Response to short comment by Dr. Andrew Sayer

We thank Dr. Sayer for his suggestions (in red).

However, I see they use the Dark Target AOD product at 470 nm, rather than 550 nm. 550 nm is the main reference wavelength for this product, the one that has been validated, and the one which is generally recommended to be used (and is indeed used by most data users).

We agree with this comment. It was realized after the original submission of the manuscript that the 470 nm product was selected unintentionally instead of the 550 nm product. Rather than withdraw the manuscript or ask for a long extension to regenerate a decade of MATLAB *.mat files required as input for our validation and mapping software, the Dark Target AOD at 470 nm was retained temporarily with the full intention of redoing the map in Fig. 1, the validation results (Table 4), *et cetera*, at the next stage in the review process.

We now write at p2L28:

Specifically, the Corrected_Optical_Depth_Land (550 nm) and the Deep_Blue_Aerosol_Optical_Depth_550_Land datasets were used and confidence for both datasets was extracted from the Quality_Assurance_Land dataset.

Similarly, the Deep Blue AOD quality flag is in Deep_Blue_Aerosol_Optical_Depth_550_Land_QA_Flag, but we also provide a data set which already has the quality flag mask applied (Deep_Blue_Aerosol_Optical_Depth_550_Land_Best_Estimate) so the user does not have to do the filtering themselves. It is not clear to me from the paper which SDS was used to QA-filter the Deep Blue data but I am assuming it is the above. More information can be found in the MODIS aerosol file spec document (http://modisatmos.gsfc.nasa.gov/_specs_c6/MOD04_L2_CDL_2013_03_21.txt) or on our website, http://deepblue.gsfc.nasa.gov. Could this be clarified?

We agree that the ACPD manuscript fails to name the SDS used to QA-filter the Deep Blue data. 'Quality_Assurance_Land' is the SDS used.

The change to the manuscript is contained in the sentence mentioned above at p2L28, in response to the previous comment.

Also, which ATSR product is used? There are at least 3 being produced in Europe in the framework of the ESA CCI project, and they all have different approaches and results (see Popp et al, Remote Sensing, 2016, doi:10.3390/rs8050421 for an overview). My inference is that this is the Swansea algorithm (Peter North's group) but I think this should be stated more clearly.

The selected ATSR product is stated clearly in the appendix of the existing manuscript (p13L17) and the appendix is referenced at p2L26 in connection with the satellite data products. Information on the ATSR product is in the following sentence of the appendix (p13L17)

AATSR and ATSR-2 version 4.1 data are from Swansea University and can be obtained from the Aerosol CCI website (http://www.esa-aerosol-cci.org/) following registration.

Perhaps the others could be added to the analysis as well, if this is not too much effort. Similar to Dark Target vs. Deep Blue for MODIS, the various ATSR algorithms have different coverage.

There are three different algorithms for both AATSR and ATSR-2, and at least two POLDER algorithms, several MODIS products (Terra vs. Aqua, Deep Blue vs. Dark Target), plus MISR. That is eleven, and it is not an exhaustive list of available products from these satellite-based sensors. The primary focus of this paper is not on algorithms but on the different aerosol sensors. The Swansea University algorithm was chosen since initially they had, by far, the longest AATSR data record available.
To make this decision clear, we now write at p3L9:

The focus in this paper is primarily on the different aerosol sensors, rather than the different retrieval algorithms applied to the same satellite data (e.g. Popp et al., 2016), with the exception of the widely used Deep Blue and Dark Target algorithms for MODIS.

For POLDER, the data product the authors have used reports AOD at 865 nm. Due to the wavelength dependence of AOD, in most cases this means that the AOD will be much lower at 865 nm than 550 nm. The smaller signal will probably cause problems for relationships constructed using this AOD, plus one would not expect a close match between AOD at 550 nm (given by the other sensors) and 865 nm since the spectral dependence of AOD is determined by the aerosol composition. I wonder if another POLDER data product like GRASP (see e.g. http://www.grasp-open.com/products/) which does report AOD at 550 nm would be more useful here (and also allow for a more direct comparison between the various data sets).

We have used the only POLDER AOD data product that was available at CNES's POLDER website. We did not search the web or the literature for alternate POLDER products.

The different satellite AOD data sets are essentially not compared in a quantitative way. The quantitative comparison is essentially against AERONET and thus the different wavelength (865 versus 550 nm) is not a major issue since AERONET measures at 870 nm and many wavelengths in the visible. The smaller aerosol signal at 865 nm does not cause problems for the linear regression relationship constructed between POLDER and AERONET AODs. This is obvious from the high correlation coefficients for POLDER in Tables 3-5. Also POLDER reports AOD at 865 nm, but uses measurements at 670 nm in the AOD retrieval.

I had also been under the impression that the particular POLDER AOD retrieval data set the authors are using is intended to be only a fine-mode AOD retrieval, rather than a total-AOD retrieval, which further complicates things. However, I may be mistaken about that as I have not used POLDER data myself for a few years now.

Dr. Sayer makes an interesting point here. This is not a fine-mode AOD product; total AOD is retrieved and reported. See:
http://www.icare.univ-lille1.fr//projects_data/parasol/docs/Parasol_Level-2_format.pdf.
However, the use of polarized radiances in the POLDER retrieval greatly reduces the sensitivity of the retrieval to coarse particles. Thus, it is possible that a coarse-mode aerosol plume could, to some extent, mask the polarization signal from underlying fine-mode particles if such an arrangement occurred. Ultimately, the low sensitivity of POLDER to coarse-mode particles appears to be a minor issue at the two AERONET sites (Fort McMurray and Fort McKay) given the lack of bias and the high degree of correlation with AERONET AOD, in spite of the fact that coarse-mode dust is known to be significant contributor in this region, particularly at Fort McKay (based on AATSR dust fraction, not shown).

I note in the text that AERONET AOD was interpolated to the satellite wavelengths (which is the standard practice), but Table 4's caption says that AERONET data at 500 nm were used. I guess that this is an error in the caption, but can this be clarified?

Dr. Sayer is correct that this needs clarification, even though there is not an error. AERONET 500 nm AOD is used, however it is scaled to the satellite wavelengths.
In the caption, we now write:

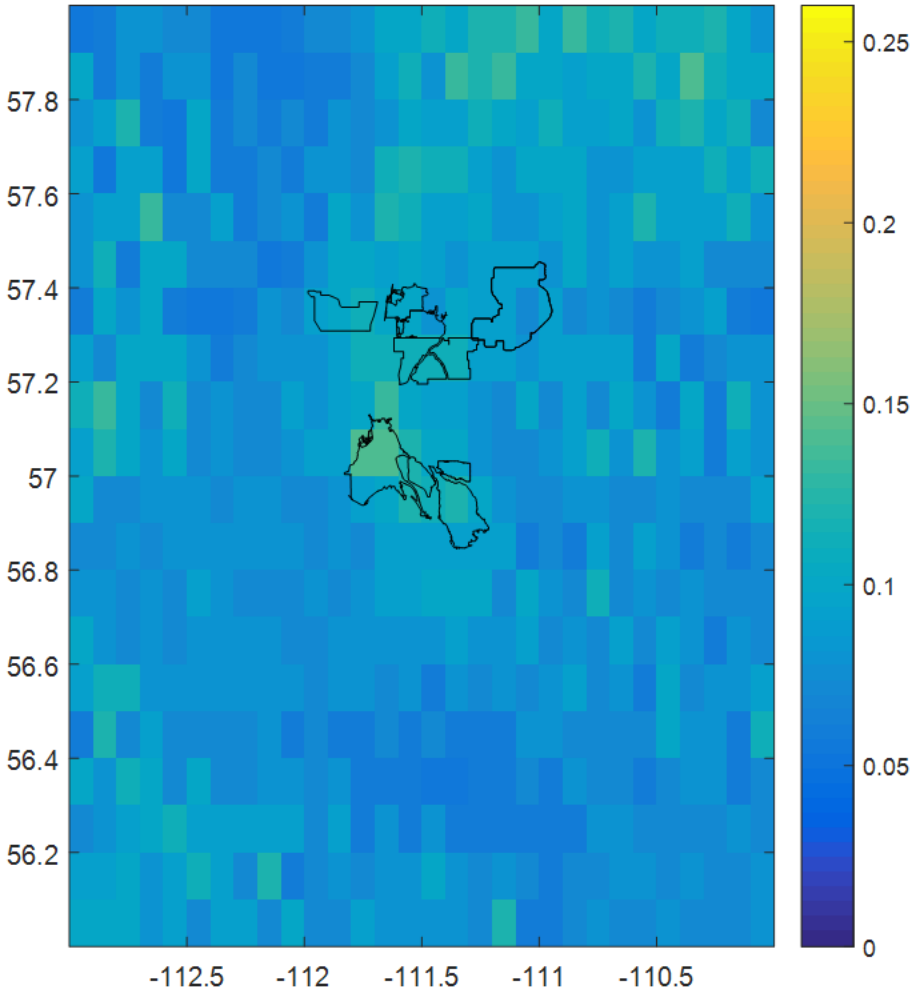
The Cimel 500 nm AOD, scaled to the satellite AOD wavelength (see Sect. 2), is used for comparison with all satellite sensors except POLDER/PARASOL, for which the Cimel 870 nm AOD is more appropriate (see Table 1).

In that case it might be better to allow the level 2 data to occupy multiple grid cells (corresponding to the actual retrieval footprint) than to snap them to the grid cell nearest to the pixel centre (which is what I assume is being done here). If the retrieval pixels are larger than the grid size (which is the case here) then it does not really make sense to assign a pixel to one grid cell, when it occupies multiple grid cells.

The orientation of actual footprint would need to be known and, for POLDER, this information is not available for each observation: only the latitude and longitude at the center of the AOD superpixel is provided. In general, we disagree that it does not make sense to assign a pixel to one grid cell. This is referred to as spatial oversampling and can be very revealing about localized sources of aerosols.

As a general comment on this figure, I would recommend keeping the colour scales the same (and ideally start at zero) to allow a direct comparison between the different data products. Right now it is hard to compare them because the colour bars are different. I realise POLDER is the odd one out here since it is at a longer wavelength, but the other data sets (at or near 550 nm) should be on a consistent scale. I'd also suggest mentioning again in the caption that POLDER is at 865 nm, hence the lower AODs.

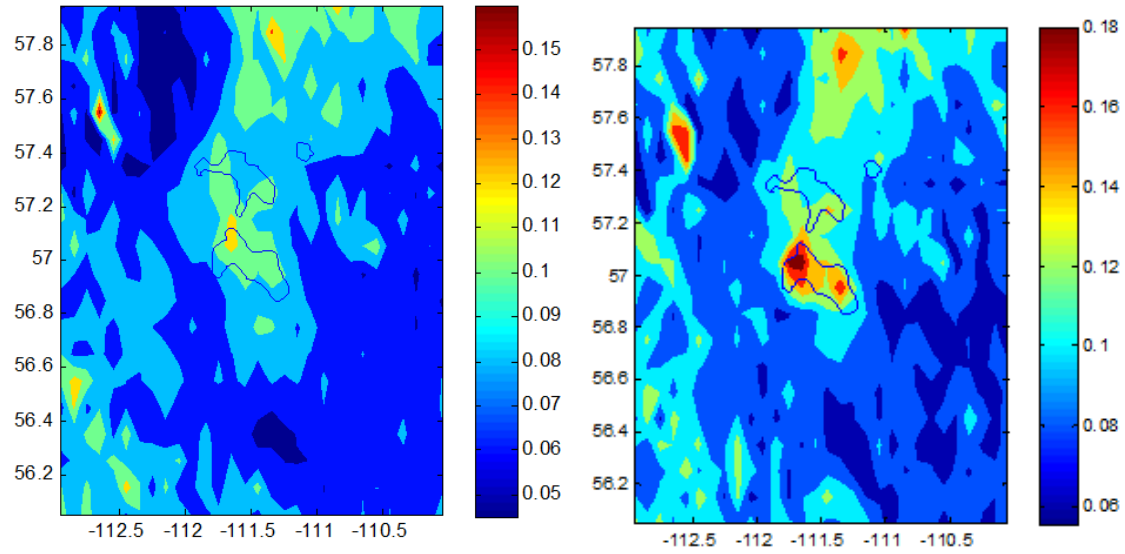
This recommendation initially seemed like a good one, but even the AOD differences between the MODIS products using the respective confidence values suggested by Dr. Sayer near the Syncrude facility are quite large, as shown in this Deep Blue climatological mean AOD map using confidence 2-3, but with the AOD range of the colour bar extending to 0.26 to cover the maximum climatological AOD of Dark Target (confidence=3).



Including such a figure would severely compromise our primary goal for Fig. 1, which is to show the spatial gradients in AOD in this region. The colour scales have been changed to have a common lower limit of 0.

We already mentioned in the caption that POLDER is at 865 nm: "(top left) POLDER 865 nm (1996-2013)".

Just as a point of information, the Deep Blue climatological AOD for confidence=3 has a hotspot near the Suncrude facility with AOD of 0.12, yet we find that higher climatological maximum AODs occur (0.18) when only confidences of 1-2 are retained, again with the hotspot being the Syncrude Mildred Lake facility, as shown in the following maps to the left and right, respectively.



As another general comment on the above figure: we know there is seasonal variation in AOD, as well as variation in things that affect sampling (e.g. cloud and snow cover). So presenting an annual mean here conflates these issues together with the issue of retrieval uncertainty. My suggestion would be to make separate maps for each season. They don't all necessarily need to be included in the paper if length is a concern. This way the seasonal aspect at least can be removed and it may bring the different data sets into closer agreement (or it might not). The next stage would be to compare the points only where they have common retrievals on the same days, but I suspect that due to the large number of data sets there would probably be few mutual points. So, making seasonal means rather than annual means is probably a good balance in terms of seeing how the data look compared to each other.

We tried plotting AODs for May through September for the MODIS and MISR products (Figures A-C below). These are the months when all five aerosol products have high measurement frequency. But again, Dr. Sayer's purpose is evidently different than ours: we are not trying to bring the different data sets into closer agreement; as stated up front (p4L33), we are mostly trying to see what each is capturing spatially over the long term, so annual means are preferable. Anyway, as shown in Figures A-C below, limiting to these 'warm season' months does not bring the data sets into closer agreement. Limiting to the warm season was mostly expected to benefit the MISR AOD map since MISR has an unusual spatiotemporal sampling pattern, but as shown in Figure C, limiting to May-September does not produce a more coherent AOD map. In the revised manuscript, all available months are retained for Figure 1.

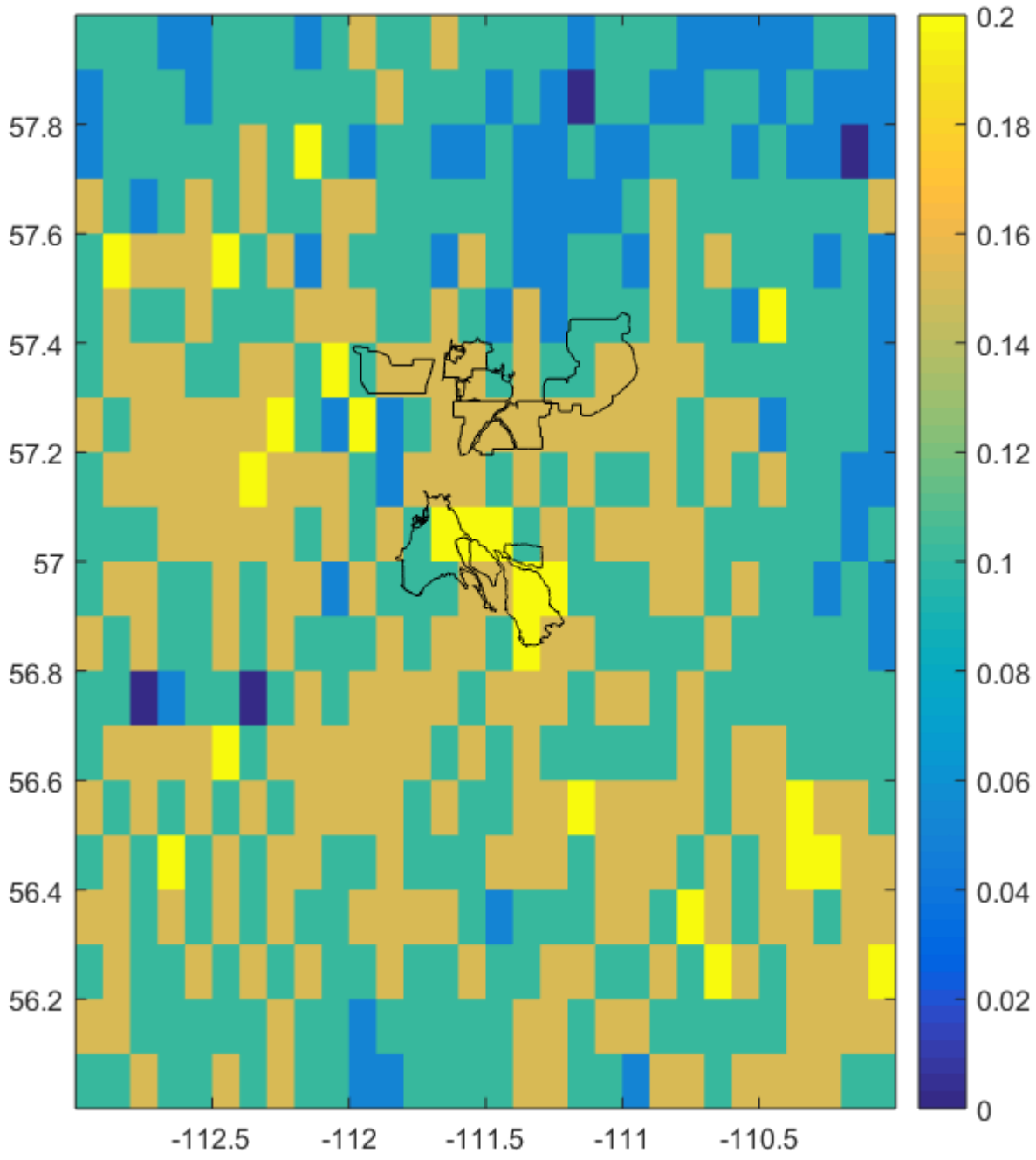


Figure A: MODIS DT 550 nm climatological AOD for May to September (confidence=3).

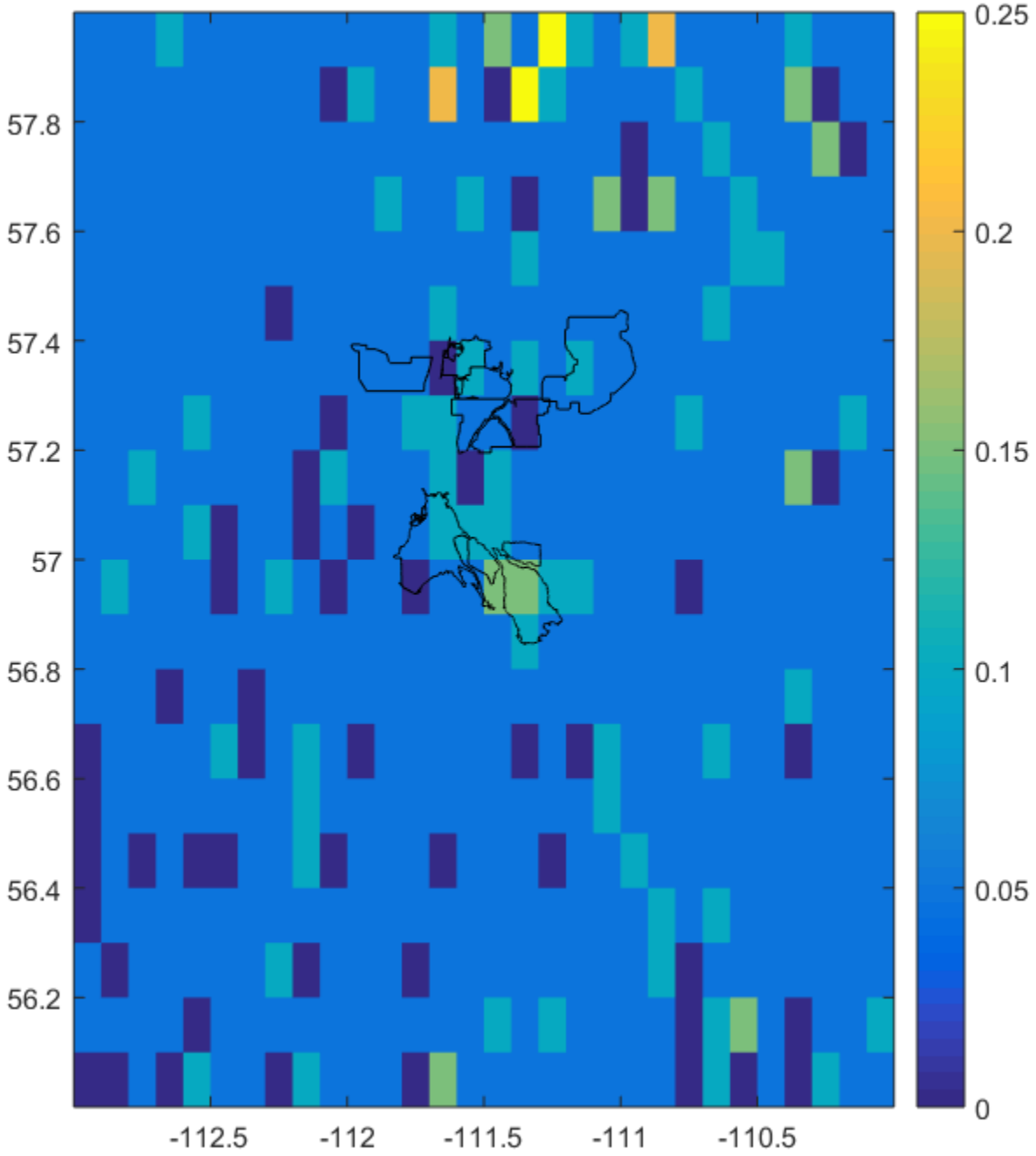


Figure B: MODIS DB 550 nm climatological AOD for May to September for confidence≥2.

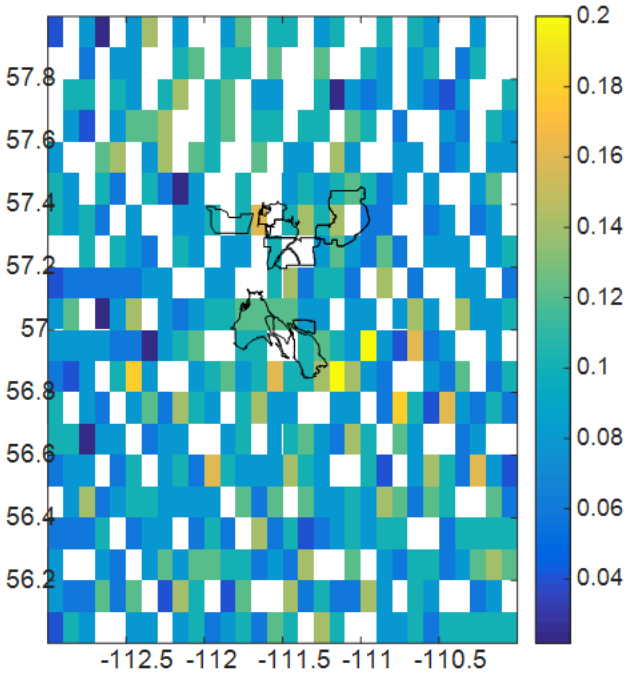


Figure C: MISR 558 nm climatological AOD for May to September.

Figure 2: If I understand correctly, this is the mean of the MODIS Deep Blue and Dark Target QA values. I understand the intent behind this figure (illustrate where the algorithms have confidence) but I think the execution is problematic. By taking the mean of the QA flag, it is being treated as a quantitative variable. However it is not – it is a categorical variable that is stored as an integer because it is easy to store integers in the hdf files. QA=0 has a fundamentally different meaning (no retrieval) from the other values, and the QA from 1 to 3 does not represent linear progression in terms of quantitative retrieval quality or uncertainty. So, taking the mean value is a bit misleading since it is conflating lack of retrievals (due to e.g. clouds) with other algorithm factors and giving a number as a mean for the grid cell which doesn't really relate to the underlying QA flags. For example if the mean QA calculated in this way is 1, it does not mean that the retrievals here have low confidence. It means either that the retrievals have low confidence, or that there is some combination of high confidence retrievals and data gaps due to clouds, etc.

This comment by Dr. Sayer is correct, and we were aware of all of these logical points. The main purpose of both panels of Fig. 2 was to show that QA is tending very close to 0 (i.e. <0.45) at the two grid cells near the Syncrude Mildred Lake facility, implying that the retrieval has no confidence (or provides a fill value) more than 55% of the time.

So, I think this figure should be updated, and we might get some more insight into what is going on if the metric here is calculated differently. In Deep Blue we recommend QA=2 and QA=3 can both be used for quantitative analyses as they have similar error characteristics (Sayer et al., JGR 2013, doi: 10.1002/jgrd.50600) while for Dark Target land retrievals they recommend QA=3 only (e.g. Levy et al, ACP 2010, doi:10.5194/acp-10-10399-2010). This is another example of the fact that QA flags

have different specific meanings for different data products. What I would suggest is making maps showing the fraction of overpasses where there is no retrieval (i.e. QA=0), the fraction where there is a poor-QA retrieval (i.e. QA=1 for Deep Blue, QA=1 or 2 for Dark Target), and the fraction where there is a good-QA retrieval (i.e. QA=2 or 3 for Deep Blue, QA=3 for Dark Target).

This suggestion is accepted. A new six-panel Fig. 2 has been generated.

Some of the data holes in the MODIS Dark Target product will be from the fact that neither their land nor ocean algorithms treat pixels which are identified as 'coastal' as valid for AOD retrieval. (Note that Deep Blue treats such pixels as land, but excludes pixels next to water frequently for other reasons.) This limits coverage in many parts of Canada and elsewhere in the world, as pixels containing lake shores are frequently identified as coastal. See Carroll et al. (IJDE, 2016, doi: 10.1080/17538947.2016.1232756).

This cause of data holes has been added to the list of causes. We now write at p5L16:

The number of pixels used in the AOD retrieval is reduced by the inland water mask (Carroll et al., 2016), …

Figure 3: This shows that in areas where there are few AATSR retrievals, those retrievals that are performed tend to have a higher sub-pixel cloud fraction. The implication is that sampling in this area is influenced by cloud cover, whether real cloud or misidentified cloud (which is reasonable). However what might make a better right panel would be the cloud fraction for ALL observations, not just for those observations where an AOD retrieval is performed. This would look more directly at where the AATSR algorithm thinks there is a cloud. Right now what the panel is showing is subtly different since pixels which are cloudy above the threshold for retrieval (I am not sure if this is 100% cloudy or some lower fraction) are exclude from the analysis.

Additional cloud tests (Bevan et al., 2012 and reference therein) were used for this AATSR aerosol retrieval algorithm that are not used in AATSR Instrument Processing Facility (IPF) v6.01 cloud product. Thus, we feel it is more appropriate to look at the cloud fractions in the successful AOD retrievals. This suggestion might have been worth pursuing if the spatial anti-correlation was not strong between cloud fraction in successful AOD retrievals and AOD sample size, but that is not the case.

Table 1: Again, the MODIS standard AOD wavelengths for both Deep Blue and Dark Target are 550 nm. Deep Blue also provides 412, 470, and 650 nm and Dark Target also provides 470 and 650 nm. Source radiances are not all at 0.5 km pixel sizes, it depends on band, so it would be better to say 0.25-1 km here. Also, due to its scan design and wide swath with, MODIS level 1 and level 2 pixel size and shape get heavily distorted from nadir to scan edge (quoted values are all for nadir pixels), which is not an issue for AATSR or MISR to the same degree due to their designs and narrower swaths. See e.g. Sayer et al (AMT, 2015, doi:10.5194/amt-8-5277-2015) for more information.

In Table 1, regarding the spatial resolution of MODIS radiances, we now write: 0.25×0.25 to 1×1.
We have also changed one column heading to: "Spatial resolution of AOD superpixel at nadir".

Table 4 and discussion: I would delete the analysis of linear least-squares regressions from the table and discussion. AOD data violate most/all the assumptions required for this technique to be valid, and so the results are misleading and fits/confidence envelopes are quantitatively incorrect. See e.g. http://people.duke.edu/~rnau/testing.htm for more discussion. (I know it is a frequently-used technique in our community, but it is fundamentally incorrect for this particular application.)

Dr. Sayer's most recent paper (Carroll et al., 2016) cites Levy et al. (2013) for AOD validation, and Dr. Sayer is also a co-author in the latter work. This latter work includes linear least-squares regression of MODIS AOD and AERONET AOD (their Fig. 11), which is precisely what we have done. It is clear that our Table 4 adheres to the established convention in this field in terms of validation statistics. As an alternative, we tested two non-parametric methods (Theil's complete and incomplete methods) to obtain the values in the first three columns of values in Table 4. None of the assumptions are violated when using Theil's incomplete method (1950). Also, application of Spearman's rank correlation is valid for this application (see Table 4). As shown in the table below, the non-parametric methods yielded slopes that were small (~0.6) and ordinary least-squares ('OLS') yielded a slope that was clearly of the wrong sign due to one small cluster of outliers at high AOD. We tested a number of robust regression methods compared in Holland and Welsch (1977), which all use a weighted least-squares (WLS) approach to reduce the sensitivity to anomalous data pairs (i.e. coincidences). Some of these robust regression methods are expected to perform better than OLS on data with non-Gaussian distributions (e.g. Andrews, 1974). The outliers affect whether the AERONET and satellite AOD data conform to a normal distribution. The table below presents the slope and offset from various robust methods using the POLDER/PARASOL and AERONET coincident data at Fort McMurray:

Method	**offset**	**slope**
Andrews	-0.017	0.787
bisquare	-0.017	0.788
Cauchy	-0.017	0.797
Fair	-0.019	0.859
Huber	-0.018	0.831
logistic	-0.018	0.835
Talwar	-0.017	0.787
Welsch	-0.017	0.791
OLS	-0.030	1.10
Theil's "incomplete"	-0.009	0.590
Theil's "complete"	-0.010	0.620

It is clear that POLDER has a negative offset, but the magnitude of the offset falls into three groups: OLS, robust WLS methods (first eight rows of table above) and robust non-parametric methods. Identical groupings of regression methods are found upon examining the slope values. Furthermore, omitting the small cluster of points with AERONET AOD>0.8, which were all measured on one day, namely 16 July 2012, the OLS slope becomes 0.7904 and offset is -0.014. This slope and offset are both very close to the slope and offset values from the various WLS fits. In the revised manuscript, we select to weight the fit residuals with Huber's function, for the following reason given by Bergstrom and Edlund (2013):

"while it still is robust, it does not completely disregard highly deviating points".

The table above shows that neither the offset nor the slope obtained with the Huber weights are outliers within the WLS group of robust regression methods. The tuning constant is assumed to be 1.345 following Holland and Welsch (1977).
At p3L29 of the revised manuscript, we now write:

Since individual AERONET and satellite AODs are not normally distributed, we use linear least-squares weighted by Huber's function to determine the slope and offset since this is a robust method that does not completely disregard highly deviating points (Bergström and Edlund, 2014). The slope and offset values determined using Huber's weighting function are encompassed by the values obtained with seven alternative weighting functions. Similarly, due to the non-normal distribution of the individual AOD data, Spearman's rank correlation coefficient (r_s) is chosen to study the site-specific AOD correlation based on individual AERONET-satellite coincidences.

References

Andrews, D. A., A robust method for multiple linear regression, Technometrics, 16(4), 523-531, 1974.

Bergström, P., and Edlund, O.: Robust registration of point sets using iteratively reweighted least squares, Comput. Optim. Appl., 58, 543–561, 2014.

Theil, H.: A rank-invariant method for linear and polynomial regression analysis: I, Proc. Kon. Ned. Akad. Wetensch., 53, 386-392, 1950.