Reply to the 1st reviewer

We thank the referee for the positive and helpful comments that have improved the manuscript. They have all been taken on board and addressed in the revised version of our manuscript.

General Comments:

The paper provides an exhaustive analysis of the multi-model ensembling in air quality problems, using the data from AQMEII exercises. The authors give clear theoretical introduction based on various error decomposition, and in the following sections compare the predictive skill of three ensemble products with well-defined mathematical properties, namely: - the arithmetic mean of the entire ensemble, - the arithmetic mean of an ensemble subset, linked to the error decompositions, - the weighted mean of the entire ensemble, linked to the analytical optimization. For the selection of ensemble subsets the authors consider several clustering methods. In the analysis different indices and skills are used – the choice seems to be sufficient for the purpose, however – to some extent – this is a question of taste (but "de gustibus non est disputandum"). The analysis is based on AQMEII dataset, which is an appropriate and representative for this purpose. The authors raised the important issues of ensemble training and predictability – this part seems to be particularly valuable. Of course drawing any final conclusions from the analysis relying on large but one dataset is uncertain, nevertheless some reasonable hints have been formulated. The paper is a step forward towards better understanding of how to build good ensembles. Specific and technical comments are included in supplement file.

We do appreciate the positive comments.

Specific Comments:

• Page 8 lines 17-18 (remark in brackets): in principle the models can have different distributions. No such an assumption is needed to obtain formulas in Table 1. Actually the only assumption made is that the models are treated as random variables with known variances (distributions doesn't have to be known and can vary from model to model).

We have removed the remark as it generated confusion. It was mistakenly pointing to the statistical distribution while the intention was to emphasize the randomness of the distribution in the i.i.d. sense (independent identically distributed).

• Page 11 line 2: while selecting the subset theoretically optimal sequence is obtained if the models are ordered starting from the one with the smallest variance and the ensemble is built by adding step by step the next with smallest variance. This is however theoretical as it works for independent models – nevertheless one can consider this also as a possible approach. This procedure could be extended to the case of correlated models by making use of eigenvalue analysis.

Indeed, this can be seen in the Example. We have included another paragraph to summarize the effects of the various perturbations.

"mme: its RMSE is reduced, compared to the i.i.d. case, if within the sample exist few members with lower variance or negative correlation."

• Page 23 lines 21-24, 30-31 and page 29 lines 9-12: the fact that static weights applied for the entire ensemble based on analytical optimization outscore other products is really noticeable. In my opinion there are several reasons for that:

- analytical optimization, in principle, relies on good statistics (as it optimizes average behavior described by the mean square error) which can be obtained in the considered case after using enough long period; on the other hand in case of dynamic weights the period is shorter thus it has worse statistics, which in case of the whole ensemble is more sensitive and the ensemble behavior can be easier worsen by few models;

- for subset of the ensemble it is still the chance that applying dynamic weights can give good results provided this subset is properly chosen. Hence the difficulty is shifted to the optimal selection of the ensemble subset, which can be cumbersome;

- theoretically the whole ensemble with proper treatment (bias corrections, good estimate of variance and covariance) should always provide minimum error provided – again – that enough good statistics exists.

Indeed, the stabilization of the statistics required a 60-day hourly time-series. The results demonstrated the superiority of the analytically optimized full ensemble at all available monitoring stations, in predictive mode. Using shorter periods, the statistics behind the weights were not robust. On the other hand, the robustness of the ensemble subset in dynamic mode is due to (a) the persistence of ozone levels and (b) the successful modelling of its extremes by only few members.

• Page 27 line 24: as I understand correctly the model's variance correction is made by using multiplicative factor. Could you shortly explain on what grounds this is based? For bias corrections there are techniques based – in general – on statistics, however the role of variance is different, and it can be treated as a kind of measure of model's uncertainty. Hence, variance correction would mean also uncertainty correction which sounds for me a bit suspiciously.

Definitely, we were comparing a 1st order bias correction that only removes the systematic errors with a 2nd order correction that also adjusts the spread. We rephrased the terms in the manuscript to the general term 'bias correction'.

Technical Remarks:

• Page 12 line 4: abbreviation JJA is not explained in the text.

Done as suggested.

• The quality of the figures could be improved - in pdf version they are not sharp – this concerns, first of all, the following pictures: 3, 7, 11, 14.

Done as suggested.

• Table 2: There are no definitions of MME, R and e_{m^*}.*

Done as suggested.