We thank the reviewers for their detailed and helpful comments and give our responses to each of these comments below:

Responses to Anonymous Referee #1

C1: Page 6993, line 4: Remove "at a pressure of ~1100mbar". The cell volumes are 5cc irrespective of the gas pressure.
R1: We have removed "at a pressure of ~1100mbar" from this sentence in the revised version.

C2: Page 6993, lines 25-27: Does this mean that the calibration sequence is zero/low- zero/high-zero/low etc. or zero-zero/low-zero-zero/high etc.?
R2: It is zero-zero/low-zero-zero/high etc.. we have clarified this in the revised version.

C3: Page 6994, line 6: Can you be more specific about these corrections? How large are the applied corrections?
R3: An experimentally determined exponential curve has been used to correct the bias, and the bias corrections range from 0.7 ppm to 0.1 ppm. We have modified the texts.

C4: Page 6994, line 13: "...one year or even a couple of years..."
R4: Corrected.

C5: Page 6996, line 6: Drifts should be given in "per time"-units, i.e. specify during which time period the observed change occurred.
R5: We have changed to "The observed drift for 0.75 L cylinders is -0.15 ± 0.06 ppm / 100 days during these tests".

C6: Page 6996, line 22: How did you derive this number (0.2ppm)? Is it based on the laboratory tests? Or on the comparison with the flask samples?
R6: The number is derived from laboratory tests. When the suggested rules are followed, deviations ranging from -0.2 to +0.1 ppm have been observed. The reason for the deviations is mainly due to deviations from the linear long-term drift, and non-ideally flushed regulators (for ideal stable signals the flushing time would have to be unreasonably long). Here we take the maximum deviation as a conservative estimate.
We have changed the sentence "When these rules are followed, such a calibration system can be said to supply the measurement system with a stable CO_2 mixing ratio within 0.2 ppm" to "When these rules are followed, deviations ranging from -0.2 to +0.1 ppm have been observed in the laboratory tests. Therefore, our laboratory experiments suggest that such a calibration system can supply the measurement system with a stable CO_2 mixing ratio within 0.2 ppm"

C7. Page 6996, lines 23-26: This is not clear. What internal and external standards are meant here? Which heat flow do you refer to?
R7: The internal standards mean the small calibration cylinders inside the NDIR analyzer system, and the external standards are 50 L laboratory working standards. The heat flow refers to the exact thermal impact discussed in Sect. 2.1. We have rephrased the sentences as follows:

"In addition, we compute the CO_2 mixing ratios of the small calibration cylinders inside the NDIR analyzer system by measuring three calibrated working standards as sampling air on the same NDIR analyzer system. This mimics the atmospheric sampling, and can compensate for known biases, e.g. the thermal impact on measurements of calibration gases (similar impact on measurements of sample air immediately after calibrations has been discussed in Sect. 2.2)."

C8. Page 6998, lines 20-21: For clarity either use P or p as the symbol for pressure (also in the equations).
R8: We use lower case p as the symbol for pressure at any time, whereas capital P for the pressure at particular times. Here should be capital P, instead of lower case. We have added a sentence in Section A.1 to clarify this. "throughout the text, we use lower case p as the symbol for pressure at any time, whereas capital P for the pressure at particular times."

C9. Equation (1): The fraction bar before (1-exp(-ts/tau)) is missing.
R9: Yes. It was a typesetting error.

C10. Section 3.3: I agree that a very likely explanation for the discrepancies is insufficient drying, and you present an elegant way of correcting it, but at the same time it is also a bit shaky. If you do not force the linear regression through zero, then for many flights (in particular in Fig. 10a) it looks like there is no significant relationship at all with the water mixing ratio. So one could just as well correct for the mean difference per flight and then obviously the mean difference becomes zero for the corrected data...
R10: It is true that correcting for the mean difference per flight will result in zero mean difference and even slightly smaller standard deviation for the differences between flask and corrected in situ measurements. However, this method would correct the in situ data with a relatively large offset even when the ambient water vapor mixing ratio is low (e.g. in the free troposphere), and could lead to an overestimated vertical gradient. The chemical dryer we used was magnesium perchlorate, and it won't humidify the sampling air if it was already dry. Therefore, we prefer to correct the in situ data based on water vapor mixing ratios other than based on a mean bias per flight. It is worth pointing out that our method does not result in zero, but a small mean difference (~0.1 ppm), and the residual difference adds to our estimated uncertainty. We have also added one sentence to clarify the uncertainty of flask analysis prior to the last sentence in the first paragraph of section 3: "...verified. The typical analytical precision of the flask measurements at MPI-BGC is smaller than 0.06 ppm. Therefore..."

C11. Page 7002, lines 5-7: What kind of uncertainties are given? 95% confidence intervals?
Could the difference of the trends also be explained by the uneven distribution of the data (most data 2007-2009) leading to a bias in the linear trends?
R11: The uncertainties are given as standard errors of the estimated trends. It is true that the data prior to 2005 are sparse, and the slope is 2.15 ± 0.14 ppm/year when using data at 2500m after 2005, and 1.77 ± 0.08 ppm/year when using data from the whole period. As for the marine reference, the slope is 1.81 ± 0.03 ppm/year when using data after 2005, and is 2.00 ± 0.02 ppm/year when using data from the whole period. These indicate that for recent years, the increase rate of CO_2 at the Bialystok site is larger than the marine reference. I have updated Fig. 11, in which the trends are computed using data after 2005 only.

C12. Section 4.2: What about the winter months? Is there a difference between ascending and descending profiles in winter? If not, would this indicate that the special difference for the peak growing season is rather caused by the sinks than by the sources?
R12: This is a good point. However, we do not have enough in situ data to perform this analysis. Due to the reduced flight frequency after 2009, very few in situ profiles were made in winter months.

Responses to Anonymous Referee #2

This paper addresses instrument calibrations, inter-comparison and the field measurements approach, although classical and practiced for a while by the community, could provide valuable guidance to the community. Some of the points noted in the paper are common knowledge to the measurements folks and will not necessarily add value, but they may be useful to the newcomers application oriented scientists. Furthermore it should be made very clear that the measurement technologies have moved forward significantly with more stable and robust laser based cavity ring-down instruments, that do not need as frequent calibrations.

C1: The motivation is particularly broad and should be more concise. For example TCCON FTS validation requires high altitude flights up-to 12+km to sample the column. Does the rental aircraft go that high? Was it used to validate or fly over the Bialystok FTS? If so then mention if not there are other platforms and campaigns that have already done this and please cite them..Wunch et al ACP and Phil Trans.. Similarly discussions on emissions verification should site recent papers or reports (e.g.National Academies US study 2010)
R1: Our profiles presented in this paper are only up to 3 km; however, combined with model results, these profiles were used to compare with FTS CO_2 retrievals in Messerschmidt et al, ACPD, 2011. The temporal coverage of these profiles made them especially useful to study the seasonal cycle of column averages. Wunch et al 2011 was already cited, and we have added a reference to Wunch et al. 2010.

C2: I did not see detailed discussion of flask sample species analysis besides CO_2, that was compared to in situ data. I would not have this in the abstract.
R2: We have changed the sentence to "Besides the in situ measurements, air samples have been collected in glass flasks and analyzed in the laboratory for CO_2 and other trace gases.

C3: The way it is written the paper almost points to the obvious, and this could be said in a more concise and clear manner.
R3: We have tried to clearly describe the experiments and the instrument characterization to allow readers to reproduce our results. We regard this as specifically important in the case where the instrument is less ideal as is the case here. We feel that significantly shortening these sections will be cutting corners.

C4: I am very concerned about the water contamination and other air contamination issues and the statistics of this should be declared in a transparent manner, at least in an appendix
R4: We have added more discussions about the water contamination statistics, and the criteria for identifying other air contamination issues are added, "The flasks are flagged as contaminated

when abnormally low values of $\delta^{13}C$ measurements ($\delta^{13}C < -10$ per mil on the VPDB scale) and abnormally high values of CO (CO > 500 ppb), and H_2 (H_2 > 600 ppb) are observed."

C5: Was any Allen-Variance analysis performed on your system, as deployed or in the lab. This would be extremely valuable to determine zero/cal frequencies. It would also be useful for others to compare various instruments that offer more stable performance....as an example please see and cite in your special issue I think Field inter-comparison of two high-accuracy fast-response spectroscopic sensors of carbon dioxide B. A. Flowers, H. H. Powers, M. K. Dubey, and N. G. McDowell Atmos. Meas.Tech. Discuss., 4, 5837-5855, 2011 Please comment on alternative technologies compared to your IRGA system, in light of problems with water contamination.
R5: Allan Variance was performed on the measurements made in the laboratory, but was not shown in the paper. The reason is that the raw signals of the analyzers drift fast (~0.3 ppm/min) and are approximately linear in short periods so that the Allan variance analysis tends to give unrealistically high frequency. The frequency of every two minutes was empirically selected, and the Allan variance of detrended signals was used to verify the targeted accuracy. These are made clear in the revised version. We agree with the reviewer about the importance of instrument comparison from our experience of comparing in situ and flask measurements presented in this paper. We have modified and added in the discussion "It is worth pointing out that CO_2 measurements using the state-of-the-art techniques (O'Keefe, 1998; Bowling et al., 2003; Crosson, 2008; McManus et al., 2008) do not require calibrations as frequently as the NDIR analyzer does. Specifically, the recently available cavity ring-down spectroscopy technique (Chen, Winderlich et al. 2010) has been proven to be sufficiently stable aboard a research aircraft within a field campaign period. However, even with a stable analyzer system, an in-flight calibration system is still recommended when no other independent measurements are available or if the analyzer needs to be deployed over the long term. ", and "The comparison of in situ with flask CO_2 measurements during flight has been successfully employed to identify water contamination issues during two periods. Since CO_2 needs to be reported as dry mole fraction, water contamination is an issue for any technology that detects CO_2 in dry air, and relies on a drying system to remove water vapor from sample air to a sufficiently low level. It has been successful for the cavity ring-down spectroscopy (CRDS) technique to use simultaneously measured water vapor to correct all water vapor effects for CO_2 (Chen et al., 2010; Winderlich et al., 2010). However, this has not been achieved or reported by using other technologies."

C6: Was planetary boundary layer height proxies measured during the in situ airborne analysis e.g. RH as discussed earlier. A more rigorous analysis would be needed to draw meaningful conclusions. Same is true for the long-term time series analysis that is very idealized and simple. The statistics should be presented clearly and the figures may not reveal clear trends without the seasonal fitted profiles. I can clearly see high bias points in the 300m time series. Are these due urban pollution since you are only 20-30km from a city. Please make this clear. Also are there data on CO that can be used to explain these high biases.
R6: Yes, there are RH measurements. We actually used virtual potential temperature profiles to determine the mixed layer height, and have added it in the revised version. We have added explanations of the uncertainties in these statistical analyses. We checked the CO data, and found out that the few high bias points in the 300m time series are associated with high CO values, suggesting those are influenced by the nearby city. Therefore, we have already excluded the

significance of the trend difference between 300m and the marine reference, due to the large scatter in these data.

C7: I know you all have done a very rigorous job on the flask vs in situ comparisons. How does the real variability you see effect this, can you use your data to put some limits on when we should compare with flask and when not or how the methodology degrades with real atmospheric variability as measured by aircraft. This would be useful.
R7: The atmospheric variability can be accounted for by the weighting coefficients that incorporate the contribution of air during the sampling process to the final value. Therefore, one should compare all flasks with in situ data regardless of atmospheric variability. However, it is critical to have the exact time when the pressurizing process starts, and the flask pressure or the flow rate during the flask sampling process. When these parameters are not available, the comparison is certainly sensitive to the atmospheric variability. We added the discussions in the revised version.

C8: You have done a good job with the flask vs in situ comparison and the paper will be much improved if other sections also have a similar rigor. The contrast is very distracting to me.
R8: We feel that the modifications included in the revised manuscript as suggested by both reviewers have improved this.