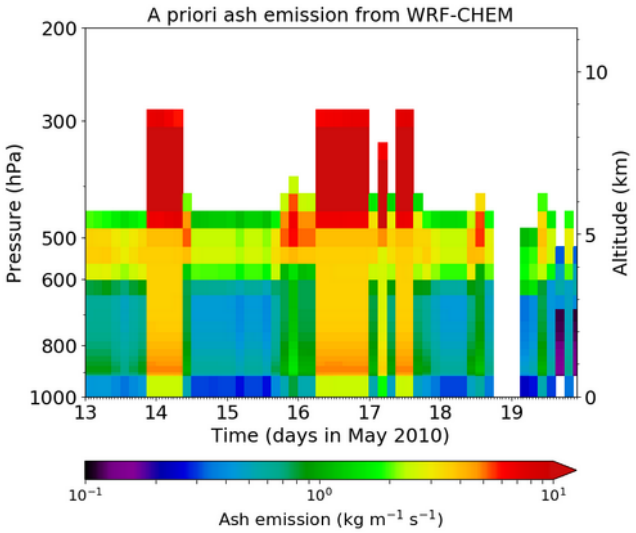
The values of MER of the a posteriori source terms are much lower than the a priori ones. We propose also to add the two figures of the MER in the manuscript, in order to discuss the performance of the models and the ensemble.

In all the figures shown in the present text and in the revised manuscript, the a priori source terms have been corrected and the model simulations and ensemble computation have been updated.

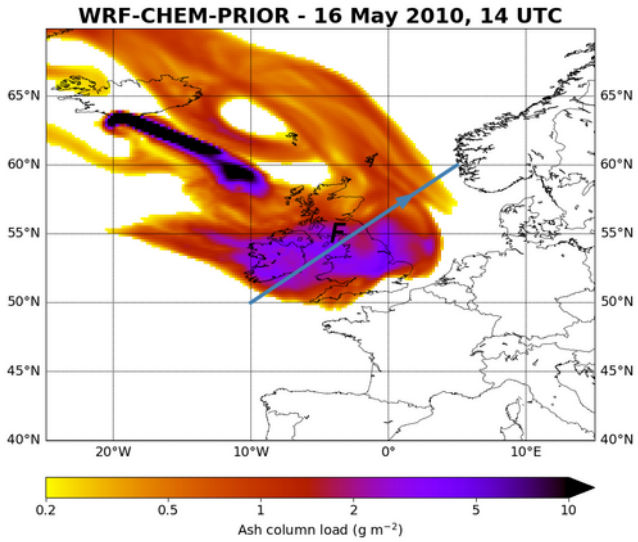
It seems unfortunate to have used different a priori source terms in the four models. Presumably this is because these are the data that were available. It would still be helpful if the authors could justify why they did not rerun the models with the same a priori. It is hardly surprising that the a posteriori ensemble outperforms the a priori ensemble as it contains less uncertainty.
Our objective was to simulate what kind of source term (in their diversity) could be used in real-time, so every centre will have its own source term. This argument is provided in the discussion section. There are indeed differences between the MER of the different a priori simulations, as well as between the aerosol size distribution. These are different kinds of uncertainties that should be taken into account. This argument has been added in the presentation of the a priori simulations and the source terms.

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.

Thanks for pointing that out. We checked every step again and found a bug in the routine that extracted the source term values out of the WRF-Chem a priori simulation. The correct representation of the source term has much higher values (see below):

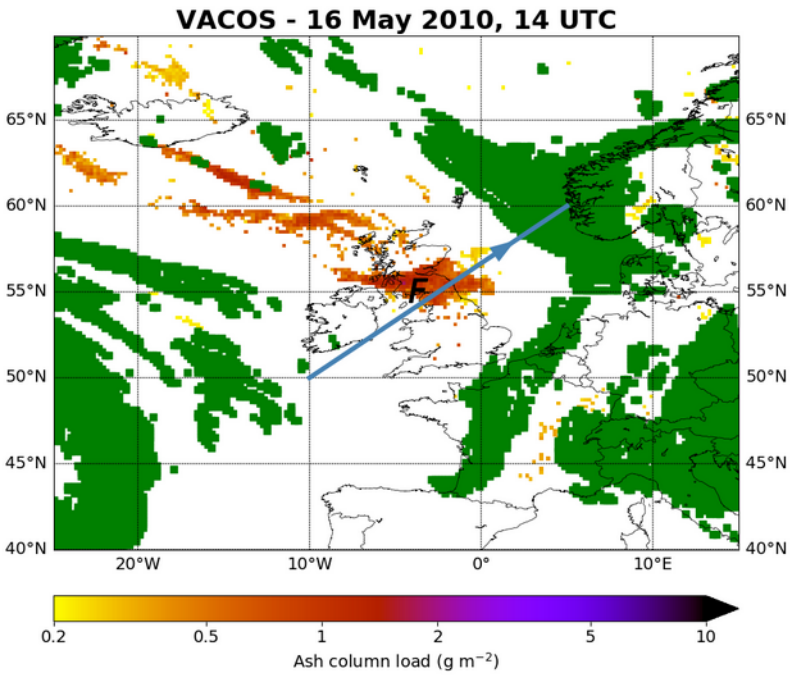


This figure is also included in Fig. 1 of the manuscript.
Besides that, the ash concentration of the model run is slightly reduced compared to the previous version because the Mastin et al. (2009) relationship is now referred to the actual volcanic vent height rather than the model orography.



These two changes make the WRF-Chem a priori simulation more consistent as the ash concentrations agree better with the amplitude of the source term.
Ash column load of this corrected WRF-Chem a priori simulation remains on the high side among all a priori simulations, with comparable values with the MOCAGE a priori simulation. After correction of the MATCH simulation, these three simulations look in a similar range of values, consistently with the MER of their source term.
All models simulated the emission of fine ash with “fine ash” defined slightly differently, e.g., with particle diameters smaller than 25 µm (FLEXPART) or smaller than 31.25 µm (WRF-Chem) (see the table describing the model configurations given as answer to RC2).
It is true that the spin-up period of the WRF-Chem simulation was longer than that of the other models (information is now also included in the table describing the model configurations given as answer to RC2). However, there is no obvious argument to attribute model differences to different spin-up lengths.
The interpretation of the results has been updated according to these changes.

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.
Figures have been updated with a new plotting scale, which clearly separates the values below 0.2 g/m2 (white) and higher than that (starting from yellow). Here is an example :



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.

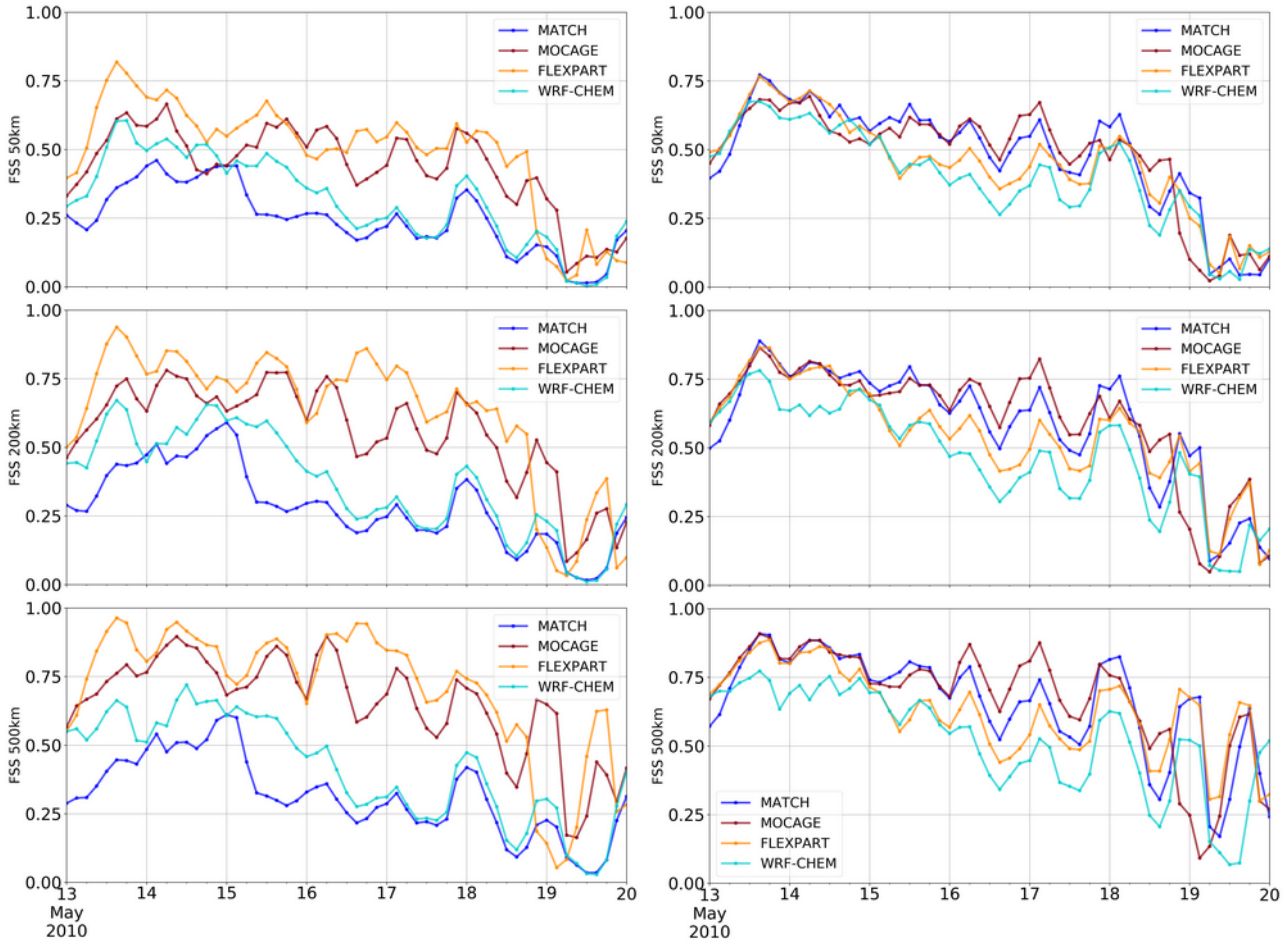
We indeed computed the FSS using this method. The text in the manuscript has been updated for a clearer description of the FSS, as: "For each model output, the G grid points with the highest ash column load in the domain are kept for further analysis and used to calculate the FSS. This implies that a different set of G grid points is derived compared to those determined from the VACOS data. After the normalization step, the FSS is a measure of the performance of the models to locate the most intense ash features, and it filters out the amplitude errors (see the supplementary material for the FSS without normalization)".

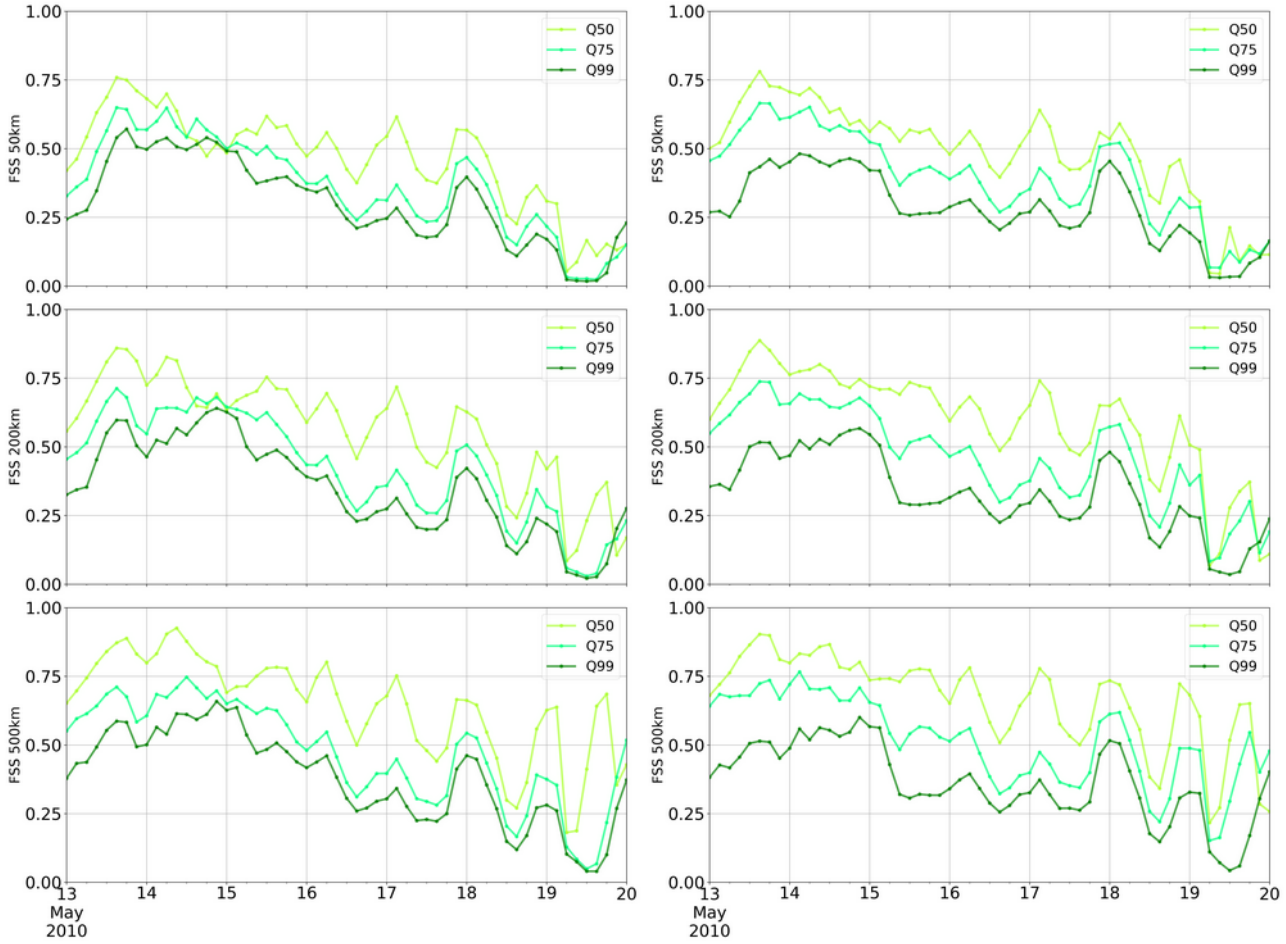
I have some serious reservations about the current application of the FSS approach to compare the different models. Particularly when the paper is aimed at quantitative model output. As noted above, the WRF-CHEM a priori data is a clear outlier, but using this approach the FSS is only slightly worse. 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.

We have computed the FSS of models and of ensemble percentile fields by applying the same 0.2 g/m2 threshold (called "without normalization"), and it provides very different results. See below.

A PRIORI

A POSTERIORI





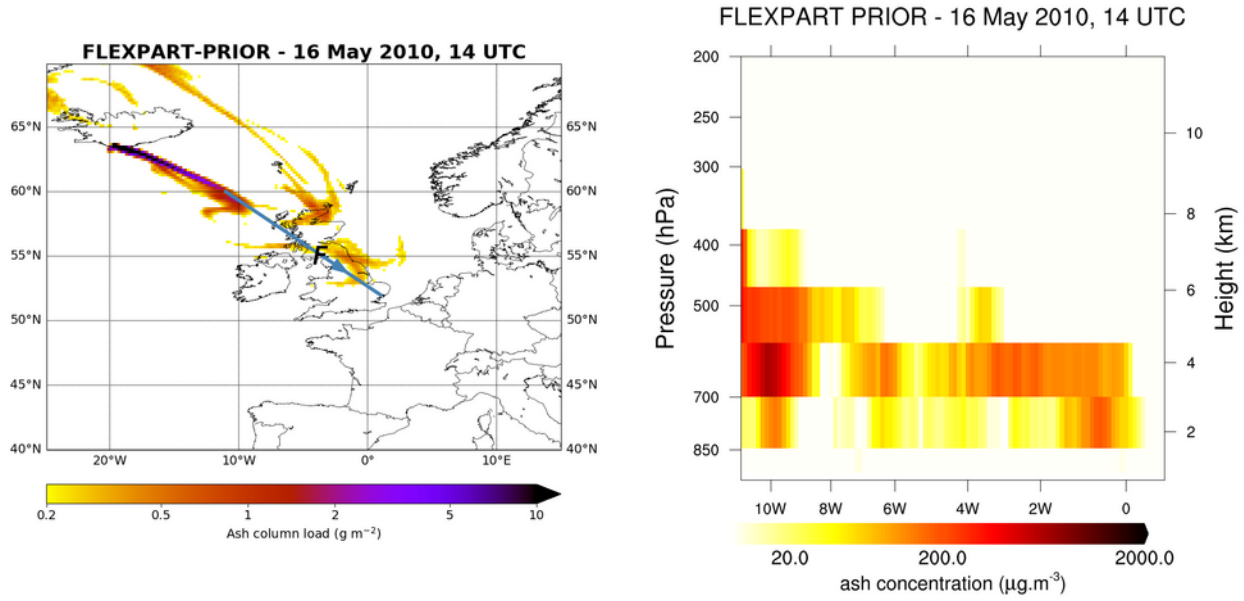
The FSS without normalization provides indeed a very different picture of model performance, compared to the FSS with normalization. Without normalization, a more clear ranking of the model performance is obvious, but it relates more to the ash load of the models (i.e., MATCH and WRF-Chem a priori runs overestimate ash load and have lower FSS than the other models).
As the reviewer writes, the normalization step helps to filter out amplitude errors, and that is the reason why the “direct” (i.e., without normalization) FSS is lower than the normalized FSS. The FSS with normalization is helpful to characterize localisation errors, complementary to amplitude errors. As a conclusion, we suggest to keep the FSS with normalization in the manuscript, and to add the FSS without normalization in the supplement. In the text of the manuscript, some discussion of this point has been added, together with addressing the explanation of FSS (see previous comment).

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.
Yes, it is true that, at this instant (16 May at 14 UTC), the FLEXPART a priori performs better than the a posteriori, but this is not always the case. This point is noted in the new manuscript. Generally speaking, the a posteriori run is not always better than the a priori run. The data and the model used for computation of the a posteriori source term are different from the ones used in the present study (different Flexpart version, different ash load satellite data).

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.
Together with the introduction of the plots of time evolution of MER, a discussion of this has been adeed in the manuscript. It is clear that the MER has direct impact on the ash load.

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.
The cross sections in the submitted version of the manuscript have been chosen to be consistent through the manuscripts and through the supplementary material, for the models (Figures 3, 5, 6, 10 and 11) and for the ensembles.

The new cross-sections for the 16th May at 14 UTC are (example shown for FLEXPART a priori simulation):



The right y-axis of the vertical cross section shows height in km. In the new figures (see above), the scale has been refined to 2-km spacing.
Placing the UK coast on the vertical plots is not so easy. The best we can say is that UK is between 5°W and 0° roughly. The flight reported by Marenco et al. at this time was between 2°W and 4°W; the horizontal scale has been updated accordingly.
To be consistent, the figures showing the ensemble quantiles have also been updated accordingly.

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.
The 99% percentile is obtained from linear interpolation between the 75% percentile and the 100% percentile, which is also the maximum value. In the representation of probabilistic forecasts, the debate is opened on what highest quantile to represent. From a user perspective, 100% or maximum may be understood as the value that can never be exceeded, so it can be misleading. To avoid such

misunderstanding, we preferred to show a high but not the maximum quantile. We note, however, that this does not make a significant difference for interpreting the 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.
This overstatement has been deleted.

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

Line 25: "type" should be "types"
Done

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.
Done (Note : this modification has not been included in response to RC1, but will be done in the final manuscript)

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"
The sentence "While medium ash concentration over long time can also be a safety issue, low ash dose can lead to longer term damage, which becomes an issue for engine maintenance as shorter intervals are needed in order to prevent performance loss (Clarkson et al., 2016)."
has been rewritten as:
"While accumulated long-term exposure of ash at lower concentrations can also be a safety issue (e.g., 1 mg m^{-3} for about 3 hours), low ash does can lead to long-term damage to the engines and may require shorter maintenance intervals in order to prevent performance loss (Clarkson et al., 2016).
(Note : this modification has not been included in response to RC1, but will be done in the final manuscript)

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

Line 334: change "in in" to "in" .
Done