Atmos. Meas. Tech. Discuss., 5, C2713–C2718, 2012
www.atmos-meas-tech-discuss.net/5/C2713/2012/

Atmospheric Measurement Techniques Discussions

Interactive comment on "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.

Anonymous Referee #1

Received and published: 30 October 2012

General Comments

The manuscript by van der Laan-Luijkx presents results from the first 4 years (2007-2011) of an ongoing atmospheric air comparison between University of Bern (UBE), University of Groningen RUG), and the Max Planck Institute for Biogeochemistry (MPI). The experiment is based on comparison of discrete atmospheric air samples collected biweekly at Jungfraujoch Station. Samples are collected in the glass flasks identical to those used by each laboratory's respective measurement program. Samples for all

3 labs were collected in series until March 2009, at which time, the sampling apparatus was modified such that UBE and RUG samples continued to be filled in series while MPI samples were collected in a parallel. Measurements of CO2, δ(O2/N2), and δ13CO2 are compared as well as average annual trends and seasonal amplitudes.

The authors find that, based on this study, CO2 measurements made by UBE and MPI meet the WMO recommended levels of compatibility when averaged over the 4-year period. However, they note that variability in the mean CO2 difference between all labs exceeds the WMO recommendations. To assess trends and seasonal patterns, the authors first remove measurement outliers using a 2.7 times 1 sigma residual filter. Their analysis shows the average annual trend and average seasonal amplitudes determine from the 3 independent records agree with the stated uncertainty. Measurements of δ(O2/N2) between the labs do not meet the WMO recommendations for compatibility. Measurements from UBE are significantly lower than those from RUG and MPI, and the authors suggest this may be due to the definition of the UBE scale. The mean difference between MPI and RUG is within 5 per meg. However, they note that variability in the observed differences between all labs is large (20-40 per meg). The observed average annual trends from RUG and MPI are similar but do not agree within the estimated uncertainty. The observed trend at UBE is considerably different; the authors can produce a more reasonable trend when the UBE record is first filtered using a more restrictive filter. The RUG and MPI observed seasonal amplitude agree within uncertainty. The observed amplitude in the UBE record is considerably lower even after using a more restrictive filter. Measurements of δ13CO2 agree to within 0.03 ‰ over the 4-year period but variability around the mean differences is an order of magnitude larger. The observed trends in the UBE and RUG records agree within the estimated uncertainties and do not agree with the observed trend from the MPI record which is consistent with the trend derived from GLOBALVIEW-CO2C13 (based on decade-long records of measurements from the University of Colorado). The authors note that because the MPI measurements are more precise than those made by UBE and RUG (based on agreement between measurements of flasks filled in series), the MPI trend

over the 4-year record can be determined with higher confidence.

The paper is well organized. The presentation is clear and concise. The authors provide a thorough description of the comparison set up. The authors present the results and discuss the challenges in meeting measurement compatibility goals recommended by the WMO. The discussion and conclusions emphasize the importance of ongoing comparison experiments and their importance in terms of understanding carbon fluxes.

The authors' use of terms representing measurement uncertainty and compatibility differ from my own understanding of these terms. My reference for these terms is Table 2 (Definitions of selected terms related to data quality (updated according to VIM3) from WMO GAW Report No. 194, "15th WMO/IAEA Meeting of Experts on Carbon Dioxide, Other Greenhouse Gases and Related Tracers Measurement Techniques", WMO/TD - No. 1553 (Jena, Germany, 7-10 September 2009). However, I find these VIM definitions difficult to understand and I must admit that I do not have a fully understand these terms as they relate to our field.

My understanding of the term "compatibility" in the context of this work is the level of agreement over time between two independent measurement records. In this paper, the authors are assessing the compatibility of measurements made by 3 independent labs based on air samples collected at a field site. Compatibility can also be assessed within a single lab using measurements derived from different detectors (e.g. comparison of co-located Picarro and Licor measurements for CO2) or from measurements derived using different methodologies (e.g., comparison of flask and quasi-continuous measurements).

Internal measurement precision, as I understand it, is assessed within each lab and is typically based on repeatability and reproducibility experiments. In my view, internal measurement precision (e.g., standard errors in the mean value of replicate samples) is not an assessment of compatibility.

My impression (again, based on my understanding of the above terms) is that the

authors are not using these terms consistently. I provide examples in Specific Comments below. The authors may argue my definitions and I may be wrong. Regardless, I strongly recommend that the authors clearly define how they will use these terms and use them consistently throughout. [The terms described in Table 2 will likely be discussed and clarified and include examples relevant to greenhouse gas and related tracer measurements at the next WMO meeting of Experts.]

There is no discussion on the comparability of these measurements. Are the CO2 measurements from MPI and RUG on the WMO X2007 scale? How does the UBE machine CO2 reference gas compare to the WMO scale? Are all δ13CO2 measurements related to the VPDB scale? The only mention of scale in this paper as it pertains to the results presented is in the discussion on δ(O2/N2). The authors are encouraged to state whether measurements from the different labs are comparable, i.e., traceable to the same reference material.

I recommend this manuscript for publication in AMT after the authors have had an opportunity to address the comments included in this review.

Specific Comments

Page 7296, line 21. This sentence is unclear. The authors claim that most labs meet the WMO recommendation for CO2 compatibility because of present-day instrumentation. Based on the definition of compatibility, this implies that most labs are able to routinely compare measurements of atmospheric CO2 using independent methods, e.g., by comparing co-located 1) independent flask air samples (collected in situ and measured in the laboratory), or 2) flask measurements and in situ quasi-continuous measurements , or 3) in situ quasi-continuous measurements using different detectors. Several laboratories that do routinely assess intra-laboratory compatibility (e.g., EC, CSIRO, NOAA, NIWA) have shown that it is difficult to establish and maintain measurement compatibility to the levels recommended by the WMO. Are the authors instead referring to measurement reproducibility or repeatability?

Present-day instrumentation may likely improve the ability to meet WMO recommendations but references are required to support this claim.

Page 7297, line 3. "...not better than +/- 5 per meg." Reference required.

Page 7297, line 13. "Specific intercomparison projects... are rare." Please clarify. Many ongoing in situ comparison experiments besides same-air comparison and super-site experiments highlighted by the authors exist. For example, long-term flask sampling is co-located with in situ quasi-continuous measurements at Izana, Cape Point, Syowa, Barbados, Zeppelin, Mauna Loa, Lampedusa, Pallas, Oschenkopf, Mace Head, Baring Head, Mt. Waliguan, American Samoa, Hegyhatsal, Barrow, and more. These efforts play a critical role in justifying the merging of data from different labs.

Page 7302, line 5. "This is well within the WMO goal for compatibility..." The authors are relating the average of standard errors in the mean of measurements of flasks sampled in series, which is a measure of internal reproducibility, to measurement compatibility as defined by the WMO recommendations. It is not clear what point the authors are trying to make here. Further, it is quite possible for independent measurement streams to be compatible at higher levels than individual participants' estimated internal measurement precision or uncertainty. A statement about each lab's internal precision is useful in the context of how long it might take for a signal in differences to be significant beyond measurement uncertainty.

Page 7302, line 11. How was the 2.7 times 1 sigma exclusion criteria determined? Likewise, the 1.9 factor from page 7305, line 23. A few words would be helpful.

Page 7302, line 27. "If we start from the obtained average..." This approach suggests that the mean difference represents a significant bias and can be removed from the distribution of differences. The statistics do not seem to support this assumption. I suggest the authors state this assumption.

Page 7303, line 17. "The average annual trend obtained from the data sets <is>..."

Page 7304, lines 8-14. "Comparing this to the required WMO ..." Please see comments above (Page 7302, line 5).

Page 7305, line 20. "Since the focus of this study..." Do the authors exclude filtered values from the actual comparison of measurements or only from the trends derived from the records? It seems defensible to exclude these values from the determination of trends and amplitudes but not from the actual comparison unless the authors are using the filter to identify possible independent experimental errors. This is a subtle but important distinction.

Page 7306, lines 10-13. Again, I question the authors use the term "compatibility". See comments above (Page 7302, line 5).

Page 7308, line 1. "Global intercomparison programs are rare ..." This is not an accurate statement. There are $\sim$24 labs making ongoing co-located direct comparisons of atmospheric measurements at more than 28 locations around the world.

Page 7308, line 19-22. "Further efforts should be made..." This is very nicely stated. Have the authors discovered and corrected any experimental problems as a result of this work that have resulted in improved consistency (confidence) in one or more of their observational records? If so, I suggest highlighting these finds as it will strengthen the point above. As an example, the persistent offset in CO2 during mid-late 2010 in which RUG measurements are approximately 1 ppm lower than UBE and MPI is tantalizing. Are there insights as to the cause, which can be supported by other experiments?

Interactive comment on Atmos. Meas. Tech. Discuss., 5, 7293, 2012.