Response to Anonymous Referee #1

Thank you for your thorough comments. Please see our responses:

1. First, no results are shown for a standard run of the model without assimilation. Hence the improvement due to assimilation is unknown and the differences between various assimilation setups cannot be properly judged.

   Response: The goal of our study was to see how the new ensemble system performs relative to the current operational prediction system (NAAPS with NAVDAS-AOD); as a result, this was considered our control. (page 28075, lines 5-8)"NAAPS with the NAVDAS-AOD data assimilation has been fully operational at FNMOC since 2010. The operational system serves as a member of the International Cooperative for Aerosol Prediction (ICAP) multi-model ensemble (Sessions et al. 2015) and is the baseline for comparison in this work." Subsequent papers will show more detailed comparison of the different methods relative to a no DA control.

2. Second, no proper attempt at filter tuning is done. In particular, ensemble size and localisation length-scale are not systematically varied and their effects studied. In this respect Fig 9 is slightly worrying: panel d (which shows differences between a 20 and 80 member run) shows similar or larger differences than the sensitivity experiments for a 20-member ensemble (a,b and c).

   Response: There was a lot of tuning that went into setting up ENAAPS-DART, we will work to make this point more clear in the manuscript. With regards to localization, several tests were run, but the results were not presented in the paper. This is discussed on page 28078 (lines 26-28)-28079 (lines 1-2). What we found with regards to localization tests was that the 1000km lengthscale performed the best. Since these results were consistent with previously published studies (Schutgens et al. 2010), we didn't feel showing these additional tests introduced anything new and would just add to an already long paper. Instead, we wanted to focus on the experiments that introduced something new to aerosol data assimilation. That is why we chose to focus on looking at constant versus adaptive inflation as well as looking at methods for generating the ensemble members. We felt these experiments were both informative and introduced something new. With regards to ensemble size, we chose 20 members because this is the size that is run operationally out to 6 days, and hence is our basis set. There are some resource limitations in place for running a system operationally that we cannot control. However, we wanted to show one test of what an increase in ensemble size could buy, and a limited time period enhanced run was acquired (single 80 member ensemble run, page 28095 (lines 1-23)). As we expected, you can get a big payoff with increasing ensemble size because we are likely doing a lot better in capturing a realistic background error covariance. I'm unsure of why the scatterplot for 20 versus 80 members is troubling, except that it shows there is a lot of room for improvement in future development of this system. With our resources, the 20 member system will serve as the base system with potential for moving to larger ensemble sizes in the future. As we mentioned in the paper, we have plans for future studies in ensemble size as well as model resolution (page 28095, lines 18-24; page 28105, lines 23-26). It should be noted that the optimization tests were conducted on the 20 member ensemble, and therefore, things like localization are not necessarily optimal for the 80 member (less localization would be needed for the larger ensemble size). We don't have the resources for these optimization runs at the moment, but expect to do more work on this in the future.

Manuscript changes: To make it clear that the optimization tests were conducted on the 20-member ensemble, we made changes to the following sentences:
It should be noted that several initial tuning experiments were conducted **with the 20 member ensemble** in which a range of constant inflation factors were tested, in a similar fashion to Schutgens et al. (2010b).

Several length scales were tested in initial tuning runs **of the 20 member ensemble** and a length scale of 1000km is selected for use in this work.

It should be noted that the single 80-member simulation uses the same localization lengthscale as the 20-member ensemble. Optimization of the 80-member ensemble was not conducted due to resource limitations and will be evaluated in future work.

3. The authors at times generalize too much from their own (limited set of) experiments: while the possible problem due to constant inflation is worth mention and analysis, no other authors have come across this and it is possible this is entirely due to very a specific system (ENAAPS-DART).

   Response: We would argue that this finding is most likely not system specific. For idealized experiments and NWP applications, similar findings with regards to constant and a varying inflation were identified. This was mentioned in the manuscript on page 28087 (lines 6-9).

   While this has never been directly discussed for aerosol applications, there have been hints to this issue. For example, Schutgens et al. (2010) ran sensitivity studies for a one month simulation (July 2005) for aerosol assimilation of AOT. One of the sensitivity experiments conducted was varying the inflation factor for a constant multiplicative inflation. They found instabilities developing for an inflation factor of 1.20 and 1.30 where unrealistic aerosol mass mixing ratios developed. This result was for a short one-month simulation, so the instability can be seen for large inflation factors. We suspect that if the simulation was run out for a longer time period, issues would have developed for smaller constant inflation factors as well. However, we will change the strength of the wording to indicate that we suspect this result is applicable to other systems.

   Manuscript changes: **Based on the results in this work,** an adaptive covariance inflation is recommended over a spatially and temporally uniform covariance inflation. The adaptive approach overcomes instability issues that arise due to spatially heterogeneous observations with the constant inflation approach **and it is expected the same finding will apply to other systems.**

4. The relative importance of source vs meteorology perturbation is hard to assess given that source perturbations are always generated with a 25% spread. This uncertainty seems optimistic at hourly and gridbox scales.

Response: We agree that the 25% uncertainty applied to the source perturbations might be optimistic as we know emissions can be highly uncertain, especially for boreal fires. However, the system behavior indicates that regardless of the perturbation applied, the spatial impact (or lack therefore) of the data assimilation using only a source-perturbed ensemble would be the same. By perturbing the sources for smoke as an example, the impact on the system is to create large correlations at all distances between smoke emissions, only limited by the localization lengthscale. While increasing the source-perturbations would increase the size of the analysis increment, it wouldn't impact the area of influence (ie. near source regions). The same problem of not being able to impact aerosol transport events (away from source-regions) as discussed in the manuscript would hold for a source-only ensemble. While we know that some changes need to be made to how the source perturbations are generated as discussed in the manuscript (page 28091, lines 13-14; page 28092, lines 13-14;page 28104, lines 16-17), our conclusion of needing both source perturbations for data assimilation near-source regions and meteorology ensemble for transport events would hold.

5. Sometimes there are quite lengthy descriptions of results, region by region, while the same results are efficiently summarised in Figures and Tables. Maybe the authors can try to make their text more concise

   Response: Thank you, we will work to make the text more concise.

6. Apparently inconsistent acronyms: AOT and NAVDAS-AOD

   Response: Aerosol optical thickness (AOT) is the more appropriate term to use for aerosol extinction in the vertical, therefore, we choose to use AOT instead of AOD throughout the manuscript. Since its development, the variational data assimilation system used with NAAPS (ie. NAVDAS-AOD) has always been referred to in this manner (Zhang et al. 2008); therefore, we choose not to change the legacy name of this system.

7. The paper by Schwartz et al JGR 2014 deserves mention as it also compares 3D-VAR and ensemble Kalman filter methods for aerosol assimilation.

   Response: We agree and will add this reference to our manuscript.

   Manuscript changes: For aerosol applications, a number of data assimilation methodologies have been tested both regionally and globally and shown to improve model performance (Collins et al. 2001; Yu et al 2003; Generoso et al. 2007; Adhikary et al. 2008; Zhang et al. 2008; Benedetti et al. 2009; Schutgens et al. 2010a,b, Zhang et al. 2011, **Schwartz et al. 2012**, Rubin et al. 2014, Sekiyama et al. 2010).

8. Introduction: a major advantage of ensemble DA systems over others is the relative ease of implementation and maintenance, especially in view of the fact that many aerosol and aerosol-cloud processes can be modelled in different ways

   Response: Thank you, we will add this point to the introduction.

   Manuscript changes: Finally, ensemble systems provide an opportunity to apply Ensemble Kalman Filter (EnKF) data assimilation technologies **which are relatively easy to implement** and

allow for flow-dependent corrections to the predicted state fields (Evensen, 1994; Houtekamer and Mitchell, 1998).

9. p 28073, l 13: "In order to increase understanding of forecast uncertainty and aerosol forecasting dependencies on underlying meteorology, a 1 resolution, 20 member ensemble version of NAAPS (ENAAPS) was created". The exact meaning eludes me. Does this refer to a one-off experiment or is it an on-going activity? What was learned from this?

   Response: As an initial exploration of forecast uncertainty, an ensemble version of NAAPS driven purely by the NOGAPS or NAVGEM meteorology ensemble was created. Forecasts using ENAAPS were initially run off of the analysis fields from the NAVDAS-AOD data assimilation system and were available on the NRL aerosol webpage. However, we wanted to take full advantage of the ensemble and set up ENAAPS forecasts to be initialized with analysis field from an ensemble data assimilation system, the focus of this work. We will clarify this point in the introduction. Thanks.

   Manuscript change:) **As an initial exploration of** aerosol forecast uncertainty and its dependencies on underlying meteorology, a 1 degree resolution, 20-member ensemble version of NAAPS (ENAAPS) **driven by the NOGAPS or NAVGEM meteorology ensemble** was created. **Forecasts using ENAAPS were initially run off of the analysis fields from the NAVDAS-AOD data assimilation system.** Encouraged by successes **using aerosol EnKF data assimilation** within an NWP framework (e.g., Sekiyama et al., 2010; Schutgens et al., 2010a,b ; Pagowski and Grell, 2012; Khade et al., 2013), **here we investigate the use of ENAAPS for operational aerosol forecasting** purposes by replacing the NAVDAS-AOD data assimilation system with the NCAR Data Assimilation Research Testbed (DART) implementation of an EnKF. **This system is referred to as the ENAAPS-DART system.**

10. p 28074, l 17: "a brief synopsis is provided here, noting a few key differences". While I agree with this level of detail, I think the text might be clearer in specifying what are the differences. E.g. "Likewise, the sea salt source is dynamic in nature with emissions as a function of surface wind speed (Witek et al., 2007)." suggests there are no differences wrt seasalt so why mention it? It doesn't help that a brief (and necessary) explanation of basic aerosol description is interjected ("A combined anthropogenic and biogenic fine aerosol species (ABF) is represented in the model which accounts for a combined sulfate, primary organic aerosol and a first order approximation of secondary organic aerosol."). I suggest to reorganise this in two paragraphs: the first a very brief overview of essential NAAPS characteristics (e.g. basic aerosol description + emission datasets and parametrisations), the second the key differences of the version used in this paper

    Response: Thank you for your feedback on this. We will edit the description of NAAPS and ENAAPS to make it clearer.

    Manuscript changes: A thorough description of basic NAAPS characteristics can be found in Witek et al., (2007) and Reid et al., (2009), but a brief synopsis is provided here, **including a few key differences between the NAAPS implementation used in this work and the literature**. Smoke emissions from biomass burning are derived from satellite-based thermal anomaly data used to construct smoke source functions via the Fire Locating and Modeling of burning Emissions-FLAMBE database (Reid et al. 2009; Hyer et al. 2013). However**, for simulations**

conducted in this work, a version of **FLAMBE that derives smoke emissions from MODIS thermal anomaly data only** is used, **consistent with the NAAPS decadal reanalysis (Lynch et al. 2015)**. Dust is emitted dynamically as a function of friction velocity, surface wetness, and surface erodibility using NAAPS standard friction velocity to the fourth power method, but with the erodibility map of Ginoux et al. 2001. The sea salt aerosol source is dynamic in nature with emissions as a function of surface wind speed **as described in Witek et al. 2007**. A combined anthropogenic and biogenic fine aerosol species (ABF) is represented in NAAPS which accounts for a combined sulfate, primary organic aerosol and a first order approximation of secondary organic aerosol. Anthropogenic emissions come from the ECMWF MACC inventory (Lamarque et al. 2010). The **Navy's** current operational aerosol forecasting system **uses NAAPS** coupled to a 2-dimensional variational (2dVAR) data assimilation system (NAVDAS-AOD, Zhang et al. 2008; 2014) for **assimilating** AOT retrievals (Zhang et al. 2005; Zhang and Reid, 2006, 2009; Hyer et al. 2011; Shi et al. 2011) to produce forecast initial conditions every 6 hours.

11. What is meant by a MODIS-only version? FLAMBE is completely ignored? Or only MODIS data are used for a specific FLAMBE version?

    Response: Here we are using a version of FLAMBE that only uses MODIS data. We use this version of FLAMBE as it is used in the NAAPS decadal reanalysis which serves as an internal benchmark. We will clarify this point in the description.

    Manuscript change: However, **for simulations conducted in this work**, a version of **FLAMBE that derives smoke emissions from MODIS thermal anomaly data only** is used, **consistent with the NAAPS decadal reanalysis (Lynch et al. 2015).**

12. ENAAPS is in principle independent of (aerosol) assimilation, no? So the "exception of data assimilation" is a bit confusing. The distinction between 'deterministic' and 'ensemble' meteorology fields is also confusing. I'm guessing this is in-house jargon? The ensemble meteorology fields are also the result of deterministic models. How is this ensemble produced (e.g. what is perturbed, a very brief description of McLay et al would be good)? What does "truncated to 1 degree" mean (is NOGAPS a spectral grid model)? Why match the deterministic (!) NAAPS reanalysis? It will beused with ENAAPS, not?

    Response: Yes, ENAAPS can be independent from data assimilation. Here we are referring to the ENAAPS-DART system versus the NAAPS/NAVDAS-AOD system. We will change ENAAPS to ENAAPS-DART and NAAPS to NAAPS/NAVDAS-AOD to make this point clear. With respect to the meteorology fields, we are referring to a single set of meteorology fields produced from the deterministic model for the 'deterministic' fields. For the ensemble meteorology fields, these are produced using an ensemble transform to perturb the analysis fields (wind, temperature, specific humidity, and surface pressure) as discussed in McLay et al. (2010). Yes, NOGAPS is a spectral model with a higher resolution than ENAAPS, therefore, the NOGAPS output is truncated to produce a one-degree resolution output for the ENAAPS simulations. We chose to match the 1 degree resolution used here in the ENAAPS-DART base system with the NAAPS reanalysis to have aerosol product lines that can be easily compared. However, as mentioned in the manuscript, we plan to do additional studies on model resolution.

    Manuscript changes: With the exception of data assimilation (Section 2.2), the architecture of **ENAAPS-DART** is very similar to the deterministic version of **NAAPS/NAVDAS-AOD**. The model

physical parameterizations are the same. However, instead of deterministic NOGAPS meteorology fields, NOGAPS ensemble meteorology fields are used. The NOGAPS ensemble meteorology fields (20 member) are produced operationally at FNMOC at 0.5 degree resolution out to six days. These fields are produced by perturbing initial conditions **(wind, temperature, specific humidity, and surface pressure)** using an ensemble transform method as discussed in McLay et al. (2010). For ENAAPS, all twenty NOGAPS ensemble members are used **for driving the model simulations**, truncated to 1 degree to match the deterministic NAAPS reanalysis **(Lynch et al. 2015)**.

13. "requires a priori assumptions". It can be argued that ensemble DA methods also require a-priori assumptions on the model forecast error, in that they assume a-priori uncertainties in meteorlogy and emissions and from that calculate the ensemble forecast.

    Response: Yes, this is true that there are assumptions in ensemble data assimilation about Gaussian distributions etc. This is certainly a limitation of ensemble data assimilation. Here we are trying to make the point that the error covariance is produced a priori and is static. While the ensemble covariance will of course not be perfect, it provides a means for allowing the uncertainty to vary with time and with processes that occur in the model simulations. We will be more specific in making this point. Thank you.

    Manuscript changes: The variational approach, which is used in the current NAVDAS-AOD system, **uses a static model forecast error**.

14. p 28076, l 4: "is considered to be a random draw from the probability distribution of the model's state given all previously used observations." This sentence completely ignores a-priori error sources in the ensemble, even though they are the essence of the system

    Response: This is the premise of ensemble prediction systems and the formulation of EnKF is based on this principle. While the analytical theory is based on this, ensemble DA systems have been found to work well even when these assumptions are violated. In particular, ensembles have been found to work well with heavily biased model forecasts when using the adaptive inflation (Anderson 2009).

15. p 28076, l 5: "The use of ensembles to sample the error allows the error to evolve non-linearly in time with the flow-dependent covariances between different state components determining how observations impact the ensemble estimate" Shouldn't there be a comma after 'in time'?

    Response: Thank you, we will add a comma.

16. p 28076, l 17: It is not entirely clear how EAKF and DART relate? EAKF is part of DART, and I think it is the only ensemble DA in DART. What does DART offer beyond EAKF?

    Response: EAKF is one of the filter options available in DART. There are several different filter types including an EnKF, Kernel filter, Particle filter and several other options as described in DART documentation. We will make this point more clearly in the text.

Manuscript change: DART has been successfully applied to a host of meteorological and atmospheric composition data assimilation problems (e.g., Arellano et al. 2007, Khade et al., 2012, Raeder et al. 2012 , Hacker et al. 2013 and many more) **and provides the option to interface to a number of different filter types, including EAKF, EnKF, kernel and particle filters**.

17. p 28076, l 20-25: Apparently DART does not include an observation operator H, but uses ENAAPS calculations of AOT. As AOT will depend on humidity (which will be different in different ENAAPS members), doesn't this imply that the effective observation operator used in DART is non-linear instead of the linear operator assumed in a Kalman filter? (That is: across the ensemble, AOT cannot be generated from a form like H x, with x the aerosol state vector and H a matrix).

    Response: This is up to the person implementing DART on whether they want to use an observation operator that acts on the state variables as they are read into DART or as done here, apply an observation operator outside of DART. Yes, there are nonlinearities due to humidity, which does vary between ensemble members. This is always an issue with data assimilation. However, DART applies forward operators sequentially, so arbitrary nonlinear h are trivial to implement.

18. p 28077, l 27: Why usually in the prior? Won't this distort any covariances that have been built up during the short-term forecast? Can't it be applied to the posterior? I thought that was the more common way to use inflation.

    Response: Priors that are unrealistically confident result in the observations having insufficient weight in the data assimilation update and over time, lead to filter divergence. Because of this, the covariance inflation is typically applied to the prior (Anderson and Anderson, 1999) and this is especially the case for EnKF systems for weather prediction. However, the inflation can be applied to the posterior as well (ie. Whitaker and Hamill, 2012). The inflation increases the spread about the mean, so it doesn't impact the sample correlations between components.

19. p 28079, l 3-5: "The effectiveness of the ensemble data assimilation system is highly dependent on having sufficient spread in the ensemble members in order for the observations to impact the model forecast." This suggests that the biggest issue is to have as large a spread as possible. I would argue instead that the spread should be an indication of forecast uncertainty (both know uncertainties, ie meteorology and emissions and unknown uncertainties, e.g. due to model errors).

    Response: Note that we aren't saying that we want the most spread possible here, we are saying there must be <u>sufficient</u> spread. This means we need adequate or enough spread (ie. to represent the uncertainty). Often times with ensembles, they are spread deficient which can lead to filter divergence and the observations won't have an impact. Here we are saying we want sufficient spread that represents the system. The adaptive inflation algorithm used in this work is designed to try and make the spread consistent with the RMSE as you suggest.

20. p 28079, l 5-15: Maybe the generation of the emission ensemble should be discussed before the inflation/localization? The latter are after all solutions to limitations in the first.

Response: We reordered this section to make it clearer. Thanks.

21. p 28079, l 13: Why 25% and not 10 or 100%? For sea-salt and dust, arguably perturbing emitted particle size/windspeeds can be just as important?

    Response: The impact of wind speed on sea salt and dust emissions is accounted for when the meteorology ensemble is used (ie. through differences in the wind fields across the ensemble members). While 25% uncertainty may be optimistic as discussed in a response to a previous comment, we thought this was a good first estimate of the source-perturbation. A means for evaluating if this is sufficient is to look at whether or not the system as a whole has enough spread. This is done in this work by evaluating how the pooled spread (combined ensemble spread and observational error) compare to the RMSE of the prior relative to the observations. These should be approximately the same if the system is well tuned. What we found is that the system was pretty well tuned with the exception of fire-impacted regions with not enough spread for high AOT events. This indicates we don't have enough spread and we need to potentially change how the fire emissions are represented in the ensemble (page 28101, lines 27-29). This could be done by increasing the source perturbations to the fire emissions (page 28092, lines 13-15). So in conclusion, we selected a conservative perturbation for the sources and based on the results from this study, have recommendations on how to move forward and improve the system.

22. p 28080, l 6: It would be good to have a brief explanation how rank histograms are created and what their purpose is? They are not a standard test in aerosol ensemble DA (but possibly should be).

    Response: Yes, we can add a few sentences to better explain the purpose of the rank histogram and how it is generated.

    Manuscript change: (page 28076, lines 34-36) The first method is through examination of the prior 6-hour forecast against MODIS AOT observations, before assimilation occurs, using diagnostics such as RMSE, bias, ensemble and total spread, number of assimilated observations, and rank histograms. **Rank histograms are generated by repeatedly tallying the rank of the observation relative to values from the ensemble sorted from lowest to highest and can be used for diagnosing errors in the mean and spread of the ensemble forecast (Hamill 2001).**

23. p 28080, l 8: Why is the prior a stronger indication of assimilation? I guess because they show how well a previous analysis pulled the system to the truth. An analysis will agree (fairly) well with observations by construction. Still, a bit more explanation or references are welcome. Do your data actually bear this out: i.e. does the prior show stronger signal to variation in experimental setup than the posterior? This would be very interesting to show.

    Response: It is much harder to compare MODIS AOT observations to the posterior AOT because they are no longer independent. It has been assimilated and therefore, you would expect better agreement. Here we are saying to use the 6-hour forecast AOT (ie. Prior) and compare that against MODIS AOT before assimilation. This gives us an indication if the model is doing a better job in predicting the state relative to the observations (before they are combined) and provides a means for evaluating how well the system is doing in representing forecast uncertainty. This is

common practice in evaluating a data assimilation system. This section was updated to clarify the points being made.

Manuscript change: The performance of the **2-month** experimental simulations is evaluated in several ways. The first method is through examination of the prior 6-hour forecast against MODIS AOT observations, before assimilation occurs, using diagnostics such as RMSE, bias, ensemble and total spread, number of assimilated observations, and rank histograms. **Rank histograms are generated by repeatedly tallying the rank of the observation relative to values from the ensemble sorted from lowest to highest and can be used for diagnosing errors in the mean and spread of the ensemble forecast (Hamill 2001).** In order to account for the effect of observation error in the rank histograms, the forecast values are randomly perturbed for each ensemble members by the observation error (Anderson 1996, Hamill, 2001, Saetra et al. 2004). The focus of this observation-space evaluation **relative to MODIS AOT** is on the prior since this is a stronger indicator of how the assimilation is impacting the model predictions. Benchmarks of a good ensemble system include stability in ensemble spread, an RMSE that is small and comparable to the total spread, and rank histograms that indicate an ensemble distribution that is consistent with the observations (Anderson 1996). Since aerosol composition and characteristics are variable depending on the type of aerosol sources and the location-dependent processes that impact transport, transformation, and lifetime, the diagnostics are evaluated regionally. The experimental 6-hour AOT forecasts are evaluated over 13 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. **Additionally, it is important to evaluate the posterior fields since these serve as forecast initial conditions. The assimilation posterior fields** are examined relative to ground-based 550 nm AOT fields based on NASA AErosol RObotic NETwork (AERONET) observations (Holben et al. 1998; O'Neill et al., 2003) **since these observations are not assimilated and therefore, can be used** as an independent evaluation of the data assimilation **analysis fields**.

24. p 28081, l 8: Maybe change "incorporate" to "assimilate"?

    Response: Ok, thanks.

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

25. p 28082, l 1: So which ENAAPS-DART assimilation experiment is shown here? What has been perturbed here? Has the system been optimised or not (inflation/localization)? What is the purpose of this Section? If it is to show global aerosol features, isn't this better shown during the comparison with NAAPS/NAVDAS? It might be clearer to first discuss the optimization experiments and only then discuss the global features seen in the best setup.

    Response: This result is for the meteorology and source perturbed ensemble with adaptive inflation. The purpose of this section was to present what aerosol features are being predicted during this time period so that they can be discussed in evaluating the system optimization as well as during the comparison between the deterministic and ensemble systems. We will work to make this clearer.

Manuscript change: Average ENAAPS-DART AOT fields **(Met+Source, adaptive**) for the Boreal Spring (April, May) and Boreal Summer (June-September), 2013 are shown in Figure 2.

26. p 28084, l 13: Why now the posterior AOT? Earlier you argued that the prior AOT should be used for comparison against observations.

    Response: Evaluating the prior is a good way to evaluate and diagnose the performance of the system relative to the observations that will be assimilated (MODIS AOT). This provides a means for evaluating how well we are doing in representing forecast uncertainty in our system and the overall health of the system. It is fair game to evaluate the posterior AOT against independent observations (AERONET) that are not assimilated. This is also an important evaluation since the posterior serves as initial conditions for our aerosol forecasts. We use both methods of evaluating the system performance in this work. This is discussed in the methods section on page 28080 (lines 1 through 21). The section describing diagnostics was updated as shown in response to comment 23 to clarify.

27. p 28084, l 17: Higher dust AOT is probably due to some higher windspeeds in the meteorology ensemble and the threshold windspeed for dust emission? What drives the increased AOT over wildfires?

    Response: Yes, the higher dust AOT is due to the introduction of different wind speeds across the ensemble members with the inclusion of the meteorology ensemble. For fire-impacted regions, the model generally produces a positive bias. With more spread in the simulations that include the meteorology ensemble, the observations have more weight in the analysis and the AOT is reduced.

28. p 28085, l 11: This is an interesting discussion of the role of inflation. It seems to me that the discrepancy between prior and observations is due to either: 1) observational biases; 2) model biases. A Kalman filter assumes that both are unbiased. Your results suggests that adaptive inflation serves to camouflage such biases (unless they become too big and the syetm crashes). This warrants some discussion by the authors.

    Response: Inflation is one of several means used to help overcome errors in ensemble systems. While it is one method for improving system performance, careful evaluation of how the algorithm behaves is also a means for better understanding the system and in ways that it can be improved. Case in point is the example you pointed out on page 28085, line 11. This is an issue that indicates a potential problem with the model as you suggested and in particular, fire-dominated regions. There were several issues related to smoke-dominated regions highlighted and the case is made throughout the manuscript (page 28092, lines 12-18; page 28096, lines 23-28) that issues in smoke-dominated regions indicate a need for re-tuning of the smoke emissions which we expect would alleviate the problems seen in the adaptive inflation algorithm for the Eurasian Boreal fire impacted region. One of our major concluding points is that work needs to be done in smoke-dominated regions to improve the system (page 28104, line 25-26 to page 28105, line 1).

29. p 28085, l 18: prior of inflation equals its posterior from previous cycle: this is also known as persistence modelling.

Response: Yes, we agree with you. In our implementation of adaptive inflation, a damping factor of 0.9 is applied to the posterior from the previous cycle to produce the prior for the next cycle (page 28085, line 21). So the damping is the time variation model for the inflation.

30. p 28086, l 2: "issues occur with the constant covariance inflation where there is limited observational coverage". See my previous comment, I believe this could be equally due to biases in observations or models than coverage.

    Response: Covariance inflation does help overcome underrepresented variance in the ensemble due to model bias and sampling error caused by the small ensemble size. We certainly agree that model bias will vary with location and time and therefore, an inflation factor at one location might not be appropriate at another. This can certainly be an issue with constant covariance inflation. However, when you have a non-uniform observing network, the result of applying a uniform inflation is that you end up with unreasonable solutions in regions that have limited observations (ie. Southern Ocean) because the ensemble is continuously inflated and there are no observations to constrain the state fields. This is a bigger issue in these under-observed regions because it can lead to the simulation crashing. This is the point we are making here.

31. p 28086, l 8: "the normalized standard deviation", that is: 1 ? Ah, Figure 4 suggests it is normalised by the mean. Please indicate this in the text as well.

    Response: Thank you, we will do that.

    Manuscript change: If the observation density is compared to the prior ensemble spread, represented as the standard deviation of the ensemble AOT **normalized by the mean**, at the end of the constant inflation experiment (Figure 4a), it is apparent that large spread develops where there is limited observational information, including high latitudes and spots over the Pacific Ocean.

32. p 280866, l 22: "The growth in spread in the Southern Pacific Ocean for the constant inflation experiment is a result of having continuous inflation with no observations to bring the ensemble back to reality". I think it is important here to note that this may be a feature solely found in DART-EAKF. To my knowledge, no other studies (e.g. Sekiyama et al, Schutgens et al, Dai et al) have found this growing ensemble spread. It may be related to the fact that in DART, inflation is applied 1) to the prior; 2) even when there is no reason for inflation (i.e. when there are no observations). p 28087, l 2: "Although spatially and temporally constant covariance inflation has been the chosen method for aerosol applications in the past, it is not recommended since aerosol observations are spatially heterogeneous. On the other hand, adaptive inflation increases ensemble spread where there is observational information available,producing stability, a desirable characteristic for an ensemble system". This statement is far too bold with little evidence to back it up. Your analysis suggests this to be true for DART-EAKF but as I said before, it hasn't be noticed by other authors. I suggest rephrasing this to something like: "It is suggested that particular attention is paid to the temporal evolution of ensemble spread in case a constant inflation factor is used, because our results suggest."

Response: There is a lot of evidence of this occurring in atmospheric data assimilation and hints of this in aerosol data assimilation as discussed in the response to comment 3. This is related to inflation without observations as previously discussed and is not specific to DART EAKF.

33. p 28087, l7: "These findings are consistent with idealized experiments and NWP applications of ensemble systems where a temporally and spatially varying inflation is recommended over a constant inflation approach (Anderson, 2009; Li et al., 2009; Miyoshi et al., 2011)." Obviously there are other reasons why AI may be preferential to a constant inflation factor. I believe the listed authors discuss the issue of model biases that are effectively dealt with by AI. Note that model biases are really the bane of DA and AI is essentially a way to sweep them under the carpet (or conversely: a way of studying them by tracking the evolution of the inflation factor).

    Response: For example, in Li et al. 2009: "we have used a globally uniform inflation factor, which is clearly not a good assumption in reality where the observations are non-uniformly distributed. With a spatially dependent inflation, we may be able to better deal with an irregularly observing network". Likewise, in Anderson 2009: "A more serious problem occurs when a single value of inflation is not appropriate for all state variables. Assimilation of in situ observations, like radiosonde and aircraft observations, in a global numerical weather prediction model provides an example. In densely observed regions like the upper troposphere over North America, ensemble variance can be inappropriately small due to model bias and sampling error. Inflation can reduce this problem. However, over the Southern Ocean, there are very few observations to constrain the model. Repeated application of inflation values large enough to correct problems over North America can systematically increase the variance of the ensemble over the Southern Ocean. Eventually, this can lead to values that are inconsistent with climatological values, and in the worst case, incompatible with the model's numerical methods. The result is ridiculous solutions, at best, and model failure, at worst. "

34. p 28087, l 23: "In particular, a large increase in spread is found at dust source regions." Presumably because of the windspeed threshold for dust emission? How much bigger than 25% is the spread?

    Response: With the meteorology ensemble, we now have different wind speeds associated with each ensemble member. This produces different amounts of dust for each ensemble member since dust emissions are a function of wind speed, therefore, increasing the ensemble spread in these regions. "In particular, a large increase in spread is found at dust source regions. For example, the spread increases from approximately 20 to 50 % in the Northern Arabian Peninsula" page 28087, lines 21-23.

35. p 28088, l 1: "the meteorology ensemble increases spread for sea salt aerosol" Seasalt emission is presumably not governed by a windspeed threshold, although it will have a non-linear dependence on windspeed. Is this effect therefore larger for dust than seasalt?

Response: We see a pretty good increase in spread for both dust and sea salt. The increase in spread is determined by how much the wind speed varies across the ensemble for a particular region and at a particular time of interest and how that difference across the ensemble translates to sources via the emission function. For dust, the emissions are represented as the surface friction velocity to the fourth power. For sea salt, the emissions are a function of the 10 meter wind speed raised to the 3.41 as described in Witek et al. 2007.

36. p 28088 l 5: "The meteorology ensemble appears to be the main driver of ensemble spread." It may be good to remind the reader that you have assumed a 25% uncertainty in emissions. I find this rather low especially because this is uncertainty on short time-scales (hourly, daily). Already at longer time-scales (months, year) Granier et al 2011 and Huneeus et al. 2011 find larger uncertainties over large regions.

    Response: We will add this point to the above sentence.

    Manuscript change: The meteorology ensemble appears to be the main driver of ensemble spread **when included with a 25% source-perturbed ensemble**.

37. p 28088, l 15: Regarding stabilization of ensemble spread, this is not obvious for WCONUS

    Response: It's hard to see in this region because there are large wildfires impacting WCONUS during the end of the simulation period. With larger AOT being produced due to the fires, the ensemble spread will increase as well. However, for longer simulations that have been conducted, we see no problems in this region with stabilization.

38. p 28088, l 20: I suggest using brackets instead of commas to delineate "the square root of the sum of the ensemble variance and the observational error variance"

    Response: Ok, thank you. We will change this.

    Manuscript change: A good means for determining how well the ensemble system represents uncertainty is a comparison of the prior total spread **(**the square root of the sum of the ensemble variance and the observational error variance**)** in AOT to the prior RMSE.

39. p 28092, l 3: couldn't this be due to insufficient ensemble spread at low AOT? Several authors have pointed out that a positive variable like AOT can only have a large spread at small values if the distribution is allowed to be very skewed (i.e. non-Gaussian, contradicting a basic assumption in a Kalman filter). The small spreads that occur in ensemble runs are a direct result of small source perturbation at low mean source values. I believe this is an unresolved issue.

    Response: Yes, we agree on this point and will include this in our discussion of the results.

    Manuscript change: 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.

40. p 28092, l 7: The case of too small a spread at high AOT may also be the result of missing causes of uncertainty. E.g. you don't perturb deposition processes. Perturbing them will have a bigger impact at high AOT than at low AOT because (again) AOT cannot go below zero.

    Response: Yes, we agree that not having enough spread means that we aren't capturing all the uncertainties.

    Manuscript change: For the case of large AOT in the North American Boreal for example, there is not enough spread and the uncertainty is underrepresented for all ENAAPS-DART experiments (Figure 7). **This may be the result of not using large enough source perturbations for smoke or the result of not accounting for uncertainties in physical processes such as deposition. However**, other regions impacted by summertime burning events such as South America, the Southern Hemisphere Atlantic Ocean (Figure 7), the Eurasian Boreal region, and the Western United States also have a tendency to underrepresent uncertainty for large AOT events. Smoke emissions have very large errors; often as large as an order of magnitude uncertainty (Reid et al. 2009, 2013; Hyer et al., 2013). As a result, a larger source perturbation (greater than the 25% standard deviation currently applied) for smoke emissions **is likely** needed to produce a better tuned system.

41. p 28093, l 17: "since they are independent." The prior and the observations are also independent so this cannot be the reason to choose the posterior.

    Response: Please see the response to comments 23 and 26.

42. p 29095, l 7: "performance gains" The authors are undoubtably aware that this comes at a hefty cost: 4x more CPU requirements. I think that 'performance' may not be the best word here as it implicitly suggests some optimal cost/benefit ratio.

    Response: Thank you, we will reword this statement.

    Manuscript change: Initial results show that **further reductions in RMSE** can be achieved by increasing the ensemble number at most AERONET sites, including Beijing in East Asia and many Eastern US, North African, European/Mediterranean, and Boreal sites (Figure 9d).

43. p 28096, l 1: It would be good at this stage to point out that NAVDAS-AOD does not include perturbed meteorology (as far as I understand it). I.e. something like Fig 10 is unlikely to be seen for NAVDAS-AOD

    Response: NAVDAS-AOD and NAAPS can't have a perturbed meteorology because it is a deterministic simulation.

44. Sect 3.3 & 3.4 and Table 3 etc: an evaluation of a base model run (control) should be part of this analysis. Is there even a substantial improvement in AOT due to assimilation (either 3DVAR or EAKF)?

Response: Please see our response to comment 1 on this topic. Our focus in this work is to see how the ENAAPS-DART system performs relative to the current operational system which serves as our baseline. Including DA does produce an improvement in AOT. We have subsequent work that will show this in more detail.

45. p 28100, l 14-19: "On the other hand, the forecasts initialized with the EAKF fields do a better job capturing the leading edge of the dust front with the ENAAPS-DART version being smoother than the deterministic counterpart along the dust front. This demonstrates that the sharpness achieved in the ensemble data assimilation propagates in the forecast and is an advantage of using the EAKF initial conditions over the variational initial conditions for the short-term forecast." The use of 'sharpness' and 'smooth' confused me initially. Unless I am mistaken, they are not juxtaposed but describe different aspects. Consider repharsing this sentence.

    Response: Yes, we can see where your confusion comes from in this statement. We were referring to different aspect of the predicted dust front, which makes it confusing. We will reword this discussion to make it clearer. Thanks.

    Manuscript change: Both of the forecasts initialized with the 2dVAR fields capture the event, but like the analysis fields, don't capture the **sharp gradient** as seen in the MODIS image. On the other hand, the forecasts initialized with the EAKF fields do a better job capturing the **AOT gradient** at the leading edge of the dust front. This demonstrates that **the sharp gradient** achieved in the ensemble data assimilation propagates in the forecast. This is an advantage of using the EAKF initial conditions over the variational initial conditions for the short-term forecast.

46. p 28100, l 5-19: I think it should be pointed out that a substantial part of the plume (eg the northern edge) is missed by all four forecasts. Please discuss possible causes.

    Response: Since this is consistent across all forecasts, this is likely attributed by model physics which is consistent across these configurations.

    Manuscript change: The MODIS visible image and MODIS AOT for the dust case is also included and shows a narrow band of high optical thickness at the leading edge of the dust front. **All four configurations produce the dust plume, although the Northern portion of the plume is missing for all cases. The missing portion of the plume is likely attributed to the model physics since this is consistent in NAAPS and ENAAPS.**

47. Section 4, Discussion: I suggest removing this Section in its entirety. It is not really a discussion but an extended summary. Its main points have already been discussed (in detail) in the main text. Important conclusions in this Discussion that are not yet in the Summary should be moved there and phrased more consisely.

    Response: Thank you, we will rework the discussion.

48. Section 5, Summary: consider my general comments.

Response: Ok

49. Fig 6: Not quite clear what is shown here. This is essentially the model forecast covariance? So it is with respect to a single location? Presumably the black dot in the top row (there are no dots in the lower rows)? It is the correlation in the AOT fields?

    Response: This is the spatial correlation in the prior AOT relative to a point indicated by the black star. This is meant to show how observational information will spread in different configurations of the ENAAPS-DART system. The figure caption will be updated and the size of the black star is now increased in the figure.

    Manuscript change: Figure 6. Ensemble correlation fields **in the prior AOT relative to a point indicated by a black star** for three different aerosol events:

50. Fig 15: What does "Not all available MODIS observations are assimilated" refer to? I realise that the NRL-MODIS dataset is a subset of the official Col 5 product. But why show here a different product than that which you have assimilated?

    Response: In this figure, we were trying to show the sharp gradient in the dust front that is produced in the ENAAPS-DART system is also seen in MODIS observations. You can see this clearly when you look at all the observations. That is why we included this figure, however, we will include an additional plot of just the assimilated observations (which are a subset of what has been shown already).

References:

Anderson, J. L. and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Mon. Weather Rev., 127, 2741–2758, 1999.

Li, H., Kalnay, E., and T. Miyoshi, T.: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, Q. J. Roy. Meteor. Soc., 135, 523–533, 2009.

McLay, J. G., Bishop, C. H., and Reynolds, C. A.: A local formulation of the ensemble transform (ET) analysis perturbation scheme. The ensemble-transform scheme adapted for the generation of stochastic forecast perturbations, Weather Forecast., 25, 985–993, 2010.

Miyoshi, T.: The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform kalman filter, Mon. Weather Rev., 139, 1519–1535, doi:10.1175/2010MWR3570.1, 2011.

Schutgens, N. A. J., Miyoshi, T., Takemura, T., and Nakajima, T.: Sensitivity tests for an ensemble Kalman filter for aerosol assimilation, Atmos. Chem. Phys., 10, 6583–6600, doi:10.5194/acp-10-6583-2010, 2010

Whitaker, J.S., and Hamill, T.M.: Evaluating Methods to Account for System Errors in Ensemble Data Assimilation. Monthly Weather Review, Volume 140, pp 3078-3089, 2012.

Witek, M., Flatau, P. J., Quinn, P. K., and Westphal, D. L.: Global sea-salt modeling: resultsand validation against multi-campaign shipboard measurements, J. Geophys. Res., 112, D08215, doi:10.1029/2006JD007779, 2007.

Zhang, J., Reid, J. S., Westphal, D. L., Baker, N. L., and Hyer, E. J.: A system for operational aerosol optical depth data assimilation over global oceans, J. Geophys. Res., 113, D10208, doi: 10.1029/2007JD009065, 2008.