

#### General Comments:

This is a straightforward study that looks at interlaboratory comparisons of eight different PM10 leaching protocols for trace element analysis. There was no effort to standardize methods across the participating labs, which often had slightly different protocols. Because the samples are subsamples of field-collected filters, the authors cannot evaluate the extent to which variability and/or lack of agreement between labs could have been a result of heterogeneity in PM10 collection across a single filter (distribution of certified reference material PM10 could have been a better approach). Given these factors, not surprisingly, the interlaboratory comparisons sometime showed good agreement, and sometimes not. The effort to link differences in element solubilization to aerosols with different transport histories is somewhat tangential, and the arguments are compromised by a small sample size. The conclusion that the adoption of best practice guidance on analytical protocols makes sense, but would have made sense even without the data presented in this paper.

#### Specific Comments:

Lines 128-135 – It is unclear what the 6 samples are. Were there 2 samples each for the 11.5, 23.5, and 35.5 hour collection times?

Line 267 – Does each dataset refer to the data for each individual element?

Lines 271-285 – Several suggestions about statistical analysis. As written, it appears that the approach was to use every possible approach and see what falls out. I'd suggest deleting the correlations and the Wilcoxon Signed Ranks tests. You are not really interested in whether two methods are correlated. What you want to know is – is the slope different from 1 and the intercept different from zero? As there is no clear independent and dependent variable here, these should be Model II regressions. Why use  $1 \pm 0.12$  when you could directly test (t-test) whether a slope is significantly different from 1.0?

Line 314 – Please define what is meant by analytical method variability. Weren't these samples all run using the same analytical method in one lab? I do not see how applying the largest median relative MAD value to all samples disentangles subsample variability from analytical method variability.

Lines 367-374 – Please clarify. Were these regressions run on the mean values from the two labs? Same for similar Figures.

Line 449 – What is meant by diverging calibration methods?

Line 467 – Figure 6 shows box and whiskers plots, so Kruskal-Wallis is appropriate. Suggest deleting "and one-way ANOVA" (lines 480 and 503, as well). Multiple comparisons results could be added to Figure 6.

Line 519 – As there are three groups, why not use a Kruskal-Wallis test instead of pairwise Mann-Whitney U tests?

Technical Corrections:

Lines 61-62 – Please rephrase. Unclear what “they” refers to.

Lines 73-74 – Suggest removing the quotation marks.

Line 74 – Change have to has.

Line 77 – Change suffers to suffer.

Line 102 – Change insignificant to nonsignificant.

Line 104 – Change which to that.

Lines 111-112 – To what does “they” refer?

Line 127 – Please use metric units.

Line 176 – I think you mean triplicate subsamples, right?

Line 182 – Change were to was.

Line 213 – Change which to that.