# The SPARC water vapour assessment II: Biases and drifts of water vapour satellite data records with respect to frost point hygrometer records

Michael Kiefer[1], Dale F. Hurst[2,3], Gabriele P. Stiller[1], Stefan Lossow[1], Holger Vömel[4], John Anderson[5], Faiza Azam[6,7], Jean-Loup Bertaux[8], Laurent Blanot[9], Klaus Bramstedt[6], John P. Burrows[6], Robert Damadeo[10], Bianca Maria Dinelli[11], Patrick Eriksson[12], Maya García-Comas[13], John C. Gille[14,15], Mark Hervig[16], Yasuko Kasai[17], Farahnaz Khosrawi[1,18], Donal Murtagh[12], Gerald E. Nedoluha[19], Stefan Noël[6], Piera Raspollini[20], William G. Read[21], Karen H. Rosenlof[22], Alexei Rozanov[6], Christopher E. Sioris[23], Takafumi Sugita[24], Thomas von Clarmann[1], Kaley A. Walker[25], and Katja Weigel[6,26]

[1]Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research, Karlsruhe, Germany
[2]Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA
[3]Global Monitoring Laboratory, NOAA Earth System Research Laboratories, Boulder, Colorado, USA
[4]Earth Observing Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA
[5]Hampton University, Hampton, Virginia, USA
[6]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany
[7]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institute of Networked Energy Systems, Oldenburg, Germany
[8]LATMOS, Sorbonne Université, Paris, France
[9]ACRI-ST, 11 Boulevard d'Alembert, 78280 Guyancourt, France
[10]NASA Langley Research Center, Hampton, VA, USA
[11]Instituto di Scienze dell'Atmosfera e del Clima del Consiglio Nazionale delle Ricerche (ISAC-CNR), Bologna, Italy
[12]Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden
[13]Instituto de Astrofísica de Andalucía, CSIC, Granada, Spain
[14]National Center for Atmospheric Research, Atmospheric Chemistry Observations & Modeling Laboratory, P.O. Box 3000, Boulder, USA
[15]Atmospheric and Oceanic Sciences, University of Colorado, Boulder, USA
[16]GATS Inc., Driggs, Idaho, USA
[17]National Institute of Information and Communications Technology (NICT), Terahertz Technology Research Center, Tokyo, Japan
[18]Institute for Advanced Simulations, Jülich Supercomputing Centre, Jülich, Germany
[19]Remote Sensing Division, Naval Research Laboratory, Washington, DC, USA
[20]Instituto di Fisica Applicata del Consiglio Nazionale delle Ricerche (IFAC-CNR), Sesto Fiorentino, Italy
[21]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA
[22]Chemical Sciences Laboratory, NOAA Earth System Research Laboratories, Boulder, Colorado, USA
[23]Center for Research in Earth and Space Science, York University, Toronto, Canada
[24]National Institute for Environmental Studies, Tsukuba (Japan)
[25]Department of Physics, University of Toronto, Toronto, Canada
[26]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany

**Correspondence:** Michael Kiefer (michael.kiefer@kit.edu)

**Abstract.**

Satellite data records of stratospheric water vapour have been compared to balloon-borne frost point hygrometer (FP) profiles that are coincident in space and time. The satellite data records of 15 different instruments cover water vapour data available from January 2000 through December 2016. The hygrometer data are from 27 stations all over the world in the same period.

5   For the comparison, real or constructed averaging kernels have been applied to the hygrometer profiles to adjust them to the measurement characteristics of the satellite instruments. For bias evaluation, we have compared satellite profiles averaged over the available temporal coverage to the means of coincident FP profiles for individual stations. For drift determinations, we analyzed timeseries of relative differences between spatiotemporally coincident satellite and hygrometer profiles at individual stations. In a synopsis we have also calculated the mean biases and drifts (and their respective uncertainties) for each satellite

10  record over all applicable hygrometer stations in three altitude ranges ($10 - 30\,\mathrm{hPa}$, $30 - 100\,\mathrm{hPa}$, and $100\,\mathrm{hPa}$ to tropopause). Most of the satellite data have biases $< 10\%$ and average drifts $< 1\,\%\,\mathrm{yr}^{-1}$ in at least one of the respective altitude ranges. Virtually all biases are significant in the sense that their uncertainty range in terms of twice the standard error of the mean does not include zero. Statistically significant drifts (95% confidence) are detected for 35% of the $\approx 1200$ timeseries of relative differences between satellites and hygrometers.

## 1   Introduction

Water vapour is the most potent greenhouse gas in the atmosphere (Kiehl and Trenberth, 1997). Its radiative effect per unit mass change is strongest around the tropical tropopause (Riese et al., 2012; Solomon et al., 2010). Trends of stratospheric water vapour are expected to be related to the temperatures of the tropical tropopause where air transporting water vapour enters the stratosphere (e.g. Fueglistaler and Haynes, 2005; Randel and Park, 2019). Rising troposphere and tropopause temperatures due

20  to global warming may lead to increasing stratospheric water vapour abundances, initiating a positive feedback-loop where global warming will be further accelerated due to increasing water vapour abundances in the lower stratosphere (e.g. Gettelman et al., 2010; Dessler et al., 2013, 2016). In addition, the major stratospheric source of water vapour is the oxidation of methane (e.g. le Texier et al., 1988) which has more than doubled since 1800 (Blunier et al., 1993) and is expected to continue rising in future (e.g. Lelieveld et al., 1998), further increasing stratospheric water vapour.

25  Since 1980, despite constant or slightly decreasing tropical tropopause temperatures (Gettelman et al., 2009; Ding-Zhu et al., 2015), an increase in stratospheric water vapour was observed over Boulder, Colorado (Oltmans and Hofmann, 1995). This cannot be explained by the high positive correlation between tropical tropopause temperatures and water vapour in the lowermost tropical stratosphere (Fueglistaler and Haynes, 2005; Randel and Park, 2019). In consequence, numerous studies have been performed to better understand the stratospheric water vapour budget and trends (e.g. Oltmans and Hofmann, 1995;

30  Oltmans et al., 2000; Rosenlof et al., 2001; Nedoluha et al., 2003; Hurst et al., 2011b; Dessler et al., 2014; Hegglin et al., 2014; Brinkop et al., 2016). Vertically resolved profiles of atmospheric water vapour have been observed around the globe by satellite-based instruments in low Earth orbits since the mid 1970s. From the year 2000 on, 15 different satellite instruments have observed vertically resolved water vapour distributions from the middle troposphere to the mesosphere and above. More

than two decades ago, a first assessment of the quality of water vapour observations including ground-based, balloon-borne and

35  satellite instrumentation was published as the WCRP/SPARC (World Climate Research Programme/Stratosphere-troposphere Processes And their Role in Climate) report no. 2 (Kley et al., 2000). The many new satellite instruments in orbit since 2000 have made it of great interest to reassess the quality and consistency of water vapour observations from space. Here we concentrate only on stratospheric measurements by satellites and balloon-borne FPs.

Many of the satellite data records included in this study are described in detail by their data providers in reports and scientific

40  papers (a compilation of information relevant to this paper is presented in Walker et al., The SPARC Water Vapour Assessment II: Data characterisation, to be submitted to AMT, 2023). These reports and scientific papers also contain, in most cases, some information about validation activities. Frost point hygrometers have often been used for satellite data validation since they are considered to be most accurate and internally consistent water vapour instruments for stratospheric measurements. Comparisons of different instruments, including their calibrations and data processing routines, were the focus of several field

45  campaigns (Vömel et al., 2007a,b; Hurst et al., 2011a; Rollins et al., 2014; Vömel et al., 2016; Hall et al., 2016). Despite differences between the measurements by instruments employing different sensing techniques, consistency was found within the data from FPs.

Each comparison of satellite data records to the FP soundings, however, has been done in a slightly different way by each validation team, resulting in a wealth of validation publications that are not consistent down to the last detail. This lack of

50  consistency hampers activities where several satellite data records need to be merged to construct a long-term time series, e.g., for trend assessments. For this reason, we decided for this WCRP/SPARC WAVAS-II (Water Vapor Assessment II) activity to perform the comparison of all available satellite data records obtained during the period of 2000 through 2016 to FP data in a fully consistent and reproducible way. We document here where the FP data came from, how we made them comparable to the satellite data, and how the comparisons were performed. Overall, all of our satellite-to-FP comparisons are done in a similar

55  way. The result of this activity is the first fully self-consistent quality assessment of vertically resolved biases and drifts in the stratospheric water vapour measurements by numerous satellite instruments and FPs, along with the respective uncertainties. In order to be consistent with the other assessments within the WCRP/SPARC WAVAS-II activity (see ACP/AMT/ESSD special issue "Water vapour in the upper troposphere and middle atmosphere: a WCRP/SPARC satellite data quality assessment including biases, variability, and drifts", https://amt.copernicus.org/articles/special_issue10_830.html), we use the same data

60  versions as used in other papers in the WAVAS-II special issue, even in cases where newer data versions have become available in the meantime.

The paper is structured as follows: In Section 2 we describe the FP data and the satellite data records, including their preparation for use within this study. Further we explain how we made the FP data comparable in terms of their vertical resolution, and how the biases and the drifts have been calculated. Section 3 presents the assessment of the biases between the

65  satellite and FP data records, starting with each individual satellite data records versus the FP data at each site, then discussing comparisons of all satellite data versus one station, as well as one satellite data record versus all stations. We summarise these findings with a synopsis of the biases and their uncertainties for each satellite data set over all its associated FP sites, in three different altitude ranges. Section 4 presents the assessment of instrumental drifts of the satellite data records against FP records,

also including a synopsis of the drifts of each satellite data set, in three altitude ranges, over all its associated FP sites. Section 5
70  summarises our findings and offers recommendations for the use of the satellite data records under assessment. The individual
bias and drift figures for pairs of satellite records and FP stations are presented in the Supplement and Appendix to this paper,
respectively.

## 2    Data and data handling

In this study, we compare the satellite data records under assessment in the WCRP/SPARC WAVAS-II activity to reference-
75  quality FP soundings at 27 stations (79°N to 45°S latitude) during 2000 through 2016. 31 data records from 15 different
satellite instruments provide a subset of measurements coincident with the FP soundings (for coincidence criteria see below)
that can be evaluated against the profile data from FP balloon soundings. In the following, we briefly describe FP and satellite
data, explain the adjustments of the vertical resolution of the FP data to each of the various satellite data records, and describe
the methods for the bias and drift assessments.

80  ### 2.1    Frost point hygrometer data

The chilled mirror technique (Brewer, 1949; Barrett et al., 1950) is based upon the well-known equilibrium thermodynamic
relationship (Clausius-Clapeyron) between an ice or liquid water surface and overlying water vapour. Frost point hygrometers
actively maintain the equilibrium of this two phase system by continuously adjusting the temperature of the condensate layer
such that it remains stable. Both the NOAA (National Oceanic and Atmospheric Administration) Global Monitoring Labora-
85  tory's Frost Point Hygrometer (NOAA FPH) and the Cryogenic Frost-Point Hygrometer (CFH) use optical detection of the
condensate layer on a small mirror. A feedback loop actively regulates the mirror temperature to maintain a stable condensate
layer, making the water vapour content of the overlying air directly calculable from the mirror temperature.

The balloon-borne NOAA FPH was first flown over Boulder, CO, in 1980 (Oltmans et al., 2000) and, to date, has produced
a 43-yr record of stratospheric water vapour mixing ratios (Hurst et al., 2011b). It has also been flown routinely at Lauder,
90  New Zealand since 2004, Hilo, Hawaii since 2010, and has been part of a number of tropical, mid-latitude and polar measure-
ment campaigns (Kley et al., 1997). The NOAA FPH makes measurements not only during ascent, but also during controlled
(5 m/s) descent of the balloon when water vapour contamination is improbable. The FPH measurement uncertainty is largely
determined by the stability of the frost layer, and under optimal performance, is 0.1–0.3 K in frost-point temperature in the
stratosphere, leading to a measurement uncertainty of $< 6\,\%$ for stratospheric mixing ratios (Hall et al., 2016).
95  The CFH (Vömel et al., 2007a,b; Vömel et al., 2016) works along the same principle as the NOAA FPH, but uses a feedback
controller with a continuously variable PID parameter schedule to make observations between the surface and the middle
stratosphere (25 km). The uncertainty of the condensate phase in the temperature range below 0°C is largely eliminated,
allowing continuous profiles over a wider range of frost-point temperatures to be measured. It suffers no artefacts in cirrus
clouds and may only be limited in wet precipitating clouds with the detector lens getting wet. The measurement uncertainty of
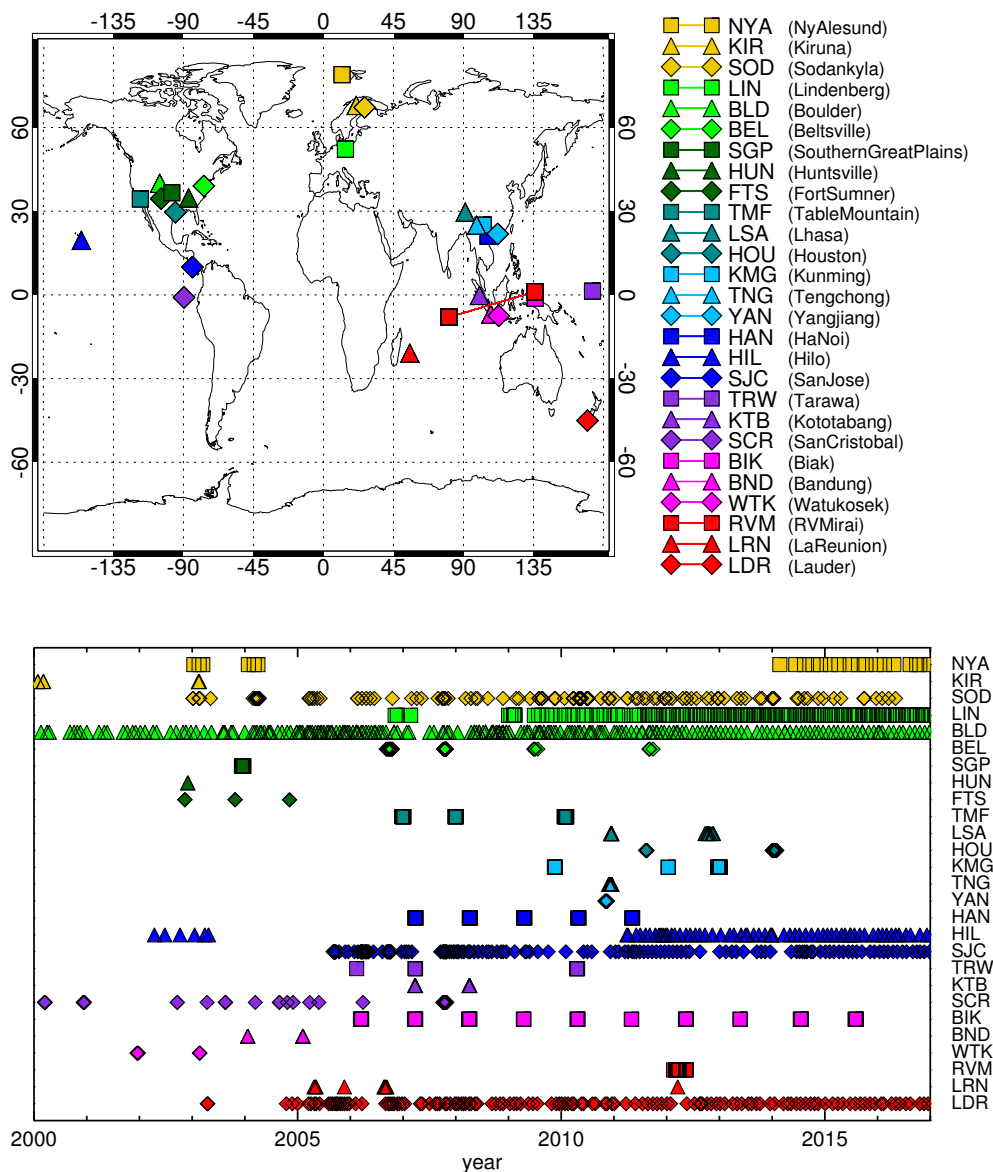
**Figure 1.** Locations of NOAA FPH and CFH stations that provided measurement data for these intercomparisons (top), and the temporal coverage of the data records at the respective stations (bottom). In the lower plot each symbol represents at least one balloon-borne FP sounding.

**Table 1.** Overview of NOAA (National Oceanic and Atmospheric Administration) frost point hygrometer (NOAA FPH) and cryogenic frost point hygrometer (CFH) stations used for comparisons with satellite data.

| # | Code | Site | Meas. period | Instrument type | Lat. / deg | Lon. / deg | Remark |
|---|------|------|--------------|-----------------|-----------|------------|--------|
| 1 | BND | Bandung | 2003 – 2004 | CFH | -6.9 | 107.6 | |
| 2 | BEL | Beltsville | 2006 – 2011 | CFH | 39.0 | -76.9 | |
| 3 | BIK | Biak[a] | 2006 – 2015 | CFH | -1.2 | 136.1 | |
| 4 | BLD | Boulder[a] | 1980 – present | CFH/NOAA FPH | 40.0 | -105.2 | |
| 5 | FTS | Fort Sumner | 1996 – 2004 | NOAA FPH | 34.5 | -104.3 | |
| 6 | HAN | Hanoi | 2007 – 2011 | CFH | 21.0 | 105.8 | |
| 7 | HIL | Hilo[a] | 2002 – present | CFH/NOAA FPH | 19.7 | -155.1 | |
| 8 | HOU | Houston | 2011, 2013 | CFH/NOAA FPH | 29.6 | -95.2 | |
| 9 | HUN | Huntsville | 2002 | NOAA FPH | 34.7 | -86.7 | |
| 10 | KIR | Kiruna | 1991 – 2003 | NOAA FPH | 67.8 | 20.2 | |
| 11 | KTB | Kototabang | 2007 – 2008 | CFH | -0.2 | 100.3 | |
| 12 | KMG | Kunming | 2009 – 2012 | CFH/NOAA FPH | 25.0 | 102.7 | |
| 13 | LRN | La Reunion | 2005 – 2011 | CFH | -20.9 | 55.5 | |
| 14 | LDR | Lauder[a] | 2003 – present | NOAA FPH | -45.0 | 169.7 | |
| 15 | LSA | Lhasa | 2010, 2013 | CFH | 29.7 | 91.1 | |
| 16 | LIN | Lindenberg[a] | 2006 – present | CFH | 52.2 | 14.1 | |
| 17 | NYA | Ny Alesund | 2002 – 2004, 2013 – present | CFH/NOAA FPH | 78.9 | 11.9 | |
| 18 | RVM | Research Vessel Mirai | 2011 | CFH | -8.0/1.2 | 80.5/136.1 | ship cruise |
| 19 | SCR | San Cristobal | 1998 – 2007 | CFH/NOAA FPH | -0.9 | -89.6 | |
| 20 | SJC | San Jose[a] | 2005 – present | CFH | 9.9 | -84.1 | incl. Alajuela, Heredia, San Pedro, and San Jose |
| 21 | SOD | Sodankyla[a] | 1995 – present | CFH/NOAA FPH | 67.4 | 26.6 | |
| 22 | SGP | Southern Great Plains | 2003 | CFH | 36.6 | -97.5 | |
| 23 | TMF | Table Mountain | 2006 – 2009, 2013 | CFH/NOAA FPH | 34.4 | -117.7 | |
| 24 | TRW | Tarawa | 2005 – 2010 | CFH | 1.4 | 172.9 | |
| 25 | TNG | Tengchong | 2010 | CFH | 25.0 | 98.5 | |
| 26 | WTK | Watukosek | 2001 – 2003 | NOAA FPH | -7.6 | 112.7 | |
| 27 | YAN | Yangjiang | 2010 | CFH | 21.9 | 112.0 | |

*a*) data from these sites were used for the drift analyses

100    the CFH is less than 0.5 K throughout the entire profile, which translates to conservative values of 4 % in the lower troposphere and 9 % in the stratosphere.

Neither the CFH nor NOAA FPH require water vapour calibration standards or a water vapour calibration scale; only the mirror thermistor must be calibrated with high accuracy and this is accomplished using traceable standards of the US National Institute of Standards and Technology (NIST).

105 Temperature and pressure measurements used to convert frost point hygrometer data into relative humidity values and volume mixing ratios, respectively, are from the accompanying radiosondes on each balloon. Measurements of temperature and pressure have been provided by different radiosonde models throughout the years: Vaisala models RS80, RS92, and RS41, InterMet models iMet-1-RSB and iMet-4-RSB, and Meisei models RS-06G and RS-11G.

Offsets in the pressure measurements of radiosondes may bias the calculation of the mixing ratio in the stratosphere (Stauffer
110 et al., 2014; Inai et al., 2015). To minimize this bias, the radiosonde pressure measurements are usually corrected using the radiosonde's acquisition of the geometric altitude by Global Navigation Satellite System (GNSS). In some radiosonde systems, the pressure is not measured directly but instead derived from the GNSS altitude. Only in older systems that precede the availability of GNSS observations on radiosondes starting in the late 1990s are pressures used without any corrections except those based on a simple pre-flight comparison at the surface with ground-based sensors. For this work we used FP mixing ratio
115 averages on a fixed 250 m altitude grid. These are typically further reduced in vertical resolution as they are convolved with real or constructed averaging kernels for the different satellite instruments (see Section 2.3).

Table 1 lists the stations from which NOAA FPH or CFH data have been used for comparison with satellite data, together with their period of operation, the type of instrument launched, and the geographical coordinates of the site. Each station is given a 3-letter code to simplify its identification in the remainder of this paper. Fig. 1 provides an overview of the geographical
120 locations and the measurement periods of the stations, together with the symbols and color codes that are used throughout this paper to mark the respective data of the stations.

In the remainder of this paper we do not distinguish between NOAA FPH and CFH so we continue to use the generic term "FP" for frostpoint hygrometer instruments and data.

## 2.2 Satellite data

125 Satellite data from all instruments providing measurements coincident with FP balloon soundings have been selected. Data quality filter criteria according to the original data descriptions from the data providers have been applied (for a summary of these data set specific criteria see Walker et al., The SPARC Water Vapour Assessment II: Data characterisation, to be submitted to AMT, 2023). No further bulk screening for data outliers surviving the previous data quality filtering has been applied. The 31 satellite data records that are used in this comparison are listed in Table 2 along with their 3-letter codes. Fig. 2 shows the
130 symbols and color codes for the satellite data sets used throughout this paper. The data versions we have assessed in this study are not the most recent ones to date for most of the satellite data sets. For reasons of consistency, the data versions used here are the same as those assessed by the other comparative studies of the SPARC WAVAS-II activity. It is left to future studies to evaluate if more recent data versions of water vapour satellite data are improved with respect to those assessed here. Such evaluations can also be done individually by comparing newer data versions to those assessed here to quantify any changes in
135 the biases and drifts reported here.
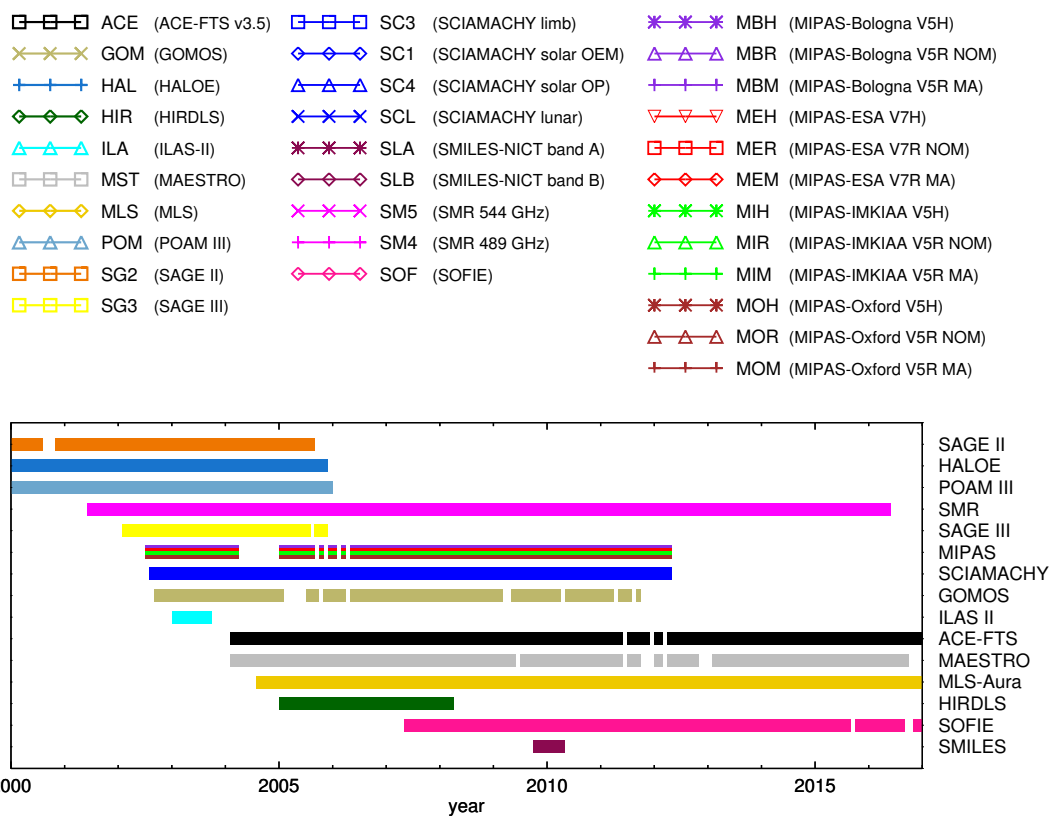
**Figure 2.** Colors, symbols, and three-letter code for the satellite data records used throughout the paper (upper part) and temporal distribution of available data of the respective satellite instruments on a monthly basis until the end of 2016 (lower part, not divided into measurement modes or data versions).

Two different sets of coincidence criteria were used in this paper: one for satellites providing data at high spatial and temporal densities, and one for lower density datasets. The criteria for dense samplers (HIRDLS, MIPAS, MLS/Aura, SCIAMACHY limb observations, SMILES, and SMR) were: time difference $\Delta t \leq 24\,h$, distance $\Delta r \leq 1000\,km$, latitudinal difference $\Delta$ lat $\leq 5°$. For less dense samplers (ACE-FTS, GOMOS, HALOE, ILAS-II, MAESTRO, POAM-III, SAGE-II, SAGE-III,

140  SCIAMACHY occultation observations, and SOFIE) we relaxed the coincidence criteria to: $\Delta t \leq 7x24\,h$, $\Delta r \leq 2000\,km$ and latitudinal difference $\Delta$ lat $\leq 15°$ to achieve enough coincidences for meaningful statistical evaluations of biases and drifts. For the troposphere, however, where high water vapour variability in smaller spatial and temporal scales is present, these criteria are too coarse. We have therefore restricted these analyses to water vapour measurements at altitudes above the local lapse rate tropopause determined from the radiosonde temperature profiles obtained simultaneously with the FP profiles using the WMO

145  criterion (World Meteorological Organization, 1957). A companion paper (Read et al., 2022) comparing satellite data to FP and radiosonde measurements in the upper troposphere uses far stricter coincidence criteria.

**Table 2.** Overview of the water vapour data sets from satellites used in this study. Column "Retr. type" indicates whether the retrieval result was number density $n_{H_2O}$ (marked ND) instead of vmr, and whether the retrieval was done in the $\log(\text{vmr})$ or $\log(n_{H_2O})$ domain. Column "Kernel type" holds the information whether a proper averaging kernel matrix (AK) or ad hoc smoothing kernels (SK) were used. The numbers in the last column indicate the FP stations that provided the data used for the drift analysis of the satellite data (compare to Table 1).

| Code | Instrument | Data set version | Label | Retr. type | Kernel type | FP # for drift analyses |
|------|------------|------------------|-------|------------|-------------|-------------------------|
| ACE | ACE-FTS | 3.5 | ACE-FTS v3.5 | | SK | 4,7,16,20,21 |
| GOM | GOMOS | LATMOS v6 | GOMOS | | SK | 4,14,21 |
| HAL | HALOE | v19 | HALOE | | SK | 4 |
| HIR | HIRDLS | v7 | HIRDLS | | SK | |
| ILA | ILAS-II | v3/3.01 | ILAS-II | | SK | |
| MST | MAESTRO | v31 | MAESTRO | | SK | 4,14,16,21 |
| MBH | MIPAS | Bologna V5H v2.3 NOM | MIPAS-Bologna V5H | | AK | |
| MBR | | Bologna V5R v2.3 NOM | MIPAS-Bologna V5R NOM | | AK | 3,4,14,20,21 |
| MBM | | Bologna V5R v2.3 MA | MIPAS-Bologna V5R MA | | AK | 4 |
| MEH | | ESA V7H v7 NOM | MIPAS-ESA V7H | | AK | |
| MER | | ESA V7R v7 NOM | MIPAS-ESA V7R NOM | | AK | 3,4,14,20,21 |
| MEM | | ESA V7R v7 MA | MIPAS-ESA V7R MA | | AK | 21 |
| MIH | | IMK/IAA V5H v20 NOM | MIPAS-IMKIAA V5H | log | AK | |
| MIR | | IMK/IAA V5R v220/1 NOM | MIPAS-IMKIAA V5R NOM | log | AK | 3,4,14,20,21 |
| MIM | | IMK/IAA V5R v522 MA | MIPAS-IMKIAA V5R MA | log | AK | 4 |
| MOH | | Oxford V5H v1.30 NOM | MIPAS-Oxford V5H | log | SK | |
| MOR | | Oxford V5R v1.30 NOM | MIPAS-Oxford V5R NOM | log | AK | 3,4,14,20,21 |
| MOM | | Oxford V5R v1.30 MA | MIPAS-Oxford V5R MA | log | SK | 21 |
| MLS | MLS | v4.2 | MLS | log | AK | 3,4,7,14,16,20,21 |
| POM | POAM III | v4 | POAM III | | SK | |
| SG2 | SAGE II | v7.00 | SAGE II | | SK | 4 |
| SG3 | SAGE III | Solar occ. v4 | SAGE III | | SK | |
| SC3 | SCIAMACHY | Limb v3.01 | SCIAMACHY limb | ND/log | AK | 4,16,21 |
| SCL | | Lunar occultation v1.0 | SCIAMACHY lunar | ND/log | SK | |
| SC1 | | Solar occ. - OEM v1.0 | SCIAMACHY solar OEM | ND/log | AK | 4,16,21 |
| SC4 | | Solar occ. - OP v4.2.1 | SCIAMACHY solar OP | ND | SK | 4,16,21 |
| SLA | SMILES | NICT v2.9.2 band A | SMILES-NICT band A | | SK | |
| SLB | | NICT v2.9.2 band B | SMILES-NICT band B | | SK | |
| SM5 | SMR | v2.0 544 GHz | SMR 544 GHz | log | AK | 3,4,14,16,20,21 |
| SM4 | | v2.1 489 GHz | SMR 489 GHz | | AK | 4,21 |
| SOF | SOFIE | v1.3 | SOFIE | | SK | 16,21 |

In our assessment of satellite measurements based on the occultation technique, we have not distinguished between sunset and sunrise measurements because the comparisons with FP profiles showed that there were only insignificant differences between sunrise and sunset measurements that could unequivocally be assigned to the respective satellite measurement mode.

150    In case of multiple coincidences of a given satellite water vapour profile with profiles of one FP data set we retain only the coincident profile pair with the lowest value of the sum of squares of spatial/temporal distances, normalized by the respective maximum allowed spatial/temporal distances from the appropriate coincidence criterion. Though this matching method slightly reduces the number of satellite profiles used for the bias assessment, there are usually enough coincidences during the 2000–2016 time period to work with. Therefore we have decided to minimize the contribution of natural variability using this method,
155    i.e. considering the closest coincidences in space and time only.

We shall use the comprehensive term "SAT" for generic statements about the satellite data.

### 2.3    Adaptation of the vertical resolution of FP profiles to the satellite data and interpolation to a common grid

The vertical grids of all satellite data sets are coarser than the vertical grids of the FP profiles. More importantly, the vertical resolution of the satellite data is never as fine as the $250\,\mathrm{m}$ averages calculated from the $5$–$10\,\mathrm{m}$ native resolution of FP
160    measurements. Therefore, prior to comparison, the vertical resolution of the FP data was necessarily adjusted to that of the satellite instrument. This was ideally done by application of the averaging kernel matrix (AK) and a priori profile of the latter for each satellite data set. However, in many cases, averaging kernels are not provided with the satellite data, so ad hoc averaging kernels were constructed. These constructed kernels were Gaussian-shaped smoothing kernels (SK) with the local vertical resolution of the satellite profile as full width at half maximum. The kernel type column of Table 2 shows whether
165    AK or SK were applied to FP profiles for each satellite dataset. The modified FP profiles, and also the satellite profiles, were then interpolated on a common vertical grid, essentially defined by the respective satellite measurement grid. Technically, the pressure grid of all involved quantities (FP profiles and SAT profiles and kernels) was used to construct an altitude grid, which essentially represents $\log P$. This pseudo-altitude grid was used as a basis for all operations. The inverse transformation, i.e. from pseudo-altitude back to pressure was then used before the plotting and comparing of data.
170    For these steps, we have followed widely the method described in Stiller et al. (2012) as is briefly summarised here. As a first step, the FP profile on the finer grid is resampled on the coarser grid of the coincident satellite profile. Resampling of a coarse profile $\boldsymbol{x}_{\mathrm{c}}$ on a fine grid can be written as

$$\boldsymbol{x}_{\mathrm{cf}} = \mathbf{W}\boldsymbol{x}_{\mathrm{c}}, \tag{1}$$

where $\mathbf{W}$ is an interpolation matrix. However, the mapping of a high-resolved profile $\boldsymbol{x}_{\mathrm{f}}$ on a less dense grid is not a unique
175    operation, but a reasonable method to achieve this is (Rodgers, 2000, Chapter 10.3.1)

$$\boldsymbol{x}_{\mathrm{fc}} = \mathbf{V}\boldsymbol{x}_{\mathrm{f}}, \tag{2}$$

where

$$\mathbf{V} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T, \tag{3}$$

which satisfies $\mathbf{VW} = \mathbf{I}$, $\mathbf{I} = $ unity, and $\mathbf{W}$ being an interpolation matrix. The application of the averaging kernel $\mathbf{A}_{\mathrm{c}}$ of the
180    low-resolved profile $\boldsymbol{x}_{\mathrm{c}}$ to the better-resolved profile $\boldsymbol{x}_{f}$ under consideration of the a priori profile $\boldsymbol{x}_{\mathrm{a}}$ of the low-resolved

retrieval is then performed on the coarse grid

$$\tilde{\boldsymbol{x}}_{\mathrm{fc}} = \mathbf{A}_{\mathrm{c}}\mathbf{V}\boldsymbol{x}_{\mathrm{f}} + (\mathbf{I} - \mathbf{A}_{\mathrm{c}})\boldsymbol{x}_{\mathrm{a}} \qquad (4)$$

For some satellite data records there is another complication: instead of mixing ratios the logarithms of water vapour mixing ratios are retrieved (see column **Ret.type** of Table 2). The averaging kernels hence refer to the logarithms of the water vapour

185 mixing ratios. The application of the averaging kernels of these specific measurements to the better resolved profile of the FP data on the basis of the coarse-grid averaging kernel $\mathbf{A}_{\mathrm{lnc}}$ of the logarithm of the water vapour mixing ratio then is

$$\tilde{\boldsymbol{x}}_{\mathrm{fc}} = \exp\left(\mathbf{A}_{\mathrm{lnc}}\mathbf{V}\ln(\boldsymbol{x}_{\mathrm{f}}) + (\mathbf{I} - \mathbf{A}_{\mathrm{lnc}})\ln(\boldsymbol{x}_{\mathrm{a}})\right). \qquad (5)$$

For the cases that use an ad hoc smoothing kernel generated from information on the vertical resolution, there is also no a priori information available. Hence the equations 4 and 5 become

190 $$\tilde{\boldsymbol{x}}_{\mathrm{fc}} = \mathbf{B}_{\mathrm{c}}\mathbf{V}\boldsymbol{x}_{\mathrm{f}} \quad \text{and} \quad \tilde{\boldsymbol{x}}_{\mathrm{fc}} = \exp\left(\mathbf{B}_{\mathrm{lnc}}\mathbf{V}\ln(\boldsymbol{x}_{\mathrm{f}})\right). \qquad (6)$$

with $\mathbf{B}_{\mathrm{c}}$ and $\mathbf{B}_{\mathrm{lnc}}$ being the smoothing kernel on the coarse grid for the linear and logarithmic retrievals, respectively.

A common technical problem in convolving measured profiles with kernels of retrieved data is that the altitude ranges do not fit. Hence we extended the FP profile above and below its upper and lower boundaries by offset-corrected, climatological water vapour data from HAMMONIA (Hamburg Model of the Neutral and Ionized Atmosphere, Schmidt et al. (2006)) as a function

195 of month and latitude. After the convolution step, the smoothed FP profile was cut to its original upper boundary. Since there is a possibly strong influence of the climatological HAMMONIA profile at the lower boundary, due to the rapidly increasing vmr-values below the hygropause, the FP profile was cut at one (local) vertical resolution distance above the original lower boundary to minimize the mapping of climatology information into the altitude range used for comparison.

Fig. 3 demonstrates the effect of the transformation on the comparison between the MIR satellite data and FP profiles at

200 the equatorial station BIK. Due to the finer 250 m vertical resolution, there is a sharper and deeper minimum in the FP profile near 90 hPa than in the satellite profile. The necessity to use the averaging kernel formalism described above is exemplarily demonstrated in the following: Comparison of the two profiles, with the FP data simply interpolated to the coarser grid of the satellite instrument, but not smoothed (black diamonds), is misleading since the MIPAS instrument is unable to resolve the sharp feature in the profile. By application of the averaging kernel and a priori profile to the FP profile according to Eq. 5

205 (in the MIR data the logarithm of water vapour mixing ratio is retrieved, and the apriori is a profile of constant nonzero value) the FP profile is transformed into the profile the satellite instrument would measure if the hygrometer profile was the truth (black squares). Convolving the FP profiles like this is the only way the two profiles can be compared in a meaningful manner. If averaging kernels and a priori profiles are not provided along with the satellite data record, the vertical resolution of the hygrometer profile can at least be adjusted using the constructed Gaussian-shaped averaging kernels. The effect of

210 this smoothing is demonstrated by the profiles with black triangles in Fig. 3, and is notably different from the application of the MIPAS-specific averaging kernel and a priori information (black squares), which is particularly obvious for the averaged profiles (lower row of Fig. 3) and the corresponding differences. Clearly the application of the correct vertical averaging
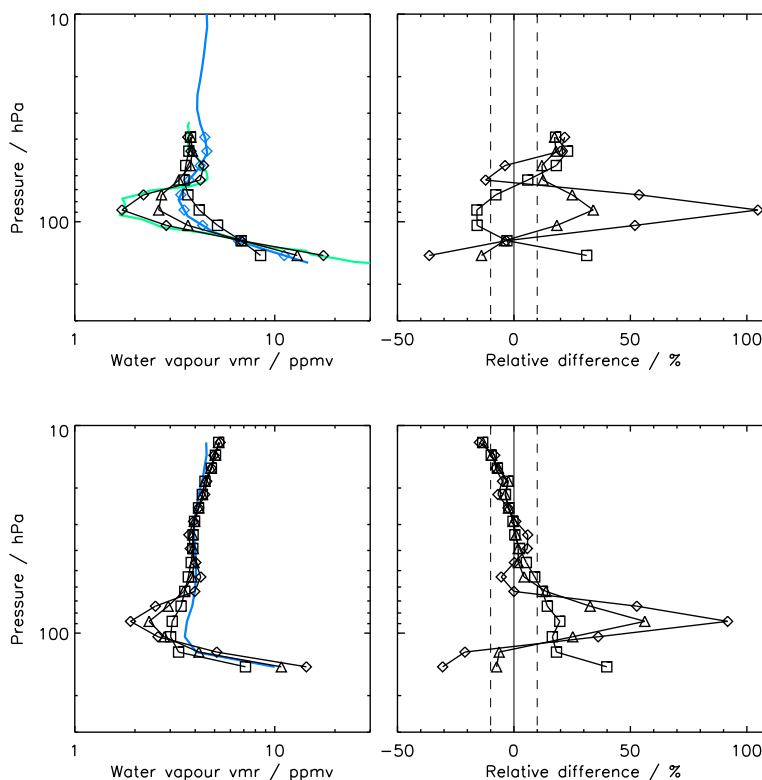
**Figure 3.** Impact of the different methods for adjustment of the vertical grid and resolution of FP profiles (here: BIK) to those of the satellite data records (here: MIR). Top left: Sample single profiles of satellite (blue) and co-located FP data (green). Black diamonds: FP profile directly interpolated onto the coarse common grid, black triangles: FP profile smoothed with a Gaussian kernel (details see text), black squares: proper averaging kernels and a priori information applied to FP profile. Top right: Corresponding relative differences for the three variants in terms of satellite profile minus FP profile. Bottom left: Averaged profiles for coincident MIR (blue) and FP data at BIK (black, symbols as for data shown in top row plot frames). Bottom right: Corresponding relative differences for the three variants in terms of satellite profile minus FP profile, divided by FP profile.

kernels and a priori profiles adds further information on the altitude displacement and the content of a priori information in the retrieved profiles to the FP profiles. In contrast, the application of ad hoc smoothing kernels alone has a much weaker effect, nevertheless it reduces the large resolution-based differences between satellite and FP measurements to a considerable degree. For this reason, we have employed the constructed smoothing kernels in all comparisons where no kernel/a priori information for the satellite profiles was available.

## 3 Bias assessment from vertical profile comparisons

### 3.1 Method of calculation of bias and standard error of the mean bias

220 We assess the bias between satellite data and the FP measurements as the mean difference between the satellite profiles and the coincident transformed FP profiles. For profile data of a given satellite instrument there is a set of $J_s$ FP locations/stations with coincident profiles. At station $j$ the bias for each grid point $i$ is:

$$b_{j,i} = \frac{\sum_{n=1}^{N_{j,i}}(x_{c;j,n,i} - \tilde{x}_{fc;j,n,i})}{N_{j,i}} = \frac{\sum_{n=1}^{N_{j,i}} x_{c;j,n,i}}{N_{j,i}} - \frac{\sum_{n=1}^{N_{j,i}} \tilde{x}_{fc;j,n,i}}{N_{j,i}},$$  (7)

i.e. it does not matter whether the individual differences are calculated first and then are averaged, or whether the difference

225 of the appropriate averages are calculated. The bias $b_{j,i}$ is calculated independently for each grid point $i$ and FP station $j \in \{1 \ldots J_s\}$ from the available $N_{j,i}$ coincident observations. For given $j$, $N_{j,i}$ can be different for different altitudes because the altitude coverage of a measurement system under assessment may vary from profile measurement to profile measurement. The standard error of the mean (SEM) of the bias, which is also the bias-corrected root mean squares (rms) difference of the profiles, is calculated as:

230 $$\sigma_{bias;j,i} = \sqrt{\frac{\sum_{n=1}^{N_{j,i}}(x_{c;j,n,i} - \tilde{x}_{fc;j,n,i} - b_{j,i})^2}{N_{j,i}(N_{j,i} - 1)}}.$$  (8)

We consider the bias $b_{j,i}$ as statistically significant if the interval $b_{j,i} \pm 2\sigma_{bias;j,i}$ does not include zero.

The mean relative bias (in percent) or percentage bias is calculated by dividing the mean bias $b_{j,i}$ by the mean of the involved FP measurements and multiplying by 100. In all of the following figures, the differences provided are satellite profiles minus FP data. The latter was adapted to the vertical resolution of the satellite data according to Sec. 2.3 and Table 2 and, as well as

235 the satellite data, brought to a common coarser vertical grid.

For a given satellite data set the mean bias over all stations and for a specific altitude range (e.g. 10–30 hPa, 30–100 hPa, and 100 hPa–tropopause as presented in Section 3.5) is calculated as follows:

$$b = \frac{\sum_{j=1}^{J_s} \sum_{i \in \mathbf{I}_j} w_{j,i} b_{j,i}}{\sum_{j=1}^{J_s} \sum_{i \in \mathbf{I}_j} w_{j,i}}.$$  (9)

Here $\mathbf{I}_j$ represents the set of indices for all the altitudes from the given altitude range and FP comparison data set. The weights

240 $w_{j,i}$ are calculated from the SEM of the bias $\sigma_{bias;j,i}$ and from the ratio of the width $\Delta z_{j,i}$ of the common coarse grid to the vertical resolution $r_{v;j,i}$ of the satellite measurement, according to

$$w_{j,i} = \frac{1}{\sigma_{bias;j,i}^2} \frac{\Delta z_{j,i}}{r_{v;j,i}}.$$  (10)

The factor $\frac{\Delta z_{j,i}}{r_{v;j,i}}$ in the weight is used to compensate for the discrepancy between actual vertical resolution and grid width. Without this factor, the SEM of the mean bias would directly depend on the grid width via the number of data points which are

245 available in a given altitude range.

Finally the SEM of this mean bias over an altitude range is given by

$$\sigma_b = \sqrt{\frac{\sum_{j=1}^{J_s} \sum_{i\in\mathbf{I}_j} w_{j,i}^2 \sigma_{\mathrm{bias};j,i}^2}{(\sum_{j=1}^{J_s} \sum_{i\in\mathbf{I}_j} w_{j,i})^2}}. \tag{11}$$

Again, the mean relative bias (in percent) or percentage bias is calculated by dividing the mean bias $b$ by the mean of the involved FP measurements and multiplying by 100.

## 3.2 Individual comparisons between satellite data records and FP stations

In this section we report on bias profiles of the satellite data records against the FP data from all the stations listed in Table 1. When using terms like, e.g., "above 100 hPa", we always refer to altitudes above 100 hPa, i.e. "above" always means "higher up in the atmosphere". The same applies mutatis mutandis to terms like, e.g., "below 30 hPa". For the selected collocations within the coincidence criteria, SAT minus FP differences have been averaged for each satellite data record's full period of coincident measurements. The comparisons for the individual SAT records vary considerably with respect to their measurement periods and numbers of coincidences. Some of the FP stations have operated their balloon soundings only during campaigns; others provide long-term measurement series based on soundings conducted at regular intervals, like LDR, SJC, BLD, and LIN. We have not separated the available comparisons into long-term and short-term series, nor have we tried to detect any temporal variation in biases for the comparisons described in this section. Drifts of satellite data sets will be tackled in Section 4.

Fig. 4 shows, as an example, the comparison of one SAT data set, namely MLS, to one balloon station, namely TMF, to demonstrate the procedure of bias determination. For this figure, every individual colocated FP profile was treated according to Eq. 5, making use of the MLS averaging kernel and a priori data; since MLS data are provided on a fixed pressure grid, further interpolation to a common vertical grid for calculation of the means over all collocations was not necessary. The comparison was limited to the vertical range above the local tropopause, with the tropopause pressure information estimated from the radiosonde temperature profiles accompanying the FP profiles. The mean SAT and FP profiles, calculated from the closest coincident data pairs, are shown in the left panel of Fig. 4, together with the standard deviations of the respective ensembles, which are displayed as shaded areas. In this example, the mean water vapour profile of the satellite data compares well to the sonde data. Even the variability of SAT and FP mixing ratios are similar, indicating that the random uncertainties of the two measurement types are similar. Further, the deviation between the two measurements due to natural variability appears to be relatively small, i.e. the chosen coincidence criteria are stringent enough to avoid unwanted large differences that could result from location and/or time mismatches between the profiles. The mean bias of the two data sets (middle panel of Fig. 4) above 100 hPa is positive or zero, except for the uppermost data point, i.e. MLS on average has a positive bias relative to the FP, which is at maximum +5 %. Below 100 hPa, close to the tropopause it has a sharp peak of -25 %. The standard error of the mean bias is very small at all satellite reporting levels, i.e. the bias assessment is quite accurate throughout the entire profile. The number of comparisons, in this case between ≈30 at the upper and lower ends of the profiles and about 350 for the central part of the profiles, is high and provides the good accuracy of the bias determination. Similar figures for other SAT–FP pairings are provided in the supplement to this paper.
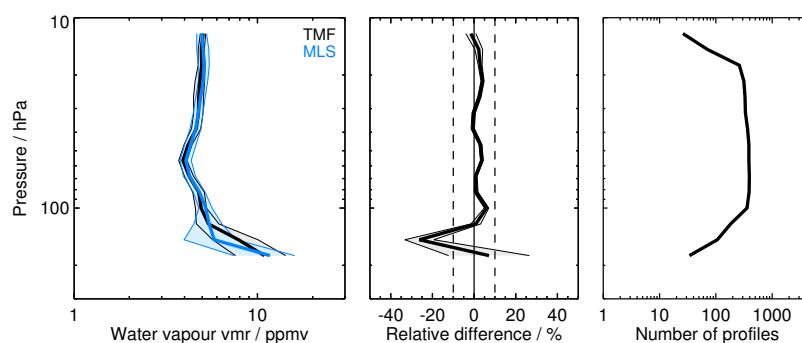
**Figure 4.** Comparison of MLS water vapour profiles with FP profiles at TMF; mean profiles over all coincidences are shown. The individual profiles were cut at the respective tropopause before averaging. Left panel: mean profiles (TMF FP: black, MLS: blue) and their standard deviations (shading with thin black/blue lines as boundaries). Middle panel: Relative mean bias and twice its standard error of the mean (shading, $2\sigma_{bias}$), calculated as the mean differences SAT-FP divided by the mean FP profile and multiplied by 100; the vertical dashed lines enclose the $\pm 10\,\%$ range. Right panel: number of data points along the vertical grid. This number can vary over the vertical range, depending on the altitude coverage of the individual coincident SAT and FP profiles, respectively.

### 3.3 Mean biases of the satellite data records by FP stations

In the following, the comparison of all available satellite data records to one specific FP station is discussed. This comparison provides some insight into potential latitudinal dependencies of the satellite data records' biases. Peculiarities specific for certain FP stations may also show up.

As an example for the comparison of all satellite data to one specific balloon sonde station, Fig. 5 presents the mean relative differences of all available satellite data records to the FP station at BLD. All satellite records having collocations with BLD balloon soundings are shown in this comparison. The presentation is similar to the middle panel of Fig. 4, but for multiple satellites. For the color coding we refer to Fig. 2. The biases of the SAT data records are shown in two panels in order not to overload the figures. We follow Nedoluha et al. (2017) and separate the data sets in non-MIPAS and MIPAS satellite data.

We find that for most of the satellite data records, the bias with respect to the BLD FP data is less than 10 % in the stratosphere between 100 hPa and the upper end of the FP soundings around 10 hPa. Between 100 hPa and the respective tropopause (i.e. the lower end of the profiles), the differences for some SATs become far larger, and for many tend to be negative. This is a typical behaviour that can be observed for many stations, mainly in the mid and high latitudes. In the tropics, however, such a systematic behaviour of the biases is not obvious. Maybe this is because the profiles cut at the local tropopause for tropical sites scarcely reach below 100 hPa (see figures in Appendix A2). It is currently unclear what causes these large deviations at the extra-tropical FP sites.
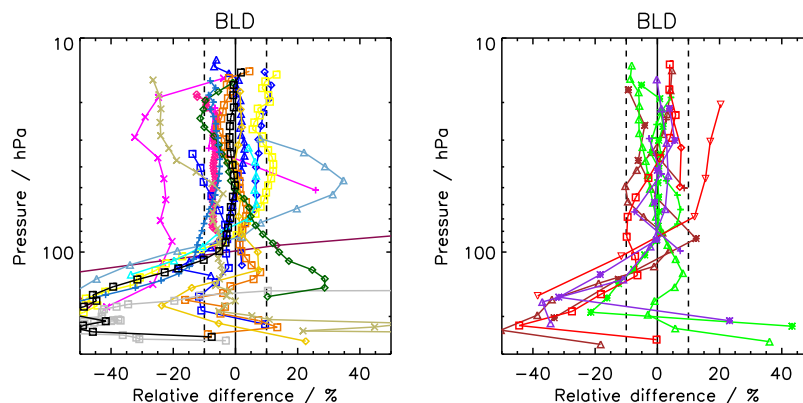
**Figure 5.** Mean relative differences between satellite and BLD FP profiles. left: all data records except MIPAS; right: all MIPAS data records. The color coding and symbols for the satellite records are provided in Fig. 2. The profiles were cut at the respective local tropopause before averaging.

In case of the comparison to BLD soundings, we identify some large biases that more consistently exceed $\pm 10\,\%$ in the
295 stratosphere above $100\,\text{hPa}$. These are SM5 and GOM, which both show negative biases, and POM and MEH, showing positive biases. All other satellite data sets are largely within $10\,\%$ relative difference to the BLD FP data above $100\,\text{hPa}$.

Fig. 6 presents an example for HIL ($20°\,\text{N}$), a subtropical northern hemisphere station. Generally, we observe some of the same characteristics as for BLD. Due to the higher tropopauses at the lower latitude sites, the profiles contain little data at pressures $> 100\,\text{hPa}$. Nevertheless, the negative deviations from the FP data again become larger at the lower end of the
300 profiles. SM5, GOM, HAL, SG2, SC3, MBR, MEH, MEM, MIH, MIM, MOH, MOM exhibit biases larger than $10\,\%$ at multiple altitudes. The biases of the MIPAS data are mostly positive above $100\,\text{hPa}$.

The SAT data records that have large and repetitive biases identified in the BLD and HIL comparisons have larger deviations in most, if not all, FP stations, too. However, the comparisons to many FP stations having a smaller number of coincidences with SAT data due to shorter and/or less dense measurement records, have a large spread (not shown in the figures), and the
305 bias determination is less accurate. In general, we can state that the satellite data records that perform well (i.e., virtually all biases $< 10\,\%$) do so with FP stations having both large and small number of collocations. These are, in alphabetical order, ACE, HAL (except for the well-known $10\,\%$ bias over the entire profile above $100\,\text{hPa}$, Randel et al. (2006); Scherer et al. (2008)), MIPAS ESA and IMK/IAA, MLS, SG2, SG3, SC1, SC4, and SOFIE (see Figs. A1–A3).

### 3.4 Mean biases of the satellite data by data record

310 Comparison of one satellite data set to several balloon sounding stations provides some insight on how the agreement between SAT and FP may be dependent on latitude. Fig. 7 provides, as an example, the comparison of HIR data to all FP stations for which coincident measurements were available (discussion, see below). Similar figures for all satellite data sets are presented in the appendix (Figs. A1–A3). In the following, we discuss the typical bias behaviour for all these satellite data sets.
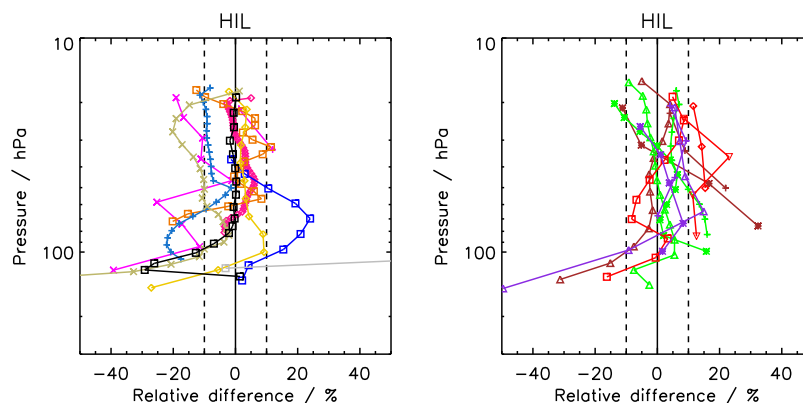
**Figure 6.** Same as Fig. 5 but FP profiles from HIL.
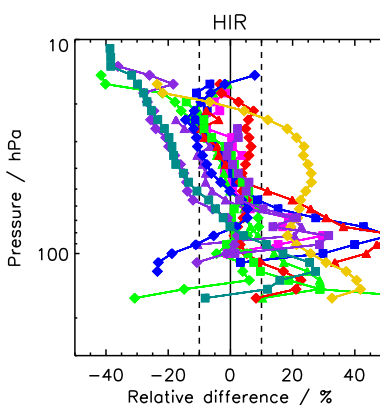
**Figure 7.** Mean relative differences between all the FP stations and the HIR satellite data record. The frost point hygrometer data were adjusted to the vertical grid and resolution of HIRDLS by a smoothing kernel, and the profiles were cut at the tropopause before averaging.

### ACE-FTS v3.5 (ACE)

315 ACE (Fig. A1) is in the $\pm 10$ % range for most of the stations above 100 hPa. Larger negative deviations in this altitude range are found in the lower stratosphere for the southeast Asian stations LSA and KMG. At the upper end of the bias profiles, deviations are negative and the satellite data deviate by more than -10 % from stations BEL and TMF. The overall impression is that the ACE biases follow a bent curve with slight negative deviations at the upper end and stronger negative deviations at the lower end of the profiles, and shows excellent agreement with the frost point hygrometer profiles between approximately

320 20 and 80 hPa. For stations in the middle to high latitudes of the Northern hemisphere, where ACE has its densest sampling, the agreement with frost point hygrometer data is, with deviations within $\pm 5$ %, excellent (e.g. LIN, SOD). The uncertainties of the mean biases (in terms of twice the standard error of the mean, 2·SEM, $2\sigma_{bias}$) are very small in the stratosphere between 100 and 30 hPa, leading to significant biases, at least in the middle to high latitudes, with the exception of BLD where the

biases are small and insignificant. In the low latitudes, the uncertainties of the biases are larger, leading to insignificant biases
325    between 60 and 30 hPa.

**GOMOS (GOM)**

Deviations from frost point hygrometer data for GOM (Fig. A1) vary strongly and cover, except for the comparison to the HAN
and the LDR stations, all the range between -40 % to 0 %. Above 100 hPa the biases have a tendency to be on the negative
side, and to show increasingly larger negative values with increasing altitude. The comparison to the station data of LDR and
330    HAN are significant outliers within this comparison, with large positive biases. The biases are all significant at the $2\sigma_{bias}$ level,
except for a few single points where the bias profiles show zero-crossings. The large spread between the stations indicate a
significant latitude dependence of the GOMOS data which might be due to the different stars used as occultation light sources
during measurement.

**HALOE (HAL)**

335    For HALOE v19 data (Fig. A1), the well-known negative bias in the order of -10 %, seen in measurements after 2001 (Randel
et al., 2006; Scherer et al., 2008) over large parts of the stratosphere is confirmed by the comparisons presented here. Except
for some altitudes, at HUN, KIR, SOD, and SGP the deviations from the frost point hygrometers always stay on the negative
side, with smallest deviations between -5 and -10 % in the 20 to 60 hPa range, and larger deviations of up to -30 % above and
below. The biases are almost all significant, the only exceptions are rare zero-crossings of the bias profiles, for example for
340    comparisons to the KIR station data. The compactness of the biases of all the stations indicates that HALOE data have a very
similar performance over all covered latitudes and related atmospheric conditions.

**HIRDLS (HIR)**

For most of the stations — with the exceptions of LDR and SOD — the comparison demonstrates a general positive to negative
tilt in the bias with increasing altitude for HIR (Fig. A2): Obviously the HIR observations have a positive bias in the order
345    between 0 to 40 % near 100 hPa and end with a similarly strong negative bias around 10 hPa. The difference with respect to the
LDR balloon soundings is more or less constant between 0 and +10 % above 100 hPa, while the difference to SOD is roughly
constant at about +25 % up to 30 hPa, and decreases to -20 % from 30 hPa to the upper end of the profile. Below 100 hPa, the
biases show a wide spread with some focus around +20 %. The collection of mean difference profiles confirms that the tilted
bias, from >20 % near the tropopause to <-20 % around 10 hPa is a distinct property of the HIRDLS water vapour observations.
350    The $2\sigma_{bias}$ uncertainties of the biases are small which makes the biases significant everywhere except near the zero-crossings
of the bias profiles.

**ILAS-II (ILA)**

ILA (Fig. A1), could be compared to three stations only; the agreement above 100 hPa is within $\pm 10$ % for BLD and SOD,
while ILAS-II deviates from the frost point hygrometer soundings at NYA by -30 to -40 %. Below 100 hPa, the deviations
355    cover the range between > -40 % and +30 %. The biases are all significant except near the zero-crossings of the bias profiles.

**MAESTRO (MST)**

MAESTRO (Fig. A1) has very few measurements of water vapour above the tropopause. The bias profiles mostly change from
-40 % at about 200 hPa to +40 % around 90 to 100 hPa, with large uncertainties, but nevertheless significant deviations. Since
we are at the upper limit of MAESTRO's measurement range, we refer to a more appropriate comparison that is provided in a
360  companion paper dealing with upper tropospheric humidity (Read et al., 2022).

**MIPAS (MBR, MER, MIR, MOR)**

For MIPAS, all observation modes and data versions from the four processors are shown in Fig. A3. Above 70 hPa, the compar-
ison to the frost point hygrometer data for the NOM RR modes remains within ± 10 % for most cases. The MOR data set is the
most compact one, with biases for most stations between +10 and -10 % for the altitude range of 80 to 20 hPa. Above 20 hPa,
365  a tendency to larger negative biases exists, while below 80 hPa, the profiles develop an increasingly negative bias. The MOH
data set is less compact, and it shows a pronounced high bias in the range of 70 to 100 hPa. The MOM data set has a bias >
10 % at the lower end of the profiles, and a smaller, almost zero bias at the upper end of the profiles. The MIR bias is also quite
compact, however it has a rather pronounced tilt from positive values around +20 % near the tropopause to -10 to -20 % above
30 hPa. Again, for MIH the biases to the various FP sites are less compact, but follow in general the characteristics of the MIR
370  data set. The biases for the MIM data set show S-shaped profiles with positive biases >+10 % at the lower and much smaller
positive biases at the upper end, similar to MOM. For MER and MBR, the biases with respect to the frost point hygrometer
stations have a somewhat larger scatter than that of MOR and MIR. Most of their biases remain between ± 10 %, however,
some prominent outliers below 60 hPa and above 20 hPa exist.

MBR biases for the Northern mid and high latitudes have mostly small uncertainties and are significant except near the
375  zero-crossings of the bias profiles. In contrast, the number of comparisons for the low latitudes is lower, therefore the bias
uncertainties are higher, leading very often to insignificant biases. For the LDR FP site (the only in the Southern mid to high
latitudes), the biases above 100 hPa are small, and despite small uncertainties, they are insignificant. MER and MOR behave
similar to MBR for all FP sites except LDR; deviations from the LDR FP measurements are larger than for MBR and they are
significant over the full altitude range of the profiles. MIR biases are more often significant (although small) than for the other
380  MIPAS data sets. In particular, in the tropics where the biases of the other MIPAS data sets often have higher uncertainties,
the MIR biases are significant except for the region of the zero crossings of the profiles. The bias profile with respect to the
LDR station is significant except in the troposphere below 200 hPa and at its zero crossing. The biases for MIPAS HR and MA
observations have, in general, larger uncertainties, mainly due to the smaller number of coincidences, and are therefore more
often insignificant. Very often, the biases are below ±10 % and the uncertainties are larger than that. Consistently significant
385  biases can be found, in general, for larger biases. This is the case for MBH in the comparisons with Northern high and mid
latitude stations, larger parts of the MEH profiles in all latitudes, and for all MIH and MIM biases that are larger than 2 or 3 %
(absolute). MBM biases are mostly not significant, while MEM biases mostly are. MOH and MOM behave very much like
MEH and MEM in terms of significance of their biases.

**MLS (MLS)**

390 MLS (Fig. A2) reveals a very compact set of biases that are almost all in the ±10 % range from 70 to 10 hPa. Exceptions are the comparisons to stations of the maritime continent, i.e. BIK, BND, Tawara, KTB, and RVM. For these stations, MLS shows a high bias up to +30 % in the altitude range of 100 to 70 hPa. Below 100 hPa, MLS tends to develop a low bias with a peak at -25 % around 200 hPa, and a better consistency to frost point hygrometer data, again in the ±10 % range, close to the local tropopause. In the Northern high and mid latitudes, the uncertainties of the biases are extremely small, leading to significant

395 but small (mostly +5 to 10%) deviations from the FP data in the stratosphere above 100 hPa. Even in the low latitudes, the uncertainties are small enough to make most of the deviations significant, in particular the deviations from FP stations in the maritime continent. The bias with respect to the LDR FP site is below 5% above 70 hPa and significant, due to the extremely small uncertainty. The larger biases below 100 hPa are mostly significant, too.

**POAM-III (POM)**

400 POM3, as an instrument covering the Northern high latitudes only, has coincidences with the most Northern stations NYA, SOD, KIR and BLD (Fig. A1). The biases to the three former stations are rather small, providing curved bias profiles with negative values around -20 % below 100 hPa and above 20 hPa, and positive values of up to +20 % between 100 and 40 hPa. The comparison with the sonde data of BLD gives a somewhat different picture: here the biases increase from extreme negative values beyond -40 % below 100 hPa to +35 % around 50 hPa and remain in the 10 to 35 % range above. The hygropause in

405 the POM3 profiles near BLD, i.e. the altitude with lowest water vapour vmr, is much lower than in the frost point hygrometer data, and the mean profile below the hygropause is displaced towards lower altitudes which both contribute to the increasingly larger negative bias below 100 hPa. All biases are significant except near the zero-crossings.

**SAGE-II (SG2)**

SG2, as another instrument besides HALOE, providing water vapour observations for decades, has been used a lot for con-

410 struction of longer-term global water vapour data time series. The comparisons to frost point hygrometer station data (Fig. A1) provide some scatter, however the most data points lie within ±20 % deviation, and a larger part also within ±10 %. The comparisons to SOD form an exception with positive deviations of approximately 25 % over a larger part of the stratosphere (90 to 35 hPa). The number of coincidences, however, is very small (around 10) for this FP site. The uncertainties of the bias profiles are often in the order of ±5% which makes some deviations in the Northern high and mid latitudes insignificant. The bias with

415 respect to BLD FP observations, however, is significant in the stratosphere above 100hPa despite the deviations being less than 5% over a large part of the profile. The same is true, although to a lesser part of the profile, for the comparison to the LDR FP site (the altitude near the zero-crossings of the profiles always excluded). For the low latitude FP sites, a general comment is difficult to make because of strongly oscillating bias profiles and sometime considerable uncertainties. Nevertheless, also in this latitude region significant biases can be found despite larger uncertainties.

420 **SAGE-III (SG3)**

SG3 data (Fig. A1) could be compared to the most Northern stations (NYA, KIR and SOD), BLD and LDR only. The comparisons agree well, indicating a bias within the ±10 % range, with two outliers in the order of +20 % at about 70 and 15 hPa.

The comparison to the BLD balloon data indicates, similar to POM3, a too low lying hygropause and displacement towards lower altitudes of the profile part below as the reason for the increasingly large negative bias below 80 hPa. Uncertainties of
425 the biases are small enough to make the biases significant over the full altitude range, except near the zero-crossings.

**SCIAMACHY (SC3, SC1, SC4, SCL)**

The SC3 observations (Fig. A2) cover the altitude range from the tropopause up to approx. 30 hPa. Above 100 hPa, the comparisons to the FP stations provide a large scatter from -40 to more than +50 %. The stations at the Maritime continent and the Indian ocean (pink, purple and partly red colours) seem to provide the highest positive biases, while for Northern midlatitude
430 stations the biases tend to be in the ±10 % range and rather on the negative side. Near to the upper end of the SCIAMACHY profiles, around 30 hPa, the deviations to the sonde station data converge to a bias range of -20 to +5 %, with only two outliers, and most of the biases within the ±10 % range. The biases with respect to the BLD FP site are significant below 150 hPa and above 70 hPa. Other sites for which the biases turn out to be significant, at least over a large part of the profiles, are SOD, LIN, and SGP, all revealing negative biases. Biases for HIL, SJC, TRW, BIK, RVM, and LRN are significant for at least a part of
435 the profile, and positive, while the comparison to LDR shows insignificant and rather small biases. The SC1 and SC4 data sets, both solar occultation observations, have in common that only stations from moderate to high northern latitudes contribute to the comparisons (Fig. A1). The SC1 relative biases are within ±10 % between 100 and 20 hPa, and above and directly below this altitude range tend towards more negative values. Biases and uncertainties are rather small and in the same order of magnitude for KIR and NYA. Therefore the biases are largely insignificant. The other stations reveal significant biases. SC4 shows
440 biases within ±10 % between 100 and 20 hPa for 4 out of 6 FP sites. Biases with respect to NYA and KIR, however, are at about 10 % at 100 hPa, but decrease to about -20 % at 30 hPa. All biases except near the zero-crossings of the bias profiles are significant. The SCIAMACHY lunar occultation data set has coincidences with the FP station of LDR only (Fig. A1). The bias for this site is between 10 and 20% and significant at all altitudes.

**SMILES (SLA, SLB)**

445 The comparison of SMILES water vapour observations from their channel A and B (Fig. A2) to frost point hygrometer sonde data shows a large scatter, and prominent biases reaching from -40 % and more below 100 hPa to +40 % and more between 40 and 30 hPa. Due to the short mission lifetime of SMILES, the number of coincidences with frost point hygrometer soundings is, however, very limited. As a consequence, uncertainties of biases are rather large. Nevertheless, the biases are significant except near the zero-crossings of the profiles.

**SMR (SM5, SM4)**

450 SMR 544 GHz comparisons with frost point hygrometer data reveal a large scatter of the biases over ±40 % and more (Fig. A2). Most of the bias profiles are negative, and there seems to be a certain concentration of biases around -30 %. There is no latitude dependence obvious. Despite the long lifetime of the SMR mission (which is still operational at the time of this writing) and its rather dense global coverage, the number of coincidences range between 10 and 100 for most of the stations
455 only. The spread of the coincident SM5 profiles is, however, far larger than the spread of the frost point hygrometer profiles,

indicating that the SM5 profiles have a considerable measurement error (see Fig. S29 in the Supplement). The SMR 489 GHz observations of water vapour are available above 50-60 hPa only. They are more compact than the 544 GHz observations, with biases between -20 and +20 %, and most data points falling into the ±10 % range. However, for this observation channel, a much smaller amount of stations providing coincident balloon soundings is available, and the number of coincidences per station is often below 10. Therefore, the biases have often large uncertainties. Nevertheless, most of the biases are significant.

**SOFIE (SOF)**

For SOFIE, comparisons with frost point hygrometer data from SOD, BLD, LIN, HIL, SJC, and LDR are available (Fig. A1). The comparisons to all frost point hygrometer data are very compact. Above 100 hPa, almost all data points of the biases fall into the ± 10 % range, and many of them are even closer than ± 5 % to the frost point hygrometer data. The uncertainties are small, making even the tiny deviations from FP measurements at BLD or LDR significant.

## 3.5   Synopsis of the bias assessment

For the overall assessment of biases of SAT records against FP profiles we have averaged the results from all stations in each of the following three pressure ranges: tropopause to 100 hPa, 100 to 30 hPa, and 30 to 10 hPa. The average biases and their standard errors were calculated with Eqs. 9 and 11, respectively. They are listed in Tables 3 and 4. In Fig. 8 thick horizontal bars show average biases plus/minus twice the respective standard errors to indicate the 95 % confidence limits. Additionally, the 5- % and 95- % percentile values are marked in the plots and listed in the tables.

In the 10–30 hPa altitude range, most of the satellite data records have mean biases within the ± 10 % range, and some of them show even better overall agreement with the FP data. ACE, MLS, SG2, SG3, SC1, SC4, and SOF have very accurately determined biases of smaller than ±5 %. Data records with mean biases larger than ±5 % but less than ±10 % are HAL, POM, and SM4. For all three, the bias accuracy is very good. Data records with biases larger than ±10 % are ILA, MST, SLA, SLB, SM5, GOM, and HIR. Except for MEM and MIH , the MIPAS data records are all within the ±10 % range regarding their bias in the 10 to 30 hPa altitude range with the majority of these being within the ±5 % range. The three ESA data products have all a positive bias, while the three IMK/IAA data records have all a negative bias. Except for the ESA product, water vapour derived from the MIPAS middle atmosphere measurement mode shows agreement with FP data of better than 5 %.

In the altitude range of 30–100 hPa, HIR, SG2, SG3, SOF, SM4, SLB, SC3, SC4, SC1, ACE, and MLS have biases less than ±5 %. However, SM4 and SLB have very large uncertainties of the biases. HAL and ILA have biases less than ±10 %. POM, GOM have biases just greater than 10 % in the 30 to 100 hPa altitude range, while the biases of SM5, SLA, and MST far exceed 10 %. Except MOM, MEM, and MEH, MIPAS data sets fall into the ±10 % bias range. MOR, MOH, MBR, MBH, MER, MIR, and MIH show even biased lower than ±5 %.

In the tropopause to 100 hPa altitude range, the biases, and especially the data spread, become large for many of the SATs. Data records with biases below ±10 % are POM, GOM, SG2, SG3, SC4, ACE, and MLS. Of these, POM, SG2, and SG3 show biases below ±5 %. ACE just misses this value. For the MIPAS data sets, MOR, MOH, MBR, MBH, MER, MIR, and
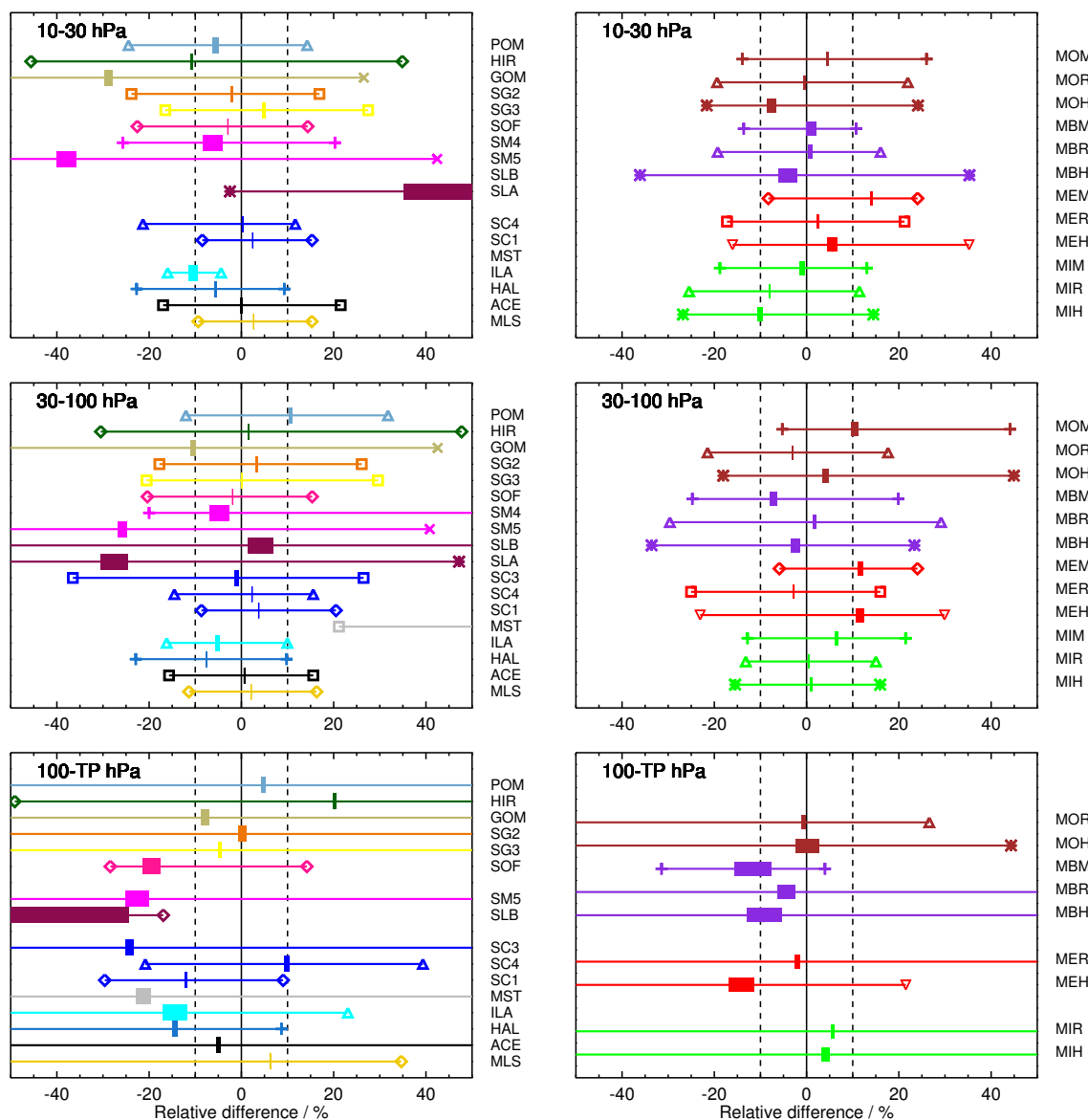
**Figure 8.** Relative differences of satellite and FP data averaged over all sites and the three pressure ranges 10–30 hPa (top row), 30–100 hPa (middle row), and 100 hPa to tropopause (bottom row). The panels on the left show comparisons for all satellite instruments except MIPAS. In the right column all MIPAS comparisons are displayed. For color coding and symbols see Fig. 2. Thin lines between symbols span the 5–95 % range of the data, while thick bars indicate the range of twice the standard errors (2×SEM) around the mean biases. The actual average bias values are given by the center of the thick bars. For the 10–30 hPa panel, SMB and MST biases and the 5–95 % range of data are beyond the relative difference scale (see Table 3 for the actual values).

**Table 3.** Tabulated data used in the left columns of Fig. 8 for the non-MIPAS data sets. 100–TP means the pressure range from 100 hPa down to the tropopause. Statistically significant biases are presented in boldface text. Note that the column "Rel. std.err." does not give the relative value for 2×SEM used for the plots in Fig. 8, but only the relative value for the SEM.

| SAT label | P range / hPa | Rel. bias / % | Rel. std.err. / % | 5-% percentile | 95-% percentile |
|---|---|---|---|---|---|
| ACE | 10–30 | -0.026 | 0.144 | -16.971 | 21.514 |
|  | 30–100 | **0.638** | 0.065 | -15.678 | 15.584 |
|  | 100–TP | **-5.013** | 0.209 | -380.964 | 117.333 |
| GOM | 10–30 | **-28.762** | 0.434 | -74.006 | 26.549 |
|  | 30–100 | **-10.539** | 0.279 | -62.126 | 42.556 |
|  | 100–TP | **-7.969** | 0.441 | -128.002 | 104.097 |
| HAL | 10–30 | **-5.606** | 0.146 | -22.702 | 9.299 |
|  | 30–100 | **-7.520** | 0.080 | -22.894 | 9.806 |
|  | 100–TP | **-14.403** | 0.284 | -267.165 | 8.682 |
| HIR | 10–30 | **-10.836** | 0.121 | -45.594 | 34.876 |
|  | 30–100 | **1.509** | 0.061 | -30.474 | 47.731 |
|  | 100–TP | **20.139** | 0.188 | -49.116 | 101.612 |
| ILA | 10–30 | **-10.512** | 0.472 | -15.973 | -4.401 |
|  | 30–100 | **-5.160** | 0.238 | -16.187 | 9.974 |
|  | 100–TP | **-14.386** | 1.304 | -56.345 | 23.072 |
| MST | 10–30 | **261.870** | 51.987 | 158.897 | 2602.118 |
|  | 30–100 | **105.646** | 1.795 | 21.137 | 383.422 |
|  | 100–TP | **-21.142** | 0.793 | -217.222 | 168.084 |
| MLS | 10–30 | **2.634** | 0.043 | -9.359 | 15.280 |
|  | 30–100 | **2.147** | 0.022 | -11.376 | 16.314 |
|  | 100–TP | **6.257** | 0.058 | -150.251 | 34.656 |
| POM | 10–30 | **-5.568** | 0.325 | -24.457 | 14.260 |
|  | 30–100 | **10.644** | 0.191 | -12.016 | 31.779 |
|  | 100–TP | **4.689** | 0.236 | -99.721 | 179.944 |
| SG2 | 10–30 | **-2.113** | 0.152 | -23.830 | 16.922 |
|  | 30–100 | **3.387** | 0.149 | -17.764 | 26.080 |
|  | 100–TP | 0.121 | 0.447 | -111.409 | 56.102 |
| SG3 | 10–30 | **4.824** | 0.160 | -16.489 | 27.473 |
|  | 30–100 | 0.138 | 0.092 | -20.543 | 29.615 |
|  | 100–TP | **-4.751** | 0.181 | -175.203 | 92.155 |
| SC3 | 30–100 | **-0.975** | 0.232 | -36.536 | 26.485 |
|  | 100–TP | **-24.304** | 0.466 | -106.064 | 64.791 |
| SC1 | 10–30 | **2.440** | 0.046 | -8.474 | 15.293 |
|  | 30–100 | **3.769** | 0.029 | -8.665 | 20.503 |
|  | 100–TP | **-12.099** | 0.103 | -29.608 | 9.048 |
| SC4 | 10–30 | **0.238** | 0.077 | -21.325 | 11.679 |
|  | 30–100 | **2.327** | 0.044 | -14.521 | 15.566 |
|  | 100–TP | **9.746** | 0.286 | -20.819 | 39.344 |
| SLA | 10–30 | **45.905** | 5.340 | -2.487 | 82.150 |
|  | 30–100 | **-27.460** | 1.469 | -231.510 | 47.187 |
| SLB | 10–30 | **160.216** | 9.478 | 131.965 | 522.533 |
|  | 30–100 | **4.142** | 1.394 | -185.750 | 202.804 |
|  | 100–TP | **-60.679** | 18.084 | -174.678 | -16.925 |
| SM4 | 10–30 | **-6.263** | 1.073 | -25.716 | 20.280 |
|  | 30–100 | **-4.710** | 1.055 | -20.002 | 64.179 |
| SM5 | 10–30 | **-37.846** | 1.047 | -79.713 | 42.433 |
|  | 30–100 | **-25.821** | 0.486 | -72.709 | 40.834 |
|  | 100–TP | **-22.650** | 1.261 | -160.116 | 107.726 |
| SOF | 10–30 | **-2.923** | 0.046 | -22.563 | 14.389 |
|  | 30–100 | **-1.906** | 0.044 | -20.371 | 15.347 |
|  | 100–TP | **-19.456** | 0.970 | -28.404 | 14.192 |

**Table 4.** Tabulated data used in the right columns of Fig. 8 for the diverse MIPAS data sets. 100–TP means the pressure range from 100 hPa down to the tropopause. Statistically significant biases are presented in boldface text. Note that the column "Rel. std.err." does not give the relative value for 2×SEM used for the plots in Fig. 8, but only the relative value for the SEM.

| SAT label | P range / hPa | Rel. bias / % | Rel. std.err. / % | 5-% percentile | 95-% percentile |
|---|---|---|---|---|---|
| MBH | 10–30 | **-4.094** | 1.003 | -36.112 | 35.307 |
| | 30–100 | **-2.386** | 0.495 | -33.585 | 23.368 |
| | 100–TP | **-9.228** | 1.900 | -236.371 | 62.386 |
| MBM | 10–30 | 1.043 | 0.566 | -13.619 | 10.737 |
| | 30–100 | **-7.233** | 0.378 | -24.762 | 19.895 |
| | 100–TP | **-11.620** | 2.010 | -31.430 | 3.986 |
| MBR | 10–30 | **0.784** | 0.250 | -19.268 | 16.029 |
| | 30–100 | **1.690** | 0.174 | -29.630 | 29.141 |
| | 100–TP | **-4.344** | 0.969 | -221.124 | 78.809 |
| MEH | 10–30 | **5.630** | 0.535 | -16.050 | 35.212 |
| | 30–100 | **11.597** | 0.413 | -23.067 | 29.887 |
| | 100–TP | **-14.034** | 1.376 | -116.286 | 21.503 |
| MEM | 10–30 | **14.065** | 0.111 | -8.291 | 24.056 |
| | 30–100 | **11.649** | 0.301 | -5.897 | 24.034 |
| MER | 10–30 | **2.424** | 0.155 | -17.250 | 21.233 |
| | 30–100 | **-2.766** | 0.095 | -25.041 | 15.995 |
| | 100–TP | **-1.983** | 0.260 | -152.016 | 51.576 |
| MIH | 10–30 | **-10.083** | 0.274 | -26.791 | 14.477 |
| | 30–100 | **1.105** | 0.151 | -15.490 | 15.954 |
| | 100–TP | **4.162** | 0.456 | -84.822 | 54.517 |
| MIM | 10–30 | **-1.044** | 0.269 | -18.790 | 13.070 |
| | 30–100 | **6.569** | 0.180 | -12.795 | 21.461 |
| MIR | 10–30 | **-7.997** | 0.102 | -25.466 | 11.439 |
| | 30–100 | **0.398** | 0.053 | -13.176 | 15.017 |
| | 100–TP | **5.608** | 0.159 | -68.337 | 78.580 |
| MOH | 10–30 | **-7.522** | 0.472 | -21.631 | 24.156 |
| | 30–100 | **4.056** | 0.358 | -18.068 | 44.929 |
| | 100–TP | 0.252 | 1.251 | -89.854 | 44.318 |
| MOM | 10–30 | **4.440** | 0.121 | -13.914 | 25.995 |
| | 30–100 | **10.517** | 0.363 | -5.259 | 44.089 |
| MOR | 10–30 | **-0.495** | 0.154 | -19.394 | 21.946 |
| | 30–100 | **-3.035** | 0.078 | -21.461 | 17.647 |
| | 100–TP | **-0.714** | 0.225 | -194.747 | 26.589 |

MIH are within ±10 % bias. MOR, MOH, MBR, MER, and MIH even stay within the ±5 % range. For all data records, the uncertainties of the biases increase compared to the other altitude ranges.

## 4 Assessment of drifts in satellite-FP differences

Linear temporal trends in the relative differences between stratospheric water vapour mixing ratios reported for satellite (SAT) and frost point hygrometer (FP) measurements are hereinafter referred to as "drifts", expressed in units of % yr$^{-1}$. Relative differences in SAT and FP mixing ratios (100 %×(SAT-FP)/FP) were calculated using FP mixing ratios as the divisors. As the bias analyses above, this investigation of drifts is based on FP and satellite-based profile measurements that are coincident in space and time according to the criteria provided in Section 2.2. Also analogous to the bias evaluations, before comparing to SAT profiles, the vertical resolution of each FP profile was degraded to that of the corresponding satellite's reporting levels and placed on its grid of reporting pressures (i.e., convolved) using satellite-specific averaging kernels or more generic Gaussian-shaped smoothing kernels (see Section 2.3 and Table 2). To minimise the influences of tropospheric air on this evaluation of stratospheric measurements, the reporting levels of SATs were limited to those above the local tropopause (Section 2.2). Similar analyses of biases and drifts between SAT and FP measurements of tropospheric water vapour have already been published by Read et al. (2022).

Several of the methods employed here to evaluate drifts are slightly different from those used above for the bias analyses. In cases where multiple profiles from a given SAT were identified as coincident with a FP profile (a "coincident cluster"), the median SAT mixing ratio and standard error of the mean SAT mixing ratio were calculated for each cluster, as was done in Hurst et al. (2014). The convolved or smoothed FP mixing ratio profile (see Section 2.3) was then subtracted from the median SAT mixing ratio profile. The advantage gained using median mixing ratios instead of averages is that they are much more resistant to skew by statistical outliers. The SAT – FP mixing ratio differences (in ppmv) were divided by the FP mixing ratios, then multiplied by 100% to produce the SAT – FP relative differences analysed here. For each unique pair of FP sites and satellites, the drift at each SAT reporting level was independently determined using a weighted linear regression fit to the time series of SAT – FP relative differences, with statistical weights based on the standard error of the mean SAT mixing ratio for each coincident cluster.

### 4.1 Evaluation of data records for drift analysis

Unlike the bias evaluations, an analysis of drift in SAT – FP relative differences requires a sufficient number of differences over an adequately long period of time to detect and determine statistically robust trends. The unique pairs of SATs and FPs evaluated for biases (see Tables 1 and 2) included many with short and/or sparse time series of relative differences. Typically, the statistical uncertainties of drifts determined for these pairs were large and the calculated drifts were very sensitive to the removal of one data point from the time series. Simple tests of drift uncertainties and sensitivities for time series with varying lengths and data densities were performed. The results revealed that robust drift statistics were consistently produced for timeseries >5 years in length and composed of at least one difference in 67% of the years covered. Of all the unique SAT – FP pairs analysed for biases, only 64 provided difference time series that met these more stringent criteria for drift analysis. This excluded several time series >5 years in length but with only short-term "bursts" of FP data from intensive measurement campaigns.

Table 1 identifies the seven FP sounding sites that paired with SATs to provide at least one time series of differences that met the record length and data density criteria. Table 2 identifies the 20 SAT retrievals that paired with FP sites to produce at least one qualifying time series. Of these 20 SAT retrievals, 8 were for MIPAS, 3 for SCIAMACHY, 2 for SMR and one each for 7 other SATs. In total, 1146 time series of SAT – FP relative differences were analysed for drift at SAT reporting pressures ranging from 275 to 13.1 hPa. These time series are based on several thousand satellite profiles and just over 900 unique FP profiles over 7 sites.

The numbers of reporting levels for each unique SAT – FP pair that met the drift analysis criteria are presented in Tables 5 and 6. Each SAT–FP pair provided difference time series at an average of 18 reporting levels over an average of 3 FP sites, though the individual coverages ranged widely from 1 to 74 reporting levels and 1 to 7 sites. For example, MIPAS Bologna V5R v2.3 MA ("MBM") retrievals produced only a single qualifying time series of differences at 43 hPa over the BLD site, while MAESTRO ("MST") retrievals produced qualifying difference time series at an average of 50 reporting levels over each of 4 FP sites. Since the records for HAL and SG2 extend only 5–6 yrs into the new millenium, they overlap adequately only with the the FP record at BLD. The MLS is the only SAT that paired with all 7 FP sites.

## 4.2 Methods for quantifying drifts

Each time series of SAT – FP differences was examined for statistical outliers by performing a preliminary standard linear regression analysis. Data points with residuals from the fit line that were >2.5 times the mean of the absolute values of the residuals were omitted from further analysis. This method of outlier filtering removed an overall average of 5.7% of the data points from the time series before they were analysed for drift.

Drifts in SAT – FP differences were determined using weighted linear regression analyses (Fig. 9). The weight $W_i$ applied to each difference was computed as the squared reciprocal of its uncertainty (Eq. 12). The uncertainty $\lambda_i$ of each difference was calculated (in quadrature) from the relative standard error ($\sigma_i$) of the mean mixing ratio of the "coincidence cluster" of satellite profiles (in %) and the ±6% uncertainty of stratospheric water vapour measurements by FPs (Hall et al., 2016; Vömel et al., 2016). Each fitting weight was then scaled to the 95% level of confidence using the student-t value for the number (n-1) of satellite profiles in each coincidence cluster. In this way, differences with smaller uncertainties had greater weights and therefore stronger influences on the linear regression fits.

$$W_i = \lambda_i{}^{-2}, \text{where } \lambda_i = t_{0.95,i} \sqrt{\sigma_i{}^2 + 0.06^2} \tag{12}$$

SAT – FP differences based on only a single coincident satellite profile (a rare occurrence) were assigned the smallest weight calculated for the entire time series. Consequently, these single profile differences had the weakest influence on the linear regression analyses of drift.

The slope of the weighted regression fit line for each time series of relative differences was utilised as the best statistical estimator of the linear temporal drift (% yr$^{-1}$) in the differences. Similarly, the 95% confidence limits of calculated slopes were considered the best estimates of drift uncertainty and thus used to evaluate the statistical significance of the drifts. In this

**Table 5.** Drift statistics in three pressure intervals for each non-MIPAS SAT across all paired FP sites.

| SAT Code | P range (hPa) | N Levels | FP site # | Drift Mean (% yr$^{-1}$) | Drift Uncertainty ($\pm$% yr$^{-1}$) | Drift 5$^{th}$ percentile (% yr$^{-1}$) | Drift 95$^{th}$ percentile (% yr$^{-1}$) |
|---|---|---|---|---|---|---|---|
| ACE | 10–30 | 18 | 4,7,16,21 | 0.03 | 0.15 | -0.40 | 0.67 |
| | 30–100 | 54 | 4,7,16,20,21 | 0.07 | 0.09 | -0.77 | 0.56 |
| | 100–TP | 48 | 4,7,16,20,21 | 0.20 | 0.32 | -2.57 | 4.80 |
| GOM | 10–30 | 10 | 4,14,21 | **-2.12** | 1.41 | -5.15 | 2.44 |
| | 30–100 | 33 | 4,14,21 | **-1.14** | 0.60 | -2.93 | 2.34 |
| | 100–TP | 45 | 4,14,21 | 0.18 | 0.69 | -3.70 | 6.46 |
| HAL | 10–30 | 3 | 4 | **1.02** | 0.56 | 0.78 | 1.27 |
| | 30–100 | 16 | 4 | **0.82** | 0.26 | 0.17 | 1.49 |
| | 100–TP | 10 | 4 | **2.49** | 1.05 | 1.16 | 4.61 |
| MST | 30–100 | 63 | 4,14,16,21 | **6.14** | 1.06 | -3.65 | 15.41 |
| | 100–TP | 137 | 4,14,16,21 | **0.74** | 0.67 | -5.58 | 9.13 |
| MLS | 10–30 | 16 | 3,4,7,14,16,20,21 | **0.38** | 0.21 | -0.98 | 0.91 |
| | 30–100 | 48 | 3,4,7,14,16,20,21 | **0.51** | 0.08 | 0.08 | 1.35 |
| | 100–TP | 21 | 4,7,14,16,20,21 | 0.11 | 0.30 | -0.75 | 2.31 |
| MLS*[a] | 10–30 | 12 | 4,14,16,20,21 | **0.95** | 0.24 | 0.29 | 1.70 |
| | 30–100 | 35 | 4,14,16,20,21 | **1.18** | 0.12 | 0.22 | 1.74 |
| | 100–TP | 20 | 4,14,16,20,21 | **1.11** | 0.56 | -0.69 | 6.80 |
| SG2 | 10–30 | 5 | 4 | 0.48 | 1.10 | -0.92 | 1.40 |
| | 30–100 | 16 | 4 | 0.33 | 0.38 | -0.91 | 1.21 |
| | 100–TP | 10 | 4 | -1.76 | 1.79 | -14.37 | 0.63 |
| SC3 | 30–100 | 22 | 4,16,21 | **-0.67** | 0.32 | -3.24 | 1.18 |
| | 100–TP | 15 | 4,16,21 | 0.25 | 1.01 | -1.11 | 9.00 |
| SC1 | 10–30 | 27 | 4,16,21 | **0.46** | 0.33 | -1.17 | 1.31 |
| | 30–100 | 50 | 4,16,21 | **0.84** | 0.18 | -0.52 | 2.32 |
| | 100–TP | 10 | 4,16,21 | **1.07** | 0.47 | -0.62 | 4.30 |
| SC4 | 10–30 | 12 | 4,16,21 | 0.19 | 0.74 | -2.03 | 1.11 |
| | 30–100 | 26 | 4,16,21 | -0.25 | 0.28 | -1.42 | 0.76 |
| | 100–TP | 3 | 4,16,21 | -0.59 | 1.38 | -1.64 | -0.28 |
| SM5 | 10–30 | 12 | 4,14,16,20,21 | -0.66 | 0.80 | -2.74 | 2.41 |
| | 30–100 | 31 | 3,4,14,16,20,21 | -0.50 | 0.57 | -3.96 | 1.36 |
| | 100–TP | 13 | 4,14,16,21 | -1.10 | 1.86 | -12.32 | 3.19 |
| SM4 | 10–30 | 2 | 4,21 | -0.75 | 2.58 | -0.87 | -0.71 |
| | 30–100 | 2 | 4,21 | 1.01 | 4.33 | 0.49 | 1.02 |
| SOF | 10–30 | 55 | 16,21 | **1.58** | 0.14 | 0.79 | 2.38 |
| | 30–100 | 83 | 16,21 | **0.86** | 0.11 | 0.14 | 2.26 |
| | 100–TP[b] | 1 | 21 | -0.53 | | | |

Statistically significant drifts are presented in boldface text.

[a] MLS* is a special case for which the MLS dataset is evaluated for drifts after a significant changepoint in each timeseries that typically occured in $\sim$ 2010 (see Sec. 4.2.1).

[b] No uncertainty or percentiles for the mean drift are presented because only one timeseries of SOF-FP differences in the 100–TP pressure interval (at 103 hPa over SOD) was available for drift analysis.

**Table 6.** Drift statistics in three pressure intervals for each MIPAS SAT across all paired FP sites.

| SAT Code | P range (hPa) | N Levels | FP site # | Drift Mean (% yr$^{-1}$) | Drift Uncertainty (±% yr$^{-1}$) | Drift 5th percentile (% yr$^{-1}$) | Drift 95th percentile (% yr$^{-1}$) |
|---|---|---|---|---|---|---|---|
| MBR | 10–30 | 6 | 3,4,14,20,21 | 0.48 | 0.89 | -0.21 | 3.11 |
| | 30–100 | 18 | 3,4,14,20,21 | -0.16 | 0.67 | -1.98 | 3.49 |
| | 100–TP | 8 | 4,14,21 | 0.19 | 2.95 | -9.91 | 5.42 |
| MBM | 30–100$^a$ | 1 | 4 | -1.82 | | | |
| MER | 10–30 | 9 | 3,4,14,20,21 | **-0.53** | 0.49 | -2.50 | 0.09 |
| | 30–100 | 26 | 3,4,14,20,21 | -0.15 | 0.30 | -1.52 | 1.62 |
| | 100–TP | 16 | 4,14,20,21 | 0.12 | 0.78 | -4.20 | 4.79 |
| MEM | 30–100 | 2 | 21 | -0.27 | 1.36 | -0.59 | -0.02 |
| MIR | 10–30 | 16 | 3,4,14,20,21 | **-0.69** | 0.39 | -3.62 | -0.06 |
| | 30–100 | 38 | 3,4,14,20,21 | **-0.25** | 0.19 | -1.17 | 1.09 |
| | 100–TP | 20 | 4,14,20,21 | **-0.75** | 0.71 | -3.23 | 3.67 |
| MIM | 30–100 | 6 | 4 | -1.00 | 1.36 | -2.83 | 0.44 |
| MOR | 10–30 | 12 | 3,4,14,20,21 | -0.35 | 0.67 | -2.79 | 1.10 |
| | 30–100 | 32 | 3,4,14,20,21 | **-0.48** | 0.33 | -1.92 | 1.56 |
| | 100–TP | 17 | 4,14,20,21 | -0.54 | 0.87 | -6.57 | 1.61 |
| MOM | 30–100 | 2 | 21 | **-2.05** | 1.43 | -2.37 | -1.77 |

Statistically significant drifts are presented in boldface text.

$^a$ No uncertainty or percentiles for the mean drift are presented because only one timeseries of MBM-FP differences in the 30-100 pressure interval (at 43 hPa over BLD) was available for drift analysis.

555 analysis, a drift in SAT – FP differences at a given reporting pressure is considered to be statistically significant if the 95% confidence interval of the regression line slope does not include zero.

The vertical profiles of drifts determined for 4 unique SAT–FP pairs (Figs. 10) and all SAT–FP pairs (Figs. B1–B5) illustrate how the 95% confidence intervals determine statistical significance. In these plots, if the 95% confidence interval (full span of the error bar) does not intersect the vertical line for zero drift, the drift is labeled significant. Red (blue) markers indicate the

560 statistical significance (non-significance) of the drifts at all reporting pressures. Fig. 10 also shows that the reporting pressures for some SAT – FP pairs span only very limited ranges (e.g., 100 – 15 hPa for SOF at LIN), while those for other pairs cover much wider intervals (e.g., 196 – 18 hPa for MIR at BLD). The ranges of reporting pressures were typically smaller over tropical FP sites because their tropopause cut-offs were at lower pressures than the extratropical sites.

### 4.2.1 Special case of MLS drifts

565 The data, trend lines and correlation coefficients produced by weighted linear regression fits of all 1146 SAT – FP difference time series were visually checked for consistency and quality. The abnormalities most often revealed were the poorer fits (and lower correlation coefficients) associated with MLS – FP differences over most of the FP sites. Visually, many of the
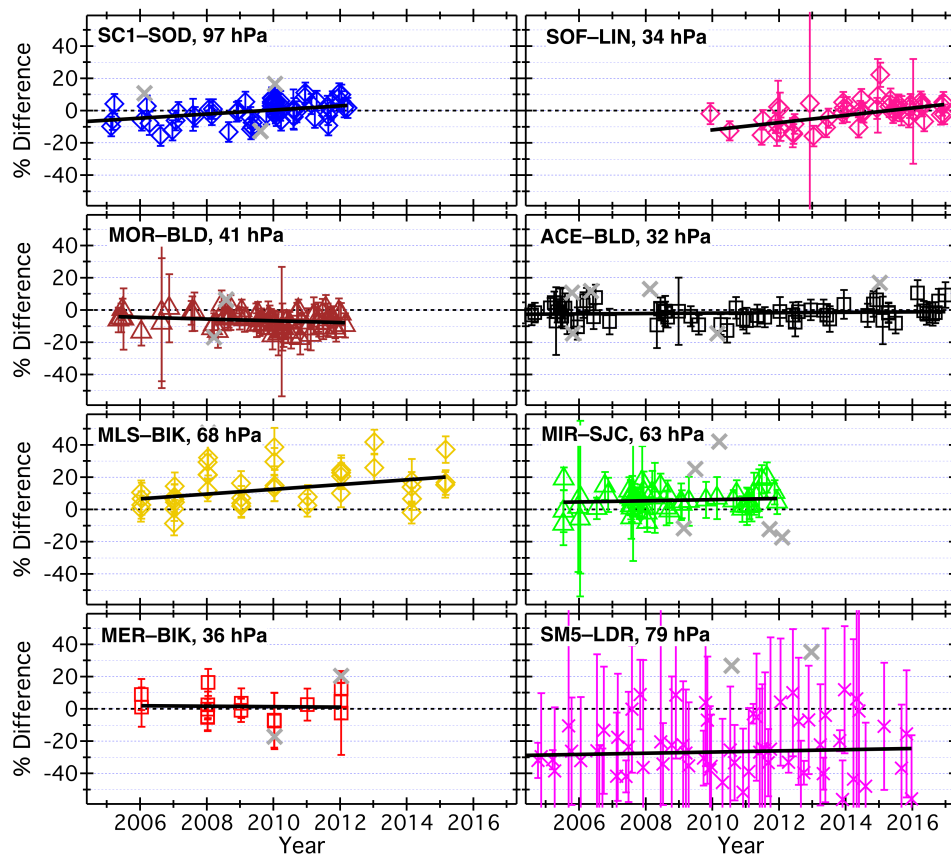
**Figure 9.** Timeseries of SAT-FP relative differences for 8 unique pairs of satellite retrievals and FP sites. See Tables 1 and 2 for the 3-letter codes that represent the relevant FP sites and satellite retrievals. The SAT reporting pressure for the timeseries shown is given in each panel. Vertical error bars depict the uncertainties in SAT-FP differences that factor into the weighted linear regressions used to calculate the black trend lines. Gray crosses are data points identified as outliers and omitted from the analyses.

MLS – FP time series show little or no evidence of drift until ~2010, after which positive trends in the differences become readily apparent (Fig. 11). These positive, post-2010 drifts in MLS – FP differences were previously reported from similar drift evaluations above the BLD, HIL, LIN, LDR and SJC sites (Hurst et al., 2016).

The alternative methodology used here, analogous to that described by Hurst et al. (2016), is a piecewise continuous weighted linear fitting procedure for each time series of relative MLS – FP differences. The fitting algorithm divides each time series into two distinct periods by identifying the point in a record when a statistically significant change in the trend occurred, the "changepoint" as described by Lund and Reeves (2002). The optimal changepoint is the date for which linear fits before and after it yield the smallest root mean square (RMS) of residuals. In the case of MLS – FP differences, the piecewise continuous weighted linear fits substantially improve the "goodness of fit" for each time series (Fig. 11). Compared to the full time series
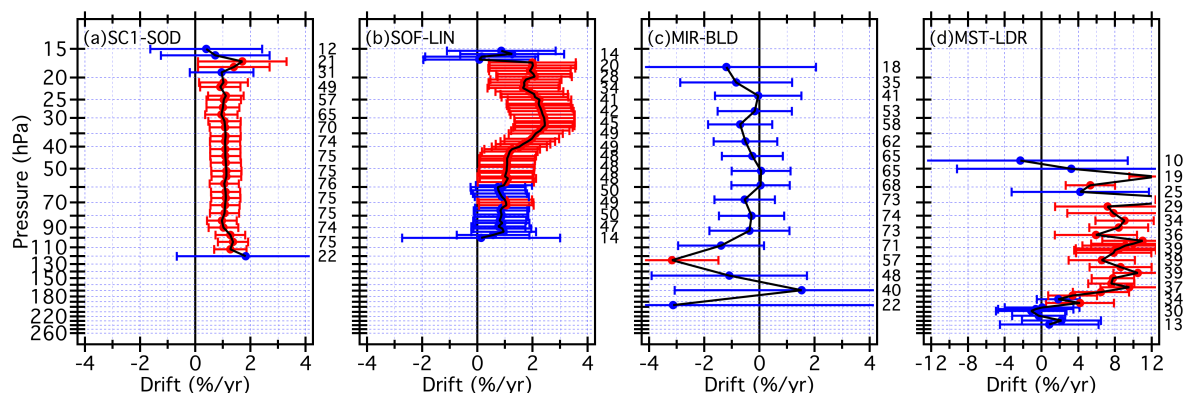
**Figure 10.** Vertical profiles of drifts (filled circles) and their 95 % confidence intervals (horizontal error bars) for 4 different SAT-FP pairs. Blue error bars denote drifts that are not significantly different from zero, while red error bars indicate statistically significant drifts. Numbers in black text to the right of each panel present the number of SAT-FP differences in the timeseries analyzed for drift at the corresponding pressure levels. Note that the x-axis scale is different for panel d.
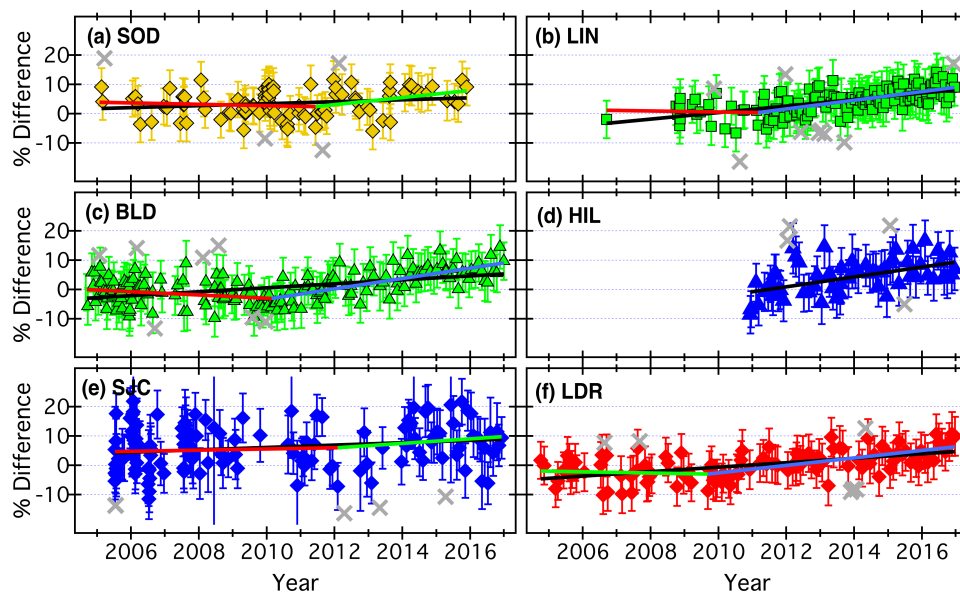
**Figure 11.** Timeseries of relative MLS-FP differences at 68 hPa over 6 different FP sites. As in Fig. 9, vertical error bars represent the uncertainties in differences and gray crosses are data points identified as outliers and omitted from the drift analyses. Black lines depict the trends determined by weighted linear regression fits to the entire records of differences. Colored lines show the linear trends in two distinct time periods (except at HIL) that are separated by a statistically significant changepoint, as determined by piecewise continuous weighted linear regression fits (see section 4.2.1).

regression fits of MLS – FP differences, the two-piece linear fits decrease the RMS of residuals at each of the 5 FP sites by 1 to 8%, with an average reduction of 4%. For the MLS – BIK timeseries, no statistically significant changepoints were identified because the FP record is dominated by data obtained during intensive but short-lived, annual measurement campaigns (Fig. 9).

580 For HIL, the records of MLS – HIL differences lack adequate data before $\approx$ 2010 for the time series to be analysed by this method (Fig. 11). The piecewise continuous fits were therefore performed only on the MLS – FP timeseries at SOD, LIN, BLD, SJC and LDR. The drifts and other statistics reported for the post-changepoint fits to MLS – FP difference time series are denoted by the SAT code MLS*.

MLS retrievals at 31 of 70 pressure levels (44%) in the 121 – 18 hPa range over 7 FP sites exhibited full-record drifts that

585 are positive and statistically significant (Fig. 12). When piecewise continuous linear fits were performed on the same MLS-FP differences, excluding those at HIL and BIK, post-changepoint drifts were positive and statistically significant at 46 of the 52 (88%) reporting levels. The vast majority (90%) of the positive post-changepoint MLS* drifts were stronger than the full record MLS drifts (Fig. 12). For MLS reported pressures from 68 to 22 hPa over all FP sites except HIL and BIK, post-changepoint drifts were an average $\pm\sigma$ of 2.4 $\pm$ 1.0 times stronger than full record drifts. Though the piecewise fits better represent the

590 MLS – FP time series and reduce the RMS of residuals compared to full- record fits, their uncertainties are larger than for the full-record drifts because the post-changepoint records have substantially smaller data populations.
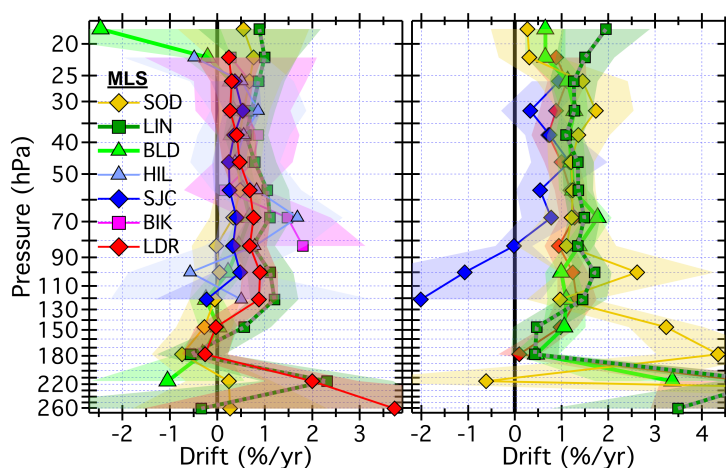
**Figure 12.** Vertical profiles of drifts in (a) MLS-FP differences spanning the entire records (2004-2016) and (b) only those MLS-FP differences after the ~2010 changepoint in each record (i.e., MLS*). The legend applies to both panels, but drifts at HIL and BIK are not available for MLS*. The colored-matched shading surrounding each profile represents the 95 % confidential intervals of the drifts over that site.

## 4.3 Drift profiles for the unique SAT-FP pairs

The drifts determined for all 21 SATs (MLS* included), each paired with 1 to 7 FP sites, are presented as vertical profiles in Fig. 13 and in Figs. B6 and B7. Similarly, Figs. 14 and B8–B9 show the drifts of all paired SATs at each of the 7 FP sites.

595 Both sets of figures are analogous to Fig. 10 except each panel of Fig. 13 shows the drifts of a unique SAT over multiple FP
sites using the coloured symbol scheme from Fig. 1, while each panel of Fig 14 shows the drift profiles of multiple SATs over
each FP site using the coloured symbol scheme from Fig. 2. In some panels of Figs. 13, B6, and B7, the colours and symbols
for some SAT – FP pairs were slightly adjusted to help differentiate between the drift profiles at specific FP sites, for example
LIN and BLD. Another modification to Figs. 13, B6, and B7 is that the 95% confidence intervals are represented by shaded,

600 symmetric envelopes that match the colours of the drift profile markers and connecting lines. For some SATs the drifts were
relatively uniform with pressure while the drifts of others were much more variable. Note that the reporting pressure for some
SATs are the same for each FP site while for other SATs they are not.
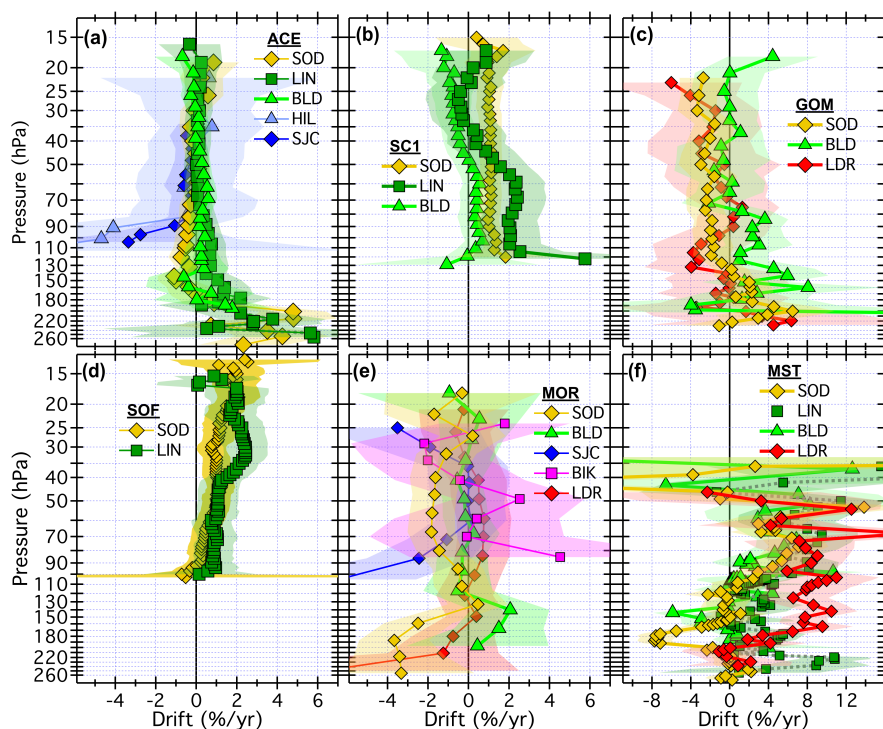
**Figure 13.** Vertical profiles of drifts in SAT-FP differences. Each panel displays the drifts for one SAT paired with 2-5 different FP sites.
Drifts over each FP site (connected colored markers) are presented with their 95 % confidential intervals (colored-matched shading). Shading
that does not cross the black vertical line at 0 % drift indicates drifts that are statistically significant. Note that the x-axis scale for the far right
column is expanded to show the drifts with greater clarity.

There is a general tendency for the uncertainties of drifts to broaden at the extremes of the reporting pressure ranges. This
often occurred because there were sharp drop-offs in data populations of the difference time series at the highest and lowest

605 pressures, and therefore fewer data points for the weighted regression fits. Two factors affecting uncertainties at the high and
low pressure ends of drift profiles are the annual cycles in extratropical tropopause pressures that reduce the data populations
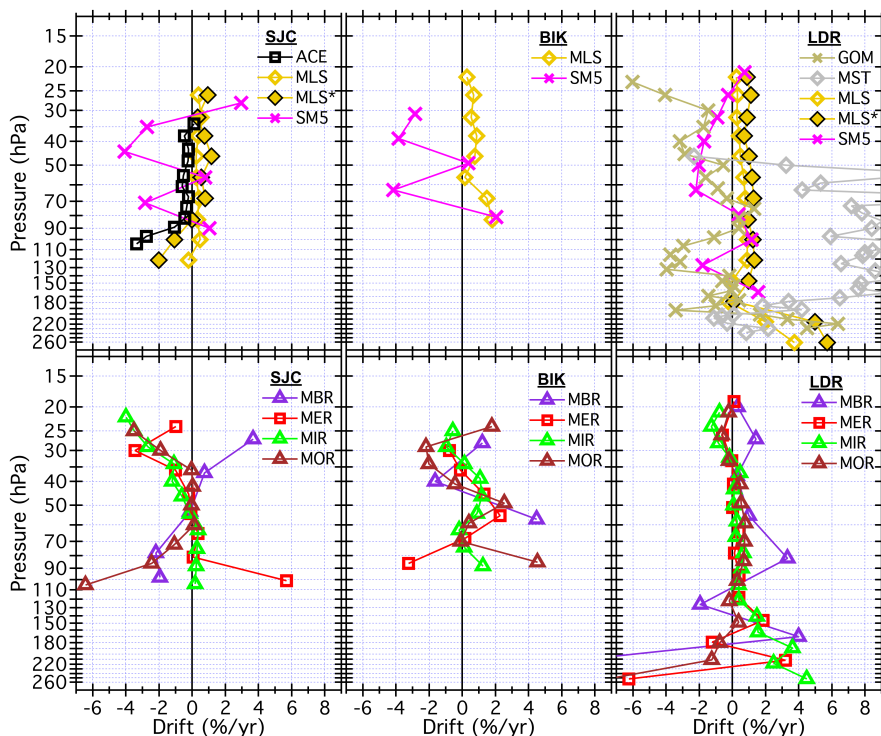
**Figure 14.** Vertical profiles of drifts in SAT–FP differences. Each panel displays the drifts for the multiple SATs paired with each FP site. Profiles for non-MIPAS SATs appear in the top row and for MIPAS SATs appear in the bottom row. Drifts for each SAT are represented by unique coloured markers according to Fig. 2. MLS* refers to drifts in MLS from ∼2010 through 2016, as discussed in Section 4.2.1.

at the highest reporting pressures during summer months, and the altitude ceilings of balloon-based FP profiles that limit data populations at the lowest reporting pressures.

Of the 21 different SATs analysed here, some have drift profiles over multiple FP sites that are statistically equivalent when

610    the drift uncertainties are considered. For example, the ACE drift profiles ≤ 85 hPa are statistically the same over all 5 FP sites (Fig. 13a). Drift profiles for SC1 (Fig. 13b) at LIN and BLD also statistically overlap over their entire pressure ranges, although the drifts ≥ 51 hPa are significant over LIN but not BLD. Drifts for SOF over SOD and LIN (Fig. 13d) also statistically overlap over their entire pressure ranges, with all drifts in the 58–17.2 hPa interval over both sites being positive and significant. Drift profiles for SATs at other FP sites are not always statistically the same, as exemplified by the absence of statistical overlap in

615    SC1 drifts at 10 reporting levels over SOD and BLD in the 45 − 23 hPa interval (Fig. 13b), MST drifts at 6 reporting levels (170–140 hPa) over SOD and BLD (Fig. 13f) and SC4 drifts at 4 levels (35–25 hPa) over SOD, LIN and BLD (Fig. B7d).

Given the vertical profiles of drifts in Figs. 12, 13 and 14, plus those provided in the appendix (Figs. B1–B9), it is evident that a few SATs exhibit statistically significant drifts at many reporting levels over multiple FP sites. Of the 1213 timeseries of SAT-FP differences analyzed here for drift (includes analyses of MLS*), 419 (35 %) of the drifts were statistically significant.

620 Of the 21 SATs examined here, MLS* and SOF had the highest percentages of reporting levels (84 % and 70 %, respectively) with statistically significant drifts over their associaed FP sites. The percentages for 4 other SATs were >40 %: MST (47 %), SC1 (46 %), SC4 (41 %) and MLS (41 %). Overall, these 6 SATs were associated with 339 (81 %) of the 419 statistically significant drifts.

Though the identification of statistically significant drifts is an important result, a more critical metric is the strength of a
625 drift, as this limits the utility of a satellite data set when trying to detect temporal trends in stratospheric water vapor. For this work, a drift of -1 % $yr^{-1}$ ($\approx$-0.5 ppmv decade$^{-1}$ in the middle stratosphere) in a data set would either completely mask a real trend of +0.5 ppmv decade$^{-1}$ or double a real trend of -0.5 ppmv decade$^{-1}$, effectively rendering the data set unusable for detecting a trend of this magnitude. With this in mind, we focus on identifying statistically significant drifts with magnitudes > 1 % $yr^{-1}$, which are hereinafter called "large significant drifts."

630 For many SATs, the numbers of large significant drifts are nearly as great as their numbers for statistically significant drifts. Overall, large significant drifts were identified for 349 (29 %) of the 1213 difference timeseries. Percentages of large significant drifts were again greatest for MLS* (63 %) and SOF (56 %), followed by MST (47 %), SC1 (43 %) and SC4 (32 %). These 5 SATs are associated with 264 (76 %) of the 349 large significant drifts.
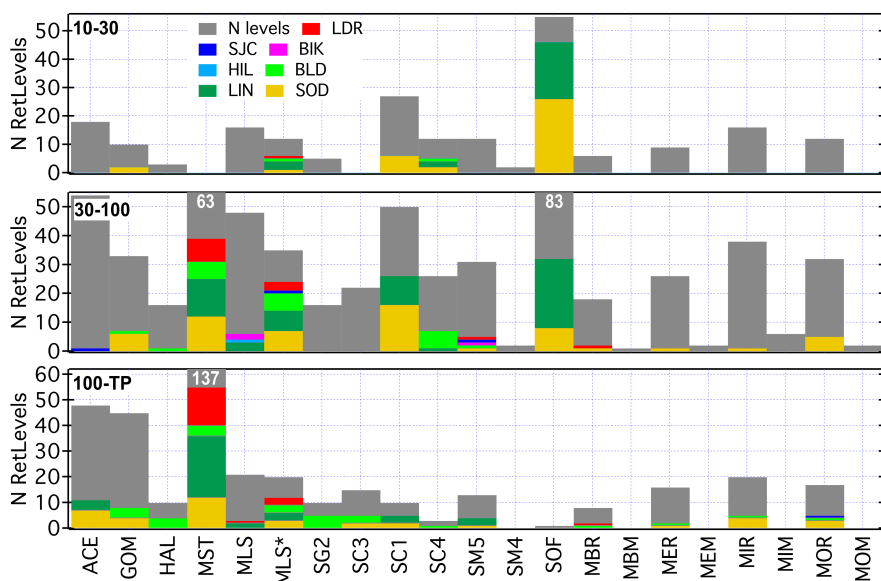
**Figure 15.** Numbers of total reporting levels (gray) and those with large (> 1 % $yr^{-1}$) significant drifts at the paired FP sites (colors) for each SAT. The counts are divided into three distinct pressure intervals: 10-30 hPa, 30–100 hPa, and 100 hPa to the local tropopause (TP) pressure. Colored bars are stacked to show the fractions of reporting levels with large significant drifts across all the FP sites relative to all reporting levels. Values for the three gray bars exceeding the y-axis limits are given in white text at the top of the bar.

The pressure dependence of the drifts of each SAT-FP was investigated by separating them into three pressure intervals:
635 10–30 hPa, 30–100 hPa and from 100 hPa down to the local tropopause pressure (100–TP). For the 21 SATs and 7 FP sites

there were a total of 215, 604 and 394 SAT-FP timeseries in the 10-30, 30–100 and 100-TP pressure ranges (Fig. 15). In these pressure intervals, 76 (35 %), 209 (35 %) and 134 (34 %) of the drifts were statistically significant and 65 (30 %), 157 (26 %) and 127 (32 %) were large and significant. The uniformity of these overall percentages across the three pressure intervals suggests a general lack of pressure dependence of the significant drifts, but these bulk statistics do not provide conclusive
640    evidence for individual SATs.

MLS* drifts in the 100-TP pressure interval were large and significant for 60 % of the reporting levels across its 5 paired FP sites (Fig. 15). In this lowest layer of the stratosphere, 50 % of the SG2 and SC1 drifts and  40 % of the HAL and MST drifts were also large and significant. Interestingly, all large significant SC1 drifts in the 100-TP interval were over SOD and LIN while none were over BLD. For MST the highest percentages of large significant drifts (100-TP) were over LIN (50 %)
645    and LDR (68 %). It is difficult to assess if the large significant drifts of a specific SAT were latitude dependent in any of the pressure intervals because most SATs paired with only 3 or fewer FP sites.

For reporting pressures 30–100 hPa, 39 % to 69 % of the drifts of SOF, SC1, MST, and MLS* (in increasing order) were large and significant. All large and significant SC1 drifts were over SOD and LIN, with none over BLD, the same as for SC1 drifts in the 100-TP interval. The greatest fractions of large significant drifts for MST were again at LIN (81 %) and LDR
650    (73 %), although there was also a high percentage at SOD (67 %). Large significant drift percentages for MLS were highest at LIN (43 %), while for MLS* they were 100 % at SOD and LIN, 86 % at BLD and 43 % at LDR. The fractions of large significant SOF drifts were considerable at LIN (61 %), one of the only two FP sites with enough SOF-coincident soundings to permit statistically robust analyses of drift.

In the 10-30 hPa interval, only two SATs had large significant drifts at ≥50 % of their reporting levels, MLS* (50 %) and SOF
655    (84 %). For MLS*, 100 % and 50 % of drifts in this pressure range were large and significant over LIN and LDR, respectively. Large significant drift percentages for SOF were again high at SOD (84 %) and LIN (83 %). Interestingly, 55 % of the drifts in SC1 retrievals at SOD were large and significant while none were at LIN or BLD.

Some SATs were associated with high percentages of large significant drifts at only one FP site. For GOM drifts at SOD, BLD and LDR, 8 of the 9 large significant drifts at 43 total reporting levels in the 10–30 hPa and 30–100 hPa intervals were
660    at SOD. In the two highest pressure intervals, 8 of 10 large significant drifts in MOR were at SOD, with one each at BLD and SJC and none at BIK or LDR. In the 100-TP interval, 7 of 11 large significant drifts for ACE were at SOD (48 reporting levels over 5 sites) and 4 of 5 large significant drifts for MIR were at SOD (20 levels over 4 sites). In total, 134 (38 %) and 122 (35 %) of 349 significant drifts were identified over SOD and LIN. However, when interpreting these high percentages, it should be considered that 359 (29 %) and 260 (21 %) of all the SAT-FP timeseries analyzed for drift were over SOD and LIN,
665    respectively.

## 4.4   Mean drifts for the unique SAT-FP pairs

Statistics for the drifts at all reporting levels in the same three pressure intervals are presented for each SAT-FP pair in Fig. 16. Each horizontal bar denotes the weighted average of drifts for a specific SAT-FP pair over all reporting levels in the given pressure interval, while the vertical bars depict the 95 % confidence intervals of the weighted means. Weights were calculated
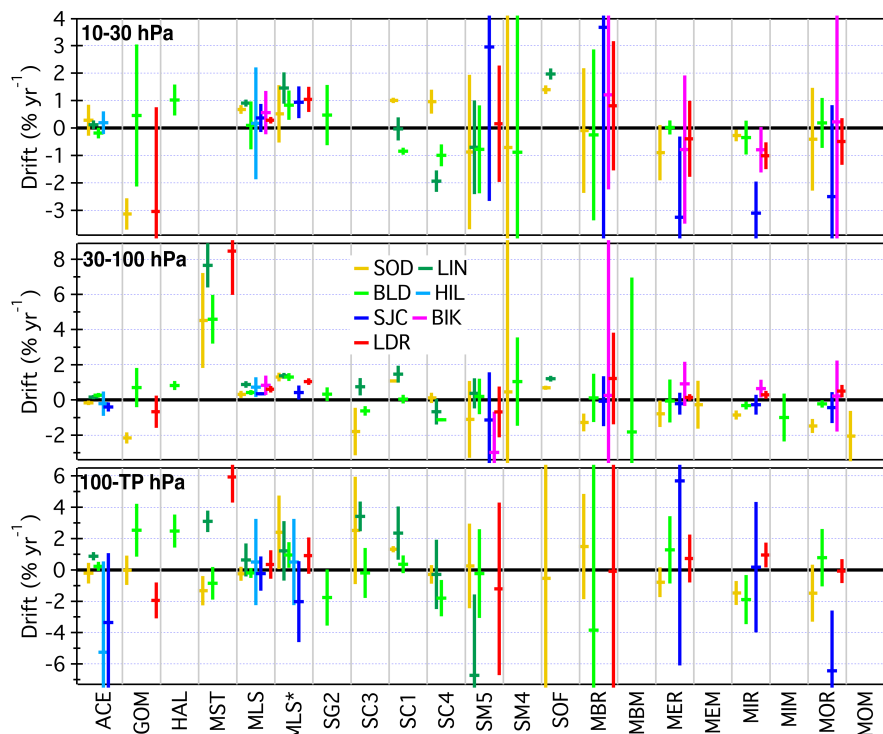
**Figure 16.** Mean drifts (horizontal bars) and their uncertainties (vertical bars) for each SAT-FP pair across all reporting levels within each of the three pressure intervals. The x-axis bins separate results for the 21 different SATs, with color-coded bars presenting statistics for the individual FP sites.

670  using Eq. 10 that scales the standard errors of drifts by the SAT-dependent compensation factors (i.e., grid width vs vertical resolution) that were employed in the bias analyses. Uncertainties of the weighted averages were determined from the weights and the standard errors of the computed drifts using Eq. 11. The mean drift for a given SAT-FP pair is statistically significant if its 95 % confidence interval (vertical bar) does not include zero (horizontal black line at zero drift).

For some SATs there are large (>1 % in magnitude) and statistically significant mean drifts in multiple pressure intervals
675  (Fig. 16). GOM has large negative mean drifts at SOD in both the 10–30 hPa and 30–100 hPa intervals. HAL has positive and significant mean drifts over BLD in all three pressure ranges but only those for 10–30 hPa and 100–TP are >1 % yr$^{-1}$. For MST, there are large significant mean drifts in the 30–100 hPa and 100–TP intervals over LDR, LIN and SOD. These mean drifts at LDR and LIN are positive, while those at SOD are opposite in sign. Large positive mean drifts in SC1 are indicated for all three pressure intervals at SOD and for the 30–100 hPa and 100–TP intervals at LIN. For SC4 there are large negative
680  mean drifts in all three pressure intervals over BLD. None of the 8 MIPAS retrievals have large significant drifts in multiple pressure intervals.

Another way to utilize Fig. 16 is to look for uniformity in the mean drifts of a SAT across all its paired FP sites, in one or more pressure intervals. Consistency in the drifts of a given SAT over multiple FP sites, especially those at widely separated latitudes in different hemispheres, provides evidence that the drifts are not latitude dependent. For example, the mean drifts of MLS* over its 5 paired FP sites, ranging in latitude from 45°S to 67°N are relatively uniform in both the 30–100 and 10-30 hPa intervals (Fig. 16). Mean drifts for ACE, SC4, SM5, MBR, MER and MIR in the 10-30 hPa range are also consistent across their paired FP sites when the uncertainties are considered. Other examples of SATs with consistent mean drifts over wide latitude ranges can be found in Fig. 16.

### 4.5    Mean drifts at each FP site across all SATs

When viewed across each panel of Fig. 16, uniformity in the large significant mean drifts of multiple SATs at a specific FP site (marker color) may be indicative of a drift in the FP timeseries at that site rather than in the SAT timeseries. However, this is conclusive only if the mean drifts for a given FP site are consistent across several different SATs, not just for multiple retrievals of the same satellite instrument (e.g., MIPAS). For example, in the 10-30 hPa interval, there are relatively uniform large negative mean drifts at SJC when paired with 3 different MIPAS retrievals (MER, MIR and MOR), but the mean drifts at SJC for another MIPAS retrieval (MBR) and 2 non-MIPAS instruments (MLS and SM5) are neither negative nor significant. Similarly, in the 30–100 hPa interval, mean drifts for SOD paired with 7 SATs are negative and statistically significant, but 5 of those 7 drifts involve MIPAS retrievals and the mean drifts for 10 other SATs at SOD are either positive or not significantly different from zero.

This interpretation of the mean drifts shown in Fig. 16 is supported by the visible inconsistencies between the drift profiles of the various SATs paired with each FP site (Figs. 14 and B8–B9). For example, the drift profiles of GOM, MST, MLS and SM5 at LDR (Fig. 14c) are all notably different. Similar inconsistencies between different SATs are present at other FP sites for both the non-MIPAS and MIPAS SATs. Preliminary calculations of mean drift profiles across all SATs, non-MIPAS and MIPAS SATs included, at a specific FP site, revealed that their large uncertainties rarely produce mean drifts with statistical significance. This is not surprising given the uniqueness of each SAT instrument and its retrieval methods for water vapor measurements, so no further analyses of the multiple SAT mean drifts at each FP site were performed. Such an analysis could be performed on a subset of the most stable SATs, but that would be somewhat subjective and beyond the scope of this paper.

### 4.6    Mean drifts for each SAT across all FP sites

Mean drifts for each SAT across all FP sites paired with it were calculated in the three pressure intervals defined above, employing the same method used to calculate the mean drifts of each unique SAT-FP pair (Fig. 16). Specifically, weighted means of drifts were computed using all the drifts in the appropriate ranges of reporting pressures, not by averaging the mean drifts at different FP sites. Averaging weights were based on the standard errors of drifts and the same SAT-dependent compensation factors employed in the site-specific mean drift computations. Uncertainties of the mean drifts of SATs across their paired FP sites were calculated from the averaging weights and standard errors of drifts using Eq. 11. Mean drifts for each SAT in the three pressure intervals are presented in Fig. 17.
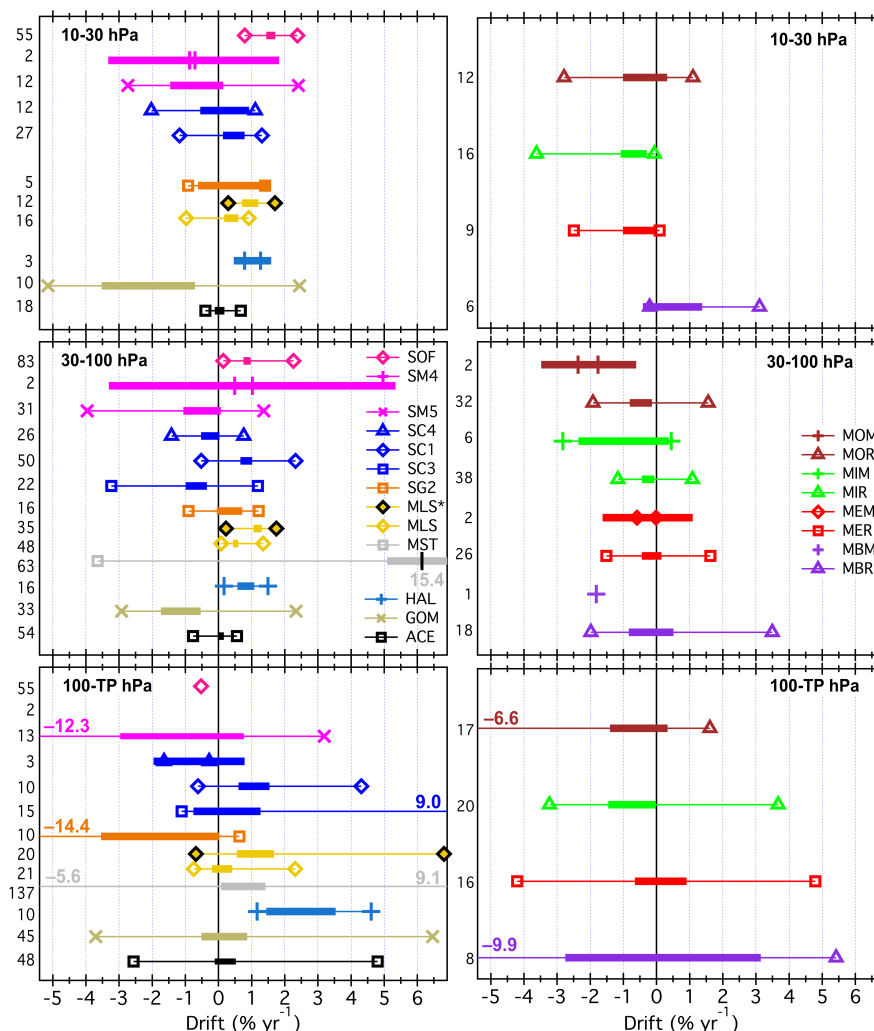
**Figure 17.** Drift statistics for each SAT across all paired FP sites, separated into the 3 pressure intervals (analogous to Fig. 8 for biases). Left and right panels are for the SATs of non-MIPAS and MIPAS instruments, respectively. Mean drifts are the values at the left-to-right centers of the 95 % confidence intervals (thick horizontal bars). Markers connected by thin horizontal lines show the 5th and 95th percentiles of the drifts for each SAT. Values of 5th and 95th percentiles that exceed the x-axis limits are enumerated as colored text. In the left 30–100 hPa panel, a small black vertical bar denotes the mean drift for MST. Only a mean drift value is presented for SOF (100–TP) and MBM (30–100) because only one retrieval pressure for each provided timeseries of differences met the criteria for drift analysis. The numbers to the left of each panel indicate the number of retrieval levels with difference timeseries that determined the statistics presented.

715     Fig. 17 is analogous to Fig. 8 except it presents statistics for drifts instead of biases. Here, each thick horizontal bar represents the 95 % confidence interval of the mean drift. Since these confidence intervals are symmetric about the mean, the value at the center of each thick horizontal bar is the mean drift. Thin horizontal lines show the inter-90 % range of drifts, with markers

denoting the 5$^{th}$ and 95$^{th}$ percentiles. In the text below, the terms "variability" and "range" are used when discussing the 95 % confidence intervals (e.g., variability of $\pm x$ % yr$^{-1}$) and inter-90 % ranges (e.g., range of y to z % yr$^{-1}$).

720 ### 4.7 Synopsis of the drift assessments

A brief statistical synopsis of the drifts in the three pressure intervals for each SAT is now presented, including relevant information about the fractions of reporting levels with large and statistically significant drifts, and where possible, statements about the uniformity, signs and magnitudes of drifts in specific pressure intervals. In most cases the information can be visually confirmed in one of more of Figs. 15, 16 and 17. Drifts for the 12 non-MIPAS retrievals are discussed first, then drifts for the

725 8 different MIPAS retrievals.

**ACE**

Drifts for ACE were large and statistically significant at 12 of the 120 (10 %) reporting levels over its 5 paired FP sites, of which 11 were at pressures >100 hPa over SOD and LIN. However, in this pressure range, only the mean drift at LIN was significant (Fig. 16). In the 30–100 hPa interval, only one drift at 54 reporting levels was large and significant over SOD, LIN,

730 BLD, HIL and SJC. Across all paired FP sites, ACE drifts spanned a wide range in the 100–TP interval, but the variability was much smaller and the mean drift was not statistically different from zero (Fig. 17). In the two pressure intervals ≤100 hPa, both the variability and ranges of drifts across all sites were small and the mean drifts were not statistically significant.

**GOM (GOMOS)**

Of the 16 large and statistically significant drifts at 88 GOM reporting levels over its 3 FP sites (SOD, BLD, LDR), 12 were over

735 SOD and 4 were over BLD. There were significant negative mean drifts in the 10-30 and 30–100 hPa intervals over SOD (Fig. 16) because drifts at all 13 reporting pressures were negative and 8 were large and significant. Of the 11 reporting levels over BLD in the 100-TP interval, 9 were uniformly positive, of which 3 were large and significant, yielding a significant positive mean drift. Several large negative drifts in the 100-TP range over LDR resulted in a significant negative mean drift. Mean drifts across all 3 sites were negative and significant in the 30–100 and 10-30 hPa intervals, but not in the 100-TP interval (Fig. 17).

740 **HAL (HALOE)**

Measurements by HAL ceased in mid-November 2005, so its record has a >5 yr overlap with only the FP record at BLD. The HAL instrument began measurements in 1991, but only the last 5.9 yr of its record (since Jan 2000) are analyzed here since the first WAVAS report covered HAL data through 1999 (Kley et al., 2000). Though large and significant drifts were determined for only 5 of 29 reporting levels, all 29 drifts were positive and relatively uniform. This uniformity resulted in positive and

745 significant mean drifts of of 0.8±0.3 % yr$^{-1}$ to 2.5±1.0 % yr$^{-1}$ over BLD in all three pressure intervals.

**MST (MAESTRO)**

The MST data analyzed here cover 200 reporting pressures over 4 FP sites, but only as high as 33 hPa. Large significant drifts were determined for 94 (47 %) reporting levels, with 39 and 55 in the 30–100 and 100-TP intervals, respectively. MST time

series were not available at pressures ≤30 hPa for drift analysis. Drifts were large and significant at 23 of 33 (70 %) reporting levels over LDR (Fig. 10d) and 37 of 64 (58 %) reporting levels over LIN (Fig. B3h). Between these two sites, 94 % of all MST drifts were positive, producing large significant mean drifts at LIN and LDR in the 100–TP and 30–100 hPa ranges (Fig. 16). Drifts over SOD and BLD were large and significant at 40 % and 23 % of all reporting levels (Figs. B3d,l), with 17 of 18 large significant drifts being positive in the 30–100 hPa range over both sites and 12 of 16 large significant drifts being negative in the 100-TP interval. Consequently, over SOD and BLD there were significant positive mean drifts in the 30–100 hPa interval and negative mean drifts in the 100-TP interval, of which the SOD mean drift was significant. Mean MST drifts across all 4 FP sites were positive and significant in both the 30–100 and 100-TP intervals (Fig. 17).

**MLS and MLS***

Drifts were calculated separately for the full MLS records (2004-2016) at all 7 FP sites and for shorter records from their changepoint dates (∼2010) through 2016 (represented by MLS*) at all sites except HIL and BIK. Across their paired FP sites, drifts for full MLS records and MLS* were large and significant at 9 (11 %) of 85 reporting levels and 42 (63 %) of 67 reporting levels, respectively. In the 30–100 hPa range, the mean drifts for MLS and MLS* across their paired FP sites were positive and significant, as were their mean drifts in the 10-30 hPa interval (Fig. 17). In the 100-TP range, mean drifts for MLS and MLS* across their paired FP sites were weakly positive and not significant, and strongly positive and significant, respectively. Figs. 16 and 17 show not only the uniformity of the MLS and MLS* mean drifts across their associated FP sites, but also how the MLS drifts in the 10-30 and 30–100 hPa intervals increase by 120-140 % when they are determined for the shorter post-changepoint records (MLS*) instead of the full records (MLS).

**SG2 (SAGE II)**

As was the case for HAL, SG2 measurements began in the mid-1980s and ended in 2005, so only the SG2-BLD timeseries spanning 2000-2005 are analyzed here for drift. Large significant negative drifts were found for 5 of the 10 reporting levels in the 100-TP interval (Fig. B3m), but the mean drift was not significant. None of the drifts at pressures ≤100 hPa were significant, and neither were the mean drifts for the 30–100 hPa and 10-30 hPa intervals (Fig. 17).

**SC3 (SCIAMACHY limb)**

Drifts were large and significant at only 5 of 37 reporting levels over SOD, LIN and BLD, and all 5 were in the 100-TP interval (Figs. B3b,f,j). However, relatively uniform positive drifts at 11 of 12 reporting levels in the 100-TP and 30–100 hPa intervals over LIN produced significant positive mean drifts (Fig. 16). SC3 timeseries were not available at pressures ≤30 hPa for drift analysis. Interestingly, drifts over SOD and BLD in the 30–100 hPa range were negative at 14 of 15 reporting levels, resulting in significant negative mean drifts. Across the three FP sites, there was a significant negative mean drift in the 30–100 hPa interval.

**SC1 (SCIAMACHY solar OEM)**

Drifts over SOD at 24 of 31 (77 %) reporting levels were uniform, positive, large and significant (Fig. 10a), as were the mean

drifts over SOD for all three pressure intervals. Drifts at all 13 SC1 reporting levels >50 hPa over LIN (Fig. B3e) were also uniform, positive, large and significant, as were the mean drifts in the 100-TP and 10-30 hPa intervals. Over BLD, none of the drifts at all 28 reporting levels were large and significant, yet the uniformly negative drifts at the 8 levels ≤30 hPa produced a significant negative mean drift in the 10-30 hPa range. SC1 mean drifts across the 3 sites were positive and significant for all three pressure intervals (Fig. 17).

**SC4 (SCIAMACHY solar OP)**

Drifts over BLD were uniformly negative at all 13 reporting pressures and large and significant at 10 levels (Fig. B3k), producing significant negative mean drifts for all three pressure intervals (Fig. 16). Mean drifts in the 10-30 hPa range over LIN and SOD were also significant, but of opposite sign. Mean drifts in the 100-TP range were based on only one reporting pressure at each site, and only the negative drift at 115 hPa over BLD was significant. The mean drift across all three sites was statistically different from zero only for the 30–100 hPa interval.

**SM5 (SMR 544 GHz)**

Only 9 of the drifts at 56 reporting levels across 6 paired FP sites were large and significant, and these were dispersed across the 6 sites. At LIN, 3 large significant negative drifts in the 100-TP interval produced a large significant negative mean drift. Mean drifts at each FP site were relatively uniform in all three pressure intervals, but none of the mean drifts across the FP sites were significant.

**SM4 (SMR 489 GHz)**

Drifts were determined at only 4 SM4 reporting pressures over 2 FP sites, SOD (23 and 35 hPa) and BLD (21 and 33 hPa). No drifts were statistically significant. With only one reporting pressure per site in each of the 10-30 and 30–100 hPa pressure ranges, the two mean drifts for each site have large uncertainties and are therefore not significant. Similarly, mean drifts across the two FP sites were highly uncertain and not statistically significant (Fig. 17).

**SOF (SOFIE)**

Vertical profiles of drifts at SOD and LIN were similar (Figs. B3n,o), with large significant positive drifts at 76 of the 100 combined reporting pressures between 14 and 60 hPa. Drifts in the 30–100 hPa and 10-30 hPa intervals over each site were fairly uniform and positive, producing significant positive mean drifts for SOD and LIN (Fig. 16). Over the two sites, there was only one reporting pressure in the 100-TP range, and this singular drift at 103 hPa over SOD was highly uncertain and not significant. Therefore, mean drifts across the 2 FP sites were not significant in the 100-TP interval but positive and significant in both the 30–100 and 10-30 hPa intervals.

**MBR and MBM (MIPAS Bologna V5R NOM and MA)**

Drifts of MBR, large and significant at only 4 of 32 reporting levels over 5 FP sites, were dispersed across 3 sites (Fig. B4). Only the mean drift at SOD in the 30–100 hPa interval was significant, the result of uniform negative drifts at all 5 reporting levels (Fig. B4a). One large, significant negative drift in the 100-TP interval over each of BLD and LDR substantially increased

the variability and range of drifts over all sites (Fig. 17). For MBM, the drift was not significant at 43 hPa over BLD, the only MBM timeseries analyzed for drift, so the resulting mean drift was highly uncertain and not significant.

### MER and MEM (MIPAS ESA V7R NOM and MA)

There were 3 large and significant drifts at 51 MER reporting levels scattered over 5 sites (Fig. B5). In general, drifts were not uniformly positive or negative in any of the pressure intervals over any site except for the 2 reporting levels ≤30 hPa above SJC, and for 4 of 5 reporting levels in the 30–100 hPa interval above SOD. Significant negative mean biases were determined for both. The variability of MER drifts across all sites was small in each pressure interval. Drift results for MEM were available for only 2 reporting pressures over SOD, 34 and 52 hPa. Neither drift was significant or resulted in a significant mean drift for the 30–100 hPa interval.

### MIR and MIM (MIPAS IMK/IAA V5R NOM and MA)

There were 6 large and significant drifts at the 74 MIR reporting levels over 5 FP sites, and all were negative (Fig. B4). Five were over SOD, of which 4 were in the 100-TP interval, producing a significant negative mean drift. In the same pressure range, there were fairly uniform positive drifts at all 7 reporting levels over LDR and negative drifts at 4 of 5 reporting pressures over BLD, including one large, significant negative drift, and these produced significant mean drifts at both sites (Fig. 16). In the 30–100 hPa range, mean drifts over SOD, BLD, BIK and LDR were significant due to relatively consistent but rarely significant negative (SOD, BLD) and positive (BIK, LDR) drifts. Similarly, consistent negative drifts at the 2-4 MIR reporting levels in the 10-30 hPa range over each of SOD, SJC and LDR resulted in significant negative mean drifts. Across the 5 FP sites, mean drifts for MIR were negative and significant in all three pressure intervals (Fig. 17). For MIM, only six timeseries of differences in the 30–100 hPa range over BLD were analyzed. None of the 6 drifts or the mean drift for 30–100 hPa were statistically significant.

### MOR and MOM (MIPAS Oxford V5R NOM and MA)

Drifts in MOR retrievals were large and significant at 10 of 61 levels over 5 FP sites, of which 8 were negative and over SOD (Fig B5). Five of the 8 significant negative drifts over SOD were in the 30–100 hPa range, producing a significant negative mean drift (Fig. 16). Consistent positive drifts at 5 of 6 reporting levels in the same pressure interval over LDR resulted in a positive significant mean drift. At SJC, there was a large significant negative drift at the lone reporting pressure >100 hPa, so the mean drift was significant and negative (Fig. 16). In the 10-30 hPa pressure interval, none of the individual or mean drifts over the 5 FP sites were significant, so neither was the mean drift across all 5 sites. There was a significant negative mean drift at SOD and across all 5 FP sites in the 30–100 hPa pressure interval due to the 5 large and significant negative drifts over SOD. For the 2 MOM reporting pressures in the 30–100 hPa range over SOD, neither of the negative drifts were significant, but they were large and consistent enough to produce a significant negative mean drift.

## 5 Summary and conclusions

We have compared satellite data records on stratospheric water vapour recorded since 2000 to FP profiles from 27 stations spanning a wide range of latitudes (45° S to 79° N). For the comparison, we applied the same approach to all satellite data

845 products. In particular, we applied a priori and averaging kernels of the satellite data to the FP profiles in order to adjust them to the retrieval characteristics of each satellite data set. If averaging kernels were not available we smoothed the FP profiles with ad hoc Gauss-shaped smoothing kernels to account for the vertical resolution of the satellite instruments. Two consistent sets of collocation criteria were utilised based on two classes of instruments: for the dense samplers we used a time difference < 24 hours, a distance < 1000 km, and a latitudinal difference < 5°, while for occultation instruments we used 7x24, 2000 km,

850 and latitudinal difference < 15°.

We determined the profiles of the biases and drifts and their uncertainties (in terms of the standard error of the mean, SEM, and of the linear regressions fits) of every instrument versus each FP station by averaging over all available collocations. By analysing the bias and drift profiles of one satellite instrument across all hygrometer stations we obtained insight in the general behaviour of the satellite instruments, including general information indicating an absence of obvious latitudinal dependencies

855 of their biases and drifts. Similarly, a comparison of all satellite instruments to each specific FP station provided some insight into any peculiarities of the FP record at each station. We have concentrated on the satellite data analysis and have, as a final synopsis of the comparisons, averaged the biases and drifts of each individual satellite instrument over all FP stations within three different pressure ranges, namely 10 to 30 hPa, 30 to 100 hPa, and 100 hPa to the local tropopause level.

In the 10 to 30 hPa range, most satellite data have a relative bias within the ± 10 % range versus the mean of all FP data.

860 Exceptions are GOM, ILA, HIR (however, the latter two are close to -10 %), MST (which does nominally not measure in this altitude range), SLA, SLB, and SM5 (nominally restricted to the UTLS). Among the well-performing instruments, ACE, MBR, MBM, MER, MIM, MOR, MOM, MLS, SG2, SC1, SC4, and SOFIE have biases within the ± 3 % range.

The 30 to 100 hPa range is where most instruments perform best. Only GOM, MST, MEH, MEM, MOM, POM, SLA, and SM5 show biases larger than 10% (while GOM and POM are close to ± 10 %, and MST does nominally not cover this altitude

865 range). The least biased water vapour data records in this altitude range are ACE, HIR, MLS, SG3, SC4, SOF, MBH, MBR, MER, MIH, and MIR, again with biases within the ± 3 % range. SG2 and MOR just miss the ± 3 % mark.

The situation is worse in the 100 hPa to tropopause range. Here, most instruments have larger biases. Nevertheless, the following SATs have biases within the ± 10 % range even in this altitude range: MLS, ACE, SC4, SG2, SG3, GOM, POM, MOR, MOH, MBR, MBH, MER, MIR, and MIH. The biases are significant for almost all data sets in the three altitude ranges

870 in the sense that the range $\pm 2\sigma_b$, i.e. twice their SEM (thick horizontal bars in Fig. 8), around the bias does not include zero. The large numbers of collocations that were available in most cases result in this high proportion of bias significance. The 5%- and 95%-percentiles, however, are very wide in most cases, which indicates a large spread in the individual bias profiles.

Of the 1213 time series of relative differences between 21 SATs and 7 FP sites that were analysed for drifts, 419 (35 %) had statistically significant drifts at the 95 % level of confidence. Of these 419 significant drifts, 349 (83 %) were also large drifts,

875 with magnitudes > 1 % yr$^{-1}$. Five SATs were together associated with 76 % of the large significant drifts: MLS*, SOF, MST,

SC1, and SC4. Eleven SATs had large significant drifts at 15 % or less of their reporting levels: ACE, MLS, SC3, SM4, MBR, MBM, MER, MEM, MIR, MIM, and MOM.

In the 10–30 hPa pressure interval there were 8 SATs with statistically significant mean drifts across all of their associated FP sites, but of these, only GOM, HAL and SOF had large significant mean drifts. Of the SATs that reported data in this pressure range, ACE, HAL, MLS, SG2, SM5, SM4, MBR, MER, MIR and MOR had no large significant drifts at any of their reporting levels.

All 21 SATs provided data for drift analysis at one or more pressure levels in the 30–100 hPa interval over at least one FP site. Overall, most SATs had their smallest mean drifts in this pressure range compared to those above and below. Mean drifts were significant for 11 SATs in this pressure range, but only GOM, MST, MLS* and MOM had large significant mean drifts.

Mean drifts in the 100–TP pressure interval were almost always associated with larger uncertainties than those for mean drifts in the two pressure ranges above. Consequently, only 34 % of the drifts in this interval were significant, but 95 % of these had magnitudes > 1 %. Overall, only 5 SATs had significant mean drifts and 3 of these (HAL, MLS*, SC1) were large significant mean drifts.

Only 5 SATs that reported in all three pressure intervals had no significant mean drifts in any of the intervals: ACE, SG2, SC4, SM5 and MBR. Together, only 18 % of their drifts at all reporting levels were significant, but of these, 84 % of the drift magnitudes were > 1 % yr$^{-1}$.

In summary, these assessments of mean biases and drifts against FP profiles from a widespread, worldwide array of FP sites demonstrate that the satellite data records are generally an extremely valuable source of information on atmospheric water vapour abundance from the tropopause to about 30 km altitude (10 hPa pressure). For this altitude range, most SAT data records have biases <10 % and drifts < 1 % yr$^{-1}$ relative to FPs that are considered the most accurate and best-characterized instruments for the measurement of stratospheric water vapour. Satellite instruments with biases below 10% over the complete altitude range analysed here are ACE, MLS, SG2, SG3, SC4, MBR, MER, MIH, MIR, MOH, and MOR. SATs with mean drifts < 1 % yr$^{-1}$ in at least two of the three pressure intervals were ACE, MLS, SG2, SC3, SC1, SC4, SM5, MBR, MER, MIR and MOR. Four SATs had mean drifts with magnitudes > 2 % yr$^{-1}$ in one pressure interval that makes their measurement time series unsuitable for the detection of stratospheric water vapor trends < 20 % decade$^{-1}$.

*Data availability.* The satellite data records used in this study are publicly available from https://publikationen.bibliothek.kit.edu/1000093970 and have the doi: 10.5445/IR/1000093970.

# Appendix A: Bias related plots

## A1    Biases per SAT dataset
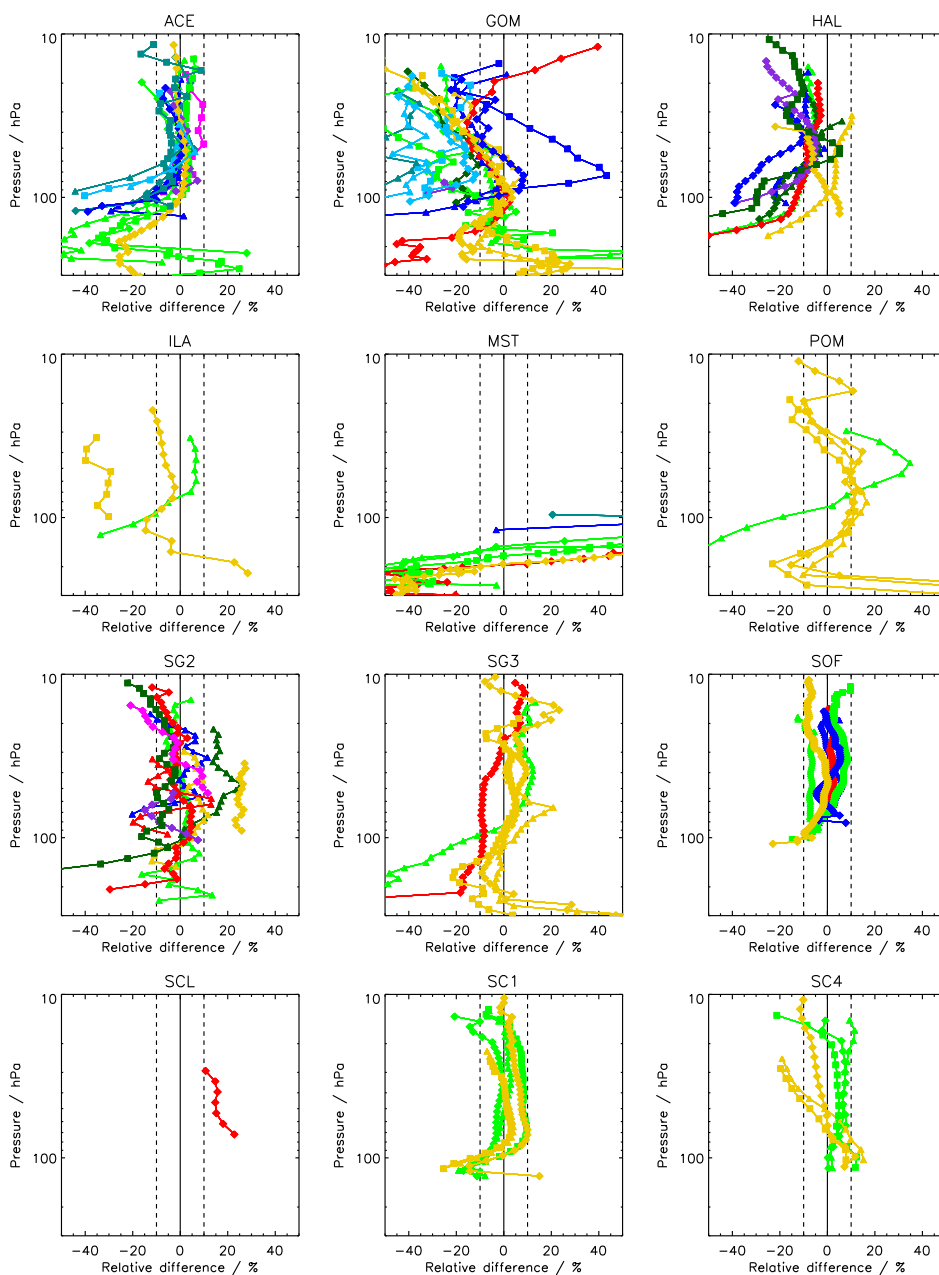
**Figure A1.** Mean relative differences between coincident SAT (no MIPAS) and FP records. Color coding of FP data according to Fig. 1.
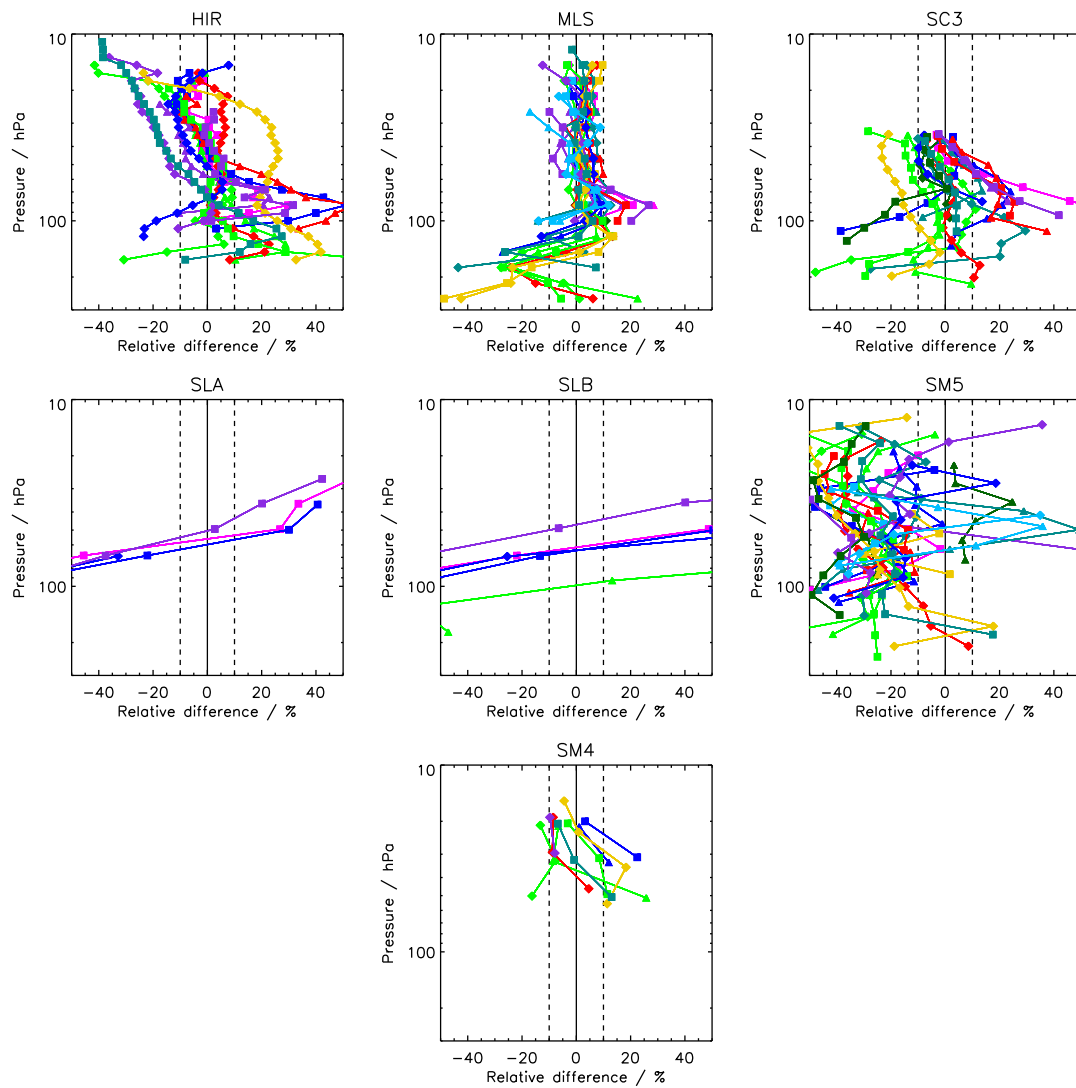
**Figure A2.** Same as Fig. A1 but for dense samplers (no MIPAS).
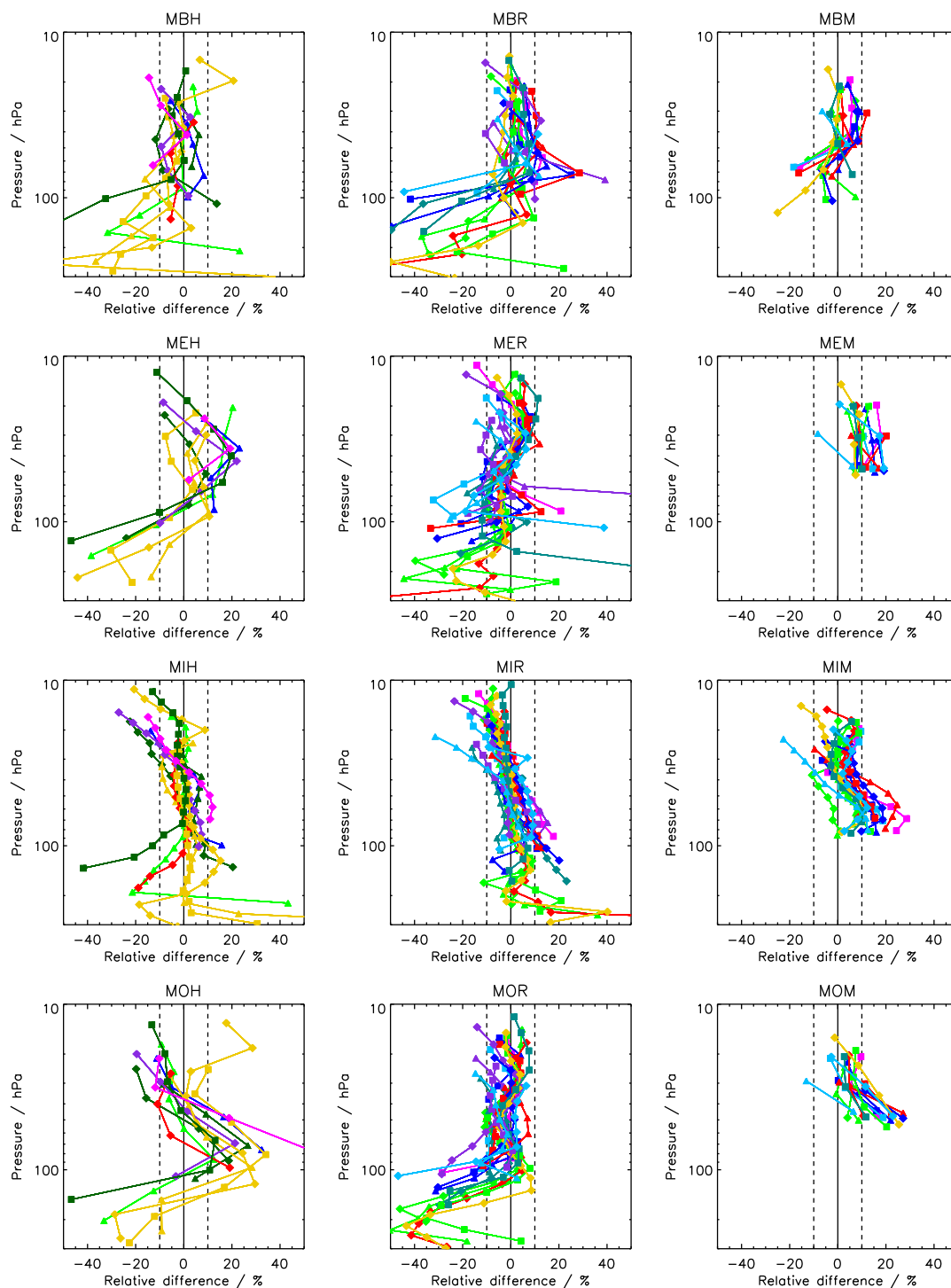
**Figure A3.** Same as Fig. A1 but MIPAS records only.

905  **A2   Biases per FP dataset, ordered by latitude**
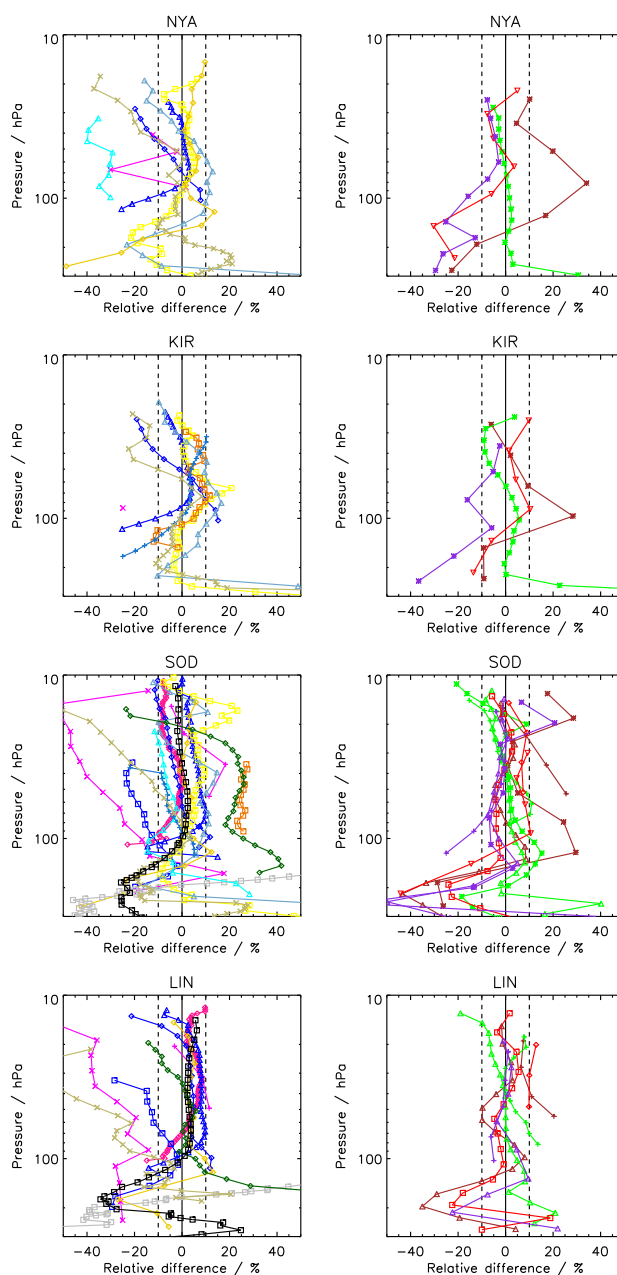
**Figure A4.** Mean relative differences between the FP stations NYA, KIR, SOD, LIN and coincident SAT records (left: data records except MIPAS; right: only MIPAS data records). Color coding of the SAT data according to Fig. 2.
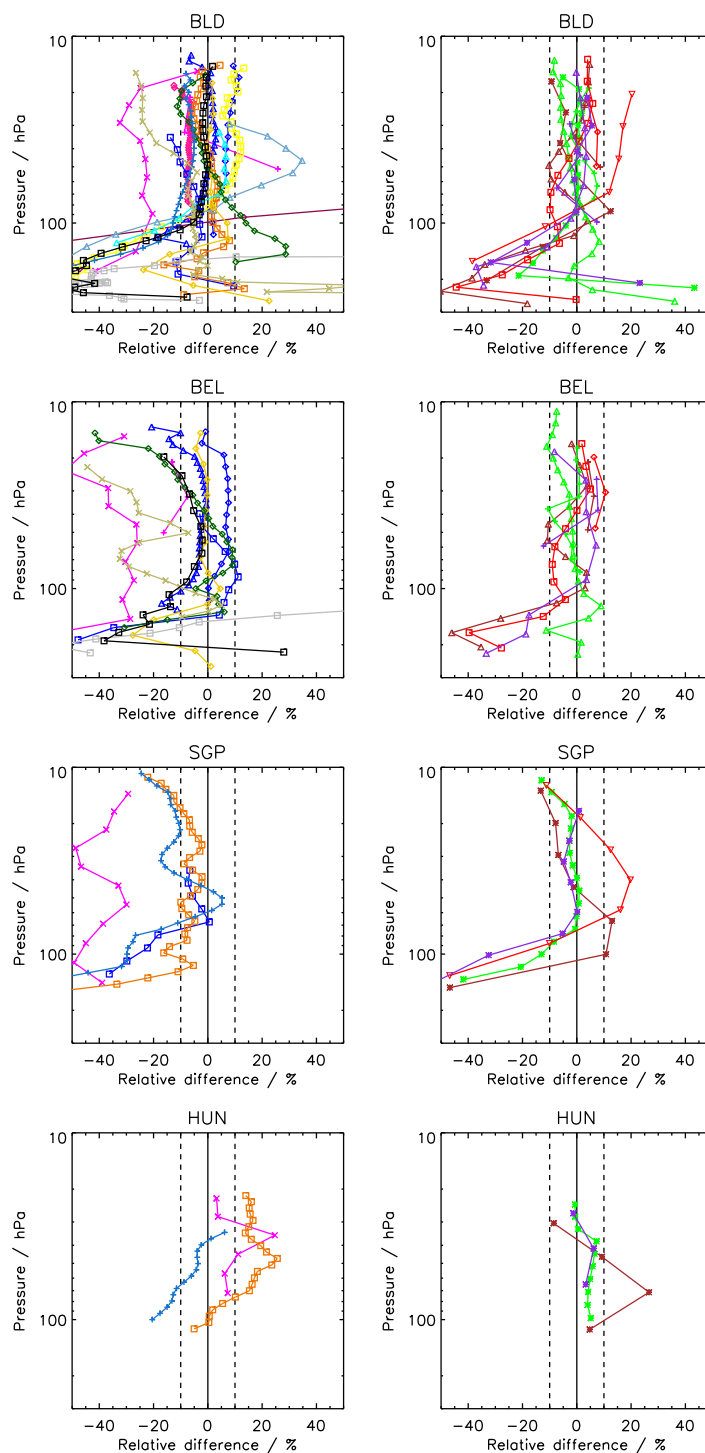
**Figure A5.** Same as Fig. A4 but for BLD, BEL, SGP, and HUN.
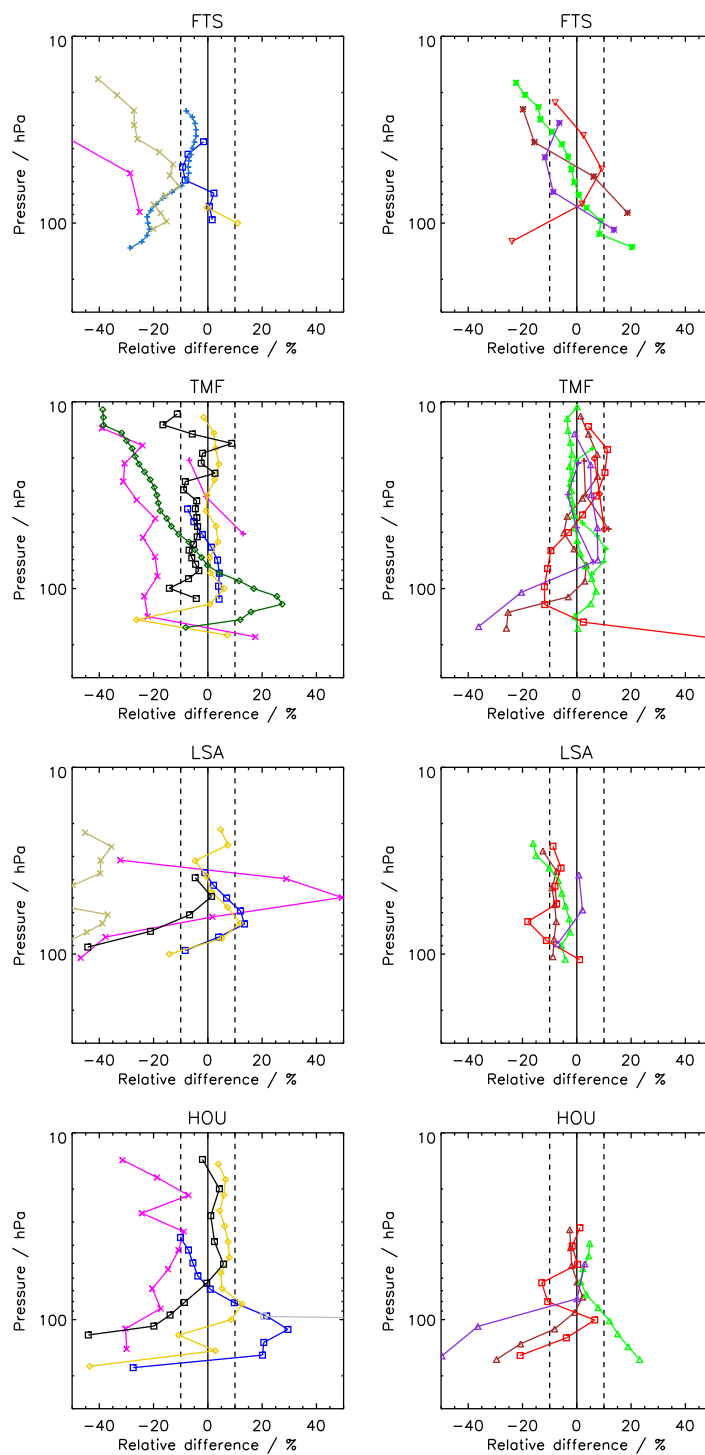
**Figure A6.** Same as Fig. A4 but for FTS, TMF, LSA, and HOU.
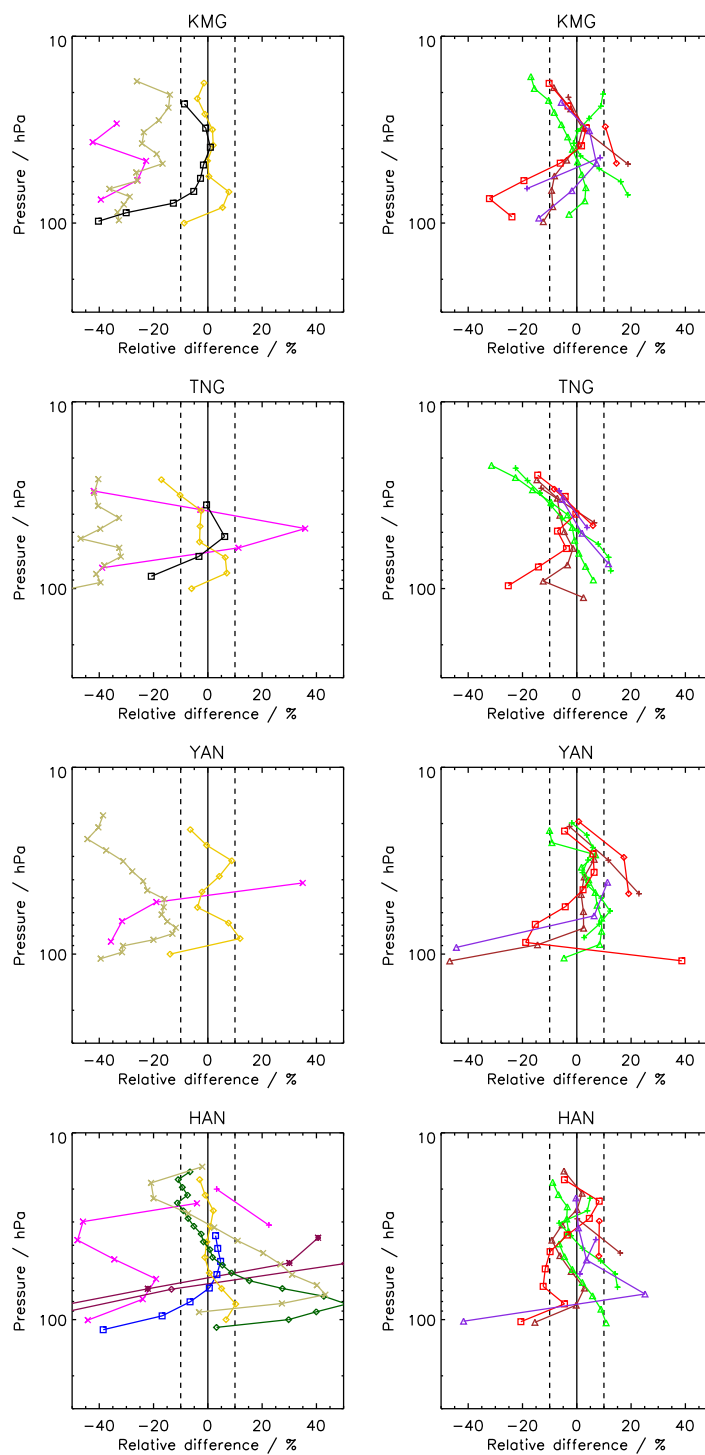
**Figure A7.** Same as Fig. A4 but for KMG, TNG, YAN, and HAN.
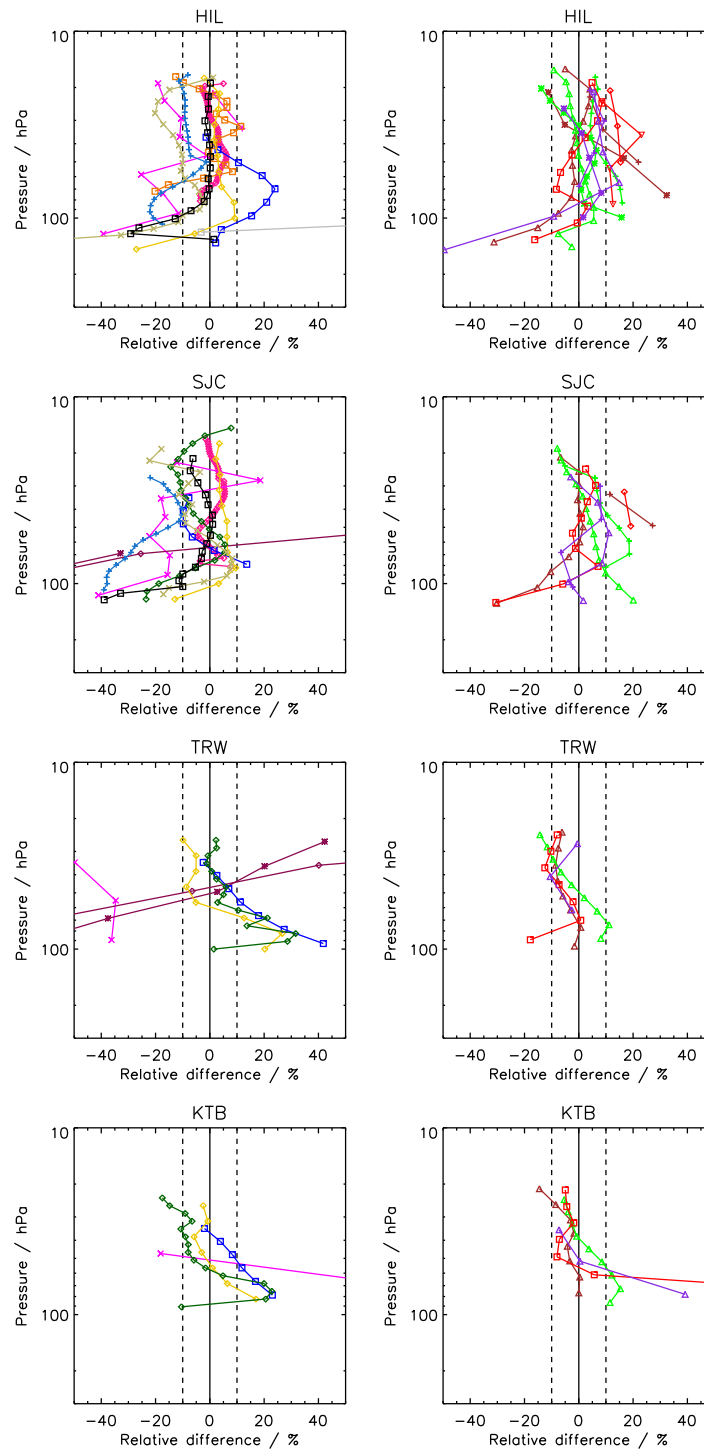
**Figure A8.** Same as Fig. A4 but for HIL, SJC, TRW, and KTB.
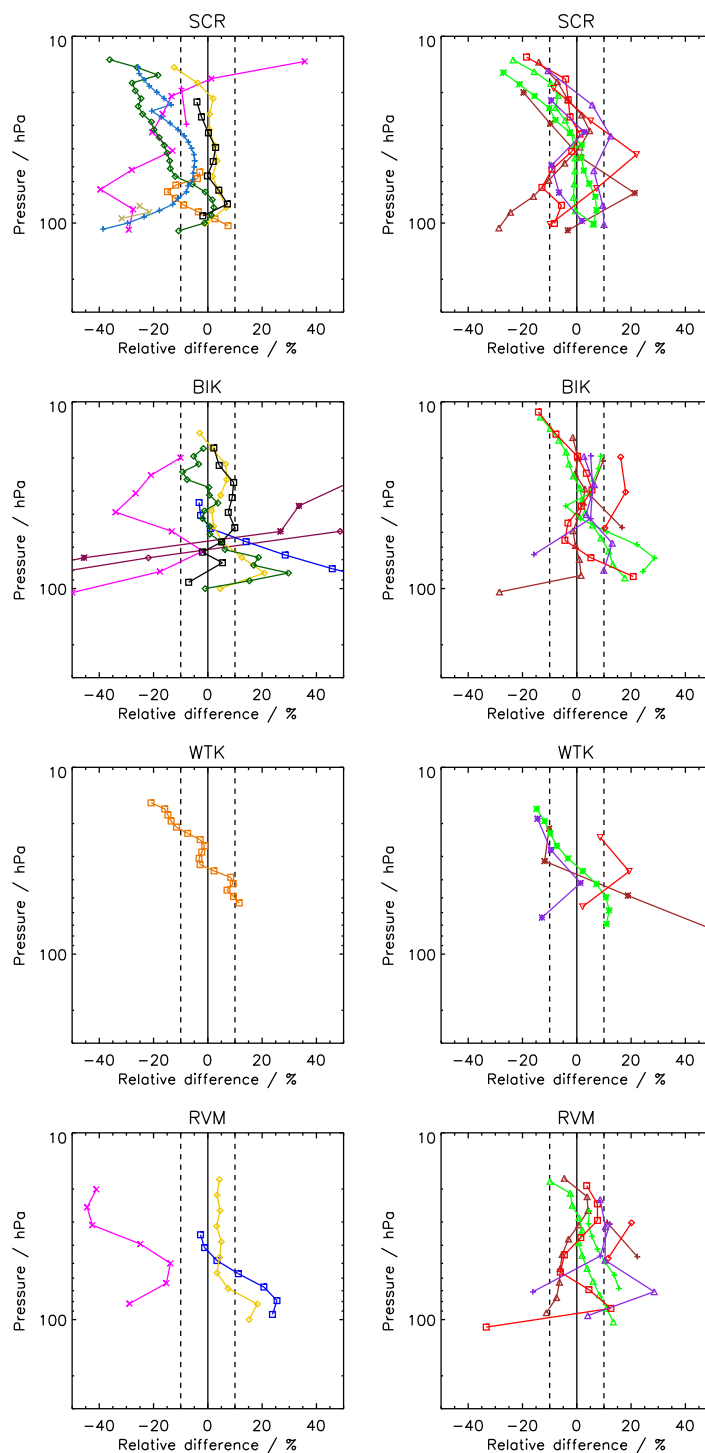
**Figure A9.** Same as Fig. A4 but for SCR, BIK, WTK, and RVM.
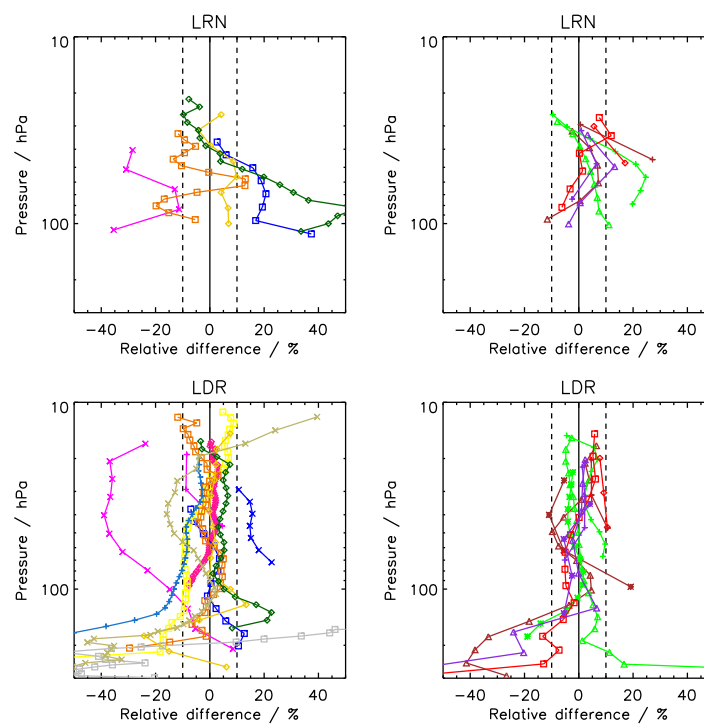
**Figure A10.** Same as Fig. A4 but for LRN and LDR.

## Appendix B: Drift related plots

### B1 Drift profiles for every SAT–FP pair
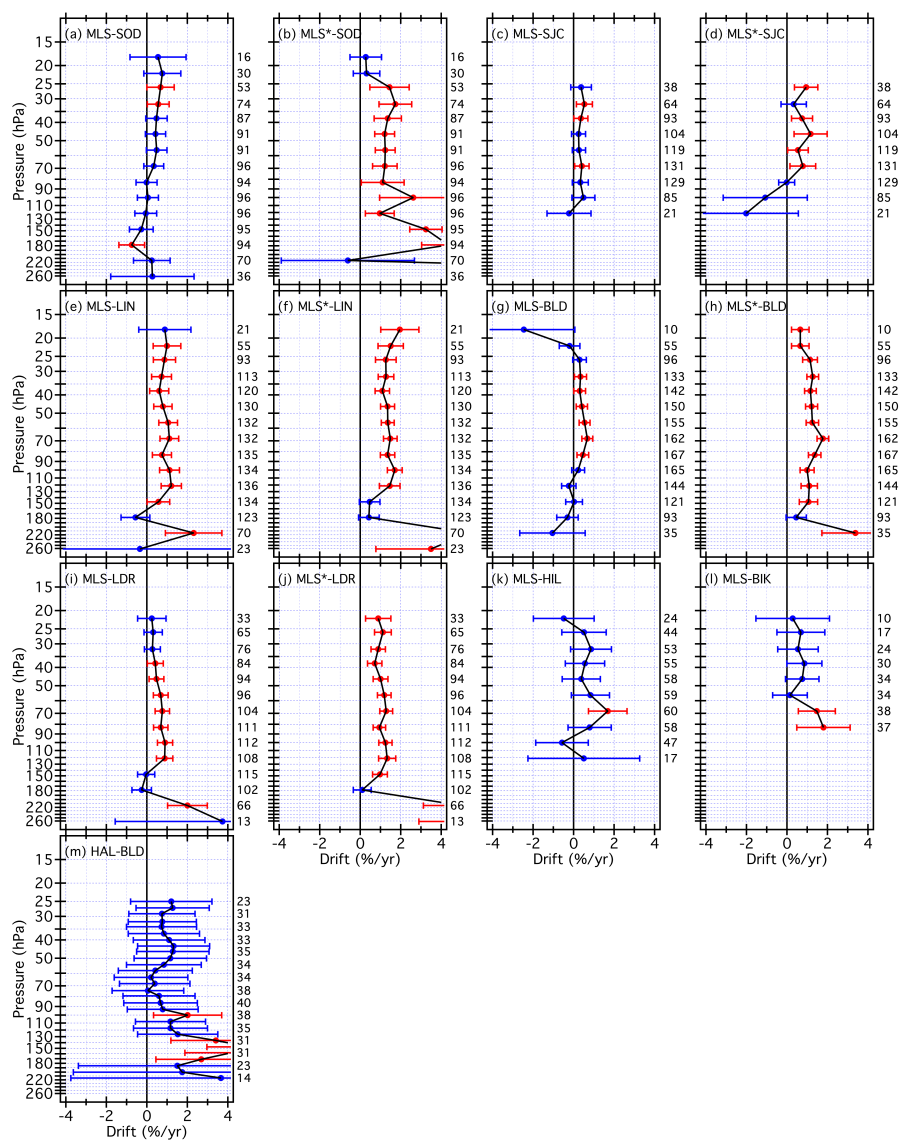
**Figure B1.** Vertical profiles of drifts (filled circles) and their 95 % confidence intervals (horizontal error bars) for 13 different SAT-FP pairs that include MLS, MLS*, and HAL. Blue error bars denote drifts that are not significantly different from zero, while red error bars indicate statistically significant drifts. Numbers in black text to the right of each panel present the number of SAT-FP differences in the timeseries analyzed for drift at the corresponding pressure levels.
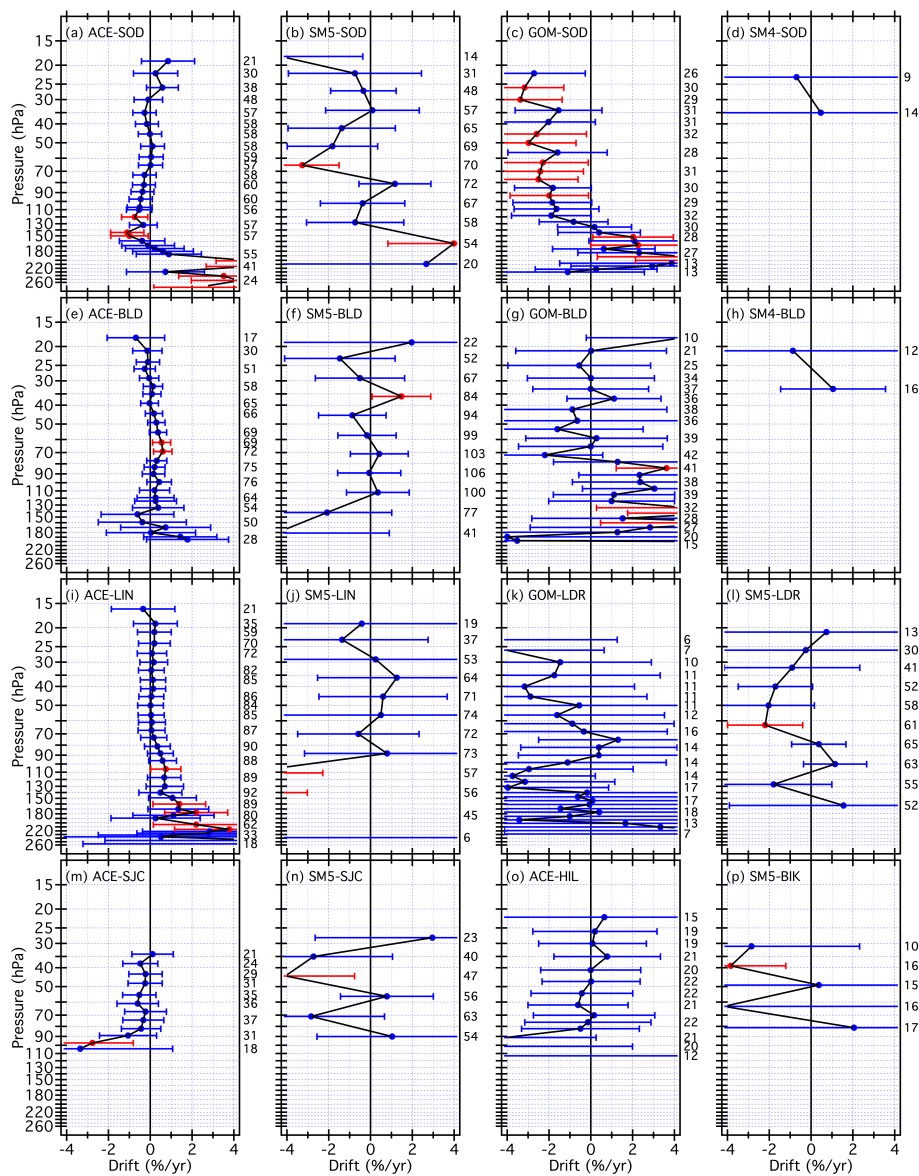
**Figure B2.** Same as Figure B1 for 16 additional SAT-FP pairings that include ACE, SM5, GOM, and SM4.
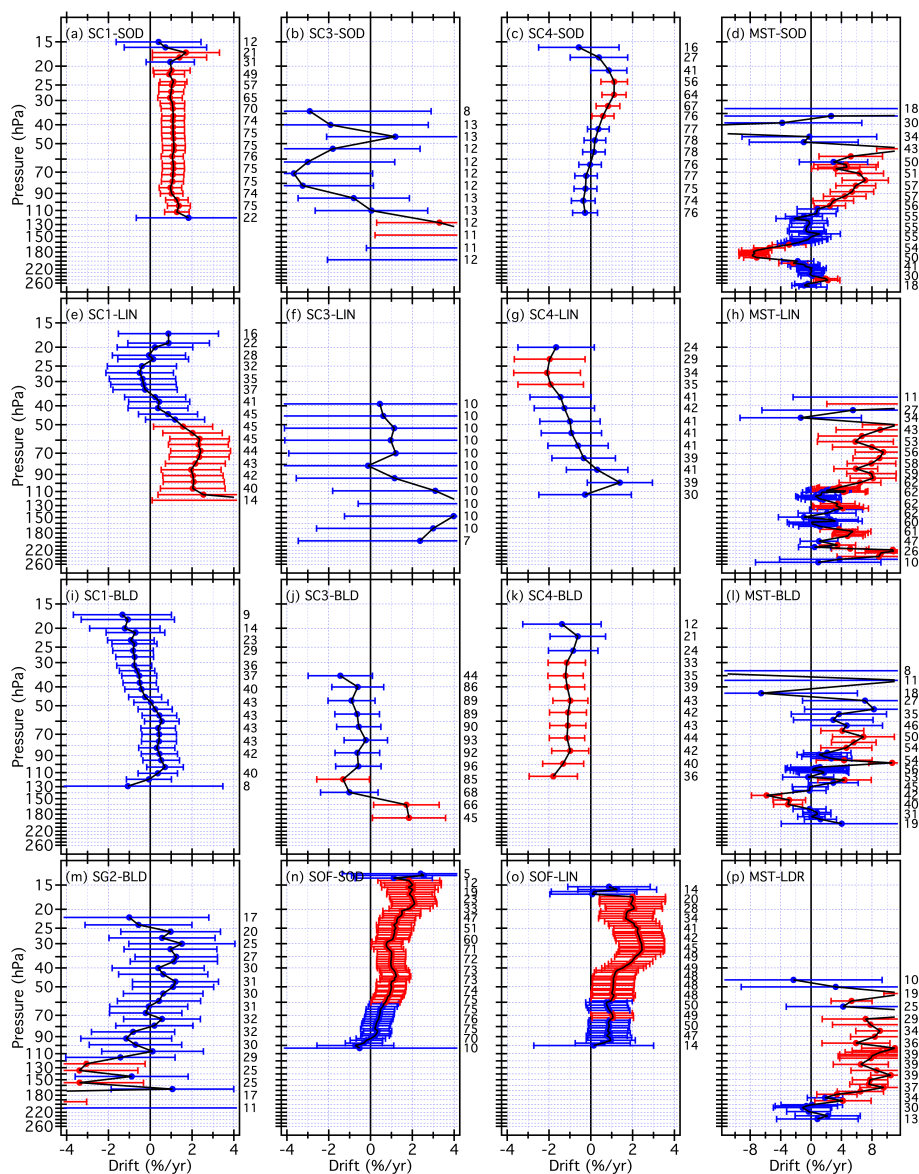
**Figure B3.** Same as Figure B1 for 16 additional SAT-FP pairings that include SC1, SC3, SC4, SG2, SOF, and MST. Note that the x-axis scale for the far right panels is substantially expanded to better show the drifts.
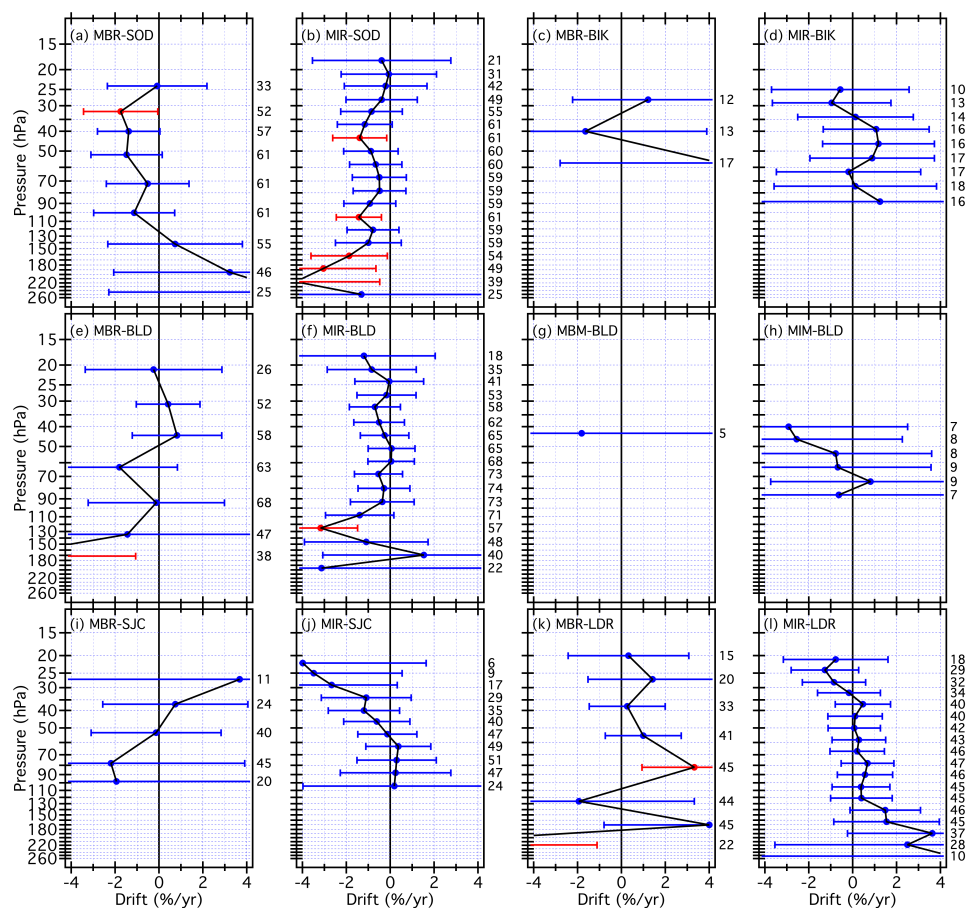
**Figure B4.** Same as Figure B1 for 12 additional SAT-FP pairings that include MBR, MIR, MBM, and MIM.
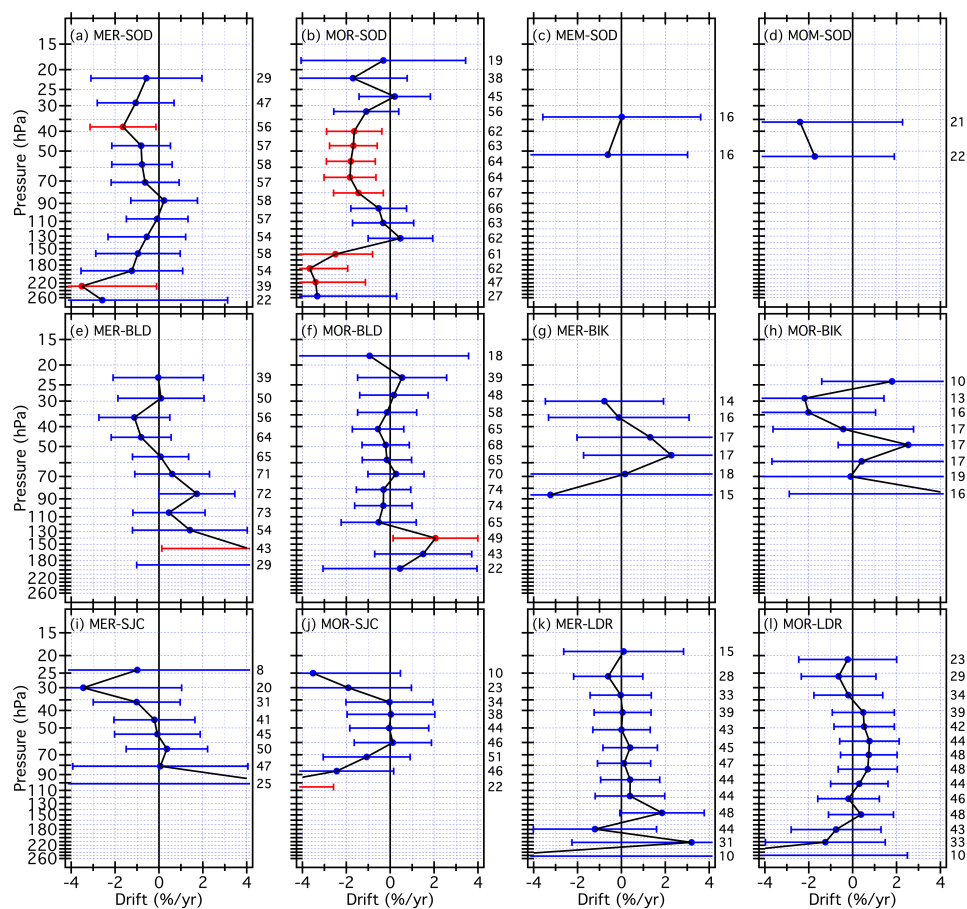
**Figure B5.** Same as Figure B1 for 12 additional SAT-FP pairings that include MER, MOR, MEM, and MOM.

## B2 Drifts per SAT dataset
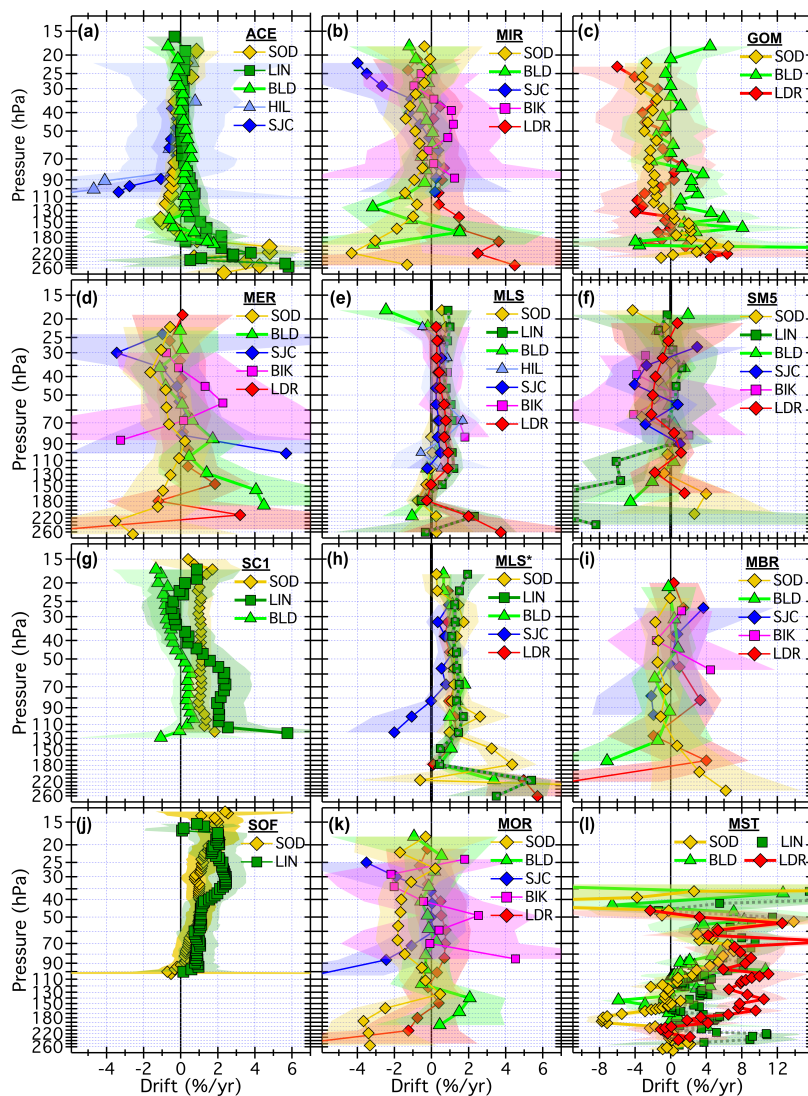
**Figure B6.** Vertical profiles of drifts in SAT-FP differences. Each panel displays the drifts for one SAT (ACE, MIR, GOM, MER, MLS, SM5, SC1, MLS*, MBR, SOF, MOR, and MST) paired with 1-7 different FP sites. Drifts over each FP site (connected colored markers) are presented with their 95 % confidential intervals (colored-matched shading). Shading that does not cross the black vertical line at 0 % drift indicates drifts that are statistically significant. Note that the x-axis scale for the panels in the far right column is expanded to show the drifts with greater clarity.
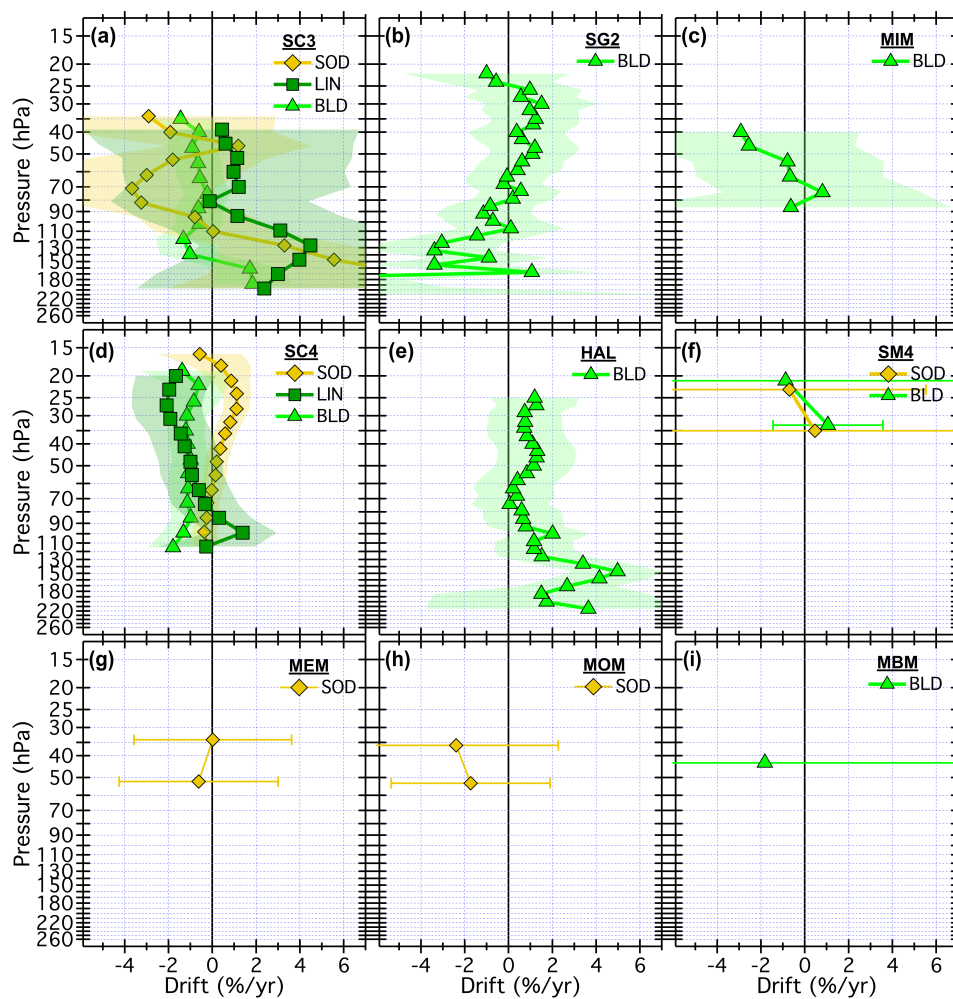
**Figure B7.** Same as Figure B6, but for SC3, SG2, MIM, SC4, HAL, SM4, MEM, MOM, and MBM.

## B3  Drifts per FP dataset, ordered by latitude
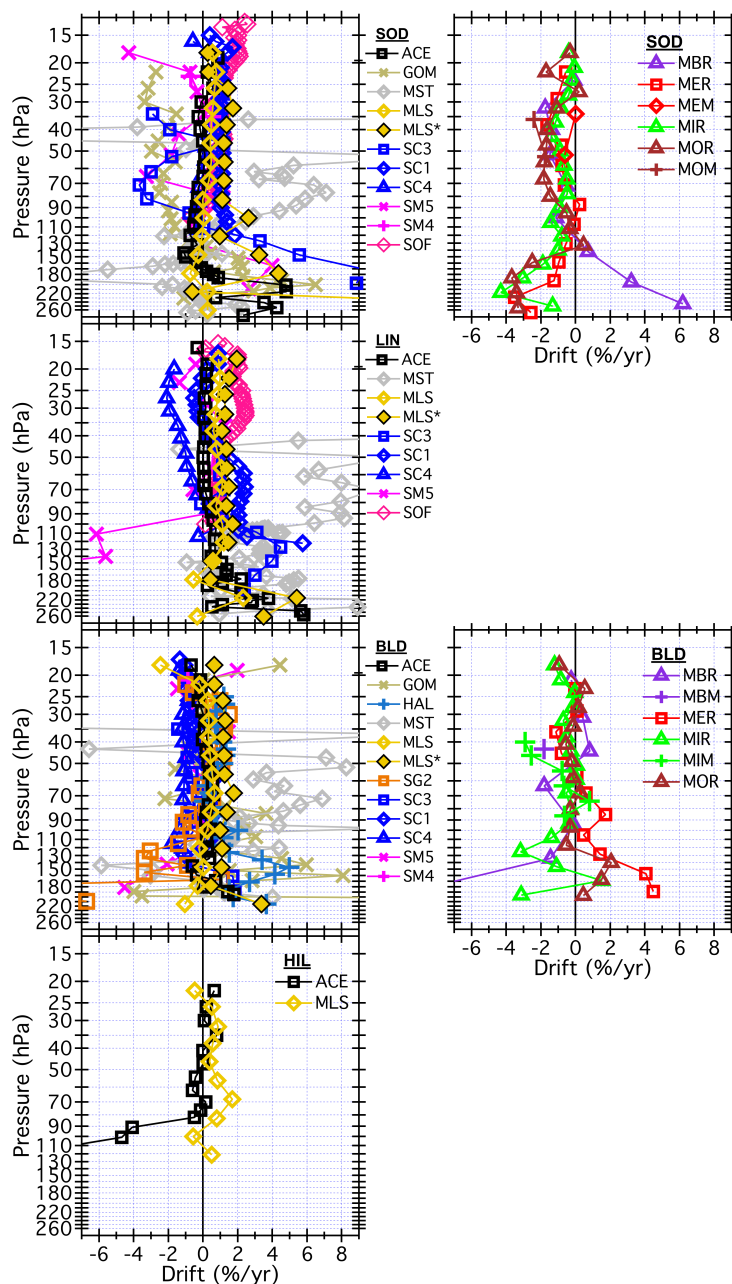
**Figure B8.** Vertical profiles of drifts in SAT-FP differences. Each panel displays the drifts for the multiple SATs paired with each FP site (SOD, LIN, BLD, and HIL). Profiles for non-MIPAS SATs appear in the left column and for MIPAS SATs appear in the right column. For LIN and HIL there were no pairings with MIPAS SATs.
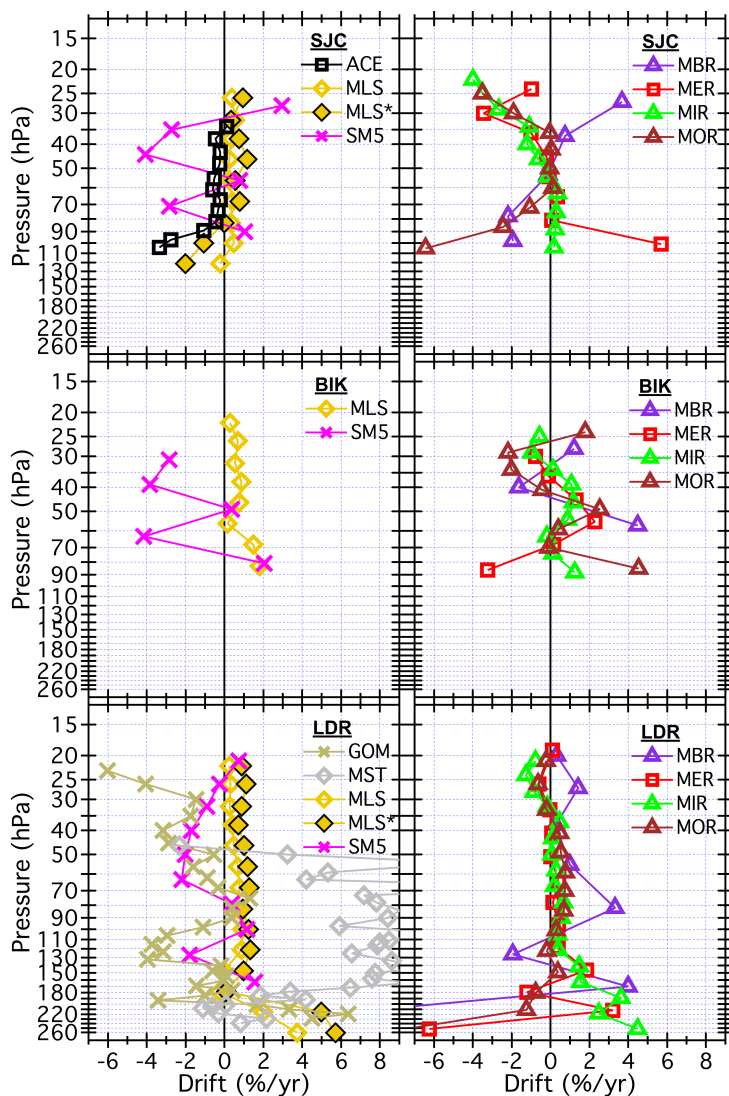
**Figure B9.** Same as Figure B8, but for SJC, BIK, and LDR.

910  *Competing interests.* Several of the (co-)authors are members of the editorial board of Atmospheric Measurement Techniques.

*Acknowledgements.* The Atmospheric Chemistry Experiment (ACE), also known as SCISAT, is a Canadian-led mission mainly supported by the Canadian Space Agency and the Natural Sciences and Engineering Research Council of Canada. We appreciate the HALOE Science Team and the many members of the HALOE project for producing and characterising the high-quality HALOE data set. We would like to thank the European Space Agency for making the MIPAS level-1b data set available. The Oxford MIPAS data was provided by A. Dudhia.

915     MLS data were obtained from the NASA Goddard Earth Sciences and Information Center. Work at the Jet Propulsion Laboratory, California Institute of Technology was done under contract with the National Aeronautics and Space Administration. SCIAMACHY spectral data have been provided by ESA. Stefan Lossow was funded by the Stratospheric Change and its Role for Climate Prediction (SHARP) under contract STI 210/9-2. Thanks to Hauke Schmidt for providing the HAMMONIA data used for the convolution of higher vertically resolved data sets. We want to express our gratitude to SPARC and WCRP (World Climate Research Programme) for their guidance, sponsorship and support

920     of the WAVAS-II programme.

# References

Barrett, E. W., Herndon Jr., L. R., and Carter, H. J.: Some Measurements of the Distribution of Water Vapor in the Stratosphere1, Tellus, 2, 302–311, https://doi.org/10.1111/j.2153-3490.1950.tb00340.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1950.tb00340.x, 1950.

925 Blunier, T., Chappellaz, J. A., Schwander, J., Barnola, J. M., Desperts, T., Stauffer, B., and Raynaud, D.: Atmospheric methane, record from a Greenland Ice Core over the last 1000 year, Geophysical Research Letters, 20, 2219–2222, https://doi.org/https://doi.org/10.1029/93GL02414, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93GL02414, 1993.

Brewer, A. W.: Evidence for a world circulation provided by the measurements of helium and water vapour distribution in the stratosphere, Quarterly Journal of the Royal Meteorological Society, 75, 351–363, https://doi.org/10.1002/qj.49707532603, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49707532603, 1949.

930

Brinkop, S., Dameris, M., Jöckel, P., Garny, H., Lossow, S., and Stiller, G.: The millennium water vapour drop in chemistry-climate model simulations, Atmos. Chem. Phys., 16, 8125–8140, https://doi.org/10.5194/acp-16-8125-2016, 2016.

Dessler, A. E., Schoeberl, M. R., Wang, T., Davis, S., and Rosenlof, K. H.: Stratospheric water vapor feedback, Proc Natl Acad Sci U S A, 110, 18 087–18 091, https://doi.org/10.1073/pnas.1310344110, 2013.

935 Dessler, A. E., Schoeberl, M. R., Wang, T., Davis, S. M., Rosenlof, K. H., and Vernier, J.: Variations of stratospheric water vapor over the past three decades, J. Geophys. Res., 119, 12,588–12,598, https://doi.org/10.1002/2014JD021712, 2014.

Dessler, A. E., Ye, H., Wang, T., Schoeberl, M. R., Oman, L. D., Douglass, A. R., Butler, A. H., Rosenlof, K. H., Davis, S. M., and Portmann, R. W.: Transport of ice into the stratosphere and the humidification of the stratosphere over the 2st century, Geophys. Res. Lett., 43, https://doi.org/10.1002/2016GL067991, 2016.

940 Ding-Zhu, H., Yuan-Yuan, H., Wen-Jun, S., and Fei, X.: Trends of Lower- to Mid-Stratospheric Water Vapor Simulated in Chemistr y-Climate Models, Atmospheric and Oceanic Science Letters, 8, 57–62, https://doi.org/10.3878/AOSL20140088, 2015.

Fueglistaler, S. and Haynes, P. H.: Control of interannual and longer–term variability of stratospheric water vapor, J. Geophys. Res., 110, D24108, https://doi.org/10.1029/2005JD006019, 2005.

Gettelman, A., Birner, T., Eyring, V., Akiyoshi, H., Bekki, S., Brühl, C., Dameris, M., Kinnison, D. E., Lefevre, F., Lott, F., Mancini, E.,
945 Pitari, G., Plummer, D. A., Rozanov, E., Shibata, K., Stenke, A., Struthers, H., and Tian, W.: The Tropical Tropopause Layer 1960–2100, Atmospheric Chemistry and Physics, 9, 1621–1637, https://doi.org/10.5194/acp-9-1621-2009, https://acp.copernicus.org/articles/9/1621/2009/, 2009.

Gettelman, A., Hegglin, M. I., Son, S.-W., Kim, J., Fujiwara, M., Birner, T., Kremser, S., Rex, M., Añel, J. A., Akiyoshi, H., Austin, J., Bekki, S., Braesike, P., Brühl, C., Butchart, N., Chipperfield, M., Dameris, M., Dhomse, S., Garny, H., Hardiman, S. C., Jöckel,
950 P., Kinnison, D. E., Lamarque, J. F., Mancini, E., Marchand, M., Michou, M., Morgenstern, O., Pawson, S., Pitari, G., Plummer, D., Pyle, J. A., Rozanov, E., Scinocca, J., Shepherd, T. G., Shibata, K., Smale, D., Teyssèdre, H., and Tian, W.: Multimodel assessment of the upper troposphere and lower stratosphere: Tropics and global trends, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD013638, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009JD013638, 2010.

Hall, E. G., Jordan, A. F., Hurst, D. F., Oltmans, S. J., Vömel, H., Kühnreich, B., and Ebert, V.: Advancements, measurement uncertainties,
955 and recent comparisons of the NOAA frost point hygrometer, Atmos. Meas. Tech., 9, 4295–4310, 2016.

Hegglin, M. I., Plummer, D. A., Shepherd, T. G., Scinocca, J. F., Anderson, J., Froidevaux, L., Funke, B., Hurst, D., Rozanov, A., Urban, J., von Clarmann, T., Walker, K. A., Wang, H. J., Tegtmeier, S., and Weigel, K.: Vertical structure of stratospheric water vapour trends derived from merged satellite data, Nature Geosci., 7, 768–776, https://doi.org/10.1038/ngeo2236, 2014.

Hurst, D. F., Hall, E. G., Jordan, A. F., Miloshevich, L. M., Whiteman, D. N., Leblanc, T., Walsh, D., Vömel, H., and Oltmans, S. J.: Comparisons of temperature, pressure and humidity measurements by balloon-borne radiosondes and frost point hygrometers during MOHAVE-2009, Atmospheric Measurement Techniques, 4, 2777–2793, https://doi.org/10.5194/amt-4-2777-2011, https://www.atmos-meas-tech.net/4/2777/2011/, 2011a.

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., 116, https://doi.org/10.1029/2010JD015065, 2011b.

Hurst, D. F., Lambert, A., Read, W. G., Davis, S. M., Rosenlof, K. H., Hall, E. G., Jordan, A. F., and Oltmans, S. J.: Validation of Aura Microwave Limb Sounder stratospheric water vapor measurements by the NOAA frost point hygrometer, J. Geophys. Res., 119, 1612–1625, https://doi.org/10.1002/2013JD020757, 2014.

Hurst, D. F., Read, W. G., Vömel, H., Selkirk, H. B., Rosenlof, K. H., Davis, S. M., Hall, E. G., Jordan, A. F., and Oltmans, S. J.: Recent divergences in stratospheric water vapor measurements by frost point hygrometers and the Aura Microwave Limb Sounder, Atmos. Meas. Tech., 9, 4447–4457, https://doi.org/10.5194/amt-9-4447-2016, 2016.

Inai, Y., Shiotani, M., Fujiwara, M., Hasebe, F., and Vömel, H.: Altitude misestimation caused by the Vaisala RS80 pressure bias and its impact on meteorological profiles, Atmos. Meas. Tech., 8, 4043–4054, https://doi.org/10.5194/amt-8-4043-2015, 2015.

Kiehl, J. T. and Trenberth, K. E.: Earth's Annual Global Mean Energy Budget, Bull. Amer. Meteor. Soc., 78, 197–208, 1997.

Kley, D., Smit, H. G. J., Vömel, H., Grassl, H., Ramanathan, V., Crutzen, P. J., Williams, S., Meywerk, J., and Oltmans, S. J.: Tropospheric water-vapour and ozone cross-sections in a zonal plane over the central equatorial Pacific Ocean, Quarterly Journal of the Royal Meteorological Society, 123, 2009–2040, 1997.

Kley, D., Russell III, J. M., and Phillips, C., eds.: Stratospheric Processes and their Role in Climate (SPARC) – Assessment of upper tropospheric and stratospheric water vapour, WCRP No. 113, WMO/TD - No. 1043, SPARC Report No.2, WMO/ICSU/IOC, Paris, 2000.

le Texier, H., Solomon, S., and Garcia, R. R.: The role of molecular hydrogen and methane oxidation in the water vapour budget of the stratosphere, Quarterly Journal of the Royal Meteorological Society, 114, 281–295, https://doi.org/10.1002/qj.49711448002, https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711448002, 1988.

Lelieveld, J., Crutzen, P. J., and Dentener, F. J.: Changing concentration, lifetime and climate forcing of atmospheric methane, Tellus B, 50, 128–150, https://doi.org/10.1034/j.1600-0889.1998.t01-1-00002.x, https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-0889.1998.t01-1-00002.x, 1998.

Lund, R. and Reeves, J.: Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model, J. Clim., 15, 2547–2554, 2002.

Nedoluha, G. E., Bevilacqua, R. M., Gomez, R. M., Hicks, B. C., Russell III, J. M., and Connor, B. J.: An evaluation of trends in middle atmospheric water vapor as measured by HALOE, WVMS, and POAM, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/10.1029/2002JD003332, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD003332, 2003.

Nedoluha, G. E., Kiefer, M., Lossow, S., Gomez, R. M., Kämpfer, N., Lainer, M., Forkman, P., Christensen, O. M., Oh, J. J., Hartogh, P., Anderson, J., Bramstedt, K., Dinelli, B. M., Garcia-Comas, M., Hervig, M., Murtagh, D., Raspollini, P., Read, W. G., Rosenlof, K., Stiller,

G. P., and Walker, K. A.: The SPARC water vapor assessment II: intercomparison of satellite and ground-based microwave measurements, Atmos. Chem. Phys., 17, 14 543–14 558, https://doi.org/10.5194/acp-17-14543-2017, 2017.

995 Oltmans, S. and Hofmann, D.: Increase in lower-stratospheric water vapour at a mid-latitude Northern Hemisphere site from 1981 to 1994, Nature, 374, 146–149, https://doi.org/10.1038/374146a0, 1995.

Oltmans, S. J., Vömel, H., Hofmann, D. J., Rosenlof, K. H., and Kley, D.: The increase in stratospheric water vapor from balloon-borne, frostpoint hygrometer measurements at Washington, D.C., and Boulder, Colorado, Geophysical Research Letters, 27, 3453–3456, https://doi.org/10.1029/2000GL012133, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000GL012133, 2000.

1000 Randel, W. and Park, M.: Diagnosing Observed Stratospheric Water Vapor Relationships to the Cold Point Tropical Tropopause, Journal of Geophysical Research: Atmospheres, 124, 7018–7033, https://doi.org/https://doi.org/10.1029/2019JD030648, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JD030648, 2019.

Randel, W. J., Wu, F., Vömel, H., Nedoluha, G. E., and Forster, P.: Decreases in stratospheric water vapor after 2001: Links to changes in the tropical tropopause and the Brewer–Dobson circulation, J. Geophys. Res., 111, D12312, https://doi.org/10.1029/2005JD006744, 2006.

1005 Read, W. G., Stiller, G., Lossow, S., Kiefer, M., Khosrawi, F., Hurst, D., Vömel, H., Rosenlof, K., Dinelli, B. M., Raspollini, P., Nedoluha, G. E., Gille, J. C., Kasai, Y., Eriksson, P., Sioris, C. E., Walker, K. A., Weigel, K., Burrows, J. P., and Rozanov, A.: The SPARC Water Vapor Assessment II: assessment of satellite measurements of upper tropospheric humidity, Atmos. Meas. Tech., 15, 3377–3400, https://doi.org/10.5194/amt-15-3377-2022, 2022.

Riese, M., Ploeger, F., Rap, A., Vogel, B., Konopka, P., Dameris, M., and Forster, P.: Impact of uncertainties of atmospheric mixing on
1010 simulated UTLS composition and related radiative effects, J. Geophys. Res., 117, D16305, https://doi.org/10.1029/2012JD017751, 2012.

Rodgers, C. D.: Inverse Methods for Atmospheric Sounding: Theory and Practice, vol. 2 of *Series on Atmospheric, Oceanic and Planetary Physics, F. W. Taylor, ed.*, World Scientific, Singapore, New Jersey, London, Hong Kong, 2000.

Rollins, A. W., Thornberry, T. D., Gao, R. S., Smith, J. B., Sayres, D. S., Sargent, M. R., Schiller, C., Krämer, M., Spelten, N., Hurst, D. F., Jordan, A. F., Hall, E. G., Vömel, H., Diskin, G. S., Podolske, J. R., Christensen, L. E., Rosenlof, K. H., Jensen, E. J., and Fahey, D. W.:
1015 Evaluation of UT/LS hygrometer accuracy by intercomparison during the NASA MACPEX mission, Journal of Geophysical Research: Atmospheres, 119, 1915–1935, https://doi.org/10.1002/2013JD020817, 2014.

Rosenlof, K. H., Oltmans, S. J., Kley, D., Russell III, J. M., Chiou, E.-W., Chu, W. P., Johnson, D. G., Kelly, K. K., Michelsen, H. A., Nedoluha, G. E., Remsberg, E. E., Toon, G. C., and McCormick, M. P.: Stratospheric water vapor increases over the past half-century, Geophysical Research Letters, 28, 1195–1198, https://doi.org/10.1029/2000GL012502, https://agupubs.onlinelibrary.wiley.com/doi/abs/
1020 10.1029/2000GL012502, 2001.

Scherer, M., Vömel, H., Fueglistaler, S., Oltmans, S. J., and Staehelin, J.: Trends and variability of midlatitude stratospheric water vapour deduced from the re–evaluated Boulder balloon series and HALOE, Atmos. Chem. Phys., 8, 1391–1402, https://doi.org/10.5194/acp-8-1391-2008, https://acp.copernicus.org/articles/8/1391/2008/, 2008.

Schmidt, H., Brasseur, G. P., Charron, M., Manzini, E., Giorgetta, M. A., Diehl, T., Fomichev, V. I., Kinnison, D., Marsch, D., and Walters,
1025 S.: The HAMMONIA Chemistry Climate Model: Sensitivity of the Mesopause Region to the 11-year solar cycle and $CO_2$ doubling, J. Clim., 19, 3903–3931, https://doi.org/10.1175/JCLI3829.1, 2006.

Solomon, S., Rosenlof, K. H., Portmann, R. W., Daniel, J. S., Davis, S. M., Sanford, T. J., and Plattner, G.-K.: Contributions of Stratospheric Water Vapor to Decadal Changes in the Rate of Global Warming, Science, 327, 1219–1223, https://doi.org/10.1126/science.1182488, 2010.

1030    Stauffer, R. M., Morris, G. A., Thompson, A. M., Joseph, E., Coetzee, G. J. R., and Nalli, N. R.: Propagation of radiosonde pressure sensor errors to ozonesonde measurements, Atmos. Meas. Tech., 7, 65–79, https://doi.org/10.5194/amt-7-65-2014, 2014.

Stiller, G. P., Kiefer, M., Eckert, E., von Clarmann, T., Kellmann, S., García-Comas, M., Funke, B., Leblanc, T., Fetzer, E., Froidevaux, L., Gomez, M., Hall, E., Hurst, D., Jordan, A., Kämpfer, N., Lambert, A., McDermid, I. S., McGee, T., Miloshevich, L., Nedoluha, G., Read, W., Schneider, M., Schwartz, M., Straub, C., Toon, G., Twigg, L. W., Walker, K., and Whiteman, D. N.: Validation of MIPAS

1035    IMK/IAA temperature, water vapor, and ozone profiles with MOHAVE-2009 campaign measurements, Atmos. Meas. Tech., 5, 289–320, https://doi.org/10.5194/amt-5-289-2012, 2012.

Vömel, H., David, D. E., and Smith, K.: Accuracy of tropospheric and stratospheric water vapor measurements by the cryogenic frost point hygrometer: Instrumental details and observations, J. Geophys. Res., 112, https://doi.org/10.1029/2006JD007224, 2007a.

Vömel, H., Yushkov, V., Khaykin, S., Korshunov, L., Kyro, E., and Kivi, R.: Intercomparison of stratospheric water vapor sensors: FLASH-b

1040    and NOAA/CMDL frost point hygrometer, J. Atmos. Ocean. Tech., 27, 941–952, https://doi.org/10.1175/JTECH2007.1, 2007b.

Vömel, H., Naebert, T., Dirksen, R., and Sommer, M.: An update on the uncertainties of water vapor measurements using cryogenic frost point hygrometers, Atmos. Meas. Tech., 9, 3755–3768, https://doi.org/10.5194/amt-9-3755-2016, 2016.

World Meteorological Organization: Definition of the tropopause, Bulletin of the World Meteorological Organization, 6, 136–137, 1957.