General comments

This manuscript provides a first characterization of the CARIBIC-AMS, an airborne instrument based on commercial Aerodyne mini-AMS that is part of the IAGOS-CARIBIC research payload. Unlike other airborne particle mass spectrometers that can rely on human input during or at least after each flight, this particular instrument is required to operate in a fully autonomous manner over the course of several flights. Given that IAGOS-CARIBIC operates exclusively on commercial airliners, the reliability of power and communications is typically a lot worse than on most research platforms. While the instrument testing reported here was cut short by COVID, this work does report successful operation for a good portion of that testing period (36 flights, plus a recent campaign in 2024 on a different platform, which is unambiguously a great achievement, for which I want to congratulate the authors

As noted above, this is not a new instrument, but the adaptation of a commercial instrument that has/will operate on a long-term mission to sample the global UT/LS (and some random snippets of the lower troposphere). Hence for a characterization paper like this I would expect it to describe, in addition to the actual modifications, both the operations and analytical performance in enough detail to serve as a guide for future data users/modelers to use the data properly and have a handle on the uncertainties. This paper is fairly detailed on both the motivation and modifications but has a few gaps when describing the operations (the reader learns a great deal on the general CARIBIC schedule of operations, but rather less on the specifics of the CARIBIC-AMS operation). However, it does not describe the analytical performance. While the pre-deployment calibrations are described in detail, no field calibrations are shown. And in lieu of instrumental validation by multiflight instrumental comparisons and performance metrics, the paper chooses to narratively highlight the sulfate measurement on two flight segments. Detection limits are shown and mentioned in the abstract, but it is unclear if these are specific to the single flight segment shown (so "best of") or an actual average of the usable flight segments.

The authors explicitly write in the conclusions that the "metrology paper" will be next. And I certainly agree that it makes sense to leave most of these details for a paper written on the new version of the instrument that is going to start flying shortly on the new CARIBIC platform. Nevertheless, in this manuscript some metrics need to be shown to support that despite all the operational challenges, the instrument actually quantifies PM consistently. Showing, as the authors do in Section 3, examples of AMS sulfate tracking stratospheric markers does not really do that, it just shows that the instrument recorded data and there were no inlet leaks. What needs to be shown is a) how well did the CARIBIC-AMS data agree with collocated instruments and b) are there systematic biases due to the way it is operated (e.g. does the instrument do worse or better on Flight #6 vs Flight #1 of a sequence?). This does not need to be shown for every possible metric (that's paper #2); e.g. just focusing on the (mentioned in the text) agreement between the AMS and the filter sampler would likely be sufficient. But without that, it is impossible to evaluate if the modifications were successful in terms of making the CARIBIC-AMS a quantitative instrument.

I also find some of the AMS specific details to be confusing, and hope that the authors can clarify them, given that this paper is probably going to be read widely.

This paper describes the first deployment of an important new analytical resource to investigate the UT/LS, and therefore is certainly a good fit for AMT.

Major comments:

- Characterization of the particle range sampled by the instrument: For any model/measurement or two instrument comparison, a well characterized operational particle range is key. For the CARIBIC-AMS, that would be a product of the aircraft inlet transmission, the transmission of transfer plumbing and of the AMS inlet (both their custom CPI and the PM1 lens). For the aircraft inlet, the authors reference Herrman et al 2001 and Brenninkmeijer 2007, which given the submicron focus of the instrument is perfectly adequate. However, there is no characterization of either plumbing losses or AMS inlet losses. It is mentioned in L128 that the plumbing losses were calculated, but no results are shown. Regarding the AMS inlet losses:
  - For the actual AMS lens, Liu et al, 2007 is referenced. As discussed in e.g. Knote et al, 2011, there are significant discrepancies both at the high and low end of the curve between Liu et al, 2007 and other reports of AMS transmission. Given that the accumulation mode in the stratosphere goes well above the PM_1 lens cut (e.g. Brock et al, 2021, Fig. 11), while in the upper troposphere the sampling of particle growth events on the low end of the transmission range is expected, both of these uncertainties will matter for accurate quantification. And they are somewhat dependent on instrument-specific parameters (quality of the alignment, lens vintage, lens pressure), hence an instrument specific characterization is needed, or , in the short term, at least some spot-check confirmation that the actual transmission does indeed follow one of the reference curves.
  - Then there is the issue of the relatively novel CPI used in this work. While Molleker et al 2020 does characterize the transmission performance of the $CPI+PM_{2.5}$ lens in detail, I am not aware of a similar characterization for this $CPI+PM_1$ lens, so the exact impact of the CPI on the lens transmission above is unclear. The authors do report that the comparisons in Mei et al, 2020 suggest serious losses at low altitudes. While I agree that this is unlikely to matter for the portions of the atmosphere sampled by the CARIBIC-AMS, it does (as the authors acknowledge) matter for understanding the sensitivity calibrations on the ground. The authors write: "For this purpose, the mass-based IE calibration as described above is done at various pressures and a pressure-dependent IE is used for the data evaluation" (L287-289). Since said calibration is done at a constant calibrant size, this seems to imply that the authors assume the losses in the CPI at high pressure to be size-independent. This is to me a counterintuitive assumption

that needs more support/explanation, since it's central to the instrument's performance.

  - Again, while these might seem like minor points, if we take Fig 11-l in Brock et al, 2021, there is a ~40% difference in accumulation sulfate mass if one were to apply the Liu vs Knote lens transmission. This uncertainty is not reflected in the 35% accuracy estimate by Bahreini et al, 2009, is dependent on the specific airmass sampled and hence should be treated (and discussed) separately.
    If none of this calibration data is available, then at least the effect of the various literature transmission curves and the "pressure dependent IE" on the final reported concentration needs to be discussed when compared to other instruments.

- Instrumental Collection Efficiency: The current manuscript does not mention collection efficiency (CE) (Canagaratna et al, 2007) at all, which for an AMS-style instrument sampling the free troposphere is a strange omission. This might be warranted if the instrument had a capture vaporizer (Xu et al, 2017, Hu et al, 2017) that mostly obviates the need for a particle bounce correction factor. Now, the authors do not specify what vaporizer was used (please add this!) but both Fig 1 and the mention of a "tungsten vaporizer" strongly implies that this is a standard vaporizer as described in Canagaratna et al 2007. Hence a bounce correction (Middlebrook et al, 2012) is needed for this instrument (and for all the data presented in Section 3) and should be described and discussed. The authors might also want to explain in this context the reasons for that particular vaporizer choice, which seems counterintuitive to me.
  One item I would like the authors to specifically address in this context is how the high ammonium detection limits of the CARIBIC-AMS will affect the accuracy of the CE correction at different altitudes. While averaging can be used to improve the CE accuracy in the stratosphere, this is probably not the case during ascent/descent. If so, what are the consequences for data reporting?
- Sensitivity Calibrations: Section 2.3.2. summarizes some pre-deployment calibrations. Fig 5 and 6a certainly have value in terms of showing that the instrument post reconfiguration was working well. But as noted, for a field instrument it would be much more valuable to show the timeseries (or set of regressions in Fig 5) for the in-field calibrations between 2018 and 2020 to assess instrument stability (particularly the calibration pairs before/after one set of flights). If the testing was too rough on the instrument, this might not be possible, but again, with 36 flights there should be some good data to show.
  The purpose of Fig 6b is not clear. First of all, the uncertainty bands shown seem incorrect. This is a test of RIE fidelity, and the uncertainty for those is 15% (2 sigma) per Bahreini et al, not 35% (which is the total uncertainty). Secondly, previous work has shown no such discrepancy for mixtures (e.g. Jimenez et al, 2016, Xu et al, 2018). Hence the reasons for this disagreement, if real, need to be explained better. There could be issues with different transmission efficiencies for both instruments, different choices in

acquisition cycle, or something else entirely. I think it could also be removed, since it does not really provide much context on the overall instrument performance (a correlation/dual timeseries of ambient measurements on the ground might in any case be more appropiate).

- Size resolved measurements and AMS operation. The description of the ePToF custom implementation and subsequent calibrations is good, although I would suggest to the authors to consider using an alternative figure that shows the (significant) increase in S/N and resolution a bit more convincingly. Also, it should probably be noted somewhere that the theoretical resolution is 1/127 (Hadamard sequence length). However, as the paper is written it is completely unclear if any size dependent measurements were taken in flight. The larger point is that there could be a bit more detail on exactly how the AMS acquisition is run. The authors write that the typical MS acquisition is 30 seconds long, but it is unclear how many open/closed cycles this encompasses and what the effective dutycycle for sampling is. When are the ePToF runs scheduled? And does the instrument operate on a fixed time basis or not?
- Detection limits: The estimation of the detection limits from the closed signal seems very reasonable, but since the instrument takes automated blanks it would be good to show a comparison of the calculated DLs with the ones derived from the statistics of the blanks. More generally, while it is great that the authors reported these DLs, as noted above it is very unclear for what conditions these are specified Do these only apply to Flight 508? I think it would be a lot more informative to:
    a) Report average DLs for the stable stratospheric periods of all available flights
    b) Show the change of these DLs for a range of flights. Ideally other species besides sulfate should be shown, since e.g. ammonium and OA are much more likely to greatly improve with additional pumping over the course of the flight/circuit.

  For the abstract, again it would be good to specify what these DLs refer to.It might be simpler and easier to follow to just write that the DLs will scale with 1/sqrt(N) for longer averaging times, instead of quoting 5 min DLs
- Section 3: Validation. Section 3.1 puts the measurements in an appropriate context and helps the reader re-digest the information from the Ops section, but I would suggest adding some AMS context from my previous comments (e.g. discussion of CE and particle size range). Section 3.2 on the other hand does not really introduce anything new. Instead, as noted in the general comments, this would be the place where some instrument comparisons over several flights are shown and discussed (either filters or volume size distributions). Given the teething problems of the instrument, this data is likely going to be noisy and will suggest low accuracies that will hopefully soon be superseded by version 2 of the instrument. This is understandable and can be properly contextualized in the text, but the data should still be shown.

Minor comments:

- Single Ion Calibration: This is a minor point, but I am not following the need for periodic single ion calibrations during flight. The single ion calibration at this point (post-2014) just supplies a scaling factor (the SI) to convert signal to counts per second. Historically, this calibration also included the spectral baseline fit, but that has been automated since 2016 and is taken on a per-run basis. Now, the in-flight SI is needed to scale it to the SI of the sensitivity calibration. BUT once the instrument is running and the SI has been characterized once, the variability of the SI is fully captured by the airbeam variability (especially in an instrument with a well-working CPI). So while the ability to do this calibration on the fly e.g. in case of in-flight instrument reboot is important and impressive to have implemented, I would really appreciate it if the authors could explain their rationale for why periodic SI calibrations are needed. As a side note, it would be nice to see a figure with either the airbeam or SI change for a couple of flights.
- Figure 3: It would be very helpful if the different operation states described in the text are added to the figure (maybe as shaded bars). Also, please clarify the units for the Y-axis. Are these A at 24 V? It might be clearer to use W(atts) here.
- Section 2.2.3 is not a straightforward read, since it tries to combine both topological and process details into one narrative. One suggestion to make it more accessible is to add a simple diagram showing the network/software topology/hierarchy and referring to it in the text.
- The units used in the paper are μg m^{-3} STP. STP is not defined anywhere, and since the atmospheric community is not exactly known for its general adherence to SI units I would strongly urge the authors to spell out what they mean in the text.
- Line 19: Instead of "part" maybe "module" would read better?
- Line 25: "Due to **the** short time"
- Line 86: I would recommend adding the mission description paper for Atom here as well for completeness:

  C. R. Thompson, S. C. Wofsy, M. J. Prather, P. A. Newman, T. F. Hanisco, T. B. Ryerson, D. W. Fahey, E. C. Apel, C. A. Brock, W. H. Brune, K. Froyd, J. M. Katich, J. M. Nicely, J. Peischl, E. Ray, P. R. Veres, S. Wang, H. M. Allen, E. Asher, H. Bian, D. Blake, I. Bourgeois, J. Budney, T. Paul Bui, A. Butler, P. Campuzano-Jost, C. Chang, M. Chin, R. Commane, G. Correa, J. D. Crounse, B. Daube, J. E. Dibb, J. P. Digangi, G. S. Diskin, M. Dollner, J. W. Elkins, A. M. Fiore, C. M. Flynn, H. Guo, S. R. Hall, R. A. Hannun, A. Hills, E. J. Hintsa, A. Hodzic, R. S. Hornbrook, L. Greg Huey, J. L. Jimenez, R. F. Keeling, M. J. Kim, A. Kupc, F. Lacey, L. R. Lait, J.-F. Lamarque, J. Liu, K. Mckain, S. Meinardi, D. O. Miller, S. A. Montzka, F. L. Moore, E. J. Morgan, D. M. Murphy, L. T. Murray, B. A. Nault, J. Andrew Neuman, L. Nguyen, Y. Gonzalez, A. Rollins, K. Rosenlof, M. Sargent, G. Schill, J. P. Schwarz, J. M. St. Clair, S. D. Steenrod, B. B. Stephens, S. E. Strahan, S. A. Strode, C. Sweeney, A. B. Thames, K. Ullmann, N. Wagner, R. Weber, B. Weinzierl, P. O. Wennberg, C. J. Williamson, G. M. Wolfe, L. Zeng, THE NASA ATMOSPHERIC TOMOGRAPHY (ATom) MISSION: Imaging the Chemistry of the Global Atmosphere. *Bull. Am. Meteorol. Soc.* **1**, 1–53 (2021).
- L161: "**into** one new housing"

- L244: “Thus, flow calibrations and size calibration will not change, need to be calibrated only once, and later on only have to be checked.”. Agreed. But can you elaborate on how often these checks are done in practice?
- L387: “…than **at** the end of the flight”

References:

P. S. K. Liu, R. Deng, K. A. Smith, L. R. Williams, J. T. Jayne, M. R. Canagaratna, K. Moore, T. B. Onasch, D. R. Worsnop, T. Deshler, Transmission efficiency of an aerodynamic focusing lens system: Comparison of model calculations and laboratory measurements for the Aerodyne Aerosol Mass Spectrometer. *Aerosol Sci. Technol.* **41**, 721–733 (2007).

C. Knote, D. Brunner, H. Vogel, J. Allan, A. Asmi, M. Äijälä, S. Carbone, H. D. van der Gon, J. L. Jimenez, A. Kiendler-Scharr, C. Mohr, L. Poulain, A. S. H. Prévôt, E. Swietlicki, B. Vogel, Towards an online-coupled chemistry-climate model: evaluation of trace gases and aerosols in COSMO-ART. *Geoscientific Model Development* **4**, 1077–1102 (2011).

C. A. Brock, K. D. Froyd, M. Dollner, C. J. Williamson, G. Schill, D. M. Murphy, N. J. Wagner, A. Kupc, J. L. Jimenez, P. Campuzano-Jost, B. A. Nault, J. C. Schroder, D. A. Day, D. J. Price, B. Weinzierl, J. P. Schwarz, J. M. Katich, S. Wang, L. Zeng, R. Weber, J. Dibb, E. Scheuer, G. S. Diskin, J. P. DiGangi, T. Bui, J. M. Dean-Day, C. R. Thompson, J. Peischl, T. B. Ryerson, I. Bourgeois, B. C. Daube, R. Commane, S. C. Wofsy, Ambient aerosol properties in the remote atmosphere from global-scale in situ measurements. *Atmos. Chem. Phys.* **21**, 15023–15063 (2021).

R. Bahreini, B. Ervens, A. M. Middlebrook, C. Warneke, J. A. de Gouw, P. F. DeCarlo, J. L. Jimenez, C. A. Brock, J. A. Neuman, T. B. Ryerson, H. Stark, Atlas, E, J. Brioude, A. Fried, J. S. Holloway, J. Peischl, D. Richter, J. Walega, P. Weibring, A. G. Wollny, F. C. Fehsenfeld, Organic aerosol formation in urban and industrial plumes near Houston and Dallas, Texas. *J. Geophys. Res.* **114**, D00F16-D00F16 (2009).

S. Molleker, F. Helleis, T. Klimach, O. Appel, H.-C. Clemen, A. Dragoneas, C. Gurk, A. Hünig, F. Köllner, F. Rubach, C. Schulz, J. Schneider, S. Borrmann, Application of an O-ring pinch device as a constant-pressure inlet (CPI) for airborne sampling. *Atmospheric Measurement Techniques* **13**, 3651–3660 (2020).

F. Mei, J. Wang, J. M. Comstock, R. Weigel, M. Krämer, C. Mahnke, J. E. Shilling, J. Schneider, C. Schulz, C. N. Long, M. Wendisch, L. A. T. MacHado, B. Schmid, T. Krisna, M. Pekour, J. Hubbe, A. Giez, B. Weinzierl, M. Zoeger, M. L. Pöhlker, H. Schlager, M. A. Cecchini, M. O. Andreae, S. T. Martin, S. S. De Sá, J. Fan, J. Tomlinson, S. Springston, U. Pöschl, P. Artaxo, C. Pöhlker, T. Klimach, A. Minikin, A. Afchine, S. Borrmann, Comparison of aircraft measurements during GoAmazon2014/5 and ACRIDICON-CHUVA. *Atmospheric Measurement Techniques* **13**, 661–684 (2020).

M. R. Canagaratna, J. T. Jayne, J. L. Jimenez, J. D. Allan, M. R. Alfarra, Q. Zhang, T. B. Onasch, F. Drewnick, H. Coe, A. Middlebrook, A. Delia, L. R. Williams, A. M. Trimborn, M. J. Northway, P. F. DeCarlo, C. E. Kolb, P. Davidovits, D. R. Worsnop, Chemical and microphysical characterization of ambient aerosols with the aerodyne aerosol mass spectrometer. *Mass Spectrom. Rev.* **26**, 185–222 (2007).

A. M. Middlebrook, R. Bahreini, J. L. Jimenez, M. R. Canagaratna, Evaluation of Composition-Dependent Collection Efficiencies for the Aerodyne Aerosol Mass Spectrometer using Field Data. *Aerosol Sci. Technol.* **46**, 258–271 (2012).

J. L. Jimenez, M. R. Canagaratna, F. Drewnick, J. D. Allan, M. R. Alfarra, A. M. Middlebrook, J. G. Slowik, Q. Zhang, H. Coe, J. T. Jayne, D. R. Worsnop, Comment on "The effects of molecular weight and thermal decomposition on the sensitivity of a thermal desorption aerosol mass spectrometer." *Aerosol Sci. Technol.* **50**, i–xv (2016).

W. Xu, A. Lambe, P. Silva, W. Hu, T. Onasch, L. Williams, P. Croteau, X. Zhang, L. Renbaum-Wolff, E. Fortner, J. L. Jimenez, J. Jayne, D. Worsnop, M. Canagaratna, Laboratory evaluation of species-dependent relative ionization efficiencies in the Aerodyne Aerosol Mass Spectrometer. *Aerosol Sci. Technol.* **52**, 626–641 (2018).