The authors appreciate the helpful feedback from the reviewers. Reviewer comments are in **black** and responses are in **blue** in the Response to Reviewer document.

Response to reviewer comment #1

This is an excellent and well-prepared paper that presents a valuable intercomparison of airborne formaldehyde (HCHO) measurements from four instruments (TOGA-TOF, CAMS, PTR-ToF, and ISAF) during the FIREX-AQ campaign. The study is well designed and supported by a large dataset of co-located observations, allowing robust statistical analysis across a broad concentration range. The close agreement between TOGA-TOF, CAMS, and PTR-ToF confirms that TOGA-TOF provides what the author's describe as reliable and accurate in situ HCHO data.

The TOGA-TOF measurements seem especially useful because it can measure both HCHO plus a large suite of precursor VOCs. As they are on the same instrument this can eliminate some of the sources of instrumental and sampling error. It would be nice if some recommendations could be made with regards to maintaining sample integrity and calibration best practise.

The authors' use of volume-weighted averaging demonstrates careful consideration of the different temporal resolutions among the instruments, rather than simple averaging. This approach can be considered for all instruments taking discrete samples, such as a sample pre-concentration or quite commonly, the capture of whole air samples (where fill rate is not constant across the entire sampling period) and is a useful demonstration of this type of analysis.

The manuscript is clearly written, easy to follow, and provides an important contribution to airborne trace gas measurement science. The results are directly relevant to ongoing efforts to improve instrument consistency and model–measurement comparison

Specific Comments

1. Calibration methodology

Different instruments have different inlets, pumping speeds, pressures etc. Can the authors suggest what might have an effect on HCHO sampling integrity? Is there an understanding of the favourable parameters from each instrument and what would be the optimal sampling strategy for HCHO?

The sampling strategies for each instrument can be quite different. The most important thing is to efficiently move the air sample (containing formaldehyde) to the analytical system quickly and without losses or delays during transfer.

We have found that the best tubing to use for sampling OVOCs including formaldehyde is Sulfinert^{TM}-treated stainless steel (SS) tubing (Restek Corporation, Bellafonte, PA). A near-equivalent product (we believe) is Silonite^{TM}-treated SS tubing (Entech Instruments, Simi Valley, CA), but we have less experience with this product.

Each of these products has an inert silicon coating deposited on the inside tubing surface to prevent surface adsorption of volatile compounds. Deming et al. (2019, https://doi.org/10.5194/amt-12-3453-2019) have demonstrated the efficacy of Silonite™-treated tubing for minimizing surface adsorption of VOCs. Eliminating or minimizing surface adsorption is important as retained organic material on surfaces can cause artifact formation when exposed to oxidants such as ozone.

The TOGA inlet system is a relatively simple custom-built design that was mounted on the DC-8 window plate. Ram air flow is transmitted through a highly electropolished inlet tube that is collinear with the DC-8 fuselage and supported via an orthogonal inlet strut beyond the aircraft boundary layer. The sample is drawn into the instrument through a heated ¼" Sulfinert™ line (40°C) at 2.5 LPM contained within the strut. During ASIA-AQ, the residence time in the inlet was 0.5 seconds.

The authors note that differences among instruments (TOGA-TOF, CAMS, ISAF, PTR-ToF) may stem partly from different calibration approaches and standards. What are the pros and cons of the different methods and is there a preferred technique? Is there one instrumental technique that shows the best accuracy, or are certain techniques limited by practicalities?

Calibrating for formaldehyde is nontrivial and has several aspects to it that need to be considered. Recent progress in producing stable mixtures in aluminum cylinders at high pressure has resulted in this being the most common calibration source among airborne research scientists. All of the instruments in the paper were calibrated this way. Henry’s Law and permeation tubes have also been used to calibrate for formaldehyde but were not used here. The mixtures can be prepared gravimetrically - and were here - but it is imperative that the mixtures be analyzed, preferably with multiple techniques, to ensure stability and to have confidence in the concentration prior to calibrating the instrument. The TOGA-TOF instrument is typically calibrated both on-line during flights for multiple VOCs/OVOCs, etc. and also pre- and post-mission in the lab. An additional challenge when calibrating for HCHO from a high-pressure cylinder is that the regulator and lines have to be highly passivated and this is best accomplished in the lab. For this reason, we didn't calibrate on-line for HCHO during flights but did pre- and post-mission calibrations. ISAF and the PTR-TOF also did not employ on-line calibrations but calibrated in the lab, similar to TOGA-TOF, whereas CAMS did both on-line and lab calibrations.

It is stated that calibration differences likely explain the observed offsets, if possible, it would also be useful to indicate how much of the stated total uncertainty (≈35%) is attributed to calibration versus instrumental variability or sample integrity?

Yes, we speculate in the paper that calibration differences may be a primary reason for the observed offsets between CAMS, TOGA-TOF, PTR-TOF and ISAF. The stated total uncertainty (≈35%) is a conservative estimate based in large part on the precision of repeated calibrations on different days. See previous response addressing calibration challenges. During the flights the inlet lines are highly passivated so we believe the sample integrity is excellent. The instrument variability was assessed during multiple repeat runs at specific calibration levels when we believed that the lines were well passivated and this folded into our uncertainty estimate. Our total uncertainty for many of the other VOCs that we measure is (≈15%) so the major uncertainty difference with HCHO compared to other measured VOCs is due to variability in calibrations

2. Sampling frequency and averaging

The use of volume-weighted averaging is a good way to align the slower TOGA-TOF data (33 s) with the faster instruments (1 Hz). It might be worth adding one or two sentences to acknowledge that while this approach slightly limits the instrument's ability to resolve rapid plume variability, the time resolution and inherent averaging is actually well suited for model comparison and regional-scale studies, if of course this is the conclusion.

The following revised sentence has been added in the conclusion:

> (Lines 541-543) "In addition, the time resolution and inherent averaging is well-suited for verifying satellite observations, three-dimensional modeling, and to compare with other instrumental techniques."

3. Use of CO as a strictly conservative tracer of dilution. CO is useful because it is as separate measurement to all the instruments being compared, but I believe it can have secondary production in the plume from VOC oxidation and could be oxidised itself? perhaps a VOC that TOGA measures could be used for estimating plume aging, perhaps acetonitrile? This would be interesting in future analysis because by being measured by the same instrument, some systematic errors could be eliminated.

The secondary production of CO is likely very small in comparison to the direct emissions, and thus we do not expect there to be a large impact from VOC oxidation in the overall CO levels in the plume. Further, emission factors of CO vary from 63 ± 17 to 135 ± 38 g kg^{-1} (approximately a factor of 2) for the seven primary biomass burning biomes, while acetonitrile emission factors vary from 0.11 ± 0.058 to 0.61 g kg^{-1}, i.e., with much lower emissions that vary by up to a factor of 6 (Akagi et al., 2011). Similarly, the emission factors for HCN vary from 0.29 ± 0.38 to 1.52 ± 0.82 g kg^{-1}. While using a VOC co-measured by the TOGA-TOF may eliminate some systematic errors in the volume-weighted averaging, it would introduce more overall uncertainty from both the variability of the emissions of HCN or acetonitrile from different fuel sources, and the lower overall mixing ratios being used in the comparisons.

Nevertheless, for future analyses, it would be interesting to compare NEMRs from both CO and acetonitrile or HCN. However, since the main purpose of this paper is to compare the NEMRs between different HCHO measurements, our conclusions are not expected to change as all NEMRs were normalized to CO.

technical corrections:

Excellent with just a few inconsistencies such as hypens in situ/ in-situ, 1 Hz /1-Hz and some spacing around parentheses.

We thank the reviewer for pointing out the inconsistencies. Based on the recommendations in the AMT style guide, which states *Latin phrases should not be hyphenated (e.g. "in situ", not "in-situ"). We recommend not to hyphenate modifiers containing abbreviated units (e.g. "3-m stick" should be "3 m stick"). This also applies to the other side of the hyphenated term (e.g. "3 m long rope", not "3-m-long rope"),* we have removed the hyphens from all instances of "in situ" and from all modifiers containing abbreviated units.

Reviewer comment 2

The manuscript "Trace Organic Gas Analyzer Time-of-Flight mass spectrometer (TOGA-TOF) system for airborne observations of formaldehyde." presents measurements from a newly designed analytical system that couples a fast GC with a time-of-flight mass analyzer to perform minutes-scale measurements of formaldehyde and near about 100 volatile organic compounds alongside. The instrument was tested during the FIREX-AQ campaign aboard the NASA DC-8 aircraft flying transacts across plumes. The measurements were compared against two other specialized instruments onboard the aircraft that performed targeted measurements of formaldehyde.

I really enjoyed reading this paper. It is very well written, structured and thorough in its comparison of the data obtained from the three instruments. The figures showcase comprehensive analyses and are easy to understand. I do not have any major comments but just a few minor ones. Following their resolution, I am happy to recommend the paper for publication:

1. Lines 192-193: The authors correct the fluctuations in sensitivity by normalizing with ambient tetrachloromethane. It would be good to add a citation for this or provide a rationale for selecting this compound from the broader suite of persistent ambient background species. Would this method work everywhere on the planet with the selected compound?

We added the following reference detailing the recommendation to normalize VOC measurements using CCl_4 to the text. However, caution does need to be exercised as many long-lived chlorinated compound mixing ratios are slowly decreasing due to being banned via the Montreal Protocol, stratospheric influenced air generally has lower values, and intermittent sources are occasionally identified. We have found for our airborne measurements in Asia, that CFC-11 is a more stable VOC to use for normalization.

Karbiwnyk, C. M., Mills, C. S., Helmig, D., and Birks, J. W.: Use of Chlorofluorocarbons as Internal Standards for the Measurement of Atmospheric Non-Methane Volatile Organic Compounds

Sampled onto Solid Adsorbent Cartridges, Environ. Sci. Technol., 37, 1002–1007, https://doi.org/10.1021/es025910q, 2003.

2. I would like the authors to expand a little more on the 35% uncertainty for HCHO measurements via TOGA-TOF. They attribute this to variability in repeat calibrations. However, a commercially available standard cylinder is used, the outflow from which is dynamically diluted with clean air or N2. So, what should create a variability this high in calibrations?

As we stated in our response to Reviewer 1 above, calibrating for HCHO is nontrivial. We do have a commercially made low ppm HCHO standard as well as a dilution of that standard as a mid-ppb level standard, both of which can be used with our calibration dilution system to generate a large range of HCHO mixing ratios relevant to ambient observations. However, independent spectroscopy-based measurements of these cylinders were only able to constrain the standard HCHO concentration to ± 5%, and repeated FTIR measurements over a 5-year period suggest that the HCHO concentrations in both standards are slowly decreasing. Thus, there is uncertainty from the standards themselves, as well as the dilution capability. Further, we note variability in the repeated dilution calibration series both associated with the use of humidified N2, dry N2, or clean air, and in single calibration dilution series. Lastly, we also need to factor in a small system background correction for HCHO from our blank Helium runs. With all of these uncertainties factored in, we believe that our stated ± 35% is a reasonable assessment of our measurement uncertainty.

3. Looking at figure 3a, I would like authors to acknowledge somehow also the limitations of this instrument for measurements on airborne platforms. Minute-scale time-resolution for a GC-TOF setup is great but figure 3a shows that it may still be slow for capturing chemically dynamic environments often experienced on fast moving platforms.

The following revision has been made in the conclusion:

(Lines 581 - 585) Due to its nature of taking discrete subsamples of air masses, highly heterogeneous features may not be captured by the TOGA-TOF system in comparison to fast HCHO measurement techniques, especially in chemically dynamic environments near strong sources with highly heterogeneous emissions and rapidly varying HCHO production and loss often detected during airborne measurements.

4. I suggest authors explain the volume-weighting in a bit more detail either in the main text or the SI. The current description somehow did not come across clearly to me as a reader.

We have edited the text that describes the volume-weighted averaging methodology to make it more clear how it was achieved:

(Lines 313-321) "The second averaging method is the volume-weighted averaging method based on the volume of air sampled. A post-mission laboratory analysis of the normalized volume vs. sample collection duration demonstrated that the flow rate through the sample trap was not constant during the sampling time period, despite the mass flow controller (MFC) being set to a fixed value. Specifically, relative flow rates were higher during the initial seconds of sampling before decreasing to the MFC setpoint. Through a series of experiments, we quantified the total sample volume as a function of collection time and derived an empirical relationship to determine the sample-volume midpoint, defined as the time at which 50% of the total collected volume had been accumulated during the sampling cycle (Fig. 3). Midpoints were calculated for each sample using this method."

Other minor points:

- Line 342: Lay out the mathematical description in equations-format for ease of understanding.

The sentence has been revised accordingly:

> "within the combined measurement uncertainties (2 ppb + 2% for LGR, and typically ± 2% for DACOM.)" --> (LGR = 2 ppb + 2% and DACOM = ± 2%).

- Line 349: "(described in Sect. 2.4). The note is already in section 2.4.

Changed accordingly.

- Lines 487-488: Do the authors mean a reduction in plume heterogeneity with aging is what suppresses the bias? "evolution" is a capture-all sort of a term but does not really tell what the authors intend to say. I acknowledge the explanation provided in subsequent lines.

(Line 497) The wording has been changed from "evolution" to "reduced".

References:

Akagi, S. K., Yokelson, R. J., Wiedinmyer, C., Alvarado, M. J., Reid, J. S., Karl, T., Crounse, J. D., Wennberg P. O.: Emission factors for open and domestic biomass burning for use in atmospheric models, Atmos. Chem. Phys., 11, 2039-4072, https://doi.org/10.5194/acp-11-4039-2011, 2011.