We want to thank Reviewer 1 for the effort and the valuable critic, which helped us strengthen the discussion of our results.

We improved the method. The revised version is described in short here:

- Preprocessing:
  - we average the data from 1°x1° to 5°x5° first to get a more robust dataset with a near-global coverage
  - Following the Algorithm Theoretical Baseline Documents (C3S2_312a_Lot2_ATBD_AER_main_latest.pdf) of the algorithms we correct the uncertainties delivered with the datasets
- Median:
  - We added the calculation of the uncertainty median.
  - The datasets are defined to be consistent regarding the median metric if the uncertainty intervals (median ± uncertainty) of the individual median values overlap.
- Trend:
  - We use Monte Carlo realization and each of 1000 realizations are perturbated depended on the uncertainties (details see new version of the manuscript attached)
  - The datasets are defined to be consistent regarding the trend metric if the intervals (trend median ± trend standard deviation – calculated from the 1000 realizations) of the individual trend values overlap.
- Seasonal amplitude:
  - Changed to annual range
  - Calculated for each year:
    - annual range: difference between the maximum and the minimum value
    - uncertainty as the sum of the uncertainty of the maximum and the minimum value
  - Median calculation of the annual range and the uncertainties individually
  - Consistent within the uncertainty regarding the annual range metric if these median with the uncertainty intervals overlap with each other.
- Correlation
  - We use Monte Carlo realization and each of 1000 realizations are perturbated depended on the uncertainties (details see new version of the manuscript attached).
  - Calculate Spearman's rank correlation coefficients between pairs of datasets for each of the realizations
  - Calculate median and standard deviation of all 1000 realizations for each of the pairs
  - To get the lowest consistency, so the worst case, we take for each pair the medium value subtracted by the standard deviation. If the minimum of all

the correlation coefficients calculated for the pairs of datasets is higher than 0.7 the datasets are defined as consistent regarding their temporal correlation.

In addition, the code for performing the described consistency assessment will be shared as a Jupyter Notebook with the new version of the manuscript, to give the reader the option to use the method easily by themselves. The answers to the specific comments are in blue directly after the comment in the following and the new version of the manuscript is attached:

Review:

Overall Notes

This paper aims to assess dataset consistency by comparing four metrics applied to the data records from three instruments in the Along-Track Scanning Radiometer family, using three different algorithms. It is important to note that the results do not necessarily represent uncertainty; validation against a "ground truth" dataset, to the degree possible with known, quantified uncertainty, is required to assess uncertainty against standard values. This could be made clearer in the introduction and conclusions – it does not reduce the value of the work, but I think it places the results in proper context.

We added a part on that to the introduction and the conclusion, see in the attached revised version of the manuscript.

Another aspect that might be worth mentioning is the degree to which different instruments or algorithms might be sensitive to different aspects of the environment. In such cases, "consistency" might not be as meaningful as it might seem. Specifically, even if different products claim to be reporting the same quantity (e.g., AOD at a particular wavelength), some methods might be more sensitive to larger or more light-absorbing particles than others, e.g., depending on the actual wavelengths measured. Also, sensitivity might vary with observing geometry or surface properties differently for different instruments or algorithms. These differences might also be convolved with differences in pixel resolution, such that the "right value" for the variable in question depends to some degree on the measurement or retrieval technique. This type of ambiguity is an issue, for example, with the variable reported as "aerosol layer height," where widely used products give different answers, all of which might be correct (within uncertainly limits), but they sample different aspects of the aerosol vertical distribution. This issue also applies, though less prominently, to AOD.

In the case of the concrete example shown in this paper, this issue is only of low relevance, as we are analysing the consistency of AOD 550 nm results from 3 different algorithms which are in each case applied to an identical sensor (3 of them in separate analysis) so that sensitivities for AOD are identical. In the discussion we already touch upon this topic when comparing the consistency results between the 3 sensors (in particular, the impact on AOD sensitivities of the different oblique viewing directions of ATSR and SLTR).

In general, consistency is meant to analyse the agreement between different results either for the same variable or even for different physically connected variables in the Earth system. This means for example that consistency analysis has i. a. the purpose of checking how far monthly AOD at 550 nm obtained from different instruments with different sensitivities and / or sampling agrees. In case of consistency analysis of different variables, the main point is to check the agreement within physical understanding of results obtained from different measurement systems. In essence, we think that these provide added value of such a consistency analysis. When it comes to variables with the same name but different definitions (such as layer height for example), any consistency analysis should threat them as different variables and check whether their agreement is appropriate within physical understanding (e.g. an extinction weighted versus a geometry weighted layer height). We have added this point to the introduction and conclusion.

That said, this paper presents useful results; it provides important insight into the strengths and limitations of the AOD products studied, as well as providing an illustration of a "consistency" approach for data-record evaluation. And the conditions under which the observed AOD consistency is greater are likely to represent places where uncertainties at least in the interpretation of the observations are diminished. In places where consistency is low, the approaches might be sensitive to different things, or there might simply be less information in the measurements about the variable of interest.

More Specific Notes

- Lines 3-4. "median values" appears twice in this list. - corrected
- Lines 28-30. It might be worth emphasizing that "consistency" in general is distinct from on uncertainty, as ground truth data are required to assess actual uncertainty. - added
- FYI, the word "exemplary" means "exceptional." It is used throughout the paper to mean "for example" or "as an example," which is not what the word means. - corrected
- Equation 1. I know it seems obvious, but for clarity, you might add "where n is the total number of points." - added
- Line 90. Perhaps: "... connecting line between all pairs of points in the dataset..." - corrected

- Line 99. “To determine the multiyear mean annual cycle, the mean of all data for each calendar month …”. Also Line 106. “multiyear”- corrected
- Equation 3. I think this is the mean annual range or amplitude, not the mean annual “cycle.” – it is renamed and changed in the new version of the manuscript
- Line 114. See Overall Notes above -- there is a degree to which the “same variable” is not exactly the same, so it might be worth at least mentioning that differences in the statistics might be considered relative to inherent differences in the measured variables themselves.

  While this paper analyses the consistency of 3 different algorithms applied to an identical sensor, there are remaining differences in the sampling of them due to different cloud masking and quality filtering. It is the explicit purpose of a consistency analysis to check the agreement of monthly AOD results between those 3 algorithms including possible sampling-induced differences. In this sense good consistency means that the sampling differences did not create any significant disagreement. We have added this point to the introduction and conclusion.
- Lines 114-115. Rather than an arbitrary 0.7 for assessing agreement, you could use one of several possible measures of the uncertainty in the derived trends to define when the calculated correlation represents agreement. – We want to keep the individual metrics separated so that it is possible to place greater emphasis on a metric depending on the considered variable. We adopted 0.7 as a threshold from Braunisch et al. 2013 (**https://doi.org/10.1111/j.1600-0587.2013.00138.x**). This we added also in the newer version of the manuscript.
- Lines 123-126. Somewhere, perhaps in the Introduction, it might be worth mentioning that other instruments provide multi-angle AOD retrievals, such as HARP-2 and SPECS on the PACE satellite, and MISR on the Terra satellite. - Added in Section 3.1
- Lines 143-144. The “consistency” of ATSR-2 must be assessed relative to something. So, when discussing consistency here, you might make clear that consistency is assessed relative to I guess the per-pixel medians of the other two satellite AOD products (or is it just the standard deviation of the per-pixel medians of only the ATSR-2 AOD values). It is the consistency of the three datasets (SU, ORAC and ADV) regarding the median metric. We specified the sentence in the revised manuscript.
- Lines 206-207. Actually, quite a bit of work has been published already comparing some of these datasets to AERONET, including, but not limited to, Sogacheva et al. (2020). The results of these studies might warrant some comparison with the observations made in the current paper.

In our view, the main result of the individual comparisons of algorithms is the conclusion stated that no single algorithm is best under all conditions. Furthermore, our consistency analysis is independent from AERONET station measurements and thus is not limited to their (somewhat sparse) locations but is able to check the agreement of the results from 3 different algorithms globally wherever they provide output. We have added this point to the conclusion.

- I think it would improve the paper considerably to look a bit more deeply into the differences in the algorithms that lead to the consistency differences shown in this paper. For example, among the leading factors producing diversity among satellite AOD retrievals is cloud masking algorithm used (e.g., Li et al., doi:10.5194/ANGEO-27-2755-2009).

  We added some possible reasons for the observed inconsistencies in the discussion of the AOD consistency analysis section. But as the aim of this consistency analysis is to check the agreement of different results from the end point, i.e. to study the results as AOD at 550 nm in this example case including but not resolving all aspects which may lead to inconsistencies. This is meant as an analysis for users who need an overall “quality stamp”. We have added this point to the discussion of the AOD consistency analysis section and in the conclusion.

- Line 213. Should be: “...the consistency approach works well in regions...” - corrected