Dear Prof. Brandt,

we have uploaded the revised version of the paper entitled "*De praeceptis ferendis*: good practice in multi-model ensembles (acp-2014-200)". We would like to thank you for all the detailed comments concerning the mathematical/physical rigor and clarity of the paper, and we tried our best to comply with the reviewers' comments. Accordingly, we believe we have been able to significantly enhance both the content of the paper and quality of English.

We have re-written and re-organized the majority of the presented material. Overall, the size of the paper was reduced by at least 15%. 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)

We believe, thanks to the reviewers, we have been able to accomplish a worthwhile improvement of the quality of the paper and have, we hope, clarified the points raised by the reviewers.

Sincerely yours,

Ioannis Kioutsioukis and Stefano Galmarini

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.

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.

Reply to the 3rd reviewer

We thank the referee for the positive, helpful and comprehensive review that has improved the manuscript and its overall readability. All comments have all been taken on board and addressed in the revised version of our manuscript.

Summary & Verdict

This article presents a detailed analysis of statistical properties of ensembles, their decompositions and demonstrates methods for constructing optimal ensembles in forecast/hindcast purposes modes.

The articles appears to be an important contribution to an area where ensemble methods are becomingly increasingly important. The article is well-written and I believe it should be published. Howevever there are a small number of major issues, alongside a large number of minor points on which the manuscript could be improved. If these points are met, then I can recommend the article for publication. I congratulate the authors on their efforts - it is a good piece of work, and I am glad I had the chance to review it.

We do appreciate the positive comments

General points

Major

• While I enjoyed reading the manuscript, I think it was unnecessarily long. The authors are encouraged to lay out a clear "take-home" message for the whole paper, and the individual sections. The connections between successive sections should be clear, and so should their importance to the main message of the article. Once this is done, it may be obvious that certain sections can be abbreviated, merged or removed. This may improve the overall readability of the manuscript.

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%.

• Related to the above point, the overall goal of the exercise should be emphasised, both in the introduction and the conclusions. Is it to obtain better predictions at non-observed sites, better predictions at future times, or a better historical reanalysis?

The overall goal is to produce a single improved forecast out of an ensemble. We have further emphasized it in the introduction and the conclusions.

"The overall goal of the study is to highlight the properties, through model selection or weighting, that guarantee a symmetric distribution of errors and eventually produce a single improved forecast out of an ensemble."

"The principal objective addressed is the emergence of ways to produce a single improved forecast out of an ensemble that potentially outscores the traditional arithmetic mean as well as the best numerical model."

"The goal of this work is to evaluate potential schemes to produce a single improved forecast out of an ensemble."

• The manuscript includes general findings (mathematically derived), which hold for any ensemble data-set, as well as findings specific to particular data-sets. I suggest clarifying the extent to which the findings from the particular data-sets examined are general properties of geophysical ensemble modelling, rather than specific to these data-sets.

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."*

• The authors need to argue the case that the AQMEII data-set is an ideal data-set for use in this context. The data-set covers one year only, so important cyclical (i.e. annual) features cannot be examined. It may have been preferable to use a longer time-series (e.g. an ensemble of multi-year climate simulations). The reasoning would certainly be stronger if additional data-sets were examined alongside the AQMEII data-set, or if there were extra references to similar analyses for longer time-series (e.g. as arising in the climate literature).

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."

• If I understand correctly, the mme< and mmW require observations to choose optimal weights/members. This means that the data must be divided somehow into a "training set" and a "test set". In the manuscript, it was often unclear how the data-sets were partitioned between testing and training. Also, for a fair comparison, the same test set should be used for all methods, including the mme, even though this does not involve a training set.

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"

• Please clarify where bias-correction has been applied, and where it has not. Is it applied everywhere? If so, how? For each individual time-series?

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.

"Bias correction. According to the bias-variance-covariance decomposition, bias is an additive factor to the MSE and model outputs should be corrected for their bias before any ensemble treatment. The analytical optimization of the ensemble error and the defined weights (Table 2) also assume bias corrected simulations. Here we do not intend to review the available algorithms for the statistical bias correction (e.g., Dosio and Paurolo, 2011; Delle Monache et al., 2008; Kang et al., 2008; McKeen et al., 2005; Galmarini et al, 2013); the correction applied in this work refers to a simple shift of the whole distribution within the examined temporal window, without any scaling or multiplicative transfer function"

"… two bias-correction schemes (namely, the ideal for the test set and the one calculated from the training set)."

Minor

• I recommend the use of upper case acronyms (e.g. MME instead of mme, ID instead of id, KZ instead of kz, I.I.D. instead of i.i.d.). Upper and lower-case acronyms are used inconsistently (e.g. MSE and RMSE were generally capitalized in the manuscript).

We generally use upper-case only for the indices of skill.

• I recommend finding another name for the mme< ensemble. The < in mme< can be misleading, because the term mme is also used, and since < has a well- understood meaning. See pp. 15825, line 25 for an example of how the term mme< can appear confusing - its placement alongside / makes it

look like a typing error. Perhaps MES for "mean of the ensemble subset" could be used instead of mme<.

We have changed mme< and mmW to mmeS and mmeW respectively.

• In your multi-plot figures, if the range of the x- or y-axes differ between the different panels (e.g. in Figure 1), make a note in the caption that the axes differ.

Added the requested clarification for the different range at plots 1, 2, 4, 6, 9 and 13.

Specific points

Major

• pp. 15811, l. 14-16: There is an important statement missing here, and in the rest of the article, as far as I can see: mme is a special case of mm< (since it uses the full subset), and mm< is a special case of mmW (since in mm< some weights are zero and the others are equal and sum to 1).

We have added the following sentence:

"Note that mmeS is a general case of mme and a special case of mmeW (if weights can only take two discrete values)."

• pp. 15812, l. 5-6: If the optimal weights can be negative, then the ensemble estimator is no longer bound by the ensemble. In the case of the AQMEII datasets, the ensemble data-set used pertains to ozone concentrations. In theory, negative weights could result in negative concentration estimates. Did this occur? And if so, how was this dealt with? Were they truncated at zero or was the negative value simply used in the analyses?

We did not filter out any value to avoid artificial skill. Indeed, negative weights were observed for mmeW using dynamic weights in predictive mode. The application of static weights did not produce any negative concentrations for the examined pollutants and season.

• pp. 15829, l. 11-12: "On the other hand, static weights outscore all other products".I think this is one of the most interesting findings of the paper. I think it deserves emphasis elsewhere in the manuscript. Table 3 shows that the mmW gives by far the best results of the F(s) column. This is a very important finding. I would like to know if this is a common feature, or specific to the data-sets considered.

We have re-calculated Table 3 (test set) including additional lengths for the training set. The mmeW has the potential to outscore all other products given a training time-series with sufficient length. For the atmospheric pollutants such as O_3, NO_2 and PM_{10}, the required period to stabilize the error covariance (and hence the weights) was found to be around sixty days. This length should be re-evaluated for every different dataset, as mentioned in the general roadmap at the conclusions. Also, the finding has been emphasized in the key results (conclusions section).

"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. Specifically:

- *mmeW: for short-range forecasting (horizon<4days), it achieves lower error than the theoretical one for uncorrelated equally weighted ensemble. However, the variability of those weights at that scale is beyond any predictability. Nevertheless, learning over long time-periods (~2 months) and using those weights in predictive mode proved robust and accurate. Its 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."*

Minor

As a native English speaker, I have noted a range of minor grammatical or spelling errors. These are noted, for example, as:

• pp. 15820, l. 28: "the the" -> "the"

However in cases where the authors have used a phrase that I think sounds odd, it is noted as a suggestion, for example:

• pp. 15805, l. 26: "object of" -> "subject to"

So here are the comments:

• pp. 15804, l. 7: suggestion: cut "for which one cannot be gained without expense of the other"

The whole sentence has been removed

• pp. 15804, l. 10: the -> these

Corrected

• pp. 15805, l. 2: suggestion: "The availability of computing means in recent ..." -> "The availability of increasingly powerful computing in recent ..."

Changed as suggested

• pp. 15805, l. 3: suggestion: "application" -> "feasibility and use"

Changed as suggested

• pp. 15805, l. 11: what is meant by "mathematical bibliography"?

Changed to: "statistical methods"

• pp. 15805, l. 14: suggestion: "driven by the initial conditions uncertainty" ->"driven by uncertainty in the initial conditions"

Changed as suggested

• *pp. 15805, l. 19: ”condition” -> ”conditions”*

Corrected

• *pp. 15805, l. 23: suggestion: ”the one of” -> ”the error of”*

Changed as suggested

• *pp. 15805, l. 26: ”object of” -> ”subject to”*

Corrected

• *pp. 15806, l. 1: suggestion: ”risky” -> ”less reliable”*

Changed as suggested

• *pp. 15806, l. 14: suggestion: ”independent members.” -> ”independent members only.”*

Changed as suggested

• *pp. 15806, l. 15-16: suggestion: add to this sentence a note that this will be demonstrated later in the article. Otherwise, the reader may want a reference.*

Changed as suggested

• *pp. 15806, l. 17: IID around observations - do you simply mean unbiased?*

Changed ‘identically’ to ‘randomly (i.e. unbiased and uncorrelated)’.

• *pp. 15806, l. 17-19: I suggest justifying this statement*

Changed to: *“This ‘randomness’ in the model outputs of an ensemble is not a pragmatic condition. Nevertheless, an optimal ensemble can be constructed a posteriori by inducing this property in the members.”*

• *pp. 15806, l. 17: ”this property could not be” -> ”this property can not be”*

Corrected

• *pp. 15806, l. 22: suggestion: delete ”to exploit ways”*

Changed as suggested

• *pp. 15808, l. 14: one of the termse between = signs is repeated and hence unnecessary*

It is not repeated actually, there is a parenthesis in one of them implying a factorization.

• *pp. 15809, l. 7-8: I am not convinced by the statement that ”the covariance term indicates the diversity or disparity”. Covariance is a joint metric of variance and correlation. Correlation is a better measure of diversity/disparity than covariance, per se.*

All statements about diversity have been removed. The links between diversity, variance and covariance are left for the next paragraph with the comparison of the two decompositions.

• pp. 15809, l. 12: ”as little as possible” -> ”as low as possible”

Corrected

• pp. 15809, l. 10-12: This claim is not self-evident. Even though two of the terms are necessarily positive, it does not follow that one should focus only on the term which may be negative. It depends on the scale of the individual terms.If the necessarily positive terms dominate, then error-minimisation may be more effective by focussing on these terms.

It has been rephrased: *“Given the positive nature of the bias and variance terms and the decreasing importance of the variance term as we include more members, the minimization of the quadratic ensemble error ideally suggests unbiased (or bias-corrected) members with low error correlation amongst them (to lower the covariance term).”*

• pp. 15810, l. 1: suggestion: ”we have no criterion” -> ”we have no a priori criterion”

Changed to: *“… we have no criterion for identifying a priori that best individual …”*

• pp. 15810, l. 1: It is claimed that ”we have no criterion for identifying the best individual”, there are the observations and a wealth of literature on verification - or is something else meant here?

Rephrased to: *“But, given that we have no criterion for identifying a priori that best individual (i.e. which ensemble member will best match the observations at future time-steps), all we could do is pick one at random. In other words, taking the combination of several models would be better on average over several patterns, than a method which selected one of the models at random. The last statement is not self-evident for non-random sampling of the best member (e.g. conditioned to past errors from the models).“*

• pp. 15810, l. 14: suggestion: ”we need to” -> ”it is necessary to”

Changed as suggested

• pp. 15811, l. 1: Please define omega

We have added a reference for Ω and we explain its terms in the text. We believe this is sufficient for the context. Nevertheless, its formula is: $\frac{1}{M}\sum_i E\{(f_i - E\{f_i\})^2\} + \frac{1}{M}\sum_i \left(E\{f_i\} - E\{\bar{f}\}\right)^2$

• pp. 15811, l. 10: ”presented decompositions” -> ”decompositions presented”

Corrected

• pp. 15811, l. 24: ”the models are assumed as random variables (i.e. their distribution is identical)”. The statement in parentheses does not follow.

We have removed the remark in parenthesis 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.

• pp. 15812, l. 9: ”as more models as possible” -> ”as many models as possible”

Corrected

• pp. 15812, l. 11: suggestion: ”it provided” -> ”it provides”

Changed as suggested

• pp. 15812, l. 11: ”At the same time, it provided” - what does ”it” refer to? This section? This method? The arithmetic mean? Something else?

Changed to “this section”.

• pp. 15812, l. 15: suggestion: ”through ..., points” -> ”points, through ..., ”

Changed as suggested

• pp. 15812, l. 18: suggestion: ”through ..., relies” -> ”relies, through ..., ”

Changed as suggested

• pp. 15812, l. 20: suggestion: ”through ..., provides” -> ”provides, through ..., ”

Changed as suggested

• pp. 15813, l. 5: Suggestion: add line ”Note that 2 is a general case of 1, and 3 is a general case of 2.”, as noted above.

Changed as suggested

• pp. 15813, l. 18: The factorial term should not have a horizontal line separating the M and the k, otherwise this may be read as a fraction.

Corrected

• pp. 15813, l. 21: suggestion: ”the optimal weights do not deviate” -> ”the optimal weights show little deviation”

Changed as suggested

• pp. 15814, l. 1: ”two-third” -> ”two-thirds”

Corrected

• pp. 15814, l. 5: suggestion: ”particular” -> ”notable”

Changed as suggested

• pp. 15814, l. 6-7: Rewrite ”the sentence ”This upper bound ... equal sign” for clarity

Rewritten as: *“The upper bound of the RMSE values is defined from the ensemble combinations consisting of biased members of equal sign.”*

• *pp. 15814, l. 10: "the optimal combination" - meaning with the lowest RMSE?*

Changed to *"optimal combination (i.e. lowest RMSE)".*

• *pp. 15814, l. 11: What is meant by "no clue"?*

Changed to *"no conclusion".*

• *pp. 15815, l. 6: suggestion: remove "well", or replace "well" with "largely"*

Changed as suggested

• *pp. 15815, l. 16: "was organized which consisted in having the two communities" -> "was organized, involving the two communities"*

Corrected

• *pp. 15815, l. 20-21: "meteorological driver, air quality model, emission" -> "meteorological drivers, air quality models, emissions"*

Corrected

• *pp. 15815, l. 25: "JJA" -> "JJA (the period of June-July-August)"*

Corrected

• *pp. 15815, l. 26: "thirteen models that give rise to" -> "thirteen models, which give rise to " or "thirteen models, giving rise to "*

Corrected

• *pp. 15816, l. 15-16: "The RMSE of each possible combination .... obtained theoretically" - this part of the sentence was unclear and could be rewritten for clarity*

Re-written as: *"The RMSE of the mean of all possible combinations as a function of the ensemble size justifies the statement obtained theoretically …"*

• *pp. 15816, l. 19: "EU4r". This acronym needs to be introduced properly. It appears later, and the reader can figure it out from the context, but I would advise explaining these terms at this point in the manuscript.*

We added the missing definitions at the beginning of the section and super-imposed the domains on an existing spatial plot: *"*In section 4, we make use of spatially aggregated time-series (EU1 to EU4, illustrated in Figure 9a) *… The evaluation of the examined ensemble products (mme, mmeW, mmeS) will rely on several indices of error statistics calculated at rural receptors."*

• *pp. 15817, l. 5: "sub-regions" - this terms should be introduced along with the above acronym*

Corrected in the abovementioned point

• *pp. 15817, l. 4: suggestion: "quasi constant" -> "roughly constant"*

Following the restructuring of the manuscript, this part has been removed.

• *pp. 15817, l. 7: suggestion: "This number is small" -> "This fraction is small"*

Following the restructuring of the manuscript, this part has been removed.

• *pp. 15817, l. 11: "normalization" - how?*

Rephrased to: *"For k=6, the 13 models give rise to 1716 combinations; each model participates at 792 of them. The fractional contribution of individual models (for k=6) to skilled sub-groups (portion of skilled combinations per model) is given with the red numbers."*

• *pp. 15817, l. 15: Model 4 has negative weights - what does this mean?*

We have added the definitions at the end of section 2.3. *"There is no physical interpretation for the negative weights; if they arise for some models, it is simply a result of the optimization of the cancelling out of the individual errors. For example, models with highly correlated errors may be given weights of opposite sign." It is also explained in section 4 "The* d_m *dendrogram also explains the reasoning behind the negative weights calculated analytically. The model pairs identified with highly correlated errors [like 4 and 12 or 11 and 13] are given weights of opposite sign."*

• *pp. 15817, l. 18: Suggestion: "Definitely" -> "Clearly"*

Changed as suggested

• *pp. 15817, l. 22: "Low skill cluster" -> "A low skill cluster"*

Corrected

• *pp. 15817, l. 22: "1, 2 and 10 that" -> "1, 2 and 10, which"*

Corrected

• *pp. 15817, l. 24: "improved variance" - improved how? closer to the variance of the obserations?*

Changed to: *"its variance is closer to the variance of the observations"*

• *pp. 15817, l. 25: suggestion: "error, correlation" -> "error, and correlation"*

Changed as suggested

• *pp. 15817, l. 23: "light" -> "slight"*

Corrected

• *pp. 15817, l. 23: Suggestion: "Compared with" -> "Considering"*

Changed as suggested

• *pp. 15817, l. 23: "participation statistics" - are these shown?*

Re-written as: *"Considering the participation statistics of the previous graph (given by the red numbers) …"*

• *pp. 15818, l. 2: suggestion: "to form good ensemble groups" -> "to form part of skilful ensemble groups"*

Changed as suggested

• *pp. 15818, l. 4: suggestion: "under proper combination scheme" -> "in the right combination"*

Changed as suggested

• *pp. 15818, l. 13: "will not produce the best ensemble output" - best how? As measured by the overall MSE?*

Added explanation in parenthesis: *"will not produce the best (i.e. minimum MSE) ensemble output."*

• *pp. 15818, l. 19: "form skilled ensemble products" -> "form skilful ensemble products"*

Corrected

• *pp. 15818, l. 22: "independent to" -> "independent of"*

Corrected

• *pp. 15818, l. 23: "part implying" -> "part, implying"*

Corrected

• *pp. 15819, l. 1-3: This refers to the variance-covariance plot in Fig 2d. I would suggest scaling the covariance term somehow by the variance, so that the horizontal and vertial axes display vertical information.*

We would prefer to keep the chart unscaled since in this form, adding the two components directly yields the MSE displayed by the colorbar (RMSE).

• *pp. 15819, l. 5-6: "error formula is becoming lower but the covariance term is deteriorated" -> "error formula falls while the covariance term deteriorates"*

Corrected

• *pp. 15819, l. 7: suggestion: "highly correlated" -> "strongly positively correlated"*

Changed as suggested

• *pp. 15819, l. 8: suggestion: "bigger" -> "larger" or "greater"*

Changed as suggested

• *pp. 15819, l. 10: "granting" -> "leading to"*

Corrected

• pp. 15819, l. 10: suggestion: ”attempted” -> ”illustrated”

Changed as suggested

• pp. 15819, l. 13: suggestion: ”conditions for being lower than the one of” -> ”conditions for the MSE being lower than that of”

Following the restructuring of the manuscript, this part has been removed.

• pp. 15819, l. 14: ”constrain” -> ”constraint”

Corrected

• pp. 15819, l. 16-17: ”the ensemble, i.e. the error dependence” -> ”the ensemble (i.e. the error dependence)”

Corrected

• pp. 15819, l. 17: ”explained variation” -> ”variation explained”

Corrected

• pp. 15819, l. 18: ”The pairwise plot” - Is this part of figure 2? If so, which plot?

Added explanations: *“The pairwise plot (Figure 5a) of the skill difference (measured by <MSE>/MSE(best)) versus the ensemble redundancy (measured by the explained variation by the maximum eigenvalue) as a function of …”*

• pp. 15819, l. 25: suggestion: remove ”thought”, or replace with ”concept”, ”principle”, ”idea” or similar

Following the restructuring of the manuscript, this part has been removed.

• pp. 15819, l. 28: suggestion: replace ”candidate for the interquartile range” -> ”estimator for commonly observed values”. This is because ”interquartile range” means something specific in statistics, namely the 75th percentile minus the 25th percentile.

Following the restructuring of the manuscript, this part has been removed.

• pp. 15820, l. 1: Figure 3 shows that the ensemble mean for the optimal combination performs much better at the extremes compared to the mean of the full ensemble. This could be noted here.

Added comment: *“The distribution around the truth in all those ensemble products has always higher symmetry compared to mme, as can be seen in Figure 2. In addition, they all perform much better at the extremes compared to the mean of the full ensemble.”*

• pp. 15820, l. 10: suggestion: ”behaving as an i.i.d. sample” -> ”under the i.i.d. assumption”

Changed to: *“behaving as a random sample (i.e. under the i.i.d. assumption)”*

• pp. 15820, l. 26: suggestion: ”case by case” -> ”for the particular data-set”

Changed as suggested

• pp. 15820, l. 28: ”the the” -> ”the”

Corrected

• pp. 15821, l. 2: At least give a basic explanation of what is meant here. Ei and di are presented in table 2 are these the same as presented here? Also, explain which of the two correlation matrices is most relevant in the contexts considered.

Added explanations: *“The clustering algorithm has been utilized against the d_m matrix defined in Table 1, namely the corr(d_i,d_j), which generates more dissimilar errors compared to the e_m metric (for details see Solazzo et al., 2013)”*

• pp. 15821, l. 8: suggestion: ”the above produced” -> ”the method described above produced” or ”the above procedure yielded”

Changed to: “*The application of the clustering procedure yielded five disjointed clusters*”

• pp. 15821, l. 11: ”inspection at” -> ”inspection of”

Corrected

• pp. 15821, l. 13: suggestion: ”the visual clusters of” -> ”the clusters visible in”

Changed as suggested

• pp. 15821, l. 14: ”plor” - ”plot”

Corrected

• pp. 15821, l. 16: suggestion: ”marked” -> ”noted” or ”clear” or ”obvious”

Changed to: “*noted*”

• pp. 15821, l. 19: ”ID, DU, SY, LT” - at least give a basic explanation of these acronym. Details can be left to the references.

Following the restructuring of the manuscript, this part has been removed. Nevertheless, their meaning was ID: intra-diurnal, DU: diurnal, SY: synoptic, LT: long-term.

• pp. 15821, l. 24: ”not fundamentally correct” - do you mean that it may lead to a higher error in the sub-ensemble

Following the restructuring of the manuscript, this part has been removed. Indeed, the meaning was that the unconditional use of models with systematic errors within an ensemble may well mask the benefits of ensemble averaging.

• pp. 15822, l. 2: ”trace” - what is meant here? The sum of the diagonal of a matrix?

Following the restructuring of the manuscript, this part has been removed. Nevertheless, the meaning was ‘location’.

• pp. 15822, l. 11, 13: ”uncertainty” - what is meant here? error? if so, they have different meanings.

Following the restructuring of the manuscript, this part has been removed. Nevertheless, the meaning was forecast uncertainty, as measured by the standard deviation of the errors.

• pp. 15822, l. 19: ”in principle uncorrelated errors” - is this really to be expected? They are based on the same limited set of ensemble members.

Following the restructuring of the manuscript, this part has been removed.

• pp. 15823, l. 8: ”diagnostic” - meaning?

Following the restructuring of the manuscript, this part has been removed. Nevertheless, the meaning was ‘analysis over historic periods’.

• pp. 15823, l. 20: ”medium-range” - what time-scales?

Added: *“… forecasts at daily to weekly time-scales”*

• pp. 15823, l. 22: ”evidences” -> ”evidence”

Corrected

• pp. 15824, l. 10: ”if models were uncorrelated” - is this so? has this been tested?

It is derived from the bias-variance-covariance decomposition. For uncorrelated and bias-corrected models, the ensemble error originates from the variance term only, which defines the lower bound for the multi-model mean of uniform ensembles. This decomposition however is not valid for the weighted ensemble mean. It has been rephrased as: *“the mmeW error is superior to the theoretically derived lower bound for the mme error (2nd term in the bias-variance-covariance decomposition) if models were uncorrelated”.*

• pp. 15824, l. 17: ”As the window size decreases” - do you mean ”As the window size increases”?

The x-axis (NoSegments) is the number of chunks in which the 3-month time-series is sliced. Therefore, NoSegments is inversely-proportional to the window size (‘1’ segment stands for 92 days, ‘2’ segments stands for 92/2 days, ‘4’ segments stands for 92/4 days etc.). The following sentence has been added:

“The x-axis is the number of chunks in which the JJA time-series is sliced; hence it is inversely proportional to the window size.”

• pp. 15824, l. 17: ”RMSE” - since the ensemble members have been bias corrected, should this metric be the bias-corrected RMSE? If so, why not call it BCRMSE to distinguish it from the normal use of RMSE, which has non-zero bias.

We would prefer to avoid additional terminology that might end up in generating more confusion than clarification, due to the length of the manuscript.

• pp. 15824, l. 19-20: ”inversely proportional .... lower variance” - clarify this statement

Following the restructuring of the manuscript, the whole paragraph has been removed.

• *pp. 15824, l. 22: "The cases with high variability, where the majority of models fail to simulate well" - note that this is a fundamental problem with high-resolution forecasting, relating to phase vs. amplitude accuracy.*

Following the restructuring of the manuscript, this part has been removed.

• *pp. 15825, l. 1: "it is variable as it" - what is "it"?*

Changed to: *"The ensemble error gain (i.e. the difference between the ensemble error and the average error of the models) is variable as it depends significantly on the individual model distributions around the truth."*

• *pp. 15825, l. 6: "ratios" - what is meant here?*

Changed to: *"If models were uncorrelated (see Table 2), the mme error would always be lower than any single model since the MSE ratios (worst/best) are smaller than 14 (=M+1)."*

• *pp. 15825, l. 10: "joint restrictions" - what is meant here?*

Added sentence: *"The joint conditions for the skill difference and the redundancy, for correlated models, granting an ensemble with mme error lower than the best model are presented in Figure 5b."*

• *pp. 15825, l. 13: what are the values in the round braces?*

Added clarification: *"[(explained variation by the highest eigenvalue, skill difference) …"*

• *pp. 15825, l. 24: what are the training/testing sets used here?*

This part is exploratory analysis (hindcast), using all data from JJA. The division into train/test set is done in sections 4.1 and 5.

• *pp. 15825, l. 25: this illustrates why the term mme< is potentially confusing*

We have changed mme< to mmeS.

• *pp. 15825, l. 28: "error minimisation through mme<" - what is meant here?*

Changed to: *"Using dissimilar time-series from the four examined sub-regions, we observe that the optimal sub-ensemble combination (mmeS) compared to the full ensemble (mme) generally:"*

• *pp. 15826, l. 2: "distorted" - what is meant here?*

Following the restructuring of the manuscript, the remark in parenthesis has been removed. Nevertheless, the meaning was 'misrepresented'.

• *pp. 15826, l. 7: "variance" - measured/defined how?*

Added clarification: *"variance (term in Eq. 2)"*

• *pp. 15826, l. 20: "matrix whose skeleton" - what is meant here?*

Following the restructuring of the manuscript, the sentence has been removed. Nevertheless, the meaning was: *"matrix whose effective number of degrees of freedom"*

• *pp. 15826, l. 22: "normalized" - how?*

We have removed the remark between commas ", regardless if normalized or not,". Nevertheless, the meaning was 'standardization' (i.e. subtract mean and divide with the standard deviation).

• *pp. 15827, l. 2: "conceived" -> "interpreted as meaning"*

Following the restructuring of the manuscript, the sentence has been removed.

• *pp. 15827, l. 9: "real" - do you mean non-negative? In mathematics, a distinction is made between real and imaginary/complex numbers, although I think somethign else is meant here.*

Following the restructuring of the manuscript, this part has been removed. Nevertheless, the meaning was 'not strictly positive'.

• *pp. 15827, l. 21: "coherent" - what is meant here?*

Changed 'coherent' to: *"have similarities (i.e. models 3, 5 and 6 that receive on average the highest weights are also the ones used most frequently in mmeS)"*

• *pp. 15828, l. 17: "remaining" - do you mean "subsequent"?*

Following the prior definition of the test dataset, the remark in parenthesis is changed to: *"(e.g. weights calculated over a sample of 62 days"*

• *pp. 15828, l. 22: "real" - do you mean non-negative or not with a zero imaginary component?*

Changed to "real numbers".

• *pp. 15828, l. 27: I have never seen this symbol before (looks like a percentage sign, but with an extra o) - what does it mean?*

Following the restructuring of the manuscript, this part has been removed. Nevertheless, the meaning was: "*Per mille (parts per thousand)*".

• *pp. 15829, l. 23: "phenomenological" is probably not the right word here. Perhaps "fundamentally" is better.*

Rephrased sentence to: *"An improvement similar to the one obtained through the mmeW scheme (bias correction, model weighting) has been documented in weather forecasting with MME (Krishnamurti et al., 1999), where weights were estimated from multiple regression."*

• *pp. 15829, l. 16, 18: "diagnostic mode", "prognostic mode" - do you mean training and testing?*

Added: (training phase) and (testing phase) respectively.

• *pp. 15836, l. 6-7: "The learning algorithms ... of it (e.g. diversity)" - please clarify*

Rephrased to: *"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."*

• *pp. 15836, l. 11: "can be seen as an application of flow depedendent error covariance" - I disagree. This has nothing to do with flow, which is only a feature of temporally dependent 1-, 2-, and 3-dimensional models. There is a substantial literature on flow dependent error covariances in the field of data assimilation, and this is quite a distinct problem.*

Following the restructuring of the manuscript, this part has been removed.

• *pp. 15836, l. 26: suggestion: "era" -> "range"*

Changed as suggested

• *pp. 15836, l. 29: suggestion: "threshold point ... progress further" -> "benchmark against which all other weighting schemes should be evaluated"*

Changed as suggested

• *pp. 15837, l. 9: "extent" - what is meant here?*

Changed to "length"

• *pp. 15837, l. 9: "confined from" - do you mean "distinct from"?*

Changed to "determined"

• *pp. 15842, table 1: reference for the derivation*

Added reference: "(from Potempski and Galmarini, 2009)"

• *pp. 15843, table 2: what do the stars mean in em, dm*

Added: A '*' indicates standardized vectors

• *pp. 15844, table 3: relate "hindcast" and "forecast" to "training" and "testing"*

Changed "hindcast" and "forecast" to "training" and "testing"

• *pp. 15847, l. 3: "Multi aspects" -> "Multiple aspects"*

Corrected

• *pp. 15837: "sm" - do you mean single model?*

Changed to: *"(explained variation by the maximum eigenvalue s_m)"*

• *pp. 15849, bottom-right panel: what is point R? Reference?*

Added in the caption: *"The point R on the x-axis represents the reference field (i.e. observations)"*

• *pp. 15850: what are "dm" and "em"?*

Following the restructuring of the manuscript, this part has been removed. Nevertheless, e_m and d_m are indices defined in Table 1.

- *pp. 15850: The caption needs further work to clarify all the terms. Please review.*

Following the restructuring of the manuscript, those plots have been removed.

- *pp. 15851: Would these curves appear linear if the x-axis were plotted on a logarithmic scale?*

Theoretically, if models were not bias-corrected, the curves would appear more linear if the x-axis was changed with the inverse of its square root.