Responses to referre#1

We thank Referee#1 for his/her useful comments. Following the editor's recommendation, each response to comments will be organized as follows: (1) comment from Referee in bold, (2) author's response in italics and (3) author's change in manuscript. Some responses are given to several comments at the same time when these comments are related to each other. The changes in the revised manuscript, except the small edit corrections, are highlighted in blue colour in the revised manuscript.

Following comments from referee#2, we made several changes in the manuscript. In particular, to show more clearly the agreement between forecast and observations, the 4 panels in Fig. 4 have been replaced by only one showing a zoom over the areas of interest for the 10th of June 2014 at 15UTC when the ozone episodes are both peaking. Doing this, we found a small error in the plotting procedure. A few stations were missing in Fig. 4a. This has been fixed in the revised version.

General comments

Given that this is labeled a "model experiment" paper, I would like to see a clear presentation and a more in-depth analysis of some scientific questions. This paper presents ensemble output statistics without going into much analysis of the underlying reasons for observed patters. A deeper analysis would lend weight to the paper. I see that this paper is for a special issue, so perhaps the above concerns are less relevant. However, even if the paper is intended to be taken in context with other papers in this issue, a clear statement of purpose of this particular paper is needed.

We agree that there was not enough in-depth analysis of scientific questions in the paper. We have added results and discussions on the following subjects:

- *information on the differences, strengths and weaknesses of the 7 models (sections 2.2 to 2.8) which helps understanding the differences in the forecasts scores,*
- *a more complete analysis of the ozone episode in June 2014 (new section 3.3),*
- *an analysis of the three-monthly performances of the 7 models which serves for understanding the ensemble scores (Section 3.4),*
- *an analysis of additional tests on the influence of missing models in the ensemble median scores over a three-months period (section 3.4).*

This led us to change the organisation of Section 3. In the revised manuscript, section 3.2 is now "Availability statistics". It only includes the information on the production reliability of the daily forecasts and analyses (text unchanged). All the discussion about the ozone episode in June 2014 is now grouped into section 3.3. It includes the observation plots (Fig. 2 in GMDD manuscript), the EPSgrams (Fig.3 in GMDD manuscript), the map of the forecast with superimposed observations (Fig. 4 in GMDD manuscript), the 7 models and ENSEMBLE performances over the selected week (Fig.5 in GMDD manuscript) and the tests with less than 7 models in the ENSEMBLE (Fig.6 in GMDD manuscript). Please note that the numbering of the figures has been changed.

Additional and more detailed information is given in responses to the specific comments.

The corresponding changes in the manuscript are given for each specific comment.

Specific comments

Introduction. As mentioned above, make it clear why this paper is needed, given that there are already a series of 6-monthly reports being published.

The first aim of the paper is to make a description of the current state of the daily air quality production. In order to improve the paper and following your comments, we have added in the introduction a second objective which is to analyse the performance of multi-model ensemble.

This has been changed in the revised manuscript. There is now a second objective of the paper stated in the introduction which is "to document and to analyse the performance of the multi-model ensemble".

In the introduction, it would also be nice to add information on how interested users can access the forecasts (I assume they are publicly available).

We agree that the fact that the data are publicly available was not clearly stated.

This information has been added in the introduction. Also details are given on how to access the data in Section 2.1.

Section 2.2-2.8. This section takes up a lot of space re-describing individual models that are described elsewhere. It would be far more interesting to read a critical analysis of the differences between the various models based on the experience with the forecasting ensemble to date. For instance, what are the strengths and weaknesses of the different models? Which differences between the models are most decisive in leading to differing model forecasts?

Although the 7 models are already described in publications, there are specificities in the MACC-II configuration for some of the models. This is why the main features of each model are given in section 2. The authors agree that information on the strength and weaknesses of each model is useful and helps to understand the multi-model ensemble performances.

Sections 2.2 to 2.8 now include this information which is used for the analysis of the performances of the 7 models and Ensemble in Section 3.

Figures 2 and 3. Can the model forecasts shown in Figure 3 be superimposed on the observations shown in Figure 2, so that the reader can more easily compare models and measurements?

This is not possible for us to easily superimposed the two figures but to help the reader we have merged the two figures in one with the left panels being the observations and the right panels being the Epsgrams, and we use the same range for the vertical scale. Also in response to referee#2 we have added comparisons to other stations to support more strongly our analysis of the ENSEMBLE behaviour for this case study.

The new figure labeled Figure 3 (merging previous figures 2 and 3) has been inserted in the manuscript and now includes comparison with two other stations, one in Germany and one in the South of France. Also a more comprehensive analysis of the case study has been done.

Section 3.4. Here there is a discussion of whether indicators for O3 and PM have gotten better from 2013 to 2014. It would be interesting to see what the trend looks like if you start with an earlier year.

Following this suggestion on section 3.4 and remarks from Referee#2 on the score significance, and also in order to make a more comprehensive analysis of the Ensemble performances, we show and discuss in the revised manuscript the statistical indicators not only of the Ensemble median but also of all 7 individual models. Following your suggestion, we were also aiming at extending the analysis over a longer time period (from 2011 to 2014). For this, we gave a close look at the time series of observations used to calculate the statistics. This work showed that there is a high variability in time and space of the surface station data available. This means that the statistical indicators calculated for a particular season and year are based on a set of observations that is significantly different from those used for other seasons and years. Therefore, the changes in the scores between years for a chosen season come partly from the set of observations used. This does not allow us to make a fair interpretation of these differences. This is why we have decided to only show and discuss the scores of the MACC-II forecasts in 2014, which illustrate the state of the multi-model ensemble performance at the end of MACC-II project.

In the revised manuscript, figures 7 and 8 show only scores for 2014 (the left column has been removed) but include all seven models in addition to the ENSEMBLE.

Diurnal patterns in statistical indicators. It is striking in Figures 3, 5, 6, 7 and 8 that there are diurnal patters associated with the forecast bias and correlation. This is an interesting feature that is not really explored. Why are these diurnal patterns seen? Is it an issue with daytime vs. nighttime boundary layer height? Or something else? I realize the authors might not have a complete answer for this, but it deserves more investigation than it is given here.

We agree that there was not enough analysis of the Ensemble statistical indicators in the manuscript, and in particular of the diurnal cycle feature. A more comprehensive analysis has been done in the revised manuscript focusing on the seasonal scores since they are more representative than the case study. To do this, we have plotted in Figures 7 and 8 (GMDD numbering), in addition to the Ensemble, the results of the 7 models and we have included a discussion on the possible reasons of the diurnal patterns in statistical indicators. We also add results (one figure and corresponding comments) in order to show the robustness of the median ensemble method for ozone with regard to the number of models available over the three months of summer 2014.

About the diurnal cycle:
In the revised manuscript, figures 7 and 8 (now named figure 6 and 8) show scores for 2014 for the Ensemble but also for the 7 models. We have included a detailed analysis of these figures. Concerning the diurnal cycle shape of the ensemble both for ozone in summer and for PM10 in winter, it is consistent with the diurnal variations of most individual models.
For ozone, MNMB, FGE and R show best performances peaking at 15UTC and worst peaking at 06UTC for each of the 4 days of the forecast. This means that all models are able to simulate the ozone daytime photochemistry with the given setup of MACC-II (IFS forecasts for meteorology, C-IFS for chemical boundary conditions and GFAS and TNO emissions). For all models, the diurnal cycle in the statistical indicators can be at least partly explained by uncertainties in the diurnal cycle of the emissions of ozone precursors used in the individual models. This is illustrated by CHIMERE correlation at night which is better than most of the other models. CHIMERE has developed diurnal factors for traffic emissions based on an objective analysis of NO_2 measurements in the different countries in Europe which improves ozone titration at night (Menut et al., 2012). Other reasons of

the diurnal cycle in the model scores could also be errors in the diurnal cycle of the boundary layer height and associated vertical diffusion. For instance, the boundary layer in the LOTOS-EUROS simulations is described with a single model level, with a diurnal variation in the boundary layer height obtained 3-hourly from the ECMWF forecasts. This differs from the description of vertical mixing in the other models and may be responsible for the low correlation feature at around 9 UTC. MATCH shows the largest diurnal variability that can be partly related to a combination of chemistry, deposition and the vertical resolution, where the latter is inherited from the IFS model with a rather shallow lowest model layer (~20m). The ozone depletion processes at the surface appears too strong and not enough compensated by the vertical diffusion. The MB is then more pronounced during night time, and a modification of the vertical diffusion has shown to improve MATCH skill.
For PM10, MB, MNMB, RMSE and FGE are best during daytime (generally around 06-07UTC and 15UTC) with diurnal variations fairly similar for all models. This is related to the fact that PM_{10} are dominated by primary anthropogenic emissions of black and organic carbon which are prescribed in all model by the same TNO inventories and which have maxima in the morning and in the afternoon. Worst MB, MNMB, RMSE and FGE are at night, as for ozone. This may be linked to uncertainties in the boundary layer height at night, in vertical diffusion and/or to an underestimation of emissions.

About the robustness of the ENSEMBLE with regard to the number of models available:
For this, we performed tests in which we remove one or more models randomly on each of the daily forecasts. The new figure 7 shows the statistical results against observations for the ENSEMBLE (7 models) and the other ensemble medians calculated by removing randomly 1, 2, 3 or 4 models. For MB, MNMB and FGE, there is hardly any difference between all ensembles. Only RMSE and R (correlation) give significant changes. As expected, decreasing the number of models used in the ensemble tends to degrade its performances. Using 6 models gives RMSE and R close to the full ensemble based on 7 models. The scores for ensembles with 4 and 5 models are close to each other but are degraded compared to when 7 or 6 models are used. When only 3 models are used, RMSE and R are worse compared to the other configurations by ~0.5 $\mu g/m^3$ and ~0.05, respectively. This shows that, the multi-model ENSEMBLE at the end of MACC-II, which is based on the median of 7 models, is robust even if 2 to 3 models are unavailable. These results are consistent with the results discussed in Section 3.3 that were calculated on one week and with a different method for the model removal.

Figures 7 and 8. Why not have the same y-axis scale for both columns, so the reader can easily compare values year to year?

We agree that the y-axis should have been the same for both columns. Since results for 2013 have been removed based on the argument given above, there is no more need to change the y-axis.

Figures 7 and 8 have been modified in the revised manuscript by removing the left column. Only the results for 2014 are shown.

Figure 9. To complement Figure 9, it would be interesting to see time series of predicted (ENSEMBLE AND AEMET) and observed ozone for a selection of stations. I think such a visualization would provide a better feeling for the differences between model predictions and observations.

We agree with this suggestion.

We have added in the revised manuscript a comparison between the two models and measurements at 3 EMEP stations for summer 2013. We have also discussed the reasons for the differences between models and observations for each station.

Technical corrections

All technical corrections have been taken into account in the revised manuscript.