Response to Reviewer 2 for AMT-2020-396

Thank you very much for your comments on our paper. You have raised several important 'big picture' questions regarding NDACC intercomparison campaigns. I have tried to respond to several specific points. Many of your comments provide us with good motivations for the next paper.

This is a generally well written paper about an NDACC intercomparison campaign, augmented with results from other NDACC campaigns. One of the main points in the paper is that the original Hohenpreissenberg lidar HOHO is compared to the NASA travelling reference lidar STROZ for the second time (2009 HOPE and 2018-2019 HOPS) and the new Hohenpreissenberg lidar HOH is now also compared to STROZ, as well as to the HOHO system. So the consistency of HOH with HOHO can be established on site for the overlapping parts of the profiles. It is shown in Fig.2 that HOH has much greater performance in terms of range and SNR, but in the overlapping regions the systems are consistent. In order to validate HOH, STROZ is needed.

General Comments:

Now, the general structure of the paper becomes sometimes a bit hard to follow, since the HOPS campaign are intertwined with LAVANDE results that have been published separately, which confronts the reader with a few storylines that have to be kept separate. The points brought in are certainly relevant, but it complicates the structure of the paper. In fact, this becomes clear in the conclusions: "The cross-comparison of NDACC campaign at Hohenpeißenberg Meteorological Observatory (HOPS) and at Observatoire de Haute Provence (LAVANDE) has allowed for the unique opportunity to assess potential biases in the NASA-STROZ reference lidar."

My suggestion is to re-structure some of the sections to clarify this and move this material as much as possible to Sec.7. Since the NDACC intercomparisons with a travelling reference lidar have been undertaken for some time (e.g. the references mentioned date back to 1995) it could be clearer described how the intercomparisons are generally carried out, according to an NDACC protocol, and perhaps explain how the HOPE and LAVANDE campaigns may be deviating from that protocol.

We sympathize with the complexity for the reader - perhaps we have attempted too much for a single article. However, we are hesitant to make major structural changes to the paper as doing so would likely add text to an already long article.

There are some instances in the text that suggest there are different variants of the protocol. It would be interesting from the network design point of view to know why these variants exist.

The last review of NDACC lidar intercomparison practices was done 17 years ago (Keckhut et al., 2004). There is a great deal of work that *should* be done to develop and modernize NDACC "standard practices".

Specific Comments:

Abstract
- Remove the sentences "The previous 2017-2018 ... are reported in the companion article."
Done.

- Add the main conclusions "The intercomparison exercise has confirmed that the original DWD lidar, HOHO continues to meet NDACC standards for ozone profiles at the 3% level between 16.5 and 43 km and at the 10% level between 10 and 44 km. The HOHO lidar meets the NDACC temperature standards for accuracy at the ±1 K level between 18 and 70 km. The new DWD lidar, HOH, meets the 3% ozone standard between 17 and 41 km, the 10% ozone standard between 15 and 41 km, and the ±1 K temperature standard 555 between 17 and 78 km."
Added text to L13:
“Both the original and new DWD lidars continue to meet the NDACC standard for lidar ozone profiles by exceeding 3% accuracy between 16.5 and 43 km.”

A similar statement for temperature is in the following paragraph.

- Add "The cross-comparison of NDACC campaign at Hohenpeißenberg Meteorological ObC2 servatory (HOPS) and at Observatoire de Haute Provence (LAVANDE) has allowed for the unique opportunity to assess potential biases in the NASA-STROZ reference lidar. Possible biases may arise from algorithm initialisation choices and serve as strong motivation for another NDACC temperature algorithm paper."
Last two paragraphs of the abstract are rewritten as:
“There was good agreement between all ozone lidar measurements in the range of 15 to 41 km with relative differences between co-located ozone profiles of less than $\pm$10\%. Differences in the measured ozone numbers densities between the lidars and the locally launched ozone sondes were also generally less than 5\% below 30 km. The satellite ozone profiles demonstrated some differences with respect to the ground based lidars which are due to sampling differences and geophysical variation. Both the original and new DWD lidars continue to meet the NDACC standard for lidar ozone profiles by exceeding 3\% accuracy between 16.5 and 43 km. Temperature differences for all instruments were less than $\pm 5 K$ below 60 km, with larger differences present in the lidar-satellite comparisons above this region. Temperature differences between the DWD lidars met the NDACC accuracy requirements of $\pm$1 K between 17 and 78 km.

A unique cross-comparison between the HOPS campaign and a similar, recent campaign at Observatoire de Haute Provence (LAVANDE) allowed for an investigation into potential biases in the NASA-STROZ reference lidar. The reference lidar may slightly underestimate ozone number densities above 43 km with respect to the French and German NDACC lidars. Below 20 km the reference lidar temperatures profiles are 5 to 10 K cooler than the temperatures which are reported by the other instruments. The differences in both temperature and ozone are likely due to choices in data treatment.“

Section 2

- The differences of the original Hohenpreissenberg lidar HOHO and the new lidar HOH are described. In the description of the travelling standard STROZ, it is not clear if instrument changes have been applied since the HOPE campaign in 2009. This is relevant since the consistency of the performance of HOHO is essentially compared again now in the HOPS campaign against the same travelling standard.
Components of the NASA lidar have been replaced over time but the experimental design of the travelling standard lidar is the same. It is impossible to guarantee that the standard remains absolutely unchanged over 25 years as there is no external reference. It is encouraging that in both the 2008 HOPE campaign and the 2019 HOPS campaign NASA measured slightly less ozone above approx. 45 km than the DWD lidar.

Section 6
- Earlier in the paper reference is made to Leblanc et al., 2016a, b, c. Are the results presented obtained using the methods described in those papers? Are results, following the blind intercomparison, processed by the proprietary processing algorithms of each group, or are they processed by a common processing code that is endorsed by NDACC? How would using a common code impact the intercomparison results for HOPE, HOPS and LAVANDE?
The three Leblanc papers provide a common set of definitions and metrics for NDACC ozone and temperature profiles. However, each group uses their own codes and analysis procedures on their lidar measurements. The blind intercomparison only identifies the absolute differences between the final data products.

A common NDACC processing code for use during intercomparison campaigns is a very good idea. It would remove a source of uncertainty from the interpretation of our results. Additionally, comparing the measurements from each lidar as processed by both codes would allow for the separation of instrumental and computational sources of bias.

We hope that this article offers some motivation for the NDACC community to agree to the development of such a “standard intercomparison code”.

Small comments:
- Not all readers may be familiar with Pearson’s correlation coefficient. Please briefly explain.
Added to L370:
“(a measure of linear correlation between two datasets)”

- The names of the colours in the figures are a bit strange; e.g. "burnt orange", and "mustard". Why not just orange and yellow?
I used the names for the standard colours in the Matlab RBG colour palette. “Burnt orange” and “mustard” have more brown in them (look darker) than orange or yellow.