Response to Reviewer 1 Sessions et al.

We thank the reviewer very much for the thorough review. We have done our best to update the paper. Certainly all points of clarification have been made. Some major conceptual changes suggested were not possible however because of the nature of the multi-model ensemble . This work is based on operational models for operational customers, so we are constrained by what operational centers provide. We have done our best to further clarify.

1. <size resolved model request>
ICAP participants and the multi model components are all operationally based. Thus, we need to take them as they come-which is all mass based. Some do have finer mass bins (like fine and coarse mode dust), but here is no evidence in AOT space that these are any better from a bulk score point of view. We choose to not go down this path at this time, as each center has its own reasons for its configuration. Further, with 4 core models and 7 dust models, there is simply not enough information in the time series to explore size resolving issues. There are so many other issues that dominate-like source functions, data assimilation protocols, and scavenging. AEROCOM has been comparing such simulations for years, and there is more to do. The point that we are making is that the multi-model ensemble is better than any single model. That is, you can get forecast improvements (e.g., in means and biases) using a multi-model ensemble vice using a single model.

2.The results in Fig.10 indicate that in general the model creates large biases and RMSE for the cases with AOT>0.6. Have you ever attempted to increase the spatial resolution to see if it makes any differences?
No, as we are attempting to leverage currently available products to produce a better forecast, and those available products are limited by each centers resources. One degree is the current largest common denominator for all models, which is why we chose it. There has been discussion to increase the ICAP MME to half a degree resolution in the future. When all models reach that benchmark, we could redo the experiment. If we could hazard a guess for high AOT events (which we do in the final section), the biggest problem is nonlinearity in the meteorological source functions or boundary layer processes.

3.It would be interesting to show some results on the heavy air pollution events. Compared to the individual models, how does the ensemble model perform in capturing such
Originally we chose not to do pollution specifically but it is not clean cut and deserves a separate paper. But as we show Cape Verde (which is the most clean cut), the reviewer is right that we should show a site where things fail. The two worst sites for verification are Kanpur and Beijing. Kanpur has a much more contiguous data record, so we added a final plot and discussion to that section to show where models all go wrong .

Line 8: what is bulk error statistics? Please give some explanations.
The statistics used are explained in text. Bulk here just refers to them being taken over numerous observations.

The advantage of using multi-model ensemble relative to the individual model is NOT

shown in abstract. Model bias in specific regimes, e.g. biomass burning, have been previously found by many aerosol models. It would be more interesting to highlight what the ICAP-MME model can do better than individual models.

What is highlighted here is that those specific regime bias within the individual models are lessened through utilization of the ensemble as a consensus model. Indeed, the conclusion of the abstract clearly states the benefit of using an ensemble "However, there is an overall AOT low bias among models, particularly for high AOT events. Biomass burning regions have the most diversity in seasonal average AOT. The southern oceans, though low in AOT, nevertheless also have high diversity. In regard to root mean square error, as expected the ICAP-MME placed first over all models worldwide, and was typically first or second in ranking against all models at individual sites. "

This is the first time in the literature that operational models have been compared. While these models often have roots in climate systems, they fundamentally are run differently, and serve different purposes (e.g., use of data assimilation, concentration on events versus seasonal bias). For RMSE, the most common metric for operations (rather than bias for climate models), the ensemble product is clearly better than any single model. We expected this result because of similar findings in other geophysical parameters, but for the first time we have shown it is true for any atmospheric composition variable. Regarding bias, despite the use of data assimilation, there is still a low bias. We thought that interesting.

Introduction:
Line 24: 'as an artifact' what does this mean?
It is an artifact of the radiance observations via SEVERI in Merchant et al 2006, as opposed to the model output from Evans.

Line 6-9: 'while single-model probabilistic ensemble forecasting is clearly enhancing model solutions (particularly in data sparse regions), multi-model ensembles are an ever increasing tool for forecasters'. As you stated that the former can clearly enhance the model solution, why the latter is becoming an increasing tool?
This varies by center, but the first order answer is that single-model ensembles are computationally resource intensive, while ICAP relies upon readily available products that are already budgeted.. Moreover, multi model ensembles have a degree of model independence which improves accuracy. Single model ensembles are useful for some applications (e.g., probabilities), but they still have their own built in biases and lack independence. The strength in different groups developing their own models is that they come up with algorithms that are more independent of each other than members of a single model ensemble.

2.1 input models: I would suggest separating this sub-section into two parts, with one for aerosol models and other for dust-only model, to make the description more clearly.
This has since been moved to an appendix.

Does the same emission inventory is employed in all the aerosol models? It is not clear. Please include the information on the emissions, especially anthropogenic emissions.

The models are largely independent in dust and smoke, but there is some cross usage on pollution. We have added this to the appendix text.

p.14940, line 5: why not simply examine the seasonal mean like JJA, DJF, MAM, and SON?
This is a one year time series, so we are short on data for statistical significance. Even so, the character of DJF+MAM and JJA+SON was not significantly distinguished to divide further. This makes sense for when one looks at aerosol meteorology, and how many aerosol events are in transitional periods, two super seasons based nominally on solstice months (D-M and Jun-N) makes a lot more sense. We have been meaning to write a paper about that-probably will soon!

p.14943, line 15: 'MODS', should be 'MODIS'?
fixed!

sec. 2.3 line 27: 'Instances of cirrus contamination (Chew et al., 2011) were evident in the level 1.5 (and to a lesser extent) level 2 products'. This sentence seems incomplete. Do you want to say 'Instances . . . level 1.5 (. . .) COMPARED TO level 2'?
No, it is present in both. The 'and' in the parentheses should be outside of them to state, "... in the level 1.5 and (to a lesser extent) level 2." They have been removed in favor of commas.

Bottom line in p. 14951: when you removed the top five percent of coarse observations in NA and EA. Is that possible to remove the data in clear sky but with the heavy dust event?
The wording of "in clear sky but with heavy dust event" is unclear to me. If you are asking if it is possible that actual dust events were removed when our clearing assumes that the observations are spurious, then the answer is yes-but we think it unlikely. We did do a hand analysis to make sure this was not the case.

Fig. 2: AOD in Winter/Spring is higher than in Summer/Fall over North Africa, which is contrary to what we expect, any explanations?
It depends on the site. Cape Verde it is slightly higher in summer in AERONET (also figure). What also may being throwing things off conceptually are:

1) Our two season scheme includes the low dust months of Oct-November
2) Statistics when there was QAed AERONET data.
3) Dust events for the second half of 2012 and into 2013 were potentially suppressed by anomalously wet conditions that year (http://www.cpc.ncep.noaa.gov/products/fews/AFR_CLIM/arch/arc2_wam_2012_anom.jpg) while there was a strong May dust season.

Please keep in mind, this is a verification and ensemble paper over a limited period, not a climatology paper.

As for other sites like Ilorin, wintertime frontal activity brings dust to the region. In the summer, winds are more barotropic and easterly and thus the dust does not get that far south. Besides, the ITCZ is parked over it at that time of year. Just check out the AERONET website http://aeronet.gsfc.nasa.gov/cgi-bin/type_one_station_opera_v2_new

Ilorin is more dusty and smoky in Dec-April. This said, we always worry about cirrus contamination in this part of the world. We hope the next version of AERONET will be able to detect it.

Table 1 and 2: It would be good to have a map plot showing the locations of these sites.
Site map added as Figure 1b.

Fig. 5 and top line on p.14957: bias in Baegnyeonng improves with time, why? Does that imply there is problematic in the analysis?
Analysis issues are potentially the causes. (p14957, L3-5)

Table 4 and line 13 on p.14957: what metric shown in Table 4? biases or RMSE?
We made a typo in the header, not the caption. These are RMSE values.

Fig. 6 and line 14 on p.14957: small dots are each model's value. What value? AOT?
These are model AOT values, with marker face indicated which model.

Line 3-5 on p. 'On average, the RMSE's of the 1 day forecasts of ICAP-MME run approximately 50% of the climatological mean. Dust AOT forecasting is superior to overall fine and coarse mode 5 AOT, running approximately 1/3rd of climatological AOT.' How do you come up to this conclusion? Please show more details.
We have modified the language a bit for clarity. "Based on the slope of RMSE against mean AOT value for each site in Figure 7, the RMSE's of the 1 day forecasts of ICAP-MME run approximately 50% of the climatological mean AOT value. Dust AOT forecasting is superior to overall fine and coarse mode AOT, running approximately $1/3^{rd}$ of climatological AOT. Again, this is part reflects the importance of the dust species by centers. Further, the AERONET Cape Verde site (in which RMSE is particularly skillful) is a common benchmark site for Saharan dust-hence models are typically tuned for the region.**"**

Line 13 on p.14958: please give the definition of fractional gross error.
Fixed.

Line 20 on p.14958: 'although the dominance of the ICAP-MME generally increases in time, particularly for dust'. What do you mean?
We changed the verbage a bit for clarification: "Like bias, forecasting skill for all models and the ensemble mean degrades in time. Although the relative performance of the ICAP-MME mean relative to the member models increases in time, particularly for dust."

Fig.9. how is the rank and histogram computed here? Did you put all the results from the individual models and the ensemble together for this calculation? Please add more details on how you obtain the results in order to help the readers to understand.
We added a bit more to the explanation. It is the result of taking all the member models at a valid time for a particular observation, then ranking that observation as if it were another member.

Line 20 on p.14966: if 'Experience has shown however that equal weighting in a con-

sensus style appears to provide the most robust results overall, and this is backed up on both practical and theoretical grounds (DelSole et al., 2013)', Then why 'we intend to convert the ICAP-MME to a super ensemble where models are weighted by their scores (e.g., Krishnamurti et al., 1999; Casanova and Ahrens, 2009)'. It is confusing here. Also, can you give some explanations on why equal weighting is most robust?

You have cut off the important first half of the "we intend" sentence. The statement is that we receive questions about future intent, including whether or not to produce a weighted super-ensemble, which we answer by stating it would be less robust. The short answer is that while there have been many expectations from super ensembles, in reality they rarely stand the test of time as they require prior knowledge of error characteristics. This is difficult to obtain as a) models are always improving, b) even "bad" models can result in an improved consensus, and c) one training data set often is not applicable to another season. But at some point we need to address the issue more directly. For now, we clarified this statement.

Response to Reviewer 2 Sessions et al.
We appreciate the time and thought into performing this review. Response to your comments are below.

Primary Notes:

Improve Abstract
We removed the first several sentences. We believe they are covered in the introduction and we agree that it makes the abstract more readable. We did not remove the 'unnecessary details.' That part refers to the omission of BSC and UKMO from the analysis, but they should be kept in the paper/abstract as a political nicety.

Place model details in appendix
Made an appendix.

Secondary Notes:
P14939, L4: modal -> model

P14939, L5: No action. Data distribution method is in flux.

P14940, L24: ICAp -> ICAP

P14941-14948, sections 2.1.1-2.1.7: Fixed

P14949, L15: How did you initialize the model at the beginning of the forecast? Do all participating models (whose results are shown in this paper) have data assimilation system for aerosol properties? How different are the AOT fields from individual models at +0h?

Each model has its own method of data assimilation or model initialization as is discussed in Appendix A. Differences in the models at +0h can surely be ascribed to these different initializations. However, we are not able to plot the analysis fields at t-0 as the MACC model does not have a 0 hour analysis of AOT due to the nature of its 4D var assimilations scheme. AOT for the MACC model is a diagnostic, while the analysis is performed in aerosol mixing ratio. This may be changed in the future through post-processing the aerosol mixing ratio at t=0 to obtain the corresponding AOD. This lack of AOT at t=0 for MACC is stated in 14952 L10.

P14949, L17: suggest to separate Fig.1a from the rest of panels in Fig1, since they are not related.
Split figures, relabeled b-e to a-d, fixed in text.

P14950, L9-10: what is the size cutoff for the fine and coarse modes?
There are no “size cuts” in the models, as we break out species by their fine (sulfate, biomass burning) or coarse mode (dust, sea salt) nature. For those models with multiple size bins they are integrated by specie. Fortunately, the SDA method for partitioning fina and coarse mode AOT accounts for the tails of the distributions, thus we have an apples to apples comparison here.

P14951, L9: please define "gross fractional error".
Gross fractional error -> fractional gross error, added equations.

P14952, L15-16: please check the grammar here.
Removed 'Although/do,' separated clause.

P14952, paragraph 2: do you use 6h mean values or instantaneous values?
As in all forecast models, the X hour forecast is the instantaneous value at that time. We accept AERONET comparisons to +/- 3 hours of this time however. This is now stated more clearly.

P14952, L28-30: This sentence is not clear to me.
Sentence removed. While accurate, it isn't necessary to explain rank histograms.

fP14956, L7-8: Where do they have lower biases?
Added another reference to table 1. The table is referenced prior in relation to biases, but was at least a page or so back.

P14956, L14: Why at some sites the model biases are even smaller as forecast day increases? Does this indicate these sites are mostly affected by local/nearby sources and less affected by meteorology and the aerosol transport from remote areas?
This is answered as part of the response to the following comment. But it does happen for some locations and parameters that sometimes even for meteorological forecasts to be more accurate day5 than day 2. This is likely connected to compensating model errors, and it's not necessarily an analysis problem. I am not sure it means that the sites are more affected by local sources, and less by transport

P14957, L3-4: why this implicates biases in the analysis? Meteorological analysis?
The reviewer has a point that we need to be clear on language. We state that the bias is bigger earlier in the forecast cycle as a statistical outcome of our comparison.. This immediately implies the analysis is biased. Thus stating "perhaps…" is repetitive and confusing. Thus we have reworded this to "This could implicate bias in the analysis, as the free running forecasts relax into lower error states before being erroneously jarred into high error by the assimilation process."

P14959, L12: How reliable is the AERONET dust AOT data? And which number in fig.9 is the ensemble mean?
Please see P14951 paragraph 2 for information regarding Level1.5 and Level 2 errors and reliability. AERONET verification for the coarse mode and AOT is of great to concern to us. While 90% of the time, the stated AERONET verification standards hold, there are some circumstances when severe cloud contamination occurs. It is hoped that version 3 will correct a number of these issues. Figure 9 is the rank histogram, so all values represent number of times the observation obtained 'ranks' in relation to the members + ens. No number in the figure represents the ensemble mean, instead the bin heights represent the number of times the observations fell into a particular rank amongst the member models.