Reply to Anonymous Review #3

We thank the reviewer for the careful review and associated comments and suggestions. We have also added text that, we hope, improves readability and provides a more quantitative analysis.

General Comments

The main concern is with respect to the non-quantitative nature of language used in the manuscript. In many cases statements are not backed up with quantitative evidence such as whether a "good agreement" is significant within (and along with) a certain level of confidence. Definitions of statistical terms used are missing, e.g. how are "scale propagation errors" defined and derived? Both the abstract and the introduction are rather vague. Given the quantity of trace gases in the atmosphere it would help the reader to know, which ones are being addressed here. Some guidance on the relative importance/abundance would be useful, too. A general structural comment is, that large blocks of text make the manuscript hard to read in places – more sub-sections might help here. Also, the discussion order between compounds is inconsistent in places. Finally, and with regard to methods, little analytical detail is provided, especially for the primary reference lab (e.g. which detector was the NOAA analysis done on for the individual compounds)? A number of more specific points can be found below.
Our Response: We have made extensive revisions in response to the points above.
While it is true that few analytical details are presented, this is due to the difficulties of introducing analytical detail across 19 laboratories. Analytical methods are defined in Tables 3 and 4 for cases in which data were submitted from multiple instruments, and the full complement of results and laboratory information is supplied as Supplemental Material.

Specific Points

Abstract: Which scale relationships were found to be consistent with those derived from bi-lateral experiments and which were not? And why are there no quantitative statements at all in this abstract? How big are the discrepancies and the "substantial improvements"?
Our Response: We added some quantitative information to the abstract. A complete list of scale relationships is not practical for the abstract.

P8024, l24 please give formulas
Our Response: Chemical formulas have been added upon first appearance.

P8025, l8 The "existence of independent calibration scales" is initially important for verifying abundances with trends being secondary. The "relationships between independent scales" are probably also of importance for trends.
Our Response: We agree and have made a minor change. "While the existence of independent calibration scales is important for verifying trends, abundances, and estimating uncertainties, it is critically important to understand the relationships between independent scales."

P8025, l11 Which greenhouse gases? Preferably not halogenated ones?
Our Response: This paragraph has been revised. "We focus primarily on the results for halogenated species (CFCs, HCFCs, HFCs, etc.) and other greenhouse gases such as N_2O*,* SF_6*, and* CH_4*."*

P8025, l20 please explain abbreviation
Our Response: Revised for clarity.

P8025, l22 And small temporal gradients. It might be worth mentioning a clarifying example here, e.g. the importance of small changes for ozone recovery dates.
Our Response: Revised for clarity.

P8026, l14 define "dry"; Was a drying agent used at this stage?
Our Response: Revised as suggested. "dewpoint ~ -78 deg "

P8027, l27 please quantify "large amounts"
Our Response: Revised as suggested. ">70% of the initial air"

P8028, l1 please quantify "minor differences"
Our Response: Revised as: "Initial and final mole fractions determined by NOAA agreed within analytical uncertainties for all gases except CH_2Br_2*. Hence, initial mole fractions assigned by NOAA were used for comparison, except for* CH_2Br_2 *(adjusted for drift) and CFC-12 (described later). For gases not measured by NOAA, we assume no drift. This is a reasonable assumption given previous experience with these cylinders under similar conditions. Differences between initial and final mole fractions (when available) can be seen in Figures 1-6 and Tables 3 and 4 (columns 1A, 1B)."*

P8028, l8 Does that mean there were objections? If so, how were these addressed?
Our Response: Some participants raised questions prior to the experiment, but approved of the experiment plan. This sentence has been revised.

P8029, l8 Why is Table 5 the third table referred to in the manuscript?
Our Response: Table 5 is now the 5th table referenced.

P8029, l13 How are these 2 sigma defined?
Our Response: Revised for clarity. Scale propagation errors are a subject of research. We felt that a representative error should be included, and estimated the uncertainty as two times the analytical uncertainty of the scale origin.

P8029, l23 Which compounds exactly? And which compounds are included but have been considered to be of less importance (and why)?
Our Response: This sentence has been revised. CHBr2Cl and CH2BrCl were excluded because only one laboratory provided results. C2Cl4 was omitted by mistake, but has been added to Tables 3 and 4.

P8030, l11-14 Do these scales actually compare better (and if they do, is there a reason?) or is this merely a statistical coincidence?
Our Response: Given the uncertainties provided with the NIST data, we conclude that there is no difference between the 0.1% IHALACE result and the previous 0.9% result of Rhoderick and Dorko. The text has been revised.

P8030, l23 and 26 "nearly the same" and "1 % larger " Are these statistically significant differences?
Our Response: 1% is outside the range of reported 1-sigma uncertainties for the IHALACE experiment for these two laboratories.

P8030, l24-26 This is a discussion of unpublished data that is not included in this paper.

Our Response: A convergence of NOAA and SIO CFC-11 global mean data can be seen in Montzka and Reimann et al (2011). We have added this citation. Further, both NOAA and SIO data are publicly available, and this statement is easily verified.

P8031, l2-6 This is a difference of 0.15 %. Even if it is significant within the 2 sigma uncertainties of the NOAA analysis the initial results should not be rejected without evidence that they were flawed and that the cause for the discrepancy has been eliminated.
Our Response: We should have said "and this was confirmed by analysis <u>of additional standards</u> at NOAA" The initial NOAA CFC-12 results were indeed flawed. Including them would lead to erroneous differences.

P8031, l14-19 Are 10 ppt significant? Generally, and also relevant for the whole manuscript: It is impossible to decide whether measurements are compatible or not without quantifying the analysis uncertainties for each of the instruments involved in this comparison.
Our Response: 10 ppt is significant. Here we are comparing two scales at two points in time. Relative differences of ~2% should persist if the scales have remained stable, and they do not. We feel it is important to update results that have appeared in the literature.

P8032, l4-6 It is important to know, that differences should be smaller between some of the labs, but it would also be good to know, whether they actually are.
Our Response: We agree, and have added some text in this regard. "In general, transfer errors are indeed smaller for AGAGE-affiliated laboratories, but not in all cases."

P8032, l21 Is this significant?
Our Response: Yes, it persists throughout various data records.

P8032, l22 "co-elution" An interfering compound perhaps? Also, why are differences between ECD and MS results not mentioned/discussed for all compounds?
Yes, it is likely an interfering compound. ECD/MS results show differences for few other compounds, and we have added text to reflect this.

P8032, l26 The discussion is sometimes jumping back and forth between percentage and ppt differences making it very hard to follow for the reader.
Our Response: We have tried to be more consistent by using %, or including both ppt and %. We prefer ppt in some instances because the Figs. 1-6 are all in ppt.

P8033, l1-4 "Fewer" "small" "larger" Please quantify.
Our Response: Revised to be more specific.

P8033, l5-8 It might be worth explaining that most labs actually report the sum of CFC-114 and 114a. Also, why are CFC-114 and 115 not listed in Table 3?
Our Response: CFC-114 and CFC-115 were not listed because Set 2 was not measured by SIO, and thus we lack a set of measurements that can be used as a consistent reference. In hindsight, we should have sent both sets of cylinders to SIO.

P8033, l17-18 "bottom-up" inventories might also be based on measurements (of emission factors).
Our Response: We do not understand how this applies.

P8033, l21-24 This is only true if there is no temporal drift between scales which could significantly affect growth rates and thus have a large effect on emission estimates.
Our Response: There is evidence in the preceding paragraph that 2 scales (NOAA and SIO) have

shown the same relative differences from 2004-2009.

P8034, l8 Jumping back to CCl4 at the end of the CH3CCl3 discussion.
Our Response: Here we address an analytical issue that is common to CCl4 and CH3CCl3.

P8034, l18 “minor” please quantify
Our Response: This section now reads: “Differences due to scale transfer and analytical methods are on the order of 3%, except for the MD (ECD) measurements from laboratory 9, which are 10% lower than laboratory 2 (scale origin).

P8034, l25 “drift” Is this a “drift in mole fraction”, and if yes, what is causing it?
Our Response: We have revised to: *“This could be caused by downward drift of CCl_4 in standards used by laboratory 3, since CCl_4 at ppt-levels may decrease with time in some types of cylinders.*

P8035, l2-3 Could the type of detector play a role here?
Our Response: These differences are not likely related to the type of detector. As shown in Fig. 8b, non-linearities for ECD and MS detectors are typically < 5% for CH3CCl3. Detector-related errors of 100% or more would be highly unusual.

P8035, l5-6 This is not true. HCFC-22 is almost as abundant as CFC-11 and HFC-134a very close to CFC-113
Our Response: Thank you. We have removed this statement.

P8035, l19-20 Could you please make this statement clearer? Why is UB-98 included in the figure but not in the table?
Our Response: We have removed this statement, and added text to the table caption (see underlined text that follows).
“Table 5: Analysis statistics for undiluted samples, based on results from laboratories that develop calibrations scales. Results from laboratories that adopt scales are not included here because we do not want to introduce potential scale transfer errors. Gases with large differences between winter (W) and summer (S) fillings were treated separately. All mole fractions are pmol mol^{-1} (ppt) unless indicated. “

P8036, l8-10 The first two sentences contradict each other on the number of indepen- dent scales.
Our Response: We have fixed the error.

P8036, l20-23 It is not made clear, why “good agreement” means 1 % for CFC-11 but 10 % for Halon-2402.
Our Response: We agree. We have revised the text to remove many non-quantitative statements.

P8036, l23-27 This is the first place in the manuscript where the reader actually learns about differences in calibration procedures. I understand that giving details of all the different calibration procedures would make this a very long manuscript. But only few experts will understand the term “non-traditional manner” if it is not explained at all.
Our Response: We have added definitions earlier in the paper to explain the common approaches, so that we can differentiate between common and less common methods at this point in the paper.

P8037, l4, 9, 15, 18 “a few percent”, “a broad range”, “large” Please quantify.
Our Response: We have made revisions to be more specific.

P8039, l13 An uncertainty is only given for one of the two ratios discussed.
Our Response: An uncertainty was not reported by Huang et al 2007.

P8040, l15 These laboratories would only be "capable of resolving very small mole fraction differences" if precisions were sufficient, but these are not provided here.
Our Response: Precision (1 s.d.) for all results is available in the Supplemental Material. Even though some of these differences are within the reported uncertainties, we feel it is important to point out that consistent summer/winter differences, all positive, were reported by many labs.

P8041, l1-2 Why are data shown for all other compounds (including CH4 which "was not a focus of IHALACE") but not for COS? Also, a more correct abbreviation would be OCS, as carbon is the central atom in this molecule.
Our Response: COS was also not a focus of this experiment, and results for COS were reported after the conclusion of the experiment. Therefore, we include COS, but in a limited way.
Both OCS and COS are acceptable abbreviations for carbonyl sulfide. OCS is more commonly used in spectroscopy, and COS is commonly used by industry.

P8041, l10-12 Other molecules (such as CH3Br or CHCl3) show large seasonal differences. Why are linear relationships only inferred for OCS? And why is only one of these relationships mentioned?
Our Response: Data sufficient to produce linear scale relationships are limited to a few gases. We focused on gases for which a large seasonal cycle in the N.H. and the dilution of one third of the samples provided a large range of mole fractions that could be compared on multiple scales. Most gases only showed a range of ~20% that could be compared on multiple scales. The comparable range for COS is 30%. CH2Cl2 results span a large mole fraction range, but data available for scale comparison only span a range of 20%. The only other gas which was observed over a range > 20% on multiple scales is CHCl3 (range spans ~50%). The linear relationships expressed for COS are included as an example. Linear relationships are also identified for CH3Br in figure 7. All data are available as Supplementary Material.

P8041, l16 This section could use a brief introduction to linearity and connected problems (e.g. possibly explaining differences between detectors in the beginning. Few readers know about the non-linear character of an ECD measurement.).
Our Response: We agree, and have added introductory text.

P8041, l17 "majority of" should be "all"
Our Response: Agreed

P8041, l20 "the three major" might be better than "some"; "CFC replacements" is a repetition
Our Response: Agreed.

P8042, l4-5 Is this factor meant to evaluate different scales as well as linearity within one scale?
We can only evaluate the relationship between two scales. Is it independent of mole fraction over the
Our Response: range tested?
This is described as follows: "This factor provides an indication of whether or not a constant scale factor might be applied over a 20–30 % mole fraction range. A LF of 1.0 results when scales differ by a constant factor at both ambient and sub-ambient mole fractions."

P8042, l6 Why does the discussion start with CFC-113?
Our Response: We start with compounds shown in Figure 8a, which includes CFC-113.

P8043, l6 Why does this only suggest “a slight non-linearity” and not “a co-elution” as for CFC-113?
Our Response: We have revised to include both the possibility of non-linearity and co-elution.

P8043, l14 “small” Please quantify. Were any of these differences significant?
Our Response: Here, “small” is defined later in the paragraph.

P8044, l10-12 Why is this only important for SF6? It is not the only compound that is increasing.
Our Response: True, as is stated in the introduction to this section.

P8044, l19 This section is not very well structured. Important differences should be clearly identified, quantified and put into context with the respective atmospheric importance, abundance, and trend. Parts of this summary are repetitive of the abstract.
Our Response: We have revised this section and hope that it is clearer now.

P8044, l20, 21,22, 24 “numerous”, “modest”, “substantial”, “large” Please quantify
Our Response: See above.

Table 3 It is not possible to evaluate these differences without knowledge of the scale uncertainties and analysis uncertainties.

Our Response: We disagree. These differences can be evaluated with respect to potential data use. For example, if one wanted to use data from a particular laboratory to evaluate emission signals in a region, one could use differences from the table, along with analysis uncertainties shown in the figures (and in the supplemental material) to perform an initial evaluation of the relative error introduced by using data from another laboratory at a different site as an upwind boundary condition. These differences can provide information to guide the data user on whether or not adjustments to a common scale might be necessary. These data also provide information on data comparability (actual differences between labs linked to the same scale). They also provide information on differences among methods (data from the same lab, measured using different methods). One would also hope that the availability of these differences would encourage data users to contact data providers for more information and updates. A full assessment of scale uncertainties for all compounds is beyond the scope of this paper.

Figure 1 Axis labels are too small. It would help if labs on the same scale were adjacent to each other (making clustering effects more apparent). Error bars, and Set 1 and 2 are not explained in the caption. Although mentioned in the caption, there are no “non- integer laboratory numbers” in the figures.

Our Response: It is not practical to put same-scale labs adjacent to each other. This might cause more confusion because then the relative placements of labs would vary across the figures. Scales are coded by color, such that labs on the same scale can easily be identified, and we have tried to keep the colors consistent throughout (eg. NOAA scales always black, SIO scales always red). We have revised the caption to emphasize that symbols in the figures identify laboratories that serve as “scale origins” and those that rely on adopted scales. There is a lot of information in figures 1-6, and we hope that the revised caption improves clarity.

Final Note: Other reviewers also found that axis labels were too small. We will work with production editors to improve label readability. The caption has been revised to include a definition of error bars.