Reply to RC 3

Manuscript information:

- Title: Non-target analysis using a gas chromatography with time-of-flight mass spectrometry: application to time series of fourth generation synthetic halocarbons at Taunus Observatory (Germany).
- Authors: Fides Lefrancois, Markus Jesswein, Markus Thoma, Andreas Engel, Kieran Stanley, and Tanja Schuck
- MS No.: amt-2020-488
- MS type: Research article
- Iteration: Final response (AMT Discussions)

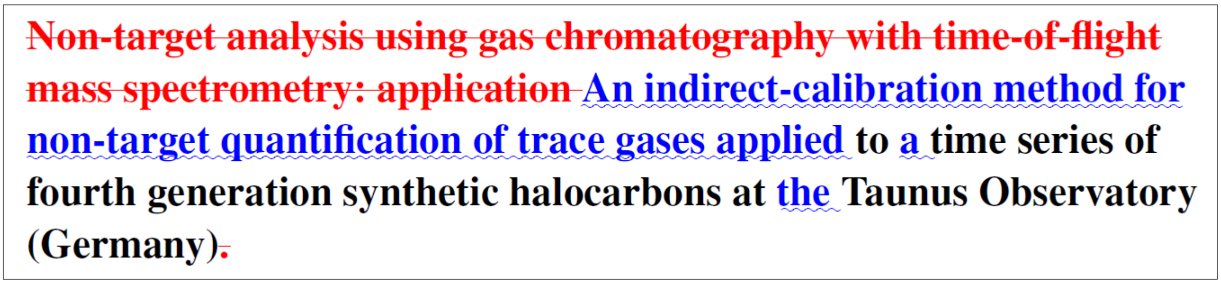
We would like to thank the reviewer for the constructive and detailed comments.

Point-by-Point reply:

1. *I would suggest to use a title that contains the main purpose of the article, which is to make non-target screening quantitative. Usually, articles with "non-target screening" in their title contain discovery of new substances, which is not the scope here. Suggestions for title: "Quantitative non-target ...."*

We changed the title as follows:

“An indirect-calibration method for non-target quantification applied to time series of fourth generation synthetic halocarbons at Taunus Observatory (Germany) using gas chromatography mass spectrometry measurements.”



~~Non-target analysis using gas chromatography with time-of-flight mass spectrometry: application~~ An indirect-calibration method for non-target quantification of trace gases applied to a time series of fourth generation synthetic halocarbons at the Taunus Observatory (Germany)~~.~~

2. *l. 8: I would simplify the grammar: "This archive can be used if", or even "This archive can be used for retrospective screening"*

Done, as suggested in the first sentence.

3. *l. 11: "or the amounts in the calibration gas may not have been quantified."*

Done.

4. *Introduction, l. 18-20: "The application of the indirect calibration method on several test cases can result into accuracies around 13% to 20 %. For H(C)FOs accuracies up to 25% are 20 achieved." I would be good to reformulate these sentences to really convey the meaning that low values represent a better accuracy. Maybe you can replace the word "accuracy" by "uncertainty" here: "The application of the indirect calibration method on several test cases can result into uncertainties around 13% to 20 %. For H(C)FOs, of particularly low mole fraction values, uncertainties up to 25% are observed.".*

Changed as suggested by the reviewer.

5. *l. 26 "which are part of or affiliated to" l.31, maybe: "is not well covered"*

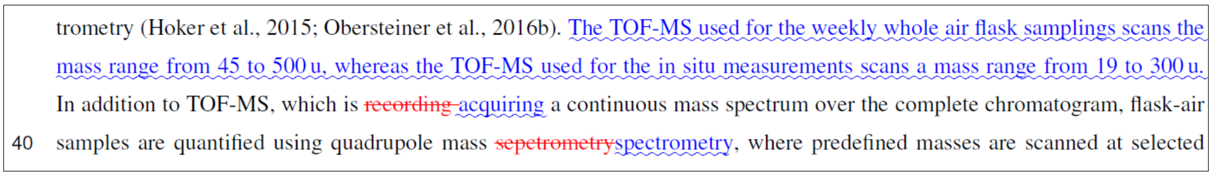
Done.

6. *l. 33: "by the installation"*

Done.

7. *l. 35: maybe you want to specify the mass range coverage of the TOF instrument (minimum and maximum measured masses).*

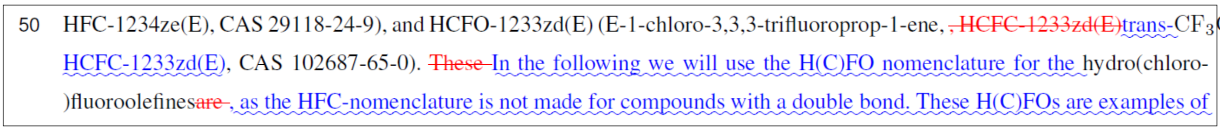
The mass range coverage was added.



trometry (Hoker et al., 2015; Obersteiner et al., 2016b). The TOF-MS used for the weekly whole air flask samplings scans the mass range from 45 to 500 u, whereas the TOF-MS used for the in situ measurements scans a mass range from 19 to 300 u. In addition to TOF-MS, which is ~~recording~~ acquiring a continuous mass spectrum over the complete chromatogram, flask-air samples are quantified using quadrupole mass ~~sepetrometry~~spectrometry, where predefined masses are scanned at selected

8. *l. 45-47: may you can consider if you would like to leave out "HFC-1234yf" and "HFC-1234ze(E)". The HFC nomenclature is actually not made for compounds with a double bond.*

Because HFC-1234yf and HFC-1234ze(E) have been widely used in the scientific literature, we leave them here among the list of possible substance names. To make clear that these substances do not fully represent the class of HFOs we modified the following sentence:



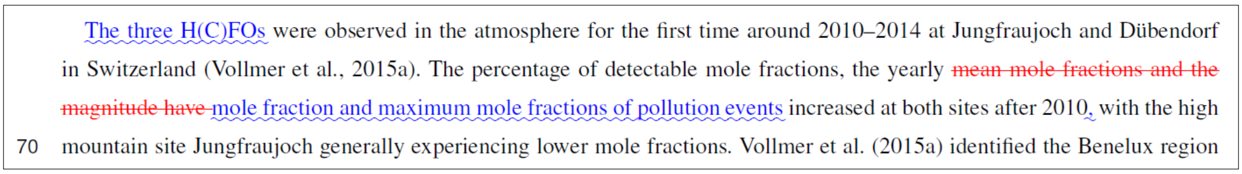
HFC-1234ze(E), CAS 29118-24-9), and HCFO-1233zd(E) (E-1-chloro-3,3,3-trifluoroprop-1-ene, ~~, HCFC-1233zd(E)~~trans-CF$_3$ HCFC-1233zd(E), CAS 102687-65-0). ~~These~~ In the following we will use the H(C)FO nomenclature for the hydro(chloro-)fluoroolefines~~are~~ , as the HFC-nomenclature is not made for compounds with a double bond. These H(C)FOs are examples of

9. *l.51 : "have no ODP": "have an ODP value of zero." "have no ODP" suggests that the computation of the ODP is impossible.*

Changed as suggested.

10. *l. 57: the magnitude of what? Amplitude of annual cycles, magnitude of mole fraction of pollution events?*

Changed wording to:



The three H(C)FOs were observed in the atmosphere for the first time around 2010–2014 at Jungfraujoch and Dübendorf in Switzerland (Vollmer et al., 2015a). The percentage of detectable mole fractions, the yearly ~~mean mole fractions and the magnitude have~~ mole fraction and maximum mole fractions of pollution events increased at both sites after 2010, with the high mountain site Jungfraujoch generally experiencing lower mole fractions. Vollmer et al. (2015a) identified the Benelux region

11. *l. 80: "and each pair of measurements is bracketed"*

Done.

12. *l. 81: "range of parts per trillion (ppt)": range of picomole per mole, pmol/mol or hereafter part per trillion (ppt)"*

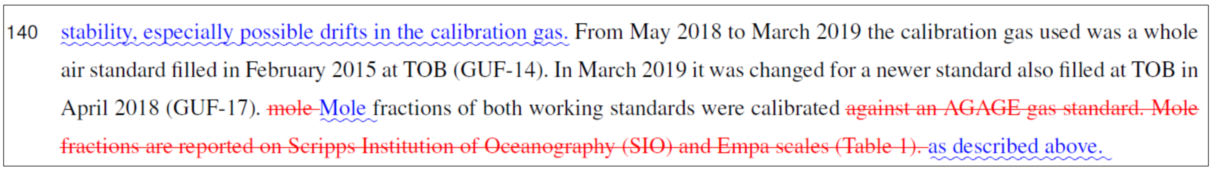
Done.

13. *l. 102: add comma: "For each measurement, approximately"*

Done.

14. *l. 117: you can leave out the sentence about calibration scales, it is already mentioned l. 92- 94.*

Changed to:



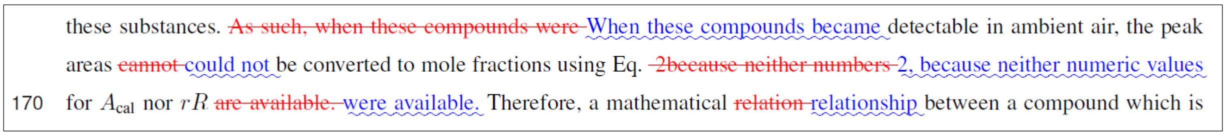
stability, especially possible drifts in the calibration gas. From May 2018 to March 2019 the calibration gas used was a whole air standard filled in February 2015 at TOB (GUF-14). In March 2019 it was changed for a newer standard also filled at TOB in April 2018 (GUF-17). ~~mole~~ Mole fractions of both working standards were calibrated ~~against an AGAGE gas standard. Mole fractions are reported on Scripps Institution of Oceanography (SIO) and Empa scales (Table 1).~~ as described above.

15. *l. 139 : "before calibration standards containing measurable amounts of these substances were used".*

Done.

16. *l. 140: tense concordance, not sure, check with native speaker. "When these compounds were detectable in ambient air, the peak areas could not be converted to mole fractions using Eq. 2 because neither numeric values for Acal nor rR were available."*

Changed to:



these substances. ~~As such, when these compounds were~~ When these compounds became detectable in ambient air, the peak areas ~~cannot~~ could not be converted to mole fractions using Eq. ~~2because neither numbers~~ 2, because neither numeric values for A_{cal} nor rR ~~are available.~~ were available. Therefore, a mathematical ~~relation~~ relationship between a compound which is

17. *l. 141: you surely mean: "between another compound which is measurable in the standard"*

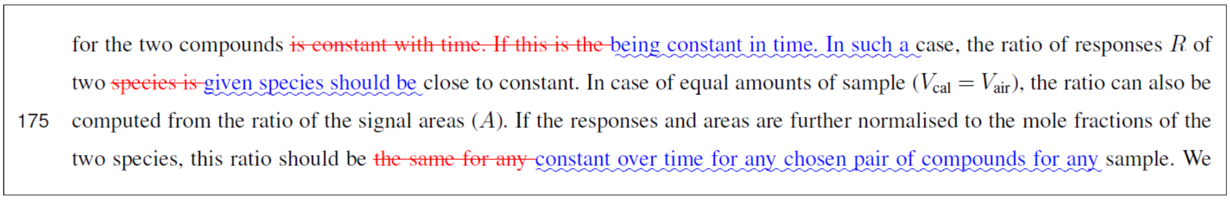
The reviewer is correct. Changed as suggested.

18. *l. 144: "that means that the ratio of signal per amount of analyte for the two compounds is constant with time." I'm not sure about the meaning of this sentence. We know that the response of a MS instrument may vary strongly over time, for example the instrument response increases after source cleaning. However what is important here is that the instrument response behaviour should vary similarly over time for all substances, as you clearly write afterwards. I would rephrase as: "Ideally, the sensitivity of the analytical system for two different species should behave similarly over time. In such a case, the ratio of responses R of two given species should be close to constant."*

Changed that as suggested.

19. *l. 146: "this ratio should be the same for any sample.": maybe too general. Suggestion to write more specifically: "this ratio should be constant over time for any chosen pair of compounds".*

Changed as follows:



for the two compounds ~~is constant with time. If this is the~~ being constant in time. In such a case, the ratio of responses R of two ~~species is~~ given species should be close to constant. In case of equal amounts of sample ($V_{\mathrm{cal}} = V_{\mathrm{air}}$), the ratio can also be computed from the ratio of the signal areas (A). If the responses and areas are further normalised to the mole fractions of the two species, this ratio should be ~~the same for any~~ constant over time for any chosen pair of compounds for any sample. We

20. *l. 155: "It must be stable over time". Check entire manuscript.*

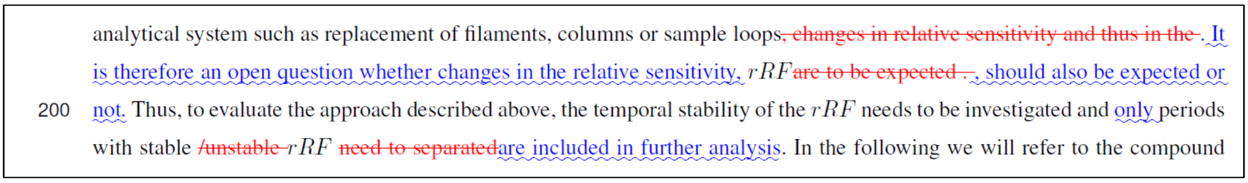
Changed as suggested.

21. *l. 164-166: meaning not clear. A non-stable sensitivity does not necessarily imply a non-stable relative sensitivity, this is something you are going to investigate next. Suggestion to rephrase: "The methodology outlined in 3.1 is based on the assumption of a constant rRF in Eq. 4. In reality, the absolute sensitivity of a mass spectrometer is known to vary over time, in particular after tuning the mass spectrometer or after modifications of the analytical system such as replacement of filaments, columns or sample loops. It is therefore an open question whether changes in the relative sensitivity rRF should also be expected or not. Thus, to evaluate [...]"*

Changed as suggested.

22. *l. 169: "need to separated": "need to be evaluated separately".*

Periods of stable/unstable rRF are not exactly evaluated separately. In fact, periods with unstable rRF are excluded from further analysis. We reworded the sentence correspondingly:



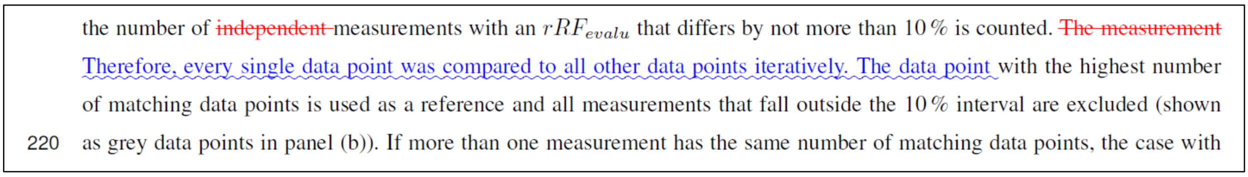
analytical system such as replacement of filaments, columns or sample loops~~, changes in relative sensitivity and thus in the .~~. It is therefore an open question whether changes in the relative sensitivity, rRF ~~are to be expected .~~, should also be expected or not. Thus, to evaluate the approach described above, the temporal stability of the rRF needs to be investigated and only periods with stable ~~/unstable~~ rRF ~~need to separated~~are included in further analysis. In the following we will refer to the compound

23. *l. 181-185: difficult to understand, suggestion to rephrase: "To identify periods of stable rRFevalu for a given pair of compounds, timeseries of rRFevalu are reviewed. To do so, for each measurement or data point of rRFevalu in the timeseries, we compute the sum of other rRFevalu data points that do not deviate from the chosen data point by more than 10%. The data point with the highest number of matching data points is used as a reference (shown with red cicle in Figure 1, panel (b)) and all data points that fall outside the 10% interval are excluded (shown as grey data points in panel (b)).*

Note: I would not use "independant measurement", since the measuring instrument is the same of course the results are not fully statistically independent, and we actually need the results not to be independent for this method to work.

To make it more clear, on Fig. 1 please mark with e.g. a red circle the data point that was selected as most likely rRF value.

We agree that the measurements are not fully independent and modify the wording accordingly. With regards to Figure 6, we fear that the suggested marking of a single data point would be confusing. All data points are compared step by step to all others and this iterative procedure would not become clear by indicating the selected data points. To make the procedure more clear we added the following statement in the iterative comparison of all data points:



the number of ~~independent~~ measurements with an rRF_{evalu} that differs by not more than 10 % is counted. ~~The measurement~~ Therefore, every single data point was compared to all other data points iteratively. The data point with the highest number of matching data points is used as a reference and all measurements that fall outside the 10 % interval are excluded (shown as grey data points in panel (b)). If more than one measurement has the same number of matching data points, the case with

24. *Table 1: add bibliographic reference to all scales where needed.*

METAS-2017: Guillevic et al., 2018 (ok, already done).

EMPA-2013: for HCFC-133a: Vollmer, M. K., Rigby, M., Laube, J. C., Henne, S., Rhee, T. S., Gooch, L. J., Wenger, A., Young, D., Steele, L. P., Langenfelds, R. L., et al. (2015), Abrupt reversal in emissions and atmospheric abundance of HCFC-133a (CF3CH2Cl),

Geophys. Res. Lett., 42, 8702– 8710, doi:10.1002/2015GL065846.

EMPA-2013 for HFOs: Vollmer et al., Environ. Sci. Technol. 2015, 49, 5, 2703–2708.

SIO-05, SIO-14: Prinn et al., J. Geophys. Res., 105, 17,751-17,792, 2000, and Prinn et al, Earth Syst. Sci. Data, 10, 985–1018, 2018.

The references were added to all scales in table 1.

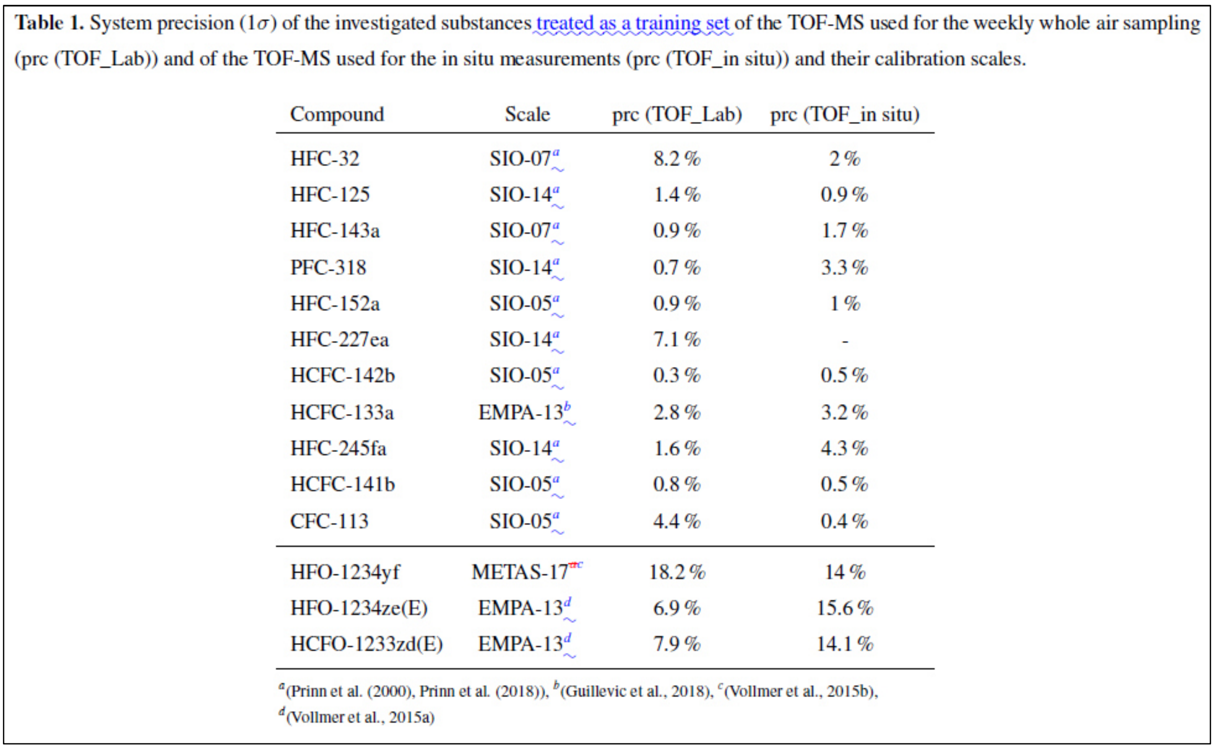


Table 1. System precision (1σ) of the investigated substances treated as a training set of the TOF-MS used for the weekly whole air sampling (prc (TOF_Lab)) and of the TOF-MS used for the in situ measurements (prc (TOF_in situ)) and their calibration scales.

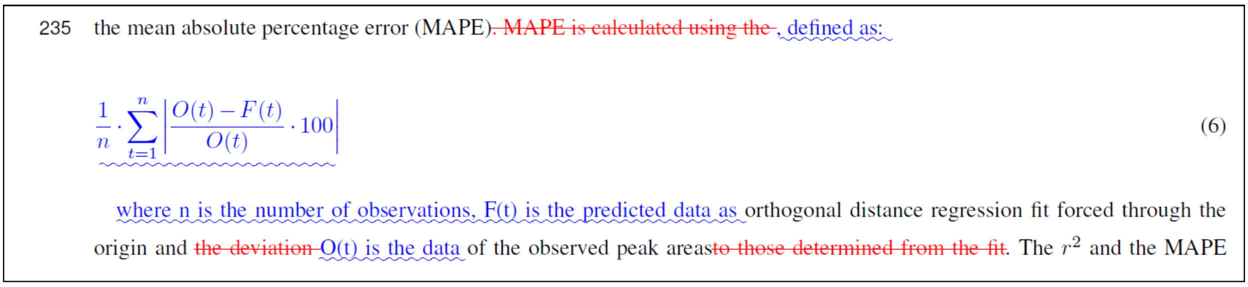
Compound	Scale	prc (TOF_Lab)	prc (TOF_in situ)
HFC-32	SIO-07[a]	8.2 %	2 %
HFC-125	SIO-14[a]	1.4 %	0.9 %
HFC-143a	SIO-07[a]	0.9 %	1.7 %
PFC-318	SIO-14[a]	0.7 %	3.3 %
HFC-152a	SIO-05[a]	0.9 %	1 %
HFC-227ea	SIO-14[a]	7.1 %	-
HCFC-142b	SIO-05[a]	0.3 %	0.5 %
HCFC-133a	EMPA-13[b]	2.8 %	3.2 %
HFC-245fa	SIO-14[a]	1.6 %	4.3 %
HCFC-141b	SIO-05[a]	0.8 %	0.5 %
CFC-113	SIO-05[a]	4.4 %	0.4 %
HFO-1234yf	METAS-17~~[a]~~[c]	18.2 %	14 %
HFO-1234ze(E)	EMPA-13[d]	6.9 %	15.6 %
HCFO-1233zd(E)	EMPA-13[d]	7.9 %	14.1 %

[a](Prinn et al. (2000), Prinn et al. (2018)), [b](Guillevic et al., 2018), [c](Vollmer et al., 2015b), [d](Vollmer et al., 2015a)

25. *l. 195: you probably need "the" in front of all "MAPE", check through the manuscript. I would add the equation for the computation or a reference (e.g. the Wiki page).*

Changed that and added article where necessary.

We added the equation for the calculation of the MAPE as suggested.



the mean absolute percentage error (MAPE)~~. MAPE is calculated using the~~, defined as:

$$\frac{1}{n}\cdot\sum_{t=1}^{n}\left|\frac{O(t)-F(t)}{O(t)}\cdot 100\right| \qquad (6)$$

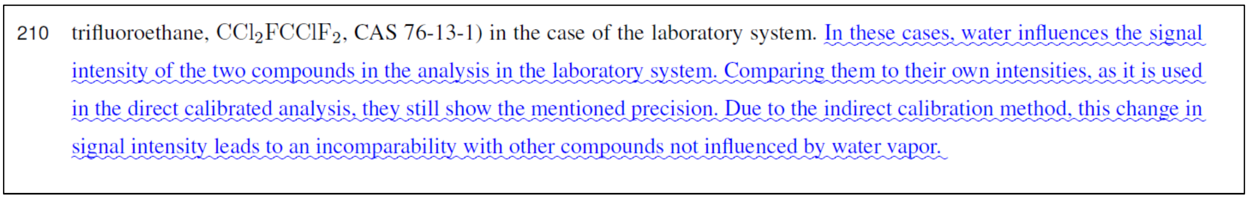
where n is the number of observations, F(t) is the predicted data as orthogonal distance regression fit forced through the origin and ~~the deviation~~ O(t) is the data of the observed peak areas~~to those determined from the fit~~. The r^2 and the MAPE

26. *l. 199: "Except for HFC-227ea"*

Done.

27. *Section 3.2.1, general question: could you find explanations for the outlier rRFevalu data points?*

One possible explanation is the influence of water, which is the case for HCFC-141b using data of the whole air flask sampling. We commented on that in the revised version of manuscript. Other explanations do need further investigations.



trifluoroethane, CCl_2FCClF_2, CAS 76-13-1) in the case of the laboratory system. In these cases, water influences the signal intensity of the two compounds in the analysis in the laboratory system. Comparing them to their own intensities, as it is used in the direct calibrated analysis, they still show the mentioned precision. Due to the indirect calibration method, this change in signal intensity leads to an incomparability with other compounds not influenced by water vapor.

28. *l. 200-201, I would try to reformulate in an easier way. E.g.: "To test which pairs of substances produce the highest correlations, all possible pairs of substances have been tested. The obtained values for r2 and MAPE are shown in Fig 3".*

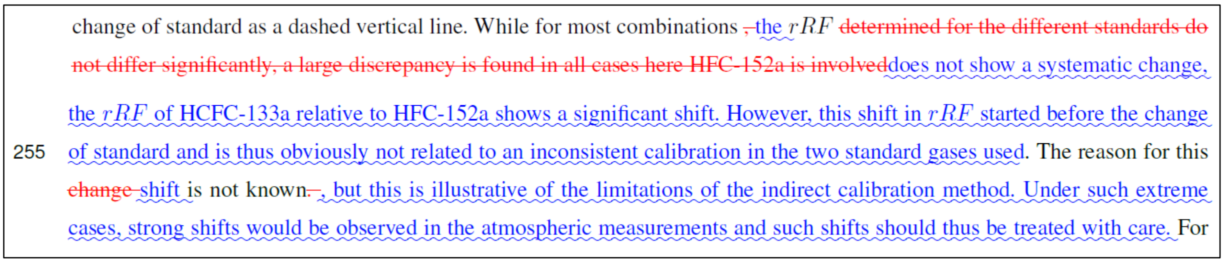
Changed as suggested.

29. *l. 208, typo: "all cases where HFC-152a is involved."*

Done.

30. *Also, it seems to me more to be a drift in the rRF value, that started before the change in standard tank, and stabilised after some runs of the new standard. Such a drift (albeit much smaller) can also be seen in the HFC-125 data points. So I'm really not sure that you can link this for the standard tank change. I would remove this sentence.*

The reviewer is correct. The change in rRF occurs before the change of standard. We have no explanation, why these drifts occur. We have made this clearer by changing the wording as follows:



change of standard as a dashed vertical line. While for most combinations ~~,~~ the rRF ~~determined for the different standards do not differ significantly, a large discrepancy is found in all cases here HFC-152a is involved~~does not show a systematic change, the rRF of HCFC-133a relative to HFC-152a shows a significant shift. However, this shift in rRF started before the change of standard and is thus obviously not related to an inconsistent calibration in the two standard gases used. The reason for this ~~change~~ shift is not known~~.~~, but this is illustrative of the limitations of the indirect calibration method. Under such extreme cases, strong shifts would be observed in the atmospheric measurements and such shifts should thus be treated with care. For

31. *l. 211: maybe you can comment on why HFC-227ea and HFC-245fa? HFC-227ea seems logical to be a bad one, as its measurement standard deviation given in Table 1 is one the highest. However why HFC-245fa? Or, alternatively, you can explain later why some are good ones?*

We chose to discuss HFC-227ea and HFC-245fa as substances less suited for the approach to demonstrate what the possible effects are. We agree that the poorer performance of HFC-227ea seemed likely given the poorer measurement precision, while this was somewhat unexpected in the case of HFC-245fa. However, based on the training substances included, we currently cannot explain why some pairs of substances are "good ones". The example thus shows that measurement precision is not a sufficient selection criterion which is the reason why we performed the additional statistical analyses.

32. *l. 220-222: "To quantify the differences between the selection of data of main reference and test substance via main reference substance and an evaluation substance we compared the relative standard deviations of the resulting filtered data sets." I don't understand this sentence. Please clarify. You may also want to cut into smaller sentences. Maybe, adding the equation you use will help to understand what you compute here.*

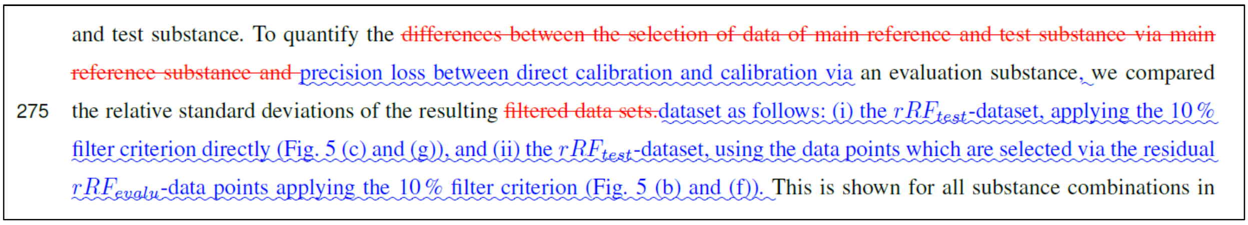
Usually there are two quantitative values to characterise a result: its standard deviation, which reflect the random noise, and the average difference between two values (usually a test value and a reference value), which is a systematic bias. A bias not equal to zero means that the method causes a systematic error.

Now based on Fig 5, maybe what you want to express here is a precision loss, that you express via the difference in standard deviation? If this is really the case, here is my suggestion:

"To quantify the precision loss between direct calibration and calibration via a transfer substance, we compare the relative standard deviations of the resulting filtered data sets.", or something similar.

The reviewer is right that we want to express is the loss in precision and we reworded the statement as suggested.

We rephrased the sentence as follows:



and test substance. To quantify the ~~differences between the selection of data of main reference and test substance via main reference substance and~~ precision loss between direct calibration and calibration via an evaluation substance, we compared the relative standard deviations of the resulting ~~filtered data sets.~~dataset as follows: (i) the rRF_{test}-dataset, applying the 10 % filter criterion directly (Fig. 5 (c) and (g)), and (ii) the rRF_{test}-dataset, using the data points which are selected via the residual rRF_{evalu}-data points applying the 10 % filter criterion (Fig. 5 (b) and (f)). This is shown for all substance combinations in

33. *Another important quantity to evaluate is if your method creates a bias or not? i.e. what is the average value of the distance (or difference) between the true and reconstructed value? It should be (close to) zero to show no bias. (cf see below comment on Table 3)*

See point 37.

34. *l. 237: if you mean precision loss, use: "the difference between the standard deviations".*

Done.

35. *l. 241: "As test cases to apply the indirect calibration method, we chose..." or "As test cases to be applied the indirect calibration method, ...".*

Done.

36. *l. 243: "mole fractions of HFC-227ea show..."*

Done.

37. *Table 3: average relative difference: this is your metric for the bias, right? Please write the equation somewhere in the text (e.g. around l. 245). Also: usually, if the bias or systematic offset value is within the 2 sigma standard deviation, it means within uncertainty, there is no bias. This is an important point to show here. But in Table 3, the "av. rel. difference" value is systematically more than the value of "standard deviation". Can you comment on this?*

Addition: We noticed an error. Due to new calculations, we decide to choose here the MAPE, as well. Even it is the calculation of the average relative difference we mentioned here before (sum($|(X_{\text{direkt}} - X_{\text{indirect}})/X_{\text{direct}}|$) * 100). Regarding to equation 6, O(t) is the direct calculated data and F(t) is the indirect calculated data.. The standard deviations shall explain, how much the relative deviation of each data point spread. So it is possible that they exhibit a larger amount as the MAPE (if the relative deviations show a high variability), or they show a small amount if it is a constant systematically error. Also the standard deviation can increase over too large chosen time periods, where the rRF exhibits maybe long term drifts.

38. *l. 275: typo: "HFO-1234yf"*

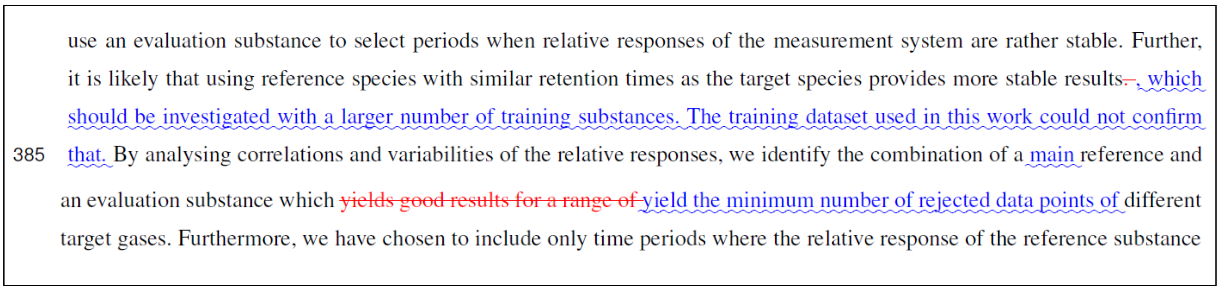
Done.

39. *l. 276: concordance of tenses, "increased continuously up to 100%"*

Done.

40. l. 311-312: "Further, it is likely that using reference species with similar retention times as the target species provides more stable results." Can you give an example here? No retention time data are provided.

While we tried to select reference species with similar retention times as the target species, we did not find any evidence that the quality of the indirect calibration is affected by the differences in retention times. We added a comment on that:



use an evaluation substance to select periods when relative responses of the measurement system are rather stable. Further, it is likely that using reference species with similar retention times as the target species provides more stable results~~.~~, which should be investigated with a larger number of training substances. The training dataset used in this work could not confirm that. By analysing correlations and variabilities of the relative responses, we identify the combination of a main reference and an evaluation substance which ~~yields good results for a range of~~ yield the minimum number of rejected data points of different target gases. Furthermore, we have chosen to include only time periods where the relative response of the reference substance

41. *l. 313: "good results" is subjective. Maybe use a quantitative value instead, e.g. "which yield the minimum number of rejected data points".*

Changed as suggested.

42. *l. 330, typo: "is the measurement", "which are expected".*

Done.

43. *Your data show a rRF that is mostly not stable over time. Can you discuss the possibility to use a running-mean rRF value over time, instead of assuming a constant value over a short time period? Also, at least for some time periods, could you assign a (hardware?) cause to the non-stable rRF?*

Using a running mean rRF over time is no option, due to systematical changes or drifts on the system. This would allow no qualification. We assume that non-stable rRFs are mostly caused by hardware issues. So this would be a good initiation to compare the method on other systems, which do maybe preserve other hardware constitutions.

Hardware issues can be assumed, when receiving a large standard deviation for the single relative deviations. A small standard deviation will mean that the system shows a systematically error.

44. *Figure 5, legend: "Illustration of data selection for the weekly flask sampling measurements..."*

Done.