ACPD 14, C11038-C11041, 2015.

Thanks to the reviewer's comments we have improved the papers in many aspects: text, references, and figures. Specifically in light of the comments of reviewer 3, we have spent considerable efforts trying to clarify the aim of the study which is to warn about the misuse of multimodel ensemble and to be more careful prior to infer conclusions out of non-inspected MM ensembles.

Response to Reviewer 1

Lines 41-42: The statement: "An inspected ensemble should always produce a result that is more accurate than the simple average of the multi model results" seems to me as a bit too strong. I can imagine the situation (for example when the models are independent and accurate) that each new model in the ensemble improves, at least slightly, the accuracy.

Response: We have reworded the sentence there

Lines 54-55: "Under this condition, biases of opposite signs cancel out …". In fact independent models can have various biases – all of them can be positive or negative or partially positive/negative. Hence the statement above is not necessarily true.

Response: We have reworded the sentence there

Lines 199-202: As the authors indicated the Talagrand diagram is created by sorting the ensemble results to define bins and counting the number of measurement within each bin. Then in order to have any reasonable statistics the number of measurement should be much greater than the number of ensemble members. Otherwise rank histogram is simply not a proper tool for the analysis and should not be used at all. I suggest to put clearly such a statement.

Response: indeed we use the Talagrand diagram to show that the ensemble is not properly generated, as there are more members than variability to span. We have added the remark in the conclusions section.

Lines 268-269: In principle measurement errors should be also taken into account in the procedure for reducing the ensemble, but in case where they are significantly smaller than the model ones, RMSE is sufficient measure.

Response: We have added the remark in the revised text

Lines 295-297: I agree with the conclusion on the importance of the inspection of the available results prior to their use in further analysis. However, it would be very nice to make this conclusion more practical, for example, by proposing an algorithm for such screening process. In fact the authors described it (lines 139-144) but I suggest to include a diagram that could in clear way show all the steps that should be done in analyzing any ensemble results. Inspection would be a part of this procedure. One of the aspects is that prior to any analysis it is seldom when one knows from scratch which models should be selected for the ensemble. This means that it is better to start with more models, and then to

reduce the ensemble basing on the comparison with measurements. This process, however also depends on what kind of analysis is supposed to be performed i.e. for which purpose the ensemble is created, and which measures or indicators should be applied. That's why I suggest to include a kind of diagram presenting all these elements. The diagram could serve as a starting point for defining good practices in using ensemble methodology in air quality problems.

Response: We feel that such a diagram is a little out of the scope of the present study, which takes one existing example to show the misuse of ensemble modelling, and we think that the one example would not support a 'best practice guideline' for all ensemble applications. Indeed, such a stepwise suggestions as to how ensemble of models should be generated is the focus of a previous paper (Kioutsioukis and Galmarini, 2014) and we wish not to duplicate the conclusions here.

Typographical errors:

Line 62: Potempsky -> Potempski DONE

Line 187: to me -> to be DONE

Response to Reviewer 3

GENERAL COMMENT

Response: There seems to be a relevant misunderstanding here. Possibly, we did not convey rightfully the message in the way we intended to. The main point of our investigation is to put ourselves in the same conditions of the Fetal09 work, thus using the same available information (which were in fact shared by the author) to show that results are different when, few, fundamental principles derived analytically are considered. We have stressed this point in several part of the revised manuscript (see, e.g., beginning of section 2.1).

What we intend to communicate is a more sensible use of ensemble practice and more cautious interpretation of the results. The common practices adopted in the past in ensemble modeling are indeed wrong, as proved by a substantial number studies in the climate community (see, e.g. recommendations advanced by Knutti et al (2010) for the CMIP3 ensemble). What is novel in our study is the application to an ensemble (that of Fetal09) from which results and a journal publication have been produced with no regard to the statistical significance and/or the opportunity to build an ensemble out of the available models. To this respect, we find surprising that criticisms are raised to our work rather than to the simplistic approach presented by Fetal09.

The work we present here is a sort of 'wrapper' that comes after a series of works on the opportunity and risk to build ensemble of models and that allowed us to develop a methodology that is more robust than usual practices in multi model ensemble treatments (Potempski and Galmarini, 2009; Riccio et al., 2012; Solazzo et al., 2012; Solazzo et al., 2013; Kioutsioukis and Galmarini 2014; Solazzo and Galmarini 2014)

We have rephrased the paper to make clearer the scopes of our investigations, although the reviewer #1 did not recommend for any major changes.

"Even if inspecting a multi-model ensemble of this kind is essential for many reasons, the methodology proposed is not robust and is not scientific rigorous (see all the major comments)."

The comment above is somehow arbitrary and not supported by evidence. In the paper we present several previous publications by the authors and several other supporting ones in which it is demonstrated analytically that our conclusions are supported by robust analysis.

MAJOR COMMENTS:

*Line 55: The conclusion about the BIAS of two independent models is false. The two models m1 and m2 can be statistically independent but may have biases that don't cancel out, i.e. the sum of the two model biases: mean(m1)+mean(m2)- 2*mean(observations) may not be equal to zero. This sentence should be reformulated.*

Response. We have reworded the sentence there

Line 56-59: The definition of spread has not been given. Even if it could be considered straight forward it should be specified (see [1]). I guess that the authors refer to the standard deviation about the ensemble mean. Furthermore, the statistical independence between the members doesn't guarantee that the spread is a reliable measure of the model uncertainty. There are many other factors that may influence this skill of the ensemble, such as the number of members and particularly the ability of the ensemble members in reproducing the PDF of the observations. The statistical consistency of an ensemble is verified if an observation being forecast by a dynamical ensemble is statistically indistinguishable from the ensemble members [2]. Line 148-151: A more formal approach should be followed. The definition of variability should be given.

Response. We have removed the sentence there to avoid misinterpretation. The fact that it is not a guarantee is disputable. In fact in principle it is, all the caveats presented are in fact taken into account in our previous publication (Potempski and Galmarini, 2009) where the analytical derivation of the aforementioned properties was proposed. The reviewer seems to miss the point here. We want to show that the methodology adopted by Fetal09 is not suitable for the conclusions reached, as many prerequisites to the analysis have been overlooked. Such scope can be only reached using the exactly the information used in F09. The definition of statistical consistency proposed is very nice and effective but clashes a bit with the reality of the data available for this study and Fetal09 before.

Line 158-187: The whole section lacks of a rigorous formal approach to allow better understanding the procedure, even without reading the others cited papers. How many data are used to compute the covariance matrix? Just using 12 data of the monthly means? If that were the case, the statistical significance of each element of the covariance matrix would be very low. The bootstrap confidence intervals could help in assessing such significance. How the mentioned projection of the so called

"observation anomalies" is used? Because my understanding from the text is that only the explained variance of the first Eigen-vectors is used to draw conclusions on how the ensemble is "wise?

Response. The comment here seems too strong. The papers cited in support of the methodology do answer the questions posed by the reviewer. The explanation of the method is exactly the same as that originally given by of Annan and Hargreaves (2010). The estimate is computed over all eigenvectors (we have added it in the revised text). As for the little data used - we agree, that is indeed the whole point of the work. If we want to reach the scope of demonstrating the inadequacy we have to start from the same premises. The support of observational information is there to support one or the other approach quality.

Line 188-214: The rank histogram, as the authors correctly mention, is meaningful if the number of pairs (forecast, observation) is much larger than the number of the ensemble members. In this paper, the former is 12 the latter is 21. [….].

Response. Sorry to insist, but the scope is to demonstrate that the data presented are too few for deriving scientific conclusions. We have replaced the figure by showing the exact number of bins. For reason of clarity we'd like to keep the discussion as plain as possible, without overcrowding the plots with extra information which, to our opinion, while adding some extra statistics, diverts from the message we want to communicate.

Line 270-272: The statement is not very clear; my understanding is that a better precision (as defined by the authors) also implies a lower RMSE.

Response Accuracy implies precision, the way round is not always true. We have removed the sentence there, anyway.

Table 2 and Figure 2: Which is the statistical significance of the values reported? The "best" models are selected computing the RMSE on 12 data? A bootstrap analysis would probably show several combinations of models exhibiting a RMSE with the same level of statistical significance.

Response. This is a standard practice in air quality, as well as climate modelling (Hanna and Hargraves, 2010; Knutti et al., 2010). What would be the added value of such a test anyway? We are not trying to match any known statistical distribution by our minimization procedure. The level of significance of other combinations might match that of the minimum one, but why the original FetA09 work did not pick any of those and used the full ensemble instead? Again, the point is not how many combinations of models reached the same level of significance: the whole point here is that such combinations exist and that should be contemplated before running prediction scenarios based on the whole ensemble mean.

Line 320-326: Considering what mentioned in the previous comment, how the authors can be sure that the "best" combinations of models will provide the best performances also with a new emission scenario? Especially considering that the numerical models haven't a linear response to a change of

emission scenario. To prove that the best combinations remain the same in different conditions (meteorological or emissions), the data-set should be divided into two parts. One should be used to find the best combinations, the other to verify that the best combinations remain the same.

Response. Since this is the most recurrent question we receive on this aspect, we reply by posing a question. If one takes one model and runs it over a real case study, verifies that the latter manages to capture the observed reality to a certain degree of satisfaction and then changes some conditions like for example the emission (an emission reduction scenario), how come that nobody ever disputes the legitimacy of such a practice? How come that none disputes the legitimacy of having taken bluntly the average of model scenarios and having driven conclusions as done by Fetal09? What is the difference between one model run and verified ensemble like the one we have put together? In our view if the scientific community is prepared to accept the average of modeled scenarios like those presented by Fetal09 as viable, even more the ensemble produced by our analyses should be considered as such, since it is screened, checked and validated according to a more rigorous procedure. If not we should also stop to use models for scenario analysis since nobody can predict the response of a modeling system to conditions different than those for which it was build and whether the latter will be a realistic response. How can we be sure, well we are as much is a single model user or the user of an average of model results pick out at random with no screening. Like the single model user and differently from the poor-man ensemble user, however we rely on the physics that drives the models and the fact that the treatment of the various results has taken into account the original contribution of models to the ensemble rather than redundant information.

MINOR COMMENTS*:*

line 187: to me -> to be DONE

line 190 and -> an DONE

line 318 must be :"described in section 2" line 320 "four monde" ??? DONE