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 annexes (four) should not be changed. However, following this comment and similar comments provided by the other anonymous referees, we will implement 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.

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 will add 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. 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?

"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.