Long-term Intercomparison of Commercial Air Quality Sensors: An Overview from the QUANT Study

Sebastian Diez, Stuart Lacy, Hugh Coe, Josefina Urquiza, Max Priestman, Michael Flynn, Nicholas Marsden, Nicholas A. Martin, Stefan Gillott, Thomas Bannan, Pete Edwards

This study is a comprehensive, long-term study of a wide range of air sensors currently available in the market. Because of the collocation in time and space during the comparison, environmental variables are lessened thereby focusing the comparison on the performance of the air sensors relative to one another and in comparison, to a "reference" monitor. This adds great value to the study. The sensors were deployed and collocated in a real-world environment that they will most likely be used, so the conditions at which the sensors are compared were not biased because of the environment (i.e., compared to if it were collocated in a "pristine" environment, or under laboratory conditions). This study includes both gas and PM sensors which adds to the novelty of the study.

This paper is recommended for publication in AMT with revisions as outlined in this discussion.

<u>GENERAL COMMENTS</u>

- The paper can benefit from tightening up language and being more succinct and concise in its statements.
- A glossary of terminologies for commonly used yet widely misused or confused terms in the field (e.g., "sensor" vs "sensor systems" vs "sensing unit", "manufacturer" vs "company", "model"/"unit"/"type"/"device") may be useful to the reader, also to help guide the authors in using consistent terminology all throughout the paper.
- A major issue is the description and reliance on usual metrics like R^2 in comparing two instrumental methods. With a goal of prediction and calibration in mind, R^2 is an appropriate statistical metric; however, in plainly comparing the correlation (specifically: the concordance or agreement between two measurements), the authors are recommended to use more appropriate statistical metrics that measure ***concordance*** such as the concordance correlation coefficient. The paper can also greatly benefit a wider audience if the authors expanded on the statistical discussion and provide a separate discussion of the statistical metrics used, thus serving as a technical guidance that outlines metrics that can be used in such an intercomparison or calibration exercise.

<u>SPECIFIC COMMENTS</u>

- For the title: A more specific term than "Evaluation" can be used. Suggestions: "Intercomparison"; "Precision Analysis"
- The abstract is missing some key findings and results, e.g. how many air sensors and reference sensors were quantified, statistical metrics used to quantify the performance of the sensors, etc. For example, Lines 89-90 can be added to the abstract.
- A separate section on the methodology summarizing performance metrics used, and explaining each under a subheader, e.g., "Bias" as a subheader and explaining R2, RMSE, etc. under this

heading would be useful to the reader and also makes this a good reference paper for intercomparison studies.

- Section 3.3 explores reference instrumentation. Authors need to make what is meant by "reference", e.g. a reference method designated by an authority (EU, US EPA, etc.) or a self-defined or agreed-upon reference method.
- In one section, the manufacturers / models of air sensors were referred to; however, in Figure 10, it was anonymized. What was the rationale? Can it be consistently anonymized or named? And if not, explain why and make sure that the transition is clearly explained within each section.
- It is useful from a consumer perspective to mention which devices are available readily as-is (without add-ons) and/or which ones require customization from the manufacturer's end. This can possibly be added to the summary in Table 1 and/or Supplementary S1, with a short reference (a sentence or two) in the main text.
- Might be useful to add in the conclusions / recommendations section for future researchers: quantify inter-location variability.
- Might be useful to explain and emphasize (including in the abstract) why correction with satellite data was not explored in this study.
- Can employ the terms "inter-device" and "inter-location" for succinctness of ideas.

Line by line comments

- Line 45: "cross-sensitivity" seems to be a term usually used in the medical context. It might be beneficial to define this term in this context, and differentiate it from "interference". Levy Zamora (2022) used "cross-sensitivity" in the title of their article and Bittner (2022) defined it, so it might be helpful if these two comes as the first articles cited in this instance)
- Lines 48-49: citations for temperature and humidity might be combined since they are usually explained in cited references in combination.
- Lines 63-68: This paragraph could benefit from differentiating "calibration" from "correction" and how these two terms are sometimes interchangeably used (albeit incorrectly). A reference to an article that explains this difference will also be helpful. The Liang (2022) paper cited explains some of these nuances, including mathematical equations for calibrations, but does not fully differentiate "correction" from "calibration".
- Lines 67-68: True for gases. Mention examples of acceptable calibration method(s) for PM?
- Line 73: See also Raheja, et al (2023) https://doi.org/10.1021/acs.est.2c09264
- Lines 81-83: Another reason is that there are a lot of sensors/sensor systems with different configurations commercially available, and also individual sensing units are sold and can be "DIY"-ed—the market is diluted with many options and many different iterations of the same underlying technology with marginal differences.
- Line 94: Clarify or add examples of "data products" e.g., APIs, mobile apps, etc.
- Lines 105-106 and 116: Useful to add a subsection that describes the UK urban environment including seasonality, sources of pollution (transportation? Household commercial products use?) in the three locations (London, Manchester and York)
- Line 106: "replicates" or "units" are more appropriate terms than "duplicates" if you are talking of the units of the same model
- Line 109: define what is meant by "near real-time" in this context.
- Line 113: Were the units tested together before deploying separately? Clarify.

- Line 121: A sentence or two succinctly describing the sites will also be useful in this line. Then you can refer to the Supplementary.
- Lines 122-126: Consider moving up before Lines 113-121.
- Line 125: "inter-device consistency" may also be rewritten as "precision".
- Lines 134-136: "vendors were invited to contribute multiple sensor devices throughout the WPS study". How does a "sensor device" differ from a "sensor" or "sensor system"? Does this mean that manufacturers can contribute different sensor models? Also, does vendor = company = manufacturer? Note consistent terminology all throughout the manuscript (might be useful to have a glossary or footnote, like that for "sensor" and "sensor system" on page 2.
- Lines 139-141: Table 1. Does AQMesh AQM, Kunak AP, and SCS Prax have *all* of the sensors listed (from NO to PM10) in one unit? This table might benefit from a clarification (can be added to the Table caption). Also, as mentioned in a previous comment, add in the description if these are consumer-ready (eg already sold in the market as that unit), or customizations available from the manufacturer.
- Lines 148-149: I understand that PurpleAir does not have a mobile data connection, only WiFi, but WiFi was not good in the location so you opted to download the data from the device memory instead. The text can be enhanced by better explaining the issue as described. (i.e., differentiating from WiFi and mobile data connectivity)
- Line 150: and harmonize? In the methodology section, it might be useful to mention that temporal and spatial scales of the sensor systems was important to match, thus aggregation and harmonization was necessary. How was incomplete data treated? Were there imputed data? Might be useful to add it in the supplementary.
- Line 150: by data format, do you mean datetime / time and date?
- Line 158: "calibrated data products": is this referring to API? Measurements? As with my previous comment – clarify what "data products" mean.
- Lines 160-166: What is cal1? cal2? Clearly define / describe these in the text and/or supplementary. This section may benefit from a subsection explaining / describing these.
- Lines 170-174: Is this a caveat / weakness of using these statistical metrics used herein (R^2, MAE, etc)? What is the alternative? I suggest concordance (agreement) metrics, such as the Concordance Correlation Coefficient: See Lin, Biometrics (1989): https://doi.org/10.2307/2532051. The reader might also benefit from a separate subsection and/or supplementary section describing the metrics or including a glossary of the metrics used.
- Line 183: "multiple devices of the same type" when you mean "type" do you mean similar underlying principles of measurement? Model? Be consistent in terminology. Also, it might be useful to cite an example of which devices you are considering a same "type", e.g. AQM and Clarity—are these of the same "type" as described?
- Lines 190-196. See also deSouza (2023): An analysis of degradation in low-cost particulate matter sensors https://doi.org/10.1039/D2EA00142J
- Lines 202-204: Good point.
- Line: 217: 75% inclusion criteria is common—but perhaps not for readers not familiar with this data type. Readers might benefit from a citation, explanation in methodology or supplementary. Suggested section to add it in: Section 2.1, lines 150-151.
- Line 225: Clarify: did you mean closer together spatially / physical location? How "close" is close?
- Line 232: could benefit more from a further explanation of the bias-variance tradeoff.

- Section 3.3. and Supplementary S4. Cite the authorities that consider the instruments mentioned as reference. e.g., are they considered “reference” because they are listed in an EU directive or US EPA documentation? If so, cite these. If not, provide a rationale or a citation as to how these instruments were categorized as “reference”.
- Lines 241-242: Expound on the significance and advantages of REU as opposed to the other metrics.
- Lines 268-269: clarify / reiterate the said minimum requirements in this text.
- Lines 283-284: R^2 can many times be subjective, e.g. how can we say that an R^2 of 0.87 means “it does not fully agree” and a slope of 0.80 is considered a ***pronounced*** bias? This is where the definition of concordance and using concordance metrics might be useful. If two measurements are concordant (in agreement), then slope is expected to be unity (=1). Also, a high R^2 does not necessarily mean agreement between the two instruments. Also, clarify what is meant by “limiting the linearity”. The authors are cautioned against using R^2 in quantifying agreement between two instruments that are being compared.
- Lines 288-289: Specify criterion stipulated by EU DQOs
- Line 308: “reasonably consistent” – reasonably is subjective and qualitative. Suggest dropping the word, or provide a percentage of the time that the RMSE is consistent (e.g. provide a qualitative measure).
- Line 312: “local conditions”: give or name some examples. Do you mean weather conditions? Traffic? etc.
- Line 334: quickly define (add a phrase) that describes “overfitting”
- Line 336: “linear correction” – “linear regression” might be the appropriate term.
- Line 347: RMSE also showed seasonality.
- Lines 345-351: Add more explanations about the seasonality. Add recommendations.
- Line 355: Clarify what a “1-day slide” means – it can be added in the supplementary or a quick description in the figure caption.
- Lines 363-366: This can be said more succinctly. Also, what sort of information can be provided? Be specific, based on your results so far—what sort of information can you recommend be provided?
- Line 373: Inconsistency in the use of the term “sensor”, “system” and “sensor system”.
- Lines 377-379: This discussion can benefit from a more detailed explanation of the tiers (Classes) assigned and what was the basis of the assignment to different classes. The figure caption for Figure 10 offers an explanation, which should be repeated and explained in more detail in-text.
- Line 388 onwards can benefit from a separate subheader / subsection.
- Lines 390-391: Specify an example of “simpler methods”
- Line 393: explain more by what instrumental method is NO_2 measured/detected from these diffusion tubes. Cite a reference as well.
- Line 411: “change points”: does this mean inflection points? Periods of biggest slopes? Peak concentration periods? Consider changing verbiage.
- Lines 414-415: Is it applied in this paper? If so, this line should be described in the methodology and explained further.
- Line 421: Expound what “unsupervised analysis” means in this context. General verbiage related to machine learning may sometimes be unnecessary to use in this text, and can be avoided,

because fundamental/rudimentary statistical metrics (as opposed to complex "black-box" machine learning algorithms) are used.

- Line 422: consistency in terminology. Do the authors mean "sensor system" when they mention "devices"?
- Line 433: "..use of these devices has been primarily limited…" I would disagree, because consumers and many users still use these devices (sensor systems) and they aren't necessarily limited by accuracy concerns, e.g. many users are willing to accept a large margin of error for awareness purposes.
- Line 439: suggested addition: (limitations in) technical ability in post-processing of data
- Lines 460-461: Will this be done by the authors in a future study, or is this a call/recommendation for other researchers?
- Line 466: suggestion for a future study: explore different VOC-NO_x regimes (see Wennberg, ES&T Air: https://doi.org/10.1021/acsestair.3c00055)

TECHNICAL CORRECTIONS

Grammatical, Typographical, Figure and Formatting comments

Throughout the text

- Note the usage of "data" as a plural noun, e.g. "data were" rather than "data was"
- "Manufacturer" rather than "Company" might be a more descriptive noun for the intended usage.
- "co-location" vs "collocation"? Stay consistent.
- Many links in the "References" section of the supplementary point to a Zotero page that is meant for Google docs, thus rendering the links inaccessible

Figure comments

In general, labeling the figure panels with letters (e.g. (a), (b), (c), (d)) allows for easier and clearer reference in text and in figure captions. (e.g. Line 327 mentions the "top row" in Figure 8)

- Figure 1. Good visual—a nice representation of the timeline of events.
- Figure 7. Which sensors are being compared here? Why the anonymity compared to the other section(s)? Also, the readers may benefit from a colorblind-friendly and more contrasting color palette. "Class 1" and "Class 2" sensors are not actually described until page 15 (line 377 onwards) – it might be useful to refer to this section (i.e. Section 3.6) in the figure caption or the accompanying text (paragraphs) that describes this figure, and mention that it will be thoroughly explained in that section.
- Figures 8 and 10. Explain the colorations, e.g. is it meant to be a heat map? What do the specific colors mean? Figure 10 may also benefit from higher contrasting – difficult to see the contrast especially in the lower left panel, and when the plots are printed. Dashing is also difficult to see—might benefit from greater color contrast.

Line by Line

- Line 80: Suggestion: “academia” or “academic research” instead of “academic arena”.
- Line 104: Suggestion: reword “transparent”. Suggested synonyms: open, comprehensive (this changes the meaning a bit)
- Line 118: Typo: “influenced”
- Line 155: Suggestion: reword “ratified” to “validated”
- Line 160: is “time-line” the correct term? Perhaps “comparison” or “matrix” would be more apt for Figure S1; Figure S2 is a scatter plot or a bivariate plot.
- Line 162: change “to use this data” to “to use **these** data”
- Line 206: “Mean Bias Error” rather than “Mean Error Bias”
- Line 209: enclose “out-of-box” in quotation marks; typically “out-of-the-box”
  Line 231: semi-colon after “MBE”, comma after “machine learning”
- Line 257: actual “metrological” as in measurements and units, or “meteorological” as in RH and Temp?
- Line 271: “…hypothetical scenario where it…” does “it” refer to T200U? T500?
- Line 275: “All of this” to “all of **these**”
- Line 276: Add comma after “monitoring”
- Line 278: “equivalent-to-reference” – consistency in hyphenation
- Line 282: “obtained with a BAM **at the** AURN York site, located on a busy avenue” – delete parentheses
- Line 289: Omit “of course”
- Line 299: capitalize “FIDAS”
- Lines 300-301: the choice of the reference measurement
- Line 309: Paraphrase “saw its slope change”. Suggested: …”a slope change from 0.69 to 0.86 was observed…”
- Line 310: change “when” to “while”
- Line 321: “despite” might not be the correct conjunction here.
- Line 331: change “akin to **this** latter” to “akin to **the** latter”
- Line 268: Redundant. Change “with a measurement instrument” to “with an instrument”
- Lines 368-369: Can be paraphrased to be more succinct.
- Line 383: “4-system” rather than “4 systems companies”
- Line 386: add “dashed”, i.e., green dashed rectangle
- Line 398-399. “high time-resolution” (note hyphen placement)
- Line 400: subscript on NO_2
- Line 400: Is “DEFRA” all capitalized, or is it “Defra” as mentioned in the acknowledgement (Line 489)?
- Line 407: Consider using a different word from “digestible”
- Line 433: change “uptake” to “uptick”
- Line 452: “high level” seems unnecessary.
- Line 455: “accuracy with respect to reference methods”
- Line 471: Lacks the link to supplementary information (online version link is accessible).