Atmos. Meas. Tech. Discuss., 8, C2565–C2574, 2015
www.atmos-meas-tech-discuss.net/8/C2565/2015/


Atmospheric Measurement Techniques Discussions

# *Interactive comment on* "Known and unknown unknowns: the application of ensemble techniques to uncertainty estimation in satellite remote sensing data" *by* A. C. Povey and R. G. Grainger

**A. M. Sayer (Referee)**

andrew.sayer@nasa.gov

Received and published: 13 August 2015

I am posting this review under my own name (Andrew Sayer) because I have worked with the authors closely, including on some of the data products discussed within the manuscript.



## 1 Summary and recommendation

This paper discusses different types of uncertainties in Earth Science data sets, and how they are represented. Although presented in a broadly-applicable manner, the focus of discussions is on aerosol and cloud property retrievals. These issues are ones which data product creators and users are (hopefully) aware of, to greater or lesser extents, but may not have fully thought through. Discussions of this type are often omitted in the literature, in part due to length concerns in journal articles when performing evaluations of the data. The subject matter is also important because many data producers and users are not as familiar as would be desired with some of the statistical terms and tests, and assumptions built into them, which are used regularly when dealing with these data sets and uncertainties.

The paper is therefore important. It is also timely because in recent years many such data providers have either begun to include, or are being required to include, uncertainty estimates in new data product versions. So, it is good to have something like this as a point of reference, and help improve the representation of uncertainty within these data sets going forwards. The paper is laid out well, but by the nature of the subject matter it is quite mathematical, which may put off some readers who are less familiar with the intricacies of retrievals and just want to use a data set. The authors have done quite a good job at keeping things accessible for those readers with figures to illustrate different concepts.

Overall, this is a good paper. Although I agree with the bulk of what the authors are saying I have a number of specific comments, below. Some of them include additional references which may be relevant to the discussion – I realise that it is not practical to list every single relevant example on a generalised topic like this, so will leave it up to the authors which one(s) to include. I favour publication in AMT after revisions to

address these comments. I am happy to review a revised version if requested.

## 2 Comments

I feel that the title does not really reflect the contents well, and would suggest removing "*the application of ensemble techniques to*" from it. Ensemble techniques are suggested as a way to propagate uncertainties which don't fit standard techniques through the retrieval system. However their discussion in the manuscript is limited to about 4 pages out of about 34 in the main body. No examples of using those ensemble techniques for Earth Science applications are shown. I will also point out that sensitivity studies sometimes contained within ATBDs (more rarely journal articles due to space concerns) are often of a form similar to the ensemble techniques discussed here (i.e. perturb input states and examine the dispersion of retrievals). Really, the paper is mostly about different components contributing to uncertainty, and difficulties in comparing different data sets (16 pages, counting sections 3 and 4). Despite this, in my view the paper stands on its own as uncertainty is an important topic. I would like to see a follow-up study in which these ensemble techniques are applied to aerosol/cloud retrieval data sets, and maybe act as a blueprint for others to follow.

P8516, start of section 2.4: "*The standard error propagation techniques*" – I would prepend "As illustrated above," to this to improve the continuity of the text, and also to direct a reader who may have jumped in at this point (looking for information on ensemble techniques) to the importance of the previous section.

P8519: As far as I could see the acronyms MISR, ORAC, SeaWIFS, and NWP are



first used on this page so should be defined at the appropriate points.

P8520: Same comment about SST.

P8521, L16: and onwards. The authors point out that on-board calibration can degrade over time, and then state that vicarious techniques can be used as well. I think that it would be better to further and make it explicit that it is not an either/or, but rather both on-board and vicarious techniques are used routinely for many sensors. Some people may not know. Additionally, the discussion mentions Earth targets. However, the Moon is often used as well, because of its stability. One relevant reference for this which could be added is Eplee *et al*. (2011) (see references at the end of this review).

P8522, section 3.3: Here, the authors define approximation errors, giving as two examples the plane parallel approximation, and the use of lookup tables (LUTs) of radiative transfer calculations (to speed up retrievals). While it is certainly correct to call these approximations, in my view they are different from each other. The plane parallel assumption is a physical approximation of how the world works, while LUTs (and interpolating them during retrievals) are more of a numerical approximation based on that physical model of the world. For the example of aerosol optical depth retrievals, would the assumed aerosol optical model (from e.g. AERONET or OPAC) count as a parameter error or an approximation error in this formulation? What about the use of Mie or T-matrix theory to calculate aerosol optical properties when the "real" aerosols might not conform to the assumed shapes or internal homogeneity; are these parameter or approximation errors? So I wonder if there should be some additional/alternative distinction drawn here as I do not feel the distinction drawn by the authors is clear-cut. At least one could note that there are different kinds of approximation error (e.g. the physical vs. numerical distinction I make).

Section 3.4: there is an additional aspect of superpixelling which may be relevant to discuss here. The current discussion talks about averaging data at the L1 stage, which will give you a "radiative average" retrieval. Most aerosol retrieval algorithms do some variant of this. On the other hand, Deep Blue does the retrieval at sensor-pixel level, and then averages the retrieved quantities (such as aerosol optical depth) to the coarser level 2 resolution. This will give a "state space average" retrieval. If the radiometric noise is small and linear, and the scene is homogeneous, the two should be equivalent; if not, they may not. Which one is more desirable may be dependent on the particular application. See e.g. Hsu *et al.* (2013), Sayer *et al.* (2014) for more information.

In the same section, I think there are a few other points which could be mentioned. Relating to Level 3 data, some sensors (e.g. AVHRR, MODIS, VIIRS) have swaths which are sufficiently broad that there is overlap between consecutive orbits. If Level 3 data are created then there is some (albeit probably small) element of the diurnal cycle aliased in as well. For example consecutive MODIS orbits are about 90 minutes apart. The nominal local Equatorial crossing times quoted for sensors are for their sub-satellite tracks, which is not the local time across the whole swath (particularly for a broad-swath imager). So across the swath the local time is variable, and at the edges data from different local times may be being averaged in the same place.

For those same sensors, there is also often a distortion of pixel size, shape, and overlap near the edge of the swath (e.g. the MODIS "bowtie effect"). This then aliases into changes in the spatial resolution of the data as a function of scan angle, which can influence statistics of retrievals (variance of distributions decreases as pixel size increases, due to smoothing), and also the independence of adjacent retrievals (as sensor pixels can overlap). We have a paper discussing the implications of this for MODIS aerosol retrievals which is being typeset for ATMD at present, and may be of



interest (hopefully will appear soon). The main point here is that the spatial resolution of these data sets is not the same everywhere. This also has implications for Level 3 data, e.g. when you have pixels of different sizes from overlapping swaths should you take the simple mean or a weighted mean or only the finer-resolution orbit parts or what?

P8524, lines 21-25: I am not sure that I agree with this definition of the independent pixel approximation (IPA). Perhaps it is a matter of wording. What the authors describe ("*a constant quantity at the pixel scale*") sounds more like a homogeneous pixel assumption. My understanding of the IPA (from the Chambers paper cited, and earlier work by Cahalan and others) was more like each pixel being considered independent of (i.e. not influenced by the conditions within) other nearby pixels – for example 1D vs. 3D radiative transfer. Of course the assumption that pixels are homogeneous is important and is buried in retrievals as well, but that's not what I consider the IPA to be. In this context the IPA is also maybe more like an approximation error from section 3.3 ("I approximate by making the assumption that the radiative transfer required to represent a pixel is independent of the state of other nearby pixels") than a resolution error.

P8259 (top). It may or may not be worth mentioning that vertical weighting functions aren't just an issue for things like CTH. The weighting function in solar bands can affect the apparent cloud effective radius as well, in clouds where droplet size varies with depth into cloud, due to differences in weighting functions in swIR bands (e.g. Platnick, 2000).

P8529 (end)-8530 (top): a practical application of this type of correction can be found in Sayer *et al.* (2011), validating CTH from satellites against ground-based cloud radar. We applied a correction to the radar data to convert the radar CTH to an estimated IR effective radiative CTH. The difference between the two, as the authors point out here,

can be non-negligible.

Section 4.1.2: This is an important point about comparing different data sets. I wonder, though, if some readers may skip over it because of unfamiliarity with the mathematics. So it would be good to make it less (potentially) intimidating. Figure 6 is good in this regard. However I think a panel (c) should be added to more explicitly illustrate a weighting function changing in a radiometer with a wider band (discussed in the text).

Section 4.1.3: I agree with many of the points the authors make here, however, I feel that the discussion is a little one-sided. The authors say that these comparisons (e.g. error envelopes from satellite vs. AERONET) are not an estimation of uncertainty, but why not? It isn't clear to me whether this is a largely semantic argument (based on the universal/consistent/transferable criteria) or whether the authors are stating that this is a fundamentally flawed way to figure out how good a data set it. I think that this argument should be clarified/backed up. I would argue that, with caveats (related to e.g. sampling and validation protocol), they do allow an estimation of uncertainty (since you are comparing to something which can be considered a reasonable approximation of truth). The fact that there are problems using these estimates for some applications means that it is not always a good estimate, but I don't think that it means that the technique lacks value. We know that AERONET and other Sun photometers can give much better estimates of AOD than spaceborne instruments. So while these comparisons can't characterise the uncertainty fully, is it not correct that they can give us an estimate of it? I do agree it is better to approach uncertainty from a theoretical standpoint and propagate errors through the system, and then compare these to results from validation exercises, but I don't see that this means the validation results aren't an estimate of your uncertainty.

For the specific aerosol examples mentioned (MISR and MODIS Dark Target AOD



error envelopes), these are defined relative to AERONET "truth" AOD which makes them diagnostic. (I feel it is reasonable to consider AERONET a truth, with some caveats, since the direct-Sun measurements have uncertainty comparable to that we'd generally expect to incur from the satellite measurements based on radiometric calibration alone, even if we had a "perfect" forward model.) Since we know that a large portion (most?) of the uncertainty in satellite AOD observations is systematic (contextual) rather than random, using these diagnostic error envelopes as a basis to compare products will run into problems because when trying to merge data sets you don't know the "true" AOD to compute the error envelope from. That's why efforts like the NRL Data Assimilation Grade MODIS aerosol product perform additional corrections and filtering before computing prognostic (i.e. defined relative to retrieved AOD) uncertainty estimates for assimilation. In short, diagnostic and prognostic uncertainty estimates are different things and should be used for different applications. This also motivated the decision to provide prognostic pixel-level AOD uncertainty estimates in the recent MODIS Collection 6 Deep Blue aerosol products. See for example Zhang and Reid (2006), Sayer *et al.* (2013), and others (several more papers by the NRL group).

I also don't think it is appropriate to refer to these as "*single uncertainty values*", as the authors do on P8534, because they are state-dependent (e.g. in this example they are envelopes which depend explicitly on the AOD, and for Deep Blue, also on solar/observational geometry). I will agree that they need improvement because they are not sufficiently context-dependent to represent the uncertainty of retrievals, but I think that it is misleading to imply that products are providing a "single uncertainty value". Maybe "simple uncertainty expression" would be a better phrase.

I will readily agree, however, that it is unfortunate that easy-to-write formulations like this can be taken to mean more that they do. For example I often see these error

envelopes referred to in descriptions of Level 3 data uncertainty, which completely ignores issues relating to e.g. random vs. systematic error and sampling.

P8537, lines 3-10: I believe the statement about low magnitude retrievals being assigned quality flag 1 refers to the MODIS ocean aerosol retrieval. In that case, this is no longer correct for MODIS Collection 6, which was released in 2014 (although the general point remains valid).

P8537, lines 15-16: I like this turn of phrase about "*the disinterested user*".

Section 5.3: I would suggest renaming this to something like "Distinction between maturity and uncertainty" or similar, since the discussion seems to focus on how the system maturity matrix may not be relevant to satellite remote sensing data.

Figure 3: I wonder if this might be clearer as a set of 5 panels rather than a set of stacked plots.

General grammar: the authors use "*data is*"; I am not sure whether Copernicus style requires "*data are*". Also, there are a few cases of split infinitives (e.g. "*widely used*" on P8511, L25).

## 3 References


- R. E. Eplee, Jr., J.-Q. Sun, G. Meister, F. S. Patt, X. Xiong, and C. R. McClain (2011), Cross calibration of SeaWiFS and MODIS using on-orbit observations of





the Moon, Appl. Opt. 50, 120-133, doi: 10.1364/AO.50.000120.

- Hsu, N. C., M.-J. Jeong, C. Bettenhausen, A. M. Sayer, R. Hansell, C. S. Seftor, J. Huang, and S.-C. Tsay (2013), Enhanced Deep Blue aerosol retrieval algorithm: The second generation, J. Geophys. Res. Atmos., 118, 9296–9315, doi:10.1002/jgrd.50712.
- Platnick, S. (2000), Vertical photon transport in cloud remote sensing problems, J. Geophys. Res., 105(D18), 22919–22935, doi:10.1029/2000JD900333.
- Sayer, A. M., Poulsen, C. A., Arnold, C., Campmany, E., Dean, S., Ewen, G. B. L., Grainger, R. G., Lawrence, B. N., Siddans, R., Thomas, G. E., and Watts, P. D. (2011), Global retrieval of ATSR cloud parameters and evaluation (GRAPE): dataset assessment, Atmos. Chem. Phys., 11, 3913-3936, doi:10.5194/acp-11-3913-2011.
- Sayer, A. M., N. C. Hsu, C. Bettenhausen, and M.-J. Jeong (2013), Validation and uncertainty estimates for MODIS Collection 6 "Deep Blue" aerosol data, J. Geophys. Res. Atmos., 118, 7864–7872, doi:10.1002/jgrd.50600.
- Sayer, A. M., L. A. Munchak, N. C. Hsu, R. C. Levy, C. Bettenhausen, and M.-J. Jeong (2014), MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and "merged" data sets, and usage recommendations, J. Geophys. Res. Atmos., 119, 13,965–13,989, doi:10.1002/2014JD022453.
- Zhang, J., and J. S. Reid (2006), MODIS aerosol product analysis for data assimilation: Assessment of over-ocean level 2 aerosol optical thickness retrievals, J. Geophys. Res., 111, D22207, doi:10.1029/2005JD006898.


Interactive comment on Atmos. Meas. Tech. Discuss., 8, 8509, 2015.