Review of manuscript "atmospheric CO_2, $\delta(O_2/N_2)$ and $\delta\,^{13}CO_2$ measurements at Jungfraujoch, Switzerland: results from a flask sampling intercomparison program" by I. T. van der Laan-Luijkx et al.

General comments:

This paper presents results from a flasks inter-comparison program conducted at Jungfraujoch atmospheric observatory (high elevation station in the Swiss Alps) for CO_2, $\delta(O_2/N_2)$ and $\delta\,^{13}CO_2$ since 2007. The first parts of the paper present the context (involved laboratories), the site and the measurements technics used to retrieve the measurements discussed in the second part. In the later one, 4 years records of CO_2, $\delta(O_2/N_2)$ and $\delta^{13}CO_2$ obtained from flasks analysis in three laboratories (UBE, RUG and MPI) are presented, compared and analyzed in term of inter-comparability and WMO goal compatibility for each specie. The ultimate goal of the paper is to access the inter-laboratory measurement compatibility.

In general the paper is well written and well structured. It presents a nice and valuable set of data and an uncommon long triple inter-comparison record, including $\delta(O_2/N_2)$ measurements which are very spares. One weak point of this manuscript in my opinion is the lack of detail and deepening in the interpretation of the differences pointed out by the different records (see below for more details). There are most of the time described but little or no explanations are given or suggested to explain them.
I would therefore recommend publication of this paper after taking into account the comments below.

Specific comments:

In the following text, I will refer to page and numbering as appearing in the printing version of the discussion paper.

Page 7297, line 3: "The consistency for $\delta(O_2/N_2)$ …. Not better than +-5 per meg". Could you please add a reference or source for this estimate?

Page 7297, line 5: Is there a reference available to this SIO scale?

Page 7297, line 14: "Specific intercomparison … are rare". I would suggest to delete this sentence as there are indeed several other inter-comparison sites running and existing, and including several different laboratories, for example Mace Head in Ireland, Cap Grim in Tasmania or Alert in Canada. There are also several other dual sampling sites running over the world including at least two laboratories. This is also described in the sentence following in the text of the paper. It is of no use in my opinion.

Page 7298, line 3: Does LT stands for Local Time. Please explicit; this is not necessary straightforward for the reader.

Page 7299, Flask sampling (line 15-27):
This paragraph is a little bit confusing. On line 15 the authors state that the flasks are sampled with "dedicated sampling units". Then on line 20-21 we learn that "all the flasks are connected in series, using a single pump". This is not coherent neither really clear. What about the drying system, is it the same for all flasks, or is there "individual U-shaped glass tubes" (especially after March 2009). Also for the inlet lines, are there individual lines or one single large line with individual line aspiring air into this large initial line? Please clarify this paragraph. I would suggest adding a full schematic of the inlet and sampling systems that would help.

Page 7300, line 3-4: The flasks are then pressurized; do you use individual pumps/compressor for each sampling system? How is it done?

Page 7300, line 4: I suggest replacing "to the respective lab" by "to **their** respective lab".

Page 7300, line 6-11: Have you been able to quantify the pressure effect for the flasks of RUG and UBE during storage? It would be interesting to do so in order to compare with the final differences obtained (especially for $\delta(O_2/N_2)$ and $\delta\,^{13}CO_2$) and see if this can explain part of it (try to quantify it). From my personal experience it doesn't seem that there are such important effects for MPI type flasks with Kel-F O-ring for $\delta(O_2/N_2)$. Any comment on the MPI flasks? No information is given on those flasks, are they sent back to MPI after each sampling event or are there no effect on those flasks?

Page 7301, line 12: "… is extracted from the air sample with **liquid air** …"I suppose it is a tipping error and should be liquid **Nitrogen**?

Page 7301-7302, CO_2, first paragraph (line19-26/1-6):
My concern here is that there is no information on how the "valid flasks results" have been retained. Nothing is said about data quality control and/or instrumental precisions. What does "flasks results that were **obviously** influenced by measurement problems or leakages have been removed from the data set" mean? How are the leakage detected, what are the criteria retained to valid or invalid a data? When there are only one or two samples retained out of pairs or triplicate, how is it done, how do the authors know which value is the good one (except when there are atmospheric outliers)?
Could the authors please give precisions on that issue?

Page 7302, line 9-12: On the figure there are only the double harmonic seasonal fits which are plotted. Reading the text I would also wait to see the linear trend which is not the case. Is there a mistake there or is it a misunderstanding from my side? Please clarify this point. This is also the case for the other species (fig. 3 and 5). Also the legend of those figures state for linear trends.
Is there any reason to choose the value 2.7 sigma in the exclusive filter of the residual? Why 2.7?

Page 7302, line 16-18: When there are large differences shown by one of the three laboratories, is there any way to decipher between all labs and to flag the bad values with reference to the others? This is usually one goal of the inter-comparison exercises. In that case why don't you flag the bad values and remove them? Could you please give more precision on that issue or detail what you mean by measurements issues or small flasks leakages?
For example are there systematic shift or bias between labs, is there always one lab with frequent outlier … This is shown more or less in figure 2, 4,6 but not commented so much.

Page 7303, line 3-14. From this paragraph and table 2, it is clear that the change of sampling set up has a large impact on the CO_2 results presented here. But no explanations are given here to try to explain this problem. It is obvious that the calculation implying MPI have been deeply modified by the change conducting to a better agreement between MPI and UBE but a worse one between MPI and RUG. The difference between MPI and RUG has not been affected? So who is right? How can you explain this? Is it linked with the pumping unit used, the flasks, and/or the inlet lines?

I also wander why the set up was kept like this after March 2009 whereas the results were not fully satisfying?

Page 7303, last paragraph: As stated in the beginning of the paper, JFJ is a high altitude station representing back-ground air from Europe. How does the results for seasonal and inter-annual trends presented here compare with other back ground European stations (e.g. Monte Rosa, Puy de Dôme) or global worldwide values? This kind of comparison could help checking the validity/quality of each data set.

Page 7304, line 26-27: Why could the problem be attributed to the UBE scale? What kind of problem? What is the difference in the calibration scale in UBE compared to the two other labs? Please give more details and information on that point.

Page 7305, first line: I'm not fully convinced that this is true; it is only 3-10 per meg difference here in the std dev. The biggest difference in Std dev is for the UBE-RUG results.

Page 7305, line 2-5: The average values are of no significance here except for UBE-RUG because there are too many discrepancies between the results for both periods. This is also clearly illustrated in the percentage given in the following sentences of the text. Therefore I would not state that the difference is within 5 per meg as this is purely artificial from averaging procedure.

Page 7305, line 5-7: Could the authors give more precision on the improvements suggested. What kinds of improvements are recommended? Is it really only a matter of sampling procedures, storage or are there some limits on the instrumentation and measurements procedures as well? Please give more details.

Page 7305, line 9: I disagree with this statement. In table two there are significant changes for both periods when looking at UBE-MPI and MPI-RUG average differences (at least more than a factor of 2 change).

Page 7305, line 14-to the end of paragraph page 7306: The main differences arise when UBE data are taken into account. There is a pretty good agreement between MPI and

RUG. The conclusion of this analysis seems to be that the UBE data set is much noisier than the two others, is there an explanation for this?
Could you please add a reference/value which could support the fact the estimated trend obtained using UBE data set is unrealistic?
Is there other literature values representative of the free troposphere that are available for comparison with these results?

Page 7306, line 20: Linear trends are not shown on the figure.

Page 7307, line 2-3: I don't fully agree with the statement of the authors. The data could be used for interpretation depending on the scientific focus we have. If the idea is to go into detailed fractionation and partitioning processes then caution should be taken and the authors are right, but the trends and seasonal variation are there, the quantification is more difficult but still there is information to get out. I would suggest rewording this sentence in a more "optimistic" manner.

Page 7307, line 16: replace data from these flask **is** by flasks **are.**

Page 7308, line 3-4: Yes, but there are also in-situ inter-comparison program running at several other stations (see above).

Page 7308, line 5-10: There are also "flasks" inter-comparison programs running in Europe as for example the so called "sausage" and "Grapefruit". Even though these exercises are different from in-situ inter-comparison program it should have been interesting to compared both sets of results. This would then tell us about how far or close we are from standard laboratory results (as obtained in the later good conditions measuring programs).

Page 7308, line 14-15: please give a comparison number from the Cucumber program (e.g. see suggestion above going in the same way).

Page 7308, line 20-22: I agree but this paper lacks conclusion and recommendation about the labs that have been evaluated. Which one is the one to trust best for which species, which way to work on to improve the results in each lab, sampling procedure… As

stated earlier this is one of the goals of this kind of inter-comparison and there is no dedicated conclusion here.

Page 7315: Table 2. In my opinion, the average values for CO_2 and $\delta(O_2/N_2)$ are not significant because there is too much discrepancy between the two considered periods. Each period has to be considered separately.

On figure 1, 3, 5: There is no trends fit represented however the text and legend refers to them.