Replies to the comments to the manuscript '*Advanced error diagnostics of the CMAQ and Chimere modelling systems within the AQMEII3 model evaluation framework*' by Solazzo et al., 2017(doi:10.5194/acp-2017-257)

Reviewer 1

In this work authors pursue their efforts in devising more sophisticated methods for error analysis in air quality models. In the light of experience gained during the AQMEIIs phases, they analyze results for ozone simulation for North America (CMAQ) and Europe (CHIMERE). Given that a statistical analysis of model results can only be the first step toward a more-in-depth diagnosis of model deficiencies (it is difficult at this level to disentangle the impact of NOx/VOC chemistry, radiation, boundary layer dynamics and biogenic emissions, dry deposition, etc.), they report some interesting results which would be help in orienting modelers during model development or gaining confidence in using model predictions.

They suggest a combination of spectral (wavelet), time series (ACF, PACF and Kolmogorov-Zurberko filter) statistical tools and simple linear regression analysis to apportion model errors, applied to the decomposition of mean square error into its components (square bias, variance and covariance). In my opinion, an intelligent approach is to subdivide the AQMEII domains into sub-regions and highlights the differences in model performances for each domain. For example a striking feature is the CHIMERE bust for the Po valley (EU3), associated with the diurnal component, opposed to the better behaviour for EU1 and EU2 domains. Why CHIMERE performs better for north Europe? In a certain sense this work opens more questions than answers, but (I think) this is exactly the aim of authors. Though their analysis does not provide a solution to the problems raised during their evaluation, the combination of these statistical tools allows a better understanding of model deficiencies.
My major remark is that authors condense a great mass of information, difficult to assimilate without rereading back and back again. Moreover, they present results for different sensitivity scenarios, e.g. zero BC, const BC, 20% red, . . . Given a so large mass of information and different statistical analysis, I suggest, if possible, to re-organize the paper. I prefer a more in depth discussion of what may be the physical reasons for model deficiencies. A suggestion may be to try to highlight the role of physical mechanisms (this has already be done here and there, throughout the paper, but I prefer more emphasis). The logical course could be to start from model components (dry deposition, PBL dynamics, etc.), show what are model deficiencies and how your analysis is able to highlight these deficiencies. In this manner your ideas could are introduced as a "proof of concept" applied to a concrete example.

We thank the reviewer for the supportive comments. We are aware that we condense a large amount of information in the paper. The paper is a follow up of an even more lengthy manuscript, where all of the models participating in AQMEII were analysed for multiple pollutants. Here we focus on two models only and only on ozone. In this respect, the work we present here needs to be considered as a complementary, deeper, analysis of what was already presented in the previous study. Also the focus is different, in that here we wish to introduce methods that can help diagnose the cause of errors. Explaining the causes of model error, on the other hand, was left as an open question from the overview analysis in the previous publication (actually since the first phase of AQMEII!) and it still is. We have managed to frame the time scale of the error and, by excluding other plausible causes, we can conclude that the dynamics of the boundary layer (which in turn depend on the representation of radiation, surface characteristics, surface energy balance, heat exchange processes, development or suppression of convection, shear generated turbulence, and entrainment and detrainment processes at the boundary layer top for heat and any other scalar) are, at least partially, responsible of the daily model error. Since '*the main aim of this study is to move towards tools devised to enable diagnostic interpretation*' as clearly stated in the introduction, we prefer to base the discussion on the methods used to analyse the processes/variables rather than on the processes/variables.
We have added a new section (section 5) where all the relevant information is summarised in support of the goals of our analysis. In section 5 we describe how the methods we have proposed, in conjunction with the sensitivity runs, can help isolate the causes of model errors.

What is the role of dry deposition? Could the PBL dynamics better analyzed? The mean square error decomposition into its component and spectral components could identify where model need a deeper analysis? A plainer analysis of these aspects, moving secondary results to the supplementary section, would help the reader to track the pieces of information better.

One of the main criticisms raised by reviewer #3 is that we do not provide enough evidence to support our claim that the PBL dynamics is responsible for (part of) the daily error. We have now deepened the discussion about PBL errors throughout the text and summarised the main findings in the new section 5. In particular, since we are not in a position to estimate *directly* the effect of the PBL error(s) on the ozone error, we try to estimate it indirectly, by excluding the role of emissions and chemistry. We have strengthened this point in the revised manuscript by including the following additional results and analyses:

- The error structure (ACF and PACF) of surface wind speed and temperature (also with diurnal fluctuations removed);
- The error structure of the ozone (ACF and PACF) for the case with zeroed anthropogenic emissions;
- The error structure of the ozone (ACF and PACF) for the case with zeroed anthropogenic emissions repeated at receptors sited in areas with negligible isoprene emissions. These stations have been selected by using the map of isoprene cumulated emissions over the months of June, July, and August as provided by the two models analysed here. We have selected three stations per continent;
- The error structure for primary species (ACF and PACF), such as PM_{10} for Europe and CO for North America.

The results of these analyses (discussed in the revised manuscript) reveal a daily periodicity of the error, thus clearly pointing toward the role of PBL dynamics (in a broad sense, including e.g. radiation) in contributing to the generation of the daily error. The absence of a spatial or emission dependence and the persistence of the daily periodicity should give sufficient backing to our hypothesis.

We also note that the timing of the error of the base case discussed in section 4.3 (time shift of the diurnal component) is very similar to timing of the error of the zero emission case (reported in the Supplementary material, figure S13, and S14) regardless of the space location where this is investigated. A new figure has been added to clarify this point (Figure S15).

We are unable to isolate individual processes at this stage, given the observational and modelling data at hands. The high non-linearity of the model's components heavily complicates the interpretation of the results. The MSE decomposition has allowed us, for instance, to isolate the error due to bias from the error due to timing. We were then able to further analyse the time dependency and quantify it by cross-covariance analysis of the diurnal component, up to the conclusion that the timing accounts for an error of about the magnitude of the observed ozone variance. This example shows that the methods we have devised are complementary and that their combined use can contribute to novel insights. Again, then, our preference is to focus on how the methods can help understand the error in the processes.

Overall, I recommend the publication of this work, since it summarize the efforts made during AQMEII3 phase and suggest useful statistical analysis, well beyond the 'standard' statistical metrics, often used to qualify model results.