

Emmal Safi, Senior Scientist
National Physical Laboratory
Teddington, UK

12th October 2026

Dr Andreas Zahn, Editor
Atmospheric Measurement Techniques

Dear Dr Zahn,

Please find below our response to the comments from the anonymous referee #2 on our paper: "Assessing site-to-site greenhouse gas measurement consistency in the UK and Ireland atmospheric monitoring network using whole and synthetic air reference materials".

We would like to thank the reviewer for taking the time to look over our manuscript and provide us with comprehensive and constructive feedback.

Best regards,

Emmal Safi

(on behalf of all co-authors)

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.

Thank you for your positive assessment of the manuscript and for recognising the value of the RM development and intercomparison presented in this work. We appreciate the constructive suggestions for improving the readability and utility of the manuscript and we have addressed each of the specific comments 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.

Thank you for this comment. We agree that the manuscript addresses two related aspects you mention. We have therefore revised the title to better reflect both aspects of the study, while keeping it sufficiently concise. The revised title is: “Assessing greenhouse gas measurement consistency and traceability in the UK and Ireland atmospheric monitoring network using whole and synthetic air reference materials.”

Much of the manuscript goes to show that quality of SRMs is already quite high for some gas species, approaching "good enough" for CH₄. 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.

Thank you for the comment. We have expanded the discussions to clarify where these RMs may have value and where further development is still required. Specifically, we now distinguish between CH₄, for which SI-traceable uncertainties are approaching atmospheric monitoring requirements, and CO₂ and N₂O, for which the uncertainties remain too large for use as high precision calibration standards. We also clarify that this currently limits their application as multi-species calibration standards at monitoring stations (paragraph added to the manuscript at the end of section 4.4). We have also expanded the discussion of future development requirements, including a focus not only on reducing SI-traceable assigned value uncertainties, but also on ensuring that these values can be reliably transferred to atmospheric measurements on the order of magnitude required for high precision monitoring (paragraph added to the manuscript at the end of section 4.4.2). We have also modified the conclusions section slightly to reflect the added text in the manuscript.

Is this reviewer correct in concluding that this works implies that the WMO CH₄ scale is consistent (to within the uncertainty) with the SI? If so, I believe that's a statement worth making explicitly.

We agree with this conclusion, with a caveat. In lines 479-484 of the original manuscript, we do state that “...the two traceability chains, the SI realisation and the WMO-NOAA scale, for these RMs are statistically consistent within their stated uncertainties.” However, we then go on to state that “However, the monotonic increase in the magnitude of the offset with increasing mole fraction (see CH₄ on the right-hand side (RHS) of Fig. 3) could indicate the presence of a systematic difference between the two approaches. Comparison against a larger suite of SI-traceable RMs would be required to begin to confirm the existence of this effect.” Therefore, we do not feel like we should make this statement too explicit, particularly if we take into account this is only one case study. We have however, revised the statement in lines 479-484 of the original manuscript, and it now states “...suggesting that, over the mole fraction range investigated here, measurements traceable to the WMO-CH₄-X2004A scale are consistent with the SI-traceable assigned values within their stated uncertainties. This provides evidence of consistency between the WMO-NOAA and SI traceability routes for CH₄ within the scope of this study”.

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.

Thank you for this suggestion. We considered consolidating the monitoring site and instrument information into an additional summary table. However, we have retained the current presentation because the relevant information on monitoring sites, instrument specifications and calibration procedures is already provided in the corresponding methods section, where necessary context and methodological detail can be given. Combining site characteristics, instrument specifications and calibration procedures into a single additional table would therefore substantially duplicate information already present in the manuscript. We have also retained identification of sites and instruments in the figure captions so that the figures remain sufficiently self-contained and can be interpreted without requiring repeated cross-reference to a separate table.

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.

Thank you for highlighting this. We agree that the use of the multiple terms you have identified for the same statistical concept makes the manuscript unnecessarily difficult to follow. We have therefore reviewed the terminology and removed a number variants identified and replacing them with consistent terminology where they refer to the same quantity or concept, this includes removing: coefficient of variability, internal agreement, the site-to-site agreement, the degree of site-to-site agreement, the spread, the spread of the measurements across the sites, the spread of results across the sites, the spread of measurements between sites, internal network consistency, average network offset, systematic network bias relative to the calibration scale, uncertainty associate with site measurements, the uncertainty associated with scale propagation and scale uncertainty. We have retained some distinct terms where these have specific and different meanings within the analysis, for example to distinguish between variability across sites, bias relative to an assigned value, and the different sources and components of uncertainty. Following these revisions, the number of terms used has been substantially reduced and each remaining term is used consistently for its intended meaning. As the remaining terminology is limited and defined in context, we have not introduced an additional terminology table or statistical notation, which we felt would add complexity rather than further improve clarity.

Technical corrections & suggestions:

L35. The compatibility goals are in absolute terms (for N₂O: 0.1 ppb / 0.3 ppb (ext.)). The limited natural range is not relevant to meeting that goal, right? Please reconsider.

Thank you for highlighting this. We agree that the limited natural variability of atmospheric N₂O does not itself make the absolute compatibility goal more difficult to meet. Rather, the small atmospheric variations that measurements are intended to resolve provide the scientific motivation for requiring stringent compatibility, while the practical challenge in meeting the goal relates to measurement precision and stability. We have revised the sentence accordingly to avoid implying the relationship between atmospheric variability and the ability to meet the absolute compatibility goal.

L40 characterization => characterization and compensation

Added.

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

The word has been removed from the sentence.

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

Thank you for the suggestion. We have retained the term “high-precision” here, as this is intended specifically to refer to the level of measurement precision required for atmospheric monitoring. We distinguish precision from accuracy throughout the manuscript, with accuracy relating to agreement with an assigned reference value and precision relating the measurement variability. We therefore consider “high-precision” to be the more appropriate term in this context.

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

We have checked the value against the IPCC AR6. The reported warming of approximately 1.6°C refers specifically to the increase in surface temperature over land (1.59°C [1.34-1.83°C]), whereas the corresponding increase in global surface temperature was 1.09°C [0.95°C-1.20°C]. both values refer to the 2011-2020 mean relative to 1850-1900. We have revised the sentence to make this more explicit.

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

Done.

L105 coordinateS

Added.

L187 SpectronUs or SpectronOUs?

It is Spectronus.

L229 SpectronUs or SpectronOUs?

It is Spectronus.

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

Thank you for the suggestion. We have added the relevant calibration scale names to the table.

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

We have added “specialty gases” to the sentence.

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.

Thank you for highlighting this. Section 4.1 is intended primarily to assess the consistency of measurements between sites and instruments using the whole air RMs, rather than the offset of each site from the assigned value. The latter is considered separately in section 4.2, where the network weighted mean is compared with the assigned calibration scale traceable value. We agree however, that the term “site-to-site measurement offsets” in the original section title could imply that offsets between individual sites were being calculated. We have therefore revised the title to “site-to-site measurement consistency using whole air reference materials” to better reflect the analysis presented. The corresponding title for the synthetic air results in section 4.3 has also been revised.

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.

This wording was intended to indicate that the RMs were arranged from lower to higher mole fractions, but we agree this information is not necessary for interpretation of the figure and that the wording could be confusing. We have therefore removed the reference to "ascending mole fraction order" from both the main text and figure caption of figures 2 and 4.

L326 Is there not a figure 2 for the SRMs?

Yes, the equivalent results for the synthetic air RMs are presented in figure 4.

L366: why specify the WARMs again?

We have removed reference to the cylinder IDs.

L368: why specify the scales again?

We have removed reference to the specific calibration scale names.

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

Thank you for the suggestion. We have retained the term "residual" as this refers specifically to the calculated difference between the network weighted mean and the assigned value.

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

We have specified this to mean whole air RMs.

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.

Thank you for highlighting this. We agree the original wording was ambiguous. We have revised the text to clarify that the N₂O residuals are larger (further from zero) than those observed for CO₂ and CH₄, indicating weaker agreement between network average measurements and assigned values. We have also clarified that the uncertainties associated with the N₂O residuals are large relative to the stringent WMO/GAW compatibility goals for N₂O.

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.

Thank you for the suggestion. We agree that the magnitude of the combined uncertainty associated to the residuals is difficult to assess quantitatively from the figure alone. To provide the information more clearly, we have added the combined uncertainty associated with each residual to table 5. We considered additionally expressing these values as percentages of the respective compatibility goals, however, as there are both compatibility and extended compatibility goals for each gas, this would introduce multiple additional percentages metrics and make the presentation unnecessarily complex.

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.

Thank you for this suggestion. We have substantially shortened the captions for figures 2, 3 and 4, removing information already provided in the text. We have retained the site and instrument identifiers, together with explanations of the plotted lines and error bars, to ensure the figure remains interpretable without unnecessary cross-referencing the main text.

We have modified the pink dotted line so that it is now black.

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

Modified.

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?

Thank you for highlighting this. We agree that the original sentence conflated two distinct aspects of the analysis: then standard deviation of the measurements between sites/instruments, and the difference between the measurements and the assigned values of the RMs. The intention here was to describe the former, which is quantified using the standard deviation and coefficient of variation across sites/instruments. We have therefore revised the sentence to refer only to the site-to-site consistency and removed comparison with the assigned values from this sentence.

L468 sentence doesn't flow.

We have modified the sentence.

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".

Thank you for the comment. The two references to uncertainty referred to the same quantity. We have separated the statement into two and clarified the role of key comparison exercises involving independent SI-traceable realisations.

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 CH4) 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).

Thank you for this perspective. We agree that the comparison can also be interpreted in the reverse direction. Since the synthetic air RMs have independently assigned SI-traceable values, agreement between these values and measurements calibrated using whole air RMs traceable to the WMO-NOAA scales provides evidence of consistency between the two traceability routes.

We have revised the discussion to make this interpretation more explicit. The most robust evidence for this comes from the CIPM key comparisons, but the results of this study are consistent with those conclusions.

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 ###."

Thank you for this suggestion. We agree the original figure caption was unnecessary detailed, and we have substantially shortened it. We have retained the station and instrument identifiers within the caption to ensure that the figure remains interpretable independently of the main text. As the relevant information on monitoring sites and instrumentation is already provided in the methods section, we did not consider an additional table necessary, as this would largely duplicate existing information.

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

We have removed the assigned values from the figure caption.

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

Thank you for spotting this. We have modified.

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

Modified.

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

Thank you for highlighting this. We agree that the primary purpose of the key comparisons is to assess agreement between laboratories and their independently realised RMs. Our use of "internal consistency" was intended to refer to the consistency of the results obtained for multiple standards submitted by the same participant, rather than the primary between-laboratory comparison. We have revised the text to make this distinction explicit and now use the term "within-laboratory consistency" where appropriate. This distinction is explicitly considered in CCQM-K82 through its assessment of within- and between-laboratory variability, and in CCQM-K68 the greater within-participant consistency is specifically identified as supporting a scale-based approach. We have revised the discussion accordingly and clarified the interpretation of the K120 results.

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?

Thank you for highlighting this, we agree that that the original wording did not make clear what we intended to say. The argument is not that the key comparison demonstrates that synthetic air RMs are less effective than whole air RMs. Rather, CCQM-K68 found greater consistency between RMs originating from the same laboratory realisation than between independent laboratory

realisations, and the K68 report identifies this as supporting a scale-based approach. A scale similarly disseminates values from a common central realisation, thereby reducing the influence of differences between independent realisations. We have revised the manuscript to make this interpretation explicit and avoid implying that the key comparison itself demonstrates an advantage of whole-air over synthetic air RMs.

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?

Thank you for the comment. We would like to clarify that assessing differences between sites and instruments was one of the aims of the study from the outset. The original abstract stated that measurements were made at six sites using a range of instruments and that the results were used to assess biases between sites (which we have now modified to "consistency between sites"). This is also stated at the end of the introduction, where the round robin exercise is described as allowing direct assessment of site-to-site and instrument-to-instrument differences under field conditions. The comparison of the measurements with the WMO-NOAA and SI-traceable values represent a second aspect of the study. We have revised the abstract and conclusions to more clearly reflect both objectives and to align with one another.

L572: "For the synthetic RMs, CH₄ 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 CH₄ 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.

Thank you for the comment. We believe there may have been some ambiguity regarding the source of the uncertainties referred to in this sentence. These are not uncertainties obtained from the key comparisons. They are uncertainties associated with the certified SI-traceable assigned values of the synthetic RMs used in the present study. The key comparisons are discussed separately in section 4.4.2 to provide context on the performance of independent SI-traceable realisations. In the present study, we compared assigned value uncertainties of our synthetic RMs with the WMM/GAW compatibility goals; we do not determine those assigned value uncertainties from the round robin measurements. We have revised the sentence in the conclusions to make the distinction explicit and avoid confusion with the uncertainties discussed in relation to key comparisons.