Reviewer #1. General Comments

The paper addresses the problem of redundancy in multi-model ensembles for air quality issues, hence it falls under the scope of ACP. The multi-model approach for air quality is extensively used and with increasing computing capabilities, large ensembles with huge datasets can be built. Therefore the question of models' diversity or similarity and redundancy of information in ensemble data is getting more and more important for practical reasons. According to my knowledge this is one of the first papers in which the authors made an attempt to compare different techniques used for reducing redundant information in air quality multi-model ensembles. The authors used data from the experiments undertaken within AQMEII community, which has been one of the biggest initiatives of air quality modellers in recent years. Hence, it seems that the choice of data for performing such an analysis was correct. Nevertheless, basing on the results presented, it is rather difficult to draw final conclusions on the best algorithm that should be used for the selection of the models from which the efficient ensemble (in the sense of redundancy) could be built. Of course, this problem is not easy and needs more intensive investigation. The paper can be considered as a starting point for further works in this field.

Indeed our main aim is to offer a starting point to the AQ community, with an overview of terminology and methods already in use in other communities. To our analysis, the issue of diversity and redundancy has proven problem-specific and general conclusions were hard to be drawn.

Specific comments

1. In order to describe similarities of the models the authors decided to use the metric based on bias proposed by Pennel and Reichler (2011). This approach is related to the idea that the models having similar systematic errors, in a sense, can be considered also similar. Of course, the reasons why these systematic errors are comparable can be very different. However, without going deeply into model details it is impossible to find these reasons. Therefore having only results of the simulations performed by the models, this approach is justified. The other elements, which can be also exploited for finding model similarities are models variances, as they are related to models uncertainties (reflecting the statement: "models are similar as their uncertainties are") – in fact one could imagine that more complex metric based both on bias and variance can be introduced. The variance is, however, much more difficult to apply in practise, as such information is simply not available in typical model simulations. In order to obtain it one would have to make sensitivity analysis or use other techniques. Anyway a short comment why such metric has been chosen by the authors could be included.

Reply. We completely agree with the reviewer. The focus of the paper is on bias and its redundancy and this is the main reason why the metric we choose deals only with bias. The variance is more difficult to assess having at disposal only model's outputs. We have added a few lines at the end of the first paragraph of section 3 to explain our point.

2. The results show that the ensemble is redundant even after removing multi-model error. Comparison of different methods has not given clear picture which one is more efficient in quantifying ensemble redundancy – this is probably problem specific. What would be interesting to know is what the weak and *strong points of each techniques are, in particular for air quality problems. This kind of comment would be highly appreciated*.

Reply. Section 4.5 offers a discussion about the results obtained by applying the different methods. We have added a few lines at the end of that section to point out that the issue of quantifying diversity is indeed problem specific.

3. The same as above comment could be said for the problem of identifying the efficient reduced ensemble and its members.

Reply. Section 5.5 and the two subsections there offer a discussion on the performance of the various methods. Section 5.5.1 in particular has been largely revised to extend the discussion and justify some results, especially about the ranking of members and the combination of minimum error.

4. The skill scores of reduced ensembles produced by different methods has not clearly shown that one of the applied techniques is superior to others, although the one based on the minimization of the mean square error on the average seems to outperform the others. This however depends on the skill to be applied and maybe on the specific problem under consideration. The interesting point here is the fact (see Table 5) that PCA technique in all the cases outperforms minMSE in RMSE values. What can be the reason – is it related to the fact that in PCA analysis additional information from observational data is provided ? Maybe the authors could comment it.

Reply. Section 5.5.2 is entirely devoted to discussing the skill of the reduced ensemble sets and indeed the beginning of the second paragraph states that the superior skill of the PCA weighted ensemble are due to the use of the observational data. Our aim is not really to find the “best” methodology, but to show that subsets selected by methods promoting e.g. diversity or accuracy can outperform the full members ensemble.

Technical comments

In general in order to make reading the paper easier I suggest to include precise mathematical formulas for the quantities used in the paper. This would clarify the presentation of the material.

Reply. We observe that it was our precise intention not to include equations, or at least to keep their number to a minimum. Our aim is in fact to make the paper as readable as possible to readers form diverse communities, not only by specialists. We note that we do not propose any new methodology here but we use several ones already well-established in the literature. We have provided a detailed and rich list of references to published works in many fields of science where details can be found. While we do not compromise on mathematical rigor in the text, we believe that keeping the (already long) presentation and discussion of the results free of heavy mathematical formulae contributes to the clarity of the concepts we wish to communicate.

1. In Section 3 the authors defined via Eq. 3 “standardized deviation of models from observation”. This term is very unlucky as it associates with the standard deviation which is something different. In fact what is defined in Eq. 3 is simply a bias normalized by the standard deviation of the species observed.

Reply. We have reworded the sentence there and avoided the term “standardization”.

2. In the same section in Eq. 5 correlation coefficient R is used – for the sake of clarification it would be better to slightly improve notation by introducing subscripts saying for which quantities this correlation is taken (for example: Rm,MME) and consequently use this notation.

Reply. Done as suggested

3. The mathematical formula for standardized quantities (Eq. 5), marked with “” could be included.*

Reply. We have provided the reference to the original work and we would prefer to avoid including the equation for this relationship

4. In Figure 3 (described in Sec. 3.1) it is difficult to distinguish colours – particularly green and blue.

Reply. We have changed colors and enlarged the symbols

5. Section 4.3 Eq. 8: if all the weights are 1, maybe it is better to remove them from the formula.

Reply. Done as suggested

6. Section 4.4 (hierarchical clustering): for better presentation one can introduce explicit mathematical formula for the level of similarity.

Reply. See reply to the main technical comment.

7. Section 5.2 and Figure 7: “mutual distance in 2D” – how is it defined ?

Reply. We have reworded the paragraph avoiding the term “mutual” which created confusion.

8. Section 5.4 (Eq. 9): to be in accordance with mathematical formalism one should rather write: if we denote by ΠPCm the projection operator on subspace PCm then dm,red=ΠPCmdm.

Reply. Done as suggested

9. Section 5.5.2: the formula in line 571 seems to be not correct: var(obs) should be with + sign.

Reply. We have corrected that.

Typographical errors:

1. Line 143: Czeck -> Czech
2. Line 298: Table 4 -> Table 3
3. Line 349 (Eq. 8): sij -> sij
4. Line 359: m disjoint …. -> r disjoint ….
5. Line 363: gropued -> grouped
6. Lines 455, 457, 458, 462: MSE -> RMSE: actually one can use either MSE or RMSE, it does not matter, but to be consistent with the statement that the decay is proportional to the square root of m, one should use RMSE.

7. Line 479 (Eq. 9): dm -> dm.
8. Lines 477, 479, 484: dmred -> dm,red (to be consistent with dm).
9. Table 5: red colour should be for the value 0.04: for ozone, PCA and NMB

Reply. We have addressed all of the listed typos.

Reviewer #1. General Comments

The paper addresses the fundamental issue of member diversity in multi-model ensembles. As the authors state, to date no attempts in this direction are documented within the air quality (AQ) community. Especially the issues of common biases and redundancy deriving from lack of independence, undermining the significance of a multimodel ensemble is handled by the paper. As stated in the abstract: Shared biases among models will determine a biased ensemble, making therefore essential the errors of the ensemble members to be independent so that bias can cancel out. Redundancy derives from having too large a portion of common variance among the members of the ensemble, producing overconfidence in the predictions and underestimation of the uncertainty. This is a very important issue within multi-model ensembles and to my knowledge it has not been handled before for AQ models. The basis for the analysis is the data from the AQMEII inter-comparison study. The AQMEII data constitutes model results from the present state-of-the-art models in North America and Europe, and most be considered the best dataset available for the study. The paper is within the focus of ACP. The paper is well written and well structured. The paper contains a smaller review of what have been done in the area until now and make references to appropriate literature. The basic starting point in the paper is that atmospheric models contains the same kind of errors since the models are based on identical level of understanding of physical, chemical processes and numerical methods. Therefore, the results from the models are not independent. This is an interesting and important issue when making model ensembles using a multi-model approach. The methodology used in the paper is scientifically sound. I therefore recommend publication in ACP.

It is kind of funny that the authors chose a Latin expression in the title. Normally, I would say that a title should be easy to understand, and as most people do not understand Latin, this criterion is not fulfilled. However, the Latin expression also potentially turns on the curiosity of the potential reader and therefore I will not suggest the authors to avoid Latin expressions in a title. Especially, when the expression is relevant. I have not found any part of the paper for specific comments. The argumentation and methodology in the paper is convincing.

Technical corrections

P 4995, line 13: we apply a number of member selection criterion - > we apply a number of member selection criteria

Line 15: non- redundant -> non-redundant

Reply. We have addressed the two typos.