Replies to reviewer 2

My only major concern is the lack of quantitative conclusions. The authors state in their response that “More care has been given to the abstract an to adding more quantitative conclusions”. However, in comparing the original to the revised version, the abstract does not appear to have any additional quantitative conclusions. Instead, all of the additional material is qualitative (“slight” increase of diversity, “marked” improvement of accuracy). I again ask that the author quantify their conclusions, using the data they already have to hand. This recommendation also applies to the conclusions section (section 5) of the paper, which I note present only one quantitative result, a “1-5% improvement”. I also recommend that the authors be more specific about what has improved, as here the actual quantity which has improved is not clearly stated.

We thank the reviewer for the detailed suggestions. We have updated the manuscript with quantitative information, introduced a new table (table 3) and added new figures (2S, 4S) in the supplementary file. The comparisons refer to the optimal ensembles build from subsets (mmeS).

[abstract]“**Abstract**
In this study we introduce a *hybrid ensem*ble consisting of air quality  models operating at both the global and regional scale. The work is motivated by the fact that these different types of models treat specific portions of the atmospheric spectrum with different levels of detail and it is hypothesized that their combination can generate an ensemble that performs better than mono-scale ensembles. A detailed analysis of the hybrid ensemble is carried out in the attempt to investigate this hypothesis and determine the real benefit it produces compared to ensembles constructed from only global scale or only regional scale models. The study utilizes 13 regional and 7 global models participating in the HTAP2/AQMEII3 activity and focuses on surface ozone concentrations over Europe for the year 2010. Observations from 405 monitoring rural stations are used for the evaluation of the ensemble performance. The analysis first compares the modelled and measured power spectra of all models and then assesses the properties of the mono-scale ensembles, particularly their level of redundancy, in order to inform the process of constructing the hybrid ensemble. This study has been conducted in the attempt to identify that the improvements obtained by the hybrid ensemble relative to the mono-scale ensembles can be attributed to its hybrid nature. The improvements are visible in a slight increase of the diversity (4% for the hourly timeseries, 10% for the daily maximum timeseries) and a smaller improvement of the accuracy compared to diversity. RMSE improved by 13-16% compared to G and by 2-3% compared to R. POD/FAR show a remarkable improvement, with a steep increase in the largest POD values, though comparable to the other for the hybrid ensemble and comparatively smallest values of FAR across the concentration ranges. The results show that the optimal set is constructed from an equal number of global and regional models at only 15% of the stations. This implies that for the majority of the cases the regional scale set of models governs the ensemble. However given high degree of redundancy that characterises the regional scale models, no further improvement could be expected in the ensemble performance by adding yet more regional models to it. Therefore the improvement obtained with the hybrid set can confidently be attributed to the different nature of the global models. The study strongly reaffirms the importance of an in-depth inspection of any ensemble of opportunity in order to extract the maximum amount of information and to have full control over the data used in the construction of the ensemble.“

[L927 section 4.3]: “To answer the question whether the multi-scale ensemble is more skillful, we consider the two optimal single scale ensembles of rank 6, namely the global (mmeS-G) and the regional (mmeS-R), and the optimal multi-scale ensemble of rank 6 (mmeS-GR) that is constructed from elements of the optimal single scale ensembles. The multi-scale ensemble achieves an improved diversity by at least 4% compared to the single-scale ensembles, reaching even 10% for the daily maximum timeseries (Table 3). It reflects the independent development of global and regional models. The change in accuracy is generally smaller since the optimal single-scale pools contains models with not very different errors. When the two pools are combined, the mmeS-GR achieves a better RMSE by 13-16% compared to mmeS-G and by 2-3% compared to mmeS-R. Further, the mean of the distributions of diversity, accuracy and RMSE from mmeS-GR differ from the corresponding mean of mmeS-G and mmeS-R (they passed the t-test at the 5% significance level). The same holds for the distributions (they passed the Kolmogorov-Smirnov test). Improvements are also revealed for the POD and the FAR, where the mmeS-R does better than mmeS-G, especially at high thresholds. The mmeS-GR generally improves the indices compared to mmeS-R, even though global models are included. Like before, the improvements are seen in all datasets, despite their temporal aggregation.”

[L980 conclusions] “In terms of quantitative conclusions, comparing the **optimal** multi scale (GR) ensemble with the **optimal** single scale (G and R) ensembles :

* Diversity improved at least by 4% for the hourly timeseries, becoming 10% for the daily maximum timeseries
* Accuracy generally improved less than diversity
* RMSE improved by 13-16% compared to G and by 2-3% compared to R

POD/FAR show a remarkable improvement, with a steep increase in the largest POD values, though comparable to the other for the hybrid ensemble and comparatively smallest values of FAR across the concentration ranges.”

All remaining concerns are minor, and listed below:

1) In my previous review, I stated:

“In the conclusions the authors refer to analysis of annual hourly, J-J-A hourly, and annual daily maximum records. However, until that point there seems to be no discussion of the latter two metrics. The authors should consider elaborating in the previous sections on the analysis they performed using these metrics.”

Based on the response given by the authors, I fear that I was not clear. My meaning is that, whenever data on “accuracy” or any other outcome is presented, the exact quantity being considered is often not clearly stated. For example, I note that section 4.3 (previously section 7) states that “[t]he comparison of the ensemble performances will be restricted to the months of June -August”. However, the authors do not state whether they are considering the average of the daily maximum over this period, or the average of all hourly readings over this period. While it may be clear to the authors, I strongly recommend that they clearly state the exact quantity being analyzed at the opening of each figure’s discussion. For example, in the discussion of POD and FAR on lines 334 to 349, are they referring to exceedances in all hourly data, or in the daily maximum?

We have updated the text:

[L559]: “In almost all subsequent results, the measured time series should be interpreted as ensemble averages of all available rural monitoring stations with 1h temporal resolution. The analysis was not performed with spatially aggregated timeseries only in Figures 7, 9 and 11 while a subset of the annual hourly timeseries was used in Figure 8 (June-August).”

2) Further to the previous comment, I cannot find any dedicated analysis in the manuscript of the daily maximum records. The closest I can find is in section 3.2, but this is prior to any discussion of ensemble results, and only addresses the maximum as a limiting case for model diversity. The ability of models to accurately predict the one-hour daily maximum is of significant policy relevance due to the fact that it is a monitored quantity under some regulations, and the fact that it was used as the predictor in a widely-cited epidemiological study which underpins much of the literature on ozone exposure impacts (Jerrett et al 2009). The ability of models to achieve a certain level of accuracy when predicting the daily mean is not shown here to be a reliable predictor of their ability to predict the daily maximum, and there are reasonable grounds to expect this not to be true (e.g. a model may represent night-time ozone chemistry well but have a poor representation of daytime behavior). I therefore suggest that the authors either:

• Clearly demonstrate which of their analyses are specifically relevant for the daily-maximum record as opposed to including all hourly records, if indeed some of the analyses are specific to the daily maximum; or
• Provide analysis in the results section of the ensemble performance specifically for daily maximum ozone, prior to the assertions on lines 629-630; or
• Remove the assertion on line 630 that the annual daily maximum records are analyzed.

As stated in the paper the models have been extensively evaluated in other publications as part of the HTAP AQMEII activities. It is also stated that not only the same models but the same sets of data evaluated are those used here. We are after an analysis of model results in ensemble form for the whole range of data not only maximum concentrations this is why we have ALSO analysed the maxima.

Results for the maximum values are provided in the supplement (Figure 4S) and in the new table 3. We have also updated the text with the appropriate links:

[L883] “… annual daily maximum timeseries (figure 4S) …”

[L927-957] New paragraph introduced around Table 3.

3) Figure 11 is critical to the paper’s conclusions, but remains somewhat difficult to interpret due to unnecessary clutter and lack of appropriate labeling. I would suggest that:

• The four colar bars are replaced by one;
• Units are added for the x- and y-axis labels;
• A smaller marker is used to prevent (or at least minimize) the degree to which markers block one another;
• Units are added for the in-figure notations.

To be clear, I recognize that Figure 11 is complex and is attempting to compactly represent a lot of data. My hope is that these suggestions will help the reader to more easily interpret an important but complicated diagram. Units should also be used during the discussion of Figure 11 (lines 577-607).

We thank the reviewer for the thorough suggestions on how to change fig 11. We have updated the figure according to his last two suggestions. Units cannot be added since the variables are dimensionless.

4) On line 279, the authors state that the spectrum behaves “remarkably similarly at scales smaller than the daily peak”. I was surprised at the lack of analysis regarding the mismatch between all models and the observations in the high frequency components, with seemingly all models showing strong resonant peaks while the observations show no such effect. Do the authors have a hypothesis for this? Is this because of shortcomings in the observations or model error?

The resonant peaks are present also for the measurement just with a lower amplitude as it is visible in the figure

5) Lines 196 to 205 have out-of-date section numbers. This paragraph should be re-written to better reflect the new paper structure.

Corrected thanks

6) Line 337: “whereas the latter is the proportion of non-occurrences that were incorrectly identified as happening” – this sentence is confusing and I suggest the authors reword it.

Corrected thanks

7) Lines 445-451 seem to be referring to Figure 11, but at this point only Figures 1-6 have been introduced. The authors should either defer this discussion to the relevant part of the manuscript or at least make clear that they are talking about a different figure.

It is referring to the accuracy diversity equation not necessarily to Figure 11, in fact not at all

8) On line 496, the terms LT, SY, and DU are used, but these are no longer defined explicitly in the manuscript.

Definition added

9) On line 499-502, would it be accurate to say that a larger number of global models is needed in locations with complex topography because we are using unrelated differences between models to substitute for an accurate representation of the effect of topography on the result? Or do some of the global models have subgrid parameterizations capable of capturing this effect?

If models are deficient in representing a process, even an infinite number of them will not compensate the deficiency. This if we understand what the reviewer is implying. Models treat topographies according to the horizontal grid spacing.

10) On lines 569-571, the authors assert that, where a small number of global models are used in the optimum ensemble for a given location, this “clearly indicat[es] a deficiency”. Doesn’t this depend on both the accuracy and diversity of that small number of global models (eg high accuracy but low diversity of the globals + low accuracy but high diversity of the regionals could produce this result)?

We have included additional material in the supplement and updated the text accordingly:

[L879] The second panel of Figure 9 also gives the spatial distribution of the number of global models contributing to the hybrid ensemble clearly indicating a preference of regional models in the northeastern part of the domain. This “spatial” preference is not observed in the jja hourly timeseries or the annual daily maximum timeseries (figure 4S), both being high-ozone datasets. This is in line with the relatively higher RMSE of the global models at low concentrations (figure 2S).

11) Figure 10 has only three lines of discussion. I recommend that it be moved to the SI, as the outcomes could easily be presented without it.

We do not agree we’d rather keep it, it is further food for thought for our readers that we consider relevant.

12) There are still some typos in the paper. I have listed some below but recommend that the authors perform another sweep:
• Line 114: “of which where global” should be “of which were global”
• Line 254: “selected as the capture” should be “selected as these capture” (or possibly “they”)
• Lines 264, 396, 573: “Figures” should be “Figure”
• Line 408: This should be Figure 5, not Figure 6
• Line 570: “Global” should be “global”
• Line 599: “+” should be “and”
• Line 600: “form” should be “from”

Corrected thanks

13) The assertion on line 613 to 615 is that “regardless of the treatment, the ensemble data capture the ozone power spectrum with no notable deviation from the measured spectrum”. This is far too strong and not supported by the data. Whether because of shortcomings in the observational technique or in the models (see comment 4 above), there is significant disagreement at frequencies below 6 hours, and this deserves explanation. This statement also seems to be almost immediately contradicted by the observation of the “large power deficit in the range from 0.8 days to 100 days”.

Corrected thanks

14) Line 924: what is meant by “most present”?

Corrected thanks

15) Table 2 still seems quite inconsistent, especially in their vertical grid descriptions. For example, the CHASER models report “32 VL up to 40 km” – no detail on spacing – while the IFS model reports “60 VL from surface up to 0.1 hPa – lowest level 15 m”. Meanwhile the GEOS-Chem adjoint model still reports “47 levels up to 0.066 hPa (bottom of last grid)”. While I appreciate that the authors did follow up on my previous question, their explanation was not convincing; GEOS-Chem does not generate values “at the base of the cell” but instead almost always calculates cell-average quantities. Meanwhile the other models report the correct upper limit, i.e. the pressure at the “ceiling” of the uppermost grid cell – this includes for example the IFS model. In the interests of consistency, I recommend that the authors replace the “vertical grid” column with “Number of layers”, “Lowest layer thickness”, and “Model top (hPa)” columns, and that they either use the actual model top for GEOS-Chem or that they change the description of the other models to also use the bottom of the uppermost layer (e.g. 0.02 hPa for IFS, 136.5 hPa for the EMEP model, and so on).

This information is sufficient for the scope of the paper; more detail on the model numerics can be found in the references present in the table,. Adding the information requested by the reviewer is not bringing any additionally relevant element instrumental for the analysis. Those interested in digging more in the details are referred to the published material. As for the non convincing information we advice the reviewer to contact the original source of the information that is GEOSCHEM group referenced. The question of the reviewer has been passed directly to the modeller in charge of the system whose reply has been reported verbatim. As a matter of fact the reviewer himself says “ almost always” therefore not always. We trust the modellers with whom we work who finally bare the responsibility for the info they provide us.