Answers to anonymous referee 1

This is a difficult amount of information to include in a single publication. As written, the details overpower to overall message of the article. Details regarding the preparation of in-house standards (preparation methods, overall mixture stability, surface treatment) could be moved to an appendix section to allow readers to focus on the direct measurement and value assignment.

The comment is appreciated. This manuscript was outlined as a methodological paper because of the relevance of the topic for the atmospheric monitoring community. Within the main text, only relevant information needed to reproduce the methodology used to generate the traceable working standards was kept. The highly detailed description was laid out as a kind of guideline for end-users. The rest of the information was moved to one of the appendix sections already included in the manuscript. We believe that the balance between the main text and appendixes (four) should not be changed. However, following this comment and similar comments provided by the other anonymous referees, we implemented the recommendations to improve the readability of the manuscript by adding schemes illustrating the processes involved in the preparation and assessment of the SI- traceable working standards (Figures 1–3).

Authors were very careful to warn readers of the broad distribution of value assignment and instability in these mixtures. This does raise questions about the application of these results within participating institutions, discussions of next steps or guidance in analyses of these compounds would be constructive.

We thank the referee for the suggestion. We added further discussions of the next steps and guidance in OVOC analyses to show how and to what degree the approaches described in the manuscript can be applied at atmospheric monitoring station measuring OVOCs (Lines 672–678). Even the "negative" results of this work (i.e. whole air working standards in canisters) may contribute to defining the next steps needed to overcome the challenges associated to these compounds (e.g., low amount fractions, reactivity, large uncertainties, surface effects, artefacts).

An additional question, did those with large offsets identify the cause and perform follow up analyses to demonstrate capabilities?

In the case of EMPA, the offset was due to a combination of different issues that included a leak in the heated valved of their GC-FID system and an initial flow overshoot when measuring with their Stirling cooling unit (used to cool the samples to -43 °C to dry them). These issues were overcome by replacing the faulty heated valve and by using an alternative method to dry the samples (Kori-xr thermal desorption (Markes International, UK) instead of the Stirling cooling unit). The error associated to these issues was estimated in 30 % (an error that was included in the uncertainty of the measurements). Concerning the large offsets of UU for methanol and MEK, these were due to the extremely low amount fractions at which the working standards were generated (far low from the target values for this work (10 nmol/mol; values closer to the atmospheric ones for the selected OVOCs) and not to the analytical system. Both participants demonstrated their capabilities when performing follow-up measurements after solving the technical issues (EMPA) or at target amount fractions (UU), as well as the data quality evaluation of their observations following ACTRIS network requirements.

Questions to authors:

Line 195- 200: "recovery due to initial loss", what is recovery? Was there an initial decrease and then increase on observed amount fraction for particular analytes? Can that be attributed to instrument performance as described later on or is this based on wall effects for the analyte of interest?

Lines 210–217: "Recovery due to initial loss" is indeed somewhat confusing. To avoid confusion, we eliminated the term "recovery". The "initial loss" is the amount fraction lost immediately after preparation of the gas mixture and it is most likely due to adsorption/interaction with the cylinder surface material (wall effects), rather than to the instrument performance.

Lines 235-240: Cylinder wall passivation is a significant challenge. Further detail regarding the testing of the passivation approach, with water in particular, would be helpful. While this section described the amount of water injected into the cylinder there is no quantification or stability of water vapor described. Additionally, there is potential for offsets or bias of analytes like MVK in the presence of water.

We agree with this comment: cylinder wall passivation is a challenging and broad research topic. Because of the manuscript length and because wall passivation was not the main focus of this work, we did not include further detail regarding cylinder wall passivation using water. We acknowledge that a separate study on the effect of matrix gas composition and passivation on whole air working standards is needed to estimate how water and oxygen contribute to the integrity/stability of the OVOC amount fractions in vessels (i.e., cylinders and canisters).

General question: How will this publication impact the CMC's for participating institutions? In cases like 31% coverage factor, this would not be useful within the user community.

Most of the participating institutions in this work are universities and research institutes out of the metrological community. The internationally recognized Calibration and Measurement Capabilities (CMC's) of the CIPM Mutual Recognition Arrangement (CIPM MRA) do not apply to them and therefore this publication will not have any impact on them. In the case of the participating NMIs, this publication will not have impact on their CMCs as that was not the objective of this study.

All the uncertainties included in this work have a coverage factor of two (95.4 %). Thus, we are unsure if the second part of the referee's comment is referring to the coverage interval or to the relative expanded uncertainty. If the latter, we would like to stress the fact that through this study it is acknowledged that measurements of the selected OVOCs at atmospheric levels were less accurate than the target expectations. For this work, the best materials and procedures were selected. However, the big uncertainties reflect how challenging is still the measurement of the selected OVOCs at monitoring stations because of their reactivity and/or low amount of substance fractions. Further research is clearly needed to understand the effects of cylinder materials, matrix gas, surface effects and analytical methods have on the measurements of the selected VOCs, as well as to develop suitable materials for sampling and storage of these compounds and passivation strategies. We are confident that improved assessment protocols, for example, using the same calibration standards among the participants (e.g., the newly developed SI-traceable working standards) and generating the same level of amount of substance fraction around 10 nmol mol^{-1} (e.g., by setting the same dilution flows of nitrogen to dilute the RGMs), would already contribute to decrease the uncertainty of the assessment measurements.

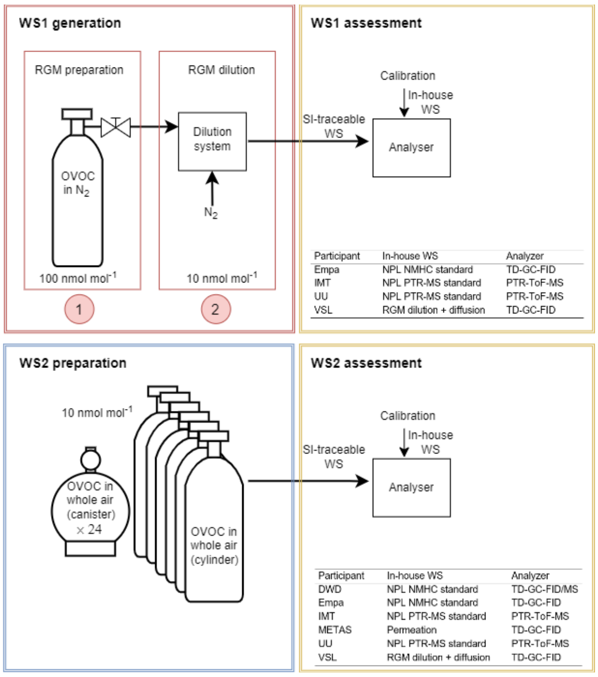
Answers to anonymous referee 2

The manuscript entitled "Towards a high quality in-situ observation network for oxygenated volatile organic compounds (OVOCs) in Europe: transferring traceability to the International System of Units (SI) to the field" by Iturrate-Garcia et al. presents the developed and assessed protocols to generate two different types of SI-traceable working standards for acetaldehyde, acetone, methanol, and methyl ethyl ketone. This method enables the calibration of OVOCs at lower concentration levels and ensures the comparability of OVOC measurements both within and across monitoring networks. The manuscript would benefit from clearer articulation, particularly in the explanation of the methodology. Additionally, a more thorough discussion of the method's strengths, limitations, and inherent uncertainties is needed to enhance its scientific rigor. The manuscript may be suitable for publication after major revision.

General comments:

1. The manuscript contains an excessive amount of detailed content, making it difficult for readers to grasp the key points. I recommend improving the readability by incorporating structured steps or flowcharts where applicable. For example, provide a flowchart illustrating the process from standard sample preparation to dilution and measurement.

Lines 132 – 137, Lines 151–153, Lines 237–238: We thank the referee for the suggestion. We added three flowcharts illustrating the different processes involved in the assessment of the two different types of SI-working standards to the manuscript (Figures 1–3).



Participant	In-house WS	Analyzer
Empa	NPL NMHC standard	TD-GC-FID
IMT	NPL PTR-MS standard	PTR-ToF-MS
UU	NPL PTR-MS standard	PTR-ToF-MS
VSL	RGM dilution + diffusion	TD-GC-FID

Participant	In-house WS	Analyzer
DWD	NPL NMHC standard	TD-GC-FID/MS
Empa	NPL NMHC standard	TD-GC-FID
IMT	NPL PTR-MS standard	PTR-ToF-MS
METAS	Permeation	TD-GC-FID
UU	NPL PTR-MS standard	PTR-ToF-MS
VSL	RGM dilution + diffusion	TD-GC-FID

Figure 1: Scheme showing the two types working standards traceable to the international system of units (SI) prepared in this work, based on the dilution of reference gas mixtures (RGM) of oxygenated volatile organic compounds (OVOC) in nitrogen (N_2) (WS1; for details, see section 2.1) and on certified spiked whole air samples (WS2; for details, see section 2.2). Participants in the assessment, analysers (thermal desorption (TD)-gas chromatography (GC)-flame ionization detector (FID) and proton transfer reaction (PTR)-Time of Flight (ToF)-mass spectrometry (MS) systems) and in-house working standards used to calibrate them are indicated.

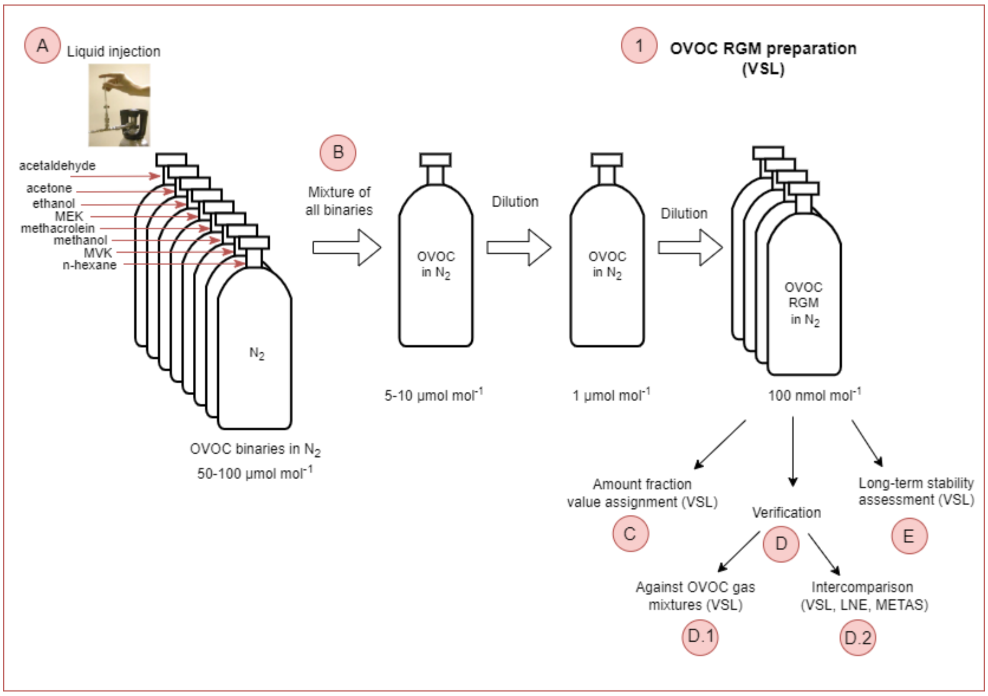


Figure 2: Schematic diagram illustrating the steps needed to prepare the reference gas mixtures (RGMs) of the selected oxygenated volatile organic compounds (OVOCs).

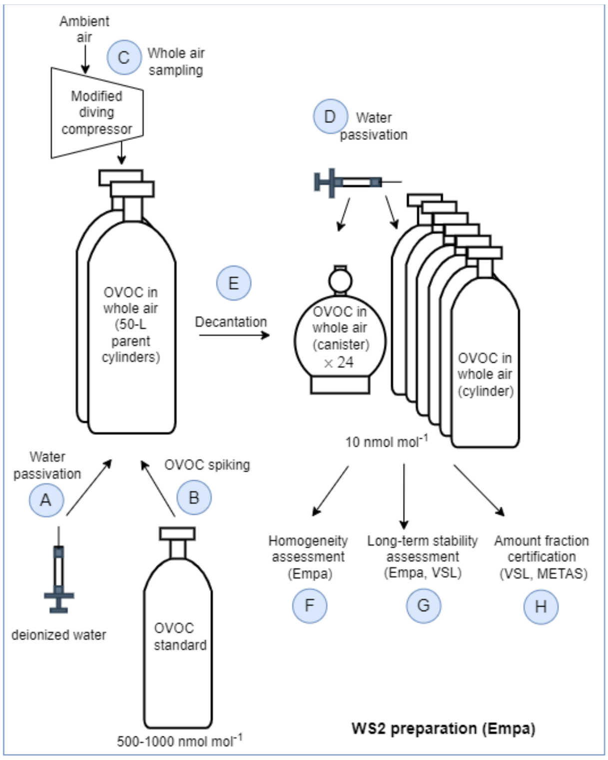


Figure 3: Schematic diagram illustrating the steps needed to prepare the SI-traceable working standard based on certified spiked whole air samples (WS2).

2. Abbreviation Issues:

a) The experiments in the manuscript were conducted across multiple participant laboratories (e.g., VSL, METAS, LNE, IMT, UU, and Empa). The full names of these laboratories should be clearly defined when the abbreviations first appear.

For the National Metrology Institutes, instead of writing the full name the first time the abbreviation appeared, we indicated clearly which national metrology institute (NMI) was the first time the abbreviation appeared in the main text (e.g. NMI of the Netherlands (Line 147), NMI of France (Line 194) and NMI of Switzerland (Line 194)), as this information is more relevant than the definition of the abbreviation. For the other participants, the definition was already included in the main text: Line 243 (Empa), Line 313 (DWD) and Line 314 (UU, IMT). To facilitate the reading of the appendix section, we added this information the first time the abbreviations appeared in each of the appendixes (e.g., Lines 716, 730, 744, 750, 759, 770 and 776). We did the same for other abbreviations, such as RGM and TD-GC-FID, among others.

b) There are inconsistencies in the capitalization of "Empa" throughout the text. Please ensure uniform usage.

Lines 910–915: Corrected (Table C2).

c) Additionally, the description of which laboratories conducted specific experiments using particular methods is somewhat unclear. A brief introduction of this information should be included either at the end of the introduction or at the beginning of Sections 2, 3, and 4 for clarity.

Lines 131–137 (Figure 1), Lines 313–314 and Lines 318–322 (Table 2): Information on which laboratories conducted specific experiments and the methods used is included in the manuscript annexes. However, the information is split into several tables. For clarity, we implemented the referee's suggestion.

3. The authors provided two different types of SI-traceable working standards, accompanied by two different dilution systems and distinct analytical methods. Although the manuscript outlines the method for calculating uncertainty, it lacks specific uncertainty values for each OVOC corresponding to each method. It is necessary to supplement the manuscript with detailed uncertainty values, the advantages and limitations of each method, and recommendations on preferred methods, processes, and considerations. This will enhance the applicability of the methods within the atmospheric measurement community.

We thank the referee for these recommendations. Regarding the detailed uncertainty values, the uncertainties of the SI-traceable working standards were already included in the manuscript (Table 3 and Table D1). We added the uncertainty of the distinct analytical methods to Table D2 (Lines 1082–1085) and Table D3 (Lines 1087–1091) in Appendix D. It is acknowledged that the comment about advantages and limitations could be better formulated in the conclusions. It must also be noted that the use of different analytical techniques (as it is normal practice in the climate community) is also playing a big role on the results of this study and on how the working standards should be used. Concerning the recommendations on preferred methods, with the current results, is difficult to evaluate which of the methods is more suitable for OVOC monitoring, because of the different calibration standards used (not possible to identify if the differences in uncertainties for all the instruments are only due to the calibration standards used or to the analytical system). To be able to give right recommendations about the most suitable method, measurements should be performed using the same calibration standards, for example, using one of the described SI-traceable working standards.

Specific Comments:

1. Line 50: This sentence contains ambiguity: while ozone has a much lower abundance compared to N2 in the atmosphere, its concentration is still relatively high compared to VOCs, and many countries and cities suffer from ozone pollution. It is recommended to remove the statement about low abundance and rewrite the sentence to emphasize ozone's contribution to OH production.

Lines 50–51: Statement removed and ozone's contribution to OH production emphasized.

2. Line 67: Typo on “play and important role”.

Line 67: Corrected.

3. Line 255: The description and quantity of "cylinders and canisters" in Table C2 differ from those provided in the main text and Appendix C. Please review and verify for consistency.

Lines 910–914: corrected; information on two canisters was missing in Table C2 (MVOC151-005D and MVOC151-008B).

4. Line 388: The tables in the Appendix B should be presented in order in the manuscript.

Corrected. Moreover, the order of the Appendix B subsections was changed, so they are introduced in sequential order in the main text (B.2 RGM verification and B.3 RGM stability evaluation).

5. Line 412-418: In some cases, images can be more vivid and readable than tables. Moreover, the manuscript emphasizes the assessment of four species: acetaldehyde, acetone, methanol, and methyl ethyl ketone (MEK). However, Figure 1 only presents data for acetone. Please include the results for the other species as well.

Lines 434–438: Figure 4 (former Figure 1) shows the results of an interlaboratory comparison that was used to complement the RGM verification process based on comparison against OVOC standards. This is not, therefore, part of the assessment of the working standards described in the manuscript. For the comparison, one of the laboratories could only assess three of the OVOCs added to the RGM cylinders: acetone, ethanol and methanol. The results for acetone were shown as an example (Figure 4). In Appendix B (Figure B.1), results for methanol are shown. For ethanol, due to technical issues with one of the analytical systems, the uncertainty of one of the participants (METAS) is too big to be able to drive any conclusion based on these results. Acetaldehyde and MEK was only measured by one laboratory (VSL). Although MVK was measured by two laboratories (VSL and METAS), the low long-term stability of the MVK permeation unit used as reference by METAS resulted in a big uncertainty that prevented us to use them to drive any conclusion.

6. Line 427: Please specify the data source for this paragraph and list it in the Appendix.

Line 553 and Lines 916–919: The data source for this paragraph is the analysis of variance performed on the data from the homogeneity test. Homogeneity data was included in Appendix C (Table C3).

7. Line 441: In the manuscript, the authors state that "certification results obtained for whole air samples contained in pressurised 10 L aluminium cylinders showed good consistency between the two laboratories performing the certification (i.e., VSL, METAS)." However, there is no data provided to support this claim. Please include the relevant data that demonstrates this consistency and add it to the Appendix for verification.

Lines 959–963: Data supporting this claim is included in Table C5 (former Table C4). We added a reference to this table to facilitate the tracking of these data. We also added a Figure in the Appendix representing the consistency of the measurements between VSL and METAS through the criterion defined in Eq. (4) (Figure C1; Lines 964–969).

8. Line 608 : The conclusions should be drawn cautiously. Based on Figures 2 through 6, it appears that the two different types of SI-traceable working standards may be more suitable for OVOC amount fractions in the range of 4–20 nmol mol^{-1}.

Lines 673–674: the figures included in these lines (4–10 nmol mol^{-1}) refer to OVOC ambient levels, which were provided by different monitoring stations during a stakeholder survey performed at the beginning of the EURAMET research project 19ENV06 "MetClimVOC". In former Figures 5 and 6 (they have been changed to relative differences in the new version following the recommendation of anonymous referee 3), the whole air samples certified and measured are above 10 nmol mol^{-1} (even close to 20 nmol mol^{-1} for acetone), but this was due to the OVOC spiking.

Answers to anonymous referee 3

This manuscript, "Towards a high quality in-situ observation network for oxygenated volatile organic compounds (OVOCs) in Europe: transferring traceability to the International System of Units (SI) to the field" describes 2 approaches to create SI-traceable standards for acetaldehyde, acetone, methanol and methyl ethyl ketone in detail. The motivation for this work is described nicely in the introduction. There are three scientific goals of the paper: to create reference gas mixtures (RGMs) in dry nitrogen at amount fractions of ~100 nmol mol^{-1} that are stable in time, to create working standards at amount fractions on the order of 10 nmol mol^{-1} by diluting the RGMs, and to create working standards at amount fractions on the order of 10 nmol mol^{-1} by spiking whole air samples. This work is highly collaborative and detailed. The sections are well-organized.

General Comments:

1. In general, the paper is highly detailed. I think a schematic describing the overarching process for creating these working standards could assist with the clarity of the paper, and anywhere a schematic could reduce details included as text would be appreciated (i.e. for the filling of the parent and subsample cylinders, or the diluting system).

   Following the valuable recommendations of referees 1, 2 and 3, we added three schematics (Figures 1–3) illustrating the different processes involved in this work: from generation or preparation of the SI-traceable working standards to their assessment (see answers to referee 2).

2. It is difficult to compare the two methods used to create working standards. For diluted RGMs, the results are shown in percent difference compared to the assigned amount fractions, and for the spiked whole air samples the results are shown in amount fractions plotted alongside the assigned value. Do the authors have a reason for providing the results in this way? It may be more helpful to future users of these techniques to present both in units of percent differences.

   Lines 610–614, Lines 633–637 and Lines 652–656: We thank the referee for the comment. The reason for providing the results in that way was to show the low amount of fraction levels measured as well as the difference between the spiked and the certified OVOC amount fractions. To help readers and users, we changed y-axis of the figures for the SI-traceable working standards based on certified spiked whole air samples to units of percent differences (Figures 8–10).

3. In section 4.1, the authors describe the methods used to assess the working standards based on whole air samples. My understanding is that the cylinders were large enough to be sent to all the laboratories, but canisters only had enough volume "for one analysis". Could the authors comment on their recommendations for going forward with this technique? Would monitoring stations be expected to purchase a cylinder (for

long term use) or a canister (for one time use) of working standard? Does the cost of manufacturing these standards seem worth the trouble for monitoring stations?

Lines 691–698: The referee's understanding is correct. The main limiting factor for the use of canister at monitoring stations is the instability of the whole air sample within the canisters (particularly the Siltek® stainless steel canisters) rather than the small volume of working standard within the canister (i.e., for one time use). Further research is needed but whole air samples within pressurised cylinders seems to be the option for going forward with this technique. The advantage of this kind of SI-traceable working standards is that they will allow monitoring stations calibrate their instruments using standards that use similar matrix gases than their samples (ambient air). Although RGMs are more stable, the matrix gas used is nitrogen, differing therefore from the ambient air. Matrix gas effects on the analytical systems are not fully understood yet, thus these standards might provide some insights on the topic. Although the authors don't have an answer based on solid proofs, the cost of manufacturing and certifying these standards might be worth the trouble for monitoring stations. The discussion was extended to cover this point.

4. Could the authors expand on what they mean by "molar masses" as a source of uncertainty during the preparation of RGMs (line 192)?

   Line 205: For the assignment of the amount fraction value of the selected OVOCs within the RGMs, the molar mass of each of the selected OVOCs is needed. This molar mass is based on the atomic weights of each of the atoms that form the compound (e.g., the molar mass of methanol is the sum of the atomic weight of carbon (C) and four times the atomic weight of hydrogen (H)). Atomic weight values are provided with an uncertainty (Coplen et al., 2020; van der Veen et al., 2021) that needs to be considered in the molar mass of each OVOC. Because, as said, the molar mass is part of the equation used to estimate the amount fraction of the OVOCs and it has an uncertainty associated propagated from the atomic weights, molar mass is considered a source of uncertainty of the prepared RGMs. We added these references to the manuscript.

   Coplen, T.B., Holden, N.E., Ding, T., Meijer, H.A.J., Vogl, J., Zhu, X. (2020). The Table of Standard Atomic Weights – An exercise in consensus. *Rap. Commun Mass Spectrom,* 36, e8864. https://doi.org/10.1002/rcm.8864

   van der Veen, A.M.H, Meija, J., Possolo, A., Hibbert, D.B. (2021). Interpretation and use of standard atomic weights (IUPAC Technical Report). *Pure Appl. Chemi.,* 93, 629–646. https://doi.org/10.1515/pac-2017-1002.

5. On line 397 the authors write, "Compound loss after preparation due to surface effects might explain relative differences of around -5 % for ethanol." Do the authors mean that ethanol had relative differences of around -5 %, and the most likely reason for that is surface effects? On first read, I thought the authors were suggesting they had done a calculation for surface effects and could explain -5 % of the ethanol differences.

   Lines 423–424: We rephrased the sentence to make clear that the relative differences for ethanol were around -5% and the most likely reason for that is surface effects.

6. On line 430 the authors write, "vessel material may play a role in the lack of homogeneity…". Was there a material that performed better than others? This is an interesting result.

   Vessel surface material performance was not the focus of this research. Experiments designed specifically to explore effects of the surface material on the stability and homogeneity of the working standards based on certified spiked whole air samples would be needed to be able to answer this question. Because the pressurised cylinders differ on their volume (10 L for cylinder type 1 and 3.6 L for the type 2) and these, in turn, differ from canisters in volume and pressure, it is difficult to state which material performed better. Differences in performance might be due to the other factors (e.g., differences on pressure, surface to volume ratio and cylinder histories) and not to the material. When comparing canisters, with same volume and similar pressures, it seems that Silonite™ stainless steel canisters perform better than Siltek® stainless stee canisters, but this cannot be confirmed because of the different history of the canisters, which might have contributed to these observations. Using new canisters and the right experimental design would allow to draw a conclusion regarding the material with the best performance.

7. On line 549 the authors write, "Assessment results for amount fraction levels around 10 nmol mol^{-1} suggests that SI-traceable working standards based on dilution of RGMs can be used as calibration standard at monitoring stations for key OVOCs, such as acetone. For other OVOCs, like methanol, acetaldehyde and MEK, suitability is not so clear due to the large uncertainty and result dispersion, particularly at amount fractions < 5 nmol mol^{-1}." However, in the conclusion, line 610 is, "The first type, based on RGMs diluted with dry nitrogen, using for that a portable dilution system that ensures SI-traceability after dilution, seems to be suitable for calibration of acetone, MEK and methanol at amount fractions around 10 nmol mol^{-1}." These statements seem conflicting as written.

   Lines 537–540: With the second of the first sentences indicated by the referee, we meant that at lower amount fractions (< 5 nmol mol^{-1}) this type of working standards for MEK and methanol doesn't seem to be suitable. We rewrote this statement to make the information clearer, avoiding the apparent conflict among the statements: "Assessment results for amount fraction levels around 10 nmol mol^{-1} suggest that SI-traceable working standards based on dilution of RGMs can be used as calibration standard at monitoring stations for key OVOCs, such as acetone. For other OVOCs, like acetaldehyde, suitability is not so clear due to the large uncertainty, particularly at amount fractions < 5 nmol mol^{-1}. At this low amount fraction level, suitability of the SI-traceable working standards for MEK and methanol is also questionable".

Technical Notes:

Line 67 - typo "atmosphere and play and important" should be "atmosphere and play **an** important"

Line 67: Corrected.

Line 68 - "the main sinks are the oxidation with the OH radical and…" would be better worded as "the main sinks are oxidation with OH radicals and…"

Lines 68–69: Changed.

Line 81 - typo "emphasises" should be "emphasise" or "emphasize"

Line 81: Corrected.

Line 118 - should "metrological accepted term" be "metrologica**lly** accepted term"?

Line 118: Thanks for the suggestion. We rephrased as follows to keep the meaning: "the accepted metrological term".

Line 214 - “PTFE” and “SS” are used for the first time, but not defined as acronyms

Line 797: Definitions of both acronyms added to the text.

Line 233 - typo “consisted on certified” should be “consisted **of** certified”

Line 234: Corrected.

Line 245 - typo, no need for comma after “both”

Line 254: Deleted.

Lines 320 and 799 - could not find an Eq. (B1); should a different equation be referenced?

Corrected. Line 333 and Line 942 refer to Eq. (C1) and not Eq. (B1).

Line 326 - CRF is used for the first time, but not defined as an acronym

Lines 344–345: Acronym definition added.

Line 553 - “contributed to **that**”, replace ‘that’ with the true meaning

Lines 540–542: Done.

Line 671 - typo “the carrier was helium” should be “the carrier **gas** was helium”

Lines 741–742: Corrected.

Table B1 - the subscripts appear in the order “2,3,1”, consider rearranging to “1,2,3”

Lines 827–830: Thanks for the suggestion. We rearranged the subscripts accordingly.

The current Table B3 appears in section B2, and the current Table B2 appears in section B3. It seems like renaming the tables (and their text references) would fix this.

Line 195, Line 414, Line 840, Line 841, Line 850 and Line 870: Tables were renamed and references in the text corrected.

Line 738 - typo, missing a close-parenthesis

Line 860: Close-parenthesis added.

Line 828 - typo “Prior instrument calibration, …” should be “Prior **to** instrument calibration, …”

Line 985: Corrected.