Reviewer 2's comments:

Review:
This paper discussed the OCEC measurements inter-comparison results obtained from the Chemical Speciation Network (CSN). There are two sets of inter-comparison, but in general, they are comparison between OCEC results obtained from a Sunset instrument vs. DRI analyzer. The main difference between the two sets of measurements lies in the technical details, which are the different analyzer model and the date coverage of the measurements. This is supposed to be a useful inter-comparison exercise and will help the network to understand the impact when switching instrument and how the long-term measurements can be interpreted when using it for trend analysis. This work certainly is valuable; however, I find it needs a major revision to re-organize the paper better before it can be published. This is rather confusing, and I do not think there is enough interpretation regarding the actual comparison between the two sets of measurements in a reader perspective. There are many aspects the authors are presenting in this paper, but they are not presented in a way that is easy to follow. I have summarized a list of things that can help improve this manuscript before it is published in AMT.

We thank the reviewer for the comments. In the following, we provide our point-to-point responses to these comments and list our revisions to the text in order to address these comments.

General comment:
In this paper, the authors rely a lot of statistics and try to interpret the difference (or bias) between the two data sets looking at all data as a population. It is one way of looking at things and to get a sense of the magnitude of this bias in all measurements as a whole. But for ambient measurements, sometimes it is also important to understand the time series of this deviation and to really understand the reason behind the bias. For example, does the bias is higher in summer than winter? Or is always consistent over the course of the sampling period? During the analysis, the authors discovered a significant number of samples that have no observable or detectable "OP". More investigation is needed to understand if this may impact the data differently for example over different seasons.

We understand and respect the perspective offered by the reviewer on a comprehensive way to carry out and interpret a comparison study like ours. It would be ideal if available data and resources could allow for a thorough investigation of various factors driving the differences, such as seasonality of chemistry, as mentioned by the reviewer, as well as source characteristics of the samples, etc. The samples analyzed here were collected during September to October of 2007 (Set 1) and May to September of 2017 (Set 2), which do not offer much information on the seasonal characteristics of inter-instrument differences. However, the dataset does cover a great variety of emission sources and meteorological conditions, given the wide spatial coverage of the CSN network, which makes it a comprehensive and well suitable statistical sample for us to achieve our primary goal of the study, i.e., to compare results obtained from three (Sunset vs. two DRI models) analyzers using the *same* analysis protocol, IMPROVE A with reflectance charring correction from the same sample. It should be mentioned that as valuable as such a comparison is for the community to make better use of these instruments and interpretations of the results, it has not yet been done for the CSN network until this study, mainly because of the large amount of work involved (e.g., a reanalysis of the archived samples). Given the comment by the reviewer, we have added a statement in the introduction section (Line 95-99) to emphasize the goal and contribution of this paper, i.e., statistical analysis of inter-instrument differences, that reads *"These samples, which were collected during September to October of 2007 (Set 1) and May to September of 2017 (Set 2), covered a great variety of emission sources and meteorological conditions, given the wide spatial coverage of the CSN network, ensuring statistically robust comparison among the three instrument models."*

The wide spatial coverage and hence variety of sources sampled are also mentioned in the revised experimental section 2.2.1 (Line 155-156): "*Both sets cover a variety of emission sources given the wide spatial coverage of CSN network."*

We also add a statement in the conclusion section (Line 500-502) to recognize the potential value of a seasonal comparison: "*While data used in this study were primarily collected during the summer/fall season, future comparisons with data covering longer sampling period will paint a fuller picture of all seasons."*

Last but not the least, it is our understanding that some of the reviewer's questions/concerns could be addressed by clarifying and emphasizing that all analyzers are using reflectance correction for charring. We have made such clarification throughout the revised manuscript where applicable.

In order to obtain the correct OC and EC, charred OC has to be taken into account. The Sunset instrument laser detector detects the transmittance whereas the DRI analyzer has the option to use either transmittance or reflectance. I could not find where the authors discuss about whether they are using transmittance or reflectance to determine charred OC (or OP).

The Sunset Model 5L analyzers used in our study are equipped with dual optical units, allowing for concurrent detection of both filter transmittance and reflectance (https://sunlab.com/wp-content/uploads/Lab-Instrument-brochure.pdf). All OC and EC data presented in this paper are corrected using reflectance. To make this clearer we have added the abovementioned detail about Sunset's dual optical units (Line 90) and clarified in multiple places where confusion might occur that the optical correction is by reflectance.

Also, in multiple places when OC and EC were discussed, are they both corrected for charred OC? Any comparison for OC and EC between the two data sets without accounting for charred OC is meaningless.

We acknowledge that the terms OC and EC are commonly referred to those carbon fractions with charring correction applied. In the original manuscript we have been careful with the different terms by using OC_{sum} and EC_{sum} when referring to the uncorrected OC and EC. It is our understanding that OC and EC subfractions are used widely by source apportionment studies as PM source markers [e.g., Liu et al., 2005], making us believe it is useful to emphasize the physical meanings of these subfractions to inform future use of these data. To avoid any confusion, in the revised paper we introduce these terms earlier in Sect 2.2.2 and use "$OC_{1+2+3+4}$"and "EC_{1+2+3}" instead of OC_{sum} and EC_{sum} to exclusively refer to the uncorrected OC and EC.

The authors introduce a new term, referred to as "SRD" or scaled relative difference. I understand this may have some statistical value, but this is not a typical common term that all readers will connect to when they read. This is particularly an issue when readers looking at various graphs generated from this value and try to understand or interpret its meaning. One way to improve this (if the authors insist of keeping this analysis) is to do a better job in explaining this term other than by defining it again. In other words, how readers should interpret this term for its magnitude. For example, when one express absolute difference between two numbers, readers understand the value of 0 means there is no bias. Or large positive or negative values means certain set is higher or lower. In comparison, SRD (because of the way it is defined), it does not easily translate.

Also, is this SRD term mostly used for interpreting the comparison between measurements with replicates? It is not clear from the heading of the sub sections to me.

We clarify that SRD is not a new term and actually has been used in many of our publications (Hyslop and White, 2009, Spada and Hyslop, 2018; Gorham et al., 2020). The merit of SRD is that it is the normalized difference between two measurements in %, and it considers measurement uncertainty from both measurements in a paired analysis. By using it for both within-analyzer and inter-analyzer comparisons, it is possible for one to easily compare the magnitude of the measurement uncertainty, determined from the replicate measurement, to the inter-analyzer difference. Considering the reviewer's comments, we do feel that it is worthwhile to further emphasize the premise and purpose of using SRD for the readers to better understand its meanings. We have expanded equation 5 to show that SRD = RD/$\sqrt{2}$ and revised the related text (Line 231-234) to emphasize its merit that reads "*The SRD, which equals relative difference (RD) divided by*$\sqrt{2}$*, is chosen over RD because it is the normalized relative difference between two measurements, accounting for the presence of equal and independent errors in both original and replicate measurements (Hyslop and White, 2009).*" We've also added a statement in the revised text (Line 269-270) that reads "*In both cases, a positive AD or SRD value occurs if the Sunset measurement is higher than the DRI measurement*."

Specific comment:
p.7, line 275-285. It is observed that Sunset TC is generally agree with DRI TC, however, Sunset OC was found lower than DRI OC (and the opposite is true for EC). This is obviously due to the difference in OCEC split but is could also due to the fact that one instrument used transmittance to determine charred OC whereas the other one use reflectance. I couldn't find where this is discussed or whether the authors are using data to ensure the consistency (i.e., all corrected data are based on transmittance or reflectance).

As discussed in the response to the earlier comments, both analyzers are using reflectance charring correction.

p.9, line 339. Determining the correct OC and EC really requires the correction of OP. So what is the purpose of showing the "uncorrected" OC and EC when you know OP has to be taken into consideration?

The purpose is to point out the large difference in uncorrected OC and EC by the two analyzers using the same TOA protocol, IMPROVE A with reflectance correction, as this discrepancy likely arises from differences in instrument configuration and setting. We look at the difference between analyzers for both uncorrected and corrected OC and EC to examine and understand this difference.

Also, by definition, OP is subtracted from EC and added to OC to determine the final EC and OC. Therefore, in comparing the "uncorrected" and "corrected" OC, one should expect OC and EC be shifted in the same magnitude but in opposite direction. How come this does not seem to be the case in Fig. 4? Is this because it is plotted as a "relative" difference rather than absolute difference? Would absolute difference be more meaningful than the relative difference in this case?

The reviewer is correct about corrected OC and EC being shifted in the same magnitude, when the absolute difference is plotted as the y-axis. In Figure 4, the (scaled) relative difference is plotted instead of the absolute difference, thus the magnitude of change before and after charring correction is different for OC and EC. We choose to plot the relative difference over the absolute difference to show the different impact that the optical correction has on OC and EC because EC is a much smaller fraction of TC compared to OC.

The data in Fig. 4 are sorted according to the "mass loading percentage". How is this defined? Do the authors mean the amount of OC or EC relative to TC? So what is the advantage of plotting the data this way rather than sorting them by the absolute EC and OC mass?

The x-axis of Figure 4 is the sorted absolute OC and EC mass loadings, not their mass fractions in TC. The distribution of SRDs is plotted for each of the 20 mass loading bins (5th percentiles). The caption of Figure 4 and related text are revised to clarify this point.

p.9, line 348-350. The authors suggested for EC measurements with large SRD were samples with no instrumentally detected OP. What is considered "large"? Better to give a range or value or threshold. Does this represent bars with high "relative difference" in Fig 4?

As described in Sect 2.2.3, inter-model uncertainty derived from the replicate analysis is used as a benchmark for inter-model comparison. Both estimates are given in SRD, enabling a direct comparison. We have added the following statement to the revised text (Line 361-364): "*In investigating this anomaly, we found that EC SRDs were larger for samples with no instrumentally detected OP (i.e., OP = 0) by both Sunset and DRI-2015 analyzers (Figure 5a), with a median value of 20.9%, which far exceeds the inter-model uncertainty for EC determined from the replicate analysis (Table 3).*"

In addition, the authors said there are consider number of samples with no detectable OP, if I understand this correctly, no "OP" means the "corrected" OC (or EC) and "uncorrected" OC (or EC) are equal, correct? However, I don't ever see any evidence that there are any samples with "non-detectable OP" in Fig. 4. Fig. 4 shows that there are always a considerable amount of OP in all samples!

The reviewer's understanding is correct that corrected and uncorrected OC and EC are the same when OP is non-detectable. The reason the reviewer doesn't easily identify the OP=0 samples in Figure 4 is that the data are binned by mass loadings and what's shown in that figure is a distribution of samples in each bin. As Figure 5b shows, samples with OP = 0 are always a subset of samples in comparison in all EC mass loadings percentile bins, even at the highest EC mass loading percentile where OP = 0 samples comprise ~30% of the total samples in that bin.

p.9, line 355-359. The authors suggested that the EC bias almost solely originated from the thermal effect and I don't understand the argument supporting this statement. In Fig. 5, I only see the sub group of measurements with "no OP" (i.e., no thermal correction) has even higher relative difference than the group with considerable amount of OP. Or may be I don't understand how to interpret this graph. Plotting in relative difference could be one issue. If the authors' statement is true, I would expect for the sub-group of measurements with no OP should give you no bias between the two data set. Is this the case?

The sub-group of measurements with no OP has the largest bias in EC because of the large difference in the quantified OC subfractions (discussed in Section 3.1.2) by the two instrument models. To better guide the reader to interpret this figure, we have revised the related text and the new text (Line 371-377) reads: "*As shown in Figure 5c, for the 179 samples with no charring correction from both models, considerable correlation was found between the inter-model differences of EC and* $OC_{1+2+3+4}$*. This suggests that, in the absence of charring correction, much of the observed bias in EC between the two models is essentially coming from the inconsistency in the quantified OC subfractions by the two models. In contrast, samples with charring correction (i.e., OP > 0) showed little correlation between the inter-model biases of EC and* $OC_{1+2+3+4}$*.*"

p.9, line 360-386. This whole paragraph may be should not belong in main text. Probably is better to be included in supplementary information and then just refer to it in the main text. This paragraph is mostly to explain Fig. 6a, however, I would suggest the authors also include the detector signal, so the authors

would better understand how the OP peak compared to OC or EC and aid the interpretation of the thermograms. Only the laser response in Fig. 6 does not help much.

Thanks for the suggestion. We agree that FID signal would be a good addition to Figure 6 as the reviewer suggested. Thus, we have modified Figure 6 to include FID signal along with the laser response in the example thermograms from Sunset. However, after consideration, we believe that Figure 6 and the paragraphs discussing it are an important element of the main text. We have also modified Figure 6 to retain only the empirical cumulative plots (new Figure 7a and b) while removing the histograms, as the latter is in essence another angle looking at the statistical distributions but does not add much more information. The new Figure 6, Figure 7, and text of discussions are now more concise and more suitable to be shown in the main text.

Fig. 6b to e. What is the purpose of these? How does the different shapes in Fig. b and c are supposed to mean and what we should expect? I really don't think these graphs really aid much in terms of understanding what is going on regarding the "OP = 0" issue. The authors can however keep them in supplementary info if they want. What would really help is to actually include a completed thermograms including both laser response and detector response. Only that will allow the readers to understand how the samples evolve over the course of the analysis in a typical situation when OP=0 and OP>0 and may also include the case for blank as reference.

The purpose of Figure 6 (new Figure 7) is to offer a revelation of some unique characteristics of samples with OP=0, by closely examining the instrument signals and statistical populations of three types of samples, i.e., OP=0, OP>0 and blanks. Those OP=0 samples uniquely stand out with low initial and final laser readings, which are evident in both the example thermogram (Figure 6a) and the statistical distribution of laser readings (Figure 6b and 6c). After considerations of the last two comments on Figure 6, we have made two revisions to this figure, i.e., to add FID signals and to remove the histograms, which make it better emphasize these key characteristics of those OP=0 samples.