

**General comments:**

Safi et al present an assessment of quality and usability of synthetic reference materials, which are being prepared to increasingly high quality over recent years. The development of SRM is of interest to the atmospheric trace gas measurement community. Despite not presenting novel methodology or unexpected findings, the impressive (and improving) results of an apparently well-coordinated effort to prepare and assess these SRMs presented in this paper warrants publication in this journal. To improve the readability and utility of the material presented, I would ask the authors to consider making minor revisions (as in, elaborate but textual only), as detailed below.

**Specific comments:**

I'm not sure the main title of the ms covers the contents. One could argue the ms is not primarily about site-to-site consistency but rather is about consistency between SI and WMO scales. L115 states "This study presents a coordinated round-robin intercomparison designed to assess atmospheric GHG measurement consistency within the UK and Ireland monitoring network", which is in line with the title, but L123 reads "these analyses provide insight into the extent to which SI-traceable RMs can complement established calibration scales, while maintaining the levels of compatibility required for atmospheric GHG monitoring networks", which is different but also valid. Consider changing the title to reflect both aspects of the work.

Much of the manuscript goes to show that quality of SRMs is already quite high for some gas species, approaching "good enough" for CH<sub>4</sub>. However, the implications thereof are not made overly explicit. In what form could such SRMs be deployed? As multi-gas standards for monitoring stations SRMs are not of much use for now. Or are they? Can you speculate on where the value lies and does not yet lie? What may this field look forward to? Please provide outlook, constraints (do we need better gravimetric prep or something else?), future work (avoiding or compensating for matrix effects), etc.

Is this reviewer correct in concluding that this works implies that the WMO CH<sub>4</sub> scale is consistent (to within the uncertainty) with the SI? If so, I believe that's a statement worth making explicitly.

I recommend that authors summarize into a table the relevant details about the stations in this study. Name, abbreviation, operator, analyzer type, basic analyzer specifications, calibration scheme, etc. etc.). Please refer to that table, and do not re-enumerate everything in every figure caption.

This manuscript is made unnecessarily confusing to read by non-formalistic use of terminology. The statistics are intricate enough that they requires strict adherence to an objective and unambiguous phrasing. However (to half-jokingly illustrate my point) a quick search in this paper finds mention of ...

variability

average

the average standard deviation

coefficient of variation

coefficient of variability

internal agreement

agreement between the network-average weighted means and the SI-traceable assigned values

agreement between the network-average measurements and the assigned values

the site-to-site agreement

the degree of site-to-site agreement

the spread

the spread across the network

the spread of the measurements across the sites

the spread of results across the network  
 the spread of measurements between sites  
 the spread of measurements between instruments and sites  
 absolute uncertainties on assigned values  
 network consistency  
 internal consistency  
 internal network consistency  
 network-average biases  
 average network offset  
 biases between sites  
 systematic network bias relative to the calibration scale  
 network-wide bias relative to the scale  
 uncertainty associated with site measurements  
 the uncertainty associated with scale propagation  
 scale uncertainty  
 scale propagation uncertainty  
 uncertainty  
 uncertainty of the result  
 complete uncertainty  
 total uncertainty  
 calibration scale uncertainty  
 uncertainty from the measurements  
 uncertainties of the analytical instrumentation  
 certified SI-traceable uncertainty  
 uncertainty of the SI-assigned values  
 uncertainty of the assigned values of the cylinders  
 calibration scale uncertainty  
 ... and likely additional variants of a – crucially – much smaller number of concepts. Consider consolidating these variants into the smallest possible number of clearly defined and unambiguously used designators. Consider using subscripts, superscripts, abbreviations etc. for brevity and consider providing a list or table summarizing your choices.

#### **Technical corrections & suggestions:**

L35. The compatibility goals are in absolute terms (for N<sub>2</sub>O: 0.1 ppb / 0.3 ppb (ext.)). The limited natural range is not relevant to meeting that goal, right? Please reconsider.

L40 characterization => characterization and compensation

L41 "dependant" ==> "independent" (right? you want to achieve INdependence, not dependence -- and it's with an e, not an a)

L42 high-precision => high-accuracy (?)

L46 1.6 sounds like the global mean SURFACE temperature deviation, not the LAND temperature deviation. Please check.

L91 Consider starting a new paragraph at "International ..."

L105 coordinateS

L187 SpectronUs or SpectronOUs?

L229 SpectronUs or SpectronOUs?

L276 For ease of reading, consider adding the scale names into a second or last row of the table

L285 "BOC" -> "BOC specialty gases, city, country"

L319 "4.1 Site-to-site measurement offsets". I don't think this section deals with offsets between sites ("station x is mostly lower than the others"), but rather deals with, for each site, the offset between its measurements and the assigned values. The title should reflect this.

L324 I'm not sure what you mean by "shown in ascending mole fraction order" (also in figure caption). Lowest concentration of co2 in the top row? That is confusing more than relevant.

L326 Is there not a figure 2 for the SRMs?

L366: why specify the WARMs again?

L368: why specify the scales again?

L384: "residual" could be "network bias"?

L420 "RMs". Is that "WARMs" or "SRMs"? Be specific.

L423 I don't get this sentence. The residuals are close to the assigned values??? Or are they close to 0? Residuals of 330 ppb would be quite bad. And they're noticeably larger than the CO2 residuals? In an absolute sense? And the residuals are accompanied by smaller relative uncertainties?? Please rephrase using unambiguous terminology.

L424 That is hard to glean from the tiny figs, if true at all. Consider a table that summarizes the attained quality as a fractional adherence to GAW goals. at 120% of the GAW limit, 60% of the extended limit, etc.

L436 Figure 3 caption restates a whole part of the main text. Consider picking one of the two. Not sure there's need to \*pink\* dot the zero axis, dotted is fine.

L437 not each site but ALL sites. Or: "For each RM, the average of all sites and instruments"

L452 "Across the four synthetic air RMs, the degree of site-to-site agreement varies by species, typically showing slightly larger differences from the assigned value than observed for whole air". What is "site-to-site agreement"? Is that network-wide uncertainty? Is the latter half of the sentence meant to illustrate for former? If so, I don't follow. The "differences from assigned values" is something else than the "degree of site-to-site agreement". What is the point you try to make?

L468 sentence doesn't flow.

L470 "values and ... uncertainty assigned ... allow[s] for traceability ... with an uncertainty that validated by x and y." The two uncertainties in sentence, are they the same entity? If so, consider breaking the sentence up in two, and start the 2nd sentence with "That uncertainty is validated by x and y".

L476. "Consequently, while the measurements generally demonstrate good internal agreement amongst the [WA and S] RMs, they do not provide a direct assessment of the absolute accuracy of the synthetic air standards themselves, but rather of their alignment to the existing whole air calibration scale." Consider presenting this argument also in the reverse (and maybe more interesting) order: the fact that WARMS and SRMs match well indicates that the arguably slightly arbitrary/consensus WMO scale is congruent with the SI! I get that for practical matters it's cool that SI doesn't dictate a scale change to what our field has been using for decades now, but that's only a convenience argument – the fact that (at least for CH<sub>4</sub>) WARMS are good against SI is much more fundamental (maybe – I'm claiming no expertise here – also for compatibility of our work with concentrations that are otherwise derived from SI, such as – maybe – TCONN-style FTIR).

L495 This caption is nicely exhaustive in its detail, but for brevity consider shortening it dramatically and just referring to a (new) table listing stations, analyzers, etc. "x-axis labels indicate station name and analyzer, as specified in table ###."

L499 Don't restate the assigned values but refer to the respective table

L517 Should that read "figures 2 and 4" (not 2 and 3)?

L525 consider writing Key Comparisons with capitals, to stress they are a specific, defined concept (and a name)

L543 "internal consistency between standards" what is that? KC's compare between labs right, not between standards?

L544 "with differences well below the spread observed across all participants". Is that "differences were observed between all participants" or "all participants observed small differences"? This sentence may be clear to those who read KC's, but not to the uninitiated. Also, I do not follow how, from that, follows "This strong internal agreement supports the effectiveness of a scale-based approach". If KCs conclude that SRMs are very good, that doesn't support "the effectiveness of WARMS", correct?

L560 "This study uses WARM and SRM to assess UK network performance" is rather different from how the ms started, namely "we use the UK network to assess compatibility between WARM and SRM". Then you write that you do this study to compare between sites and instruments, but that was never the outset, no?

L572: "For the synthetic RMs, CH<sub>4</sub> was shown to have uncertainties on the SI-traceable assigned values that are comparable to the extended compatibility goals". I don't follow. If I understand things correctly, YOU are not the party demonstrating this, correct? As in, The "uncertainties on the SI-traceable CH<sub>4</sub> values assigned to SRMs" are given in the Key Comparisons. Or have you discussed different "uncertainties" without clearly differentiating between them? This again begs for more rigorous naming, likely with a lot of subscripts etc.