The manuscript "Balloon-borne Stratospheric Vertical Profiling of Carbonyl Sulfide and Evaluation of Ozone Scrubbers" by Alessandro Zanchetta et al. compare several new methods to measure carbonyl sulfide (COS) concentrations along vertical profiles into the stratosphere using balloon-based air sampling platforms. The authors compare measured COS profiles between the different sampling instruments, from three locations and assess their results in comparison with modelled COS profiles. The methods show good overall agreement and will undoubtedly be able to provide new and insightful information to the growing science community interested in COS.

The manuscript furthermore includes an assessment of the efficiency of ozone scrubbers, based on experiments undertaken on a purpose-built setup. The authors tested cotton, a material that has been suggested some 30 years ago and found that it is not as reliable in ozone elimination as it might have been expected. They tested an alternative material in laboratory experiments, which has not yet been tested on the balloon-based air samplers, but shows very promising results.

The manuscript is very well written, the experiments are clearly described, and the overall methods are sound. In my opinion, this is a fantastic study, and the manuscript is an excellent fit for publication in AMT.

*Authors' Reply (**AR):** thank you for the very positive feedback about our study. The answers to the provided comments will be presented in italic black text and the modifications to the manuscript will be presented as **italic bold** text, citing page numbers and lines when applicable.*

There are, however, a few minor points that I suggest considering prior to publication to improve the clarity of the manuscript:

1) Instrument calibration. The observations span a period of four years, from June 2019 to August 2023. COS measurements are challenged by instability of COS in reference gas cylinders. While the authors provide information on the measurement precision, information on the long-term stability of the instrument is not presented. As this study potentially requires long-term stability of the instrument over four years, a demonstration or small assessment on the stability to underpin the robustness of the presented method would be desirable.

***AR:** we appreciate this remark and have clarified our efforts to achieve long-term stability of our observations. The long-term calibration was performed identically to previous studies about COS measurements on the same instrument (Kooijmans et al., 2016; Zanchetta et al., 2023). In addition, by using dilution of a cylinder with a known mixture of ~1 ppm COS and ~0.4% CH_4, we assessed our 2015-era standard in 2023 and 2024 and found no appreciable drift (<0.3%, or <2 ppt COS) over that period.*

In 2018, our lab acquired from NOAA (Boulder, USA) a calibration standard at ~1000 times elevated and accurately known concentrations of CH_4, COS and CO. High-concentration cylinders are known to show much reduced drift over time than 'natural concentration' cylinders (NOAA, personal communication). This cylinder may be used to set our QCLS scale by using the accurately known CH_4/COS-ratio of that high-concentration standard. Pulsed admixture of a small amount of the highly concentrated gas into a trace-gas free carrier gas (either N_2 or an O_2/N_2/Ar mixture) allows determination of the slope (and linearity) of the COS to CH_4 ratio of the QCLS. After calibrating the CH_4 measurements to standards, we can in principle calibrate its COS measurements, infer drift in standards, or even re-assign COS values to them.

This assessment has been attempted in 2023 and 2024 for our 2015-era standard, which was found not to have drifted appreciably (<0.3%, or <2 ppt COS) over that period. However, currently unclear deficiencies and unknowns in our procedure (among other: repeatability, effects of zero gas, unresolved uncertainty of the COS scale of the standard; Bradley Hall, NOAA, personal communication) have so far made us hesitate to re-assign COS values to that cylinder. Tentatively, such an updated COS value may be ~ 5% (~25 ppt) lower than 10 years ago, with proportional implications to the AirCore results reported in this publication. However, in light of the uncertainty of this assessment, for this publication, we will assume that no appreciable drift has occurred in the COS of our standards.

The following paragraph has been added in Sect. 2.2, lines 213-219:

As reported in Zanchetta et al. (2023), field standard cylinders are calibrated against NOAA standards (NOAA-2004 COS scale) in the laboratory before and after each measurement period to test for drift in the molar fraction of gas species. The COS mole fraction measurements of nine cylinders are available, and five cylinders changed by less than 2.5 ppt yr^{-1}, two cylinders decreased by ~ 10 ppt yr^{-1} and two cylinders decreased by ~ 30 ppt yr^{-1}. The four cylinders that drifted more than 10 ppt yr^{-1} were not used as reference cylinders in the data processing. All of the cylinders were uncoated aluminium cylinders, which, according to experience at NOAA, are more prone to COS mole fraction drift than Aculife-treated aluminium cylinders. More details on the instrumental calibration, precision and stability can be found in Kooijmans et al. (2016)."

A dedicated, and more refined, attempt is currently being undertaken to use the high-concentration standard for calibration/monitoring of our standards. Results thereof may be reported in an upcoming publication, which may additionally allow for an assessment of the accuracy of the AirCore work presented here.

2) Figures. Many of the figures include a lot of data. Several figures are so crowded that I find it difficult to comprehend (for example 3 a). Figures often include very small symbols (for example 3 b and c), which are too small for me to tell apart by colour, blurring the figure and the story the data tell. Other examples include Figure 1 left panel, Figure 2, Figure 3 left panel, symbol size

in Figure 3 middle and right, Figure 4, Figure 5. Can Figures 6 and 7 have larger symbols? Maybe split some of them and have Figures on single panels with full width (i.e., Figure 3 a)? Could log scales be useful to decompress the large numbers of symbols in the lower y-axis range? Or could panels b and c of Figure 3 be on top of each other as a second column in the Figure, allowing 3 a to be larger? I am not sure how to improve, but I think the present state doesn't make the most out of the fantastic data and it would be well worth to improve their clarity.

Some multi-panel Figures have letters as identifiers (i.e., Figure 3 a-c), others (i.e., Figure 1) don't, while some are referred to as second/third (p16, l377). Furthermore, the text refers to the Figures in a sequence that is different from the appearance and enumeration of the Figures, i.e., the text first refers to Figure 3, then Figure 5, then Figure 1. This might be a matter of personal taste, but greater consistency on those points would make the manuscript more accessible to me as the reader. It can be difficult to follow at times, especially when the text is jumping between different figures across different pages with focus on different events in those figures, i.e., P17, l386, where it might be useful to have markers in the figures (arrows?) to mark these events.

AR: *thank you for this feedback, and we have made the following changes in the figures.*

The panels of Fig. 3 have been separated and Fig. 3a is now a standalone Fig. 3, while Fig, 3b and Fig. 3c are now shown as Fig. 4a and Fig. 4b, respectively*.*

Following the referee's suggestions, ***all multi-panels figures have been labelled with letters*** *(e.g., panel a, b, ...) and* ***the markers have been enlarged for all figures, with the exceptions of Fig. 3 (previously, Fig. 3a)****. We believe that for Fig. 3, given the presence of error bars, bigger markers would have made the figure more busy and confusing.*

The numbering of the figures was corrected and the in-text references to figures were adapted to the labels given to the figures' panels.

Fig. 7 was moved to Sect. 4.2 and consequently became Fig. 5. Accordingly, Table 4 was divided in Table 4 and Table 5 reporting the SPIRALE regression results and the ACE-FTS regression results, respectively*.*

We are grateful for the very useful and constructive comments, and believe that these changes have improved the paper's structure and generally improved its readability.

3) Structure. The main part of the manuscript includes an Appendix, which talks about the setup to assess different O3 scrubber materials, as well as the results of experiments made using gas from one cylinder with 0 ppt COS. There are also supplementary information to the manuscript, which include more test results on O3 scrubbers using a different cylinder with a different test gas,

containing some 750 ppt COS. It is not clear to me why the O3 scrubber assessment is separated between appendix and supplementary information? It would seem to me that they could be merged into one assessment in either appendix OR supplementary information, which can then be referred to in the main text. As the assessment of the scrubbers in the appendix is made on a gas containing 0 ppt COS, I believe this doesn't allow to assess whether the scrubbers could potentially reduce COS in the sample, as it is already at 0 ppt. To me, this is an important point that could potentially have affected the outcome of the study, if tests hadn't been made with the 750 ppt COS gas as well. Here, the combined information from tests made with the 0 ppt COS and 750 ppt COS provide the robustness of the results that is needed. Therefore, the separation of those sections into Appendix and Supplements seems confusing to me and I'd suggest keep together in the manuscript. This would enable presenting the analysis in a more robust and compact way, which I think would improve the clarity of the manuscript.

AR: *Thank you for your comments on the paper structure and pinpointing on the separation of O_3 scrubbers between Appendix and the SI. We have made the following changes:* ***the text and figures of Sect. S2 of the Supplement has been moved to the Results and Discussion sections of the Appendix****. The tables reporting the p-values resulting from the ANOVA tests were left in the Supplement, and Sect. S2 became Sect. S3 to resemble the order in which the information is presented in the Appendix. This should also allow the reader to obtain a more thorough overview of the performed experiments and provide more detailed insights of the possible effects related to the interaction of COS, O_3 and the tested scrubbing materials.*

Specific comments:

Title: "Ozone Scrubber Materials" instead of "Ozone Scrubbers"?

AR: *changed to "Ozone Scrubbing Materials".*

P2, l42: ", with a mole fraction range of 350-520 parts per trillion (ppt) in the unpolluted free troposphere (Berry 2013, Remaud 2023)."

AR: *added.*

P2, l47: "to CO2 and sulfur dioxide (SO2), a precursor…"

AR: *added.*

P5, l142: What is the COS mole fraction in that gas?

AR: *this air mixture was made in the laboratory and contained 0 N_2O and 0 COS. It has now been specified in the text: "[…]* with air from a cylinder of synthetic air mixed with low

mole fractions of CH_4, CO_2 and CO, which was meant to simulate stratospheric air conditions ***and contained 0 ppb N_2O and 0 ppt COS."***

P6, l 169: Are these valves used for heating?

AR: *yes, as a coincidental feature the activation of these valves causes them to heat up, as the activation is triggered by an electrical coil. This coil creates the magnetic field to operate the valve and generates heat as a byproduct of its operation.*

P6, l181: Is the number of 1800 s correct, i.e., 30 minutes?

AR: *yes. BigLISA flew on a gondola attached to a large balloon. The gondola underwent slow descent following the balloon's burst. This allowed for longer sample collection times, particularly for the highest-altitude samples.*

P9, l217: Should this be section number 2.3.1 SPIRALE, instead of 2.4.1 SPIRALE?

AR: *thank you for noting this. It has now been corrected.*

P10, l223: Should this be section number 2.3.2 ACE-FTS, instead of 2.4.2 ACE-FTS?

AR: *corrected.*

P12, l272: Can you provide details of the differential pressure sensors? Could be quite useful to know for interested audience, in terms of materials to avoid.

AR: *the specific brand and model have been included (now Line 293):* ***Amsys, model AMS5915_0050_D_B****.*

P14, l327: Is the "self-consistent" statement fully applicable when comparing systems that sample either during ascend or during descend?

AR: *thank you for raising the good question. For regular weather balloon flights, the ascent and descent phases take a similar amount of time. Furthermore, the payload does not travel a significant horizontal distance during the flight, which is particularly true for the stratospheric sampling. Therefore, results from both phases should not differ greatly. For clarity, "self-consistent" was changed to "consistent".*

P14, l331: I'd suggest making the statement in that sentence more clear and spell out the conclusion with more clarity, for example: "However, we assume that differences due to instrumental effects remain marginal, while we believe that the day-to-day variability and long-term trends in COS mole fractions are the most important cause for the observed differences."

AR: *thank you for this suggestion, and we have changed the sentence to* ***"Therefore, we assume that differences due to instrumental effects remain marginal, while we believe that***

the day-to-day variability and long-term trends in COS mole fractions are the most important cause for the observed differences."

P14, Table caption: Should this be expanded a bit? It is not clear to me what this exactly means. Are there data gaps at specific heights, and are these suggested to be due to contamination from glue, surface effects, and differential pressure sensors? If so, how could this only affect short sections/a small number of the sample, but not all samples?

AR: *the caption has been changed to "**data gaps caused by clear contamination effects and their respective causes in different AirCore profiles**". During both sampling and analysis, air flows into and out of the AirCore in nearly a plug flow, causing very little mixing between adjacent portions of sampled air (Karion et al., 2010; Membrive et al., 2017; Tans, 2022). Therefore, if a contamination source is present along the coil, the contamination is likely to show as a "localised" feature rather than spreading through all the sample.*

P15, l334: Maybe add "as" to "and as we have no O3 measurements"?

AR: *done.*

P15, l358: Can you be more specific on "the variability" you refer to? Could you please explain what difference have in mind here?

AR: *"variability" was changed to "[…] in spite of the observed **differences between the retrieved profiles**, all profiles […]".*

P17, l386: Example where a marker/arrow within the data figure to identify/highlight that event could be helpful to better understand what the discussion is about.

AR: *we recognise that adding markers or arrows may generally help the reader in identifying specific features in a figure. However, the authors believe that in this case the introduction of an arrow may collaterally cover other data, moving the focus away from the main theme of the discussion.*

P20, Figure captions: Here and in other figures where relevant, can the time interval represented by the modelled COS profiles be included into the caption as well?

AR: *the time interval over which the remote sensing COS observations of ACE-FTS were averaged has been included in the captions.*

P22, Figure 6: Here and in other figures where relevant, can the regression functions be included into the figures? This would be much easier to follow than having them in a separate table. If they are needed in a table, I suggest showing in both forms.

AR: *the authors would prefer to exclude additional text from figures that are already rather busy. Tables have now been moved to be adjacent to their respective figures – the authors*

believe this solution to be clearer and more informative, since in the Tables it is also possible to report the uncertainties related to the results of each regression.

P24, l470: Say which laboratory.

AR: *"[…] with a QCLS in* ***the CIO laboratory at the University of Groningen*** *[…]"*

P25, l480: Change to "different sorts of inlets with different O3 scrubber materials."

AR: *done.*

P25, l482: "… variability, a hypothesis…"

AR: *done.*

Supplements, Figures 7-15:

- Where indicated in the figure legend, I am unable to tell ascent data and descent data apart; all I see is one colour.

AR: *the plots were checked to be colour blindness proof. For the TRN flights, Fig. S7-S9 (now Fig. S5-S7) ascent data was not available. The figures have been corrected.*

- Include in captions that the tropopause height is indicated by the green line.

AR: *done.*