Reply to the 2nd 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.

Summary & Assessment:

This work provides a detailed (and somehow unnecessarily extensive) mathematical analysis of the main properties of an ensemble being focused on air quality issues utilizing AQMEII data sets. Constructing an optimal ensemble is the main target of this work and a variety of error decomposition methods are being used for this purpose.

Results based on (a) the ensemble mean of all ensemble members, (b) the ensemble mean of certain subsets and (c) the weighted ensemble mean of the total population of the ensemble, are presented and well documented. A variety of different cluster methods are being utilized for this purpose.

For the final assessment a set of different indices and skill scores are being utilized although some additional - quite helpful indices in building an optimal ensemble - are missing (as the Talagrand bin score for example).

This is a well-written and well-documented work and I trust it should be published although some major and quite a few more minor issues should be taken care beforehand.

We thank the Referee for the helpful comments. We have incorporated them into the revised manuscript.

Points of Major Importance

(a) *The paper seems to be quite long. By skipping some "unnecessary" details the paper could be abbreviated and become easier to be comprehended by non-specialist readers as well.*

We have re-written the Sections 2-6 according to reviewer's comments. To this end, we have:

- merged the old sections 2 and 3,
- created a new section 3 (Data and Methodology) with elements from the former sections 2.4, 4 as well as additional material requested for AQMEII,
- abbreviated section 4 under the new title "Interpretation of the ensemble error in light of its terms" with more clear connections to sections 3 & 5,
- emphasized the importance of the various parts in section 5,
- expanded the conclusions with the 'take-home' messages
- reduced the total number of figures to 13 (from 18)

Overall, the size of the paper was reduced by at least 15%.

(b) *It has to be documented why the specific data set being used for this study (namely AQMEII) has been an appropriate data set since its time span covers only a year. This becomes even more demanding since there is a clear tendency to generalize the results of this study beyond this "limited" data set.*

We have added the reasoning behind the appropriateness of the AQMEII database in the frame of the regional-scale operational air-quality ensembles in the new section 3 (Data and Methodology).

"The direct comparison of the simulated fields with the air quality measurements available from monitoring stations across the continent, at large temporal and spatial scales, is considered essential to assess model performance and identify model deficiencies (Dennis et al., 2010). This analysis falls within the context of operational evaluation of regional-scale chemical weather systems where most of the peaks in the energy spectrum are in the high-frequency era (hour, day, week). Together with the fact that the monitoring network extends over the whole continent, it emerges that the AQMEII database is suitable to capture the core temporal and spatial dependencies of the examined pollutants."

(c) Certain clarifications for the selection and utilization of the training (training set) and predictability modes are necessary for better understanding final results.

We have emphasized the split of the data in train/test sets in section 4.1 (spatially aggregated time-series) and 5 (point time-series).

"All ensemble products have been evaluated against the same test set, consisting of 30 equally spaced days from JJA (3rd June, 6th June, 9th June, etc)."

"We split records into a test dataset (30 equally spaced days from JJA: 3rd June, 6th June, 9th June, etc) and a train dataset (remaining two-third of the records). Using the train dataset, we first bias correct the time-series and then we estimate the mmeW weights and mmeS subset. Last, we apply the estimated parameters from the training dataset (weights, bias, N_{EFF}, clusters) into the test dataset"

(d) There should be a clear distinction between results that are true for any ensemble and what has found to be different for any special data-set ensemble as the one being used.

We have merged the key results in the conclusions section, ordered according to their 'generality'.

"The key results, obtained from the application of two general-purpose ensemble models to a representative air-quality dataset, can be summarized as follows (in order of decreasing generality):

1. *The unconditional averaging of ensemble members is highly unlikely to systematically generate a forecast with higher skill than its members across all percentiles as models generally depart significantly from behaving as a random sample (i.e. under the i.i.d. assumption). Further, the ensemble mean is superior to the best single model given conditions that relate to the skill difference of the members and the ensemble redundancy.*
2. *The relative skill of the deterministic models radically varies with location. The error of the ensemble mean is not necessarily better than the skill of the "locally" best model, but its expectation over multiple locations is, making the ensemble mean a skilled product on*

average. A continuous spatial superiority over all single models is feasible in ensemble products such as mmeW (error optimization through model weighting; keep all models) and mmeS (error optimization through trade-off between accuracy and diversity or variance and covariance; average on selected subset of models).

3. *Unlike mme, mmeW and mmeS require some training phase to find robust weights or clusters. The mmeW skill was more sensitive to its controlling factors than mmeS. A 2-month period was found necessary for the stabilization of the mmeW weights. On the other hand, mmeS was robust using both static/dynamic modes. In prognostic mode, if the training data have sufficient extent (at least 30 days), the minimum error is obtained with mmeW while for the case of limited training data, the minimum error is obtained with mmeS. Specifically:*
    - *mmeW: the weights were rather sensitive to the length of the training period, requiring at least 30 days to approach an asymptotic consensus. Nevertheless, learning over long time-periods (~2 months) and using those weights in predictive mode proved robust and accurate. Under proper training, its forecast skill outperformed all other ensemble products as well as individual models. The improvement across all stations over the mme was up to 35% for the RMSE and around 85% for the median hit rate.*
    - *mmeS: for the 13 member ensemble, the effective number of models was in the range 2-8, with the peak between 3 and 4. Its skill was significantly better over mme and individual models and it demonstrated the highest robustness with respect to the length of the training period. For training data of limited length (< 1 month), its skill was also better than mmeW. For ozone, switching from mme to mmeS, the properties that were relatively corrected more were accuracy (over diversity), error covariance (over error variance) and skill difference (over error correlation). The learning algorithms for subset selection, based on a sole dependent function of the error (e.g., diversity) rather than the error, did not achieve higher skill than mme. The improvement across all stations over the mme was up to 25% for the RMSE and 57% for the median hit rate.*
4. *The gross improvement in the RMSE of the multi-model ensemble mean achieved through the first and second moment correction of the modelled time-series, compared to only first moment correction was 0.6% for O_3, 2.1% for NO_2 and 11.8% for PM10. On the other hand, the improvement in the RMSE achieved through the exploitation of the ensemble mean in the form of mmeW or mmeS was 8.6% for O_3, 14.9% for NO_2 and 13.5% for PM10. Hence, even with adjustments in the systematic error and the spread in the models of an ensemble, a portion of its potential predictability is lost by using solely full ensemble averaging; superior improvements can be achieved through the optimization of an error decomposition approach.*
5. *For i.i.d. samples, the effective number of models equals the ensemble size (members). The mmeS and mmeW improve the skill of mme by constraining the ensemble into another where participating models replicate better the properties of an i.i.d. sample. Using N_{EFF} as indicator of i.i.d. sample, the decomposition of the skill as a function of the*

effective number of models demonstrated that for ozone, the three products were converging with increasing N_{EFF}. Those cases were occurring for intermediate concentration ranges, that all models are somehow tuned to replicate. On the other end, as N_{EFF} was decreasing and the ensemble was departing from behaving as an i.i.d. sample, the error gain from mmeS or mmeW over mme was gradually increasing, reaching on average 15% and 30% respectively. The extreme records were generally found in the assymetric range of the ensemble."

(e) The effect of bias correction on the ensemble characteristics (over- or under- prediction) could be easily (and clearly) shown by utilizing a set of Talagrand bin diagrams. It has not been clear also where exactly (i.e., on which data sets or sub-sets) this bias correction has been applied.

Bias correction is a perquisite for the mmeW and for this reason it was applied to all time-series individually. It was a simple 1st order correction applied to the examined chunk of the time-series. In forecast mode (test dataset), the bias estimated from the train dataset was subtracted.

The challenge faced in this work is to produce a single improved forecast out of an ensemble. Hence, the use of Talagrand diagrams, unlike probabilistic predictions (e.g. weather forecasting), have a different interpretation within this context (error minimization). On the other hand, a series of Talagrand diagrams as a function of the effective number of models has been plotted in a separate figure, by means of comparing the statistical distributions.

Points of Minor Importance

An extensive range of minor grammatical or spelling errors can be found in the document. These errors could be easily spotted and corrected by a native English speaker.

A comprehensive grammatical review has been generated by a native English speaker (anonymous reviewer No 3). All comments have been incorporated into the revised manuscript.