Response to anonymous reviewer #1

Thank you very much for your positive review and your helpful comments – they have improved the manuscript greatly.

Major comments:

The study uses v4.3 for the comparison with MLS water vapor measurements. The MLS team has already released v5.0, which addresses some significant biases and drifts that do affect the comparison with the in situ measurements. I would strongly suggest using v5 instead of v4.3 as the results would look slightly differently.
→ This was commented on by all reviewers and the following reply has been sent to all reviewers.

Thank you for alerting the authors to the newest version of MLS. At the time this analysis was started v5 had not yet been released. We have updated the analysis to use MLS v5. The results are similar, but in the UTLS (~68 hPa), v5 is about 15% drier than v4 across the whole globe (not shown). In consultation with the MLS team at JPL, there is not a known cause of the larger than expected (from changing sideband fractions) discrepancy at these levels between the two versions, which is about twice as large as reported by Livesey et al. (2021) in section 4: https://doi.org/10.5194/acp-21-15409-2021/. For completeness we have included both v4 and v5 in Fig 9 (previously Fig. 8). Now MLS appears drier than in situ measurements (aircraft and balloon) in both the warm/wet and cold/dry halves of the campaign, but still captures the qualitative shift of drying through time observed by the aircraft instruments across these two periods. We have added a short discussion about the differences between MLS v4 and v5 to the text:

"Here we use 126 water vapor profiles spatially and temporally co-located with the StratoClim flights as a point of comparison (shown in Fig. 1a). We use version 5.0 (v5) profiles which were selected in the region between 20–30°N and 78–92°E during the campaign dates of 27 July – 10 August 2017, using screening criteria from Livesey et al. (2022). We also show MLS version 4.3 (v4) profiles (only 118) which were selected using screening criteria in Livesey et al. (2020). We interpolate the H2O profiles onto a potential temperature grid using the MLS temperature product provided at the same pressure levels. MLS v5 includes a correction on the H2O retrievals described in Livesey et al. (2021), which results in an approximately spatially uniform drying at 68 hPa of about 15%.
…
Overall MLS v4 shows a wet bias compared to v5 (of about 15% between 380-500 K), but both versions are able to discern trend across the campaign of a cooling/drying of the UTLS seen by the aircraft measurements.
…
During the warm/wet period, MLS v5 shows a significant dry bias compared to the aircraft instruments of (-19 ± 7)% and (-22 ± 6)% for ChiWIS and FLASH, respectively.

During the cold/dry period, MLS v5 shows an insignificant dry bias of (-12 ± 12)%, (-11 ± 12)%, and (-5 ± 15)% compared to ChiWIS, FLASH, and the balloon CFH, respectively. Because MLS v4 is 15% wetter compared to v5 in this altitude range, MLS v4 actually agrees more closely with the in situ measurements, reporting no statistically significant differences with any of the instruments during either period of the campaign."

Minor comments:

I would suggest moving Figure S6 from the supplement to the main document. It is quite helpful in understanding the discussion of the cloud determination.
→ Thank you, Figure S6 has been moved into the main text.

Since the authors refer to Figure S9 twice, it too could possibly be moved to the main text.
→ Figure S9 is shown to present the reader with a fuller picture of the atmospheric state, including data from ChiWIS that was excluded from the intercomparison. So as not to confuse the reader on which datapoints are being intercompared, we have kept S9 as a supplemental figure. Thank you.

Lines 246ff: It would be quite useful to have an overview table, which lists for each flight the total flight hours, total number of data points, total number of measurements in cloud and out of cloud. Without this, the number of data points given here miss perspective.
→ A new table (Table 2) has been added with a summary of the flight hours for the 6 flights discussed in this paper.

The comparisons in Figure 8 look quite good. Nevertheless, there may be some sampling biases in this comparison, since the balloon-borne measurements are typically launched only in non-precipitating conditions, and MLS is sensitive to cloud contamination in the UTLS. A short discussion about how this potential sampling bias may influence the comparison could be helpful.
→ Thanks for this suggestion. There are a variety of possible sampling biases that may be present in this comparison. This makes it all the more impressive that there is such good agreement across platforms. We have added the following discussion about Fig 9 (previously 8):

"Comparisons between these three platforms is challenging because the measurements were not perfectly coincident in space or time and the region sampled showed large day-to-day variability. The in situ measurements from the aircraft and balloons also have much higher spatial and temporal resolution than the MLS satellite instrument. Furthermore, sampling biases may be exacerbated in this comparison to do diurnal variations since the aircraft and MLS measured only during daytime while the balloons were only launched at dawn/dusk. Other sampling biases may be present such as cloud contamination, which to some extent all measurements are susceptible to."

Lines 374: Not unexpected, the comparison was not blind and there have been many interactions between the different teams during the campaign. I have confidence in each team,

but it would still be good to know to what extent if any instrument calibrations were adjusted during the campaign based on input from the other teams.
→ Thank you for this comment. We appreciate now that the sentence as written may be misleading. We did not mean that any instrument calibrations were adjusted based on other teams' data, but rather that problems were initially discovered based on a preliminary instrument intercomparison and then confirmed with laboratory experiments. Calibrations were applied based solely on those laboratory experiments and calibration runs with gas standards. We have updated the text to clarify this as follows:

"Finally, frequent communication between instrument teams and a preliminary intercomparison effort led to the early discovery of measurement problems from both the FLASH and ChiWIS instruments. The instruments did not calibrate relative to each other, but rather these issues corrected for independently based on laboratory experiments and calibration runs with gas standards. More detail on the ChiWIS laser "pedestal" (stray light) that was corrected for with calibration based on laboratory experiments can be found in Clouser et al. (2022, in prep.)."

Technical comments:

Line 150: delete the stray comma.
→ Done.

Lines 269f: Better write: "During the stair stepping, there were two moist layers around …"
→ Done.

In Figure 5: I would suggest making the vertical axes for potential temperature and H2O the same between the upper and lower group of panels.
→ Thank you for this suggestion. While we agree that it is convenient to use the same vertical axes on subplots, due to the fact that different flights sampled different altitude ranges, doing so makes the F7 data more difficult to see because it is squashed together. We have chosen to leave Figure 5 as is, but included a statement in the caption that points out the different vertical axes to the reader.