Authors' Response to Referee Comments:

Referee #1:

The paper by Andrews et al. presents the technical setup of the NOAA network of 8 tall tower sites to measure atmospheric CO2 and CO (and partly CH4) concentrations. The authors comprehensively describe the evolution of the instrumentation, its automation, and its evaluation from the 1990s to the current state to achieve the WMO comparability goals. The applied CO2 analyzers are also compared to the most recent laser based technology. The subject of the paper perfectly matches the scope of AMT; the methods need clarification on a few points mentioned below. However, the text needs considerable work to reduce lengths, avoid repetitions, and get a clearer structure before being published in AMT.

Main scientific concerns:

(1) Consistent negative bias of in situ CO2 measurement:
The flask - in situ comparison shows a dominating positive bias (Table 5).

Table 5 shows small biases initially, with the flask – in situ difference becoming increasingly positive over the course of several years. We now have evidence from laboratory tests that certain flasks are contaminated such that the presence of even low levels of H_2O in the sample (e.g., -20 deg C dewpoint at 1013 hPa) will cause elevated CO_2. We have observed positive offsets of up to 2 ppm in laboratory tests, and offsets of 0.5 ppm are evidently common. The nature of the contamination is not yet known, but the number of flasks affected seems to be growing over time. The dry air "chemical integrity" tests that have been used until now for routine testing have failed to detect this problem. We have begun systematic testing of all PFP flasks, with plans to clean all affected flasks. We do not yet know how to prevent the problem from recurring.

In turn the in situ measurement might be too low. Another indication for a low bias in the in situ CO2 is given by the tank air experiment provide to the inlet (p1506/L2).

We have done several tests and comparisons to evaluate possible biases in our systems over the years. We have been particularly concerned about the uncalibrated inlet components such as long sampling tubes, pumps, and chillers. (Nafion will be addressed below.) The referee points to a few tests where measurements with our Licor system were lower than the comparison measurement. However, there are other comparison experiments where the difference has the opposite sign, e.g., the $MgClO_4$-dried air test (Section 6.2.1) and WBI comparison with PSU CRDS (Manuscript Table 8). Since the publication of the AMTD manuscript, we have done several rounds of Licor versus Picarro comparisons in the laboratory, where humidified air was sample covering the range 0.5 – 4.5% H2O. Results from recent laboratory comparisons are shown in Figure 1 below, and laboratory and field comparisons under dry and humid conditions are summarized in Tables 1 and 2 of this document, respectively. The average bias over all tests is <0.05 ppm for both dry and humid cases (excluding the summertime values from the PSU comparison at WBI), although there have been individual tests with differences of 0.15 ppm or even larger. Admittedly, this is statistics of small numbers, especially for high humidity.

We are not confident of our ability to perform laboratory comparisons or Picarro H2O correction calibrations for H2O > 3.5%, since liquid water is likely to condense in the tubing at these values for normal laboratory temperatures. This is not of immediate concern, since humidity > 3% is unusual and > 3.5% is extremely rare at our sites. We used data from sites in the US Climate Reference Network to estimate the fraction of data with H2O > 3% and 3.5%. US CRN data for temperature and RH are more robust than our own meteorological data, and the network is sufficiently dense so that nearby sites exist for all of our sites. We found that sites in IA, SC, and MN have the highest humidity levels with 4.7, 3.5% and 1.3% of the hourly values exceeding 3.0%. Hourly values exceeding 3.5% were seen only at IA and MN, accounting for 0.2% of hours for which data was available for each of those sites.

I guess, that is due to the Nafion setup. The counter-flow of the Nafion dryer is operated with reduced pressure; thus, giving a pressure difference leading to diffusion. CO2 permeation is preferred to other gases, resulting in a CO2 depletion. As the calibration gases also flow through the Nafion dryer, this bias is partly compensated. However, a membrane with few water allows less CO2 diffusion than a wet membrane; therefore, the sample air is depleted more than the dry calibration gases (even if the air at the Nafion outlet has same humidity, the water concentration changes at the Nafion inlet).x

We have measured CO2 loss across the nafion membrane using a well-calibrated Picarro analyzer. We found that CO2 loss across the membrane is nearly identical for calibration gases and sample gas in our system. Calibration gases were either routed directly into the Picarro (bypassing the nafion) or sent through the nafion. Field conditions were closely matched: H2O exiting nafion ~180 ppm, flow ~250 sccm, internal nafion pressure ~10 psig, external pressure ~200 torr. We measured CO2 loss in the nafion of 0.125 +/- 0.05 ppm. We also simulated atmospheric sampling by routing gas from a cylinder through a bubbler and into the chillers (bypassing our pumps). H2O exiting the chiller was ~0.57% measured by the Picarro (chiller temperature ~3.5 deg C, chiller pressure ~10 psig), and CO2 loss across the nafion membrane was 0.10 +/- 0.03 ppm. Chiller temperatures are typically close to 1.6 deg C, so this is a worst case scenario.

This hypothesis is further strengthened by other arguments: Stephens et al. (2011) use a Nafion setup with almost no pressure drop in the counter-flow, and they see not a loss of CO2 (p1509/L5)!

Stephens sees a similar trend in the PFP comparison as we do, with the flasks becoming more positive relative to the in situ over the last few years. We see this at all sites and in our aircraft program, pointing to a problem with the flasks.

Moreover, there is also a slight positive offset of the Licor to the lab-calibrated Picarro (Rella et. al.,2012) (p1510/L11).

The Licor – Picarro difference of -0.04 ppm is within the uncertainty of the measurements.

Finally, the target tanks get an increased low bias, when measured independent from other dry air, i.e. a wetter membrane (p1503/L28).

This is an interesting theory, but the high target values following CO2C4 is more likely caused by hysteresis that has been underestimated (please see the response to comment 1499/10).

An open question is why Picarro and Licor show comparable results at BAO on the long term (p1511/L4ff). I suppose, both analyzer share the calibration gases lines (judging on p1512/L14f)? May you go through your experiment in section 6.2.1 again to clarify this?

The configuration with shared calibration gases described on p1512/L14f is specific to the permanent Picarro installation at WGC. For the BAO comparison, only the laboratory calibration was used, along with the laboratory water correction. Thus, the Licor and Picarro were truly independent (no shared calibration gases, separate sampling lines), and we will clarify this in the text. We originally deployed the Picarro at BAO to look for transient artifacts related to pressure changes in the PFP sampling line, and so we did not go to great lengths to calibrate it. We were surprised by the good agreement between the Licor and the un-dried Picarro, and by the stability of the difference.

This was the first definitive experiment showing that there is a problem with the PFP flasks themselves, and not with the Licor system or sampling lines. We have since acquired new lab data showing that certain PFP flasks exhibit very large (>0.5 ppm) positive biases after exposure to even low levels of humidity. Positive offsets appear when sample air has a dew point of -20 deg C at 1013 hPa, but not when sample air is dried to -70 deg C. We will begin cleaning all PFPs in an effort to mitigate the problem, but we still do not know the nature of the contamination in the PFP flasks or how to prevent it from happening. We are cautiously optimistic that lab tests will reliably identify contaminated PFPs, so that previous data can be flagged.

We devoted a lot of text to the PFP comparison in the Discussion paper, in an effort to show that the problem was most likely with the PFPs. Now that this has been confirmed, we will shorten section 6.3.1.

(2) Statistical background of the uncertainty calculations: To sum up different uncertainty terms to one error number (p1496/L14) is restricted to statistically independent error sources. However, the 7 components described here are often not independent (e.g. up and ub), or no random numbers with a normal distribution. Some are even biases that should be corrected for (ustdeq, usmpeq, uwv). The ideas presented here are valuable indicators when remotely assessing the analyzers deficiencies, but they should not be used in models for carbon flux estimates, which often strictly rely on a Gaussian error distribution.

We recognize that only independent errors can be added in quadrature. We are grateful for the constructive criticism provided by the referee. We will implement significant changes to certain terms as described below.

In practice it can be difficult or impossible to rigorously separate various contributions to the error. However, we have attempted to take particular analyzer characteristics into account in order to minimize dependencies. For example, in the case of the Picarro, we do use a baseline drift correction because the drift is negligible compared to the 30-second analyzer precision. For the Licor, baseline drift is much larger than the 30-second precision. In the case of the Thermo Electron CO analyzer, precision and baseline drift have comparable magnitude. Random errors in the raw signal and the baseline are independent. Errors due to unresolved baseline drift are difficult to separate from random errors, and we will revise the algorithm for u_b to use whichever of the u_p or the drift error is larger for any particular point.

In the revised manuscript, we will attempt to clarify where error terms may not be independent. We have done a simple Monte Carlo analysis of the uncertainties, and based on the results, we plan to revise the total uncertainty calculation to use standard errors (rather than standard deviations) for u_p and u_b, and to replace the prediction interval with the confidence interval, as was recommended by the referee.

Our target measurements do provide a check on the magnitude of the total measurement uncertainty. However, the target residuals do not provide a measure of uncertainties associated with extrapolating beyond the calibrated range or of artifacts caused by components that are not included in the calibrated path (i.e. tubing, pumps, condensers). For cases where 68% of the residuals do not fall within the estimated uncertainty, we will increase the uncertainty so that condition is met. We would like to use the comparison results shown in Tables 1 and 2 to estimate the magnitude of biases associated with uncalibrated components. A value of ±0.1 ppm encompasses 60-70% of the results, with the range depending on whether the Al valve manifold test and the PSU summertime comparison are included.

We do agree that known biases should be corrected in accordance with the guidance provided in the GUM and VIM documents (please see references at the end of this document for the full titles). . Initially, we were not confident enough in the robustness of our algorithms to apply a correction. By now, we have dealt with many of the cases that cause the algorithms to fail. The u_{smpeq} and u_{stdeq} terms are more significant than we originally thought, and we will modify our algorithms to correct for differences from equilibrium values using an exponential fit (see response to 1499/10). In the case of extrapolation errors, $u_{ex,}$ we will acknowledge that a correction would be a better approach in the revised manuscript, but we do not have sufficient extrapolation error measurements for each unit to confidently apply a correction. The extrapolation errors are only important for atypical conditions such as fire. It is trivial to implement a correction for u_{wv}, so we will do that, though this term is only significant when the analyzer is malfunctioning.

Modelers using CO2 measurements to estimate fluxes assume that measurement errors are Gaussian because they have not developed tools to deal with biases that certainly do exist. We feel that providing more detailed uncertainty information may help motivate development of such tools. The Integrated non-CO2 Greenhouse Gas Observing System (InGOS) is also working on a scheme for detailed uncertainty reporting, including separate estimation of repeatability and systematic errors for historical datasets such as CH4 from Heidelberg ("The InGOS Project: Setup and First Results", A. T. Vermeulen, S. Hammer, P. Bergamaschi, U. Karstens, O. Peltola, I. Levin, presented at the NOAA Earth System Research Laboratory's Global Monitoring Annual Conference, 21-22 May 2013, Boulder, CO).

Comments on presentation style:
Please, substantially shorten the text to do a favor to reader and reviewers. To cut the length of more than a half seems reasonable, even without losing valuable information. Only the most eye-catching examples are pointed out in the specific comments below (repetitions, unnecessary information). I prefer a more precise/scientific than narrative style. The manuscript often illustrates the full story behind the development with all its drawbacks. Even though science works like that, in my eyes, that does not match the AMT journal style. I recommend clear statements about the final setup, with some short reasoning, why the materials/methods were chosen by either citing own experiences (instead of the full story) or other references (currently done very rarely). When rewriting the text, a clearer structure can be achieved by a clear division between the setup description and its evaluation. Now, it is often confusing when technical solutions from different points in time and sites are mixed and even mixed up with future suggestions, which sometimes even leave open whether they have been already applied. I would recommend a restructuring of the manuscript, e.g.: A Introduction (chapter 1) B Instrumentation/Hardware (instrumentation chapter 2.1-2.10, incl. leak checks, lab validations 6.2) C Automation (calibration 4 + 5.1, alerts 3.1, data uncertainty 5.2) D Results (Data evaluation (target, Picarro): 6.1, 6.3.1, 2.11+6.3.2+6.3.3, 6.3.4; add time series) E Conclusions (incl. short list of future recommendations (chapter 3.2, 7, 8)) Even though the discussion of data results is out of scope of the paper and the journal, I would personally prefer a figure with the time series of e.g. CO2 from all stations. It would easily visualize Table 1 and would easily illustrate what kind of signal can be seen from the data to prove the introductory words right in the

conclusions. Moreover, a small map showing all sites in the US would nicely illustrate the network.

We are grateful for the specific suggestions to reduce length and reorganize the manuscript. The referee is clearly knowledgeable about GHG measurement techniques, and some of the details presented here may seem quite obvious. However, we are often contacted for advice by researchers who are new to GHG measurement. The anecdotes and examples were carefully chosen and are intended to prevent others repeating our mistakes (see responses to specific comments below). We will relegate some sections to an Appendix (e.g., temperature control problems, description of PFP sampling system) and will eliminate unintentional redundancy.

We will present time series and interpretation of the data in other papers. It is difficult to imagine a single figure or small set of figures that would show the data in any comprehensive way. We could include a site map, but it doesn't seem necessary, since the purpose of this manuscript is to describe the instrumentation. We are certainly not looking to make this paper any longer, and prefer to focus on the instrumentation rather than the data or the network in the current manuscript. We will make some changes in the introduction to this end.

Specific comments (page/line):

1463/22ff: It is sometimes difficult to follow the introduction, because it sometimes does not follow a clear argumentation line. The order of arguments in the paragraph starting on page 1464/22 might be put in a better way, to understand the idea to use a tall tower for atmospheric measurements. At the moment the whole work is not well motivated. I would also put paragraph 1466/12ff in front of p1465/21, as it first describes the networks, than it more and more focusses to the presented network. The last paragraph also fits better on p.1466/12. You mention already many important facts, but I miss a clear argumentation that culminates in a clear statement about the novelty/importance of your work for the science community. We will attempt to improve the introduction.

1464/9: leave out this sentence here, no connection. We disagree. The WMO recommendations for CO2 are based primarily on the goal of resolving net annual flux. If monthly fluxes were the signals would be an order of magnitude greater, and the recommendations could be substantially relaxed.

1464/28: This footprint was calculated for a tracer without diurnal cycle, for CO2 the far distance signal is diluted much faster. For CO2 or any other tracer, the contribution of the near versus far field depends on the distribution of sources and sinks. The idea that opposing day and nighttime fluxes cause nearfield amplification, is only true for the case of a very homogenous fetch. Air can travel hundreds of km in 24 hours, so in reality the vegetation encountered on days 2 and 3 may be quite different than on day 1. However, since both reviewers commented on this, we plan to drop the estimate of the footprint size altogether, since it is not essential for this paper.

1465/23: since when it expanded? 2007? Start dates for each site are provided in

Table 1 of the manuscript.

1465/26: why > 300 m, why is it representative for the planetary boundary layer? The original rationale was to use the tallest towers available, in order to collect samples well above ground level and minimize near-field influences. Early data from the NOAA tall tower sites ITN (in North Carolina) and LEF showed very little vertical gradient above 100m. We now target towers that are >300m because air above ~250 m is frequently decoupled from the surface at night, and so the highest intake nighttime data has a very different footprint from the afternoon data and from the nighttime lower level data. If models evolve to the point that nighttime simulations are reliable, then we may be able to use the nighttime data to constrain flux estimates for regions that are far upwind.

1466/17ff: What other models exist apart from Carbon Tracker? Why this model is introduced and emphasized here so comprehensively? CarbonTracker is just a convenient example, with which we are very familiar. We agree that there are too many details about CarbonTracker, and we will include a few additional modeling references.

1467/21 and Table 1: Please mention full name of the station at the first time you mention it, later you may use the 3-letter-abbreviation. Some of the sites do not have "full names", for example, the site name LEF is derived from the TV-station call letters WLEF-TV.

1467/26ff: Why so much advertisement of the system at this position, when it is not even presented yet?

It seems appropriate to introduce the instrumentation section with a summary of the main advantages over it's predecessor. The modularity and temperature stabilization are the principle differences from the Bakwin et al. [1998] design.

1468/5f: What kind of measurement technique is the Licor? please mention it before presenting all alternatives Agreed.

1468/10: "core of the tall tower system". If it is also your system remains unclear. We will clarify.

1468/11f: Please provide the exact type number and company name. Agreed.

1468/16: in section 6.3.1 nothing about the Licor calibration frequency is written The text should point to 6.3.3.

1468/16: insensitivity to environment > minimize calibration gas? This is not necessarily directly linked; an internally erratic sensor also needs many calibrations. We will reword.

1469/6: please give the type and company in an unambiguous way throughout the whole paper, e.g. (type: xyz, full Manufacturer name, country), see also 1469/16, 1470/5, 1470/8, 1470/10 etc. Agreed.

1469/19: What happens, if the flow does not return? Reword: Flow gradually returns to previous levels, typically within a few days.

1469/24: Is the pressure drop of 44 hPa realistic? It is probably a correct theoretical number on a straight tube, but I would expect a much larger pressure drop in practice. Did you ever but a pressure sensor upstream the first pump to validate this number? I cannot find a record of having done this test. Pressure transducers were not included upstream of the pumps in order to minimize the number of upstream components. We will attempt to do this test at the BAO tower

before submitting the revised manuscript.
1470/22: Which box? Only clear when reading the paper the second time. We will clarify.
1471/2: Where is the transducer in the figures? It is not so easy to follow without an overview figure. We will add the transducer to the figure.
1471/4: Why Viton? Viton was used in the Bakwin et al. [1998] system, and we adopted many of their materials choices. For clarification, EPDM is used only in the exhaust pump because it is supposed to be more durable than Viton.
1471/4: Is quick-connect fittings a general term? You mention the producer only later. We will include a brief definition at the first occurrence.
1471/25: poorly motivated, why drying of H2O is necessary (only in chapter 7.3) We will add an introductory sentence or two.
1472/7: What is the final dew point? **1472/13**: I see! OK as is.
1473/6: What laboratory and field tests? We measured the dew point in the lab using a Vaisala Dry Cap sensor. Later, when the Picarro was added to the WGC system, we were able to monitor the drier performance. Why you write here the type of the Picarro? We will add model info to first occurrence and omit thereafter.
1473/8ff: Write this description in the beginning of the paragraph, describe Fig.2b, than you can present the performance. OK.
1473/14: Improved drying, yes, but also improved CO2 diffusion! Yes, but CO2 diffusion is the same for cals and samples as shown by lab tests. We will add these details in the revised manuscript.
1474/22: Did you ever estimate the maximal tolerable leak rate of this valves? This gets even more important for more stable analyzer in the future, as the calibration gas sits longer time and gives room for CO2 diffusion. The Parker 9-series gives 10E-7 cc/sc/atm, the 99-series would have 10E-8. Standard valves have several orders of magnitude larger leaks rates. The Valco multi-position valve has 10E-10 cc/sc/atm, you may keep in mind for the future. We adopted the Galtek valves from an earlier design that was implemented at the AMT site. That system did not have comprehensive flow monitoring, and so cross-port leaks may have gone unnoticed. Only after the cross-port leaks occurred at WGC did we review the specifications for the valve, and we determined that they were inadequate. There were repeated failures with Valvo valves in the Bakwin et al. [1998] system, and in that configuration, in-field replacement was very difficult. My understanding is that reliability of the Valco valves has improved, and with the current modular system no in-field replacement would be required. So the valco would be a good option for future systems, but the steel solenoid valves work well enough that we cannot justify the expense of replacing them.
1475/9f: Why is the temperature control relative to the room temperature? Why you do not use a simple, but absolute controller? With this setup the performance relies on a good air conditioning, which is much more difficult to achieve. The temperature controller is absolute. However, the setpoint is site-dependent and is chosen to be 10-15 deg above room temperature. We will clarify in the revised manuscript.
1476/5: Where is this pressure controller located in Fig. 1 or 2? Figure 2d—we will note in text. I though you are using a MKS pressure controller (p 1475/L18)? That is

correct, the Tavco part number refers to the pressure relief valve. We will move the part information to be adjacent to "pressure relief valve". The whole section would benefit from a better figure and/or clearer structure.
1476/22: I guess, the motivation to include a filter in the setup to avoid introduction of debris is clear if you mention it once in you manuscript but not every time you mention a filter. OK.
1476/23: Why you scrub CO from air, even though you want to measure it? You explain it one sentence later. Please turn the arguments to make reading easier. That is valid for the whole section: Your argumentation line: CO measurement principle > pump removed > cell pressure sample flow > sensors removed > zero removed > external gas > ...; why not using: CO measurement principle > cell pressure sample flow > external gas > zero removed > pump removed > sensors removed > ... . It is arduous reading right now. We will rearrange.
1477/4ff: Almost no valuable information. Shorten it. At least the information about the size of the scrubber would be useful to some readers, but we will try to shorten.
1477/15: Repetition to 1475/11. Cut once. Agreed.
1477/15ff: Merge it to: Maximum operating temperature for both analyzers are 45 degC. That would be inconsistent with how the Licor specifications are reported.

1477/25f: "RTD(). We performed ..." > cut it to: ", which is optimally placed in the center of the enclosure." Of course, there were tests done etc., but a paper cannot be a log of all lab experiments. Agreed.
1478/10: The whole paragraph gives few valuable information. That the temperature control relative to ambient air can cause trouble is obvious (see my comment on p. 1475). I would summarize the full paragraph to one sentence: We use a doubled calibration frequency at the WGC site, because of a higher variability in room temperatures. As stated above, our temperature setpoint is absolute. In practice, it can be difficult to find a set point that works for the whole year (i.e. the room is very cold in winter, so the heaters cannot achieve a setpoint temperature that would be higher than the summertime maximum temperature).
1479/10ff: Shorter or completely cut. We will shorten.
1479/19ff: Shorter, as it is really no novel innovation (e.g. The PC and data logger is synchronized to a time server to limit the time drift below ... sec.) Why you are not using GPS sensors to permanently getting current time stamps? Or use a clock card for the PC to provide constant time even without internet? To synchronize with an internet time server, each operating system already has an internal routine provided. We will shorten this. The point is that PC clock drifts are significant and so time must be synched.
1479/26ff: The paragraph is quite narrative and can be shortened. Revisit.
1480/15: Why you start a paragraph with pointing to another one? Please restructure. This section is meant to describe hardware, but it seems appropriate to give some context.
1480/16: Did you test this pressure regulator for diffusion effects on CO2 and CO? Apparently not, I don't understand what is meant by diffusion effects in this context. These are the same regulators that are used in the WMO CO2 calibration system and

are used throughout our laboratory. How much flushing is required? Perhaps more than we are doing (see response to 1499/10). Calibration equilibration was a primary driver of our 5-minute sampling sequence. I recall that it takes a few minutes to drain the regulator at a flow rate of 250 sccm, so we are probably flushing only 2-3 volumes in 5 minutes.
Is this type still available, as I cannot find it in the online product cataloque? The part number should be 51-14C-CGA-590.
1481/2: It is confusing to read the metric units for the OD, since it is ordered in inch only. We agree but are under the impression that SI units are required for this journal.
1481/11: The WGC installation... > The installation at WGC site? For the reader all site names are not necessarily common. Agreed.
1481/21f: The sentence reads like an advertising booklet from the manufacturer. The sentence before already implies that the control is done by the instrument. We will omit the offending verbiage.
1481/26f: There are a lot of other sensors not included in the Picarro. This information is not necessary for the understanding of the system (might be enough to show it in an overview plot and a table of all sensors (with unique names) in the supplement). We monitor flow through the analyzer in order to ensure that the flushing is adequate and for accounting such as allowed us to find the cross-port leak problem in the sample manifold. It has been suggested by the manufacturer that the Picarro control valve voltage (or is it current?) can be used as a proxy for flow, but we find that signal is not sufficiently constant over time to enable accurate flow accounting.
1481/28f: Exhaust is described in 2.4, and in Fig. 2b. Data acquisition is described in 2.9. It is somewhat redundant information that disturbs a fluent reading. These details are unique to the WGC Picarro installation.
1482/21f: Of course the alerts developed over time. It is not worth to mention in a scientific paper. It seems worth mentioning that the software must be flexible since not all failure modes can be anticipated.
1483/2: Why not discussing all cases at one place? We will omit this sentence—not sure which cases were meant.
1484/6ff: Very narrative paragraph. I am interested in the final solution, with a short note that certain valves should be avoided. The intent is to warn so that others don't repeat this mistake. Many systems do not have adequate housekeeping data or algorithms to detect such a failure (e.g. the Bakwin et al. 1998 system did not monitor bypass flows, the Picarro is not factory equipped with a flowmeter).
1485/6: This kind of leak check is also not too novel. Did you estimate a maximum tolerable leak rate for this pressure test? Perhaps not novel, but practical and important.
1485/7ff: Very narrative and often in contrast to high-accuracy measurements. Accurate measurements are of little use when the height from which air is being drawn is unknown.
1485/24ff: Repetition to 2.8 We will omit.
1486/7: A 24 h cycle is always bad for a calibration cycle as it mimics other influences e.g. from temperature diurnal cycle. You mention this shortcoming later,

but this is experience from other research groups as well. Yes, we know this from discussions with colleagues. We would be happy to cite a reference from another group and will search for one. With 4 cals per day (originally) and multiple target measurements, we had adequate information to detect diurnal influences on the data.

1486/18: Why you show and discuss Fig. 5 at this place? You did not even mention the calibration, but already show the data. We will consider whether this can be reorganized. The vertical gradients present at night determine how much flushing time is required. The range of values 365 – 420 ppm CO2 gives an idea of what range of standards are required. Some knowledge of the expected signal is needed to design the sampling sequence.

1487/11ff: Your thoughts are correct but everybody can do them, when setting up a different system. You should justify your choice when you mention the 5-minutesampling the first time. We will consider reorganizing. These are the considerations that are relevant for our system and sites.

1487/21f: You considered it, but what happened? Why you decided against it? You may put it into the future recommendations. We were concerned about adverse effects of integrating volumes (e.g. materials effects, infiltration of water, non-uniform temporal weighting), and also the loss of high frequency information. The use of integrating volumes may be advantageous for some sites. An ideal system would have two analyzers, one to dwell on the top level, and the other for profiling.

1488/9ff: To find optimal calibration times, the trial and error method can be successful, but it does not necessarily is. Why you don’t use more sophisticated methods like Allan variance to determine the calibration time? Optimal calibration frequency may be site-dependent or time-dependent (e.g. depending on seasonally varying air conditioning or heating cycles or on time-varying analyzer performance). Allen variance estimation would require running gas from a tank for long periods, and results from a single test may not apply to all conditions.

1488/22ff: The experiences with the prototype system can be shortened. The true differential zero configuration has been used by other groups (e.g. Bakwin et al. 1998, Daube et al. 2001), and we were disturbed that we could not get it to work, and wanted to warn the community about this potential problem.

1489/13: Difficult to follow: You doubt the linearity of the fifth-order polynomial results, and you want to check it with 4 tanks only? The linearity of the (linearized) signal seems to be adequate, since we have seen good residuals (<0.05 ppm) for many units. We use 4 tanks so that we can cover the range from 350 -410 with ~30 ppm between standards, and the 460 tank was added to bracket the nighttime values from low levels.

1489/25: How do you guarantee the 0 ppb CO in the scrubbed air? We don't have a routine check for this, other than the residuals for the calibration curve, which include the zero. An early version of the system had the capability to scrub the calibration gases, with the idea that if the scrubbed 350 ppb standard was higher than the 100 ppb standard, that would indicate less than 100% scrubbing efficiency. Running the check daily consumes a lot of gas, and we were initially constrained by the 24 hour measurement cycle. This capability exists in the current system, and we should implement this check something like once per ten days.

1490/3f: You start here to discuss the measurement technique again (see sect. 2.6) The CO2 interference seems to be a calibration issue and so appropriate to discuss here.
1490/16ff: Some repetitions, and some ideas can be moved to future recommendations. The H2O/nafion sentence will be moved to the instrumentation section.
1491/13: Why you mention it here? The analyzer used is out of context here and confuses the reader. OK.
1491/28ff: What does this sentence mean? The idea is that sometimes the pre-deployment calibration gives better residuals for the field calibrations, other times the post-deployment value may provide better residuals. For CO2, pre/post deployment cal differences are rarely significant. For CO, however there have some large differences, which we are still trying to understand.
1492/5f: Reference for this statement? Why you still do it, it seems not necessary? We will clarify or omit.
1492/17ff: Redundancies to earlier descriptions. Seems worth including here.
1493/12f: The baseline is added to the raw differential CO2 data? Not subtracted, right? Perhaps the word "baseline" is somehow confusing, but we do subtract the time-interpolated signal corresponding to CO2C2 in the case of CO2 and COZER in the case of CO.
1493/20ff: merge it with information of chapter 2.8 Not sure to do this before we have discussed baseline drift. We may move the entire discussion to an Appendix.
1494/14f: Why you don't use one long calibration every 3 days? Would save calibration gas and equilibration time. Interesting idea, especially in light of the long equilibration time for our system (see 1499/10, below). If we want to have the timing drift so as not to alias diurnal cycles of room temperature, it might take quite a long time to cycle through the day—depending on implementation.
1494/21: Repetition. This single-sentence repetition seems warranted so that section 5.1.3 can be read without having to refer to previous description.
1495/7ff: This discussion fits into the introduction to motivate the paper, not here when describing the methods. OK.
1495/23: Many studies... Give references! OK.
1495/25: Repetition. True but arguably worth repeating in this section.
1496/1ff: The given information of the paragraph has no value at this place. You may mention the magnitude of the calibration scale differences that contribute to your absolute uncertainty in one sentence. The rest is either redundant or not related to your work. Yes, this paragraph is awkward.
1496/14: You missed the "square root" of the quadrature sum of the seven terms? The condition for this is statistical independency of all terms, but they aren't (see scientific concern (2)). I understood the phrase "adding in quadrature" to mean the square root of the sum of the squares, but perhaps this is incorrect.
1497/15: Standard deviation from 3 values is hardly statistics. It seems reasonable to consider the 3 nearest baseline measurements, and the standard deviation seems the best measure of consistency. It may be more appropriate to use the standard error for consistency with other terms.
1497/17f: How does this function look like? As a first estimate, it is probably linear

to the distance from the baseline measurement. But shouldn't it be sigma/2 in the middle? The weighting function is equal to zero at the time of each baseline measurement and equal to one halfway between successive baseline measurements. The whole idea might be better handled more serious from a statistic point of view (e.g. use Allan variances). At the moment the green line in Fig. 6 already includes parts of the instrumental noise up. Baseline uncertainty necessarily contains both drift and precision. As discussed above in response to scientific concern #2, we plan to revise u_b so that it is the larger or u_p (red line) and the drift term (current u_b), and report as a standard error instead of a standard deviation. Allan variance is a nice idealized concept, but rarely useful in the field where conditions change seasonally and from site to site.

1498/3f: The statement here confuses. How the target tanks are distributed is written somewhere else, and a different pattern is suggested here. Why you did not use it from the beginning? No different pattern is suggested here. In section 6.1, we describe an early version of the sampling program that included one target measurement per day adjacent to the full cal and three other target measurements that were distant from full cals and baseline checks.

1498/10f: Up to which CO2 concentration this estimates is valid? We used standards up to 650 ppm for the extrapolation tests and will state this in the revised paper. There are only a handful of events in our data record with CO2 > 650 ppm.

1498/22f: Why do you use the prediction interval instead of the confidence interval? From my understanding, the prediction interval is used to predict an unknown statistical distribution, but here you know the measurement data already and should use the confidence interval. Am I wrong here? Maybe you can give a short reasoning/references here. The prediction interval assumes that individual future will have similar error characteristics as the standards used to compute the curve (as indicated by the standard residuals). The confidence interval represents uncertainty in the calibration curve, while the prediction interval represents uncertainty in individual measured points (From the Igor Pro Manual Version 6, page III-195: "95% of measured points should fall within the prediction band. If you could repeat the experiment numerous times, 95% of the time the fit line should fall within the confidence band.") But, it seems that by using the prediction interval and separately estimating the analyzer precision, u_p and baseline errors, we are double-counting the errors in the sample. We plan to retain the u_p and u_b terms, and switch to the confidence interval for u_{fit}.

1499/10: The difference isn't an uncertainty estimate. It is a bias that can/should be corrected for, e.g by using an exponential fit. Best would be to exclude it by sufficient flushing and/or an optimized setup (dead volume, diffusion). See immediately below.

1499/10: This difference isn't an uncertainty estimate neither (not like a standard deviation). Did you try to extrapolate the function in Fig. 9 to estimate the equilibrium value? The decay time of the exponential fit should be related to the mixing time of your cell/setup (mixing time = volume / flow).

Figure 2 (this document) shows an exponential fit to the normalized calibration data. Our assumption of total equilibrium at the end of 5 minutes is clearly faulty.

We used the exponential fit to correct the standard values before computing the calibration curve and also to correct all of the sample values, and the difference in mole fraction compared to our original values is shown in the lower panel of Figure 2. The effect is significant for CO2 transitions larger than 50 ppm. The calculated e-folding time of 97.3 second is very long compared to expectations if flushing of the licor sample cell were the primary consideration. The volume of the sample cell is specified as 10.86 cm^3, and the volume flow rate through the cell is ~299 cm^3/minute, corresponding to an e-folding time of ~2.2 s. The effective volume corresponding to the observed e-folding time is 5 $cm^3 s^{-1}$ * 97.3 sec = 467 cm^3.

We will implement this type of correction in our post-processing software, and instead of reporting separate u_{stdeq} and u_{smpeq}, we will report a single u_{eq} based on the standard errors of the residuals to the exponential fit. Note that this correction should reduce the difference between target values following our high span versus following ambient samples, such as was seen at WKT (manuscript p1503, line 24). The typical difference between CO2C4 and CO2TGT is ~75 ppm, so for the WBI Aug 2009 case described above, CO2TGT values following CO2C4 would be biased high by 0.06 ppm, consistent with the difference observed at WKT.

1500/10: Please use SI units here and replace 10E6 by 1, than the formula is valid for any unit. Otherwise the H2O should also be divided by 100. OK.
1500/25: What is the relative importance of each of your 7 terms for the final time series? It would be nice to see it exemplarily on some part of a time series (similar to Fig. 11). Table 4 is meant to indicate the relative importance of the terms. Many of the terms are significant only when the system is malfunctioning (but perhaps not fatally), or when unusual atmospheric conditions are encountered (e.g. during a fire u_{ext} may be significant for CO).
1501/19: Repetition. We will try to eliminate.
1502/5: at WKT only? Yes, we will add a statement that patterns seen in the residuals at WKT are similar to those from other sites.
1502/9: If you use the Licor water corrected output, than you don't need the dilution error term. If still used, it becomes more complicated. Did you evaluate the influence of the wrong H2O measurement on your final result? Maybe the internal algorithm gives overcorrected data for negative H2O readings. We used the licor water corrected output, but since we compute a calibration curve using our own standards in the field, then it seems the correction is still accurate. The values for this term are negligible except when something malfunctions and the water changes very quickly (like in Figure 10).
1502/11ff: Shift it to Sect. 3.2 The text should point to Section 4.2. It seems best to describe the behavior of the residuals in the context of Figure 11.
1503/10f: The target tanks indeed provide an independent measure of analytical uncertainty. Why it is not used in chapter 5.2, it is a better statistical measure.
We liked the idea of keeping the target residuals separate to check the uncertainty estimates, and that has generally worked well. An earlier version of the calculation inflated the estimated total uncertainty if needed so that 68% of the target residuals fell within the uncertainty. Since we want to provide the best possible uncertainty

estimates, we will revert to that strategy takes advantage of the extra information provided by target measurements not adjacent to standards (e.g. cases like WKT during 2006 as shown in Figure 11).

1503/12: Did you ever check the influence of the pumps separately? Not the pumps separately, but the tests described in section 6.2.1 and 6.2.3 were intended to test all of the upstream components.

Why you use the pumps upstream the analyzer and not downstream like the Picarro? There were several considerations that went into the Bakwin system, and we did not re-evaluate all of their choices. Regarding the pumps, running the Licor at increased pressure improves signal to noise (not much of an issue with the Li-7000), with increased pressure, there is a tendency for air to leak out of the sample lines rather than in; when air in the chillers is compressed, a lower dewpoint is achieved; similarly nafion drier performance is improved since the volume flow rate is decreased.

1503/19ff: Very narrative again. A useful illustration of how strategically sampled target tanks can provide information about problems that could otherwise go undetected.

1504/12: 0.2 ppm are quite a high bias, in case the data is used for carbon flux estimates. Based on the results summarized in Tables 1 & 2 of this document, we plan to revise this to 0.1 ppm. The CO2 flux estimation problem is very hard for annual mean uptake. However, monthly flux estimates tare not so sensitive to small biases and can inform about processes.

1504/15ff: This section is again written like an anecdote. It is not clear why you give all this information to the reader, but without conclusions from it. The conclusion is that artifacts associated with un-calibrated components are of order 0.1 ppm – a pretty important conclusion, and the best we could do prior to the availability of Picarro analyzers and the evaluation of the Picarro H2O correction.

1505/3: May this difference be due to surface effects? When switching from wet to dry air the H2O molecules in the tubing and valves give place to the less adherent CO2. At the outlet one would see the remaining water and less CO2. The whole experiment depends a lot on time scales and materials used. In combination with my scientific concern (1), this experiment can give central answers to the observed biased, thus may need to get more attention. Please, double-check the sign of the differences (it is written the opposite way from the explanation in brackets p1505/L3f) Prior to the availability of Picarro analyzers, this was the only method we had for evaluating potential biases in our system due to wetted and uncalibrated components (pumps and chillers). We performed the test on the SCT and SNP systems prior to deployment, and repeated the tests several times over the course of a few months with very consistent results. The sign is indeed opposite from the "tank air sampled through BAO inlet tubes" experiment, i.e. wetted air sampled through the sample ports was higher than the control. The control was wetted air dried with MgClO4 and sampled through the target port, bypassing pumps and chillers. With the SCT system and using a Valco valve in place of a solenoid manifold we observed an offset of +0.04 ppm (sample –CO2TGT port). With the SNP system that was equipped with aluminum solenoid valves, we saw an offset of +0.15 ppm.

1505/11ff: The paragraph could be more precise. The described test is quite limited to stable conditions. The inlet tubes may bias the air for highly variable conditions the most. We were looking for systematic offsets caused by the tubing, and so it seemed reasonable to use the integrating volume to reduce variability. It is difficult to do this experiment any other way with a single analyzer. I don't think inlet tubes will bias air more under variable conditions, though they might dampen the response.

1506/5ff: First paragraph rather fits into introduction.

1506/14ff: Did you do a storage test for the flasks? Are Teflon O-rings ideal? How long do flasks wait until analysis? Please mention the drying (it can be only seen from the supplement). Storage tests performed with dry air in PFP flasks show that CO2 is lost at a rate of 0.007 ppm per day. We don't know if the effect is the same in humid samples. Since we currently fill one PFP over a period of two weeks, some flasks wait as long as 20 days for analysis. Early on we were able to exchange packages once per week, so delay times were typically < 14 days. We expected to see evidence of the storage offset, but never could, even early on when the flasks were apparently cleaner.

1507/27ff: No valuable information in this paragraph except the last sentence. We will improve this paragraph so that the relevance of the first three sentences is more clear.

1509/12: Why the bias should systematically should increase with time? The trend in the comparison is caused by a growing number of contaminated flasks.

1509/24: Please work on it, and include it in the paper. A bias should be excluded from every measurement system. When we have a routine testing and cleaning protocol and/or a revised sampling protocol that prevents the contamination from occurring, we will write a separate paper describing the PFP sampling system. We plan to abbreviate the discussion of the PFP setup and comparison in the body of the current paper and relegate essential details to an Appendix.

1511/12ff: This is the central sentence of the section. The rest can be condensed. Agreed.

1511/17f: Why you describe both setups here? Why you mention a firmware update? We will shorten.

1513/12ff: By far too much information. That can be written in the log file of the data, if somebody is interested in data during a certain time. It cannot be part of a publication to list every failure individually. We will shorten.

1514/22ff: Repetitions... Appropriate for this section.

1515/19: Why four tanks? An absolute and linear instrument might even live with one single tank to track the drift. These recommendations may hold true for your system,but are not necessarily true for other setups. For that reason, the manuscript cannot compete with a full review paper here. These are meant to be general recommendations based on our experience with multiple analyzers with diverse characteristics. Field calibrations should be based on a sufficient number of standards to provide meaningful residuals and to have a separate target tank. An incorrectly installed regulator could compromise any single tank—a small leak could escape detection but cause fractionation of the gas en route to the detector. The intro to section 7 clearly states that these are recommendations based on our experience—we do not claim these are based on community consensus or literature

review.
1515/25ff: Many repetition follow and by far too many sentences. Yes, you should have a cycle not equal to 24 h. That is the only recommendation here. Some of these points may be obvious to members of the Greenhouse Gas Measurement Techniques community, but they are important guidelines for novices.
1516/13 - 1519/11: Many repetitions. This part can be strongly reduced. Section 7 is a summary of important points from the manuscript and so some repetition is appropriate.
1519/12: Where do these ideas come from? Why it should be in this paper? These are merely examples of the kind of ancillary data that would be useful for interpretation of GHG measurements. It seems relevant and worth mentioning in the context of recommendations for future GHG monitoring efforts.
1520/6: The conclusions are insufficient. Please summarize, what you achieved, what accuracy you can reach, what validations have been done. Then give a more specific outlook what experiments or improvements can be done in the near future. Please also cover CO and CH4 measurements. CO seems to be the most difficult species to reach WMO specifications. We will expand the discussion to include CO and CH4 , although we were not attempting to meet WMO specifications for CO. Our target of 10 ppb was what we thought we could achieve with inexpensive sensors.
1520/11: Don't argue what you would need for hypothetical further work, but summarize the work you have done, e.g. show a resulting time series. An outlook may include some hints for further hypothetical work in the end. We will summarize the performance of our own system but we are not inclined to include a time series.
1520/21: So the whole setup is insufficient for the purpose it was built for? Our own tests and most of the comparisons in Tables 1 & 2 of this document are consistent at the level of 0.1 ppm or better. However, the comparison with the Penn State CRDS at WBI was poor with NOAA higher than PSU 0.3 under summer afternoon conditions. We do not know the source of the discrepancy. The PFP comparison is very disappointing for CO2, but we can now confidently blame PFP contamination. We desire to have ongoing independent comparisons at all of our sites so that we can reliably detect problems with uncalibrated components that may develop over time or depend on ambient conditions.
1520/21f: "Several research groups..." and what have you done? We will clarify. We have achieved repeatability of order 0.1 ppm through the calibrated portion of our system. Comparisons and lab tests indicate that bias is < 0.1 ppm. We simply wanted to acknowledge that other groups have also achieved repeatability of this level.

Table 4: why you give medians here? The standard deviation is the only valid measure that can be used for adding up independent error estimates. The table is intended to show typical values for the various terms across the various species/analyzers. The values for "Total Analytical" were not computed by summing the monthly median values, above but instead were computed for individual 30-second measurements.

Fig. 1: Confusing picture, as it is hard to follow the air stream. Additionally, the reader is sometimes lost in the text, which component is exactly meant. A clear flow diagram with unique notations would help. We will try to improve the figure, but the flow through the system is very convoluted because of practical considerations (e.g. the Nafion enclosure is chilled, and we did not want it to sit above the data acquisition system, lest water should condense and drip on the circuit boards).

Fig. 9: The unit of XCO2Difference is probably not ppm? The legend suggests unitless, as it is normalized. Why the data starts negative? It should result in a drift from the positive side, as it is normalized to the difference to the previous interval? The values have been normalized to correspond to a +1 ppm transition, so values approach equilibrium from below.

Supplement Fig. 4: How do the inlets look like? Are they ice shielded? Do you have any lightning protection? We do not have ice shields, but the inlets are housed in a fairly sturdy enclosure. We will include a picture in the revised Supplement. We do not have lighting protection beyond what is present on the towers already. We have not had any problems caused by lightning.

Minor corrections (page/line):
1467/15f: (Jeong, et al. 2012) and (Deeter et al., 2012) are missing in the reference list. Thank you.

1493/18: stored in an array. Why array? Just "stored." is enough. Agreed.

1498/12: Start a new paragraph before "The range ..." OK.

1514/2: "fantastic resource" sounds funny in a scientific publication. We will change to “unique” resource.

Fig. 7: Legend: Please, be consistent with the text: use uex instead of uext. Thank you.

Fig. 8: b) CO unit is not ppm, but ppb? Correct. Thank you.

Fig. 11: The name of the vertical axis might be changed, as standard deviations are also shown. We will address this.

Fig. 14: Legend: Description is mixed up: a) is Picarro, b) is Licor not vice versa. Thank you.

Referee #2

General comments: This paper by Andrews et al describes the NOAA tall tower measurements of CO2, CH4 and CO dry mole fractions. As the authors mention, the instrumentation to measure these gases has evolved since the beginning of the

NOAA tall tower program and now cavity ring-down spectrometers make these measurements easier. But I agree that this paper is a useful contribution to the literature and I appreciate the practical advice. For example, using a valve on the inlet to be able to test for leaks in the sample line, using dry ice to test for leaks (and using a hand-held sensor to make sure levels aren't too high for safety!), and checking for a torn diaphragm by capping the inlet and checking the flow. The length of the paper could be reduced by eliminating or moving to the supplement descriptions that are specific to NOAA or could be done many ways. For example, the first three paragraphs in Section 5.1 are not very helpful, except that three days of calibration data are used. Combining Figures 1 and 2a-e would save space.

We recognize that the paper is very long, and we will try to move some sections to Appendices. We included details about the SYSMODE and INTERVAL so that we can refer to these fields in the discussion of uncertainty algorithms, but perhaps this can be avoided.

Specific and technical comments
p. 1463: Accurate measurements of atmospheric CO2 do not, by themselves, provide an objective basis for verifying reported emissions. We need models to make that connection. Agreed.

p. 1464: Regarding the statement that the typical sampling footprint is 1/10th the area of the contiguous U.S., other studies (Lauvaux et al. 2008; Gerbig et al. 2009) emphasis the importance of the near field. Both referees commented on this issue, and we are inclined to eliminate the reference to the footprint altogether rather than getting into the details. As mentioned above in the response to Referee #1, the near-field amplification is strictly only true for homogenous fetch. However, we can save this for a modeling paper.

p. 1465: "Background values of CO2 are relatively high (currently _390 ppm) and vary with latitude, altitude, and time, so signals from individual sources are rapidly diluted, becoming faint." I don't see the connection in this sentence. We are trying to motivate the need for high precision. The signature of e.g. an urban plume in CO2 disappears faster than for a species with a low or constant background (e.g. based on persistence of plumes observed by research aircraft during field campaigns).

p. 1469: Please explain why the Reynolds number is relevant here. Also could use just the average value, since the exact number is not important and the repeated parenthe- ses are awkward. The pressure drop depends on whether flow is laminar or turbulent. Engineers like to see the Reynolds number.

p. 1484: "Leaks within the field laboratory" ... missing word. We mean to say that leaks within the sea container or building where the equipment is housed rarely develop spontaneously – as opposed to leaks that occur in the tubing on the tower that may develop as a result of falling ice, high winds, or vibration.

p. 1482: “Recent studies have shown that Picarro measurements of CO2 and CH4 can reliably be corrected for water vapor effects.” True, if you want to characterize each instrument individually and re-check periodically. Otherwise, there are problems at high water vapor values. Yes. We should include those caveats.

p. 1491: “It’s uncertainty is ~0.7 ppm.” Too casual, plus there should be no apostrophe in any case. Agreed.

p. 1491: “14 cylinders had absolute differences > 0.1 ppm.” How much larger than 0.1ppm? We will provide more details.

p. 1514: “fantastic resource” reword Agreed.

p. 1515: “We recommend deploying any analyzer with two or more additional cylinders than required to generate a calibration curve.” We will improve this sentence. See also response to Referee #1, who objected to this recommendation.

Table 2: This table is very specific and could be removed or put in the supplement. We will relegate it to an Appendix.

Also I’m not sure why (CO2, CO, CH4) follows “Water content of the sample flow”. Agreed that this should be clarified—we separately monitor the H2O content in the CO2 sample airstream and the CO airstream because they have separate nafion driers.

Figure 2: Is there any way to combine Figures 1 and 2a-e into one figure? That would take up less space and eliminate repeated legends. Also consider writing out “sample” instead of abbreviating as “sam”. Ideally, we would provide a single schematic, but how to show detail and make it fit on a single page? Maybe we could make an oversize figure and include it as a supplement.

Figure 4: no a or b in the figures. In a) this is the CO2 compared to the reference cell? It is a differential signal in approximate ppm – uncalibrated analyzer output.

Units are ppm? In b) what are the units? Seems high for CO in ppb? The CO analyzer baseline steadily drifts up. We do not want to cause discontinuities in the data by continually adjusting the baseline, so we simply subtract the measured baseline value -- ~415 ppb in this case. We will clarify this in the caption.

Figure 6: In a) the baseline is compared to the reference cell? Yes. It is a differential measurement. It is a little surprising that it is so low in this case, which was chosen at random. Typically the value is within the range -10 to 10 ppm.

Table 1: Summary of Dry Air Comparisons (H2O < 1.5%)

	Tower - Other	**Uncertainty**	**H2O %**
BAO tank through inlet tubes 2010-03	-0.12	0.1	0
BAO Picarro Comparison 2011-09	-0.04	0.06	0.6 - 1.4
BAO Comparison 2011-10	0	0.03	
BAO P3-instrument on elevator	-0.16	0.2	0.16 - 0.53
BAO P-3 instrument shared intake 29 July to 1 Aug 2008	-0.04	0.06	0.44 - 1.53
WKT P-3 on aircraft 13 & 25 September 2006	0.01	0.14	1.1
BAO P-3 on aircraft 1 April 2008 20:54:57 -21:19:57	0.01	0.27	0.4
LEF wintertime network flasks	0.01	0.02	
WBI -PSU afternoon average	0.13	0.63	
AVERAGE	-0.02	0.09	
Standard Deviation (n=9)	0.08		
AVERAGE Excluding PSU	-0.04	0.05	
Standard Deviation (n=8)	0.06		

Table 2: Summary of Humid Air Comparisons (H2O > 1.5%)

	Tower - Other	**Uncertainty**	**H2O %**
MgClO4 test -- SCT system with Valco (April 2008)	0.04	0.1	~2.8
MgClO4 test - SBP system with Al valves (April 2008)	*0.15*		~2.8
LAB Picarro Comparisons Apri-May 2013	-0.04	0.07	2-2.5
LEF summertime network flasks	-0.09	0.14	>2%
WBI PSU July - August	*0.33*	*0.83*	>2%
AVERAGE	0.078		
Standard Deviation (n=5)	0.15		
AVERAGE Excluding PSU & Al valve case	-0.03	0.06	
Standard Deviation (n=3)	0.05		

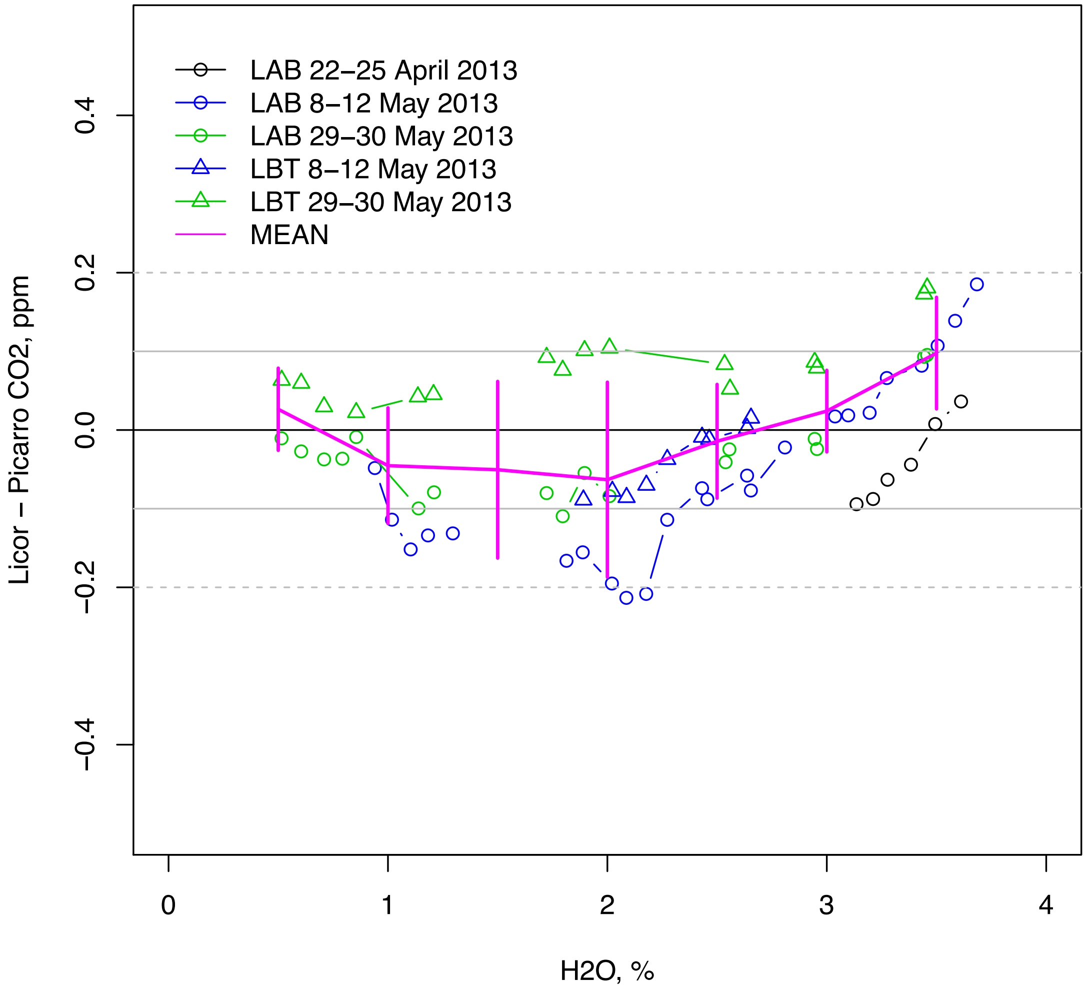

Figure 1. Summary of Laboratory Licor/Picarro Comparisons during April-May 2013. The mean curve was created by smoothing data from each experiment over the range spanned by each test, and then values were interpolated to 0.5 ppm intervals and averaged. Error bars represent one standard deviation.

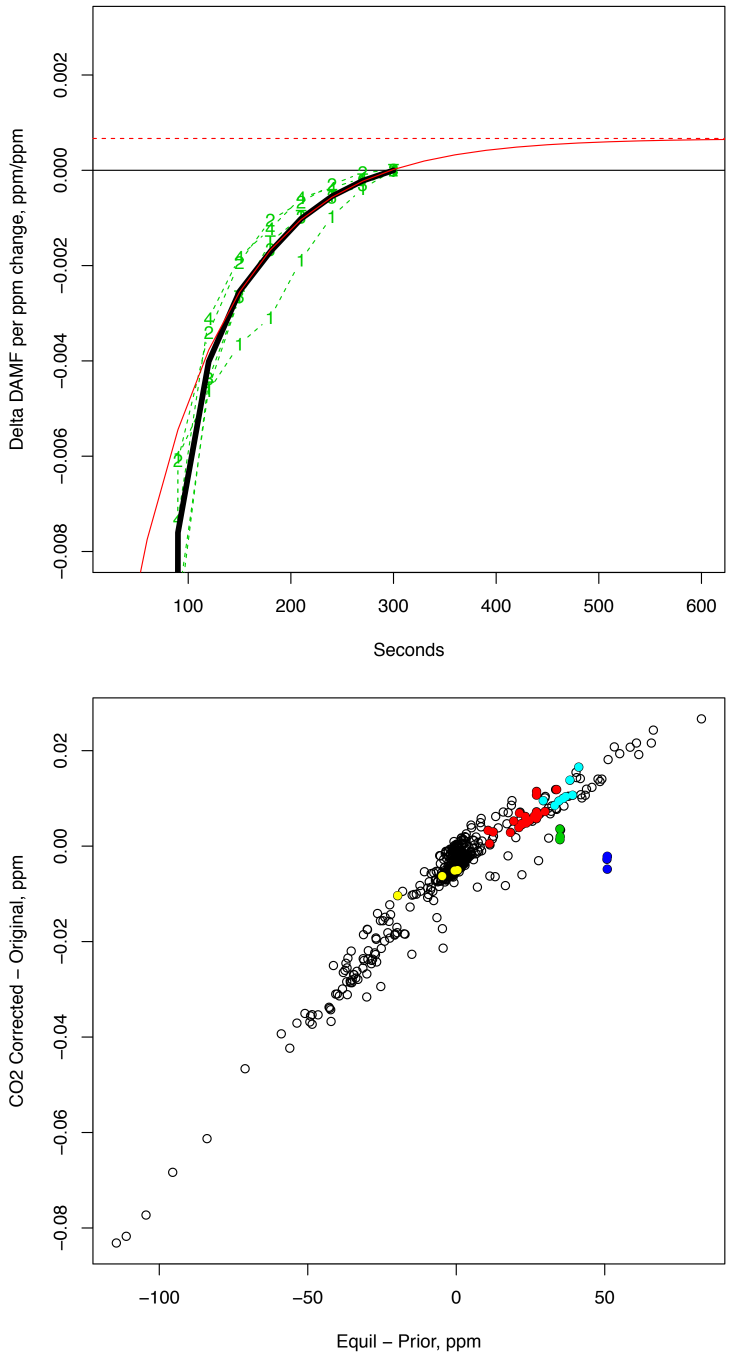

Figure 2. (Top) Revised figure 9 showing the approach to equilibrium fitted to data with x ≥ 150. The red curve is a fit to the data with the equation y=0.00066819-0.003223*exp(-(x-150)/93.759) , and the horizontal line represents the estimated equilibrium value 0.00066819 ppm/ppm. (Bottom) change in calculated CO2 when equilibrium values are used to compute the calibration curve and sample DAMF instead of the uncorrected analyzer values. Symbols corresponding to the standards are colored (CO2C1=yellow, CO2C2=red, CO2C3=green, CO2C4=blue, CO2TGT=cyan).

References:

GUM: ISO Publications, Guide to the Expression of Uncertainty in Measurement, International Organization for Standardization (Geneva, Switzerland), ISBN 92-67-10188-9, 110 p. (1995). (The abbreviation of this title is GUM). Equivalent Guide: American National Standard for Calibration - U.S. Guide to the Expression of Uncertainty in Measurement, ANSI/INCSL Z540-2-1997, NCSL International, Boulder, USA, 101 p. (1997).)

VIM: International vocabulary of metrology. Basic and general concepts and associated terms, 3rd edition, Joint Committee for Guides in Metrology (JCGM), 2008, http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf. WMO, All WMO references below are available from: http://www.wmo.int/pages/prog/arep/gaw/gaw-reports.html