Response to review comments on acp-2015-586 from reviewer 2

The original comments are provided in black, our response is given below each comment in red.

Thank you for the careful reading of our manuscript and your review.

This paper evaluates one year of a high-resolution (i.e. 12 km grid-spacing) WRF-Chem simulation over North America with observations from MODIS Aqua and Terra as well as the ground networks AERONET and EPA. The remotely sensed observations include both AOT and AE. The authors collocate the simulated data to remotely sensed data and analyse the resulting spatial patterns on monthly and yearly time-scales. The topic of the paper is entirely in line with the interests of ACP, and so publication in ACP is possible. There appears to be a serious issue though with the remotely sensed data used in the analysis: MODIS and AERONET agree even less with each other than MODIS and WRF-Chem or AERONET and WRF-Chem (Table 3, AOT column). This suggests that at least one of these remotely sensed datasets is flawed and not appropriate for the evaluation of WRF-Chem. The authors merely list this statistic but draw no conclusions from it or offer explanations of it. This issue really needs to be resolved before publication.

Thank you for your positive assessment. We have addressed the issue with the remotely sensed data in the comments below and in the manuscript.

General comments

While model evaluation with observations is very important, it is difficult to see what this paper adds besides a lot of statistics. In particular, the authors barely explore two interesting datasets: the EPA data and the Delaware gridded precip data. Some interesting questions come out of this study and addressing them might give the paper a bigger impact:

- does the model agreement with observations depend on scale? What are the length- and time-scales in the different datasets anyway? Does the model agree better after further aggregating the data over, say, 24, 48, 96 km? (Note that while pollution forecasts require spatio-temporally highly resolved simulations, forcing estimates probably can do with spatio-temporal averages)

Thanks for the useful comment. Using very limited data, prior research indicated mesoscale variability (horizontal scales of 40–400 km and temporal scales of 2–48 h) is a common and perhaps universal feature of lower-tropospheric aerosol light extinction [*Anderson et al.*, 2003]. However, to our knowledge, no prior systematic attempt has been made to quantify and test the universality of aerosol scales of coherence over the contiguous US. We have conducted some additional analyses to test the dependence of MFB on the spatio-temporal scales by aggregating the 12km grid cells (both from WRF and MODIS) to coarser resolutions (see Figure 6). When looking at monthly aggregated data we only see a slight variation of MFB during cold months when the 12km data are aggregated to a coarser resolution, possibly indicating that those months are more sensitive to biases in the chemical composition, mostly associated with underestimation of sulfate aerosols (see response to reviewer 3) and possibly also as a result of the lower data availability.

Reference:

Anderson, T. L., Charlson, R. J., Winker, D. M., Ogren, J. A., and Holmen, K.: Mesoscale variations of tropospheric aerosols, Journal of the Atmospheric Sciences, 60, 119-136, 10.1175/1520-0469(2003)060<0119:MVOTA>2.0.CO;2, 2003.

We added the following text:
"Using very limited data, prior research indicated mesoscale variability (horizontal scales of 40–400 km and temporal scales of 2–48 h) is a common and perhaps universal feature of lower-tropospheric aerosol light extinction [Anderson et al., 2003]. However, we are not aware of prior systematic attempts to quantify and test the universality of AOD scales of coherence over the contiguous US. To test the sensitivity of the MFB in simulated AOD to spatial aggregation, we excluded the first 12 cells to the left and to the top of the simulated domain and averaged the remaining 12×12 km grid cells over the following scales: 24×24, 36×36, 48×48, 72×72, 96×96, 144×144, 192×192, 216×216, 288×288, 384×384, 432×432, 576×576, 864×864, 1152×1152, 1728×1728, 3456×3456 km. The last spatial average corresponds to a single grid cell encompassing the entire domain (excluding the outer 12 cells located to the West and North of the simulation domain). Each spatial average at a coarser resolution is computed as the mean of all valid 12×12 km grid cells within the averaging area. We then computed the MFB for the regridded WRF-Chem and MODIS data pair and found that, on a yearly basis, MFB is highest at 12km (0.14 for Aqua and 0.15 for Terra) and reaches a first minimum at 72 km for Aqua (MFB=0.13) and 384 km for Terra (MFB=0.13) (see Fig. 6). However, the MFB and hence systematic error in AOD relative to MODIS exhibits only a weak dependence on the level of spatial aggregation."

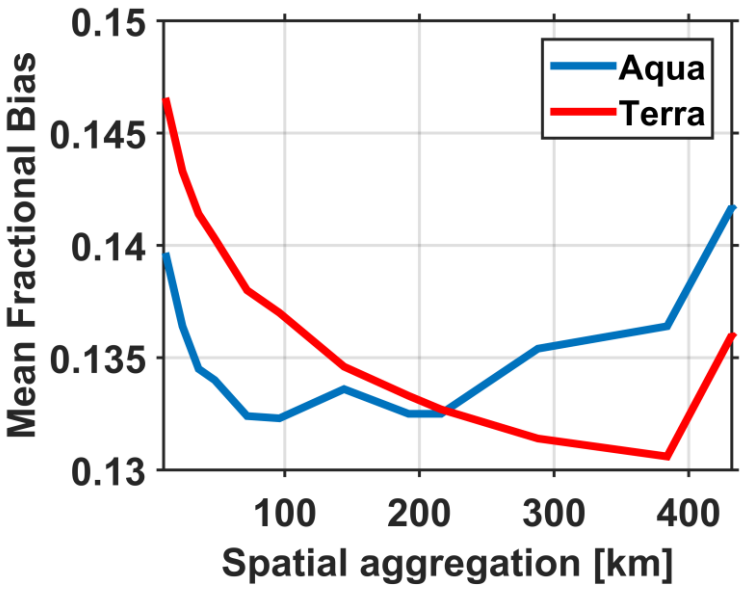

Figure 6. Mean fraction bias (MFB) on AOD from WRF-Chem as a function of spatial aggregation relative to observations from Terra (red line) and Aqua (blue line).

- Are model deviations from remotely sensed observations correlated with e.g. EPA differences or precip measurements? The paper only addresses this in the most cursory fashion. What can we learn from this about model deficiencies?

As we mentioned the AOD biases in the fall months (September and October) do appear to be linked to precipitation biases, and certainly are reflected in the near-surface $PM_{2.5}$ concentrations and composition (Fig. S4 and Fig. 4). We now elaborate on this a little further (lines 370-376; 418-420; Figures 4 and 8).

- Are AE differences somehow correlated with AOT differences (or vice versa)? Can this be used to understand model deficiencies?

As the reviewer will know AE is very difficult to derive from the MODIS measurements and the uncertainty in AE scales with AOD (AE is very uncertain at AOD < 0.2). This

and the fact that AE is derived from wavelength dependent AOD makes the uncertainties on the measurements certainly correlated. As indicated in Figure 7, for some AERONET sites there is evidence that positive bias in AOD is associated with high negative bias in AE, but this is not uniformly the case (e.g. for the site at 77.8W 55.3N WRF-Chem exhibits positive bias in AOD across the entire pdf while the simulated AE is negative biased, but the site at 84.28W 35.95N exhibits relative good accord for AOD but is negative biased in AE almost to the same amount as the northern station).

We also added the following comment at the end of Section 3.2:
"AE is very difficult to derive from the MODIS measurements and the uncertainty in AE scales with AOD (AE is very uncertain at AOD < 0.2). Further, AE is derived from wavelength dependent AOD, thus the uncertainties on the measurements are certainly correlated. As indicated in Figure 7, for some AERONET sites there is evidence that positive bias in AOD is associated with high negative bias in AE, but this does not uniformly occur over eastern North America (e.g. for the site at 77.8W 55.3N WRF-Chem exhibits positive bias in AOD across the entire pdf while the simulated AE is negative biased, but the site at 84.28W 35.95N exhibits relative good accord for AOD but is negative biased in AE almost to the same amount as the northern station)."

- Why are only 12 AERONET sites used? Surely AERONET offers more over the continental USA? Possibly this is due to a very strict interpretation of Kinne et al. 2013 recommendations?
We analysed data from 22 AERONET stations which are all stations collecting data during 2008 over our domain and satisfying the condition described in Section 2.2 for the comparison on a monthly basis:

"Where WRF-Chem output is compared with data from AERONET stations, a station is only included if there are at least 20 simultaneous estimates available."

It is worthy of note that although a large number of sites in the US have seen deployment of AERONET instrumentation, relatively few have significant data availability for 2008 as shown by the figure below:

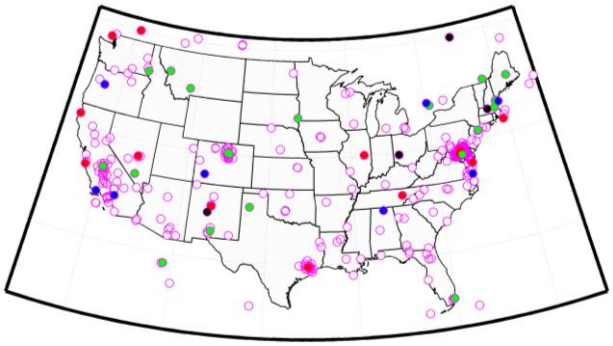

Figure. AERONET stations in/close to the contiguous US (magenta) that have been in operation as part of the network. Colors show the number of days at each station that in 2008 had > 1 observation of AOD at 551 nm (red>200, green=100-200, blue=50-100, black <50).

- Finally, the title of the paper is rather grand. A simple 'Evaluation of high-resolution WRF-Chem run over North America with remote sensing datasets' would do as well. The current title suggests a far broader canvas: multiple regional models for different domains using a set of complimentary observations beyond remote sensing data. Also, while remote sensing data

are of course appropriate for analysing forcing estimates from a model, they are by no means conclusive. The authors never really make the link to forcings.

We modified the title as follows:

Evaluating the skill of high resolution WRF-Chem simulations in describing drivers of aerosol direct climate forcing at the regional scale

Specific comments

Abstract

p 27312, l 10: MFB=0.5 is not a small bias. Even 0.17 is not a small bias, given that part of AOT is due to background and presumably constant in climate change/future predictions. Please strike 'small'.

Done

p 27312, l 15: "AE is retrieved with higher uncertainty from the remote sensing observations." does not belong here. Either strike or move one sentence.

We rephrased as follows:

"The model is biased towards simulation of coarse mode aerosols (annual MFB for AE = -0.10 relative to MODIS and -0.59 for AERONET), but the spatial correlation for AE with observations is 0.3-0.5 during most months, despite AE is retrieved with higher uncertainty from the remote sensing observations."

Introduction

p 27313, l 27: this suggests that PM10 or PM2.5 measurements have no bias and zero measurement uncertainty. This is of course not true. Please rephrase. AFAIK, IMPROVE measurements are made every 3 days, so also with PM10, PM2.5 under sampling may be an issue.

We have rephrased this:

"Long-term measurements of aerosol properties are largely confined to aerosol mass (total, PM_{10} or $PM_{2.5}$) in the near-surface layer which may or may not be representative of either the total atmospheric burden (Ford and Heald, 2013;Alston et al., 2012), or radiation extinction and hence climate forcing. Further, aerosol composition measurements are often a 24-hour integrated sample taken only 1 in 3 days and thus are subject to under sampling. Hence they provide an incomplete description of temporal variability and mean aerosol burdens for model performance evaluation."

p. 27314, l. 10: These are strange references here. E.g. Spracklen et al does not really discuss spatial scales in observed aerosol. There is quite a bit of literature on this though: Anderson et al JAS 2003; Kovacs et al JGR 2006; Santese et al JGR 2007; Sinzuka & Redemann ACP 2011; Schutgens et al AMT 2013. Several of these papers deal explicitly with spatial scales in remotely sensed properties.

Thanks for the suggestions. We replaced the reference with the following:

"However, aerosol populations (and dynamics) are known to exhibit higher spatial variability (and scales) than can be manifest in those models (Kovacs et al.,2006;Kulmala et al., 2011;Santese et al., 2007; Schutgens et al., 2013;Sinzuka and Redemann, 2011)."

p 27314, l 14: "The skill of these models in reproducing the spatio-temporal variability in the aerosol size distribution, composition, concentration and radiative properties is incompletely characterized. Accordingly, there is large model-to-model variability both in the global mean

direct aerosol forcing and in the spatial distribution". Skill characterisation and model-to-model variability are unrelated. Please rephrase as these sentences are confusing.
We rephrased as follows:
"The skill of these models in reproducing the spatio-temporal variability in the aerosol size distribution, composition, concentration and radiative properties is incompletely characterized. Further large model-to-model variability both in the global mean direct aerosol forcing and in the spatial distribution thereof exists (Kulmala et al., 2011;Myhre et al., 2013) leading to high uncertianty in quantification of aerosol climate forcing."

p 27315, l 13: "However, there are also variations in the way in which model skill is evaluated leading to ambiguity in terms of prioritizing future research directions". Even if we all use the same metric, there would still be ambiguity over e.g. what is the best way to improve models. Arguably, this is far more important than the metric itself. Please rephrase.
We rephrased as follows:
"However, there are also variations in the way in which model skill is evaluated and divergent opinions regarding prioritization of future research directions."

p 27315, l 23: "Assessment of value added (or lack thereof) from high resolution regional vs. global coarse resolution models is not quantifiable from prior studies alone." Which prior studies are referred to? What is meant by this sentence?
We rephrased as follows:
"Assessment of value added (or lack thereof) from high resolution regional versus global coarse resolution models has not been clearly quantified in previous studies (Table 1)."

p 27316, l 4: "inferential statistics". Descriptive statistics seem more appropriate here. I find little hypothesis testing or inference in this paper.
Changed to "descriptive statistics".

p 27316, l 9: "Prior analyses of Level-3 10 (10 resolution) MODIS AOD over the eastern half of North America have indicated the frequency of co-occurrence of extreme AOD values (>local 90th percentile) decreases to below 50% at 150 km from a central grid cell located in southern Indiana, but is above that expected by random chance over almost all of eastern North America (Sullivan et al., 2015)." What central grid-cell? I guess the authors are referring to a particular model evaluation? What is the importance of the 150 km distance? Instead of going into a lot of detail, maybe you can just tell in one or two sentences what the relevance of Sullivan 2015 is to your work?
We agree and rephrased as follows:

Prior analyses of Level-3 (1° resolution) MODIS AOD over the eastern half of North America have indicated extreme AOD values (> local 90th percentile) are coherent over regional scales (~ 150 km) (Sullivan et al., 2015). Thus, our evaluation exercise also includes an analysis of the spatio-temporal coherence of extreme events.

p 27316, l 27: Strictly speaking, AERONET measurements are not columnar measurements. Standard AERONET product measures attenuation of direct sun-light and so actually measures aerosol along a slant path. However, final AOT values are corrected for this to represent the vertical column.
Agree, we removed "columnar".

p 27317, l 12: It is customary to have a brief overview of the paper's structure at this point.
We added the following paragraph:

"This paper is structured as follows. We first describe the settings used in our WRF-Chem simulations and introduce the remote sensing and other data used for model evaluation (Sect. 2). A description of statistical metrics used for the evaluation is also provided. Section 3 presents results of the evaluation of simulated AOD and AE versus observations, as well as findings on extreme AOD values. In Section 4 we summarize our findings and draw conclusions."

p 27318, l 29: Don't the median diameters of MADE aerosol vary throughout the simulation, in both space and time? Or are they fixed (i.e. is a single moment scheme used, where mass only is considered)?

Yes, diameters vary throughout the simulations (the values we reported refer to the initial diameter) whereas the standard deviations are fixed within each mode. We modified the text accordingly.

p 27320, l 7: How does this official error estimate compare with Hyer et al AMT 2011? I believe official MODIS estimates are rather optimistic.

Thanks for pointing this out. We included this reference for comparison.
"The L2 AOD uncertainty is ± 0.05 ± 0.15×AOD over land relative to global sun photometer measurements from AERONET; even when no spatiotemporal averaging is used in the comparison (i.e. all combinations of MODIS retrievals within 30 km of an AERONET site and all AERONET retrievals within 30 min of the satellite overpass), 71% of MODIS retrievals fall within a ± 0.05 ± 0.2 ×AOD envelope relative to AERONET over E. CONUS (Hyer et al., 2011)."
p 27321, l 6-19: The exact procedure is not clear due to missing information and confusing sentences. The cloud screen (presumably from MODIS?) is applied to model data first and then only cells with 5 or more observations per month are retained? Cases with cloud fraction > 0 are discarded? In my experience that removes a lot of good observations as well. Which cloud screen do you use: the one that is part of the aerosol product MYD/MOD04 or another one? What do you do with MISR data or AERONET? Model data are not masked by observation availability in their case? AERONET is compared to the closest grid-cell or do you interpolate model data to the site? What about time of observations? You choose again nearest model time?
We did not apply a cloud screen to the MODIS or MISR data, beyond what is already in the algorithms to remove cloud pixels. In the NASA products 'cloudy' pixels are identified and removed; then for the remaining pixels, the 50%/20% brightest/darkest pixels are also removed (assumed to be cloud contaminated), and the remaining pixels are averaged for the retrieval (Levy et al. 2013). So we do get good retrievals when cloud fraction > 0, but the cloud pixels are screened out.
We reworded the data section as follows:

"To avoid the discontinuity in the MODIS retrieval algorithm due to different assumed aerosol types (Levy et al., 2007), we confine our analyses of model skill to longitudes east of 98°W. Only WRF grid cells with cloud fraction = 0 during the satellite over pass of each grid cell are used in comparison to MODIS/MISR observations, and only grid cells with at least 5 valid observations (both from MODIS/MISR and cloud-screened WRF) during a given month are included in the analyses presented herein. It is worth noting that setting a threshold of 10 observations does not significantly affect the results. For a uniform assessment, L2 MODIS and L3 MISR data have been interpolated from their native grids (and resolutions of 10 km and 0.5°×0.5°, respectively) to the WRF-Chem 12 km resolution grid by computing the mean of pixels with valid data within 0.1°/0.3° for

MODIS/MISR from the model centroids. The choice of averaging over a slightly larger area than model resolution is dictated by the sparsity of valid satellite retrievals. For AERONET vs. MODIS comparison, we only use the nearest MODIS data (after regridding to WRF) to each site. Where hourly WRF-Chem output is compared with data from AERONET sites, a station is only included if there are at least 20 simultaneous estimates available, and each AERONET measurement is compared to the nearest WRF-Chem time step and to the grid cell containing the station."

Reference:

Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., & Hsu, N. C. (2013). The Collection 6 MODIS aerosol products over land and ocean. Atmos. Meas. Tech. Discuss, 6, 159-259.

p 27321, l 23: While the use of MFB is warranted, its interpretation is less clear than (M-O)/O, please discuss this. Also, relative errors (like MFB) seem less appropriate than absolute errors in case of an intensive property like AE.
As the reviewer suggests, there are a range of performance metrics one can use to evaluate models. We decided to compute the MFB instead of Normalized Mean Bias (NMB) since NMB is biased towards overestimations and assumes observations are without error, while MFB gives equal weight to underestimation and overestimation. We put a reference to this at line 285.

p 27322, l 1-5: "Where MFB is reported for WRF-Chem vs. MODIS or MISR, Cm is the monthly mean AOD or AE simulated by WRF-Chem at a specific location, C0 refers to the same quantify from MODIS or MISR (Table 3) and N is the sample size. Where MFB is reported in comparisons of WRF-Chem with AERONET, the monthly average in the model grid cell containing the AERONET site is compared with monthly averaged observations (C0)." So much text suggests there is a difference in how you treat MODIS and AERONET data, yet I see no difference?
Correct, there is no big difference in the way we treat AERONET, so we reworded as follows:

"Where MFB is reported for WRF-Chem versus MODIS/ MISR/AERONET, C_m is the monthly mean AOD or AE simulated by WRF-Chem at a specific location, C_0 refers to the same quantify from remote sensing data (Table 3) and N is the sample size."

p 27323, l 10: What is type i? Which rows and columns do you refer to? Maybe it is easier to simply mention these metrics (incl EQQ and Taylor plot) and then refer to papers, books that discuss them in more detail.
We simplified the text and removed the formula. We preferred to keep some brief explanations of the methodologies applied for clarity.

p. 27323, l 25: So ME, WN and MN are frequencies of occurence? Occurence itself is not a metric.
Replaced with "frequency".

p 27324, l 10: Why are these extra metrics HR & TS useful? What do they tell you that Accuracy does not tell you? Instead of giving the functional forms (which readers can look up in books anyway) it is more useful to explain the meaning of the various metrics.

We preferred to maintain the functional forms for easier reference in the result and discussion sections. However we included a more detailed description as follows:
"The Accuracy describes the fraction of grid cells co-identified as exceeding *p75* or not in MODIS and WRF-Chem, and thus equally weights event and non-event conditions. Since the Accuracy quantifies model skill in correctly identifying both extreme and non-extreme aerosol loadings, it is thus indicative of model performance in capturing the overall AOD spatial variability."

Interpretation of the three metrics is also included in section 3.3 (first paragraph) and in the Table 6 caption.

p 27324, l 16: Why is this done for a single reference location only? Wouldn't it make more sense to use a reference location on the East coast where more pollution exists anyway?
We chose the center of WRF-Chem simulated domain as reference location for several reasons:

1) **to be comparable to Sullivan et al. (2015) where it is also shown that moving the centroid did not greatly impact the coherence estimates**
2) **to represent a grid cell that closely represents the center of gravity of the domain**

We added the following to support our choice:
"The reference location represents the center of gravity of the domain and was previously used by Sullivan et al. (2015) for assessing scales of coherence. In that work they also found the spatial scales of coherence are not sensitive to the precise choice of reference location."

p 27325, l 5: Table 3 shows that largest non-zero MFB occurs when MODIS Terra is compared to AERONET AOT. Doesn't this suggest that either Terra is really wrong (and not suited to evaluate WRF) or AERONET is already unrepresentative for scales like the 10 km MODIS pixel (unlikely)?
Thanks for this comment. We clarified in the text and Table 3 that the MFB of MODIS vs. AERONET is strongly affected by some outlier sites and the MFB decreases when we remove the three most biased sites. Further, the number of co-samples between MODIS is quite limited, thus those MFB may be not very representative. We added the following comment:
"When MODIS is compared to the 22 AERONET stations the MFB is -1.23 suggesting an underestimation of AERONET relative to MODIS. The large bias can be explained noting that the number of co-samples between MODIS is quite small and that MFB is strongly impacted by a few outliers. When we remove the three most biased sites (one land site in the North and two sites along the East coast) the MFB decreases to -0.91."

p 27326, l 6: "because WRF-Chem simulates high AOD and aerosol nitrate and sulfate concentrations". This is a sweeping statement with no evidence to support it. Please remove or elaborate.
We included more analyses on the chemical composition comparison and modified the text accordingly. Please see detailed response to reviewer 3.

p 27326, l 21: "occupy much of the same parameter space". This sentence is confusing. How can WRF-Chem comparisons with AERONET (M-O) be compared to AERONET or MODIS observations (O)?

The comparison between WRF-Chem and MODIS is done by gridding L2-MODIS data to 12km to match model grid, whereas comparison between WRF/MODIS and AERONET is done by comparing hours with simultaneous data in the grid cell including each AERONET station. We modified the manuscript accordingly in the data section as discussed before.

p 27326, l 23: "model simulations reproduce the range and probability of low uncertainty AERONET measured AOD nearly as well as MODIS." But the times and locations can be way off. It is important to comment on this aspect. EQQ plots can only take you so far.
We agree. The EQQ plots do not necessarily simultaneously compare the same MODIS-AERONET and WRF-Chem-AERONET pairs. We rephrased as follows:
"However, it is worthy of note that WRF-Chem comparisons with AERONET observations occupy much of the same observational range as simultaneous MODIS and AERONET at those sites (Fig. 9a), although the EQQ plot does not necessarily compare the same MODIS-AERONET and WRF-Chem-AERONET data pairs (i.e. the sample used to compare AERONET and MODIS may differ from that used to compare WRF-Chem and AERONET due to the cloud screening procedure)."

p 27326, l 27: "Nevertheless,". Why nevertheless? These correlations seem very low to me. Maybe that is due to observational error but I doubt it. AE MFB WRF-Chem AERONET = -0.59, so a substantial bias (note that AERONET AE have been averaged over 20 individual measurements during a month reducing measurement errors), so WRF-Chem probably has an issue in correctly simulating AE anyway.
We rephrased as follows:
Despite the low confidence in AE retrievals from MODIS, the comparison of WRF-Chem with the remote sensing estimates indicates some degree of agreement. The overall MFB of WRF-Chem vs MODIS Terra is -0.09 (-0.11 vs. Aqua) and the correlation between WRF-Chem and MODIS monthly mean AE seems to be independent of season and lies between 0.20 and 0.54 for all months except April, May and November when it is lower, whereas *r* is always < 0.14 when comparing with MISR (Fig. 7b).

p 27327, l 14: "After cloud screening". Why after cloud screening? I thought all model data used in comparison with observations are cloud-screened to start with?
Yes, it's correct. We removed "after cloud screening" to avoid confusion.

p 27328, l 12: the threshold for extreme AOT events (p75) is different for WRF-Chem and MODIS. How different is it?
Given we already focused on the quantification of the bias in AOD magnitude, now we are analysing differences in distribution and in spatial patterns. As an example, for Aqua, the p75 threshold varies by a minimum of 7% larger for WRF-Chem relative to MODIS in July to up a three times larger during the month of October when we already know the model has a larger bias in AOD due to the underestimation of precipitation.

p 27330, l 12: AOD=0.22 is a domain-average for clear grid-cells. So the orbit of MODIS was not taken into account? The MFB is thus calculated from two datasets with different spatial sampling? If so, that would be plain wrong.
No, we are still considering data over the same grid for hours of satellite overpass time. We rephrased for clarity as follows:
"After grid cells with any cloud presence are removed and considering only overpass hours, the domain averaged simulated mean AOD is 0.22."

p 27330, l 18: AERONET MFB=0.5 according to Table 1
Thanks, fixed.

p 27330, l 22: Please also discuss/mention clear north-south gradient in AOT bias vs Terra (Fig 6). Maybe relative errors do not show a gradient? Does this gradient also exist in yearly precip errors (like Fig S3)?
The figure below shows that the N-S gradient is still present when we use NMB to evaluate model performance. We explicitly note this pattern in the text:
"A clear North-South gradient in AOD bias vs MODIS is also observed."

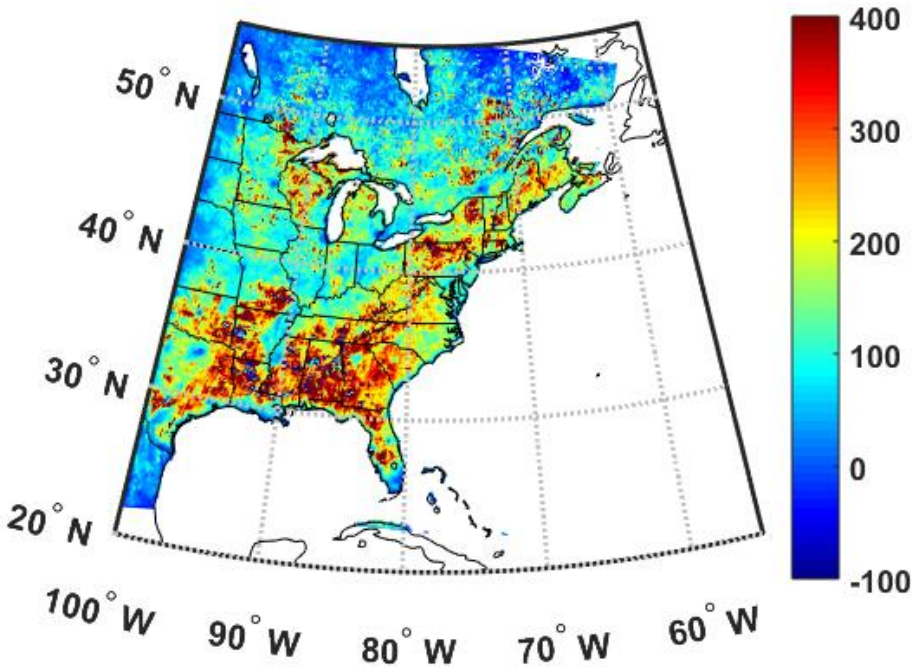

Figure. Normalized Mean Bias of AOD from WRF-Chem and Terra on a yearly basis.

p 27331, l 6: Table 3 suggests AE MFD vs AERONET is -0.59
Thanks, fixed.

p 27331, l 9: "the bias relative to AERONET is consistent with prior research (Table 1) and is symptomatic of relatively poor model performance for this metric." A non-zero bias is not symptomatic of poor model performance, it is one of the most important metrics by which we judge model performance.
We rephrased as follows:
"the large bias relative to AERONET is consistent with prior research (Table 1), and is symptomatic of substantial systematic error."

p 27331, l 22: "central tendency" -> mean or average
Changed with "mean AOD values".

p 27331, l23: Not 'maximized' but 'greater'. After all, you talk about high loadings, not the highest loadings
Done

p 27348: Larger symbols for AERONET sites would be useful

We modified Figure 1 making larger symbols and including the MFB for AOD at AERONET locations (previously in Figure 2).

p 27349: Numbers in plot hard to read and not very useful anyway because exact location of site not clear and lot of fine structure in underlying MODIS data. Consider removing AERONET data.

We removed the numbers and included those relative to AOD in Figure 1 to save the information regarding the spatial variability in model performance.

p 27350: the lack of spatial variation in the observations is striking. Is this simply because of the colorbar scale? Or does WRF-Chem show more variation?

We remade the figure setting a different colorbar scale for WRF-Chem and EPA for easier visualization of the spatial variability in the observations.

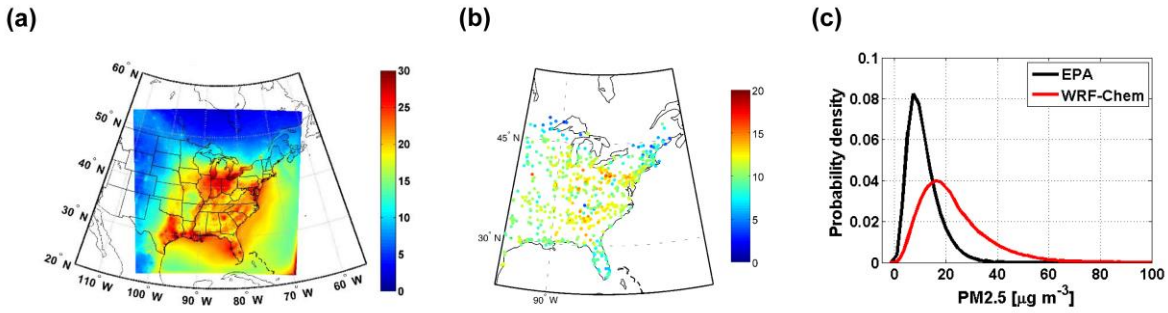

Figure 3. Mean daily $PM_{2.5}$ concentrations [µg m^{-3}] during 2008 as (a) simulated by WRF-Chem in the layer closest to the surface and (b) observed at 1230 EPA sites (note the different colorbar). Panel (c) shows the probability distribution of daily mean $PM_{2.5}$ concentrations observed (black line) and simulated (red line) at the measurement stations.

p 27351: While an interesting attempt at presenting a lot of information concisely, I find it difficult to easily separate the different coloured rings. Rather, one might try to use color (MFB, blue-red scale), symbol size (correlation) and symbol (RMSD, clearly this requires the RMSD to be binned in to 5 or so range bins) to denote the same information.

We remade this figure.

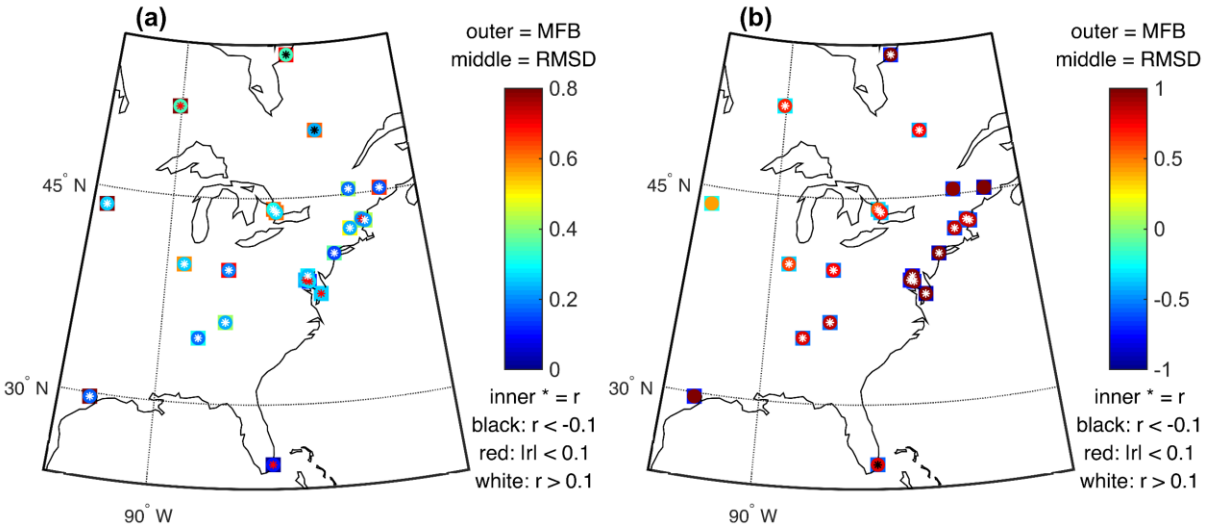

Figure 5. Summary statistics of comparisons of WRF-Chem simulations of (a) AOD and (b) AE relative to simultaneous observations at the AERONET sites. For a location to be included in this analysis at least 20 coincident observations and simulations must be available. The symbols at each AERONET station report MFB (outer square), root mean squared difference (RMSD, outer circle) and correlation coefficient (r, inner *). Note the different colorbar for MFB and RMSD between the two frames. The correlation coefficient is displayed with different colors according with 3 classes: r <-0.1 (black), $|r|$ <0.1 (red) and r>0.1 (white).

p 27352: It would be very interesting to see if these Taylor plots change when data is spatially aggregated first, i.e. what if model+obs are averaged over 12, 24, 48, 96 km before Taylor plots are made?

We performed this analysis and included a figure in the Supplementary Materials. The text was changed as follows:
"We also examined the impact of spatial aggregation (at 12, 24, 36, 48, 72 and 96 km) on the seasonality of model performance. For AOD the spatial correlations are largest for most months when data are aggregated to a resolution of 24×24 km and the ratio of spatial standard deviation is also closer to 1 when AOD are spatially aggregated, possibly indicating that the spatial patterns simulated by WRF-Chem at a fine scale do not always match those observed by MODIS (Fig. 8). For AE both spatial correlations and ratio of standard deviations do not vary significantly when data are aggregated to a coarser resolution (Fig. S5). "

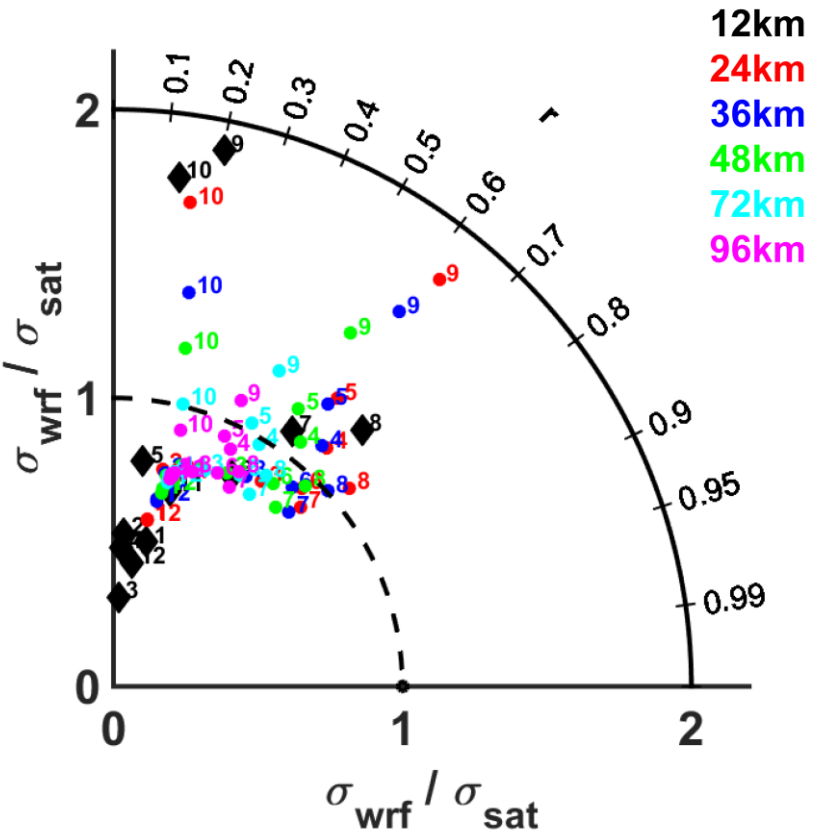

Figure 8. Taylor diagrams for AOD when MODIS observations and WRF-Chem simulations at 12 km are spatially aggregated to 24, 36, 48, 72 and 96 km. Numbers next to the colored dots/diamonds indicate different months.