https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

The SPARC water vapor assessment II: Assessment of satellite measurements of upper tropospheric water vapor

William Read[1], Gabriele Stiller[2], Stefan Lossow[2], Michael Kiefer[2], Farahnaz Khosrawi[2], Dale Hurst[3], Holger Vömel[4], Karen Rosenlof[5], Bianca M. Dinelli[6], Piera Raspollini[7], Gerald E. Nedoluha[8], John C. Gille[9,10], Yasuko Kasai[11], Patrick Eriksson[12], Christopher E. Sioris[13], Kaley A. Walker[14], Katja Weigel[15], John P. Burrows[15], and Alexei Rozanov[15]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, Ca., USA.
[2]Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research, Karlsruhe, Germany.
[3]Global Monitoring Division, NOAA, Earth System Research Laboratory, Boulder, Colorado, USA.
[4]Earth Observing Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA.
[5]Chemical Science Division, NOAA, Earth System Research Laboratory, Boulder, Colorado, USA.
[6]Instituto di Scienze dell'Atmosfera e del Clima del Consiglio Nazionale delle Ricerche (ISAC-CNR), Via Gobetti, 101, 40129 Bologna, Italy.
[7]Instituto di Fisica Applicata del Consiglio Nazionale delle Ricerche (IFAC-CNR), Via Madonna del Piano, 10, 50019 Sesto Fiorentino, Italy.
[8]Naval Research Laboratory, Remote Sensing Division, 4555 Overlook Avenue Southwest, Washington, DC 20375, USA.
[9]National Center for Atmospheric Research, Atmospheric Chemistry Observations & Modeling Laboratory, P.O. Box 3000, Boulder, CO. 80307-3000, USA.
[10]University of Colorado, Atmospheric and Oceanic Sciences, Boulder, CO 80309-0311, USA.
[11]National Institute of Information and Communications Technology (NICT), 20 THz Research Center, 4-2-1 Nukui-kita, Koganei, Tokyo 184–8795, Japan.
[12]Chalmers University of Technology, Department of Space, Earth and Environment, Hörsalsvägen 11, 41296 Göteborg, Sweden.
[13]York University, Center for Research in Earth and Space Science, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada.
[14]University of Toronto, Department of Physics, 60 St. George Street, Toronto, Ontario M5S 1A7, Canada.
[15]University of Bremen, Institute of Environmental Physics, Otto-Hahn-Allee 1, 28334 Breman, Germany.

Correspondence: Read (william.g.read@jpl.nasa.gov)

Abstract. Nineteen limb viewing (occultation and passive thermal) and two nadir humidity data sets are intercompared and also compared to frostpoint hygrometer balloon sondes. The upper troposphere considered here covers the pressure range from 300–100 hPa. Water vapor in this region is a challenging measurement because concentrations vary between 2–1000 parts per million volume with sharp changes in vertical gradients near the tropopause. The atmospheric temperature is also highly variable ranging from 180–250K. The assessment of satellite measured humidity is based on coincident comparisons with frostpoint hygrometer sondes, multi month mapped comparisons, zonal mean time series comparisons and coincident satellite to satellite comparisons. While the satellite fields show similar features in maps and time series, quantitatively, they can differ by a factor of two in concentration, with strong dependencies on the amount of H_2O. Additionally, time-lag response corrected Vaisala-RS92 radiosondes are compared to satellites and the frostpoint hygrometer measurements. In summary, most satellite data sets reviewed here show on average ∼30% agreement amongst themselves and frostpoint data but with an additional

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

~30% variability about the mean. The Vaisala-RS92 sonde even with a time-lag correction shows poor behavior for pressures less than 200 hPa.

1 Introduction

A general assessment of water vapor measurements, both from remote and in-situ sensors was undertaken within the Stratosphere-troposphere Processes and their Role in Climate (SPARC) core project of the World Climate Research Program (WCRP) prior to 2000. This activity known as the Water Vapor Assessment (WAVAS) published a report in 2000 (Kley et al., 2000). Since then, there have been a significant increase in satellite missions, ground based instruments, launches of balloon frostpoint hygrometers (BFH) and improved operational radiosonde hygrometers. Therefore a reassessment of these new resources is needed, now referred as the SPARC WAVAS-II assessment. This paper amongst several in this ACP/AMT/ESSD special issue focuses on the upper troposphere. The upper troposphere is defined, depending on the application, from 300 hPa to the NASA Goddard's Global Modeling Assimilation Office (GMAO), Modern Era-Retrospective Analysis for Research and Applications (MERRA) tropopause height or 100 hPa.

2 Data Sets

Table 1 lists the 21 satellite data sets that are considered in this report. In addition, we use BFH and corrected Vaisala-RS92 radiosondes. Although BFH have been in use for decades and launched from multiple sites, only six sites are considered here because they have a long time series of repeated launches. This helps provide some comparison statistics and temporal variability. The chosen sites are Boulder, USA (105.3W, 40.0N, 1980 to present), Lauder, New Zealand (169.7E, 45.0S, 2004 to present), Hilo, USA (155.1W, 19.7N, 2010 to present), Heredia, Costa Rica (84.1W, 10.0N, 2005 to present), Lindenberg, Germany (14.2E, 52.2N, 2006 to present), and Sodankylä, Finland (26.6E, 67.4N, 2002 to present). The Global Climate Observing System (GCOS) Reference Upper Air Network (GRUAN) launches high quality radiosondes for climate research. Here we use water vapor observations from the Vaisala-RS92 radiosondes, which have been processed by GRUAN to remove all known biases and to correct for all known effects influencing water vapor measurements (Dirksen et al., 2014). For the corrected Vaisala-RS92 radiosonde set we use data from Barrow, USA (156.3W, 71.3N, 2009 to present), Boulder, USA (105.3W, 40.0N, 2011 to present), Cabauw, Netherlands (5.5E, 52.1N, 2011 to present), Lauder, New Zealand (169.7E, 45.0S, 2012 to present), Lindenberg, Germany (14.2E, 52.2N, 2005 to present), Ny-Ålesund, Norway (12.6E, 78.9N, 2006 to present), Southern Great Plains, USA (97.0E, 36.6N, 2009 to present), and Sodankylä, Finland (26.6E, 67.4N, 2007 to present).

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

Table 1. Instruments considered for upper tropospheric H_2O quality assessment

Data Set	Version	Period[a]	Meas. Type	Lat. Coverage	Vert Range
AIRS Aqua	V6	Jul 2002—present	IR Thermal Emission	90S–90N	p $\geq$ 200 hPa
ACE-FTS	V3.5	Aug 2003—present	IR solar Occultation	90S–90N	p $\leq$ 500 hPa/CT[a]
GOMOS	V6	Mar 2002–Apr 2012	NIR Stellar Occultation	90S–90N	p $\leq$ 300 hPa
HALOE	v19	Sep 1991–Nov 2005	IR solar Occultation	80S–80N	p $\leq$ 300 hPa
HIRDLS	V7	Jan 2005–Apr 2008	IR Thermal Emission	65S–82N	p $\leq$ 200 hPa
MAESTRO	V27,28,29,30	May 2004–present	NIR Solar Occultation	86S–87N	p $<$ 500 hPa/CT[b]
MIPAS Bologna	V5H	Jul 2002–Mar 2004	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS Bologna	V5R NOM	Jan 2005–Apr 2012	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS ESA	V5H	Jul 2002–Mar 2004	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS ESA	V5R NOM	Jan 2005–Apr 2012	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS IMK	V5H	Jul 2002–Mar 2004	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS IMK	V5R NOM	Jan 2005–Apr 2012	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS Oxford	V5H	Jul 2002–Mar 2004	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MIPAS Oxford	V5R NOM	Jan 2005–Apr 2012	IR Thermal Emission	90S–90N	p $\leq$ 300 hPa
MLS Aura	V4.2	Jul 2004–present	mm Thermal Emission	82S–82N	p $\leq$ 316 hPa
MLS UARS	V490	Sep 1991–Jun 2008	mm Thermal Emission	80S–34N, 34S–80N	464–147 hPa
POAM-III	V4	Mar 1998–Dec 2005	NIR Solar Occultation	88S–63S, 55N–71N	p $\leq$ CT[a]
SAGE-II	V7	Oct 1984–Aug 2005	NIR Solar Occultation	70S–70N	p $\leq$ CT[a]
SAGE-III	V4	Dec 2001–Mar 2006	NIR Solar Occultation	50S–80N	p $\leq$ CT[a]
SCIAMACHY	V3	Mar 2002–Apr 2012	UV limb solar backscatter	85S–85N	p $\leq$ 300 hPa
SMILES NICT	V1	Sep 2009–Apr2010	smm Thermal Emission	40S–65N, 65S–40N	200–100 hPa
SMILES JPL	V2	Sep 2009–Apr2010	smm Thermal Emission	40S–65N, 65S–40N	215–83 hPa
SMILES Chalmers	V3	Sep 2009–Apr2010	smm Thermal Emission	30S–30N, 65S–40N	280–200 hPa
SMR-Odin	UT	Feb 2001–present	smm Thermal Emission	30S–30N	p $\leq$ 300 hPa
TES	V6	Jul 2004–present	IR Thermal Emission	82S–82N	p $>$ 200 hPa

a Comparisons are performed prior to 2017. b Cloud top

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

The satellite data sets are described in more detail in a companion paper by Walker and Stiller (2021). The data sets are quality screened per recommendations from each of the data set providers. These data sets were read and repackaged in a common format that contains the fields, year, UT time, longitude, latitude, day night or sunrise sunset flag, tropopause height, height, pressure, H_2O concentration, and H_2O concentration uncertainty. Figure 1 shows a list of data sets used here and the color and symbol coding being used when multiple data sets are shown in a plot.

3 Comparison Methods

Tropospheric H_2O is highly variable both temporally and spatially making accuracy assessments difficult. Three comparison methods, coincident comparisons, time series, and gridded maps are used here to assess the data. For coincident comparisons we compare measurement pairs that are within 2.5° in longitude and latitude and 3 hours in time. The spatial coincidence is roughly the weighting function width for a limb sounder and there is no benefit to using a tighter criterion. The temporal matching criterion is rather arbitrary but is well under a diurnal time difference (12 hours). In principle, coincident pair matches are the best method of comparing two data sets but have some limitations. The first having a suitable number of coincidences to obtain enough statistics, and secondly, having good global coverage of the matches. For example, comparing a limb viewer in a sun synchronous orbit with an occultation instrument means the only available coincidences will occur for specific latitudes where the local viewing time is within ± 3 hours of sunrise and sunset.

Time series comparisons are useful to see how well the instruments track temporal changes and interannual variability. These comparisons are also useful for detecting possible drifts in their H_2O retrievals. This is why we only used BFH sites that have frequent launches (monthly or more frequent) over several years.

The third comparison methodology compares gridded data maps. Many scientific studies are interested in global distributions of H_2O during the year and also how this changes interannually. One advantage of this type of comparison is that small-scale variability over time and space is averaged out. A disadvantage is that there can be significant sampling biases. For example, limb viewing infrared instruments are heavily cloud contaminated in the tropics and will show a significant dry bias compared to maps made from nadir viewing or submillimeter instruments (Millán et al., 2018). This study will show that the sampling bias is more than a factor of two in the upper troposphere. Also there will be temporal biases because sun-synchronous orbiters only sample two local times missing much of the diurnal cycle (Eriksson et al., 2010).

4 Coincident Comparisons

4.1 Frostpoint and Radiosonde hygrometers

4.1.1 Comparison Issues

A humidity measurement made remotely by a satellite is quite different from an in-situ measurement because the remote sensor sees the averaged humidity over a few 100 km whereas the in-situ humidity is a point in space. Upper tropospheric air is usually

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

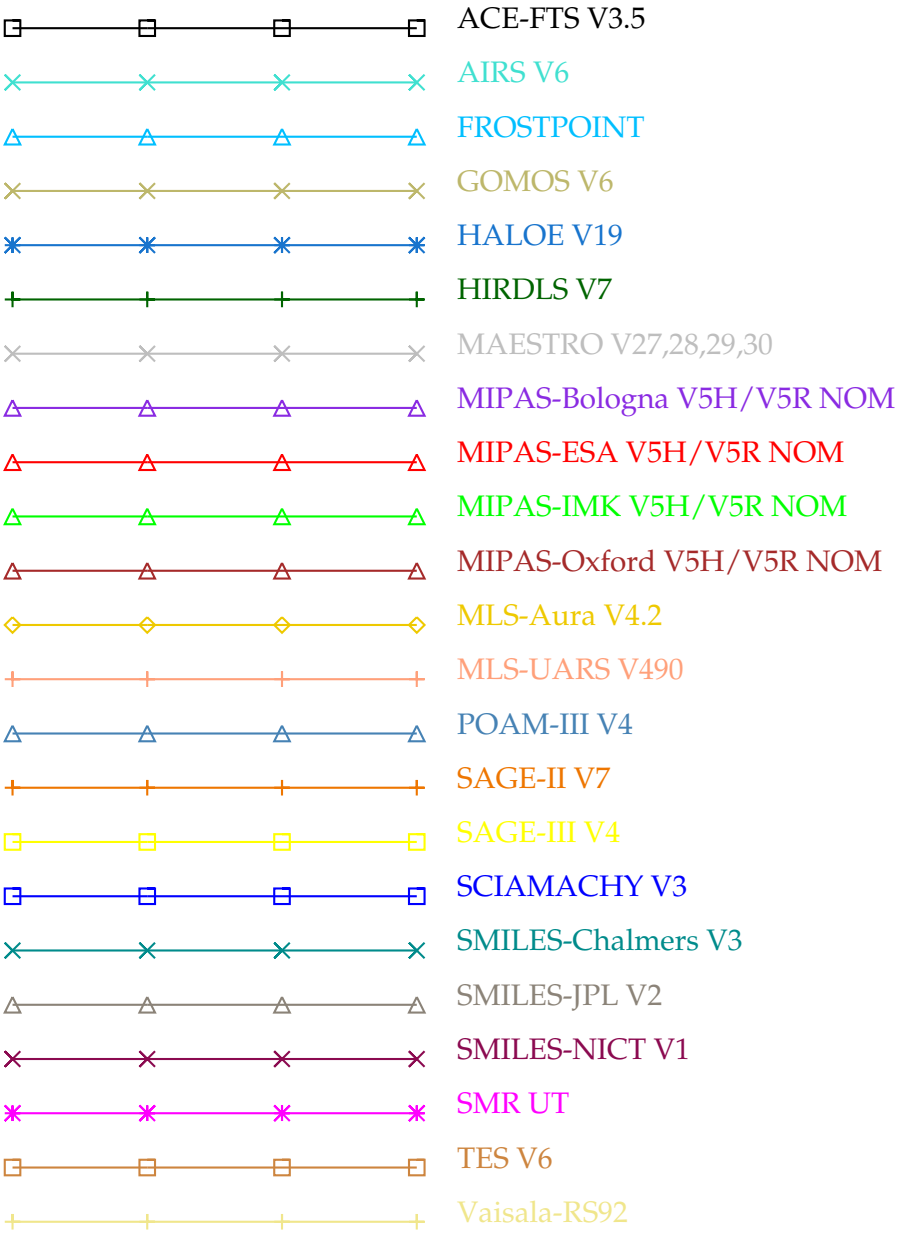


Figure 1. The colors and symbols used to identify instruments in multi-instrument plots.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

not well mixed and therefore there is typically a large humidity variation within the measured volume that will not be captured in an in-situ measurement. Using MOZAIC (Measurement of OZone and water vapor by Airbus In-service airCraft, Marenco et al., 1998) data during the UARS MLS upper tropospheric validation, it was noted that over 100 km of level flight, MOZIAC humidity typically showed 20-30% variability (Read et al., 2001). An example of how a "coincident" comparison between an in-situ and volume averaged satellite measurement might look like is shown in Figure 2. In this figure we take 200 values and generate a random sequence of values having a mean of 100 with a 1-standard deviation variability of 25 (25%) shown as black asterisks ($*$). These values are sorted and plotted. Note that the curve is nonlinear with low and high values curving away from the mean value of 100. A BFH can measure any one of these values. Although it is more likely that the in-situ hygrometer will measure a subvolume that is close to the mean value, it is possible that it may measure a value in another region of the volume that departs significantly from the mean. A remote measurement will always sample a large volume and measure an average value; however, the average derived depends on the instrument measurement response to the H_2O concentration. This is shown as the gray plus ($+$) symbols in the figure.

Upper tropospheric water vapor has a large dynamic range of values from 2 ppmv to 1000 ppmv. Therefore it makes most sense to assess the degree of agreement in terms of percent of humidity. Reporting the results of a comparison between dataset x and dataset y can be done in three ways. First one can compute percent differences relative to dataset x and calculate the mean and standard deviation of the comparison. Another way is to compute the percent difference relative to the mean of the x and y datasets. The third is to compute the mean and standerd deviation in concentration and convert the result into percent. In our example in figure 2, computing the statistics in percent relative to the x dataset has a biased mean of 9.4% and a 38.8% standard deviation. If the comparison is made in terms of the sum of the x and y datasets, the mean bias and standard deviation are reduced to 3.9% and 28.9% respectively. When the analysis is done in concentration units and converted to percent produces a mean difference of 0% and 27% for the standard deviation. Ideally the expected result should be 0% for the mean and 25% for the standard deviation; therefore, the analysis of the comparsions are done in concentration and converted to percent afterwards.

Figure 3 shows a coincident humidity comparison between AIRS and the Earth Science Research Laboratory (ESRL) BFH that is routinely launched from Boulder on a monthly basis. The comparison pressure is 261 hPa. Next to it is a comparison between MLS and the BFH at 100 hPa. Averaging kernels were not applied to either data. The BFH data have been sorted by value and shows a similar shape to that in Figure 2. Likewise, both AIRS and MLS show a generally flatter response. In addition to the spatial averaging variation, there is atmospheric variability that accounts for why the slope for the satellite measurement is not zero like it is in the demonstration plot (Figure 2). Additionally, satellite spatial averaging and the retrieval itself over the sampled volume will exhibit some nonlinearities and non Gaussian behavior. Therefore, while a comparison like that in Figure 2 may not look good, the reality is that, the agreement may actually be as good as one can expect because of the very different characteristics of the measurements themselves. It is also important to recognize that the dynamic range of measurements at 261 hPa is 10–400 ppmv versus the stratospheric 100 hPa which run from 2.5–7.5 ppmv. The smaller dynamic range for the stratospheric measurements suggests that H_2O is more tightly regulated by large scale atmospheric processes (e.g. tropical tropopause temperature, transport, and chemistry). This is shown in the MLS-Aura comparison which generally tracks

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

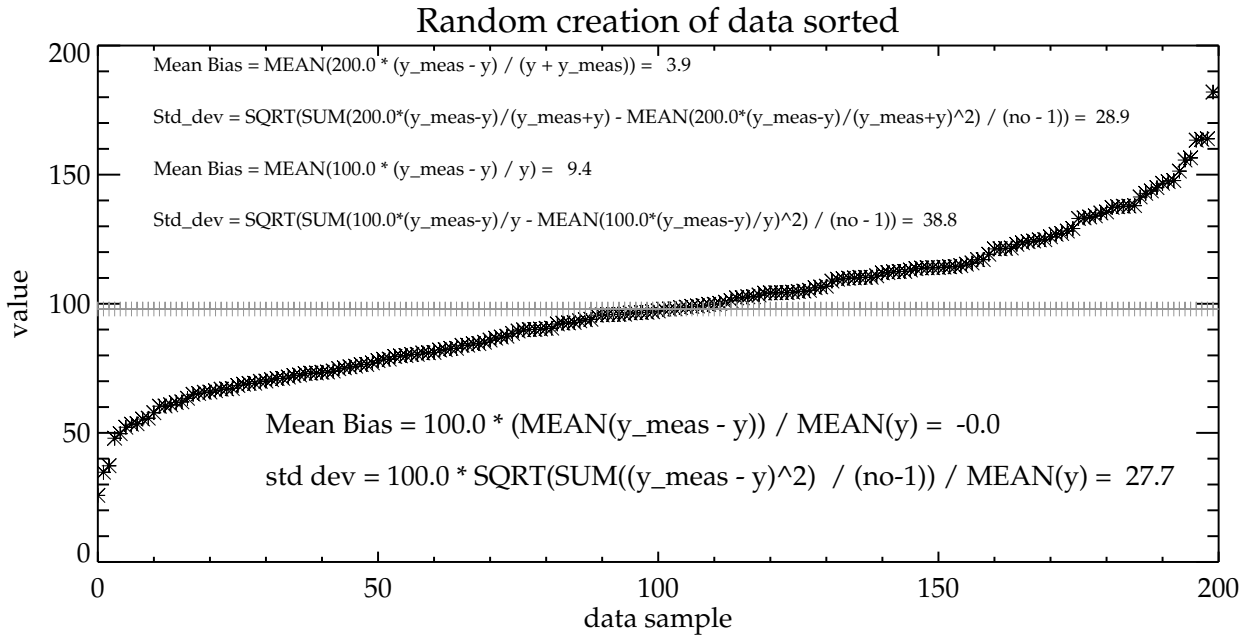


Figure 2. Figure shows how a randomly sampled measurement within a volume would compare to a whole sample volume averaged measurement. The + symbols represent the individually generated measurements (as measured by a sonde) that make up an average value of 100 (as seen by a satellite).

the BFH values even through to the highest values. MLS-Aura appears to overestimate however the extreme lower values. BFH versus MLS-Aura and the MIPAS suite comparisons at 261 hPa look similar to that shown for AIRS, and at 100 hPa for the MIPAS suite look similar to that shown for MLS-Aura.

4.1.2 Summary of Coincidences

Figure 4 summarizes results of coincident comparisons between BFH launched from Boulder and some individual satellite data sets with and without application of the averaging kernel. The averaging kernel is applied (Livesey et al., 2018) to the coincident sonde profiles for the MIPAS and MLS retrievals. Although applying the averaging kernel to a highly vertically resolved measurement when comparing to a remote sensing measurement is the proper method for comparison, in practice there are some limitations. These include neglect of non linear forward model effects (causes the averaging kernel function to be profile shape and amount dependent) and truncation effects when the balloon does not achieve high altitude. For some of the retrievals, applying the averaging kernel makes the agreement worse, in particular, the MIPAS-Oxford retrieval. For MLS-Aura, which has the most number of coincidences, applying the averaging kernel makes very little difference. The same is also true of the MIPAS-IMK retrieval. Off line simulation studies done on MLS support the above result. The averaging

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

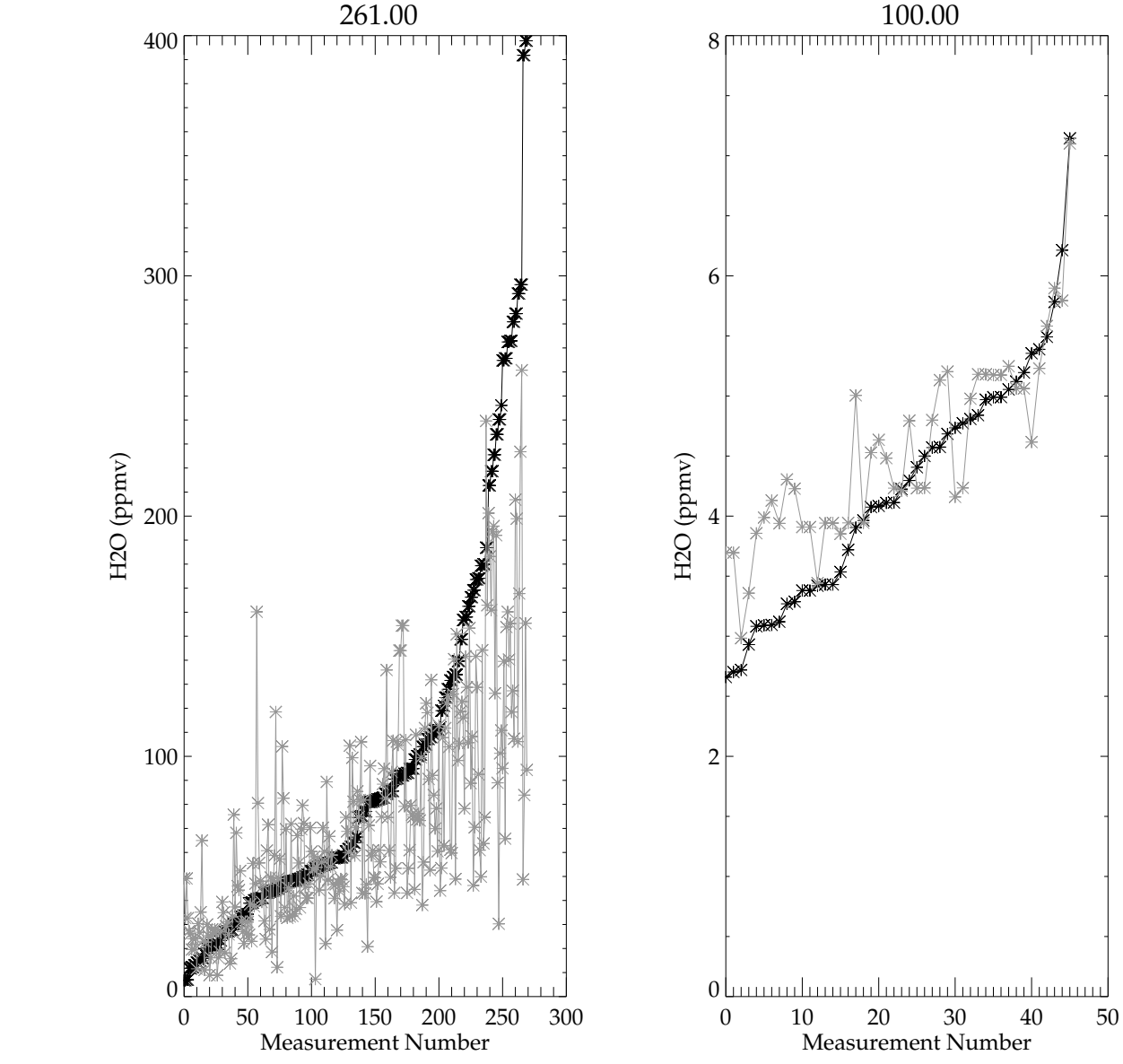


Figure 3. A comparison of coincident Boulder BFH and AIRS measurements (left panel) at 261 hPa and MLS-Aura measurements (right panel) at 100 hPa is shown. The BFH measurements are sorted by value, black, the satellite measurements in grey.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

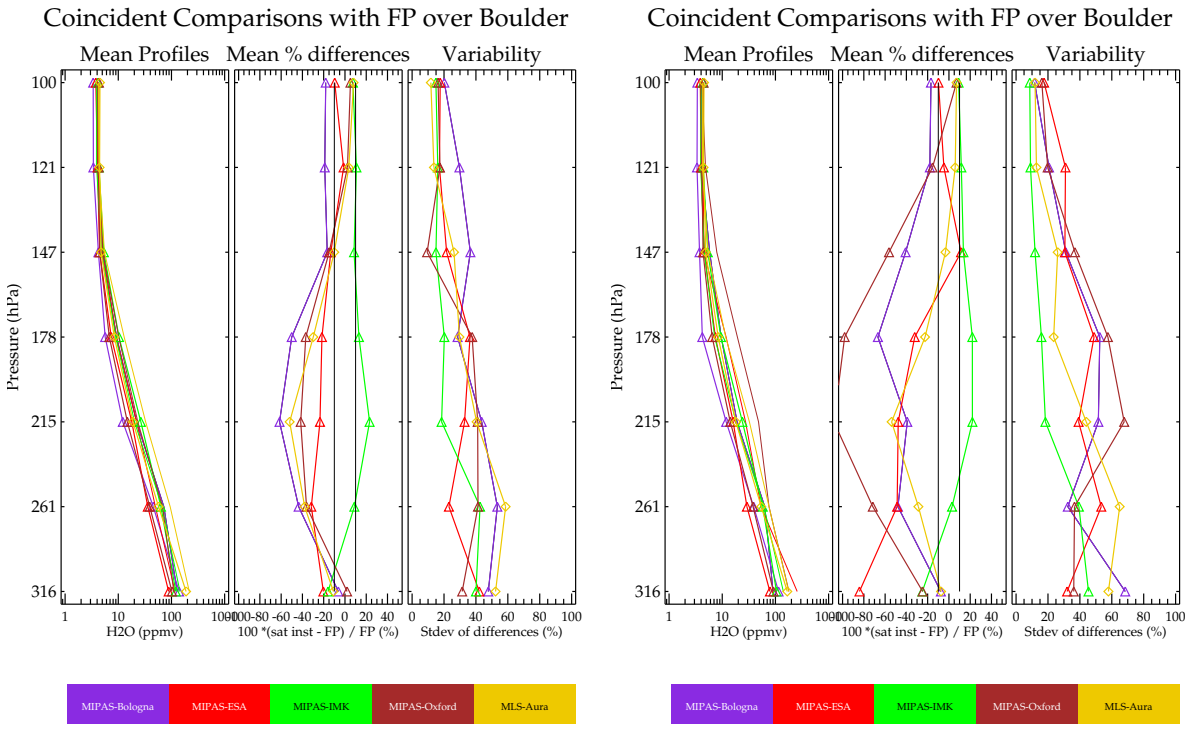


Figure 4. Summaries of coincident profiles comparisons between Boulder BFH without averaging kernel applied (left) and with averaging kernel applied (right) and the satellite data sets. The data sets are color coded according to the caption below. The leftmost in the group of three panels shows the mean of the coincident profiles (thick w/symbols) for the satellite data set, and the corresponding mean for the BFH (thin line). The center panel within the triad shows the mean bias, and the right panel within the triad is the variability about the mean.

kernel is important for the 121 hPa and lower pressure levels but was not important for the higher pressure levels (Read et al., 2008). Averaging kernels are not available for all the data sets used here, for example AIRS. Therefore there is no advantage to be gained from using the averaging kernel and for consistency in handling of the data sets here it is not used in the following analysis.

Figure 5 is a summary of coincident scatter plot comparisons between BFH and satellite retrievals where there are enough coincidences to generate some statistics (10 or more). The tight coincidence matching criterion limits these comparisons to passive limb sounders. For each sonde site, the left panels show the mean profile of the coincidences being measured with the thick line with symbols are the satellite data sets and the unsymbolled thin line is the mean of the coincident sonde profiles. Since the actual coincidences differ in number and time of measurement, the means of the sonde profiles will be different for each instrument. The center panel shows the bias in percent. The right panel shows the variability of the coincident differences

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

about the mean difference. The root mean square of the coincidently compared profiles is the square root of the sum of the bias (center panel) squared and the variability (right panel) squared.

The comparisons in Figure 5 show mean coincident agreement within several tens of percent of the BFH with a scatter about the mean value of 20–60% for most instruments. AIRS shows the best agreement overall. MIPAS-IMK shows typically 20% agreement but with a positive bias. The other MIPAS retrievals (Bologna, ESA, and Oxford) are mostly drier. MLS-Aura is also drier but consistently shows a significant dry bias for the level that is 2–3 km below the tropopause. For the mid–high latitude, this is near 215 hPa and for the tropical latitudes, it is at 147 hPa. Curiously, the MIPAS-Bologna retrieval also exhibits this behavior for the mid latitude comparisons but it is not possible to determine if this is linked to the tropopause height because there were no suitable comparisons in the tropics. HIRDLS shows moist biases for the mid latitude sites but a small dry bias at the Heredia (tropical) site.

4.2 Time Series Comparisons

Another way to look at the humidity data is through a time series. This type of comparison shows how each satellite data set will capture seasonal cycles and interannual variability. Comparisons are shown in two formats. Figure 6 shows an overlay of BFH sonde measurements at Boulder with smoothed reconstruction of satellite measurements in the vicinity of Boulder ($\pm 2.5^\circ$ longitude and latitude). Temporal coincidence with the actual Boulder sonde launches is not imposed. As is shown in the figure, most of the data sets capture similar annual cycles with varying degrees of fidelity relative to the sonde. Interannual variability is similar among the majority of the data sets and sonde. For example, 2007 shows higher values and a stronger seasonal amplitude than during the two years succeeding.

Figure 7 shows the time series over Hilo (Hawaii), a tropical site. As with Boulder, most of the satellite retrievals capture the seasonal cycles seen in the BFH sonde data. One exception is MLS-Aura at 147 hPa which shows a much weaker amplitude than the BFH sonde and is also drier (MLS-Aura is not unique in this respect though). This feature was noted in a comparison report by Hegglin et al. (2013). A possible explanation for this could be related to the tropopause height dependence of the dry bias seen in the MLS-Aura sonde comparisons. Over the tropical sites, the tropopause is rising and falling by $\sim$1.5 km over the year and thus the MLS-Aura bias also rises and falls with it causing a potential flattening of the annual cycle. Notice in Figure 5 the bias gradient with the tropopause height is rather steep. The mid-latitude locations where the tropopause is near 147 hPa shows a 50–60% dry bias for MLS-Aura at 215 hPa and much smaller 10% dry bias at 147 hPa. For the tropical sites, where the tropopause is near 100 hPa, the dry bias at 215 hPa drops to 10–15% but increases to 40% at 147 hPa. Therefore a seasonally modulating tropopause height would be expected to modulate the MLS dry bias significantly for the level that is 2–3 km below the tropopause or in this case the 147 hPa level and also the levels above and below but to a lesser extent. Subsequent investigation of this bias suggests that it is caused by a pointing difference error between the radiometer that measures water vapor and the radiometer that measures O_2 for pointing. This bias will be corrected in version 5 under development. Early version 5 testing is showing that the current pointing error in v4 is flattening the 147 hPa annual cycle (in contrast to accentuating it).

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

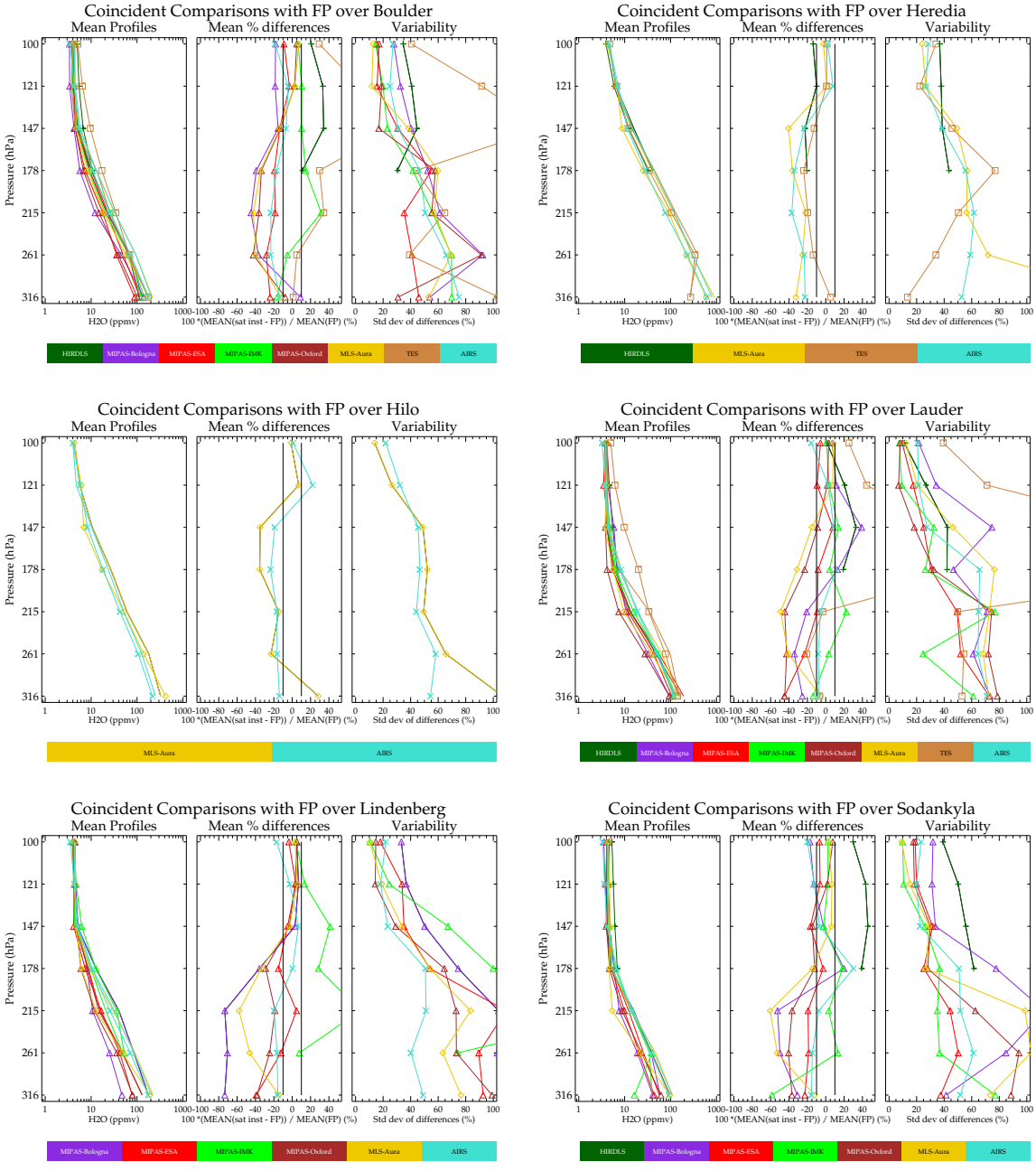


Figure 5. Same as Figure 4 but for different BFH locations. No averaging kernel applied.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

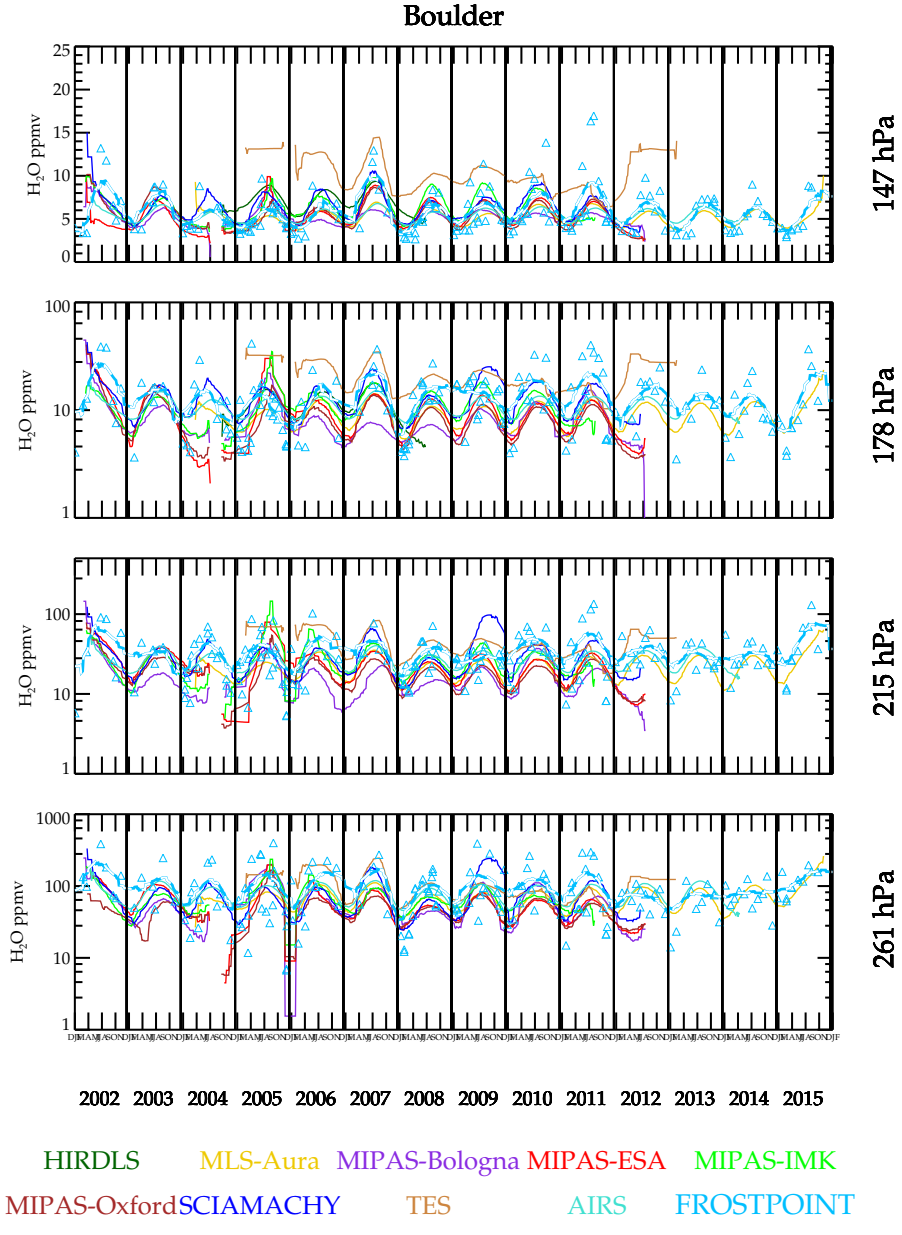


Figure 6. Time series of BFH and satellite/retrievals having enough data coincident with Boulder to produce a smoothed time series curve. The BFH data shows the individual measurements in addition to the smoothed curve. Only smoothed curves are shown for the data sets to remove excessive data clutter.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

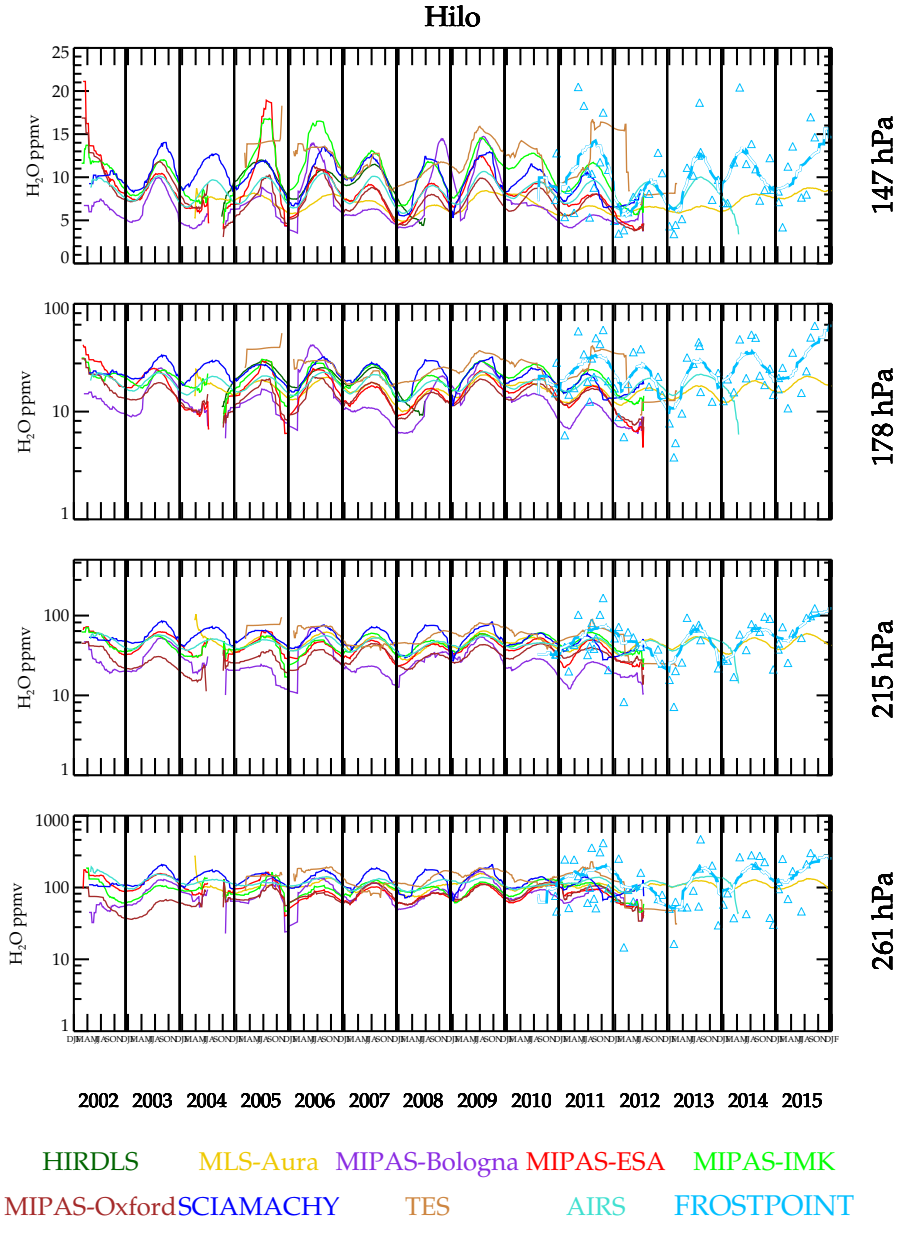


Figure 7. Same as Figure 6, but for Hilo, Hawaii.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

Figure 8 shows a data smoothed time series comparison over Sodankylä, Finland, a high-latitude northern hemisphere site. The BFH shows a weak annual cycle at 147 hPa and stronger ones at lower altitudes. This behavior is captured by most of the data sets.

Figure 9 shows a data smoothed time series comparison over Lauder, New Zealand, a mid latitude southern hemisphere site. The BFH shows an irregular seasonal cycle that in most years is weak at 147 hPa except at the beginning of 2007. Most satellite measurements show larger seasonal cycles with a more regular phasing. The phasing does differ among the satellites.

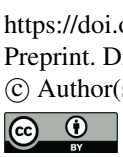
Exploring the question of seasonal amplitudes and phase further, the time series data is fitted to a periodic function that yields a mean value, annual cycle amplitude and phase. Interannual variability is ignored in this fit, thus the result should be viewed as a "climatology". The result for Boulder, USA is shown in Figure 10. The data sets capture the annual cycle with correct phase. Figure 11 shows a comparison of the fitted function to Hilo data. It is noteworthy that MLS-Aura greatly underestimates the seasonal cycle at 147 hPa relative to the other data sets and BFH sondes. This feature is present regardless of whether averaging kernels are applied or not and a likely cause has been identified.

Figure 12 summarizes the results from fitting a periodic function to the coincident data for the 6 sonde sites. Ideally, the left panels in Figure 12 should be similar to the center panel in Figure 5. The difference between these is that Figure 5 is based on location and temporal coincidences whereas Figure 12 is based only on location coincidences and uses a function to interpolate in time. While the former is the better method of comparison, because of the limited number of sonde launches, statistics are sparse and very few instruments can be compared. The fitted time series function approach improves the statistics and a few more instruments can be compared; however, differences in temporal sampling impact the comparison. At Sodankylä, the seasonal cycle at 147 hPa is weak and no instrument captures it well based on the BFH time series fit. Most instruments do much better at the lower altitudes. Over the two tropical sites as noted before, MLS-Aura significantly underestimates the annual cycle amplitude. SMR underestimates the annual cycle at all four altitudes shown here. BFH launched from Lauder also has a weak seasonal cycle at 147 hPa. Most of the satellite instruments, including MLS-Aura tend to overestimate the seasonal cycle relative to the BFH. Also, the phasing in the BFH is irregular and the fit is dominated by the large moist event that occurred in late 2006/early 2007.

Figure 13 shows mean biases between BFH and instrument data sets derived from mean differences of spatial/temporal coincidences and mean value derived from fitting to all data with a periodic function that is only spatially coincident. AIRS and MLS-Aura, the data sets with the best statistics for the coincident comparisons, show the best agreement for a mean derived from a time series fit and a mean from a coincident comparison fit. Even for these data sets, the agreement between the two methods is as large as 20%. The lack of consistency between sonde values and the direct coincidences prompts us not to use derived biases as a proxy for direct coincidences when summarizing results later in this paper.

Figures 14 and 15 show time series and scatter plot comparisons of satellite measurements with Vaisala-RS92 balloon hygrometers over Southern Great Plains, USA, and Sodankylä, Finland. The Vaisala-RS92 sonde uses a capacitance hygrometer that is pre calibrated by the manufacturer prior to launch. These hygrometers are relatively inexpensive and therefore launched more often than BFH. Unfortunately these capacitive hygrometers are not accurate near the tropopause or in the stratosphere. The response time of the capacitive element lengthens as the humidity approaches stratospheric concentrations and thus become

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021



Atmospheric Measurement Techniques Discussions

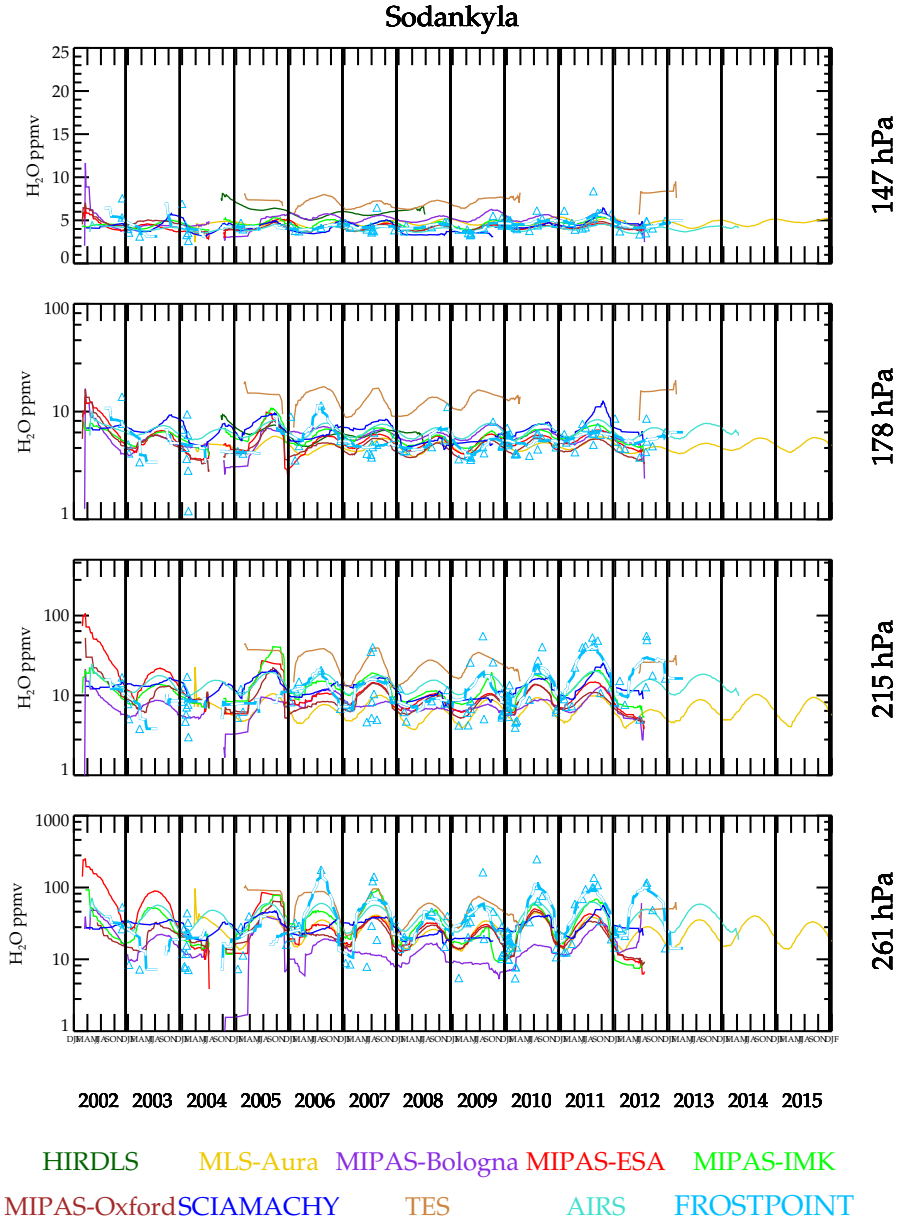


Figure 8. Same as Figure 6, but for Sodankylä, Finland.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

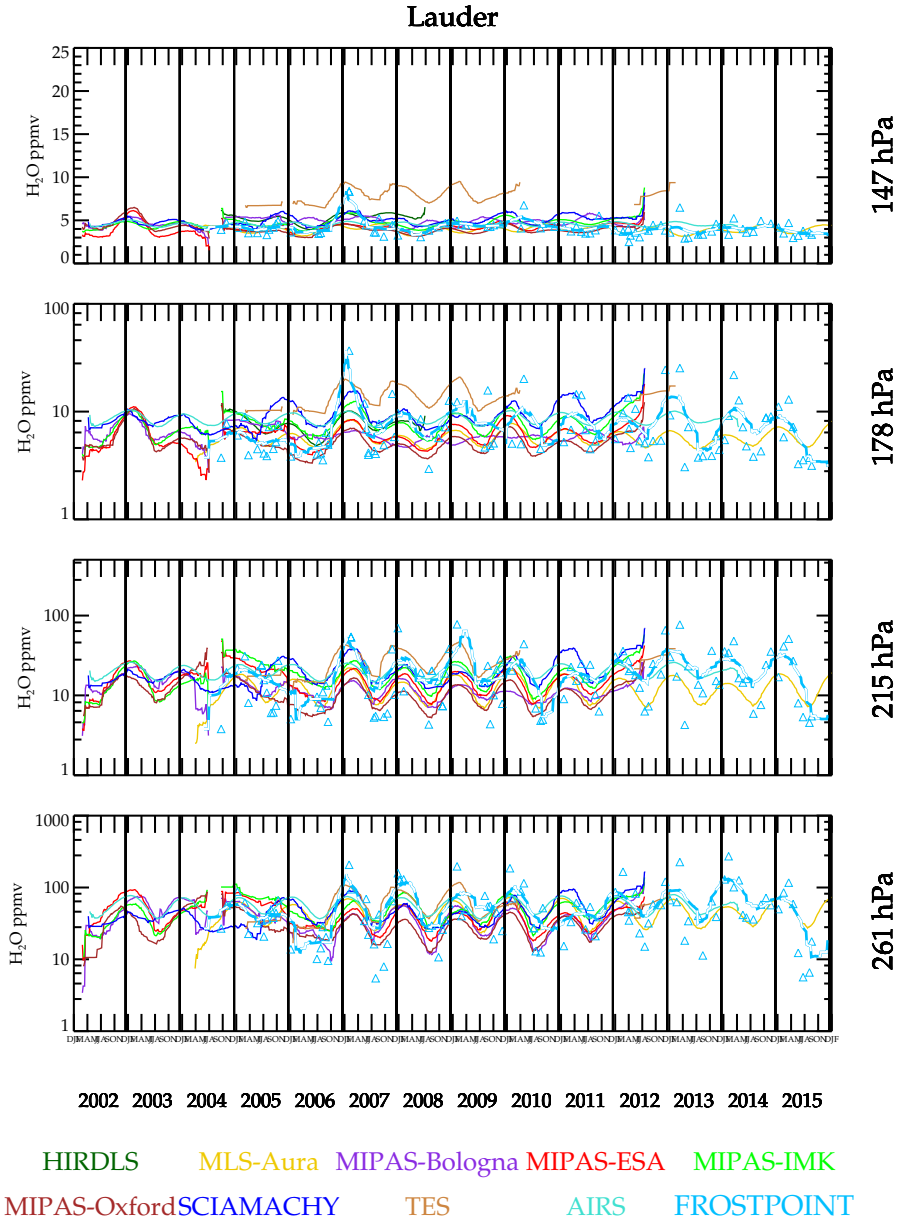


Figure 9. Same as Figure 6, but for Lauder, New Zealand.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

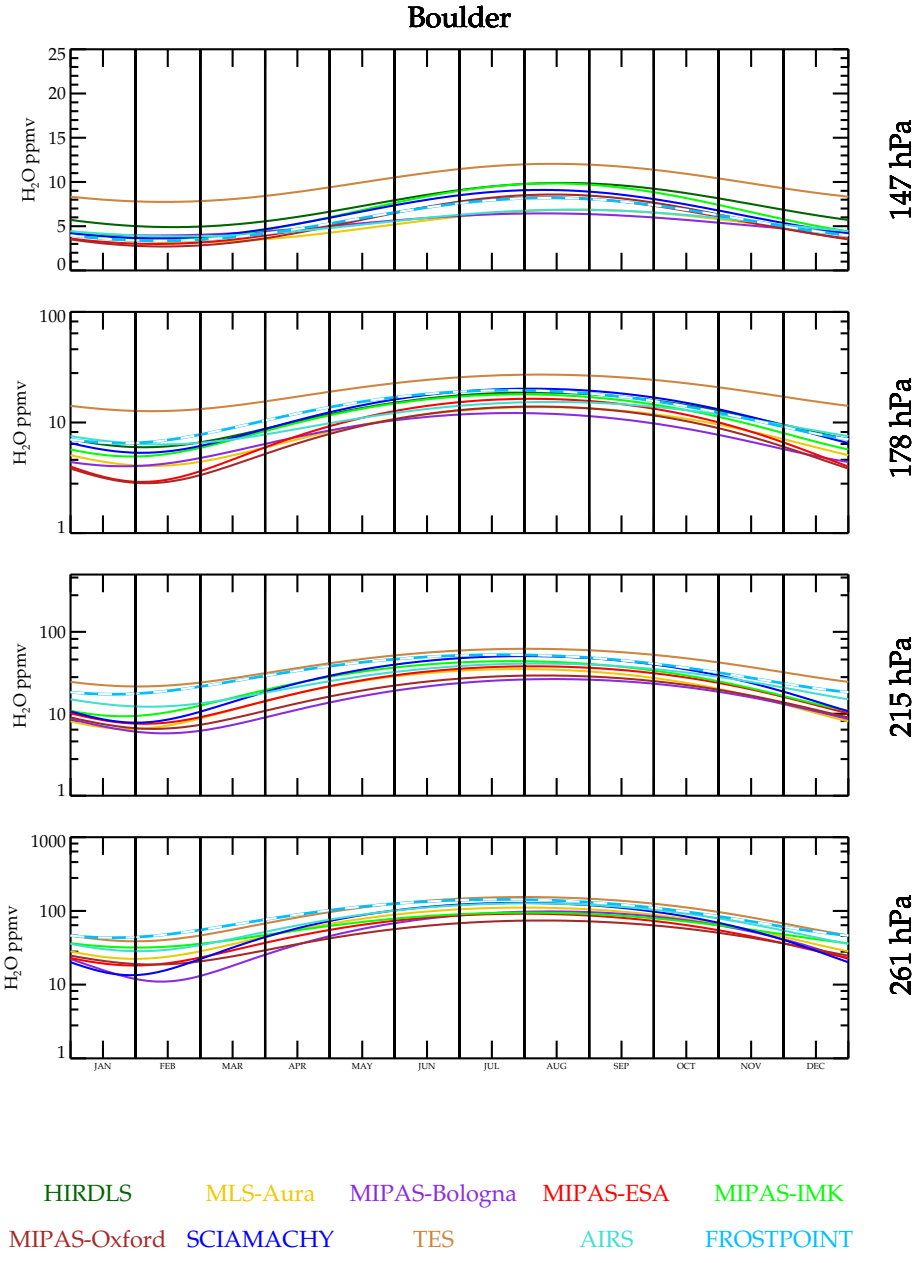


Figure 10. For each data set, a periodic one year function is fitted to the time series data near Boulder, Co. and plotted for one year. Interannual variability is averaged over for each data set. Therefore this figure is a climatology for that data set.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

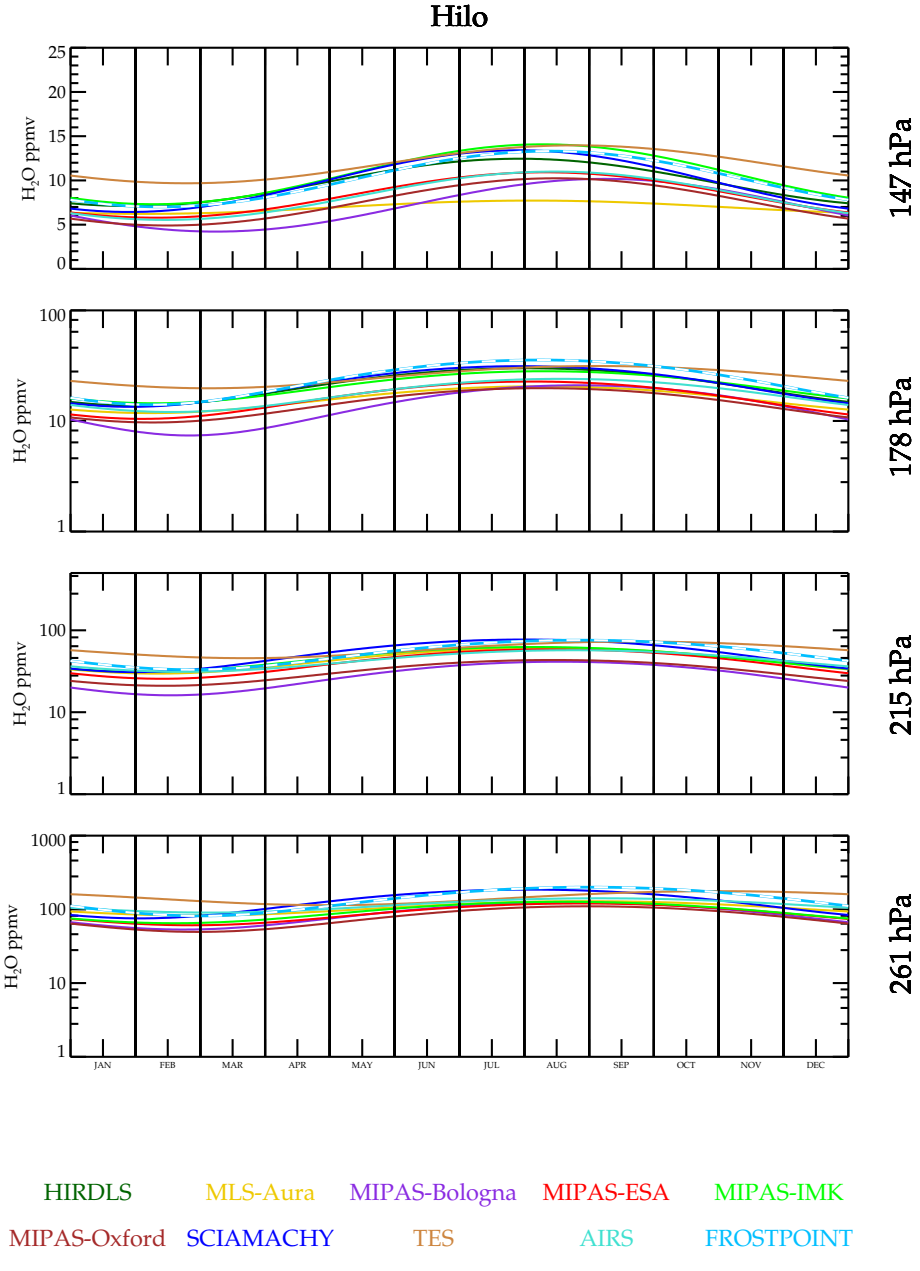


Figure 11. Same as Figure 8, but for Hilo, Hawaii.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

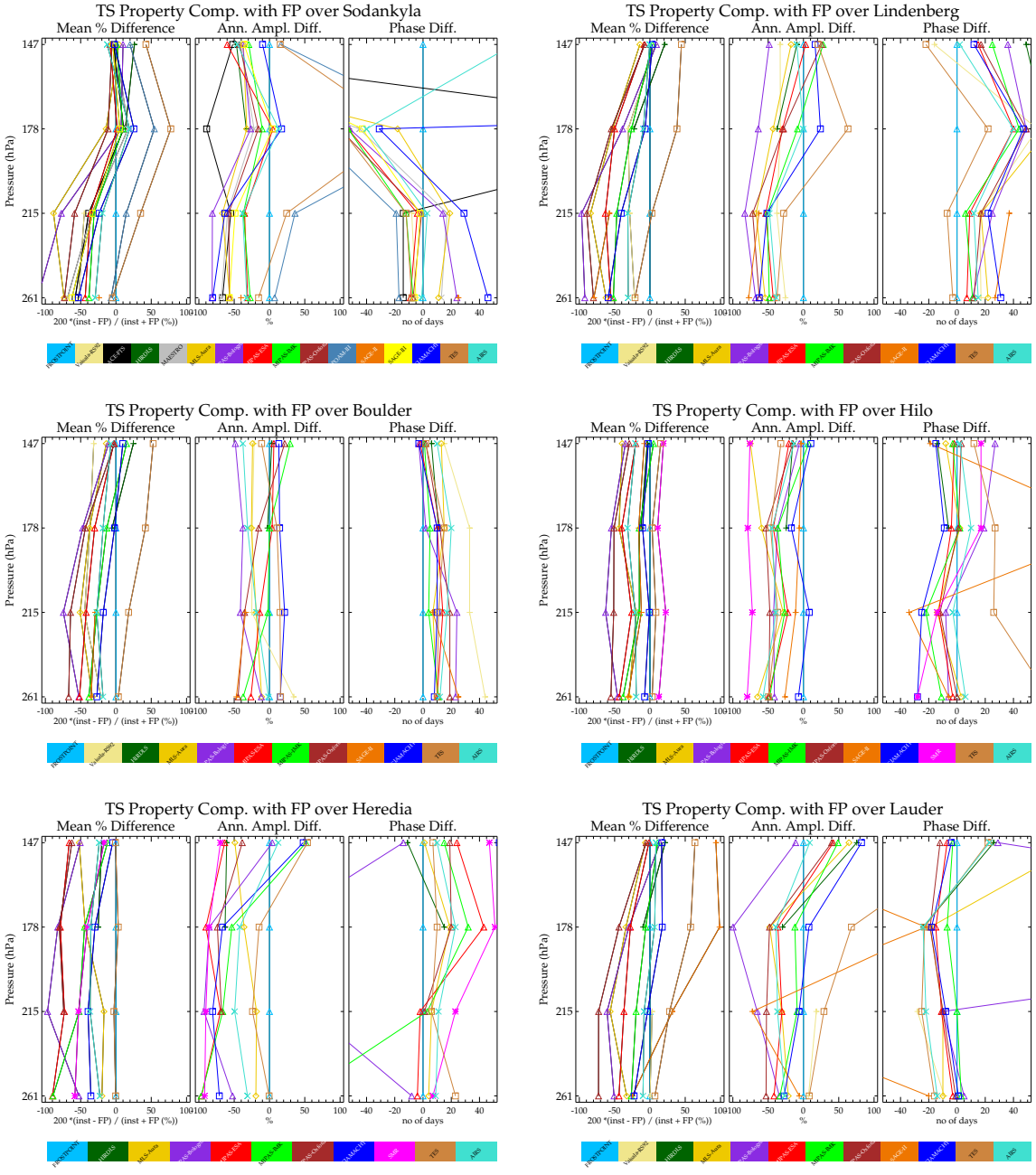


Figure 12. Comparisons of fitted periodic function parameters (mean value (left), amplitude(center), and phase(right)) between a data set (colored line) and BFH sonde as a function of altitude (pressure).

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

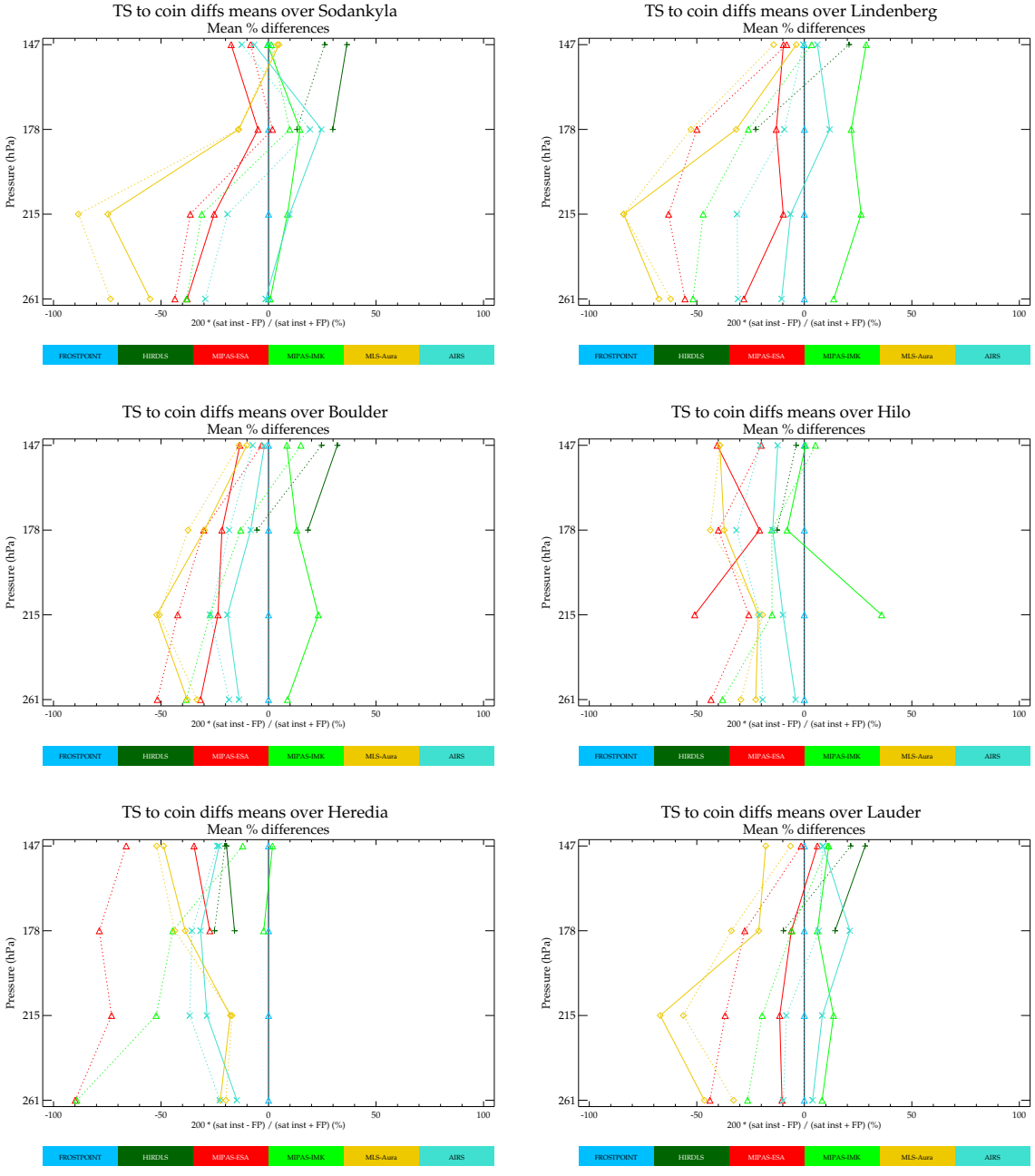


Figure 13. Comparisons of biases computed from direct location and temporal coincidences (solid lines) versus that derived from a time series function using all available data satisfying a positional coincidence criteria (dotted).

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

erroneous under extremely dry conditions. Post processing algorithms have been developed to compensate for this time lag and correct these data based on coincident BFH launches during campaigns (Dirksen et al., 2014). Comparisons here indicate that the Vaisala-RS92, even with corrections, are probably not reliable for concentrations $<$10 ppmv. Figure 14, showing a comparison over the Southern Great Plains region of the US shows reasonable correlations at all levels including 147 hPa. The time series measurements show that during Northern Hemisphere Summer, very high values ($\sim$ 20 ppmv) are often prevalent. It has been shown that summertime deep convection can indeed inject high H_2O into the stratosphere (Anderson et al., 2012; Schwartz et al., 2013) consistent with the concentrations shown here. Therefore one might conclude that the RS-92 with the correction algorithm is successful. But caution is definitely in order here. Launches over Sondankylä (Figure 15) present a different view. Like the Southern Great Plains site, the RS-92 shows a prevalence of very high humidity events occurring during the Northern Hemisphere Summer that are not seen in any Satellite data set. A scatter plot with MLS-Aura is shown because it had the most coincidences, but all the MIPAS retrievals are identical in that there are no high humidity events ($>$ 10 ppmv) seen over Sodankylä at 147 hPa. The dashed lines are an orthogonal distance regression fit to the scatter points. When both both the x and y data sets have large and unknown error, and orthogonal distance regresssion is a reasonable method for determining the correlation. In the case for the 178 and 147 hPa at Sodankylä, the correlation is so poor that a best fit line is meaningless and therefore it is omitted for clarity. The correlation at 178 hPa is also poor, again with the Vaisala-RS92 showing extremely dry and moist events whereas the satellites having measurements usually between 3–8 ppmv. Other sites with frequent launches including Lindenberg, Germany, Ny-Ålesund Norway, Barrow, USA, Cabauw, Netherlands, and Lauder New Zealand behave like Sodankylä. It probably should be assumed, that the highest altitude pressure level for which the RS-92 can be used is $\sim$200 hPa with humidity concentrations exceeding 10 ppmv. It is likely, that some of the high values measured by RS-92 over the Southern Great Plains are accurate measurements because the capacitance sensor responds best to more moist conditions, but given that equally high values are seen elsewhere where they are unlikely to be present makes such measurements suspect. A case in point is Lindenberg where there were a high number of both BFH and RS-92 launches. At 147 hPa, neither the satellite sensors nor the BFH show a measurement exceeding 10 ppmv, whereas the RS-92 shows a large number of them.

4.3 Satellite to Satellite Coincident Comparisons

Coincidences between satellite based data sets are discussed here. Figure 16 shows a coincident match scatter plot comparison between MLS-Aura and ACE-FTS as a probability density function (PDF). Only data below the MERRA-5.2 tropopause height is considered. A relative density amount for each contour is shown in the color bar. Since the number of coincidences decreases with altitude, only data below the tropopause are being compared, the scale is relative. The number in all bins is divided by the number in the bin having the greatest number of points and assigned a color. The solid circles on the thick line define the bins whose H_2O concentrations are values shown on the x axis. The bin values for the y data set are the same as those on the x data set. The thin line is the mean value of the y points within the x-bin, the thick line is the corresponding median value, and the dashed lines are 1-σ standard deviation about the mean value. As can be seen in the plots, ACE-FTS is usually more moist than MLS-Aura.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

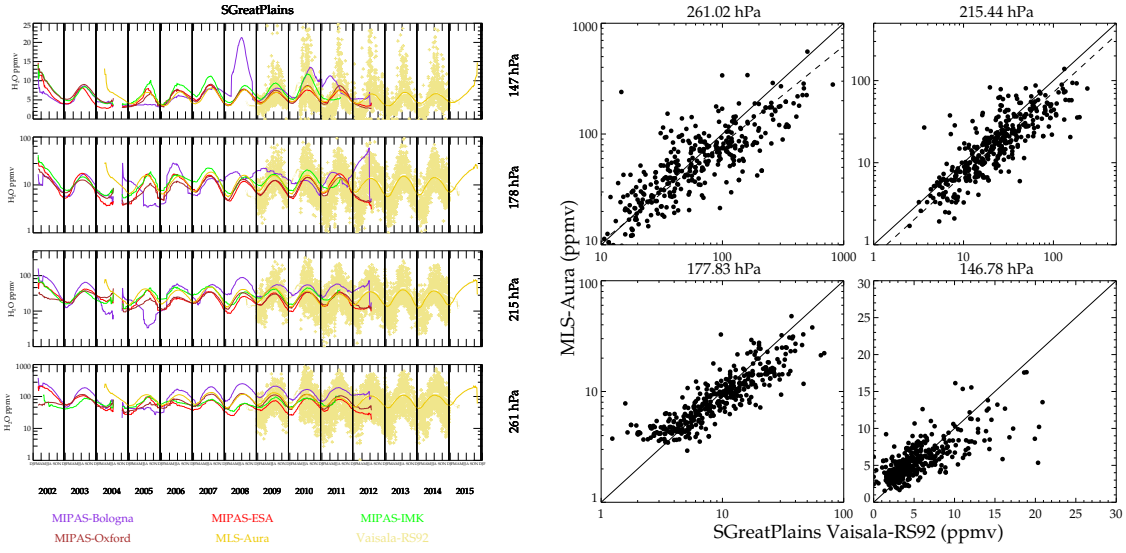


Figure 14. Smoothed time series (left) and scatter plot (right) comparing Vaisala-RS92 hygrometer versus satellite instrument/retrievals at Southern Great Plains USA.

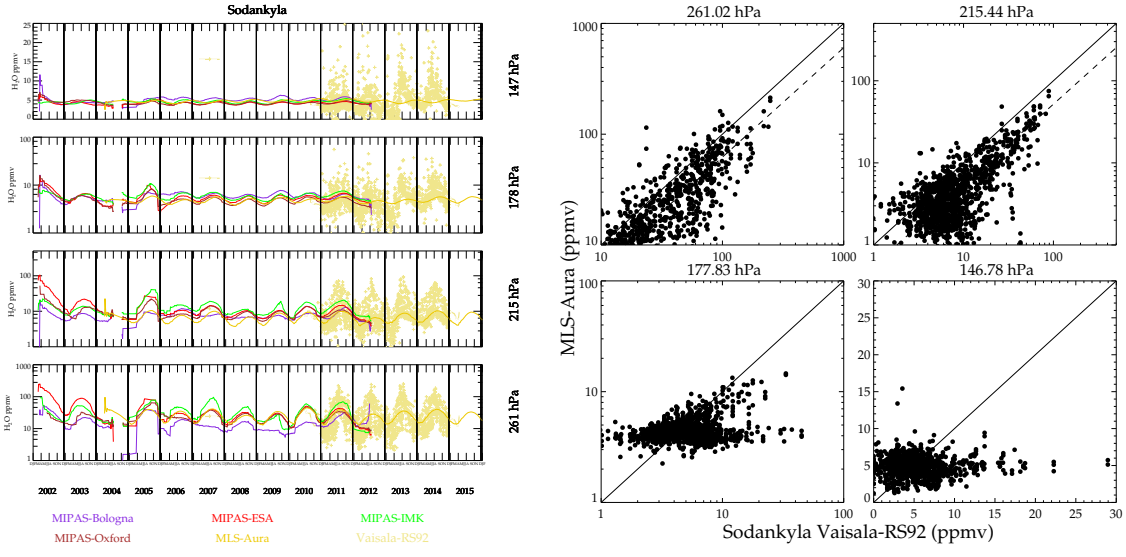


Figure 15. smoothed time series (left) and scatter plot (right) comparing Vaisala-RS92 hygrometer versus satellite instrument/retrievals at Sodanklya, Finland.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

Figure 17 Shows a comparison between MLS-Aura and MIPAS-IMK. Note that there are essentially no coincidences in the tropics due to the 3 hour coincidence criterion (the equator crossing local time for Aura is 13:45 versus 10:00 for ENVISAT). As with ACE-FTS, MLS-Aura tends to be drier. Figure 18 shows a scatter comparison between MIPAS-ESA versus MIPAS-IMK. Since these are different retrievals from the same instrument, all measurements are coincident and all latitudes are covered. This shows that MIPAS-IMK is more humid than the ESA product. One thing that is noteworthy in all these plots is the stretched S shape of the means and medians curve. The lowest x-bins have y values that are more moist and the highest x-bins have drier y values. This feature persists even if the x and y instruments are interchanged. The extreme x-axis value bins are populated with few values and are probably poor retrievals not caught by screening criteria. The corresponding y values are probably better measurements and more accurately represent the atmospheric state. Many plots of this type were generated and are deferred to the supplement. The summary of the results are presented in the conclusions section.

4.4 Gridded Map Comparisons

Gridded map comparison is another method where climatologies can be compared. It has the advantage of not requiring coincidences, and inter measurement coincident matched variability should average down. Its weakness is that sampling biases can significantly affect the comparison. Figure 19 shows 3 month gridded maps of AIRS and MLS-Aura for June 2012–August 2012. The gridding resolution is 5° longitude and 4° latitude in figure 19. This period is shown to highlight a time where MLS-Aura underestimates humidity at 147 hPa over Costa Rica and Hilo, Hawaii (Figure 7). The morphological agreement is excellent. The quantitative agreement will be shown in detail later and is mostly good. An exception here in particular, is the contrast between Central America versus Asia at 150 hPa. Although both instruments show these are moisture rich regions, MLS-Aura shows a much greater moisture difference between Asia and Central America than does AIRS. Based on two tropical BFH sondes sites, MLS-Aura should have a larger dry bias at 150 hPa than AIRS and the mapped field over Central America is consistent with this; however, this behavior doesn't extend throughout the tropics as MLS is considerably more moist than AIRS over Asia. In contrast 300 and 250 hPa fields show MLS-Aura being more humid than AIRS over Central America and Asia.

The fields track the tropopause well where the values become stratospheric like ($<$ 10ppmv) poleward of the tropopause contour. Tropopause tracking is generally better with MLS than AIRS because the limb viewing technique doesn't require an atmospheric thermal gradient to provide spectral contrast in the measurement and MLS most likely has better vertical resolution across the tropopause. Quantitatively, MLS at all levels tends to show higher humidity at the extreme wet regions and more dryness in the desiccated regions than AIRS except over Central America at 150 hPa. These kinds of plots have been made for other months and years and the characteristics of the comparisons are the same despite changing morphologies.

Figure 20 shows a scatter plot of the gridded map values between AIRS and MLS-Aura. The scatter plot provides a more quantitive assessment of the gridded values than can be seen from comparing contour plots. The correlation is good, however the slopes of the correlations are greater than one because MLS-Aura tends to exaggerate the extreme moist and dry values relative to AIRS. The Asian moist bias in the MLS-Aura measurement is evident in the 150 hPa panel.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

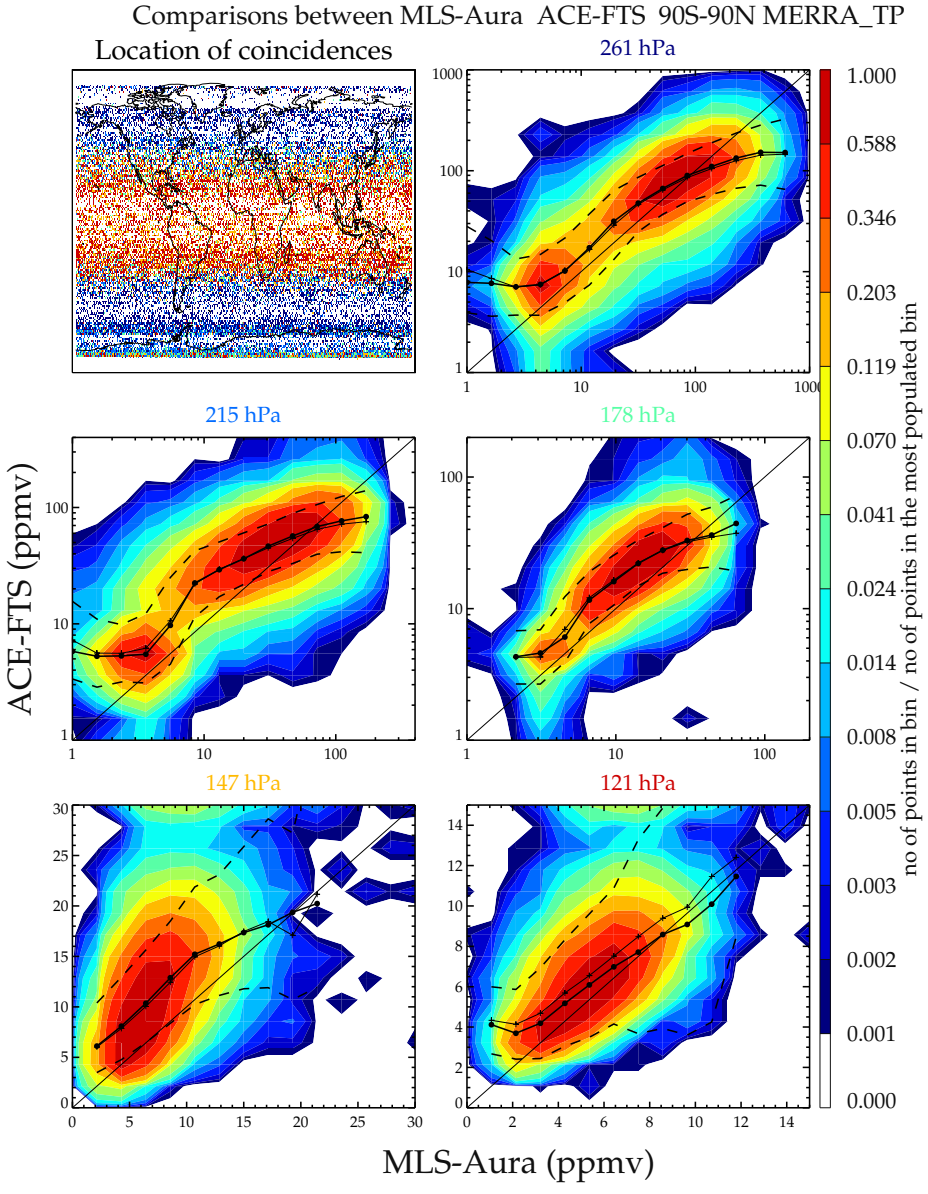


Figure 16. Contour probability density function (PDF) plots of coincident humidity measurements between MLS-Aura and ACE-FTS. Only data below the tropopause is shown. The top left panel shows the location of the coincidences and the points are color coded by the pressure level closest to but below the tropopause height. The color bar shows the ratio of points in an x-y concentration bin to bin with the maximum number of points.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

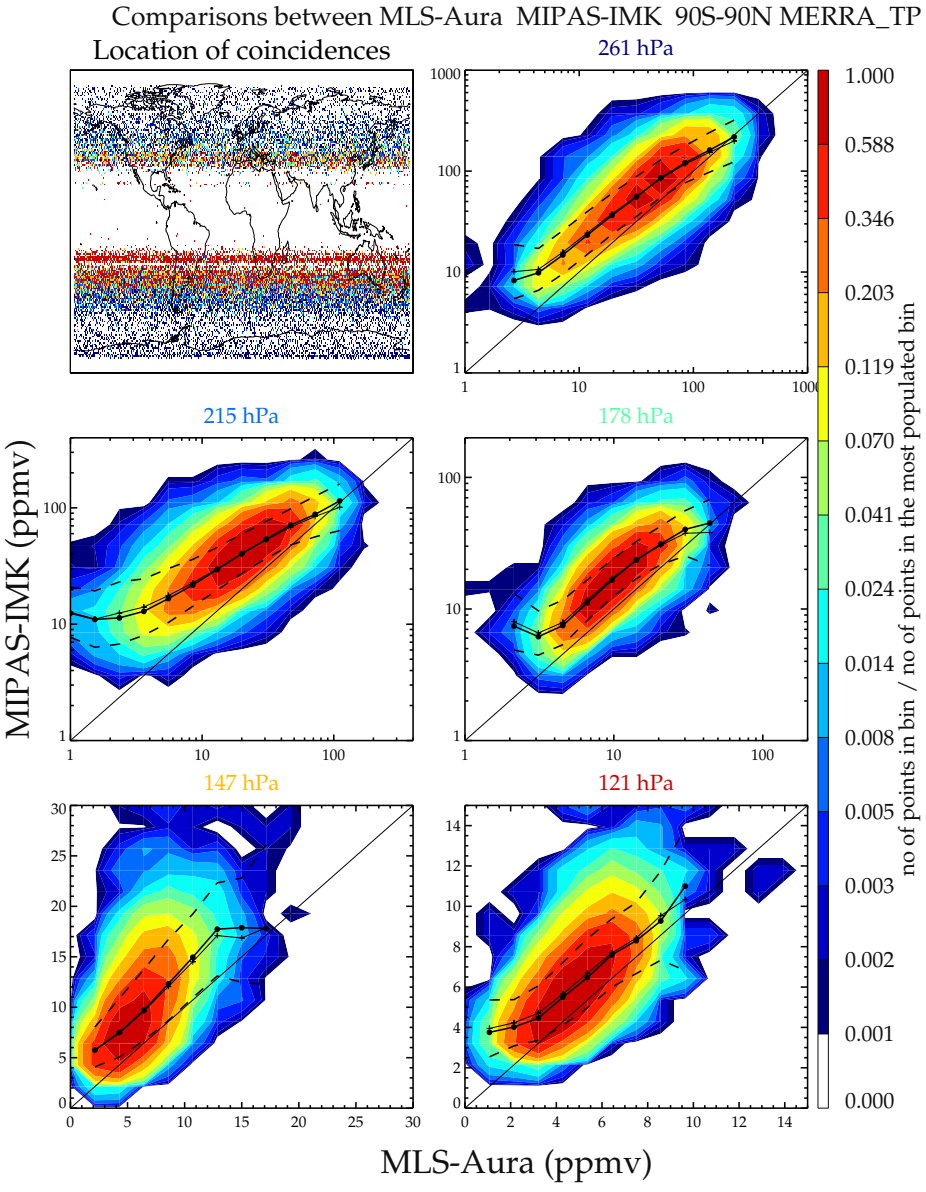


Figure 17. Same as Figure 16 comparing MLS-Aura and MIPAS-IMK.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

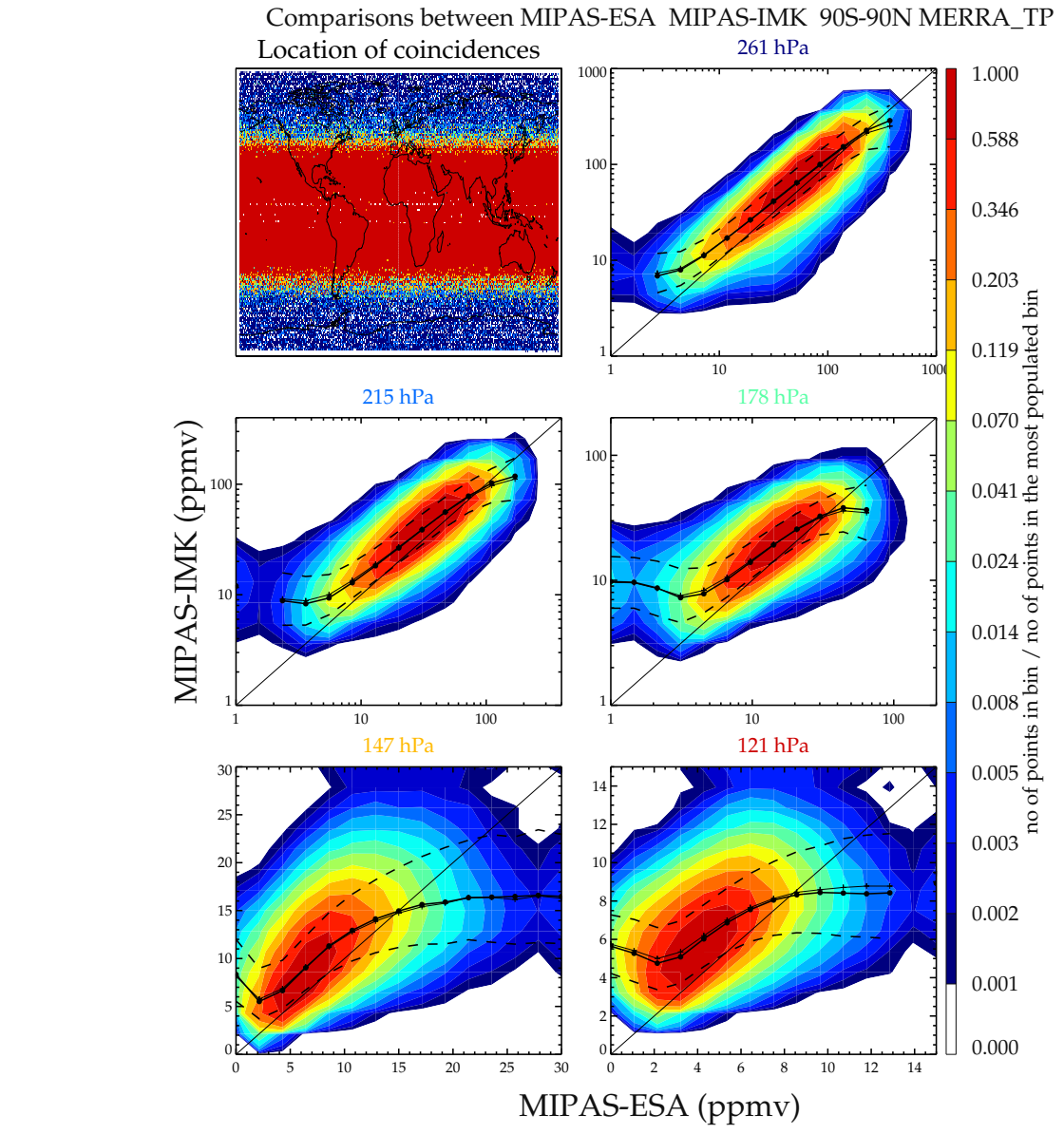


Figure 18. Same as Figure 16 comparing MIPAS-ESA and MIPAS-IMK.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

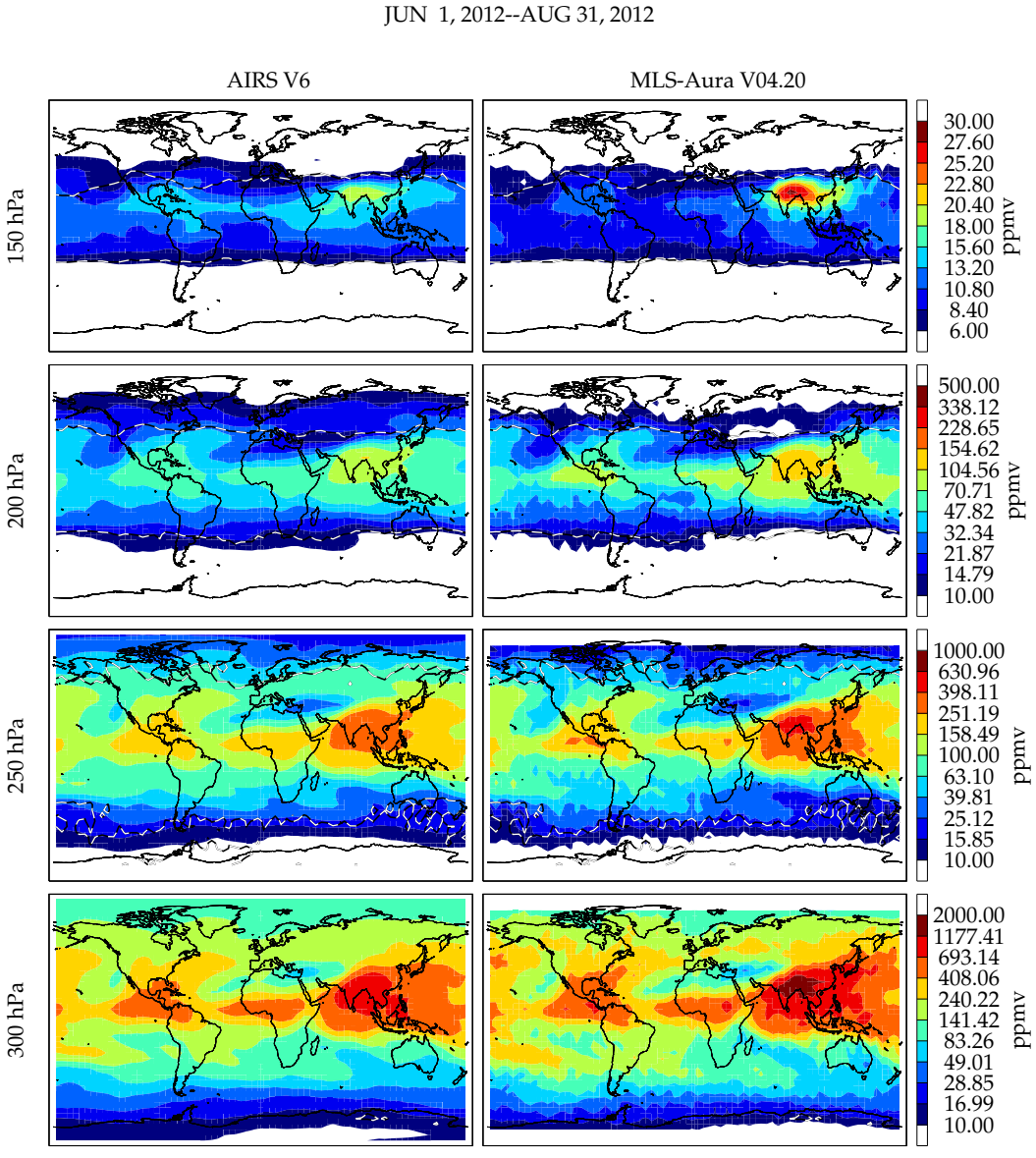
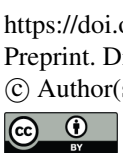


Figure 19. Gridded map comparison between MLS-Aura and AIRS V6 during June–August 2012. The tropoause is indicated by the black and white dashed contour line. Equatorward of this line is in the troposphere and poleward of this line is in the stratopsphere. The white regions are $H_2O < 10$ ppmv. MLS has no measurements poleward of 82°.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021



Atmospheric Measurement Techniques Discussions

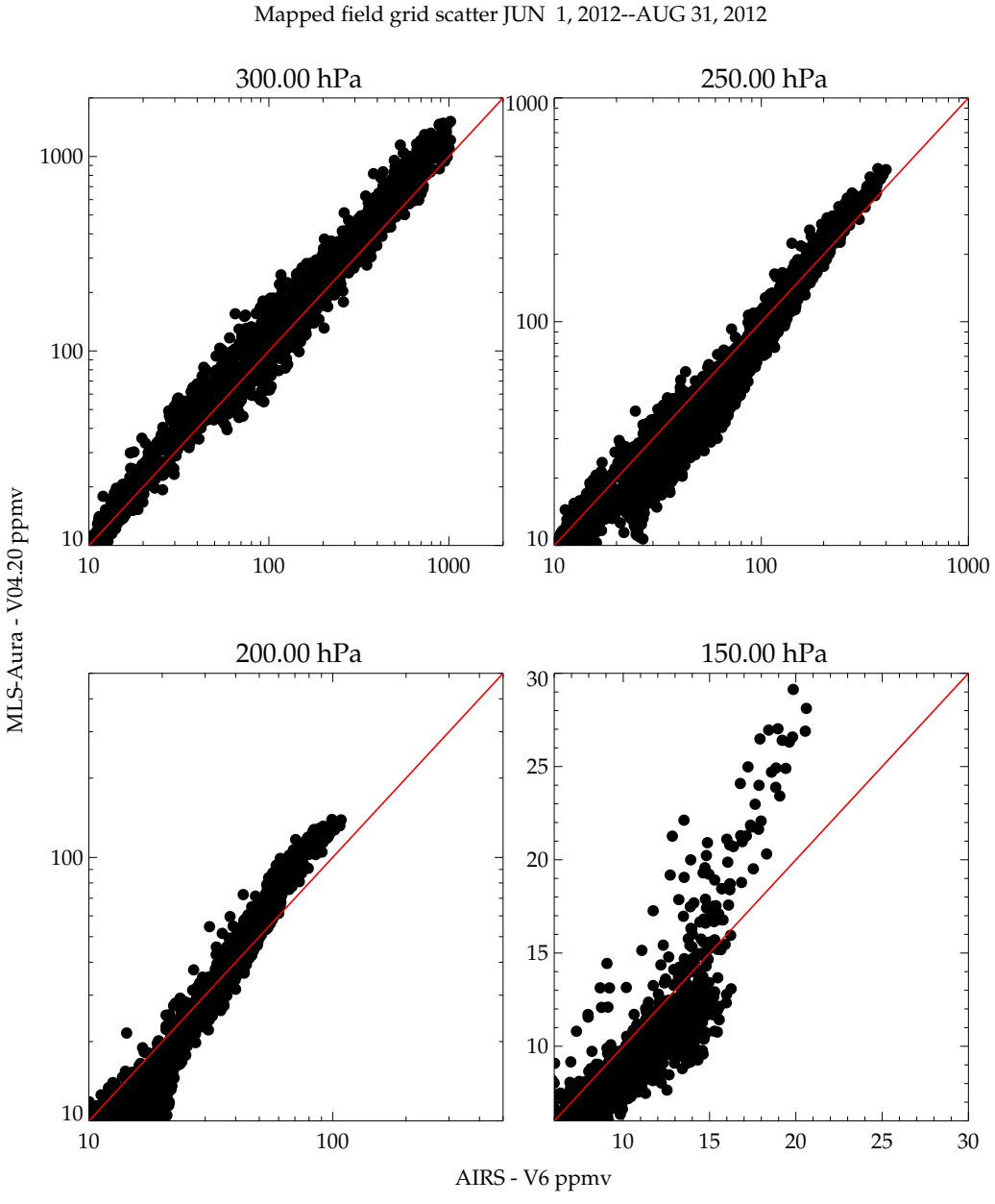


Figure 20. Scatter plot comparison of the gridded box values in Figure 19 between AIRS and MLS-Aura during June–August 2012. The red line is the one to one line.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

Figure 21 shows humidity maps at 175 hPa from 9 other data sets. HIRDLS, MLS-Aura, MIPAS, and SCIAMCHY are limb viewing instruments, SMR (in UTH retrieval mode) and TES are downward viewing instruments. These data sets sample the Earth less frequently than either AIRS or MLS, the grid box size is 10° longitude by 6° latitude. These comparisons fall into two distinct groups, MLS-Aura, SMR, and TES showing very moist tropics with moist features coincident with frequent convective activity over the tropical continents including the maritime, and HIRDLS, the MIPAS suite, and SCIAMACHY showing a more featureless and less moist tropics. These differences are all attributable to cloud impacts. HIRDLS/MIPAS and SCIAMACHY measure infrared and ultraviolet radiation in the limb and are very often cloud contaminated. Their tropical sampling is poor and only the driest, cloud free scenes can be processed. The limb geometry is especially problematic because of the long absorption pathlength in the atmosphere. The result is that the deep tropics are not well sampled for these instruments. The large missing data region in the southern Atlantic Ocean and South America in the SCIAMACHY map is caused by the south atlantic anomaly where this instrument chooses not to make retrievals. The microwave instruments (MLS-Aura and SMR) and the nadir looking infrared TES instrument can better deal with cloudy scenes and therefore show more moisture in the tropics and well defined convective features. These features must be kept in mind when making climatological maps from satellite data. Climatological maps for other heights are presented in the supplement.

Figure 22 shows a scatter plot of the mapped grid values with MLS-Aura on the x-axis and various instruments on the y-axis. The correlation is generally good between the instrument pairs. The MIPAS suite and HIRDLS are drier for moist values relative to MLS for reasons previously described. SCIAMACHY has no measurements in regions associated with active convection probably because the UV backscatter is affected by even thinner clouds than the IR. TES and SMR are more moist than MLS-Aura for all values of humidity. Scatter plots for other heights are shown in the supplement.

Another submillimeter radiometer, SMILES, has dense enough data coverage to produce climatological maps. SMILES operated for 6 months on the International Space Station (ISS). The instrument was not specifically designed to measure H_2O but its radiances are affected by it providing an opportunity for its measurement. Three independent humidity retrievals are available for SMILES using three different approaches. The NICT product retrieves H_2O from the line wing shape in its A and B radiometers. The JPL product fits the radiance growth curve in the window regions of each of its available radiometer bands (A, B, or C) relying on knowledge of the H_2O continuum function. The Chalmers product retrieves from the opaque down looking radiance, similar to its upper tropospheric humidity product on SMR. Table 1 gives the altitude ranges of these retrievals. The supplement has maps showing these comparisons and scatter plots. Using MLS-Aura as a comparison standard all these retrievals show significant biases; however, qualitatively, they do show the same patterns, but over limited altitude ranges. A quick summary shows that the Chalmers retrieval produces good qualitative results from 280–200 hPa, the JPL retrieval does so from 200–125 hPa, and the NICT retrieval from 175–125 hPa. The NICT A band retrieval is much drier than the B band retrieval. The JPL retrievals suffer from high value artifacts at ∼45°S that are not detected in quality screening. As mentioned previously, all these retrievals show significant (> factor of 2) usually moist biases relative to MLS-Aura.

Climatologies for all of 2005 for several occultation instruments are compared to MLS-Aura. The sampling of the occultation instruments is much more sparse than it is for a passive thermal emission instrument. Moreover many of these occultation instruments are set-up in orbits to emphasize coverage in high latitudes in the interest in studying polar ozone chemistry.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

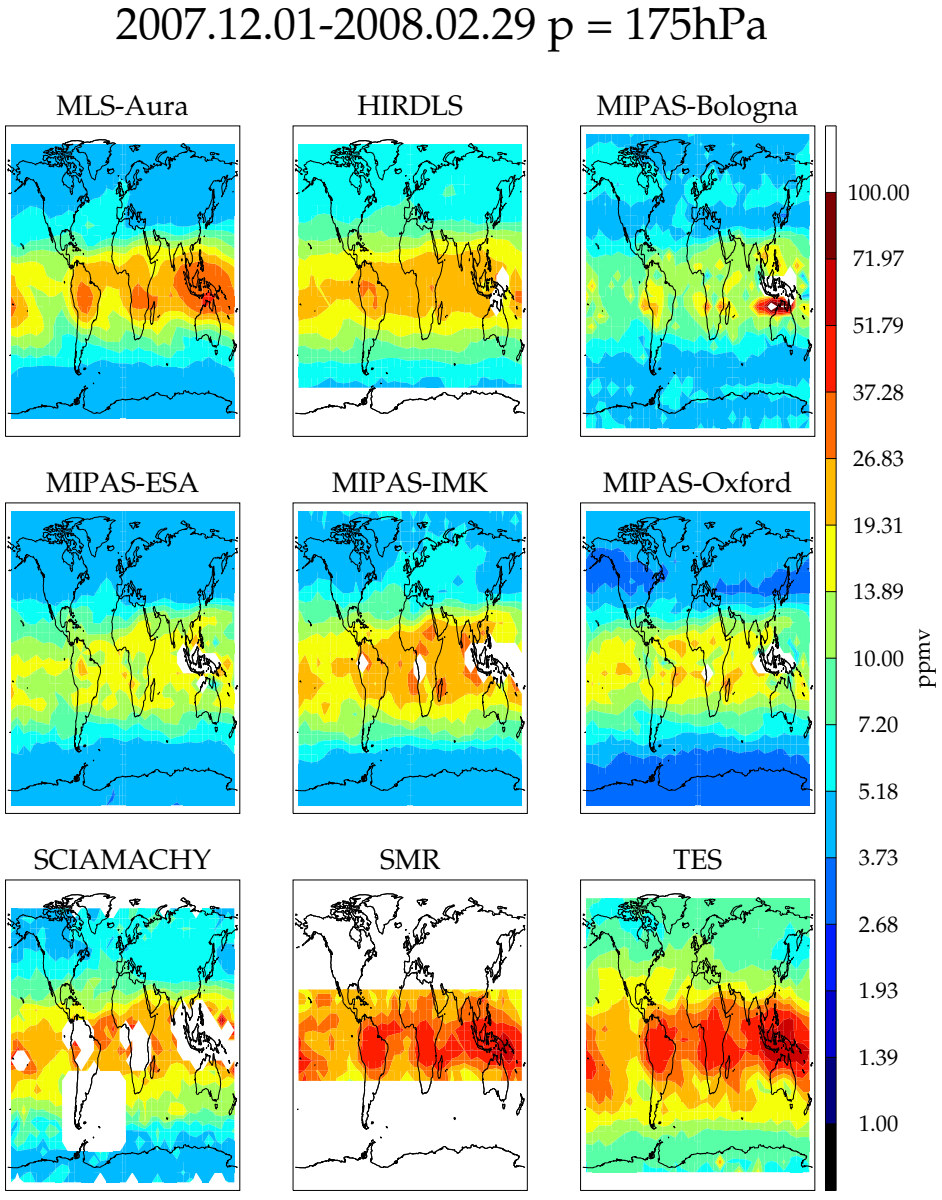


Figure 21. Gridded maps at 175 hPa generated from 9 instruments.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

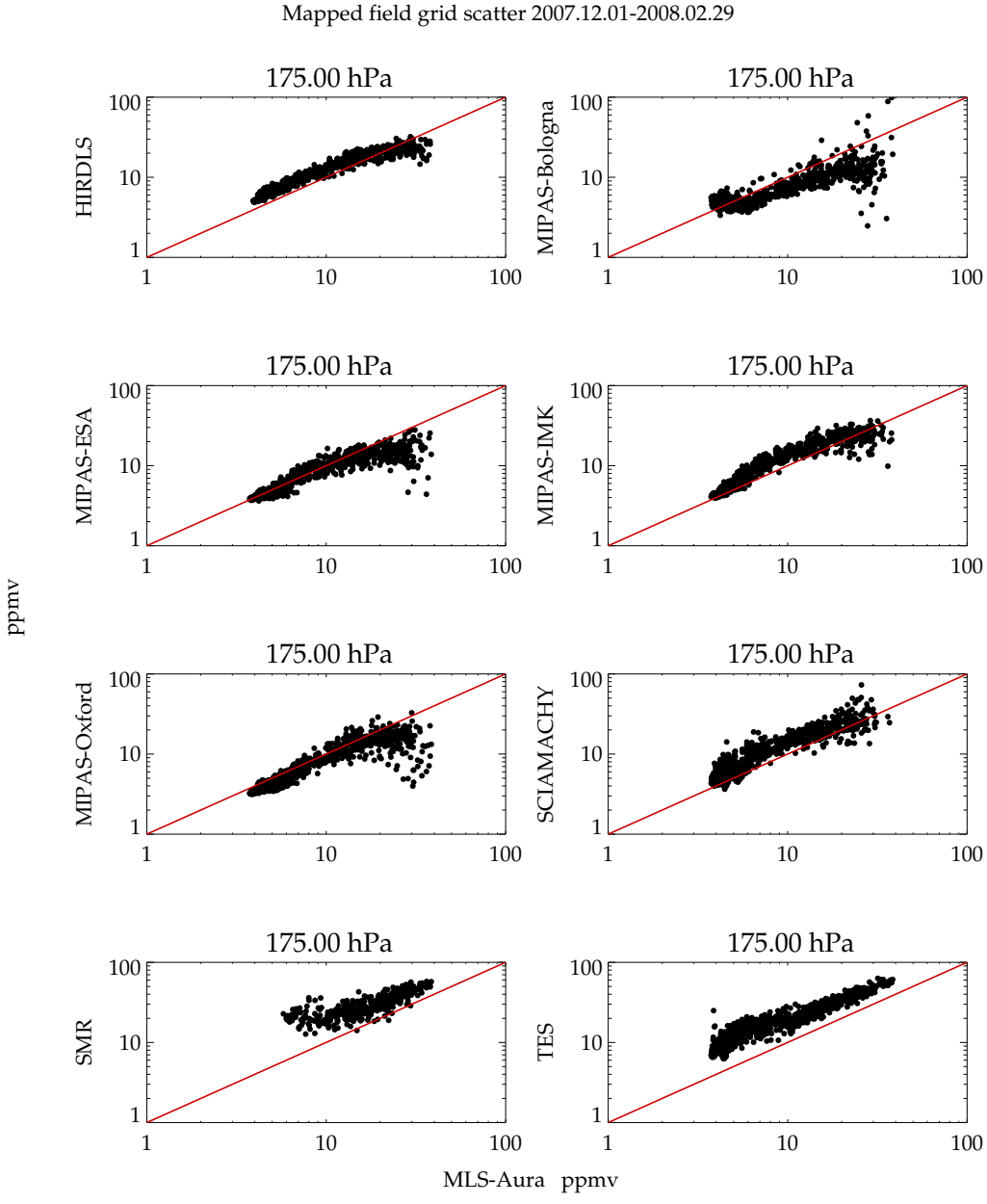


Figure 22. Scatter plot for 175 hPa and Dec. 2007–Feb. 2008 (MLS-Aura versus y-instrument) of the map grid values in figure 21

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

Therefore despite the long gathering period and more coarse grids, such maps have lots of data gaps and inadequate data coverage. Maps and scatter plots for 200 and 150 hPa are in the supplement. For the occultation instruments, the gridded map comparisons are in agreement with the other comparison methods.

5 Conclusions

Figure 23 shows a quick view summary of the comparisons done here. The figure is subsetted into two broad altitude regions, 300–200 hPa and 199–140 hPa. The figure shows results based on the three comparison methodologies, versus BFH sondes, inter satellite coincidences, and mapped grid comparisons. The advantages and disadvantages of these comparison methodologies have been discussed. Moreover, not all three types of comparisons can be done for every data set but by showing three methods, one can bridge one type of comparison (e.g. sonde) over to another (e.g. satellite coincidences). The "zero" difference reference is relative to BFH sonde. The BFH sonde is an in-situ hygrometer with a long historical operational record with an established track record and is currently accepted as an accurate ($\pm$ 10%) hygrometer (Hurst et al., 2011) for measuring humidity down to sub ppmv concentrations. Due to the tight coincidence criterion, (2.5° longitude and latitude, 3 hours time), a small number of instruments have enough coincidences (minimum of 7) to be assessed with spatial and temporal coincidences with BFH sondes. The width of the horizontal bar is a 1-σ spread of the values among the six sondes sites used in this study. Coincidences are available for a subset of the BFH sites for some instruments. Thus those instruments will have less geographical sampling in their assessment. The MIPAS retrieval suite for example has no coincidences with tropically located sondes. A summary of sondes and instruments with suitable coincidences is summarized in Figure 5.

For those instruments for which there are direct sonde comparisons available, a mean bias can be established. For example for MLS-Aura, it is -25% for p>200 hPa and -31% for p<200 hPa. When another instrument is compared to MLS-Aura in a satellite coincident or a gridded map comparison, the MLS-sonde bias is added to those comparison results in order to "correct" for the likely MLS dry bias relative to the BFHs.

The direct satellite to satellite coincidence comparisons use MLS-Aura, MIPAS-ESA, and MIPAS-IMK as the x-axis "reference instrument." Exceptions are MAESTRO which uses ACE-FTS and MLS-Aura for the reference instruments, and MLS-UARS which uses SAGE-II. Since there is no direct coincidence sonde bias estimate for SAGE-II, there is no adjustment applied to that instrument's comparison results. The comparison statistics for the inter-satellite comparisons, in addition to being screened by the tropopause height, were binned by concentration amounts. The following bins 0–10 ppmv, 10–50 ppmv, 50–100 ppmv and >100 ppmv are used. Statistics are computed for each of these binned values and across the satellite reference suite (typically, MLS-Aura, MIPAS-ESA, and MIPAS-IMK). In many cases, results from all bins are not included because doing so would greatly skew the results when it is clear that the measurement of one of the instruments is poor. For example, the x-instrument's retrieved values establishes the bin values e.g. H_2O >100 ppmv. Such values are sparse and often outliers within that instruments retrieval and are probably overestimations of the true concentrations. The y-instrument values that are coincident will be considerably less because they are better quality retrievals in those instances. These are easily identified in plots like figures 16–18, where the correlation curve tends to zero slope. The same is also true for the low value bin <10 ppmv

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

for instruments (MLS-UARS, p>200 hPa,SMILES-JPL, SMILES-Chalmers, SMR, and TES), that are not capable of measuring such low values. The paranthetical instruments are either nadir like sounders requiring a thermal gradient, or retrieve directly from the H_2O continuum. Dry stratospheric values are often near where the thermal lapse rate is small or its signal is dominated by other atmospheric continuum contributors like N_2 and O_2, both being sensitive to spectroscopic systematic errors.

Gridded map comparisons are handled similarly to the coincident satellite comparisons. MLS-Aura is used for the reference instrument in all cases except for assessing MLS-Aura itself. In that case AIRS is the reference instrument. Examples of gridded maps and their grid value scatter plots are shown in Figures 19–22. Like the coincident satellite comparisons, the scatter plots statistics are derived from concentration bins set by the reference instrument. These are $H_2O < 10$ ppmv, 10–50 ppmv, 50–100 ppmv and >100 ppmv. These comparisons are performed for 3 month climatologies for DJF 2007/8– SON 2008, except for SMILES which was done from DJFM 2009/10 because SMILES operated for a 6 month period in 2009–2010. As for the satellite coincidences, some bins were not included in the statistical assessment, Comparisons involving AIRS, SMILES-JPL, SMILES-Chalmers, and TES disregard results from the $H_2O < 10$ ppmv bin. The infrared and UV-Vis limb instruments are significantly cloud contaminated in the tropics and therefore their sampling is greatly reduced there relative to the reference instrument MLS-Aura. Measurements from MLS-Aura, TES, AIRS, and SMR, show that the cloud-impacted grids are the most moist. Therefore comparison statistics in H_2O bins > 100 ppmv at 250 hPa and H_2O bins > 50 ppmv at 200–150 hPa are disregarded for the MIPAS suite and SCIAMACHY. After the comparison statistics are computed, they are shifted by the MLS-Aura dry bias relative to BFH as shown in Figure 23. The spread of values represent a 1-sigma spread of the computed statistics for the pressure levels, H_2O bins, and seasons evaluated.

Figure 23 attempts to show possible satellite and Vaisala-RS92 biases relative to BFH sondes with the assumption that the BFH represents the best accuracy standard for measuring humidity in the upper troposphere and lower stratosphere. The BFH hygrometer itself is considered to be accurate to 10%. Figure 23 is the upper tropospheric equivalent to Figure 1 in the first assessment report (Kley et al., 2000) summarizing the stratospheric humidity sensors in the pre 2000 era. The spread in the variability bar shown arises due to many factors such as location and concentration dependencies, sampling differences, possible averaging kernel smoothing effect dependencies on profile shape and many possible systematic error contributions such as errors in atmospheric temperature and interfering species whose errors may not be uniform under all conditions that these comparisons are made. Although an attempt has been made to reference these biases relative to the BFH, there are some inconsistencies. The mapped comparisons between MLS-Aura and AIRS typically show agreement within $\pm$20% for H_2O bins, pressures and seasons considered. However, when the MLS-Aura dry bias relative to sonde is added it suggests that a climatological map produced by AIRS should have an overall dry bias of 20%. Whereas the same gridded map comparison for MLS-Aura which is based on the same comparison with AIRS except that AIRS is the reference measurement shows only a slight (<10%) dry bias. This is because the AIRS to BFH adjustment is -2% and -6% for the higher and lower pressure ranges in Figure 23. The cause of the differences relative to the BFH reference arises from MLS showing a strong bias dependence based on the height of the tropopause. In short for pressure levels considered here the MLS bias runs between near 0 to 60% when the tropopause is 2–3 km above the compared pressure level. The adjustment is roughly an average of these conditions.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

AIRS does not show this behavior and therefore its bias adjustment is not tropopause height dependent and is therefore more robustly applicable. What is not included in Figure 23 for the satellite coincident comparisons is an additional scatter resulting from the variability of paired differences and for the gridded maps, paired grid box value difference variability. These are typically ∼30% for these comparisons and therefore an additional ∼30% variability would be added to that shown in Figure 23 if one is to compare a single matched pair comparison.

Upper tropospheric H_2O is a highly variable field in space and time. In the atmosphere, H_2O can vary by a couple of orders of magnitude. Figure 23 shows that for most of the instruments, their comparisons among themselves and with BFH sondes are indicating mean agreement within ∼30% but with large spreads suggesting something like a factor of two agreement. Relative to stratospheric comparisons where H_2O is well mixed and it is possible to quantify biases to within a few percent, for the upper troposphere such a precise assessment cannot be realized. The problem is that the measurements sample atmospheric volumes differently where concentration gradients are large. The measurement systems have non linear responses to changes in water vapor amounts. The retrievals also require temperature in their inversion that also may have large vertical gradients. In short it is probable that a comparison between two satellites or with balloon sondes (discussed earlier) will show different degrees of agreement for a large ensemble of coincident data making it not possible to establish a single bias number by height and latitude.

In closing some features specific to certain instruments will be discussed. It is clear from the gridded map comparisons that high clouds in the tropical upper troposphere have a significant impact on infrared–ultra violet limb viewers (see Table 1). While the limb geometry allows low concentrations of H_2O to be measured and doesn't require a thermal lapse, the long horizontal path length makes cloud encounters much more likely. The MIPAS retrieval suite and SCIAMACHY demonstrated good agreement with mid and high latitude sondes; however, their clear sky sampling limitation causes a severe undersampling of the tropics leading to a dry bias. This limitation was so severe that for Figure 23, moist value bins were not included in the assessment summary. Of course this limitation needs to be kept in mind for science investigations. The microwave limb viewers MLS-Aura/UARS, SMR, and SMILES are more immune to clouds due to the longer measurement wavelength being less subject to cloud emission and scattering. Nadir sounding geometries work better than limb in cloudy scenes because the imaged scene is small compared to the horizontal distance covered by limb viewer and can look at scenes in close proximity to clouds without being contaminated by them. Also AIRS by using highly spatially resolved pixels can use a cloud clearing scheme to derive a cloud free signal. Therefore AIRS and TES although being infrared instruments can better observe in cloudy regions and avoid the severe sampling bias. As mentioned before, these instruments have a relatively short path length in the atmosphere and require a thermal lapse to measure humidity. Therefore they are unable to make measurements ∼3 km below the tropopause and above or where H_2O concentrations are <10-20 ppmv.

Among the limb viewers, MLS-Aura has the highest daily sampling, one of the longest running operations (still in operation) and is the least affected by clouds. Therefore it was probably one of the better instruments to use as a reference for comparing the others which was often done in this study. Having said that, the one significant feature is that MLS-Aura shows a significant dry bias (50%) in any level that is ∼2.5 km below the tropopause. This bias reduces to <20% above and below this critical level. The bias behavior is not caused by retrieval smoothing that can be corrected by including the averaging kernels. The

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

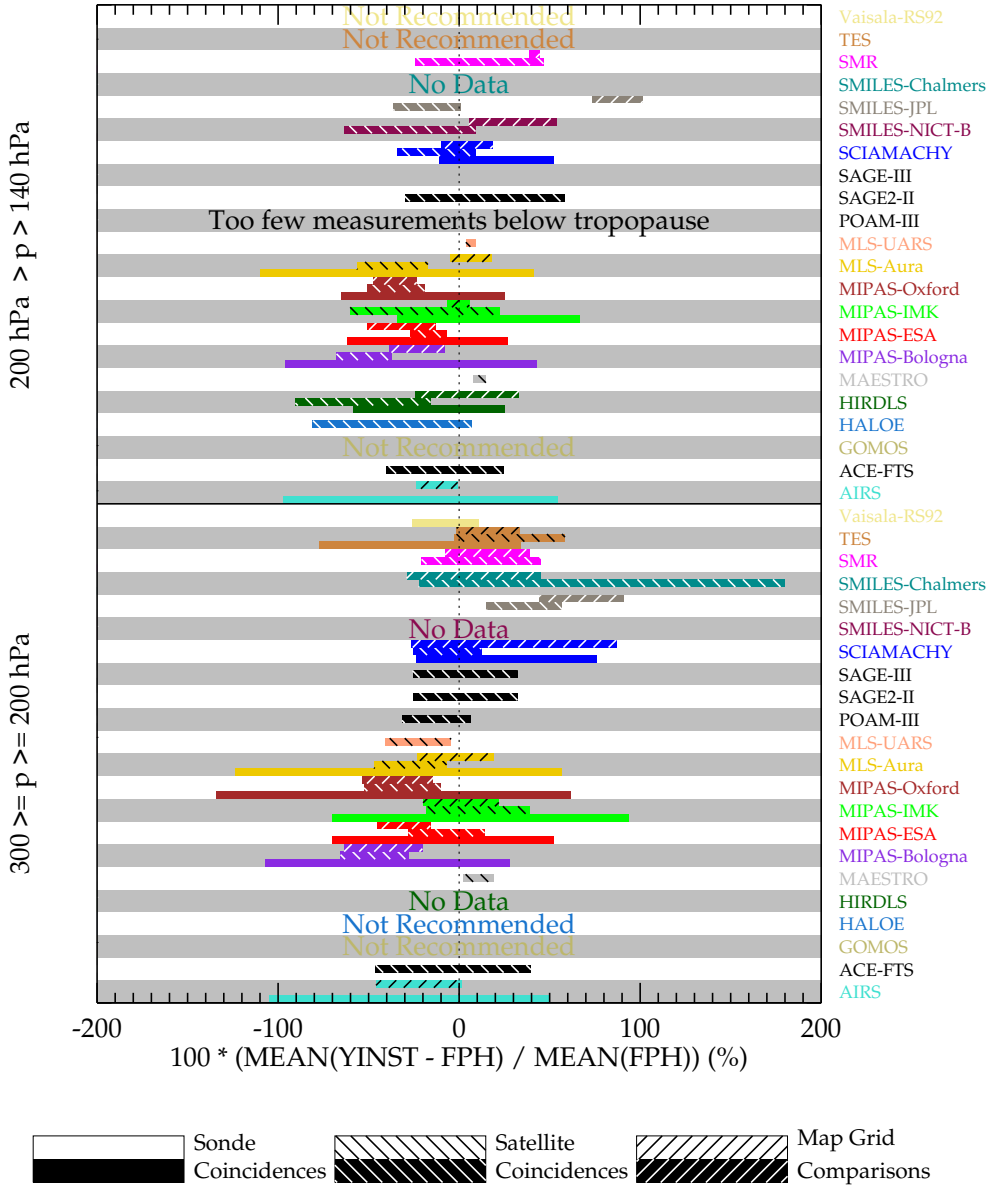


Figure 23. A summary plot of biases among the satellite and Vaisala-RS92 sensor for the upper troposphere relative to BFH measurements (zero value in the x-axis). The sonde coincidences have no hashes and shown in the bottom third of the y-axis band dedicated to the dataset. The satellite coincidences and mapped grid comparisons have either black or white diagonal hashes depending on the darkness of the data sets color and are shown in the middle and top third of the y-axis dedicated to the data set.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

cause is most likely due to a pointing error in the retrieval system. The pointing error is from a combination of two sources, one being a field-of-view alignment measurement and another from a sideband measurement. The needed adjustment for the field-of-view alignment is within its pre-launch uncertaianty, but the sideband adjustment is ∼15 times larger than its prelauch uncertainty which confounded its discovery. Version 5 currently in production corrects for these deficiencies will be shown in a future publication.

The occultation sounders can provide accurate profile measurements, ACE-FTS being the best amongst them in terms of sampling a wide range of concentration values, a long operational period (still in operation) and producing accurate measurements. However, the high temporal and spatial variability of H_2O in the upper troposphere along with the sparse sampling from occultation instruments, limits the usefulness of these measurements mostly to validation studies.

Instruments such as SMILES, HIRDLS had short operational lifetimes 6 months and 3 years respectively). The science that can be done with these measurements would be limited to features unique to that instrument. For example, HIRDLS has the best vertical resolution (1 km) among the satellite suite (typically 3 km). SMILES was mounted on the ISS and thus its measurements sample the full diurnal cycle. This was exploited in a cloud study (Jiang et al., 2015). The water vapor products from SMILES are research products for which the instrument was not specifically designed to measure. Although qualitatively the mapped fields are mostly reasonable, biases are large and artifacts present (see the supplement for more detail).

The last observation derived from this study refers to the goodness of the Vaisala-RS92 radiosonde hygrometer in the uppermost troposphere. It is well known that the response time of the humicap sensor in the Vaisala-RS92 slows as the air becomes more desiccated (Miloshevich et al., 2009). This leads to erroneous measurements. Time lag correction algorithms have been applied to some of these sondes and only corrected sondes have been used here. This is in contrast to those used by the radiosonde network that uses an algorithm provided by Vaisala that does not have the time lag correction. The motivation for including the Vaisala-RS92 profiles was to greatly expand the number of Vaisala-RS92 profiles available for more satellite datasets to be compared. Unfortunately, in the uppermost troposphere, the Vaisala-RS92 show inconsistent results and therefore best not used for pressures less than 200 hPa. The agreement is much better for pressures between 300–200 hPa but show a dry bias of 20%. The expanded Vaisala-RS92 data set does allow an assessment to be made for ACE-FTS, SMILES, and SMR. After correcting for the 20% dry bias, and only considering pressure levels $>$ 200 hPa, the mean agreement for ACE-FTS is 8%, SMILES-JPL, -10%, and SMR, 110%. The variability of the differences between the Vaisala-RS92 and the satellite instruments is quite large ∼100%.

In conclusion, with exceptions noted in the text, most of the satellite instruments do a realistic job of tracking upper tropospheric humidity changes. Precise quantitative assessment is much more difficult because the nature of these measurements coupled with the sharp vertical and horizontal gradients in H_2O leads to large variability of the coincident pair differences between the data sets. Even among the MIPAS suite of retrieval products where the four retrieval products are using the same radiance signal sampling exactly the same volume with perfect spatial and temporal coincidence show surprisingly large biases and variability underscoring significant sensitivities to the forward models. Science investigations using these data need to take these features into consideration and be aware of the sampling and measurement durations of these data sets (Table 1). Having said this, and ignoring some notable anomalies (e.g. MLS-Aura large dry bias in a 2–3 km layer below the tropoapuse

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021


Atmospheric Measurement Techniques Discussions

for example) quantitatively, for most of the instruments, agreement within 20–30% amongst each other with an additional variability of 30% is being achieved. The list of instruments that consistently produce acceptable results are: ACE-FTS, AIRS, HIRDLS, MAESTRO, MIPAS-Bologna, MIPAS-ESA, MIPAS-IMK, MIPAS-Oxford, MLS-Aura, MLS-UARS, POAM-III, SAGE-II, SAGE-III, SCIAMACHY, SMR, and TES (P$>$ 200 hPa only). The SMILES suite is borderline in that it produces realistic results but also is subject to erroneous artifacts and can have very large biases and variability. HALOE is only good when the true atmospheric H_2O composition is $<$ 10 ppmv which restricts it to the very uppermost troposphere and therefore is not generally useful for tropospheric humidity. The GOMOS upper topospheric humidity product is not recommended.

Acknowledgements. The work conducted here is done at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. We wish to express our gratitude to SPARC and the World Climate Research Programme for their guidance, sponsorship, and support of the WAVAS-II programme.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

References

Anderson, J. G., M.Wilmouth, D., Smith, J. B., and Sayres, D. S.: UV dosage levels in Summer: Increased risk of ozone loss from convectively injected water vapor, Science, 337, 835–839, doi:10.1126/science.1222 978, 2012.

Dirksen, R. J., Summer, M., Immler, F. J., Hurst, D. F., Kivi, R., and Vömel, H.: Reference quality upper-air measurements: GRUAN data processing for the Vaisala RS92 radiosonde, Atmos. Meas. Tech., 7, 4463–4490, doi:10.5194/amt–7–4463–2014, 2014.

Eriksson, P., Rydberg, B., Johnston, M., Murtagh, D. P., Struthers, H., Ferrachat, S., and Lohmann, U.: Diurnal variations of humidity and ice water content in the tropical upper troposphere, Atmos. Chem. Phys., 10, 11 519–11 533, https://doi.org/10.5194/acp-10-11519-2010, 2010.

Hegglin, M. I., Tegtmeier, S., Anderson, J., Froidevaux, L., Fuller, R., Funke, B., Jones, A., Lingenfelser, G., Lumpe, J., Pendlebury, D., Remsberg, E., Rozanov, A., Toohey, M., Urban, J., von Clarmann, T., Walker, K. A., Wang, R., and Weigel, K.: SPARC Data Initiative: Comparison of water vapor climatologies from international satellite limb sounders, J. Geophys. Res., 118, 11,824–11,846, doi:10.1002/jgrd.50 752, 2013.

Hurst, D. F., Oltmans, S. J., Vömel, H., Rosenlof, K. H., Davis, S. M., Ray, E. A., Hall, E. G., and Jordan, A. F.: Stratospheric water vapor trends over Boulder, Colorado: Analysis of the 30 year Boulder record, J. Geophys. Res. Atmos, 116, D02 306, 2011.

Jiang, J. H., Su, H., Zhai, C., Shen, T., Wu, T., Zhang, J., Cole, J., von Salzen, K., Donner, L., Seman, C., Genio, A., Nazarenko, L., Dufresne, J., Watanabe, M., Morcrette, C., Koshiro, T., Kawai, H., Gettelman, A., Millán, L., Read, W., Livesey, N., Kasai, Y., and Shiotani, M.: Evaluating the diurnal cycle of upper tropospheric ice clouds in climate models using SMILES observations, J. Atmos. Sci., 72, 1022–1044, doi:10.1175/JAS–D–14–0124.1, 2015.

Kley, D., Russell III, J. M., and Phillips, C.: SPARC assessment of upper tropospheric and stratospheric water vapour, SPARC Report No. 2 WCRP-113, WMO/ICSU/IOC, CNRS, Verrières le Buisson, 2000.

Livesey, N. J., Read, W. G., Wagner, P. A., Froidevaux, L., Lambert, A., Manney, G. L., Valle, L. F. M., Pumphrey, H. C., Santee, M. L., Schwartz, M. J., Wang, S., Fuller, R. A., Jarnot, R. F., Knosp, B. W., Martinez, E., and Lay, R. R.: Earth Observing System (EOS) Aura Microwave Limb Sounder (MLS) Version 4.2x Level 2 Data Data Quality and Description Document, Tech. Rep. JPL D-33509 Rev. D, Jet Propulsion Laboratory, 2018.

Marenco, A., Thouret, V., Nédélec, P., Smit, H., Helten, M., Kley, D., Karcher, F., Simon, P., Law, K., Pyle, J., Poschmann, G., Wrede, R. V., Hume, C., and Cook, T.: Measurement of ozone and water vapor by Airbus in-service Aircraft: The MOZAIC airborne program, An overview, J. Geophys. Res., 103, 25,631–25,642, 1998.

Millán, L. F., Livesey, N. J., Santee, M. L., and von Clarmann, T.: Characterizing sampling and quality screening biases in infrared and microwave limb sounding, Atmos. Chem. Phys., 18, 4187–4199, doi:10.5194/acp–18–4187–2018, 2018.

Miloshevich, L. M., Vömel, H., Whiteman, D. N., and Leblanc, T.: Accuracy assessment and correction of Vaisala RS92 radiosonde water vapor measurements, J. Geophys. Res., 114, D11 305, doi:10.1029/2008JD011 565, 2009.

Read, W. G., Waters, J. W., Wu, D. L., Stone, E. M., Shippony, Z., Smedley, A. C., Smallcomb, C. C., Oltmans, S., Kley, D., Smit, H. G. J., Mergenthaler, J., and Karki, M. K.: UARS MLS Upper Tropospheric Humidity Measurement: Method and Validation, J. Geophys. Res., 106, 32,207–32,258, 2001.

Read, W. G., Schwartz, M. J., Lambert, A., Su, H., Livesey, N. J., Daffer, W. H., and Boone, C. D.: The Roles of Convection, Extratropical Mixing, and In-Situ Freeze-drying in the Tropical Tropopause Layer, Atmos. Chem. Phys., 8, 6051–6067, 2008.

https://doi.org/10.5194/amt-2021-300
Preprint. Discussion started: 5 November 2021

Atmospheric Measurement Techniques Discussions

480 Schwartz, M. J., Read, W. G., Santee, M. L., Livesey, N. J., Froidevaux, L., Lambert, A., and Manney, G. L.: Convectively Injected Water Vapor in the North American Summer Lowermost Stratosphere, Geophys. Res. Lett., 40, 2316–2321, doi:1002/grl.50 421, 2013.

Walker, K. A. and Stiller, G. P.: The SPARC water vapour assessment II: Data set overview, in preparation, 2021.