Reply to comments by T. Sekiyama:

Thank you Thomas for the thoughtful review. Below are our responses to your comments.

1. I was surprised that the adaptive inflation worked well for aerosol in this study because my adaptive inflation failed and diverged when I used a method other than Anderson 2009. I have thought that the adaptive inflation for aerosol is unstable due to a large uncertainty of aerosol modeling compared to NWP. What do you think?

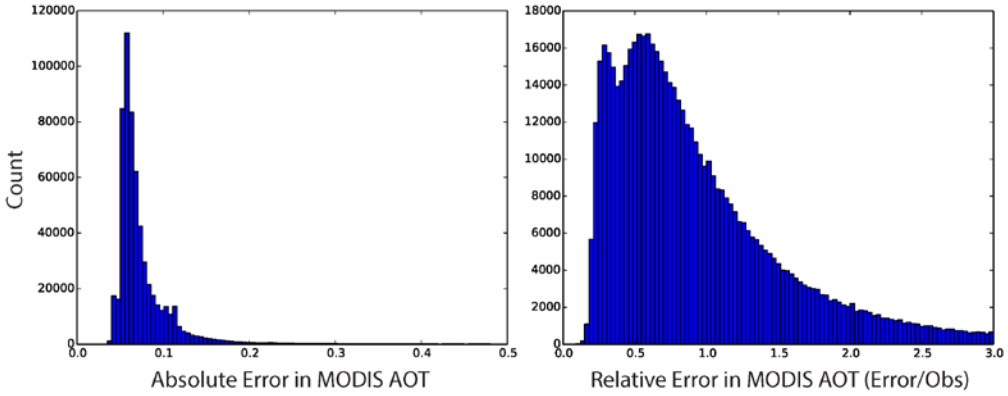
   Response: The adaptive inflation worked quite well (ie. stable) with the exception of regions impacted by fires. Without any measures to control inflation in these regions, the adaptive inflation did in fact blow-up with inflation factors exceeding 10. Eventually, unrealistic aerosol concentrations were produced and the simulations crashed. This behavior of the adaptive inflation algorithm for fire-impacted regions indicates that there is an inconsistency between the observational and the background distribution in optical thickness. Fire emissions have very large uncertainties and we thought were the likely drivers of the inconsistencies generating the unstable growth in the inflation factor since large and persistent fires were occurring during the simulation time period. In order to create stability in the simulations, we tuned the standard deviation of the inflation factor and defined a maximum inflation factor (1.5). However, we think that doing some tuning to the smoke emissions in the future would allow for the adaptive inflation to run without a maximum inflation needed. These stability problems were discussed in the results section (page 28085, lines 7-27).

2. It was not described in this paper how (and how much) the observation errors were estimated. Even though it is described in the references, the estimation method and size of observation errors are crucial for data assimilation. It is better to show the validity of the method (and error size) in the manuscript, if possible. Generally speaking, "observation errors" are underestimated because it is difficult to estimate spatial representativeness and remote-sensing bias. I am afraid that observation errors are unnaturally underestimated in this paper too.

   Response: Yes, we agree that the observational errors are a crucial component of data assimilation. The observational error estimates are based on long-term comparisons of MODIS Terra and Aqua AOT to AERONET AOT for over-ocean (Zhang and Reid, 2006, 2009) and over-land (Hyer et al. 2011). The observational error covariances are treated as diagonal matrices, so no accounting for correlated errors. We can add more discussion on this in the paper. In this study, since we are using the current operational system as a baseline for comparison, we wanted to assimilate the exact same product as is used in the NAVDAS-AOD system, so we made no changes to how the observational error is represented. However, we can include some additional plots in the supplementary material to show what the observational error looks like. In future work, we may reevaluate the observational error.

   Manuscript change: The NAAPS/NAVDAS-AOD simulations are run with a 1 degree resolution and assimilate the same MODIS AOT observational dataset **with the same observational errors**

(Zhang et al. 2005; Zhang and Reid, 2006, 2009; Hyer et al. 2011; Shi et al. 2011) for consistency.



3. All the emissions in this study were perturbed using the same factor for a given ensemble member. Actually, it is not a good way as the authors mentioned. Instead of that, there are some alternative techniques to reduce correlations between independent sources. For example, make perturbation factors one-by-one randomly each grid-point, 2) smooth out the distribution of the factors using a 3-dimensional smoothing filter, 3) and use the smoothed factors to perturb emission sources. Usually, the ensemble mean of the perturbed emission flux is not very shifted by this method.

   Response: In this study, the same perturbation factor is applied for a given ensemble member for each source type. As an example, smoke emissions for ensemble member n are all perturbed with the same randomly produced perturbation factor. This essentially creates an infinite correlation lengthscale for smoke emissions that is only limited by the localization lengthscale. However, for ensemble member n, the perturbation factor for smoke, dust, sea salt, and anthropogenic and biogenic fine are not the same. Thus, given our localization of X, we do have a regional smoothing parameter in a way built in. We preferred this methodology to the moving Gaussian method in that method predefines the maximum length scales. Here we wanted to see what naturally and reasonably covarried, and then look at how those covariance fields looked. We will make these points more clear in the manuscript. But, as noted in the manuscript (page 28079, lines 20-21), we did initially try grid-by-grid source perturbations as you suggested. We found this had no impact on ensemble spread, therefore, ruled this method out. Indeed, the strategy used in this work for perturbing source functions worked well when the emission correlation lengthscale is greater than the localization lengthscale (ie. large spatially correlated aerosol events). For source types in which the emission correlation lengthscale is less than the localization lengthscale (ie. spatially independent sources such as small boreal forest fires), we plan to test a perturbation function as you suggested. We think this should provide substantial improvement in some problem regions identified in the manuscript (ie. Eastern united states, North American boreal regions). This problem, and the point you made above regarding adaptive inflation, are of course intertwined with your comment 1.

4. The authors are using a maximum inflation limit, but I am afraid that the maximum value (=1.30) is too small because adaptive inflation factors more than 2 or 3 are acceptable for NWP without any problem.

   Response: The maximum inflation used in this work is actually 1.50 (page 28085, line 20). When we tested the free-running adaptive inflation without any constraints on the maximum inflation, the maximum inflation did not exceed 1.5 with the exception of fire-impacted regions where the adaptive inflation became unstable. We believe this instability is due to the persistent nature of the fires during the simulation time period and an inconsistency between the background (dominated by emissions for these large fire events) and the observations. This was the value for which we found the adaptive inflation was stable for fire regions; however, we think with some tuning of the fire emissions, we can let the adaptive inflation algorithm run freely without a maximum inflation constraint. This is work planned for future studies.

5. The authors say, "the ensemble isn't fully representing the distribution with an excess of observations occurring of low ranks," but when the rank histogram shows a one-side peak, it is only certain that the ensemble members have a large bias. With only the information of "one-side peak," we don't know whether the ensemble spread is small or not.

   Response: When the majority of observations are below the lowest bin, this indicates bias in the ensemble relative to the observations, just as you stated. Yes, we agree that since the observations are mostly below the ensemble members, you can't state much about the actual spread of the ensemble relative to the observational spread. What we meant with this statement is that the ensemble members aren't capturing the low AOT values of the observed distribution. We will revise this statement to make it clearer.

   Manuscript change: The Eurasian Boreal smoke region rank histogram, consistent with the evaluation of the total spread to RMSE ratio, shows that the ensemble isn't **capturing low AOT values of the observed distribution**, with an excess of observations occurring for low ranks.

6. If AOT values are small, it's no wonder AOT observational errors are relatively large because the error of remote sensing is almost independent from the AOT (=retrieved) value. On the other hand, when AOT values are small, it's impossible to make a large ensemble spread. This is a disadvantage of ensemble data assimilation.

   Response: Yes, we agree with your statement. Since we have a positive-definite state variable, the ensemble spread can only be so large for small AOT values. The result is that the observational error is much greater than the forecast error and the assimilation would weight the analysis mostly to the background. So if there is a bias present, the assimilation won't be able to correct for this. We will include the fact that this is a limitation of ensemble data assimilation as you mentioned.

   Manuscript change: At the lower end of the AOT distribution (< 0.1), the total spread (combined ensemble spread and observational error) exceeds the RMSE; however, it is found that the observational error dominates the total spread (Figure 7). This relationship is consistent across the experimental ENAAPS-DART configurations, represented by the different colors in Figure 7. It indicates that the observational error is too large **relative to the ensemble spread** for small AOT values, with similar results found for other fire-impacted regions (South America, Southern

Hemisphere Atlantic). **This relationship is likely caused by the ensemble spread being too small for small AOT values since aerosol mass is a positive-definite quantity.** For data assimilation, this translates to a reduced impact of the observation on the model state for small AOT.

7. Do the authors mean that there is a large difference between meteorological analyses since there are few meteorological observations in the Southern Ocean? If so, this sentence (Line 18-20) is a little confusing.

   Response: NAAPS and ENAAPS are offline, so here we are assimilating only aerosol-related observations. There are no AOT observations being assimilated in the Southern Ocean. Between deterministic NAAPS and ensemble NAAPS (ENAAPS), the only difference is the data assimilation system and the meteorology fields (deterministic NOGAPS and the ensemble NOGAPS fields) they are run on. Since there are no AOT observations being assimilated in the Southern Ocean, any differences between the NAAPS/NAVDAS-AOD simulation and the ENAAPS-DART simulation are due to differences in the meteorology fields used to drive the simulations. For example, sea salt emissions are parameterized as a function of wind speed. Differences in wind speed between deterministic and ensemble meteorology fields would impact sea salt emissions and therefore, the optical thickness in the region. Likewise, differences in humidity fields would impact the optical thickness. We will add to the discussion in the manuscript to make this point more clear.

   Manuscript change: Since there are very few **AOT** observations for assimilation in the Southern Ocean, any differences in this region are attributed to differences in the deterministic and ensemble meteorology fields **(winds, humidity) that drive the models**. **For example, differences in wind would impact sea salt emissions and therefore, optical thickness in the region. Likewise, differences in humidity fields would impact the optical thickness.**

8. The authors often use the term “variational” (assimilation, system, initial condition, etc.) as an inferior method to the EAKF, but the “variational” method is the 2D Var in this paper. We have another variational method, the 4D Var, which is comparable or superior to the EAKF. It is better to always use the term “2D Var” in this paper to avoid confusion

   Response: Yes, we agree with your statement and will update the manuscript to be clear that we refer to 2D Var and not all variational methods.

9. I could not understand the meaning of “the optimal combined meteorology and source ensemble”. What is optimal?

   Response: Here we meant that the combined source and meteorology ensemble performed better than source-perturbed or meteorology ensemble alone and was the chosen approach. We will revise this statement to say the chosen configuration instead of optimal.

   Manuscript change: The example, shown in Figure 15, shows the analysis increments for the NAVDAS-AOD 2DVar system as well as analysis increments for ENAAPS-DART, both for the source only and the combined meteorology and source ensemble.

10. In this study, the EAKF system captures sharp gradients while the 2D Var smooths plume distributions. However, the EAKF and 2D Var have similar RMSE and bias. That means, probably, although the EAKF result looks realistic, the plume location is slightly shifted from the real one. It is difficult to judge which is better "sharp but slightly-shifted plumes" or "blunt but broadly-covering plumes" as operational prediction/warning, I think.

    Response: Yes, we agree that it is difficult to define which is better in this instance depending on what the application of the forecast is (ie. Smoothed out event would give a larger warning region). However, we think the real advantage of the ensemble approach is that we can produce more realistic corrections to the state fields (which produce sharper gradients that are consistent with what is seen from satellite) which will become more important as additional observational information is introduced into the system, such as Lidar and other spatially limited pieces of information.

    Manuscript change: On the other hand, the **2DVar** system produces a dust plume feature that is smoothed out. This dust case demonstrates a major advantage of the EAKF system over the 2dVar in its ability to spread information in a realistic manner and as a result, capture sharp gradients**. It is anticipated that the ability of the EAKF to produce more realistic corrections to the state field will become more important as additional observational information is introduced into the system, such as Lidar and other spatially limited pieces of information.**

11. Are these RMSE global?

    Response: Yes, these are global.

    Manuscript change: The 24-hour forecast **global** RMSE against AERONET AOT with bootstrapped 95% confidence intervals

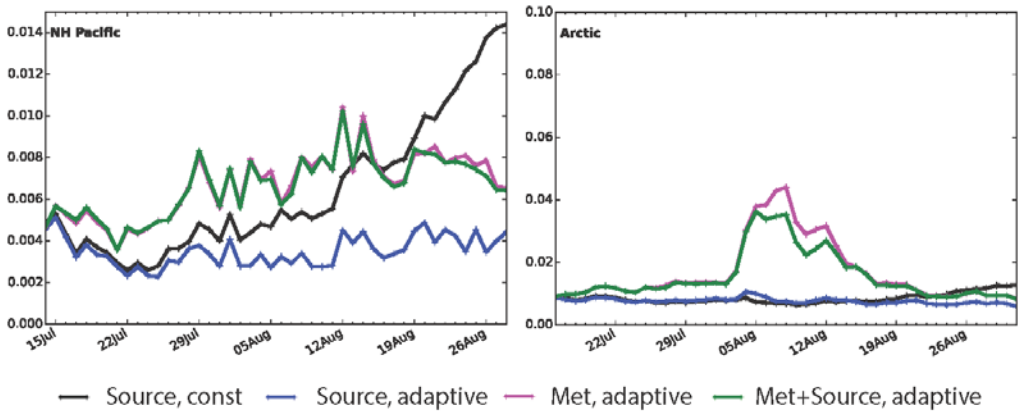
12. The authors say, "the observational error may be too large for small AOT values, which could also contribute to the positive bias", but I don't think so. Generally speaking, it is extremely difficult to assimilate zero or almost zero values like small AOT. It is because a population that contains a lot of zeros (or almost zeros) and is not allowed to be negative values (e.g., radar-measured precipitation) is not Gaussian-distributed. Fundamentally, it is nonsense to quantify the error of non-Gaussian-distributed values using a standard deviation. However, data assimilation assumes everything Gaussian. It is the reason why zero-value assimilation is difficult. The positive bias observed in smoke regions may be relevant to non-Gaussian AOT distribution and irrelevant to the size of AOT observational error.

    Response: Thank you for your input on this. We agree with your statement and will add discussion on this issue. There has been discussion here on to what extent this is a real problem, and what is the best way to cope with, ranging from complex transforms to something simple, like assimilate in log space.

    Manuscript change: The discussion and conclusion were consolidated, but we have changed our discussion in the manuscript to talk about dealing with small AOT values (such as changes to comment 6 above) as well as in the conclusions.

13. I am very interested in NH Pacific Ocean, Arctic, and Antarctic.

    Response: We can add additional plots to the supplementary material for these regions.



14. Why did the authors plot MODIS AOT that was not quality-controlled? I would like to see the comparison between assimilated observations (= quality-controlled AOT) and assimilation results.

    Response: We were trying to show the sharp-gradient present in the MODIS AOT observations. This can be seen pretty clearly when all AOT values are shown, however, we added an additional plot with assimilated AOT only.

15. I am very interested in why the AOT over the Sahara is largely changed by the 2DVar although there is almost no observation over the Sahara. The influence radius of observations in 2D Var is only 250 km or so, right?

    Response: The radius of influence for the variational system is determined through an exponential function as defined in Zhang et al. 2008. If R is the distance between observation and background location and L is the defined 200km lengthscale, the function is (1+R/L)*exp(-R/L). An influence can be present beyond the defined 200km lengthscale; however, the impact will decrease with distance.

16. Page 28080, Line 16: The description “over 13 land regions” is actually “over 15 land regions”?

    Response: Thank you, we updated this to 15.

    Manuscript change: The experimental 6-hour AOT forecasts are evaluated over **15** land regions as indicated in Figure 1 as well as six ocean regions, including the northern and southern hemisphere Pacific and Atlantic Oceans, the Indian and the Southern Ocean.

17. Page 28085, Line 11: There are two spellings “blow-up” and “blowup” in this manuscript. Choose either one

Manuscript change: We changed this to “blow up”. Thank you.

18. Page 28090, Line 18: Is it necessary after “large” to put a comma?

    Response: I think with multiple adjectives, we need to separate them with a comma. We could be wrong though.

19. Page 28092, Line 22: isn’t -> is not

    Manuscript change: The Eurasian Boreal smoke region rank histogram, consistent with the evaluation of the total spread to RMSE ratio, shows that the ensemble **is not** capturing low AOT values in the observed distribution, with an excess of observations occurring for low ranks.

20. Page 28093, Lines 10 and 12, etc.:There are two expressions “meteorology ensemble” and “NOGAPS ensemble” in this manuscript. Choose either one.

    Manuscript change: These were changed to either “meteorology ensemble” or “NOGAPS meteorology ensemble”. The NOGAPS is included at times to be specific about where the meteorology fields come from.

21. Page 28093, Line 29: There are two spellings “source-perturbed” and “source perturbed” in this manuscript. Choose either one

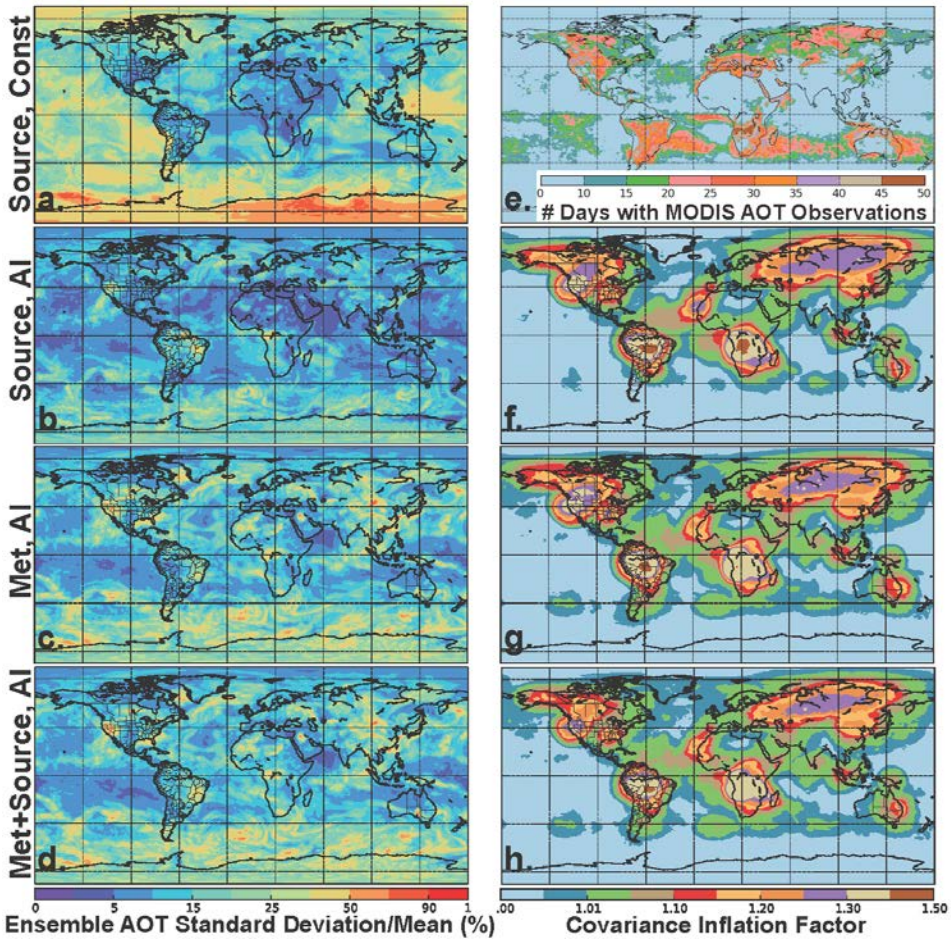
    Manuscript change: These were all updated to “source-perturbed”. Thanks.

22. Page 28094, Line 27: Putting “(Table 2)” at the end of this sentence, it becomes easy understandable.

    Response: This paragraph is talking about the evaluation of the posterior relative to AERONET AOT. Table 2 is the evaluation of the prior to MODIS AOT.

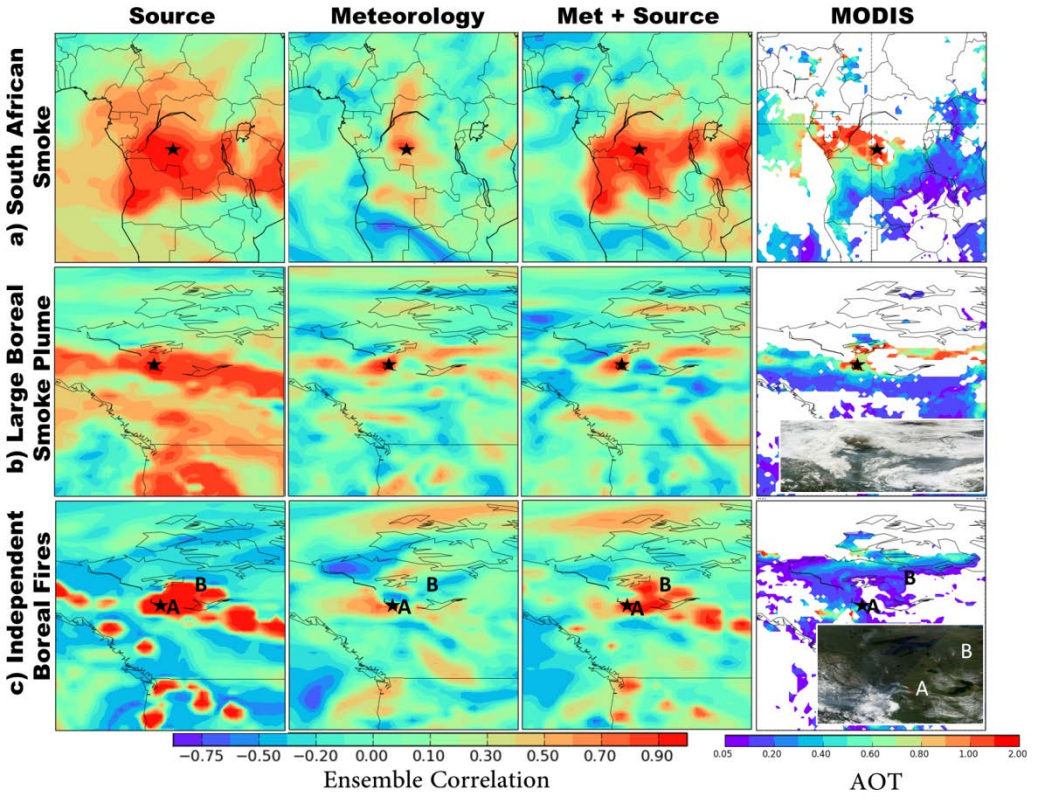
23. Figure 4: The characters “a” “h” in the figure panels are too small and extremely unreadable

    Manuscript change: This figure was updated with larger font.



24. Figure 6: It is very difficult to find a “point”, especially in (b) panels

Manuscript change: This figure was updated with larger black stars.



25. Figure 11: Some of the AOT observation plots are illegible, especially on 22 August

    Response: In the AERONET AOT timeseries plot, there aren't any observations on August 22.

26. Caption of Figures 11 and 12: The "analysis" is plotted here, I think. But the caption says, "predicted total AOT".

    Response: Yes, you are correct. We changed the caption to be more specific. Thank you.

    Manuscript change: Timeseries of **analysis** total AOT (grey

27. Figure 16: It is hard to find the area where the MODIS plot indicates, at a glance

    Response: We were trying to zoom in to the leading edge to show how sharp the gradient is. We will update the caption to make it clearer.

    Manuscript change: A **zoomed in** MODIS true color image of the leading edge of the dust plume is also shown as well as MODIS AOT (550nm) observations.