Atmospheric
Chemistry
and Physics
Discussions

Interactive comment on "Spatial and seasonal variations of aerosols over China from two decades of multi-satellite observations. Part I: ATSR (1995–2011) and MODIS C6.1 (2000–2017)" by Larisa Sogacheva et al.

A. M. Sayer (Referee)

andrew.sayer@nasa.gov

Received and published: 15 May 2018

Summary:

I am writing this review under my own name (Andrew Sayer) as I have previously discussed this research with the authors, and am on the team responsible for the MODIS aerosol data products being used in the study. I also reviewed the paper de Leeuw et al (2018), which is in some sense a predecessor to this study, and have been invited to review the follow up Part II to this paper also by Sogacheva et al and also currently in

ACPD. I feel I am able to provide an impartial review, but am signing the review in the interests of transparency.

The goal of this paper is to look at spatial and temporal (seasonal/interannual) variations of AOD over China. This is accomplished mainly by using two satellite data sets: the ADV algorithm applied to the combined ATSR2/AATSR record (1995-2012), and the combined Deep Blue/Dark Target algorithms applied to the MODIS Terra record (2000 onwards) from the latest Collection 6.1. CALIOP data are also used briefly to look at seasonality of desert dust. AERONET data are also used for validate the MODIS and ATSR retrievals where possible – the available sites are largely confined to eastern China. This paper sets up the later analysis in Part II, which is more focused on combining the time series to look at trends, although there's a bit of overlap in that trends (or at least 'tendencies' in the authors' terminology) are discussed in this paper as well. As a result I felt it best to review this Part I first. Aerosol remote sensing over China can be more difficult than over many other parts of the world, there are less publicly-available validation data than some other areas, and there are suggestions from the literature that historical trends in AOD in some areas may be changing (i.e. they're not linear and start/end dates of analyses are important). So the research question behind the authors' series of papers is relevant. There is overlap between this paper and de Leeuw et al (2018) but the analysis here is expanded and updated, and I think there is sufficient new stuff here that the overlap does not preclude publication.

My overall recommendation is for major revisions and re-review. Overall it is a good paper but I think there's a gap in establishing the reliability of the ATSR2 portion of the record, and the premise of the analysis rests of being able to treat ATSR2 and AATSR as one consistent long-term record. This has not really been assessed by the paper. In my specific comments below I have a few suggestions how that could be done.

The quality of language is overall good and any issues can probably be dealt with by Copernicus' copy-editing and typesetting process. Therefore my review mainly concentrates on technical abstracts. Here, PXLY refers to page X, line Y.

Specific comments:

Abstract: I would condense this into one paragraph if possible and shorten it to highlight the main findings. For example I'd delete the paragraph comparing validation statistics for MODIS C6 and C6.1 as probably not the most important thing (see also later comments on that). I'd also consider shortening or removing the final paragraphs about AOD tendencies over the period, since that's the focus of part II. In my mind it is enough here to talk about the overall validation results and spatial/seasonal patterns. The sentence on P2L5-6 which briefly summarises the interannual variability is useful.

P5L9-10: I would assume that "China" here is defined as any 1x1 degree grid cell with retrievals over land which crosses or is contained by the borders shown in Figure 2. It would be good to make this explicit.

Figure 2: It is hard to see the regional boxes and borders over the top of the AOD map, because there are a lot of colours and some lines are thin. I suggest plotting the national borders in black rather than blue, to make them stand out, as the blue colour used is similar to that coded in the colour bar for common values of AOD shown in the map. Subregion boundaries could also do with having thicker lines and being in a strong contrasting colour. For example although the colour scale tops out at red, red may be a suitable choice as it doesn't look like any of the grid cells reach this value. Alternatively (and probably even better), the authors could abandon the rainbow colour scale used in this Figure and some others in the paper. Rainbow scales can present difficulties for colour blind readers, do not always display or print consistently, and draw the eye to specific points in the range while smoothing variations in other parts. See for example https://www.climate-lab-book.ac.uk/2014/end-of-the-rainbow/ or https://eagereyes.org/basics/rainbow-color-map for some good discussion of why this is, and https://personal.sron.nl/~pault/ for some good alternative palettes to use.

Section 3.1: somewhere in here I would cite Sayer et al (JGR 2014) as the reference for the MODIS merged Deep Blue/Dark Target data set used. This describes the algo-

rithm, shows some examples, and provides validation. The paper is about Collection 6 but the merging logic is unchanged for Collection 6.1 (C61) so it still stands as the best reference. We don't have a paper specific to C61 at present. This study is cited elsewhere in the manuscript already, but I think it would be good to add it here too.

It would be useful to be clear about exactly which products (and which SDS names within them) were used. For example P6L11 says "L3" data – there are several different L3 aggregation time scales and it isn't clear if the daily product (MOD08_D3) was used as a basis and aggregated to a monthly scale, or if the monthly product (MOD08_M3) was used directly. Plus, presumably the later validation exercise used the L2 product (MOD04_L2).

P6L20-23: Here the authors link to a summary document about C61 about cloud detection. This is not quite satisfactory since that's not necessarily a stable url and it isn't a document with a defined author or contact point. I reached out to the MODIS calibration and cloud teams and they suggest replacing with a citation to Moeller et al (SPIE, 2017): http://spie.org/Publications/Proceedings/Paper/10.1117/12.2274340. That paper contains more examples of the effect of the crosstalk problem and the Wilson et al (2017) algorithm to fix it on the cloud detection time series. So I'd just cite the Moeller and Wilson papers here instead of linking to the C61 pdf from the MODIS Atmospheres site.

P7L5-6: Here the authors reference by url one of my review comments to de Leeuw et al (2018) on ACP. I don't necessarily think the sentence is needed since the relevant info is more or less shown in their Figure 3. But if the authors want to keep it I would note that ACPD review comments (to which the figure linked was attached) are citable, and this one was doi:10.5194/acp-2017-838-RC1, so that's a better way than giving the url.

Section 3, general (and my main substantive comment): One point that is glossed over in the validation and later discussion is the fact that almost all the available validation

data for ADV here are for the AATSR portion of the record (2002-2012), not the ATSR-2 portion (1995-2002). The instrument characteristics are very similar so one would expect similar performance. The main potential reason for difference would be offsets in the sensors' absolute calibration, which remains a thorny issue for all satellite AOD data sets. On the one hand there what the authors can do here is limited, because without validation data covering both the ATSR-2 and AATSR missions it's hard to say whether the error characteristics of the two instruments are similar or not, and it is definitely not the authors' fault that this is not available. But on the other hand, since the point of this paper and Part II is to look at a long time series, it is unfortunate that the first 5 years of the period (1995-2000) before MODIS, which are one big advantage of the ATSRs compared to many other sensors, are a period where the data set can't effectively be validated. This is doubly unfortunate since trend analyses are particularly sensitive to values at start and end points.

I think it would make the authors' case stronger if they could demonstrate the reliability of the ATSR2 portion of the record in a more concrete way. From when I used to work on the ATSRs (a different algorithm, ORAC), I recall there is about 1 year of overlap between ATSR2 and AATSR (during 2002-2003) when the two instruments were flying on the same orbit track spaced something like 30 minutes apart. Actually I think there is more than 1 year but my memory is that the ERS2 satellite on which ATSR2 flew had some technical issues with its pointing accuracy after 2003. My point is, I am pretty sure there is 1 year of overlapping data, and potentially more, which could be used here.

So one step would be to take this year of overlapping data, and do an analysis of how consistent the ATSR2 and AATSR retrievals are (in terms of AOD retrieved and also if they make similar choices about when to retrieve, i.e. data coverage). This could be a few extra figures/tables showing joint scatter density histograms or the monthly means in the various regions of interests and things like that.

As a second step, there were a few AERONET sites in China in operation during the

ATSR2 period, and during the ATSR2/AATSR overlap period (2002-2003). There will probably not be too many matches with AERONET during this year, but even if you get only a dozen or so where you have AERONET, ATSR2, and AATSR all together, that lets you say something about whether the retrieval errors are similar or not.

A third option is do this sort of 3-way comparison using ATSR2, AATSR, and MODIS Terra, using the 3-way overlapping time period, which should get more spatial coverage than AERONET. Even though MODIS can't be considered a reference truth, one can still see if the ATSR2 vs. MODIS and AATSR vs. MODIS differences are consistent.

In my view adding analyses along these lines would put the discussions later and in Part II of the manuscript on much firmer ground. Otherwise the study is assuming that ATSR2 can be used to extend the AATSR record back in time, with consistent error characteristics, without directly testing that assumption and assessing quantitatively (via these comparisons) the differences and their implications for long-term analyses. This was also missing from de Leeuw et al (2018) so adding here would further help distinguish this study from that.

One more note on this topic: the stability of the calibration of the ATSR2/AATSR sensors is well established, thanks to on board and vicarious calibration techniques. It is not that which I am concerned about so much in terms of the combined ATSR2/AATSR record. Rather, it is small differences in the absolute calibration (on-board only monitors stability i.e. relative calibration and absolute is based on pre-launch measurements), and small differences in band spectral response functions, which might lead to offsets in the retrieved AOD between the sensors, and thus affect apparent AOD trends/tendencies.

P9L3-9: The differences between C6 and C6.1 performance seem to be sufficiently small to me that I don't think the statement that MODIS C6.1 is "slightly better" is warranted. My hunch is that the differences are not statistically significant given the limited sample sizes. If I were writing this I would say that C6.1 has about 5 percent

more matchups with AERONET but performance is about the same. Also, the sentence about bias decreasing "from 0.007 to 0.06" must be incorrect because 0.06 is larger than 0.007. Figure 4 suggests 0.06 for C.61 (and does not show C6) but the Abstract says 0.007 and 0.006. So something is inconsistent here – I wonder if a decimal place has been added or subtracted in one of the cases. My guess is that it should read 0.07 and 0.06.

Section 3.3, general: I like the various breakdowns and subsets of the data. One other way I have found to be useful is to subset the satellite data by some proxy for aerosol "type" using AERONET data. I tend to classify conditions as "background" (AERONET AOD less than some number like 0.2), "fine-dominated" (AOD larger than 0.2 and AE larger than some value like 1), or "coarse-dominated" (AOD larger than 0.2 and AE smaller than 1). Then we can see if biases in the satellite retrievals are associated with particular aerosol conditions. It might be that dust is fine but smoke is biased, for example. Similar splits can be done with surface classifications. This may be instructive when later looking at regional comparisons and understanding why the data sets are different and which may be more reliable in a given situation. I am not saying this sort of analysis absolutely must be added to the paper, but it might be useful for some of the later discussions and when attempting to merge the MODIS and ATSR data sets to estimate trends in Part II.

P11L1: Here it says a 90 minute temporal window is used. Previously on P9L4, the text said 1 hour. Which is correct, or were different thresholds used for the separate analyses? This should be clarified.

P13L3: The authors tend to use the term "bias" when comparing any two data sets in the paper and describing one as lower/higher than the other. However "bias" is a loaded word with the implication that whatever is described is incorrect by that amount. I think it is appropriate to use the term "bias" when comparing something to AERONET, since AERONET is taken as our ground truth. However when comparing the satellite data sets to one another, as at P13L3, it is better to use the word "offset" instead (i.e. x

is offset from y rather than x is biased against y). I think this more precise terminology will help clarify when we are really talking about something which might be an error, and when we are talking about something that is a difference but the error is uncertain. This would also help a reader who is less familiar with the satellite data and might not realise what typical levels of uncertainty in sun photometer or satellite AOD data sets are.

Figure 7: I would remove this figure because I do not think that annual maps like this are particularly meaningful. Clouds/snow show seasonal dependence, affecting coverage. Aerosol loading and type and surface cover show seasonal dependence, affecting error characteristics. So a set of seasonal maps would be better. These are already provided and discussed later as Figure 10. So I suggest just deleting Figure 7 and the sentences about it and save that discussion for Figure 10 instead. I don't think that having both Figure 7 and Figure 10 is necessary, and Figure 10 is the more informative of the two.

Section 4 and Figures 8 and 9 and discussion: To help focus the manuscript, I would also consider removing these figures and associated discussion (or at least shortening it somewhat). From the outset the goal of this paper is to assess the MODIS and ATSR records, and Part II is meant to be about the trends. So I would try to stick to this focus to make the paper more readable, and also avoid readers of Part II either having to refer back to the trend discussion here or having duplicate material if it is reproduced. Otherwise the subject matter of the two papers is too mixed and you may as well just combine them into one paper. I feel similarly about sections 5.2 and 5.3. Reading through Part II, I feel that this material (which is intended to make the point that the two data sets can be combined, after some correction) feels much more natural if it is moved to an early place in the Part II paper. To me it is more or less part of the method for Part II, not the analysis of Part I. This will also decrease the need for repetition of content between the two papers and allow more space for a comparison between ATSR2 and AATSR during their overlapping year(s) of 2002-2003 as suggested.

I don't think that taking out the interannual variability here would weaken the conclusions or argument for Part I. I understand why the authors initially put this here and that they want to make the analysis more distinct from de Leeuw et al (2018), but I really think the interannual stuff belongs in Part II. Another risk is that a reader reads Part II and sees the tendency/trend calculations here and takes that as the final word, and never goes on to see Part II. So I really feel that switching around this material between the papers is worthwhile for both parts to maintain their focus and readability. If the authors really want to include some interannual stuff here, that's fine, but I would keep it brief and only show a few time series to illustrate the overall magnitudes and interannual variability in ATSR/MODIS without trying to fit linear trends to it. But I don't think that's needed and would rather the space was spent establishing whether ATSR2 and AATSR ADV AOD retrievals can be treated as one seamless AOD record rather than a pair of records from similar instrument.

P31L12-14: as mentioned earlier, I'd instead say that the performance of MODIS C6 and C6.1 are pretty much the same except coverage increased a bit.

C9