"*Evaluation and Error Apportionment of an Ensemble of Atmospheric Chemistry Transport Modelling Systems: Multi-variable Temporal and Spatial Breakdown*", by Efisio Solazzo et al, ACP, 2016.
We are thankful to both reviewers for the positive comments and useful suggestions. We have revised the manuscript in many parts to take on board the suggestions and improve the exposure and discussion of the results. All the figures have been replaced to accommodate change of scale and grouping of models. A new table has been added, summarising the configuration of the WRF model adopted by the modelling groups. The evaluation of model results has been extended to include also the wind direction. More than twenty new references have been added to integrate the discussion of the results.

Response to reviewer #2.

Solazzo et al. (S2016) present an interesting model-observation analysis, making use of a range of model data from the third phase of the Air Quality Model Evaluation International Initiative (AQMEII), of primary and secondary air pollutants important for our understanding of the chemistry of the atmosphere and how well models simulate it. This sort of study is vital as more and more model evidence is used to link exposure to air pollution and impacts. The model and observed time series were broken down using spectral decomposition - breaking the time series into it's spectral components - and further analysis focused on separating the mean square error between the model and observations to better characterise the sources of error with the aim of attributing them to specific sources/processes. This technique was recently published by the same authors Solazzo and Galmirini, 2016) and the work here greatly extends previous analysis.

In general I think this paper should be published following several minor revisions. In my opinion, more emphasis needs to be spent on evaluating models at the process level and this paper raises a promising avenue for others. As the authors say, this method has not completely determined the causes of model error - but I think that modifications to and expansions of the method will lead to improved insight in the future.
Response. We thank the reviewer for the comment. We agree that the direction the evaluation of models should move is indeed towards processes, but that is not feasible at the moment as would require a new design of the AQMEII activity. A follow up paper focusing on two models only (CMAQ and Chimere) is in preparation for this same special issue dealing with more in depth diagnostic analysis, making use of extra model runs to determine the degree of impact on error of fluxes at the boundaries (deposition, emission and boundary conditions). We have highlighted in the conclusion section our thoughts about future directions of AQ model evaluation.

<u>Major comments</u>
My main criticism of this paper comes about in the presentation of the results. There are multiple figures at such poor resolution that I almost had to give up looking at them. Many of the graphics look like they are plotted in R and I would encourage the authors to save the graphics in a high resolution output (pdf, eps etc).
Response. The figures have been produced as 'tiff' with high resolution and imported onto a word processor. The conversion of the document to 'pdf' for submission produces figures hardly readable. At print out stage we will provide 'pdf' figures as separate files, but for the initial submission we could not find an alternative to converting the document to pdf, which reduced the quality of the figures drastically.

I also think that the reader would benefit from a consistent set of axes limits for plots in EU and NA and the ordering of the models should follow those used in Figures 23 and 26 (i.e. clustering the WRF-Chem variants together and the WRF-CMAQ variants together). When I did this "by hand" I found that the intra model variability was large -no doubt reflecting different model options (chemistry schemes etc).
Response. We have redone all figures and used same axis limits for both continents where possible (not for CO, for example as the EU bias is one order of magnitude higher than in NA). The models have been grouped following the reviewer's suggestion.

My other minor criticism is the choice of the domains. I just wonder if country specific (or State specific in the USA) boundaries were used could we learn more about emission estimate biases? I would imagine that the averaging over states and countries in the current classification will smear out these effects to some extent depending on their heterogeneity.
Response. The reviewer is right, that is indeed the case. That of countries borders is an issue we have dealt with in the previous analyses of AQMEII, especially in the Solazzo and Galmarini (2015) paper, where we showed that not only the results of the models for Europe tend to group by countries due to country-specific

emission profiles, but also the measurements showed very robust clustering properties by network (in North America) and, again, by country in Europe (Airbase network) due to lack of harmonised reporting and instrumental settings.
Nonetheless, in the context of a multi-pollutant and multi-scale screening analysis of regional scale models such as the one presented here, country-based investigation would have required much more space in an already lengthy manuscript. One of the aims of the cluster analysis described in section 2.3.2 is to identify sub-regions where the LT and SY components do present homogeneous features, allowing to overcome country-specific discussion. For these two components, thus, we do not expect heterogeneity induced by emissions. The effect of heterogeneity of emissions (for Europe) is therefore limited to the DU component. The following text has been added to the revised section 2.3.2:
'*As noted in the introduction, unsupervised hierarchical clustering was used to determine sub-regions where the LT and SY components showed similar characteristics – spatial averaging within these sub-regions was carried out due to the similarity of the observation data within these regions implying they will experience common physical and chemical characteristics. Errors due to the heterogeneity induced by country-specific emission profiles (in EU) are therefore included in the DU component.'.*

Minor comments
line 138: Delete the first "known". Done as suggested

line 162/Section 2.2 opening paragraph: There are other spectral filtering methods (e.g. Bowdalo et al., 2016 ACP). A comment on these would be useful and why the kz approach has been selected. Done as suggested

line 191: re-phrase the sentence starting "A clear-cut...". Done

line 197: add "do" after "they". Done as suggested

line 458: what is the source of the meteorological data that you compare against? As mentioned in the Acknowledgements: 'Data from meteorological station monitoring networks were provided by NOAA and Environment Canada (for the US and Canadian meteorological network data) and the National Center for Atmospheric Research (NCAR) data support section'.

line 647: It's not clear what you mean by timescale here? Do you mean the e-folding. Removed

line 1043 lifetime? Should that not depend on the concentration of NO? 20-30 days is an averaged value to provide an overall sense of the time-scale involved.
line 699: remove duplicate round bracket. Done as suggested

line 726: insert "is" after "bias". Done as suggested

line 728, 826, 823, ...: add space between mixing ration and units (e.g. 11.5ppb -> 11.5 ppb). This is a common error so please search the document for this. Done as suggested

line 766: insert space "Figure17and". Done as suggested

line 824: Do the authors really believe that vertical mixing can be analysed through something as complex as ozone? I would suggest that vertical mixing needs sources with no chemistry to be understood (e.g. Rn or Pb). Rephrased

lines 841: Is this really good? It was overlooked. It has been clarified now that CMAQ overestimates surface observations.

line 858: "sulphates" should be "sulfates". Done as suggested

line 881: What about oxidants? Could the corr($bias_{O3}$, $bias_{SO2}$) be useful? We could not find any consistent pattern relating $bias_{O3}$ with $bias_{SO2}$

line 928: What about temperature effects? Could corr($bias_{Temp}$, $bias_{PM}$) be useful? We have summarised the

information of corr($bias_{Temp}$, $bias_{PM10}$) in the revised text (section 3.3.6). We could not detect any consistent pattern relating $bias_{Temp}$ to $bias_{PM2.5}$. We have added the following text to the revised manuscript:
'The analysis of corr($bias_{Temp}$, $bias_{PM10})_{LT}$ shows that the error of these two variables are related, especially during the spring months and more consistently in EU3 (up to 0.74 for the WRF/Chem1 model) and during autumn in EU1 (the bias of Temp and the bias of PM_{10} are anti-correlated up to -0.67 for CMAQ1). Other models (e.g. the CAMx model), on the other hand, do not show any significant correlation.'