Review's comments

Manuscript Number: *ESSD-2021-213*

Title: Two decades of flask observations of atmospheric δ(O_2/N_2), CO_2, and APO at stations Lutjewad (the Netherlands) and Mace Head (Ireland), and 3 years from Halley station (Antarctica)

Authors: Nguyen, L. N. T., Meijer, H. A. J., van Leeuwen, C., Kers, B. A. M., Scheeren, H. A., Jones, A. E., Brough, N., Barningham, T., Pickers, P. A., Manning, A. C., and Luijkx, I. T.

This paper provides two decadal data of atmospheric CO_2 and O_2 observed at Lutjewad and Mace Head and 3-year record at Halley. As is the case with the atmospheric CO_2, the atmospheric O_2 data from a variety of laboratories are also expected to be synthetically analyzed by using atmospheric transport models, biogeochemical models, and so on. However, compared with the atmospheric CO_2 measurements, the atmospheric oxygen measurements are still very challenging because we need much more efforts in the process of the air sampling, storing, analysis, and scale maintenance. Especially, it is crucially important to how to keep the O_2 scale stability. The authors describe the details of the calibration procedure and several efforts to check the O_2 scale stability. However, the authors should make much more effort to clarify the data quality and quantitatively describe the uncertainties associated with the flask measurements in the manuscript. Although I found that the paper contains material that should be published in ESSD, I recommend the manuscript to be published after following minor revisions.

General comments:

I understand authors' various effort to keep highly precise measurements of the atmospheric O_2/N_2 ratio of the flask samples. However, the δ(O_2/N_2) and APO values at Lutjewad and Mace Head plotted in Figs 6 and 7 show rather scattered plots, which don't seem to be real variations. So, I suspect that the uncertainty of the flask measurement is not so small to adequately detect the atmospheric variation. It is crucially important to clarify the total uncertainty associated with the O_2 data of the flask samples for the synthetic analyses together with the data from other laboratories. Nevertheless, I cannot find any clear description of the analytical precision and the repeatability of the flask measurements in the manuscript. In addition, the authors described the contamination of the flask samples collected at Halley during the storing period. If the same type of the flasks were used for the air sampling at Lutjewad and Mace Head, there is a possibility that the contamination would cause the positive and negative biases of the CO_2 and O_2/N_2 values for the flask samples, respectively. These potential biases should be also evaluated in the manuscript.

The O_2 scale stability is also very important as the authors also recognized. Although the evaluation of the stability is very difficult because there is no absolute scale at present, the authors should quantitatively evaluate the overall stability of the O_2 scale in this manuscript. In section 3.1, the authors described that the standard deviations of the repeated measurements of the working tanks were less than 13.5 per meg. Did it mean that the uncertainty of the CIO scale stability was estimated to be about 1 per meg/yr (=13.5 per meg/14 years)? Probably, the results of the COLLUM cylinders would also give a clue of the quantitative evaluation of the O_2 scale stability. The evaluation of the uncertainty of the scale stability is directly related to the evaluation of the uncertainty of the carbon budget evaluation described in page 20.

I think that the trend of APO at Mace Head is rather curious because the decreasing rate of APO trend gradually decrease from -15.15 per meg yr^{-1} in 2002 to -5.83 per meg yr^{-1} in 2018. The authors attributed to the O_2 emissions from North Atlantic associated with the gradual changes of the NAO. However, I cannot accept the mechanism that the O_2 emissions from the North Atlantic only influenced the O_2/N_2 and APO at Mace Head. The APO decreasing rates are computed from the fitted quadratic functions. However, taking the data variability for both sites and sparse sampling frequency for Mace Head into account, I suspect that there are no significant differences in the trends between the two sites.

Specific comments:

Page 2, line 55: “(Tohjima, 2005)” should be “(Tohjima et al., 2005)”. And please add the coauthors to the reference (Page 29, line 827).

Page 3, line 102-114: At Lutjewad, the air sample was dried by passing it through a Nafion drying tube. How about the Mace Head and Halley stations? The Nafion drying tubes were used at both sites?

Page 4, line 123: It would be better to clarify the temperature of the cryogenic drier.

Page 4, line 132-136: Were the same 2.5-liter glass flasks as Lutjewad used at Mace Head and Halley stations?

Page 8, line 252: Is it possible to describe the linear function to convert the O_2/N_2 value based on the CIO scale to that based on SIO scale? Is the conversion function fixed during the observation period of this study? In addition, I think it would be better to describe the uncertainties for the coefficients of the linear function. Such data would be useful to consider the propagation errors for the O_2/N_2 values of the flask samples and the standard cylinders.

Page 8, line 267-268: WT4845 show rather unstable O_2/N_2 values. It would be informative, if possible, to describe the reason of the instability.

Page 9, line 298-299: I believe that the conversions of the CIO values to the SIO values are based on the fixed conversion function. If so, the discrepancies in the O_2/N_2 ratio between the assigned values and the measured values suggest that systematic change in the conversion function, which correspond to the change in the CIO scale, or change in the O_2/N_2 values in the Scripps primary cylinders.

Page 9, line 303-305: Do the authors mention that the linear conversion function is often calibrated based on the measurements of the Scripps primary standards as shown in Fig. 4?

Page 10, line 317-318: “Manning et al., 2015” is not in the list of References.

Page 11, line 340: Why do the authors refer to the "WMO extended compatibility goal of 10 per meg"? The extended compatibility goal is set for the studies like urban observations that are strongly influenced by local fluxes. I believe that the authors aim to observe the background air through their three sites observation because they evaluated the global carbon budgets based on their observations in Section 5.1. Therefore, I think the authors should refer the "WMO Network compatibility goal of 2 per meg" here.

Page 11, line 349: All of the GOLLUM cylinders show the increasing drift (Fig. 5). Nevertheless, the authors described that the average overall drift rate significantly small (4 ± 6 per meg yr^{-1}). How did the authors calculate the uncertainty of 6 per meg yr^{-1}? In addition, I think that the scale drift rate of 4 per meg yr^{-1} is not small because it corresponds to bias of 1.6 PgC yr^{-1} for the carbon budget calculation.

Page 11, line 349-350: The WMO compatibility is defined as "a measure of the persistent bias between measurement records". Thus, it should not be compared with the scale drifting rate.

Page 12, line 370-372: If the authors fit a combination quadratic function and three harmonics to the data by using a least square method and do not use a digital filtering method of Thoning et al. (1989), the authors don't need to refer to Thoning et al. (1989).

Page 13, line 392-394: I don't understand the reason the exclusion of the last 2 years data. For example, the data at Mace Head in 2017 are much sparser than the data in the last year (2019). Additionally, I cannot accept the authors' idea that the sparse data in the last two years introduce biases in the fits. Since there are enough data to determine the average seasonal cycle for both Lutjewad and Mace Head, the larger number of data, even if the sparse, can cause the better fitting results.

Page 13, line 403-405: Please see the comment for Page 13, line 392-394.

Page 13, line 415: I think the longest period for the trend calculation is 17-year (from

2002 to 2018).

Page 18, line 483-486: It should be better to describe the detail of the drying method at Mace Head because there is no description.

Page 19, line 534-537: For the comparison of the observed results between Lutjewad and Mace Head, it is enough to simply compare their trends. I think there is no need for the authors to examine much about carbon budget calculations.

Page 19, line 536: The calculation method adopted in this study is not exactly same as that of Keeling and Manning (2014). In Keeling and Manning (2014), the NOAA's global mean CO_2 data was used to evaluate the accumulated CO_2 in the atmosphere and globally averaged annual mean APO estimated from the limited background observations was used to evaluate the change in the APO (not fitted trend line).

Page 20, Figure 10: There is no explanation about the red lines in the figures. In addition, the exact period for each annual average shown by the black dot is unclear. Is the period of the annual average for 2002 from January 2002 to December 2002 or from July 2001 to June 2002?

Page 21, line 583: I cannot understand the meaning of "noisier seasonal amplitude".

Page 21, line 596-597: Does the trend for Weybourne show much faster decrease than that for Lutjewad? It would be better to plot the trend for Weybourne in the figures.

Page 22, line 604: The authors described that the CO_2 discrepancies are shown in "the first half of 2016". But I think that the discrepancies are shown in July and, probably, August 2016. It should be clarified.

Page 22, line 603-607: It would be better to show the correlation plot of the differences of the flask CO_2 and O_2/N_2 from the continuous observations. The slope of the scatter plot would give us the information about the origin of the contamination.

Page 22, line 601: I cannot find “NOAA, 2021” in the list of References.

Page 22, line 618-621: The contamination of the flask samples collected at Halley is clearly shown from the comparison of CO_2 mole fractions of the flask samples with the in-situ continuous data.