Comments by referees are in blue.
Our replies are in black.
Changes to the manuscript are highlighted in red both here and in the revised manuscript.

Reply to referee #1

General Comments:

This is a straightforward study that looks at interlaboratory comparisons of eight different PM_{10} leaching protocols for trace element analysis. There was no effort to standardize methods across the participating labs, which often had slightly different protocols.

Reply: We would like to thank ref #1 for reviewing our manuscript and recommending it for publication after major revision. We have addressed these comments and revised the manuscript accordingly. Please find more details below.

The aim of this study was to assess the range of data generated using the leaching protocols which currently exist in the literature. To do so, we, indeed, had to produce and compare data obtained using 8 commonly used leaching protocols, without attempt to modify or standardize the protocols. The slight differences in protocols were sometimes shown to result in significant variability in the resulting data, while other times, the differences in resulting data were negligible. This study suggests that the use of some specific partices in aerosol leaching experiments (for example, sonication) should be avoided, and highlights the need for using a backing filter for consistent data comparison. This information is key to future studies comparing newly produced to literature data.

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 PM_{10} collection across a single filter (distribution of certified reference material PM_{10} could have been a better approach).

Reply: We do not entirely agree with this comment. We have made considerable effort to assess the heterogeneity of the aerosol samples used in our study (as described in Section 3.1). As a result of this work, we have shown that the differences between groups' results are clearly too large to be due to heterogeneity of the samples alone (for example, through the slopes of method – method regressions being significantly different to 1±0.12, Figure 2).

The reviewer is correct that a comparison based on the distribution of a reference material would have avoided the complicating influence of sample heterogeneity and would have been easier to interpret, and this was already acknowledged in the original manuscript (page 31, line 551-554; page 32, lines 587-591). However, a large part of our motivation for this study was to establish the differences obtained by the various analytical methods for ambient aerosols. This information is vital for the interpretation of aerosol soluble trace element concentrations from diverse sources and for appropriate use of such observations in the validation and interpretation of numerical models.

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.

Reply: As stated above, we consider the behaviour of ambient aerosols to be a central part of our study and we do not agree that differences between samples with different source and transport histories are "tangential". The sample size for the SW group is indeed small, but we have applied appropriate statistical tests to the data and only reported differences where they were statistically significant (Figure 7). The comparison between the two larger groups (N and

YS) showed statistically significant differences in every case where Kruskal Wallis tests indicated that differences were present between the groups.

We agree that further work is needed on this topic (as we already mentioned in Section 4 of the original manuscript), but we do not agree that this aspect of our results is compromised by the relatively small sample size (our sample size is equivalent to or greater than most studies using shipboard collected aerosols).

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.

Reply: The reviewer is correct to some extent; on the other hand, after inter-comparison of different leaching protocols using experimental data, our recomendation is more convincing. We highlight which specific protocol steps may lead to larger variability or increased inconsistency compared to other protocols (hence which steps may be better avoided when developing best practices). Our data is also key for modellers to better understand the field data they (should or should not) use to validate their model.

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?

Reply: The six samples were used to test particle distribution homogeneity. In the revised manuscript (page 6) we have made the following change to provide further explanation: “Six PM_{10} samples were collected at an urban site (113°36’E, 23°13’N) in Guangzhou from 24 November to 01 December 2021 to test aerosol trace element distribution on the filters (Table 1).” In addition, to help readers better understand our work, in the revised manuscript (page 6) we have also added one table (Table 1) to summarize the aerosol samples collected in this study.

We varied the collection times to intentionally vary the amount of particles collected over a large range. In the revised manuscript (page 6) we have made the following change to make it clearer: “…lasted for 11.5, 23.5 or 35.5 hours to intentionally vary the amount of aerosol particles collected over a large range.”

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

Reply: Ref #1 is correct. In the revised manuscript (page 13) we have made the following change to make it clearer: “…each dataset for each element and method was tested for normality using…”

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 ± 0.12 when you could directly test (t-test) whether a slope is significantly different from 1.0?

Reply: Ref #1 raised a few comments here, and we have addressed them one after another, as detailed below. There were specific reasons for our choice of statistical tests and these were explained in Section 2.6.3.

(1) We do not quite agree with ref #1 that we are not really interested in whether two methods are correlated. If fact, only when the two methods are correlated does it makes sense to test the slope and intercept of the method-method regression.

(2) We agree with ref #1 that we should use Model II regression analysis. We have conducted orthogonal distance regression (ODR) analysis, and have updated the manuscript and SI accordingly:

(2.1) In the revised manuscript (page 14), we have removed Pearsons' correlation and added one sentence to explain why we chose to use ODR: "A two-tailed t-test was used to test the slope of a method-method relationship. Method-method slopes and intercepts were determined using orthogonal distance regression (ODR), since both analytical parameters were subject to significant uncertainty and simple linear regression was therefore not suitable."

(2.2) With the results obtained using ODR analysis, we have updated Figure 2 in the revised manuscript (page 20) and Figures S4-S6 in the supplement. Using ODR analysis does not lead to significant change in our results/conclusions, and at in some cases it produces results which better support our conclusions. Changes to the text can be found at a few places in the revised manuscript: (i) Cu has been removed from the list of elements for slope differs significantly from 1±0.12, as the slope is different from 1±0.12 only for Al and V (not for Cu anymore) (page 21); (ii) "For all trace elements except Pb, the comparison of these two methods, however, resulted in a slope different from 1±0.12 and an intercept different from 0 for Ni and V (Figure 2)…" (page 23); (iii) "For Al, Cu and Mn, measurements show excellent agreement (significant correlation ($p<0.01$), and no significant differences in slopes, intercepts or soluble masses) for the two AmmAc methods (Figures 2 and 5). The other elements (Fe, Ni, Pb and V) also show good agreement between methods, with no significant differences for slopes and only Ni having a significant difference for intercept, possibly due to…" (page 25).

(3) As we explained in our original manuscript (page 14, line 277-280), in order to take in account sample distribution heterogeneity, we should test whether the slope is different from 1 ± 0.12 (instead of 1).

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.

Reply: In this work we chose a conservative approach by applying the highest median relative MAD value (i.e. 12%) to represent the uncertainty associated with trace element distribution heterogeneity over a filter, regardless of the element analyzed. In the revised manuscript (page 16-17) we have made the following changes to provide further explanation: "In this work we chose a conservative approach by applying the highest median relative MAD value (i.e. 12%) to represent the uncertainty associated with trace element distribution heterogeneity over a filter, regardless of the element analyzed. Variations between different methods greater than this 12% heterogeneity are likely to be due to differences in analytical results, as described in Section 2.6.3."

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

Reply: Figure 2 summarizes statistical analysis (mainly regression analysis) results for two given labs and for individual elements. Regression analysis cannot be run on two mean values, and we have provided the sample size (n = 26) in the caption of Figure 2. Figure 2 is the only figure to present regression results.

Line 449 – What is meant by diverging calibration methods?

Reply: In the revised manuscript (page 25) we have made the following changes to make it clearer: "These four elements (Fe, Ni, Pb and V) do show significant differences for soluble masses, possibly due to differences in the calibration methods or other analytical differences."

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.

Reply: We have conducted Kruskal Wallis and one-way ANOVA tests, both giving the same results; therefore, we prefer not to delete “one-way ANOVA”. We are not sure what “multiple comparisons results could be added to Figure 6” means, but Figure 6 already includes comparisons of multiple elements and of multiple leaching protocols.

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

Reply: As stated in the caption of Figure 7, Kruskal Wallis tests were used to establish whether there were significant differences with the three air mass types for each element and method combination. Pairwise Mann Whitney tests were only applied when appropriate according to the result of the Kruskal Wallis test.

Technical Corrections:

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

Reply: In the revised manuscript (page 3) we have changed “they” to “anthropogenic emissions”.

Lines 73-74 – Suggest removing the quotation marks.

Reply: Here we use quotation marks here deliberately, because the words are quotations.

Line 74 – Change have to has.

Reply: It is correct to use “have” instead of “has”, and thus in the revised manuscript we have not made any changes.

Line 77 – Change suffers to suffer.

Reply: It is correct to use “suffers” instead of “suffer”, and thus in the revised manuscript we have not made any changes.

Line 102 – Change insignificant to nonsignificant.

Reply: As suggested, in the revised manuscript (page 5) we have changed “insignficant” to “non-significant”.

Line 104 – Change which to that.

Reply: As suggested, in the revised manuscript (page 5) we have changed “which” to “that”.

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

Reply: In the revised manuscript (page 5) we have changed “they” to “These protocols”.

Line 127 – Please use metric units.

Reply: As suggested, in the revised manuscript (page 6) we have changed it to “229 mm × 305 mm”.

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

Reply: Ref #1 is right. In the revised manuscript (page 8) we have made the following change to make it clearer: “Each lab was provided with triplicate subsamples of each C filter.”

Line 182 – Change were to was.

Reply: It is correct to use “were” instead of “was”, and thus in the revised manuscript we have not made any changes.

Line 213 – Change which to that.

Reply: As suggested, in the revised manuscript (page 11) we have changed “which” to “that”.