

Atmos. Meas. Tech. Discuss., referee comment RC3
<https://doi.org/10.5194/amt-2022-13-RC3>, 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Comment on amt-2022-13

Anonymous Referee #3

Referee comment on "Intercomparison of upper tropospheric and lower stratospheric water vapor measurements over the Asian Summer Monsoon during the StratoClim campaign" by Clare E. Singer et al., Atmos. Meas. Tech. Discuss.,
<https://doi.org/10.5194/amt-2022-13-RC3>, 2022

Review of "Intercomparison of upper tropospheric and lower stratospheric water vapor measurements over the Asian Summer Monsoon during the StratoClim campaign", by C.E. Singer et al.

The paper directly compares in situ water vapor measurements by three instruments aboard the M-55 Geophysica aircraft during the StratoClim campaign: CWIS, FISH and FLASH-a. Of the three, CWIS is the newest instrument, making its debut on an aircraft. The data evaluation is fairly straightforward except that FISH measurements must be divided into "clear sky" and "in cloud" since it does not discriminate between water vapor and condensed-phase water entering its sampling inlet. The paper concludes finding "no intrinsic technique-dependent biases". The paper compares not only the mixing ratio data from each instrument, but also RH data calculated using ambient temperature measurements. These RH values are also used to demonstrate the "realism" of RH-ice values calculated from the mixing ratio data relative to theoretical limits.

Overall, the paper is fairly well written, and reasonably concise with 10 of 18 figures relegated to a supplement. Some of the figures in the supplement are discussed in the main body text, making it cumbersome to flip between the two documents. My only three significant concerns are the following, with minor comments to follow.

The near-zero bias values between instruments over the full campaign are misleading because a much larger positive bias for one flight effectively "cancels out" a similar magnitude negative bias for another flight. For example, for FISH-FLASH biases for F4 and F6 in Figure 3. F4 shows a negative bias of 10-15% and F6 shows positive bias of 10-15% for F6. When combined over the entire campaign, the two biases basically add to zero. I would like to suggest that summary bias statistics for each instrument pair, for each flight, including uncertainties, be provided in a table. Then the reader can see not only the large biases during specific flights, but also how these biases change in both magnitude and sign from one flight to the next. Yes, this information can be gleaned from Figure 3, but

having the statistics readily available in a table will make them much more accessible.

I'm sure others have pointed out that the MLS4.3 data used in the paper have now been replaced by v5 that reduces some of the "wet" biases that had crept into the older v4 retrievals. Although I don't find it essential to replace the MLSv4 data in the paper (e.g., profile statistics in Figure 8), it should be stated that similar profile statistics based on MLSv5 data would generally shift them to lower mixing ratios by 0.2-0.3 ppm. The "dry bias" mentioned in L348 would become even greater, while the "wet bias" (L349) would be reduced by v5. The importance of this MLS data to the comparison with the aircraft instruments would become much greater if some quantitative "coincidence" criteria with the aircraft flights were mentioned. How far apart in time and space were these MLS and Geophysica measurements made?

I am somewhat baffled by the treatment (and lack of use) of the 11 different CFH profiles that were obtained during the campaign. Was the only use of this data to consolidate 11 individual profiles down to one "statistical" (mean \pm stddev) profile to compare with the aircraft data? Were the individual CFH profiles of little use in comparing to the aircraft data for individual flights? For example, would CFH profiles in coincidence with F4 and F6 help explain why the bias between FISH and FLASH changed sign between the two flights? If there is a good reason why the CFH profiles are used only to create one "statistical" profile it would be beneficial to include such a statement, otherwise the reader is left wondering why the balloon profiles play so little role in the intercomparison.

Minor comments:

Lines 10-11 vs line 12: This confused me when I first read the abstract. Overall mean differences are on the order of 1%, but then "In general, CWIS and FLASH agreed to better than 10% for 92% ...". I thought there was probably a typographical error. However, after reading the paper, I now understand why the overall 1% bias values are so small - the topic of my first significant concern above. I suggest re-working the abstract text describing overall (full campaign) bias statistics to be less misleading, as I think full-campaign statistics are not very meaningful when there are large flight-to-flight differences in the biases.

L10-11: Need to include uncertainty values for mean differences

L13: change "to" to "within"

L15: The detection of "fine-scale spatial structures" depends not only on high precision (not accuracy), but also on instrument response time. This sentence is a bit of "grandstanding" relative to the contents of this paper.

L18: With mixing ratios > 10,000 ppm at Earth's surface, water vapor is not a "trace gas"

L20: "that roughly doubles the anthropogenic warming from carbon dioxide alone" is a very mis-interpretable statement without additional context.

L26: Mole fractions are not "concentrations".

L36: Is water vapor a "pollutant" in Earth's atmosphere?

L42: "low" is such a relative adjective. Most trace gases exist at mixing ratios <5 ppm, some < 5 ppt, so why is 5 ppm considered "low" for water vapor? It is because water vapor can also be found in the atmosphere at >10,000 ppm. Anytime an instrument is carried from the very wet lower troposphere to the UTLS it is challenge to measure the very dry air there. That's why at 5 ppm the water vapor mixing ratio is low.

L46: I would change "Although" for "Because"

L55: NOAA frost point hygrometer is also known as the NOAA FPH.

L64: Voemel and Hall published what can be described as "detailed uncertainty analysis of water vapor measurements by the CFH and FPH, respectively," rather than "an intercomparison"

L74: remove "with" before "within"

L81: Knowing that none of these three instrument directly measures RH, how can RH values provide accuracy information when the RH uncertainties are a combination of the uncertainties of simultaneous WV and T measurements?

L83: the second occurrence of "in situ measurements" can be omitted

L90: There have been previous in situ measurements in the ASM. For example, see <https://doi.org/10.1016/j.atmosres.2022.106093>

L219: "finite" is extremely vague here. Is the time needed on the order of 10s of seconds? 100s of seconds? Please be more quantitative here.

L236: "of better 1.5%" makes no sense.

Figure 3: This figure intrigues me because there are no tails on both sides of the distributions. Do the normalized PDFs really drop from viewable values to unviewable values in these panels? Why does the y-scale go from 10^{-3} to 10^{-5} in one increment when all the other increments are a single factor of 10?

L253: These mean differences (and every single one in the paper) should be accompanied by a uncertainty value, most like two standard deviations of the mean.

Figure 4: Yellow is never a good choice for colored markers in Figures. It is difficult to tell where the yellow markers begin and end in this figure. The terminology "heat map" is new to me. What does it mean?

L282: You don't need pressure to calculate RH, only WV partial pressure and ambient temperature.

Figures 6, 7: I understand the purpose of showing values of RH-ice here against the theoretical limit to assess WV measurement accuracy, but with anything but the highest accuracy T measurements used to derive RH values, calculated RH values have uncertainty contributions from the T measurements as well as the WV measurements. If RH-ice slightly exceeds the theoretical limit, is the WV accuracy to blame? or the T accuracy? or both?

What I do not understand is the point-by-point comparison of RH-ice values from the three hygrometers. How does this analysis shed any more (or different) light on the previous comparison of WV mixing ratios? RH values are a derived quantity, while WV mixing ratios are the direct measurements. What does Figure 8 add that Figure 2 hasn't already shown?

L368: This sentence is far too "grandstanding" for what is discussed in this paper. How can the attribution of "good agreement" for the full campaign data, knowing that individual flight biases are high in some cases, and flip from negative to positive from one flight to the next, be the product of these two things? Please tone this sentence down as it far too overreaching. It also makes it sound like you all colluded after the flights to make

sure the data from different instruments agree with one another. It is clear these data sets were not prepared or submitted under "blind" conditions, but "robust communication between instrument teams during subsequent data analysis" probably makes it sound more conspiratorial than it was.

L374: I think "major progress" overreaches the findings of this comparison paper. Some instrument pairs were found to have large (10-15%) mean biases during some flights. That's not new. The biases sometime changed sign from one flight to the next. Definitely not new. When all the flight are combined, the large biases observed for some individual flights were diminished by similar biases of opposite sign for other flights. This is fortunate, but not necessarily new. If you can demonstrate small flight biases for instrument pairs that are consistent over a campaign, now THAT would represent "major progress".