Overall there is not much that the reader can learn from this paper. A number of model studies already exist (as cited in the paper) that describe the difficulties in reconstructing or even forecasting the ash dispersion from a volcanic eruption. The main finding is always that the source strength of the volcanic ash is the main difficulty for an accurate forecast of the ash concentrations. The authors of this paper confirm this finding in their comparison to observations of the particle number density at Hohenpeissenberg, where the best agreement was found with a reconstructed source function (i.e. one that was already based on observations). They also confirm the finding that early observations close to the source are necessary to feed the model systems with information about the emitted mass and the size distribution of the volcanic ash. The authors finally propose that a time lagged ensemble forecast could give probability distribution functions as forecast products instead of one map derived from a deterministic forecast. This is certainly a good idea, however their ensemble includes only different starting times of the forecast. This means that only effects of the meteorological fields are considered in the probability distribution while the main uncertainty is the emission strength.

I recommend major revisions before the paper can be published. The authors should include other sources of uncertainty into their ensemble forecast and they should extend their study beyond the first eleven days after the eruption.

We thank the reviewer for his thorough review of our paper and his critical and constructive comments. After his and the other reviewers remarks on the structure and the content of our paper we have totally revised it. Moreover, we carried out additional model runs and focus on a more detailed evaluation of our time lagged ensembles.

Many aircraft and lidar observations have been done in May 2010, when the volcano was still active. These observations could serve as a test bed for the quality of the ensemble forecast that is proposed here.

We know about that data which is now available but it was not when we performed most of our simulations. Simulating the period in May 2010 would be a completely new effort. The simulation of this period is therefore reserved for future work.

Specific comments

Page 13440, l5-7: If the model is calibrated and the emissions are modified accordingly it should show good agreement with the observations. The term "nearly perfect" is inappropriate, in particular because no numbers are given for this comparison.

We now give a quantitative measure for the agreement.

Page 13440, l22: Besides the temporal and the spatial distribution, information about the mass concentrations is needed most urgently.

We added this in the revised version.

Page 13441, l3-4: As far as I followed the publications about the Eyjafjallajökull eruption, the VAAC forecasts were not quantitative either when the eruption started. Other model systems could only give quantitative information after being compared to observations. This kind of "calibration" was the same as it is done in this paper.

We have changed it in the text.

Page 13442, l21-24: How are you going to determine the reliability of the forecast if you take only uncertainties in the meteorology into account and not those in the source strength?

We have performed additional ensemble runs varying the source height and strength.

Page 13443, l 14: What was the width of the size bins? Why is coagulation neglected? Please justify this.

We used 6 different mono disperse particle classes with diameters of 1,3,5,10,15, and 30 µm. In this size range neglecting of coagulation is justified.

Page 13444, l1: Why did you scale the source strength linearly with the source height? Why didn't you use the source function from Mastin (2009) right from the beginning?

We skipped the part with the simulation with the linearly scaling of the source strength.

Page 13444, l18: What do you mean by "own" analysis?

We described the method in more detail and added a reference.

Page 13445, l7: I cannot derive from Fig.2 that the horizontal distribution of the ash plume was captured "quite well". As far as I can follow the Meteosat picture, the plume was located in Northern Germany along the coast to the Baltic Sea. In this area, the model results reveal no ash.

As the Meteosat picture does neither allow a qualitative nor a quantitative comparison with model results for several reasons (e.g. cloudcover) we skipped this part of the paper.

Page 13446, l6: You show that the model produces thin ash layers when 80 vertical levels are used. What is the spacing of the levels in the altitude of the ash layer? How thick is the layer in the observations?

We added this information in the text.

Page 13446, l11: “The simulated ash layer above the boundary layer starts to decrease”: this is unclear. What is decreasing? The concentration, the vertical thickness, the altitude?

We apologize for this confusion due to the used wording. We reformulated this section.

Page 13446, l19-23: What can I learn from this study of the modified emission profile? Couldn't you also find a modified emission profile that shows better agreement with the observations?

We skipped this part of the text.

Page 13446, l27: What does the fact that the parameterization of deep convection does not change the results tell us? Is it not important at all or just not important in this case (and for this specific location). Shouldn't you better analyze the model results at other locations, too?

We evaluated the model runs in more detail. For this time period there was not much convective activity in the model domain. Of course that does not mean that this process is always of no importance.

Page 13449, l22 - page 13450, l14: In this section I am missing comparisons of the simulated mass densities with observations, e.g. from the Falcon flights. Why do you restrict yourself to the number densities in the previous section while the mass is more relevant for the flight restrictions?

As number concentrations and mass concentrations are related via the density of the ash particles which we adopted from the Schumann et al (2011) paper this should not make any difference. We made such comparison and found a sufficient agreement.

Page 13449, l28: It is certainly a good idea to give a probability forecast but you should include all uncertainties in your ensemble, in particular the source strength.

We have performed additional model runs with modified source height and source strength. As we used observations to calibrate our model results the uncertainty due to the source strength is drastically reduced.

Page 13450, l24: I assume there is a “high” missing between “resolution is” and “compared” in this sentence.

This section was rewritten.

Page 13450, l15 - page 13451, l20: This is a summary of what has been done. I am missing real conclusions except the recommendations for the observations necessary to help the models to calculate the correct concentrations. These *observations should not include only aircraft and lidars but also passive instruments like radiometers, either ground based or on board of satellites.*

We agree and have rewritten this section.

Figures:

Fig. 2: The Meteosat picture is difficult to interpret.

We agree and skipped this Figure.

Fig. 3: The inlay in Fig. c) is hard to understand. What is shown here?

We removed this figure.

Fig. 5: What was changed in the emission profile (orange curve)? Does this correspond to Fig 3 c)?

We removed this figure.

Fig. 6: How can you justify to use the size distribution from 2 May for your simulations in an earlier period?

This was the only date were airborne size distributions were measured close to the volcano. Therefore, we had to use this data.

Fig. 8: Add the date for this simulation result

We have added the date.