Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2022-58-RC1, 2022

Comment on amt-2022-58

Anonymous Referee #1

Referee comment on "Air pollution measurement errors: is your data fit for purpose?" by Sebastian Diez et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2022-58-RC1, 2022

General Comments

Overall, the paper is well written, and makes some important points regarding the limitations of simple performance metrics and the need for more intensive investigations of measurement error. I agree with the conclusions of the paper in principle, but I think that the paper could better support these conclusions through its examples. As written, besides the contrived example of Figure 2, I don't see a clear case where differences and deficiencies in the measurements are not at least hinted at through relatively worse simple performance metric values (R-squared, RMSE). As shown in several examples of Section 4, the simple performance metrics do have utility in allowing comparisons between alternative measurement devices or techniques; for the most part, the sensors showing difficulties in the B-A and REU plots also showed relatively worse performance metric values. This is partly due to how these results are presented; they mostly compare data from common collocation experiments at the same location and covering the same time period, and therefore represent situations where it would be more appropriate to compare simple performance metrics (this is well stated in lines 394-396). The exception is Figure 6, but in that case, it isn't clear that the B-A or REU plots show any more ability to anticipate the poor observed performance at the roadside site than the simple performance metrics; rather, this is a general issue of relying on single-site collocation studies for characterization.

More directly relevant to the topic of the paper would be to show attempts to compare different collocation datasets using simple metrics only, and to illustrate the shortcomings of that approach; these shortcomings can then be addressed through the approaches you suggest. Perhaps such an example might be constructed from the existing data you present in the paper. For example, a collocation dataset could be divided across time by taking data collected in different seasons (if possible) and treating these as separate collocation experiments. In different seasons, the same sensor could have different performance metrics due to the differences in concentrations and variability in environmental conditions between seasons. These differences and their effects on errors likely would be much more apparent in B-A or REU plots (e.g., the collocation data would span different sections on the horizontal axis). Therefore, the information on error

characteristics from each collocation analyzed via B-A or REU plots would tend to complement each other, as opposed to the simple performance metrics which might seemingly contradict each other. This is just a thought; while in general I agree with the logic underlying the arguments being presented here, I don't think that the examples, as they are currently presented, do a strong enough job of backing up these arguments.

Specific Comments

Line 75: One of the commas here seems misplaced.

Lines 78-79: Might be better stated as "a linear additive model is often assumed".

Figure 1: REU should be defined before it is used in this figure.

Line 103: Should be "data are communicated".

Line 136: Remove "And" at start of sentence.

Line 174: Suggest replacing "data" with "data set".

Lines 306-310: This is background information, better included as part of the introduction, where it can be integrated with similar statements already there.

Lines 342-344: This is an important point, often used as justification for the use of LCS for applications like hotspot identification. I wonder if the authors could comment more on this, either here or elsewhere. My prompting question would be: what kinds of analysis approaches could be used to verify the ability of LCS to qualitatively identify meaningful differences between measurements, even in situations where relative uncertainties are too high to make reliable quantitative comparisons? Alternatively, is such a distinction (qualitative versus quantitative analysis) meaningful here, or is this "qualitative analysis" merely a quantitative analysis performed under higher relative uncertainty.

Line 367: Second "at" is superfluous.

Line 370: “deviations” should be “deviation”.

Line 372: “appears” should be “appear”.

Lines 376-378: This is another important point. Since air quality regulations are based on these reference instruments, the traceability of LCS to these reference instruments has been a major focus of work. However, we must acknowledge that these references themselves are imperfect. Is it thus inappropriate to hold LCS to certain performance standards which the reference instruments themselves may not meet (especially if improperly operated)? On the other hand, what is the alternative to ensuring data quality? I think that, as you suggest, comparing different reference instruments among themselves should be done more frequently, and these intercomparisons more widely used as a benchmark against which the performance of LCS can be judged (instead of establishing arbitrary performance metric targets, espeically if these targets are not connected in some way to the different conditions under which the sensors are expected to operate). However, there is of course the practical question of the cost and feasibility of doing this at the necessary scale. Generally speaking, this is a major point which could be explored further by the authors either here or elsewhere.

Line 443: “data is” should be “data are”.

Lines 450-453: The meaning of this sentence is unclear; consider breaking it into several simpler sentences.