General review:

This manuscript compares different measuring techniques for aircraft gas-turbine-emitted nvPM mass number and size, and total number and size from exhaust sampled at 43m of a V2527-A5 and a CFM56-2C1 aircraft engine burning a range of sustainable and conventional aviation fuels as part of the ECLIF 2 test campaign.

The manuscript is well written, and the data presented is novel and relevant to the scientific community. The data processing is of good quality; however, I found that the interpretation of the consistency between nvPM number and mass emission indices was not entirely addressed and sometimes misleading, particularly for mass and size. Please note that I wasn't able to access the supplementary information and table S1 which may answer some of my comments.

Major comments:

- **Loss correction:**
    - Penetration function 1 measurement: Would you be able to add the size distribution characteristics of the nebulised ammonium sulfate (it could only be in the SI)? Was it representative and the size distributions you typically measured during the test campaign (i.e., GMD ~20-40 nm)?

    - Penetration function 2: I am unsure of what you mean here. Do you mean you used the UTRC model to predict size-dependent losses in this section (as can be seen in Figure 4) or did you use the full N/M method that outputs a correction factor for nvPM number and mass? If you only used the UTRC model, which particle size distribution did you use? If you used the N/M method, how did you correct for losses to the DMS-500? The loss correction methodology you used would affect the interpretation of your results, and therefore it should be clearly explained.

- **Figure 5:**
    - 40% N1 with REF4 graph: It appears the SMPS+TD also measures part of a peak < 10 nm, which appears to be volatile given the SMPS+CS doesn't see it. Does that mean the TD is not 100% efficient at removing volatiles? This should be discussed. Also in line 49, you discuss that an nvPM mode < 10 nm was observed with the CFM56 engine. Can you confirm that it was a nvPM mode and not a volatile mode?

    - On multiple PSDs, the start of a large size mode can be observed (~200 nm) which could indicate you were measuring shed particles (unless it came from the engine) or oil. Were you performing regular cleanliness checks? Did you use in-line cyclones to all your analysers? If this was shedding, it could significantly impact nvPM mass measurements < 10 ug/m3 and would affect EIm estimation from integrated particle size measurement. This is not discussed anywhere in the manuscript; hence discussion regarding this should be added.

- **Particle size statistics; GMD and GSD**: It is not clear whether you compared the measured particle size distributions or if they were corrected for particle loss to a common sampling point (plenum or probe) using a bin-by-bin approach with measured PSD and the penetration functions (or another method?). Please clarify this in the main text and in the figure titles.

- **Section 4.3: Consistency between number-based emission indices of nvPM and vPM**
    - APC Vs SMPS number: Have you considered that both SMPSs were just under-reporting due to the large corrections performed within the SMPS software (losses in the DMA, the poor charging efficiency of the bipolar charger, non-linearity of the CPC response)? When were the SMPSs last serviced and calibrated? Were the SMPSs compared with the APC on the same source prior to the test campaign? As currently written, this section implies that SMPSs are more precise at measuring nvPM number than the standard regulatory compliant APC. I don't agree with your conclusion that the APC was likely overcorrected only because you found the two SMPSs to agree with each-other. What if there was two APC in agreement and only one SMPS?

    - L560: The DMS-500 was measuring unstripped aerosol and therefore could be picking up volatiles in comparison with the other nvPM EInum analysers, which could also explain why it was reporting higher values.

    - Were all the size/number analysers within the recommended 12 months service and calibration period? If not, that could explain some of the disparities observed between the different size analysers. For example, the DMS-500 is calibrated for number and size to a traceable standard but drifts over time and Cambustion only certifies measurement precision of 10% for size and 20% for number within 12-months. I suggest you add discussion on the calibration uncertainty associated with all analysers.

    - Figure 7: It is biased that you only used the SMPSs to calculate the mean and then you compared that mean to the DMS GMD and GSD given it wasn't included in the mean calculation. Why didn't you apply the same methodology as for the nvPM mass analyser, calculating the mean using all the different analysers?

- **Section 4.4: Consistency between mass-based emission indices**
    - Scatter < 100 mg/kg fuel: Are you sure this reflects the noise levels of the instruments? Analysers like the LII and MSS are, to my knowledge, capable of precisely measuring down to 1 ug/m3. Is the difference between the LIIs bias or scatter? Did you consider shedding from your system could impact nvPM mass measurement as mentioned in a comment above or that potentially inaccurate calibration caused this difference (see comment below)? Please discuss this in the manuscript.

    - I would expect the scatter between the three LIIs to be lower than reported given they are the same analyser, particularly the two NRC LIIs given they are next to each-other and presumably calibrated in the same manner. This is not addressed or discussed in the text, which is surprising given the detail that goes into the fluence sector. Could it have to do with the calibration performed for these analysers? Did the laboratory diffusion flame show ICAO annex16 applicability? If not, it could well explain some of the scatter you observe for EIm. This is something worth discussing in the manuscript.

  - SMPS based EIm: I find the interpretation of this section misleading, as it suggests the SMPS is nearly as good as an LII or MSS at measuring nvPM mass. First, the SMPSs generally do not capture the full VSD (as can be seen in figure 5). Secondly, you assumed unit density for the CS-SMPS but the particle density could well be below 1 g/cm3, particularly given particle effective density decreases with increasing size and the mass is carried by the larger particles. Using integrated particle size measurement to derive mass is strongly influenced by the density you select and should be discussed. Furthermore, I disagree with L613-L618; I believe the main reasons for the higher SMPS predictions is density assumption, the measurement uncertainty for the size bins >100 nm where the number count is very low, and potential shedding interference (see comments above). It is also surprising to me that in all figure 5b there is excellent agreement between the 2 stripped SMPS VSD's however it is then observed in figure 8b that there is no agreement. I am unsure how this is consistent if the same assumption regarding density is made?

  - L645: if the TAP and PSAP require a filter change at each test point to operate optimally, doesn't it make them not suitable for aircraft nvPM mass measurement? Particularly given the mass loading you've experienced were typically lower than certification measurements as you were sampling 43m downstream of the engine. If that's the case, then I suggest re-writing the abstract and conclusion to highlight this.

- **Conclusion**: I suggest re-writing the conclusion considering the comments above. For example, L574, I disagree that 10 ug/m3 is the noise level of the instruments. I suggest replacing "instrument" by "instrument calibration and sampling methodology".

Minor comments:

- General: The manuscript could do with more cross-referencing for the reader to find information more easily.

- L39 & L169: replace "sampling" by "measurement"

- L117: add "minimum" before 50%.

- L122: I don't believe it's true that the APC, MSS and LII are the only commercial instruments that satisfy the SARP. For example, a Dekati DEED and a Grimm or TSI CPC is a commercial system that satisfies the SARP. Please clarify.

- L152 & L183: I don't think you can reference something that hasn't been published yet or that that hasn't been peer reviewed yet. As a reviewer to this paper it is hard to critically appraise the statements and conclusions without being able to see the detailed experimental set-up and graphs addressing fuel effects etc.

- L219: Why was the plenum only maintained at 33°C? It seems odd to me that you first sampled via a 60°C heated line, then a 33°C plenum and then other 60°C heated lines. This would promote thermophoretic loss (although very small) and could cause water to condense. Please justify.

- L230: Why did you use a 25m line between the Dekati diluter and the NARS instruments? Was it because you couldn't get container 2 any closer to container 1? Wouldn't it have been better to use a shorter line to minimise diffusional losses and reduce your loss correction uncertainty which accounts for some of the discrepancies observed in your data? Given the 4:1 dilution ratio and sampling position the NARS system was not in compliance anyway hence could have been further optimised.

- L244: Please quantify what you mean by "good agreement".

- L255: The DMS-500 measures from 5 nm not 10 nm. Also, what is the size range of the EEPS?

- L317: drift not drifted.

- L371: Typo CO_2 not CO2

- L376: remove "from".

- L518: What inversion matrix was used for processing the DMS-500 data? Please add to manuscript.

- L550: It's not true the SMPS measures at 10 nm with a 100% efficiency as lots of corrections are applied (charging efficiency, loss through DMA and tubing). Please add "corrected" before 100%.

- Figure 7: The x-axis is labelled "mean nvPM xxx", however some total PM is also shown (as labelled in the legend). Please clarify and correct x-axis.

- L783: therefore not therefor