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.

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.
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.
3. The same as above comment could be said for the problem of identifying the efficient reduced ensemble and its members.
4. The skill scores of reduced ensembles produced by different methods has not clearly shown that one of the applied technique 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.

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.

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.
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: $R_{m,MME}$) and consequently use this notation.
3. The mathematical formula for standardized quantities (Eq. 5), marked with "*" could be included.
4. In Figure 3 (described in Sec. 3.1) it is difficult to distinguish colours – particularly green and blue.
5. Section 4.3 Eq. 8: if all the weights are 1, maybe it is better to remove them from the formula.
6. Section 4.4 (hierarchical clustering): for better presentation one can introduce explicit mathematical formula for the level of similarity.
7. Section 5.2 and Figure 7: "mutual distance in 2D" – how is it defined ?
8. Section 5.4 (Eq. 9): to be in accordance with mathematical formalism one should rather write: if we denote by Π_{PCm} the projection operator on subspace PC_m then $d_{m,red}=\Pi_{PCm}d_m$.
9. Section 5.5.2: the formula in line 571 seems to be not correct: var(obs) should be with + sign.

Typographical errors:

1. Line 143: Czeck -> Czech
2. Line 298: Table 4 -> Table 3
3. Line 349 (Eq. 8): sij -> s_{ij}
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 -> d_m.
8. Lines 477, 479, 484: dm_{red} -> $d_{m,red}$ (to be consistent with d_m).
9. Table 5: red colour should be for the value 0.04: for ozone, PCA and NMB.