Anonymous Referee #1

Received and published: 21 September 2017

This paper describes a change to the MISR aerosol retrieval algorithm. They select an ensemble of aerosol types and, for each, compute the radiances that would be observed at a range of aerosol optical depths (AOD). Previously, ensemble members were evaluated separately so each gave an AOD and cost, which were then filtered and averaged to calculate the final product. This paper proposes minimising a single cost function (being the sum of the individual cost functions) to find the AOD and it's uncertainty. The technique is rationalised based on two months of observations and is shown to produce more believable uncertainties, on average, than the previous algorithm.

I recommend this paper for publication after minor revisions. The technique proposed is definitely a step in the right direction and the paper is superbly drafted. However, the technique and description thereof could be improved by a more statistical approach. The paper justifies itself with qualitative descriptions of global averages and internal metrics rather than any validation activity, which is common but always disappointing. Specific comments on the paper are listed below, with some minor details collected at the end. The notation PxLy refers to line y of page x.

- My experience is in optimisation. One defines a cost function and selects an algorithm to efficiently search the 'surface' of that function for its global minima. The uncertainty is a measure of the 'width' of that minima in multi-dimensional space (i.e. the magnitude by which a variable could be changed without significantly increasing the cost). The cost function is usually the RMS difference between some modelled value and a measurement. If the model is accurate and the measurement suffers only random noise (of known variance), the minimal value of the cost function will sample a χ^2 distribution, from which one can determine the probability that this measurement fit that model.
  To me, this paper essentially proposes that $f(\tau)$ is a probability density function (PDF) for AOD and that it is normally distributed. It follows that the most likely AOD is the τ that maximises f and the uncertainty is the function's width. The proposed ARCI threshold can then be understood as eliminating retrievals that are exceedingly unlikely. Describing the problem with these basic statistical concepts could vastly simplify the paper, avoiding awkward phrasing like P8L5.
  Re: It is a very valuable observation. We added the following clarification below Eq. 4. "The function f can be interpreted as a probability density function (PDF) for AOD. The most likely AOD is the one that maximizes f (Eq. 4), and the retrieval uncertainty is related to the width of the PDF."
  We also modified the somehow awkward phrasing in P8L5 to read: "Large ARCI, on the other hand, means that for some models sufficiently low χ^2_{abs} were obtained, signifying good agreement with the observations."

- Because this is a fairly straightforward statistical problem, there exists a variety of tools to check that (a) f is in fact a good model of the PDF, (b) f is normally distributed, and (c) the selected aerosol models are an unbiased sampling of the complete state space of real-world aerosols. A brief discussion of some of those points could provide a standardised means to evaluate your assumptions and avoid qualitative judgements, such as the function 'closely resembles a Gaussian' (P7L12).
  Re: In the process of designing and testing the new approach, at one point we did fit a normal distribution to our PDF results. We compared most likely AODs retrieved from PDFs against those retrieved from the fitted normal distributions. The results were in

excellent agreement. This exercise gave us confidence that, at least in those cases that we considered, the PDFs closely resembled Gaussian distributions. However, we though this analysis was too technical to be included in the manuscript. As for point (c), we write in the manuscript that the resulting uncertainty is dependent on the LUT considered in the retrieval (P10L30-35) and that the 74 mixtures currently included in MISR retrieval process are not complete (P4L35-38).

- Are you tabulating f as a function of linear or log τ? Figure 1 uses both as an x-axis, which is misleading. It should be logarithmic as AOD is log-normally distributed (which is clear from the asymmetry about τ_{max} in Fig.1(3)). If you're using linear space, you will underestimate the uncertainty and overestimate the mean.
  Re: All equations in the manuscript use linear τ. In Figure 1a we use the logarithmic scale in the x-axis to better visualize the cost functions at very low τ. Because after inverting the cost functions, at low τ the signal becomes very small, the log scale is no longer necessary. We added additional clarification regarding the x-axis scale in the caption. The distribution in Fig. 1c is close to Gaussian. The misleading resemblance to a log-normal distribution comes from the fact that the PDF is truncated on the left side due to the physical constraint (τ>0.0).

- Why is there no validation of the new algorithm? It seems fairly substantial to move from averaging a few aerosol types per pixel to averaging 74. A few comparisons against AERONET or MODIS would be fine for a paper like this. A simple comparison of V22 vs. V23 would be a start, considering you did it for the uncertainty!
  Re: A validation paper is currently under preparation. It was our intention to designate external validation efforts to a separate publication. One reason for this is that, at the time of preparing this manuscript, we only had two months of data available, which is not enough to obtain sufficient number of collocations with ground based observations. Furthermore, we plan to investigate the new AODs and their pixel-level uncertainties in greater detail, which we feel justifies a separate study.
  Our analysis indicates that the new algorithm leads to AODs that are similar, but not identical, to those obtained using thresholds from V22. However, the uncertainty quantification in the new approach is sufficiently different from V22 to justify a comparison figure (Fig. 7).

- In Sec. 3, you implicitly assume that the choice of aerosol type overwhelms any measurement error. Could Fig. 1 be adapted to show the sensitivity of a χ^2 curve to typical measurement error? I'd expect it to move the curve slightly, but much less than the spread between curves.
  Re: The measurement error is embedded in the calculation of χ^2_{abs} (Eq. 2). The absolute radiometric uncertainty σ_{abs} in V22 is set to 5% of the signal itself for each camera and wavelength (P5L16). We feel that showing the sensitivity of χ^2_{abs} to different levels of σ_{abs} would decrease the clarity of the figure.

P10L24 I'm unhappy with this paragraph.

- L27 I think this is trying to distinguish between a validation activity, which you sadly aren't doing, and an uncertainty estimate, which you are. By definition, uncertainty is a parameter describing the range of values that can be reasonably ascribed to the quantity that is being measured. I believe that provides a 'measure of how far the retrieved AOD deviates from the "truth"'. The distinction is that uncertainty is a prediction of that difference while validation is a direct calculation of it.

Re: What we are trying to distinguish here is the algorithmic retrieval uncertainty on the one hand, and the uncertainty that comes from comparing a retrieved AOD with ground truth on the other hand. In both cases we are considering pixel-level information, or individual retrievals, rather than a bulk validation metric like the error envelope. Yes, we are predicting an uncertainty in our algorithm, but this prediction might not necessarily represent the real range of values that are being measured. We are trying to be cautious here and not assign undue credit and value to the algorithm's prediction. A validation activity is required to establish the relationship between the reported uncertainties and the ground truth. Because we think this is a challenging task, we left it for a separate investigation. Our initial results, however, show very promising linkage between our reported uncertainty and the standard deviation of a normally distributed error function.

- It's good to be clear that the estimated uncertainty is sensitive to the way you solve the problem. However, you don't tell the user what to do with that information. I think a rational response at the moment is to avoid MISR data as it's more sensitive to your assumptions than the environment. I can think of three approaches to remedy this:
  1. Give up and declare that your uncertainty values are uncalibrated, providing a pixel-by-pixel assessment of the relative reliability. (I'd recommend that you normalise the values to clarify that their magnitude is not inherently meaningful.)
  2. Show that, despite the algorithm's theoretical sensitivity to your assumptions, the uncertainties you produce are an approximation of the true error. This would be done through a validation activity (e.g. the distribution of $(\tau_{MISR} - \tau_{AERONET})^2 = \sigma^2_{MISR}$ is approximately normal).
  3. Demonstrate that the sensitivity to your assumptions is small. The precise choice of types is a matter for another paper, but it's important to quantify the uncertainty's sensitivity to it. A straightforward way to do so would be re-running the retrieval with a few types removed at random.

  Re: Indeed, at the moment our uncertainty values are uncalibrated. But this is a temporary position that will be resolved in a separate investigation. In order to validate our uncertainties, large comparison statistics against ground truth are required. As mentioned above, at the time of writing we did not have enough data (two months of retrievals) and enough collocations against AERONET to perform a detailed evaluation of the retrieved parameters. This activity will be performed along with the reprocessing of the MISR mission with the new V23 version of the aerosol product.

- Sec. 4 argues that this method is good because it excludes high AOD retrievals. Could you provide some evidence that, for the two months of data you've considered, there were no large aerosol events?
  Re: Figure 5 shows the global distribution of AOD with ARCI screening for January and July of 2007. There are high AOD regions visible off the west coast of Africa and off the coasts of India and China. These are associated with high-AOD events such as dust outflow from Africa, biomass burning, and anthropogenic emissions. We write in the manuscript: (P10L7) "At the same time, climatologically large AODs off the coasts of Africa and South and East Asia are retained, indicating that the new screening method does not unintentionally remove all high AODs that are likely valid."

P3L17 The spread of the MISR ensemble is providing a quantitative insight into the uncertainty in each retrieval due to the assumptions made. While the description of ensemble techniques at L9 is technically correct, ensemble techniques are used to estimate uncertainties that can't be accurately or efficiently calculated by other means. It's exceedingly rare to perturb more than one of the input data, auxiliary parameters, and

underlying assumptions. Numerical weather prediction perturbs its input data in order to estimate the sensitivity of a chaotic system. Climate models perturb the auxiliary parameters because they are unknown. MISR perturbs the assumed aerosol type because the radiances available don't fully constrain the problem. MISR doesn't need to perturb the input data as the physics of remote sensing are sufficiently linear that error propagation does a reasonable job of estimating the uncertainty due to measurement error. Hence, I wouldn't agree that extending ensembles to 'all possible sources of error' would be overly useful. Ensemble techniques are used to quantify uncertainties due to poorly understood, poorly constrained, or exceedingly non-linear error sources.

Re: We modified this sentence to read: "Such an approach, if extended to all poorly quantifiable nonlinear sources of error and physically plausible realizations of parameter space, has the potential of providing a robust and comprehensive measure of retrieval uncertainty in the manner suggested by Povey and Grainger (2015)."

P8L38 Within this paper, the only evidence that the cloud filtering is effective is showing that mean AOD is lower. MISR is on the same platform as a MODIS, so you have the ability to check if your cloud flagging spatially agrees with them. That would be rather more convincing than the distribution of a month's observations presented in Fig. 5.

Re: Yes, it could potentially be convincing to compare our screening method against MODIS. However, comparing different cloud screening techniques between satellite instruments, even on the same platform, is quite challenging and in our opinion it would extend beyond the scope of this study. MISR and MODIS have different spectral bands with different calibrations, different spatial resolutions, and the data are projected differently. The Global Energy and Water cycle Experiment (GEWEX) has an extensive report that describes such instrumental differences and compares their cloud products available online (http://climserv.ipsl.polytechnique.fr/gewexca/). The point being made in the manuscript is that the ARCI-based retrieval screening provides first line of defense against cloud-contaminated retrievals. Additional screening steps using other types of information are applied to filter out more retrievals potentially contaminated by clouds. These cloud-screening steps will be described in a separate publication.

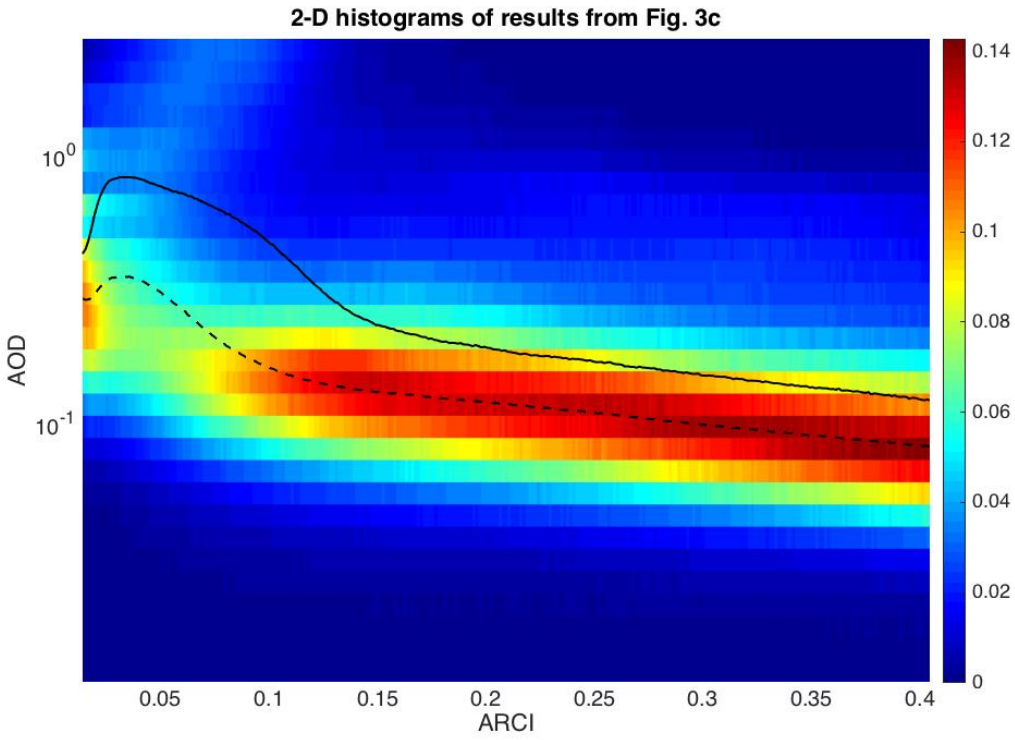
Fig.3 (b) is rather concerning. Do the peaks in retrieval count correspond to the divisions of your LUT? Also, could 3(a) and (c) be shown as 2-D histograms with the mean overplotted? Your argument would be stronger if the decrease in mean AOD with increasing ARCI is due to a decreased prevalence of large AOD (the cloud-contaminated retrievals) while the variation with χ^2 is more uniform.

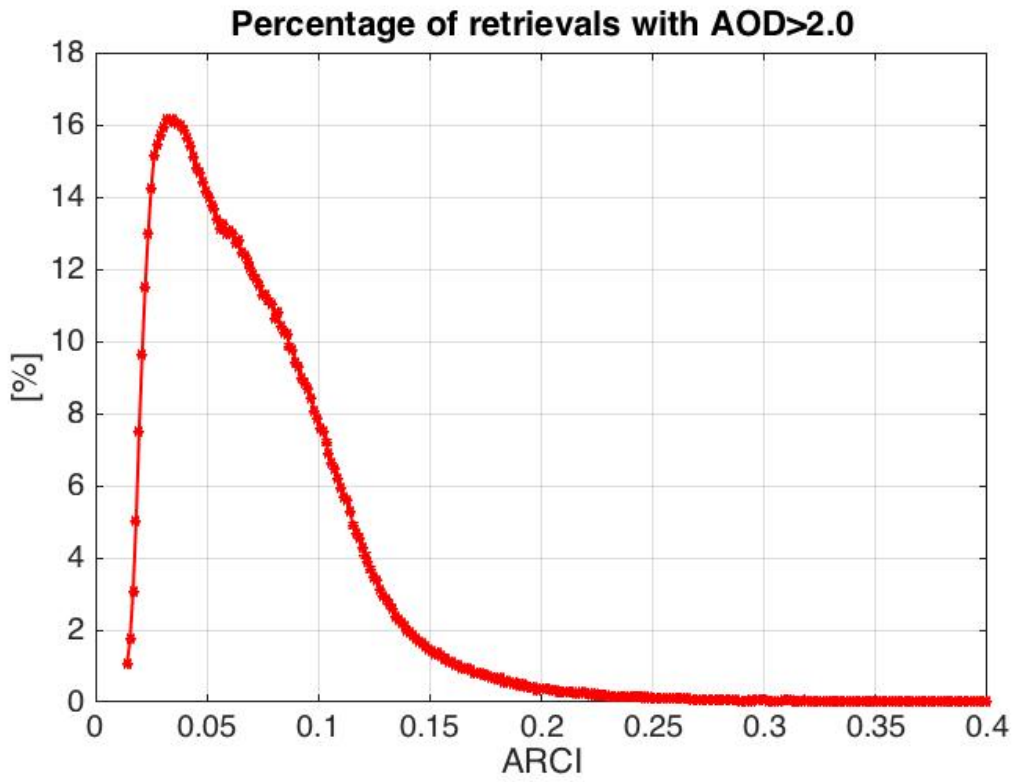
Re: We do see certain clustering around specific min(χ^2) values in our dataset, which gives rise to the small wiggles seen in Fig. 3b. This is probably related to the finite AOD gridding of our LUT, which is 0.025 throughout most of the AOD range. We plan to investigate this feature in greater detail in the future. Furthermore, the wiggles in Fig. 3b become apparent only because of very fine sampling of the min(χ^2) space. Our interval is 0.025, which results in 200 data points for min(χ^2) between 0 and 5.

We created a 2-D figure with results from Fig. 3c by plotting normalized histograms of AOD at each ARCI level. An example is presented below in Figure 1. The black solid and dashed lines are the mean and the median AODs. The figure does show decreasing number of high AODs with increasing ARCI, but the results are not as clearly visible as in Fig. 3 in the manuscript. Another useful metric showing that the number of cloud-contaminated high-AOD retrievals is decreasing with increasing ARCI is the percentage of retrievals with AODs higher than 2.0. This is presented below in Figure 2. The percentage of high-AOD retrievals decreases from the top 16% at ARCI=0.03 to about 1% at ARCI=0.15.

Figure 2 below is simple and clearly conveys the message, but we decided a description in the text was sufficient to strengthen our argument.

"In the first regime, the average AOD is highly sensitive to the specific value of ARCI, characterized by a sharp decrease in AOD with increasing ARCI between about 0.03 and 0.13. This suggests that a decreasing number of cloud-contaminated, high-AOD retrievals are included in the average as the ARCI is increased. Indeed, the percentage of retrievals with AOD higher than 2.0 reaches its peak, 16%, at ARCI equal to 0.03, and decreases to about 2% when ARCI is 0.13. In the second regime…"





P9L2 This paragraph ascribes the variations in Fig. 3 at low χ^2 or ARCI to poor sampling. That implies that there should be retrievals there but you didn't see them. Very low χ^2 implies a very close fit to observations, which is unlikely, and very low ARCI implies a very unlikely fit, which should happen infrequently if the ensemble of aerosol types was well-chosen. Hence, I'd ascribe the sharp variations in Fig. 3 in those regions to scenes that are poorly suited to this retrieval.

Re: We agree with this statement. We ascribed the variations at low χ^2 or ARCI to poor sampling without providing an explanation of why the sampling in these regimes is low. We did not want to put too much emphasis in our analysis to these low χ^2 or ARCI regimes, as they are not very relevant to our main arguments. We did, however, change the phrasing in this paragraph:

"After a rapid initial drop related to a similar rapid increase in sampling…"

"After excluding the initial fluctuation for extremely small ARCI related to poor sampling…"

P9L19 I wouldn't say that the trend in AOD is statistically robust. I'd say that the shape of 3(c) isn't evident in 3(d), so we don't ascribe the kink in the former to a change in frequency.

Re: Agreed. We changed this sentence to read: "The retrieval count decreases slowly with increasing ARCI (Fig. 3d), indicating that the observed trends in the average AOD cannot be ascribed to a change in frequency."

Fig.4 This is a superb figure and deserves more attention than Fig. 3. However, the caption is

unclear if it is plotting the same data as in Fig. 3.
Re: We clarified the data used in the caption and in the text:
"Another way to look at the difference between the two screening approaches is presented in Fig. 4a, which shows the two-dimensional distribution of average AOD as a function of min(χ^2_{abs}) and ARCI using combined data from January and July of 2007."
"Figure 4 (a) average AOD as a function of ARCI and min(χ^2_{abs}) for the combined months of January and July of 2007…"

P10L12 Any idea why cloud contamination is a function of latitude? Does the ARCI threshold need to vary with latitude?
Re: Global cloud fraction has a strong latitudinal component due to the patterns of global atmospheric circulation. We have not investigated possible variations of the ARCI threshold with latitude. We will consider this possibility in the future.

P12L19 You didn't provide a 'strong statistical foundation'. You justified the ARCI threshold by the shape of the distribution of AOD. Statistics would calculate a theoretically sensible value of ARCI based on typical noise and a very large ensemble of aerosol types.
Re: Agreed. We changed this sentence to read:
"Although this screening method does not eliminate all AOD outliers, it is superior to the previously used thresholds in V22 of the MISR aerosol product."

- Finally, I would prefer it if the paper and any data files released clearly describe the retrieved product as 'ensemble mean AOD'. Evaluating a range of aerosol types is an excellent way to sample the unconstrained parts of state space (such as refractive index). Providing an ensemble of results to the user illustrates what the data constrains and what it doesn't. However, a combination of ensemble members doesn't necessarily have a physical meaning. To use an example from a related problem, a thick but high cloud can produce the same TOA thermal radiance as a thin and low one. Giving the user both results shows that both are possible. An ensemble mean, though, gives a medium-thickness layer midway through the atmosphere, which is inconsistent with the data.
  Re: The new V23 data product labels the retrieved AOD as, simply, "Aerosol_Optical_Depth". To a sophisticated user, the idea that this is essentially a "ensemble mean AOD" is a useful concept, which is one of the reasons for writing this manuscript. However, as this AOD is the one the MISR project would like the majority of users to work with, we elected to eliminate the jargon and provide a simpler designation for this field.

A few more minor points:

P4L6 Perhaps 'The previous MISR dark water algorithm' would be a more informative title to someone skimming the paper?
Re: We changed the title of the second section to: "Previous MISR V22 dark water algorithm"

P5L2 reflectance is defines defined as
Re: Corrected.

P5L26 Considering you don't define them, and their precise definition is unimportant to this

paper, perhaps remove specific references to the now neglected χ^2, parameters?
Re: We want to make sure that the readers are aware of additional metrics and thresholds used in V22 processing. This is important as the new approach simplifies the process considerably and makes it more transparent.

P6L34 'turns out to be' is rather colloquial. Perhaps 'and will be shown to produce superior results to the original algorithm'?

Re: Agreed. We modified this sentence to read: "Furthermore, it results in a single parameter that enables screening of retrieval blunders and AOD outliers and which outperforms results derived using the original V22 thresholds."

P7L4 If these are continuous functions of τ, you are presumably interpolating as the LUT is discrete. What are you interpolating — ρ; χ^2; or *f*?
Re: We interpolate χ^2