Referee #2: Anonymous

We thank the anonymous reviewer for the many helpful suggestions that have improved the manuscript. They have all been taken into consideration and addressed in the revised version of our manuscript.

General/Specific Comments:

The paper analyses the two phases of the AQMEII initiatives to test different techniques for improving deterministic estimates from multi-model ensembles. Even though the paper is generally well written, my opinion is that the scientific novelty is scarce, and most of the conclusions are not solid. Here are my motivations:

1) As stated in the Abstract: Line 5-7 "we demonstrate. . .is far from optimum,". This has been already proved several times is previous publications. (see Solazzo et al. 2013, Riccio et al. 2007, Galmarini et al. 2013, among others). In these papers, the same concepts and techniques of reducing the dimensionality of multi-model ensembles and optimal combination have been widely and repeatedly presented.

Response: We have re-written many parts of the manuscript to make more clear the focus and originality of the study. The scientific novelty of the study includes: (a) the comparison of several ensemble methods on pollutants of different skill using different datasets, (b) the introduction of an approach based on high-dimension spectral optimization, (c) the introduction of innovative charts for the interpretation of the error of the unconditional ensemble mean with respect to indicators reflecting the skill difference and error dependence of the models as well as the effective number of models. The manuscript has been rewritten to better reflect its originality.

2) Pag 4 line 4-13. The differences between the two experiments are described. The differences in the meteorology (two different years) and stations (amount of observations and their locations) are those that undermine more the statistical significance of the results. Most of them are presented (see Table2 and Table 4) without bootstrap confidence intervals or other techniques to assess if the differences between the two phases are statistically significant. The numbers of Phase I and II are often very close, and despite that, the authors build many conclusions on the top of these small differences. Also, most of the differences (if any) could be explained by the meteorology or the underlying changes in the station network. The authors should, at least, have made an attempt to make the two experiments more homogenous, i.e. by keeping a similar kind of stations over the two phases (same amount of urban, background stations).

Response: We have tested the statistical hypotheses on the differences of the distributions and their means through the Kolmogorov-Smirnov test and the t-test respectively. Although many differences were generally significant at the 1% level, we have decided to remove the comparisons of the two phases. The two experiments were independently designed and executed and have many differences. The harmonization of the validation set would remove an uncertain factor. Even then, the attribution of the differences between the two datasets to the uncertain factors (meteorology, models, coupling, etc.) in a statistical framework would still include a considerable amount of uncertainty. Moreover, such quantitative decomposition is beyond the objectives stated in this study. Therefore, the manuscript has been rewritten as an analysis of the performance of different ensemble techniques rather than as a comparison of the results from the two phases of the AQMEII activity.

3) Section 4.1 Forecasting performances. The authors want to prove that the weighting scheme might be used in forecasting mode. There are two issues here that undermine the conclusions of this section. My understanding is that some of the models participating at the inter-comparison are not running in forecasting mode (they use meteorological reanalysis as boundary conditions). While they

should run as an operational real-time forecasting model to be considered as realistic forecasts. Running these model in forecasting mode would change the model behaviors and error structures. Hence the conclusions achieved might change as well. How the bias of the models is removed in this test? Using the bias computed over the entire period (as previously mentioned) to correct forecast issued over the same test period would not be possible in real-time forecasting. This simple bias removal technique might not be so effective especially in forecasting mode when data from the future cannot be used.

Response: With respect to the first issue, the term 'forecast' has been changed to 'simulation' throughout the text. Concerning the second issue, bias removal is beneficial to the ensemble mean according to the bias-variance-covariance decomposition. It is not necessary for the approaches relying on reduced-dimensionality ensembles but the formulas for the analytically optimized weights have been derived with the assumption of bias-free members. As for the implementation, the mean bias over the training period is removed from the time-series of the test dataset. An explanation has been added in the text to clarify that the bias calculated in the train dataset (for the examined training periods of 5-60 days) is subtracted from the test dataset.

Our results indicate that after 30-60 days, the variable biases and weights have no effect in the skill of the weighted ensemble mean. Besides that, the seasonal bias reflects the systematic errors of the single models and it is considered a known quantity for validated models. Those considerations support the possible application of the approaches in real-time forecasting.

Minor comments:

The sentence: "In addition, mathematical tools such as ensemble forecasting provide an extra channel for uncertainty quantification and eventually reduction. Such method seems similar to the Monte Carlo approach; in practice, the similarity is only phenomenological since the probability density function of the uncertainty is not sampled in any statistical context like random, latin-hypercube, etc." is not clear at all. Ensemble forecasting cannot be considered as a mathematical tool in general. What does it mean:" Similarity in only phenomenological. . .."?

Response: The sentence has been removed from the text.

"benefits from ensemble forecasting arise from the averaging out of the unpredictable components (Kalnay, 2003)." It would be correct to say that benefits arise from averaging estimates with uncorrelated errors.

Response: Done. The sentence has changed accordingly.

Pag 3 line 25 "One of the challenges in ensemble forecasting is the processing of the deterministic models". This is true only if you are talking about a multi-model ensemble.

Response: Done. The sentence has changed accordingly.

Eq1 bias, var, cov? Should be presented with a more detailed notation

Response: Done.

Eq 2 E is the mean over what?

Response: Eq 1 and Eq 2 are related through their expectations over multiple stations.

Line 6 page 8 keep the same stations over the two phases

Response: We do not compare differences between the two phases in the revised manuscript.

Line 24 page 8 indirect feedback of what? Some details should be added

Response: The sentence has been removed from the text.

Line 19 page 5 I'd say the minimum (what does it mean ideal?)

Response: Done. The sentence has changed accordingly.

Section 2.1 86 % is a general value or something related to this paper

Response: It is general, the first N_{EFF} members account for 86% of the variability.

The same bunch of authors (or most of them) appears in previous publications regarding AQMEII phase I and II. I have some doubts (but I might be wrong) that they all give an active contribution to this paper or at least original compared to what already provided in the previous publications regarding these experiments. It would be fair to include in detail a description of the contribution of each author to this paper.

Response: The authors present in many AQMEII publications, as well as this one, are from the modeling groups that performed the simulations. Without the simulations, none of the published analyses would be possible.