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 In particular, to show more clearly the agreement between forecast and observations, the 4 panels in Fig. 4 have been replaced by only one showing a zoom over the areas of interest for the 10th of June 2014 at 15UTC when the ozone episodes are both peaking.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.

Responses to referre#2

The authors thank Referee#2 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.

General comments

This manuscript presents the status of the ensemble air quality forecast system for Europe at the end of the Monitoring Atmospheric Composition and Climate: Interim Implementation (MACC-II) project in the summer of 2014. The ensemble forecast system consists of seven regional air quality models. The median of the ensemble member values is evaluated against observations from a selected set of stations using a number of statistics. The selection of stations is unfortunate and should be revisited – it appears to be eliminating most of the urban stations, which are precisely those where air quality issues are often seen.

The general objective of MACC-II regional products is not to provide air quality forecasts and analyses for local situations but at the pan-European scale. For this purpose, the horizontal resolution chosen for the 7 individual models is 10-20 km, in order to represent large scale phenomena and the background air pollution. Note that one of the main applications of these products is its use as boundary conditions for high resolution air quality models run on selected regions in Europe in order to forecast local air quality. This information was in the manuscript in Section 2.1 first paragraph but should have been made clearer earlier in the text.

Concerning the selection of observations, for a fair statistical evaluation of the MACC-II products, the authors only want to use the surface monitoring stations that are representative of the 10-20 km model resolution. For this, the authors use the classification proposed by Joly and Peuch (2012) to exclude the stations having a concentration variability that is typical of locations mainly influenced by local phenomena. More explanations are given below in responses to other comments related to this issue.

In the revised manuscript, the authors have now stated clearly in the introduction and in the conclusion that the MACC-II regional products are not intended to be used for local situations but for regional events and for background pollution. There are also additional details in Section 2.1 on the selection of stations used for the verification.

The performance of ensemble median, which is said to perform better than any of the individual models, should be examined further: how does it perform for high pollutant values, does it capture the peak concentrations as well as an individual model? The robustness of the ensemble median should be examined for the cases when one or two of the best performing models are withheld from the ensemble.

Both Referee#1 and you commented on the fact that in the paper, in its GMDD version, there is not a comprehensive analysis of the performances of the ensemble versus individual models. Therefore, the authors have added material in the revised version 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) with comments on the ability of the 7 models and the ensemble median to capture peak concentrations,
- 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). The authors preferred making these tests on a long time period rather than on the one week of the case study because this gives more robust information.

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 (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.
More details are given below in the response to the technical comments.

More specific technical comments are listed below, followed by selected edits and corrections as portions of the manuscript require English proofreading/editing.

See below the responses to the technical comments.

All the English edits and corrections have been taken into account in the revised manuscript.

Technical comments

p. 2743, l. 7: "In PMs, there is no distinction between primary (dust, sea salts, black carbon and organic carbon) and secondary aerosols formed from gaseous precursors such as SO2, DMS, H2S, NH3, NOx and VOCs." This should be reworded. There is a distinction – it is ignored when composition is not taken into account, such as when considering mass or number concentration only.

The authors agree with this comment.

The manuscript has been revised accordingly.

p. 2744, l. 4: ". . .the approach based on a multi-model ensemble of forecasts has been developed to improve their quality through statistical approaches." This should be reworded. The ensemble provides better information by combining information from different models - it does not directly improve the quality of forecasts from an individual model (although careful routine evaluation/comparisons may guide model improvement process over time).

The authors agree with this comment.

The manuscript has been revised accordingly.

p. 2744, l. 27: Introduce individual models, provide references.

To avoid making the introduction very long and since the individual models are described in detail in Section 2, we only added the main references for each model in the introduction of the revised manuscript.

p. 2745, l. 1: "Ensemble approach provides on average forecasts and analyses of better quality than any of the individual models" - Add a reference to support this statement. Demonstration in this manuscript is limited to ozone over only one week comparing the "best" individual member vs ensemble median.

The authors agree with this comment which was also raised by referee#1. To better show the performances of the ensemble approach the authors have added in figs. 7 and 8 (GMDD numbering) the scores of the 7 individual models. The authors made a thorough analysis of the ensemble based on the performances of the 7 models for ozone in summer 2014 and for PM10 in winter 2013-2014.

Additional material has been included in the revised manuscript to show that the Ensemble approach provides on average forecasts of better quality than the individual models for ozone. For PM10, results show that the ensemble median provides the best scores on some of the statistical indicators but not for all of them, as discussed now in section 3.4. The authors have changed this sentence to "Although each of these models can perform very well on particular days in particular areas, the ensemble approach aims at providing on average forecasts and analyses of better quality than any of the individual". The authors have added in Section 3.4 material and discussion on the performances of the ensemble median for ozone and PM10.

p. 2746, l. 2: Do all models have only those 8 (4 earlier) vertical levels, or are these levels the only ones used in the ensemble?

Each of the 7 models is run with its own vertical grid. The chemical species concentrations from each model are then regridded onto the 8 levels. These levels are used to calculate the ensemble. Numerical data and/or plots for the individual models and the ENSEMBLE are only provided to users on these 8 vertical levels.

The manuscript has been revised to make this clearer (Section 2.1).

p. 2746, l. 18: "All 7 models do not produce yet the analysis for all the 6 core species." - How are missing levels/species handled in ensemble construction?

This sentence was misleading because no further information was given there. It was removed in the revised version. The full explanation is given in section 2.9.

The corresponding text in section 2.9 has been revised to make it clearer: "For the analyses, the individual assimilation systems provide only analyses at the surface level and do not produce analyses for all species yet. At the end of MACC-II, ozone was the only species that is produced by 6 of the models. For other species, analyses from less than 5 models were available. This is why the ensemble analysis in MACC-II is only calculated for ozone. It has been extended to NO_2 in 2015 since more models will produce NO_2 analyses."

p. 2747, l. 8: “cut-off time at 07:00UTC on Day0 for the dataset covering Day0–1. At this time of the day, more than 90% (on average) of all data are available.” Is 90% of data for Day 0 really available by 7 UTC on Day 0, or should that be 7 UTC on Day 1? Define Day0-1, Day0-2.

The definition of Day0 was not given. For both the forecast and the analysis this is the day when the forecast and the analysis are run. Day0-1 is the day before Day0. For the analysis this means that the authors get the data for Day0-1 only in the morning (7UTC) of Day0. This is now explained in the revised version.

The authors have added in section 2.1:

- about the forecast: “Day0 is defined as the day when the forecast is run. The forecast initial time/date is Day0 at 00UTC and final time/date is Day3 at 24UTC. “
- about the analysis : “Like in the forecasts, Day0 is defined as the day when the analysis is run. Day0-1 refers to the day before Day0. The analysis initial time/date is Day0-1 at 00UTC and final time/date is Day0-1 at 23UTC.”

p. 2747, l. 17: quantify “almost complete”

‘Almost complete’ suggests that the authors can compare to a reference which would be the full set of data. The set of stations vary from one day to another because of the many possible issues before being delivered to the EEA database. Therefore it is not possible to define what would be “the full set of data”. More interestingly, the authors can say from the monitoring of the EEA database over several months that there is about 10% more data available at 23UTC than at 07 UTC on average. As illustrated in a report produced during MACC-II project (D16_3; http://www.gmes-atmosphere.eu/documents/maccii/deliverables/obs/), the additional data collected at 23UTC compared to 07UTC are mainly data from the end of the previous day. This is because there is a significant number of stations that do not send their late afternoon and evening Day0-1 data before 07UTC on Day0. This means that the 23UTC dataset used for verification is homogeneous with approximately the same number of observations in the morning, afternoon and evening.

The authors have removed this part of the sentence and provided the argumentation given above in the revised manuscript.

p. 2747, l. 18-20: Provide specifics on data selection/quality control procedures – what is blacklisted, what thresholds are used, how is representativeness determined?
p. 2747, l. 22: Is anything in data selection different from Joly and Peuch (2012)?
p. 2748, l.1: Does station selection essentially eliminate all the urban sites? Are you verifying over rural sites only? This seems to ignore model performance in urban areas, where most serious AQ issues are often seen and weakens all of the subsequent results.

All three comments above refer to the same overall subject. The authors give below only one response which addresses all three comments/questions together.

The selection/quality control of the stations is done in two steps:

1. *a blacklist which includes stations indentified as unrealistic, such as for instance stations giving the same concentration for each hour of the day. This blacklist is monitored manually and updated as often as needed. An automated sorting based on fixed thresholds for each species is currently tested.*

2. *Use the classification from Joly and Peuch (2012) to only select the stations that are representative of the resolution of the MACC models (10-20 km). More details are given in the responses below.*

The classification of Joly and Peuch (2012) attributes to each station a class number from 1 (most rural characteristics, i.e. corresponding to large scale/background air quality) to 10 (most urban characteristics, i.e. corresponding to very local scale air quality such as traffic stations) from a Linear Discriminant Analysis. This objective classification of Joly and Peuch (2012) was based on long series of validated data over Europe spanning from 2002 and 2009. It has been updated in MACC-II using version 7 of Airbase data spanning from 2002 to 2011.

As stated above, the objective is not to provide air quality forecasts and analyses for local situations but at the pan-European scale with a resolution of ~10-20 km. This is why the authors aim at selecting the surface monitoring stations that are representative of a 10-20 km resolution for the statistical evaluation of the MACC-II products. This could be done on the basis of the metadata of the stations by excluding traffic and urban stations. But there is currently no uniform and reliable metadata on site representativeness available for all regions and countries of Europe. This is why the authors chose to use the classification proposed by Joly and Peuch (2012) that sorts all the European stations in 10 classes from an objective statistical analysis. Once each station is classified the authors exclude the stations having a concentration variability that is typical of locations mainly influenced by local phenomena. This is done by only keeping stations having a class number ranging from 1 to 5 for all pollutants. The threshold of 5 allows us to remove the stations influenced by local phenomena while keeping a reasonable number of stations for calculating statistical indicators. There is ongoing work to improve the station selection, still on the basis of the Joly and Peuch's classification, by determining the best thresholds to be used individually for each of the 6 pollutants (O_3, NO_2, SO_2, CO, PM10, PM2.5), in particular based on pollutant lifetime.

These explanations have been included in the revised manuscript in section 2.1, and in the conclusion for the part concerning on-going work.

p. 2748, l.4: How are these 20% of observations that are withheld from the assimilation selected?

For each pollutant, a list of stations is kept aside for the verification of analyses. This list is the same everyday and it has been determined so that the stations are well spread inside the domain. The observations at these stations are not assimilated. A posteriori statistics over several months show that the ratio of observations that are kept aside for the verification of analyses is roughly 20% of the total amount of observations that are downloaded at 23UTC.

This information has been included in the revised version (Section 2.1).

p. 2748, l.7: This needs to be rewritten as the list provided is not complete. Uncertainties in observations and assimilation methods impact analysis uncertainty. Initial condition uncertainties impact forecast and analysis uncertainty.

The authors agree that the list of uncertainties was not complete.

The manuscript has been modified in section 2.1 following your comment. "Major sources of uncertainties in regional AQ forecast and analyses are the quality of the emissions used, the meteorological forcings, the representation of the atmospheric physical and chemical processes, the initial and boundary conditions for the chemical species and the uncertainties in observations and assimilation methods impacting the analysis."

p. 2748, l.10: The use of identical emissions and meteorology in all models minimizes the ensemble spread, not necessarily true uncertainty.

This is true that using identical emissions and meteorology minimizes the ensemble spread but not necessarily true uncertainty. Nevertheless, using the best possible emission inventories and best meteorological forecasts the authors expect to provide good quality AQ forecasts.

In the sentence the authors have removed in the revised manuscript "to minimize uncertainties".

P. 2757, l. 4: Define SIA.

SIA stands for secondary inorganic aerosols.

SIA has been defined in the revised manuscript.

p. 2762, Figs 2 and 3: While ensemble median might be capturing day-to-day trends, it clearly underestimates observed elevated ozone values during daytime, so why is this statistic used to represent ensemble predictions?

This is true that the ensemble median captures well the day-to-day trends and that it often tends to underestimate ozone peaks but this is not always the case. For example, the ensemble median at Sausset on Friday the 13th is higher than observations. This is also illustrated on the maps of the ENSEMBLE with overplotted observations. Note also that for the comparison the EPSgrams are only plotted every 3 hours while measurements are every one hour. Very transient peaks on the observations can be missed on the 3-hourly EPSgrams. To make a more detailed analysis the authors have added two stations, one in Germany and one in France.

Figures for the two additional stations are included in new figure 3 and comments on the skills of the ENSEMBLE to predict daytime peaks are given. In particular, there is an analysis of the reasons for the underestimation of the ENSEMBLE compared to observations at Sausset. This is linked to the particular location of the station combined with particular meteorological conditions.

p. 2763, l. 24-25, Figure 4: It is not clear from Fig. 4 that the median captures the two ozone episodes well. Suggest replacing by a single map with enough resolution to be able to discern the level of agreement between model and observations. In the central European area it is difficult to see the model values among the observations – suggest zooming into that area. At the same time, there are no observations shown in the area of high ozone in Italy.

The authors agree that the figures were not clear enough. The 4 panels in Figure 4 have been replaced by a figure zooming on the areas of interest for the 10th of June 2014 at 15 UTC, when the ozone episodes are both high. Doing this the authors found a small error in the plotting procedure. A few stations were missing in Fig. 4a. This has been fixed in the revised version.

Concerning the measurements in Italy, there was unfortunately no hourly data available during the studied period over Italy in the European Environment Agency NRT database. This is why there is no possibility to evaluate the ensemble ozone fields against observations in this area.

Figure 4 has been modified in the revised manuscript. The fact that no observation in Italy was available in the EEA NRT database during the period has been clearly stated in the revised manuscript (new section 3.3). Following one of referee#1's comment, the authors have merged figures 2 and 3 for any easier comparison between the observations and the EPSgrams. Therefore all the figure numbering has been changed.

P. 2765, l. 12-23: Why are "best" and "worst" models removed simultaneously? This may be misleading as an illustration of the robustness. How robust are the results to loss of one or two of the best models?

By removing at the same time the best (or two bests) and the worst (or two worst) models the authors estimate an "average situation". The authors agree that it does not give the full picture of the impact of the number of models on the performances of the median ensemble. This is why the authors have added results and discussion of new tests performed on the summer 2014 period in which the authors have removed randomly every day 1, 2, 3 or 4 models. The choice of a longer period (3 months) than the case study period ensures the robustness of the tests. The choice of a random sorting is based on the fact that the model failures are actually random. Removing one or two best models requires defining what is the best model. Over a long period is difficult to define the best model since each model has its "good days" and can be best for a specific time and location. The results of the tests show that the median ensemble with 7 or 6 models give close scores, the median ensemble with 5 models gives scores close to 4 but with lower performances than 7 or 6, and the median ensemble with 3 models gives worse scores than the other configurations.

The sentence "By removing at the same time the best (or two bests) and the worst (or two worst) models the authors estimate an average situation." has been added in section 3.3 and the new tests on the 3 month period are shown in new figure 7 and discussed in section 3.4.

P. 2765, l. 24-25: "Note that we only illustrate here the behaviour of the ENSEMBLE over a short period of time, but these results are still true over longer time periods." Please support this statement by providing quantitative statistics for a longer period of time.

The authors agree that the paper was not showing clearly enough the overall behaviour of the Ensemble. This is why the authors have added in figures 7 and 8 which (seasonal scores) the seven models in addition to the Ensemble and the authors have analysed in detail the results. This also follows recommendation by referee 1 of a more in-depth analysis of the Ensemble performances.

In the revised manuscript, this sentence has been changed to: "In this section, we illustrate the performances of the MACC-II AQ forecasts for a case study of ozone pollution events that took place between the 10th and the 13th June 2014. A more in-depth analysis of the individual model and of the ENSEMBLE performances is done over longer time periods in Section 3.4"

P. 2767, l. 23: Correlations of 0.7 at best, and often lower than 0.5 should not be called "high correlations.

The authors agree that the correlations cannot be qualified as high correlations. Nevertheless, they are good if one considers that the authors use only raw forecasts of the individual models to build the Ensemble. There is no use of bias correction or a posteriori statistical adaptation as sometimes used in operational applications. For PM10, bias correction is one of the possible way considered for improving the models and therefore the Ensemble.

The sentence has been modified by changing “high correlations” by “good”. The point made about bias corrections is also included in the conclusion.

P. 2767, l. 23: “The changes in scores between 2013 and 2014 are not large enough to be regarded as significant. . .” Please demonstrate the lack of statistical significance to support this statement p. 2768, l. 1: “show a specific behaviour that is not analysed since not significant” Please demonstrate the lack of statistical significance to support this statement.

Following these two questions on the score significance and a comment of Referee#2 on adding performances for previous years, the authors have analysed the statistical indicators not only of the Ensemble but also of all 7 individual models and over a longer time period (from 2011 to 2014). For this, the authors gave a closer look at the time series of observations used to calculate the statistics. This work showed that there is significant 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 specific set of observations. Therefore, the differences between different years come at least partly from the observations used. This does not allow us to make a fair interpretation of these differences. This is why the authors have decided to only show the scores of the MACC-II forecasts in 2014, which illustrates the state of the ensemble performance at the end of MACC-II project. The authors have made a thorough analysis of these scores on the basis of the scores of the 7 models that are now shown in the revised manuscript.

In the revised manuscript, figures 7 and 8 (GMDD version) have been changed. They now only show scores for 2014 (the left column has been removed). These scores are now not given only for the Ensemble but also for the 7 models. The authors have included a detailed analysis of these figures.

Figure 7: For each statistic, use the same vertical range for both years to allow comparison between the years.

The authors 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.

p. 2768, l. 2-4: “The MB and MNMB both indicate a low bias in the ENSEMBLE. This can be linked to the fact that not all individual models include secondary inorganic aerosols and/or secondary organic aerosols.” – Demonstrate that this is the cause of low PM10 bias by stratifying the statistics by separating the models based on inclusion/ exclusion of secondary aerosols.

The authors agree that there was not enough explanation on this point. This has been done on the basis of the new figure 8 which shows the 7 individual models in addition to the ensemble median. The authors give an interpretation of MB and MNMB for each model on the basis of which types of *aerosols are simulated in the different models and how well their associated processes are taken into account.*

Selected corrections/edits

All selected corrections/edits have been taken into account in the revised manuscript.