Atmos. Chem. Phys. Discuss., 15, C8936–C8943, 2015
www.atmos-chem-phys-discuss.net/15/C8936/2015/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "How skillfully can we simulate drivers of aerosol direct climate forcing at the regional scale?" *by* P. Crippa et al.

Anonymous Referee #2

Received and published: 4 November 2015

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 eachother 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.

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)

- 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?

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

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?

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.

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'.

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.

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 evry 3 days, so also with PM10, PM2.5 undersampling may be an issue.

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.

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 characterisationand model-to-model variability are unrelated. Please rephrase as these sentences are confusing.

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.

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?

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

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 particualr 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?

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.

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

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)?

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.

p 27321, l 6-19: The exact procdure 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?

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.

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?

p 27323, l 10: What is type i? Which rows and columns do you refer to? Maybe it is eaier to simply mention these metrics (incl EQQ and Taylor plot) and then refer to papers, books that discuss them in more detail.

p. 27323, l 25: So ME, WN and MN are frequencies of occurence? Occurence itself is

not a metric.

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 ueful to explain the meaning of the various metrics.

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?

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)?

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.

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)?

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 sofar.

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.

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?

p 27328, l 12: the threshhold for extreme AOT events (p75) is different for WRF-Chem and MODIS. How different is it?

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.

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

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)?

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

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.

p 27331, l 22: "central tendency" -> mean or average

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

p 27348: Larger symbols for AERONET sites would be useful

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.

p 27350: the lack of spatial variation in the observations is striking. Is this simply

because of the colourbar scale? Or does WRF-Chem show more variation?

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

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?

Interactive comment on Atmos. Chem. Phys. Discuss., 15, 27311, 2015.