We want to thank Reviewer 2 for the effort and the valuable critic, which helped us a lot to improve the method significantly and strengthen the discussion of our results.

The answers to the more specific comments are in blue directly after the comment in the following and the new version of the manuscript is attached:

Review:

This manuscript provides a concrete example of a consistency method, based on a previous paper discussing consistency on a higher level (Popp et al 2020). The authors propose a simple method, then apply it to 3 data sets to study their consistency. I think that the idea to derive a concrete method is useful for the community, however I think that the proposed method is too simple for a dedicated paper. I would see it as support for an analysis in a paper focusing on the use of different data sets. For a paper focusing on the method itself (with an illustration of its use), it requires to be deeper than saying one should look at medians, cycle amplitude, trends and correlation which are all obvious to me. Such a methodology paper should go deeper in how to define and use those metrics for a correct and in-depth analysis. Here are some more precise comments about what I think is missing:

We adjusted the method to include the missing use of uncertainties. The improvements of the method which we include in the new version of the manuscript are in short:

- 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 add 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, i. e. 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.

1. There is no use of the data sets uncertainties, although their importance is highlighted in the paper cited as foundation of the current manuscript (Popp et al, 2020).

Thanks for this comment. As described above, all metrics have been extended to account for measurement uncertainties. This improved the methodology significantly and made it more consistent with the definition in Popp, et al., 2020.

2. There is no discussion of the difference between a constant bias between algorithms and truly inconsistent data sets, which would both reflect in the median value but have very different meaning in terms of data use.

This is one of the reasons why we apply the different metrics. A discussion on this point is added in the revised version of the manuscript.

3. Linear trends are used, with a robust estimator, but their uncertainties and significance are not considered. I am not sure to understand how to justify the comparisons of trends without their uncertainty range, and without ensuring their

significance. I do understand the limitation of having less data if considering uncertainty and significance, but I think it is scientifically more sound to have less meaningful data than more data with uncertain meaning.

From our point of view one of the important aspects of the consistency assessment is to have a near-global coverage (e.g. not limited to sparse ground-based network locations used for validation), so that it allows to identify areas in which the datasets can be used easily from regions where they should be handled with care.

Using the new method including also the uncertainties of the CDRs we see a big improvement for the trend assessment.

4. The seasonal amplitude is calculated using means and not medians, leaving a possible strong impact of outliers.

We changed this metric completely to now look at the annual range calculated for every year individually, with a calculation of the median of theses including the uncertainties of the datasets.

In addition the mean seasonal cycle is obtained using different years for the different data sets (due to their different time coverage), but there is no discussion about the expectation of these seasonal cycles to remain constant over time.

We only assess the consistency of the different algorithms using the data of the same instrument, so also the same years. Only in a second step, we compare the consistency results for theses algorithms applied on the three different instruments.

I also find the cycle amplitude definition to be very basic and dependent on outliers. Why not fitting a curve (such as a sinus, if this is the expected behavior) on the data and take the amplitude of that fitted curve?

First, we want to address the annual range with this metric, while the temporal behavious is addressed in the correlation metrics. The second point is that the annual cycle has quite a different behavior in different regions of the world, and as the consistency framework should be able to be applied without the knowledge of the individual physical behavior of the analyzed variable, we want to avoid putting that information into the framework.

5. The correlation should not be used bluntly. The possibility to have a good correlation depends not only on the quality of the data, but also on the variability / cycles with respect to the uncertainty / noise in the data. A bad correlation between 2 data sets (in a grid cell) could simply mean that the variability (in that grid cell) is below or close to the precision of the retrievals.

By perturbating the data with their uncertainty for Monte-Carlo calculations, we reduce the influence of the described effect. In addition, if the variability is below or close to the

precision of the retrievals than the data should be treated with greater caution. So any flagging as not consistent would in fact be preferable.

Other precise questions / comments:

line 66: why average on a 5° grid?

This was to get a mean and standard deviation value which were needed for the comparison. This is changed now, we average on a 5°x5° grid before applying all the metrics so that a better coverage can lead to near-global consistency assessment. This is also added in the manuscript.

lines 69-70: what is this std dev of all three algorithms?

That was meant for each algorithm, so that there was one interval per algorithm and if they overlap it was defined as consistent. But this also changed due to the change of the methodology

eq 1: although they may seem obvious, n and m should be defined

Added in the revised manuscript.

lines 99-100: does that mean that each observation has the same weight, but each day does not have the same weight? Then the mean depends on whether there were more observations when there was significant aerosol loading, for example. Please justify this approach for the determination of a mean annual cycle.

As we use monthly level 3 data we cannot differentiate how much influence individual days have on the monthly mean. It is correct that the monthly means are dependent of the overpassed which are also not daily for the dual-view instruments. In combination with cloud masking it could be that only a few measurements contribute to a monthly mean value. As these limitations apply for all algorithms for the same instruments nearly the same, we think that the influence on the consistency assessment is not significant while the consistency analysis is a valid assessment whether this has any negative effects on their agreement.

lines 101-105 are unclear

As we changed the method here the text is changed.

line 126: how does this impact the aerosol retrieval and why is it important to discuss here?

It is the reason why we look at these more robust and reliable aerosol retrievals from the dual-view instruments in the manuscript. We added this in the manuscript.

Minor comments / suggestions / corrections:

In general, "examplary" is used incorrectly; first I do not think it exists with an “a”, and “exemplary” does not mean “for example” at all. - corrected

line 3: median is listed twice - corrected

line 15: is -> are - corrected

lines 32, 38, 211: wrong citation command - corrected

line 68: consistency “is” fulfilled ? (instead of “if”) – reformulated

lines 96-97: phrasing is a bit weird – reformulated