Chapter 3: Challenges for trend studies

This chapter aims to provide information that, in *Chapter 5*, assists the interpretation of the long-term trends derived from the ozone profile data records described earlier. Profile trends can differ for a variety of reasons and this represents a real challenge to assess the long-term evolution of ozone (*e.g.*, *WMO,* 2014; *Harris et al.,* 2015; and references therein).

One of the primary drivers of differences in trends comes from the data sets themselves. First of all, single-instrument data records can differ in terms of their stability, their accuracy, and their sampling and smoothing properties in the spatial and temporal domain. The stability of a data record can be affected by aging of or changes in instrumentation, which stems from evolutions in the operation or calibration procedures, *etc*. Secondly, inter-instrument biases lead to discontinuities in merged data records for each transition from one set of instruments to another. The ability to adjust for offsets between instrument records depends on how they are merged and especially on the sampling properties and data quality of the records (*Tummon et al.,* 2015). This merging step is unavoidable for satellite data since no single instrument provides both the spatial and the temporal coverage needed to study multi-decadal trends at the near-global scale. In *Section 3.1*, we report the results of intercomparisons of different single-sensor or multi-sensor data records. These studies aim to identify potential artefacts in the time series, which can help us understand the cause of discrepancy between trends.

Another possible cause of differences in trends from intercomparisons lies in the reduction of the single profile data to monthly zonal mean data. Changes in the sampling pattern introduce a changing bias in the presence of spatio-temporal gradients in the ozone field (*e.g.*, diurnal and seasonal cycles or meridional structure). Differences in sampling properties also affect the comparison of ground-based and satellite trends. The satellite monthly zonal mean time series are not necessarily representative of the monthly mean values observed above the station. Both issues are studied in more detail in *Section 3.2*.

3.1 Consistency of ozone profile data records

3.1.1 Homogeneity of ground-based network data

In this section, we investigate inhomogeneities in the ground-based data records; these may occur in time (*e.g.*, due to changes in instrumentation, instrument performance, or calibration) and in space (*e.g.*, due to differences in instrumentation, instrument performance, or calibration between sites). We describe the key results of exploratory work by *Hubert et al.* (2019) on the homogeneity of ozone profile observations gathered by ground-based networks between 2002 and 2016. They used an ensemble of complementary high quality satellite data records as a transfer standard to investigate the data at 60 ozonesonde, 8 lidar, and 5 microwave radiometer stations operating within the NDACC, GAW, and SHADOZ networks. A description of these ground-based instruments can be found in *Section 2.1.1* of this Report. More detailed findings, discussion and conclusions of this exploration can be found in *Hubert et al.* (2019).

The analysis of *Hubert et al.* (2019) starts off with the following steps. Profile data from a ground-based record X_L (*e.g.*, ozonesonde data at a given location *L*) and a space-based record Y_M (*e.g.*, of satellite mission *M*) are first cleared of spurious measurements, then co-located in time and space (the window is detailed in the next paragraph), then converted to the same profile representation (ozone unit and vertical coordinate), and then finally interpolated to a common vertical grid. Ozonesonde and lidar data are smoothed to the vertical resolution of the satellite data, which differs for each instrument *M* (**Table 1b** in *Hassler et al.,* 2014). The second part of the analysis consists of computing the relative difference $\Delta_{LM}(z,t_i) = 100*(X_{L,i} - Y_{M,i})/Y_{M,i}$ for each profile pair *i*. For the sake of brevity, the indices *L* and *M* are excluded from the formulae that follow. These $\Delta(z,t_i)$ form a time series of relative differences (expressed as a percentage) which are smoothed using an N-month running median filter centered around the middle of each month *j*, which we denote $\Delta^*(z,t_j)$. What are used in the final analysis are time series of relative difference anomalies δ (expressed as a percentage),

$$\delta(z,t_j) = \Delta^*(z,t_j) - \underline{\Delta}(z) \qquad (3.1).$$

Here, $\underline{\Delta}(z)$ represents the median value of $\{\Delta(z,t_i)\}$ for all t_i in the reference period, which is 2005–2011 for all instruments except OMPS-LP where 2012–2016 was chosen (see grey horizontal line in **Figures 3.1 – 3.4**). By removing the satellite- and grid level-dependent median value, the values of δ at different grid levels (*z* and *z'*) and from different satellite records (*M* and *M'*) are on a comparable scale. The δ time series represents anomalies of the relative difference of ground-based minus satellite observations with respect to their median value over the reference period. Positive anomalies indicate that the ground-based bias relative to a satellite is more positive (or less negative) than usual during the reference period and vice versa for negative anomalies.

By construction, these positive and negative anomalies average to zero over the reference period, but this is not necessarily the case outside of this period. The median was preferred over the mean for $\underline{\Delta}(z)$ to avoid the impact of single outliers on the absolute scale of the anomaly time series. The uncertainty (2σ) of the anomalies (δ) is estimated as half the 95 % interpercentile of the Δ values in the running window divided by the square root of the number of pairs in that window. The absolute scale holds valuable information but is mostly disregarded here, as it can be offset for different M due to differences in the reference period or differences in the sampling of the co-located data set.

Different filter window lengths N were investigated. We only show results for a 12-month wide window, which balances the need to reduce the noise in the comparisons in order to identify small anomalies and the desire to preserve information to localise identified anomalies sufficiently well in time. The influence of natural variability is considered negligible because of the large smoothing window but especially because the pairs are well co-located in space (< 300km for comparison to MWR; all others < 500km) and time (< 1h for MWR stations with hourly data; < 6h for MIPAS and Aura MLS; all other comparisons < 12h).

Inhomogeneities in the ground-based data reveal themselves in the temporal and vertical structure of the relative difference anomalies. Increased confidence that anomalies are caused by (originate in) the ground-based record L is found when the δ_{LM} values are significant for multiple independent satellite references {M} over the same altitude region and during the same period in time. The six limb and occultation satellite records selected for this study represent complementary measurement techniques (different spectral ranges, viewing geometries, sampling properties, calibration and retrieval methods) and can therefore be considered independent. The consideration of six complementary and independent satellite records is a vital asset in attributing common features in δ_{LM} to the ground-based data. We considered OSIRIS, GOMOS, MIPAS, SCIAMACHY, Aura MLS, and OMPS-LP as references (**Table 3.1**), which are all fairly dense samplers that have been thoroughly validated and intercompared in recent years (*Tegtmeier et al.,* 2013; *Hassler et al.,* 2014; *Rahpoe et al.,* 2015; *Hubert et al.,* 2016).

Figure 3.1 shows anomaly time series δ for the MLO lidar station in Hawaii, USA and the six satellite records. Stippled cells are not statistically different from zero at the 2-sigma level. The temporal and vertical structure is fairly featureless with only a few areas showing significant anomalies, which shows the good and stable agreement between this lidar record and the different satellite references. While significant positive anomalies are apparent from 2002 to 2004 for OSIRIS, GOMOS, and SCIAMACHY, the statistically significant region is not fully coherent between the three references, which leaves ambiguity as to whether the differences are caused by the ground-based or the satellite record. Significant negative anomalies of ~4 % are apparent at 40–45 km starting in early 2013 for both OSIRIS and Aura MLS. Unfortunately, OMPS-LP cannot confirm this finding as it only started taking measurements in 2012. However, this feature resides at the upper end of the lidar profile where the uncertainty of the measurement becomes larger and hence larger anomalies are expected (*e.g.*, see **Figure 16** of *Leblanc et al.,* 2016b).

Figures 3.2 to **3.4** show selected results for ozonesonde, lidar, and microwave radiometer station records. It is not the purpose of this Report to discuss each station record individually as these can be found in *Hubert et al.* (2019). Instead, we focus on the general performance of the ground-based networks. The majority of the ground station data records exhibit one or more temporal features in their anomaly field with a magnitude of 5 % and often more. These features manifest themselves as (a series of) sudden discontinuities or as transient events over a variety of timescales, but they do not necessarily occur over the entire vertical range.

Some observed discontinuities coincide with, and are caused by, known changes in instrumentation. For instance, the switch from the KC-96 to the ECC ozonesonde at Naha station in November 2008 (*Morris et al.,* 2013) is clearly visible as a +10 % discontinuity in **Figure 3.2** (panel B). A correction scheme has been developed (*Section 2.1.1.1*) which should

Instrument	Platform	Analysis period	Level-2 data version	Viewing geometry	Spectral range	Analysis
SAGE II	ERBS	1984-2005	v7	solar occultation	UV-VIS-NIR	S
HALOE	UARS	1991-2005	v19	solar occultation	NIR-SWIR	S
OSIRIS	Odin	2001-2016	v5.10	limb scattered	UV-VIS-NIR	G, S
GOMOS	EnviSat	2002-2011	ALGOM2s v1	stellar occultation	UV-VIS-NIR	G, S
MIPAS		2005-2012	IMK/AA v7	limb emission	MIR-TIR	G, S
SCIAMACHY		2003-2012	IUP v3.5	limb scattered	UV-VIS-NIR	G, S
ACE-FTS	SciSat-1	2004-2016	v3.5/v3.6	solar occultation	MIR-TIR	S
MLS	EOS/Aura	2005-2016	v4.2	limb emission	MW	G, S
OMPS-LP	Suomi-NPP	2012-2016	USask-2D v1.0.2	limb scattered	UV-VIS	G, S

Table 3.1: *Characteristics of nine limb/occultation satellite data records. The last column indicates whether the data were used for the study of the homogeneity of ground-based data records (G; see Section 3.1.1) and/or for the estimation of satellite drift (S; see Section 3.1.2).*

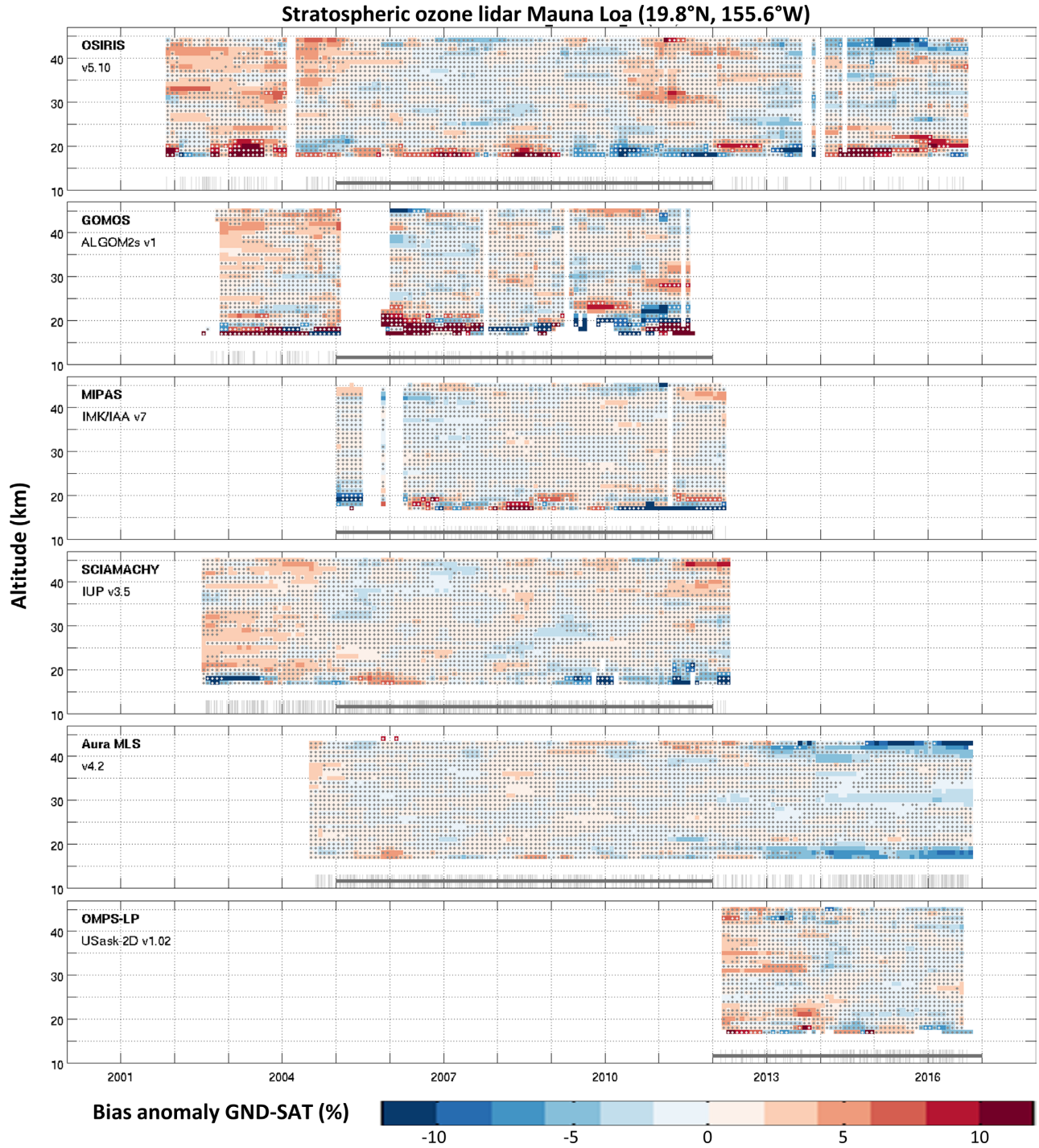

Figure 3.1: *Smoothed anomaly time series (δ, see Eq. 3.1) of the relative difference of MLO lidar and six satellite ozone profile data records (top to bottom). Red values indicate regions in which lidar measurements are biased more positive (or less negative) compared to satellite than their median value during the reference period. Stippled areas denote δ values that are not statistically different from zero at the 2-sigma level. A running average with a 12-month window was used to smooth the time series. Thin grey vertical lines show the sampling of the co-located profile data records; the grey horizontal lines indicate the reference period for each comparison. Adapted from Hubert et al. (2019).*

result in a more homogeneous time series. Another example is the positive (negative) bias excursion above (below) the ozone maximum starting in 2010 and ending in 2012 in the Hohenpeissenberg lidar record (**Figure 3.3**; panel A). These are likely related to an aging device that fired the laser until it was replaced in early 2013 (*W. Steinbrecht,* private comm.). However, in many cases further investigations are needed to understand these anomalies, to find the cause of the changes, and to ultimately develop a correction strategy.

In this respect, we comment that simply adjusting to another observational data record will lead to a loss of independence between records. Clearly, this should be avoided as much as possible.

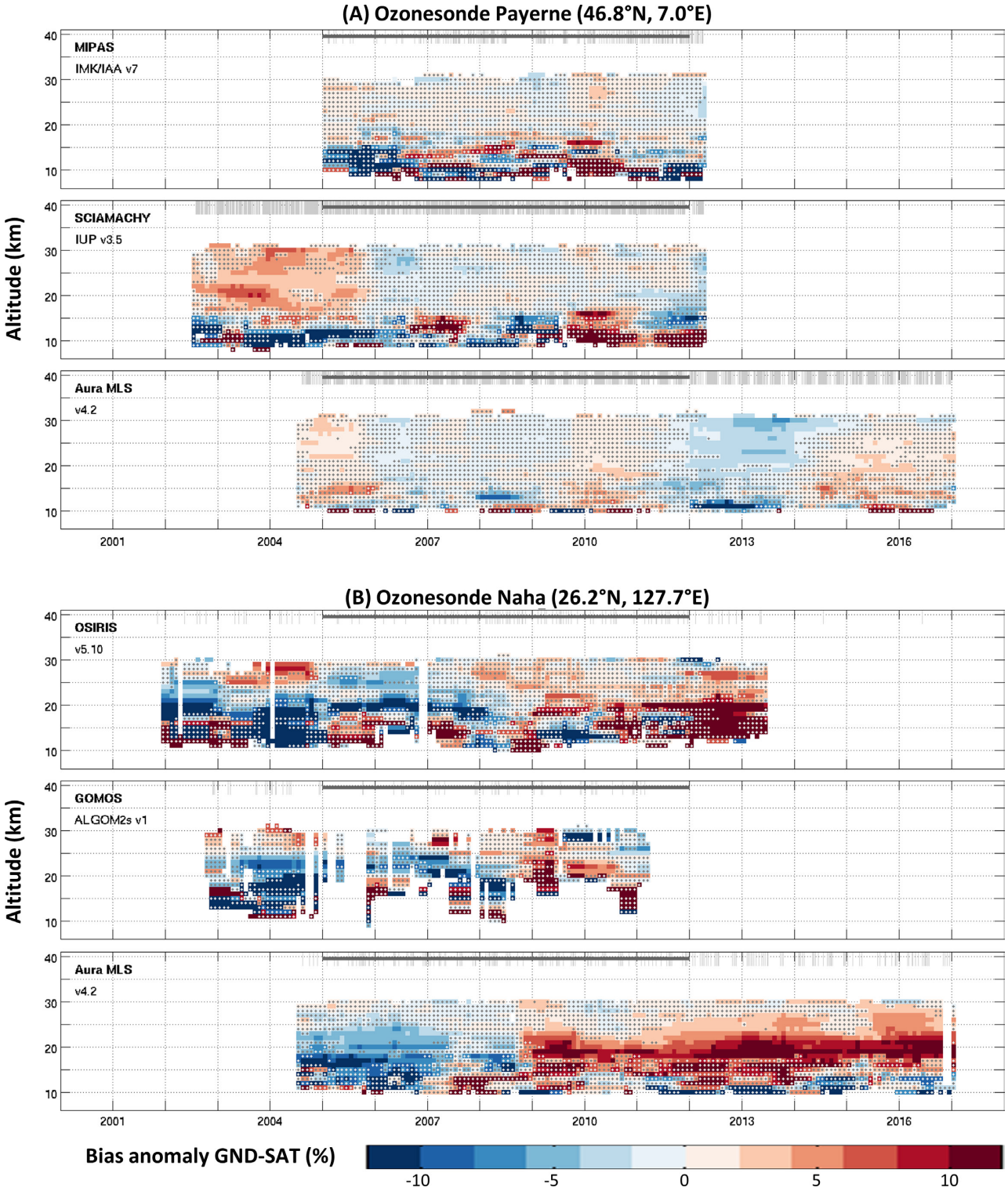

Figure 3.2: *As* ***Figure 3.1*** *but for two ozonesonde sites each with a different selection of three satellite references. Stippled areas denote non-significant δ values. Comparisons to all six satellite records for both stations are shown in* ***Figures S3.1*** *and* ***S3.2*** *in the Supplement. Adapted from Hubert et al. (2019).*

Our results indicate that most of the 73 considered station records (60 ozonesonde, 8 lidar, and 5 MWR) have one or more periods with inhomogeneities over part of the vertical range of the data record. Such artefacts are noted even in ground-based data records that are generally considered as "golden" time series for trend studies (because of their length and/or supposedly better stability). Examples are shown in **Figures 3.2** (A), **3.3** (A&B), and **3.4** (A). The magnitude of the anomalies is broadly consistent with the quoted 5–10 % systematic uncertainty of the ground-based measurement techniques (*Section 2.1*). Nonetheless, it illustrates the importance of clarifying to data users that the quoted systematic uncertainty is pertinent to every single ozone profile and that the sign and magnitude may change

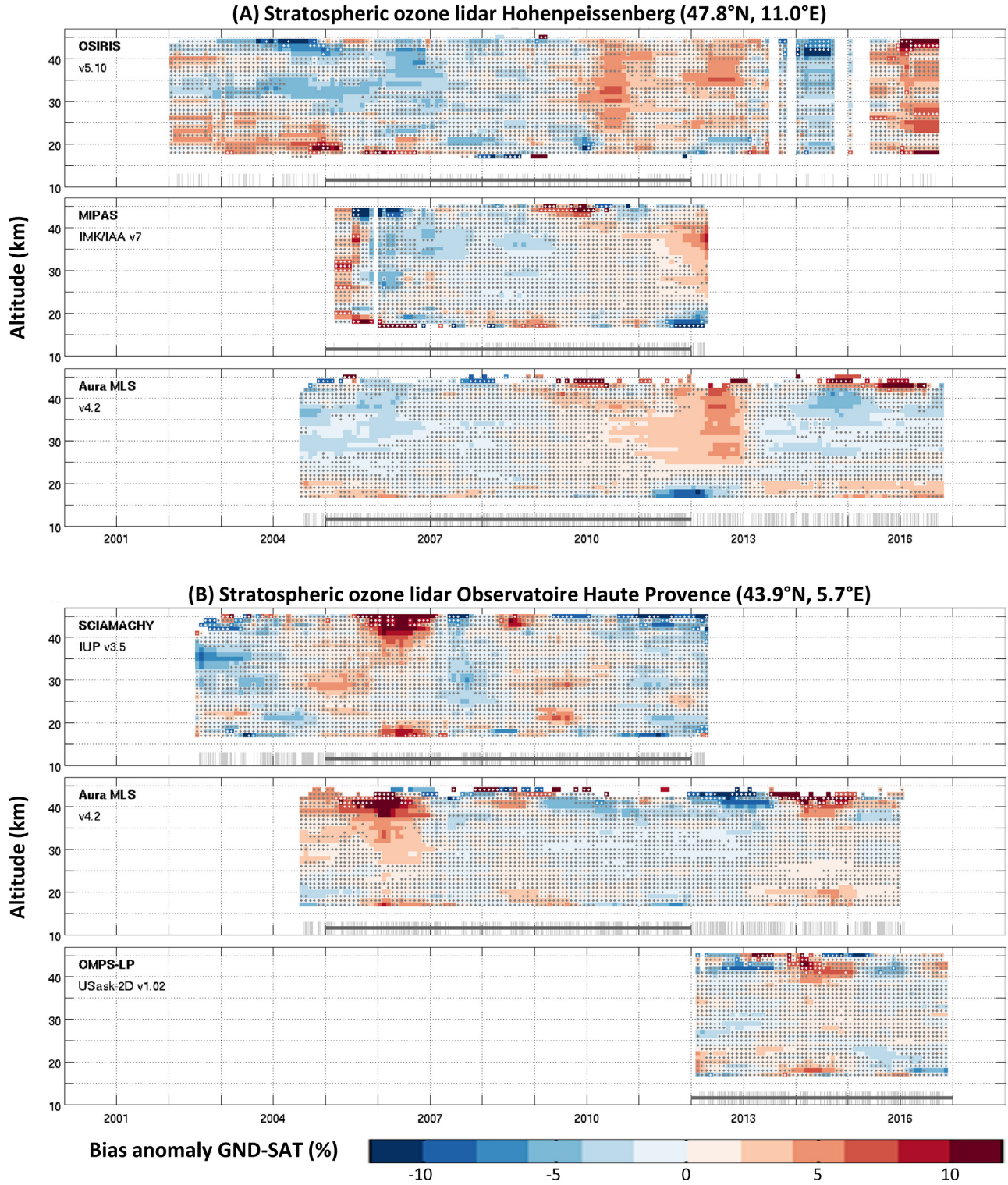

Figure 3.3: *As **Figure 3.2** but for two stratospheric ozone lidar sites. Stippled areas denote non-significant δ values. Comparisons to all six satellite records for both stations are shown in **Figures S3.3** and **S3.4** in the Supplement. Adapted from Hubert et al. (2019).*

over the data record, thereby effectively representing a non-systematic uncertainty component in the time domain.

Measurement artefacts are generally not modelled in regression analyses, thereby introducing random and systematic uncertainty in profile trends. These artefacts are furthermore dependent on the station and the vertical level. Single large discontinuities or multiple discontinuities with the same sign constitute a low frequency signal which will clearly bias the derived trend. The likelihood of such trend biases decreases with an increasing number of excursions as long as their sign and magnitude is sufficiently random in time. The random uncertainty of the trend estimate, on the other hand, will unavoidably accrue contributions from the unmodelled variance.

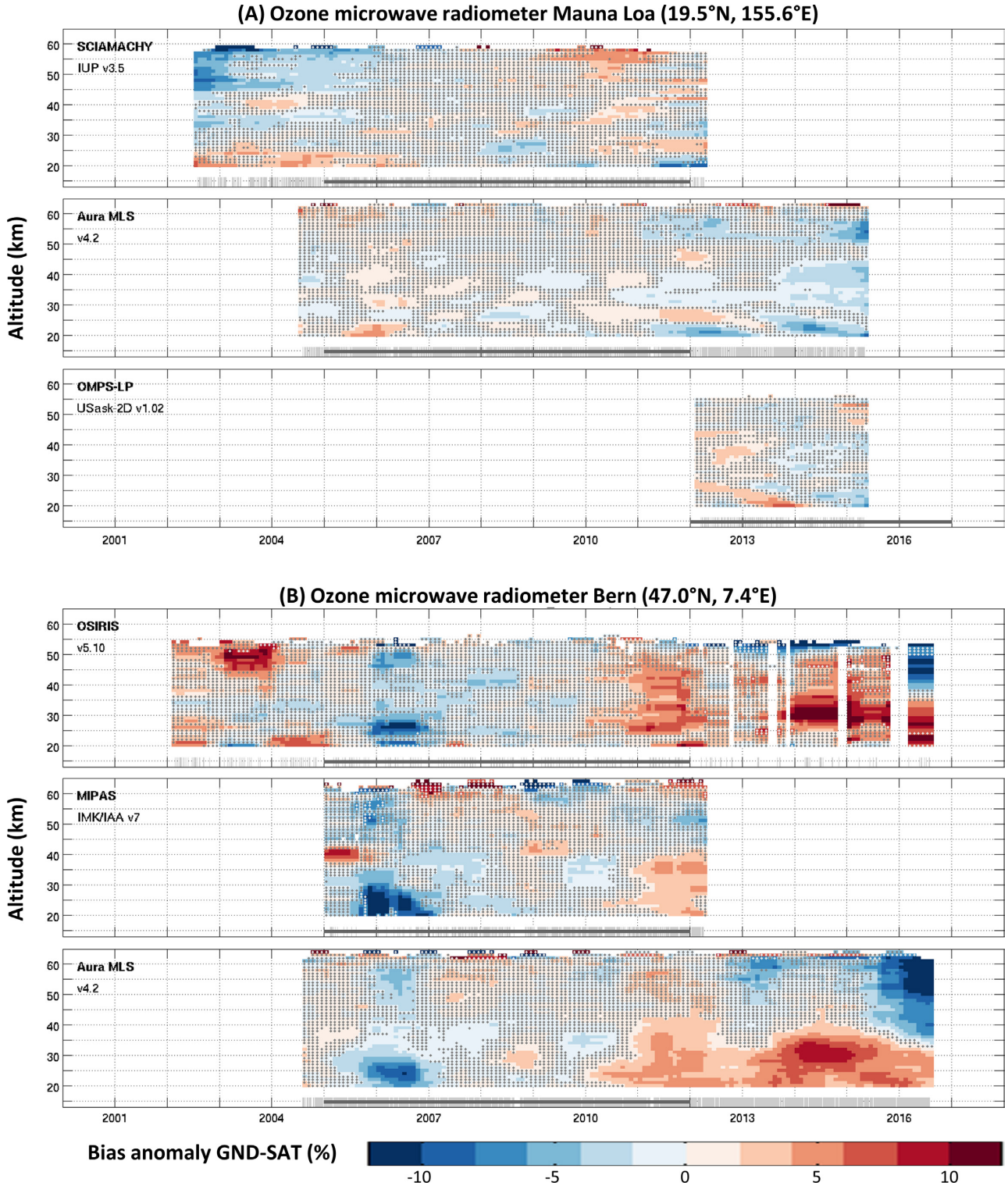

Figure 3.4: *As* ***Figure 3.2*** *but for two microwave radiometer sites. Stippled areas denote non-significant δ values. Comparisons to all six satellite records for both stations are shown in* ***Figures S3.5*** *and* ***S3.6*** *in the Supplement. Adapted from Hubert et al. (2019).*

A detailed time series analysis is needed to quantify the possible impact of measurement inhomogeneities on ozone profile trends. However, since both occurrence and magnitude of the artefacts depend on the site and the ground-based instrument, it is expected that the profile trends will (also) differ as a result of the additional bias and variance. Such differences between neighbouring sites and between ground-based instruments at one site have been reported repeatedly in past and recent analyses (*e.g.*, *Steinbrecht et al.*, 2006; *Logan et al.*, 2012; *Nair et al.*, 2013, 2015; *Tarasick et al.*, 2016; *Van Malderen et al.*, 2016).

A comparison of trends from different instruments at Lauder and Hawaii using the same regression model will be shown in *Section 5.4*.

A pragmatic approach to reduce the impact of these measurement inhomogeneities would be to average the station records over several sites, perhaps the entire ground-based network (*Logan et al.,* 1994, 1999a, 1999b; *Terao and Logan,* 2007). This will effectively decrease the relative importance of systematic effects as many measurement artefacts across the network are of varying magnitude and occur randomly in time and space. However, some artefacts can be attributed to changes that occurred fairly simultaneously across parts of the network, in particular for the ozonesonde data records. For instance, the ten sites in the Canadian subnetwork transitioned from BM to ECC ozonesonde models in the early 1980s and had a further series of simultaneous changes in the following decades (*Tarasick et al.,* 2016), the five Japanese sites switched from the KC to the ECC ozonesonde around 2009 (*Morris et al.,* 2013), and there are additional changes in the NOAA and SHADOZ subnetworks (*Witte et al.,* 2017; *Sterling et al.,* 2018). Such simultaneous changes in the sonde network will not be fully averaged out, so it is vital to have as many independent station records as possible. Measurement operations at lidar and MWR stations, on the other hand, are fairly independent, but the number of sites is much smaller than for the sonde network. This re-emphasises the need to sustain the current number of stations in the ground-based networks.

3.1.2 Stability of limb data records relative to ground-based networks

Bottom-up calculations of the stability of ozone profile data records, that is from first principles or from the propagation of low level monitoring data through the retrieval chain, are contentious as they rarely lead to a realistic perception of the long-term performance of the measurement systems. Top-down approaches compare profile measurements to a reference data record and ultimately derive estimates of the stability (also called "drift") relative to that reference (*e.g.*, *Nair et al.*, 2012; *Rahpoe et al.*, 2015; *Hubert et al.*, 2016). These estimates approximate absolute stability if the reference is sufficiently stable. However, past validation and intercomparison exercises have shown the challenges in establishing one (or more) ozone profile data records as a stable data record at the level required by profile trend assessments, which lies around 1 % per decade (*GCOS*, 2011).

Results from intercomparisons between different satellite records are described in *Sections 3.1.3–3.1.5*. In this section we use the ground-based networks of ozonesonde, lidar, and microwave radiometer measurements as a reference to assess the decadal stability of single-sensor single-profile data (Level-2) and of gridded monthly zonal mean data (Level-3) from single sensors and for one multi-sensor data record. Only limb/occultation sounders are considered here; ground-based comparisons for the SBUV nadir profilers have been reported by *Kramarova et al.* (2013a). The method follows that of *Hubert et al.* (2016) where regressions are made to the different time series of satellite and ground-based data, and the linear slope estimates, interpreted as satellite drift, are subsequently averaged over the network to obtain pseudo-global estimates. This reduces the impact of noise and inhomogeneities in the ground-based records on the satellite drift estimate (see *Section 3.1.1*). However, the uncertainties resulting from the linear fit do not fully take into account inhomogeneities across the network, so these are inflated using a χ^2-scheme (see *Section 4.1.2* in *Hubert et al.*, 2016). Finally, the uncertainty of the network-averaged satellite drift is obtained by propagating the χ^2-corrected uncertainties of the linear term through the weighted average.

The first analysis investigates single satellite ozone profiles (*i.e.*, Level-2) co-located in space (< 300 or 500 km) and time (< 1, 6, or 12 hours) to ground-based observations by ozonesonde, stratospheric lidar, and MWR networks. The co-location requirement reduces the number of compared measurements considerably in favour of a smaller mismatch uncertainty (*Verhoelst et al.*, 2015) in the comparison time series. **Figure 3.5** shows the vertical dependence of the drift relative to ozonesonde (bottom left), lidar (top left), and MWR (bottom right) for eleven limb/occultation sounder data records. Nine of these are part of a merged data record in this Report: SAGE II, HALOE, OSIRIS, GOMOS, MIPAS, SCIAMACHY, ACE-FTS, Aura MLS, and OMPS-LP (detailed version information can be found in **Table 3.1**). **Figure 3.6** shows the corresponding significance, with 2σ chosen as threshold for detection (*i.e.*, 95 % confidence level).

Longer time series are available here compared to *Hubert et al.* (2016), but revised or different satellite retrieval algorithms were also considered for most instruments (except for SAGE II and HALOE, for which no new data were available), and one entirely new instrument record was added to the analysis (OMPS-LP). The main conclusions of *Hubert et al.* (2016) still hold and a fairly consistent picture emerges from the ozonesonde, lidar, and microwave radiometer results. These show that, generally, the limb/occultation data records are stable within 5 % per decade in the middle and upper stratosphere. For some records, for example SAGE II and Aura MLS, the constraints on stability are even better, with an upper bound on drift of less than 2 % per decade. No significant drift was found for MIPAS and ACE-FTS. SCIAMACHY data prior to August 2003 were removed from the analysis (see *Section 3.1.3* and *Sofieva et al.*, 2017). The negative drift around 35 km is now no longer statistically significant and decreased from 5 % to 3 % per decade. However, statistically significant deviations from zero were found for a few instruments in different regions of the atmosphere. In chronological order, HALOE data around 25–30 km drift to lower ozone mixing ratios by 3–4 % per decade. Improvements in the pointing stability for OSIRIS have clearly reduced the positive drift from 8 % to 4 % per decade, but the latter result remains statistically significant. And GOMOS occultation data drift to lower ozone values by 5 % per decade and more below around 25 km.

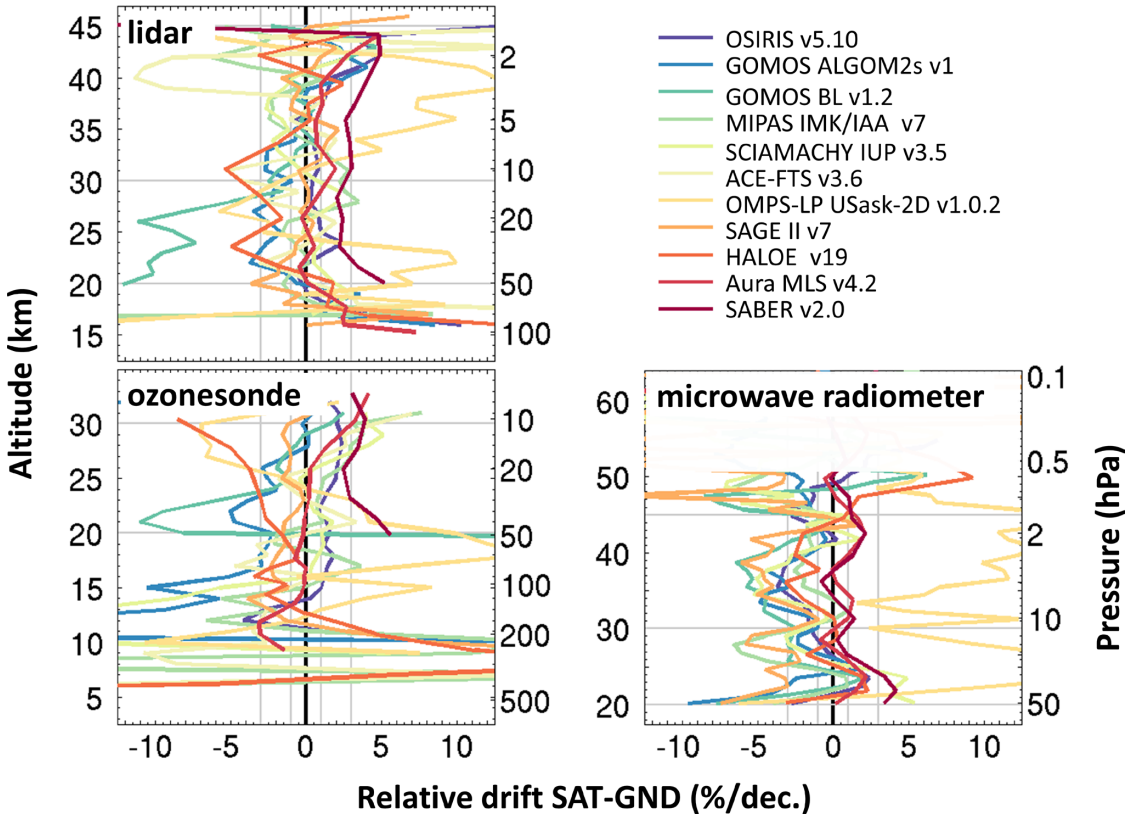

Figure 3.5: *Vertical profile of network-averaged satellite drift (Level-2) relative to co-located ground-based measurements by ozonesonde (bottom left), lidar (top left) and microwave radiometer (bottom right). Colours represent different limb/occultation data records (see legend).*

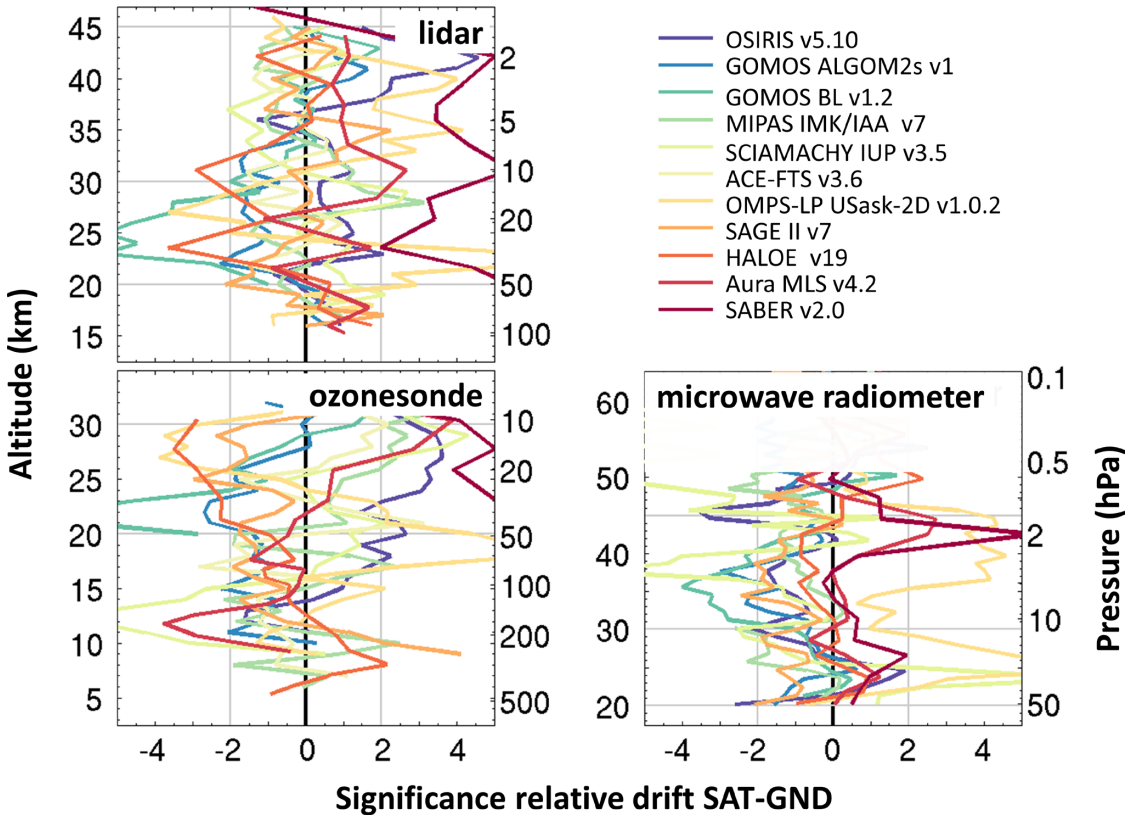

Figure 3.6: *As **Figure 3.5** but for the significance of the drift estimates. The 2σ detection threshold is indicated by grey vertical lines.*

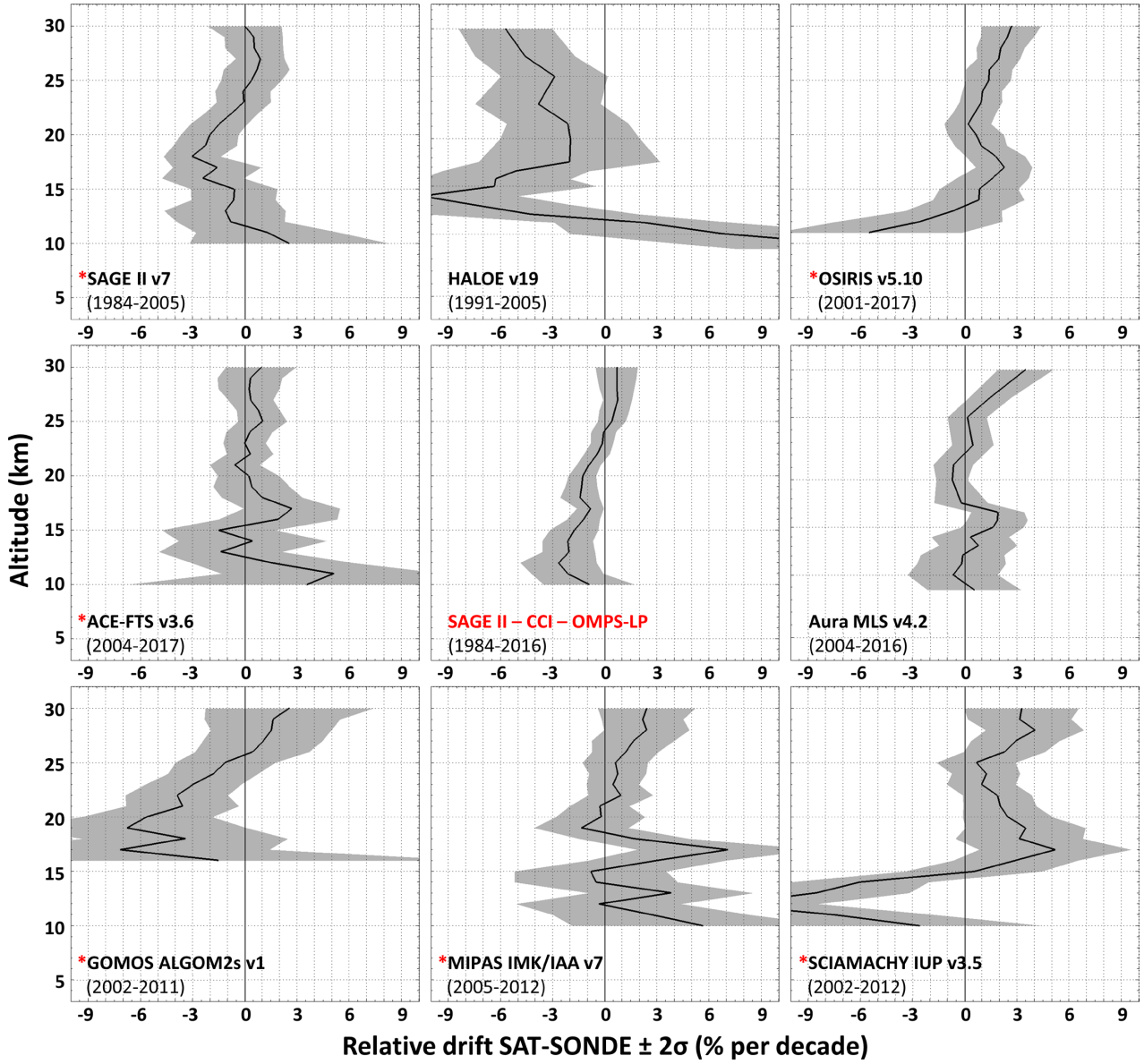

Figure 3.7: *Drift estimates and 95 % confidence interval of monthly zonal mean satellite data relative to the ground-based ozonesonde network. Eight limb/occultation records and the merged SAGE-CCI-OMPS (central panel) are shown. Satellite records contributing to the merged record are indicated with a red asterisk.*

The most striking result comes from OMPS-LP whose vertical drift profile oscillates between negative values (-6 % per decade) at 27 km and positive values around 18 km (+7 % per decade) and 40 km (+9 % per decade). These oscillations are clear from comparisons to each of the three ground-based data records. Even though the OMPS time series is only five years long, the ample co-location statistics allow for drift estimates with comparable precision to that of many other limb/occultation sounders. Instabilities in the altitude registration may be at the origin of the drift in the ozone record (*Moy et al.*, 2017; *Zawada et al.*, 2018; *Kramarova et al.*, 2018).

We stress that statistically insignificant results should not be blindly interpreted as instances where no drift in the data is guaranteed. This is because the adopted method, the available comparison statistics, and the quality of the ground-based reference data only allow us to probe satellite drift to levels that are comparable to (or larger than) the geophysical ozone profile trend expected since the mid-1990s. The lower bound to detect drift is at best 1 % per decade for just a few single-sensor records. Typically, detection thresholds are closer to 2–3 % per decade in the middle stratosphere and 3–4 % per decade elsewhere. These results do not necessarily apply to multi-sensor records since the very combination of different data sets will affect the resulting long-term stability.

The second analysis considers space- and time-gridded limb/occultation data (*i.e.*, Level-3) and gridded ozonesonde data. This approach takes advantage of the complete time series of satellite and ground-based records but at the cost of leaving mismatch or sampling uncertainty in the time series. However, the latter source of error becomes less important with increasing numbers of single profiles averaged by month in 5° latitude zones.

The time series analysis and derivation of drift follows the method described in *Hubert et al.* (2016), but the input data are different. Essentially, we consider the anomalies of monthly zonal mean data relative to its seasonal cycle in a reference period. Ozonesonde data are gridded following the procedure outlined in *Section 2.1.2.1* with slightly different selection criteria and reference period (2004–2011). Comparison time series are constructed as the absolute difference of the gridded satellite anomaly minus the gridded ozonesonde anomaly, with both terms expressed in percent. As before, drift is regressed for each latitude band that contains a ground-based record and then averaged over the entire latitudinal range of the sonde network. The central network-averaged drift estimates and 95 % confidence intervals are shown in **Figure 3.7**. These Level-3 results generally confirm what was observed in Level-2 results, but it appears that the precision of the Level-3 drift estimates is slightly better (*e.g.*, for ACE-FTS). The availability of the entire time series, instead of a subset of co-located measurements, contributes to this improvement, but the smaller impact of reducing station-to-station inhomogeneities helps as well. Indeed, the latter are mostly avoided by constructing the reference as an average of deseasonalised anomaly station data, where the seasonal cycle is derived individually from each sonde record. If ozone profiles at a certain site are, on average, 5 % higher, then this multiplicative bias will be present in both the monthly and seasonal cycle data, and, therefore not in the deseasonalised monthly relative anomaly data. This step brings the average level of the anomaly time series of all stations to zero over the reference period, which avoids artificial steps in the station-combined sonde time series where a measurement gap starts or ends for a particular station. A second benefit of this deseasonalisation procedure is that it reduces the variance in the comparison time series caused by differences in the seasonal cycle of satellite and sonde data.

The central panel of **Figure 3.7** also shows drift results for the merged SAGE-CCI-OMPS data record described in *Section 2.2.4.2* and by *Sofieva et al.* (2017). Non-significant, positive values of 0.8 % per decade are found between 25–30 km. The negative values of ~1.5 % per decade below 24 km are statistically significant, but we advise great care in interpreting significance in the lower part of the stratosphere. The variability of the ozone field, the lower ozone concentrations and the fading sensitivity of limb sounders make it very difficult to obtain precise uncertainty estimates in this part of the atmosphere. Comprehensive studies are therefore needed to further quantify these errors. For now, the result of lower stratospheric drift is inconclusive. Future work will consider other merged profile data records and extend the analysis to the lidar network data.

3.1.3 Intercomparisons of limb satellite measurements

In the context of data validation studies, intercomparisons of satellite measurements are typically performed using profile data that are co-located in space and in time. Many such analyses have been performed in recent years for limb sounders (*e.g.*, *Adams et al.*, 2013, 2014; *Laeng et al.*, 2014; *Kyrölä et al.*, 2013) including analyses of relative drifts (*Rahpoe et al.*, 2015). For the creation of merged data sets, on the other hand, the intercomparison of monthly zonal mean data is more relevant since the combination of different satellite records is usually done at the level of monthly zonal mean data or at the level of monthly deseasonalised anomalies. Such studies have been performed recently by the groups that created merged data sets from limb instruments (*e.g.*, *Bourassa et al.*, 2014; *Froidevaux et al.*, 2015; *Davis et al.*, 2016; *Sofieva et al.*, 2017). Results of intercomparison studies for the SBUV nadir profile sounders are reported in *Section 3.1.4*.

During the preparation of the merged SAGE-CCI-OMPS data set of ozone profiles, the deseasonalised anomalies of the individual instruments (SAGE II, GOMOS, MIPAS, SCIAMACHY, OSIRIS, ACE-FTS, and OMPS-LP USask 2D) have been extensively intercompared by computing and visualising the time series of the difference between the single-sensor anomalies and the median anomaly of the seven data records. This method is sensitive to detecting anomalous features (*i.e.*, large or increasing deviations from the median) in the time series of single sensors.

In particular, it was found that the deseasonalised anomalies for SCIAMACHY are larger at the beginning of the mission, for nearly all latitude bands and at many altitude levels (**Figure S3.7** in the supplement). Similarly, OMPS anomalies are lower in the first three months of the mission (**Figure S3.8** in the supplement). Note that the sampling of the OMPS-LP was significantly coarser in the first three months of the mission. The data from these early periods of SCIAMACHY and OMPS operation are therefore not included in the merged SAGE-CCI-OMPS data set. After the data selection, the anomalies from individual instruments are found to be in good agreement with each other. This is illustrated in **Figure 3.8,** which shows the deviations of deseasonalised anomalies of each instrument relative to the median anomaly of all limb records for 30°S–40°S. Deviations from the median anomalies are small, less than 5 % for the majority of data, and do not have statistically significant drifts with respect to the median anomaly (see also illustrations in the Supplement of *Sofieva et al.*, 2017).

3.1.4 Stability of limb data records relative to ground-based networks

The two nadir instrument-based merged ozone data sets used in this Report (SBUV MOD and SBUV COH; see *Chapter 2*, *Sections 2.2.1.1* and *2.2.1.2*) are constructed from the same initial set of SBUV satellite records. The SBUV series of instruments have similar design and data are retrieved using the same Version 8.6 algorithm (*McPeters et al.*, 2013). Furthermore, as part of the Version 8.6

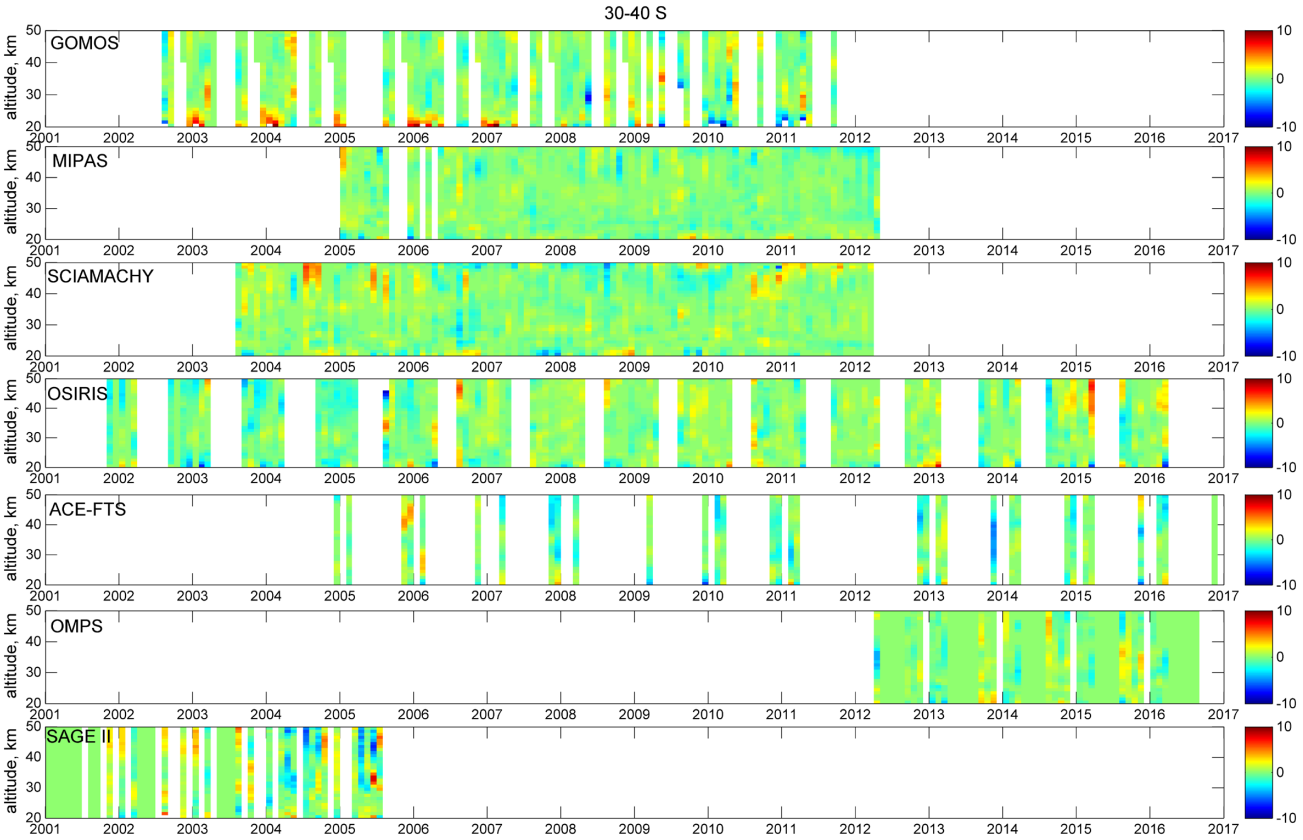

Figure 3.8: *Deviations (in %, colour) of deseasonalised anomalies for GOMOS, MIPAS, SCIAMACHY, OSIRIS, ACE-FTS, OMPS, and SAGE II (indicated in the panels) from the median deseasonalised anomalies computed using all data sets. Latitude band is 30°S–40°S. From Sofieva et al. (2017).*

processing, SBUV measurements from individual instruments were inter-calibrated at the radiance level based on comparisons during instrument overlap periods (*DeLand et al.*, 2012). However, despite the instrument similarity and common retrieval algorithm, each instrument experienced unique operational conditions (*e.g.*, instrument degradation or specific on-orbit problems) and orbital characteristics (including measurement time of day), which contribute to differences among the individual records. Therefore, differences in how the data are selected and merged in the combined records can lead to differences between the merged SBUV products.

In general, the SBUV ozone profile measurements agree to within ±5 % when compared to external satellite and ground-based instruments, with similar or better agreement among the SBUV instruments themselves (*Kramarova et al.*, 2013a; *Frith et al.*, 2017; *Wild et al.*, 2019). However, lower quality data from NOAA-9, NOAA-11 descending, and NOAA-14 lead to larger uncertainties (10–15 %) in the mid-1990s and complicate efforts to establish a long-term calibration over the full record (from 1980s to 2000s) (*DeLand et al.*, 2012; *Kramarova et al.*, 2013a; *Tummon et al.*, 2015; *Ball et al.*, 2017). The SBUV MOD merging approach is to average all available data after removing portions of individual records found to be inferior (*e.g.*, data from drifting orbits, NOAA-9 SBUV/2). This approach relies on the average of multiple measurements to mitigate the effects of small offsets and drifts in individual data sets rather than attempting to choose a single record as a reference calibration. The SBUV COH merging approach is to identify a representative satellite for each time period, thus preserving knowledge of orbital characteristics for each measurement period. Additionally, data after 2001 are adjusted directly to NOAA-18 in SBUV COH, removing small inter-satellite differences. Each approach has advantages and disadvantages. SBUV MOD is sensitive to successively increasing or decreasing biases in the instrument series that might alias into a trend. SBUV COH is sensitive to drifts in the reference instruments that might be propagated to other periods in the record. This was the case in the previous version of SBUV COH used in the SI2N report (*Tummon et al.*, 2015). The potential for unphysical drifts is greatly reduced in the current version of the SBUV COH data set, which limits inter-instrument corrections to periods where long overlaps of high quality data exist.

Frith et al. (2014; 2017) analysed the differences in monthly zonal mean time series from the individual SBUV data sets during periods of overlap in an effort to characterise the uncertainty associated with the merging process. Given the numerous instruments and overlaps, many reasonable approaches could be chosen based on different selections of data (*e.g.*, instrument and time period) and different means of determining inter-satellite adjustments (*e.g.*, mean offset, offset and drift, no adjustments).

The authors used the distribution of measured offsets and drifts between SBUV instruments during times of overlap to construct 10000 MC simulations of potential instrument error (see *Frith et al.*, 2017, **Figure 7**). In essence the collection of SBUV inter-instrument offsets and drifts were used to define an SBUV-system uncertainty, in an effort to account for both relative and absolute uncertainties. The MC simulations were structured to account for the larger observed uncertainty of instruments operating in the 1990s and time dependence of the absolute calibration procedures used within the SBUV retrieval algorithm (*DeLand et al.*, 2012). Previous studies using MC simulations suggest that the long-term drift uncertainty in a record constructed from multiple data sets is less than that for a single instrument because the introduction of new data "resets" the drift (*Stolarski and Frith*, 2006; *Weber et al.*, 2016), but adding too many data sets increases uncertainty as a result of multiple potential discontinuities in the record (*Weber et al.*, 2016). By applying the multiple regression model to the MC simulations we can test the degree to which potential time-dependent uncertainties alias into individual regression terms and assess the additional uncertainty due to the merging process itself. The 2-sigma variation of terms from a regression model fit to the MC simulations defines the "merging uncertainty" for each term.

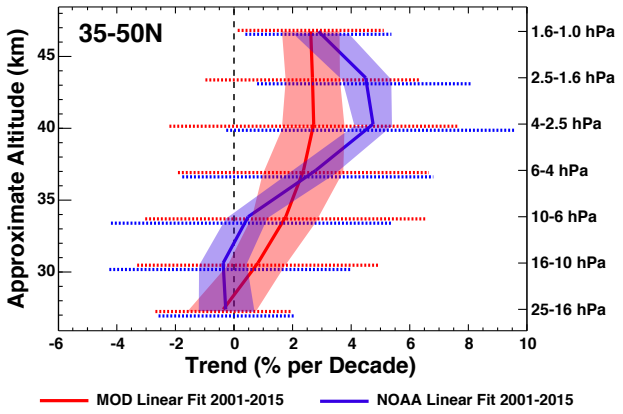

Figure 3.9: *ILT trend proxy fit to 35S–50S monthly zonal mean SBUV MOD (red) and SBUV COH (blue, referred to as "NOAA" in Figure) records over the 2001–2015 time period. The shaded regions indicate the 2-sigma statistical uncertainty estimated from the unexplained variability in the multiple regression analysis. The dotted error bars show the total trend uncertainty when the SBUV MOD 2-sigma merging uncertainty is included. The uncertainties are combined using the root sum of squares of each error term. For comparison, the estimated MOD uncertainty is also added to the SBUV COH error bars. From Frith et al. (2017).*

Frith et al. (2017) also compared regression analysis results between the SBUV MOD and SBUV COH data sets, treating each as equally valid approaches to merging the data record from the SBUV instrument suite. The authors report differences in the post-2000 trend with SBUV COH trends being generally more positive than SBUV MOD at altitudes above the 5 hPa level and less positive below 5 hPa, consistent with the results of this Report (*e.g.*, **Figures 5.1** and **5.2**). When only statistical error is included the results are statistically significantly different from each other, but when the merging uncertainty is taken into account the trend error bars overlap (see **Figure 3.9**). Direct comparisons between both data sets show the differences below 5 hPa are largely a result of a positive bias in NOAA-19 at the end of the record, which is adjusted in SBUV COH but not in SBUV MOD (**Figures 2** and **10** of *Frith et al.*, 2017). Above 5 hPa, a small positive drift in NOAA-18, used as the reference in SBUV COH, leads to a more positive trend relative to SBUV MOD (**Figures 3** and **10** of *Frith et al.*, 2017). **Figure 3.10** shows the annual drift (percent per year) relative to Aura MLS v4 for SBUV MOD and SBUV COH computed from October 2004 to December 2016. As described above, comparisons using Aura MLS as a transfer standard show SBUV COH with a more positive drift above 5 hPa and SBUV MOD with a more positive drift below 5 hPa.

3.1.5 The BASIC composite and its use for intercomparisons of merged data records

All merged (hereafter also called composite) data sets suffer from artefacts and/or drifts inherent to the instrument data used in their construction, or from absolute offsets and discontinuities when instrument data are combined. As has been extensively discussed in previous sections, the presence of these artefacts can lead to inaccurate and/or more uncertain trend estimates. The BAyeSian Integrated and Consolidated (BASIC) composite is a set of algorithms that, using only the data available, collectively merges multiple ozone composites into one. This Bayesian approach provides a principled way to incorporate prior information about data artefacts (see below), with a Gaussian mixture likelihood that together allows for a robust estimate of true ozone given both the information available and the design of the statistical model (see *Ball et al.* (2017) for details). The approach is designed to take advantage of the common variability present in all the ozone composite data sets to inform, within a probabilistic framework, the most likely ozone time series. Since each composite contains both the real ozone time series and additional composite- and instrument-specific artefacts such as drifts, spikes and discontinuities, the availability of multiple co-temporal and co-spatial time series allows BASIC to account for many of these issues that might remain in any individual ozone composite (*i.e.*, sampling differences, satellite drifts, biases between data sets merged in the composites, and resolution differences between instruments within composites).

The BASIC approach is thoroughly documented in *Ball et al.* (2017), but we briefly describe the steps here. First, errors provided with each composite are formed using different approaches and statistics and therefore cannot be directly compared. Thus for BASIC, the uncertainties for each composite time series are derived independently using singular value decomposition (SVD). The underlying assumption for using SVD is that each composite contains the true ozone time series and a set of

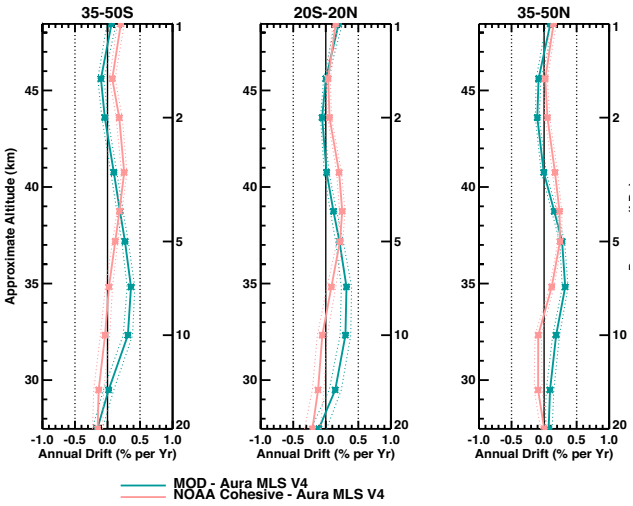

Figure 3.10: *Drift (in % per year) of zonally averaged profile data from SBUV MOD (turquoise) and SBUV COH (red) relative to Aura MLS v4 for 50°S–35°S (left), 20°S–20°N (middle) and 35°N–50°N (right).*

instrument and composite-construction artefacts that should be unique to each composite. Through SVD, the common modes of variability are identified, and if behaviour deviates from these then this is assigned as an uncertainty; ignoring the first SVD mode as the one common to all composites, the higher modes that describe deviations from the first mode are used to form an uncertainty: (see discussion in *Section 3.1.1* about the benefits of multiple datasets to identify and assign the source data causing a discrepancy). Use of SVD does not provide true uncertainties, but assigns an uncertainty based on common behaviour where a deviation of a composite, or group of composites, from the common behaviour leads to an estimated larger uncertainty. Indeed, enhanced uncertainties often correspond to known problems in specific composites or underlying instrument data, so *Ball et al.* (2017) consider this a reasonable assignment of belief, or uncertainty, in the accuracy of the composites. Often the spread in the ozone composites is well within the uncertainties provided with each composite. Second, any prior information about artefacts or drifts in the individual instrument or composite data are incorporated by inflating the SVD-estimated uncertainties; such information can be at times when instruments are known to change in each composite (leading to step-function changes in the time series), or when orbital drifts are known to induce an artificial trend in the time series (*e.g.*, some SBUV instruments in the later 1990s; see *Section 3.1.4*). We note that the choice of using an inflation factor of two is subjective, but we see little difference in the impact of choosing larger values (see *Ball et al.*, 2017). Third, we then form a Gaussian mixture likelihood for each month that allows for the probability distribution of ozone to form multiple peaks. As such, combining information about the most likely state f the ozone in the current month with information available in the preceding and following months leads to a posterior distribution that provides the most likely ozone time series given the information available. To form the posterior estimate of the monthly ozone, we must sample what is a high-dimensional problem using an efficient method such as Markov chain Monte Carlo (MCMC), which we do using Hamiltonian Monte Carlo (HMC; *Neal,* 1993; *Carpenter et al.*, 2016).

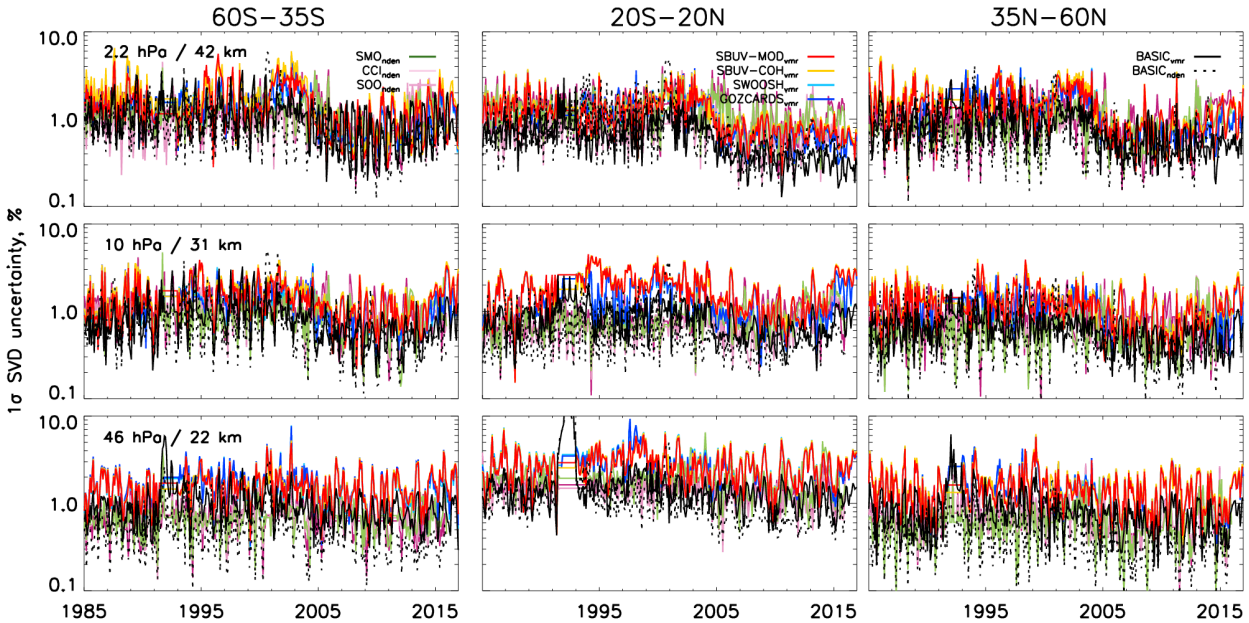

Figure 3.11: *Latitude weighted mean 1-sigma errors (%) estimated from the application of SVD for three number density composites (SAGE-MIPAS-OMPS (SMO), SAGE-CCI-OMPS (CCI), and SAGE-OSIRIS-OMPS (SOO)) and the 1-sigma uncertainty in the $BASIC_{NDEN}$ composite (dotted black line) derived from these, and for four VMR-based composites (SWOOSH, GOZCARDS, SBUV MOD, and SBUV COH) and the $BASIC_{VMR}$ composite (solid black line). Note that number density is on altitude, and VMR on pressure level, so comparing between the VMR and number density data sets is only indicative.*

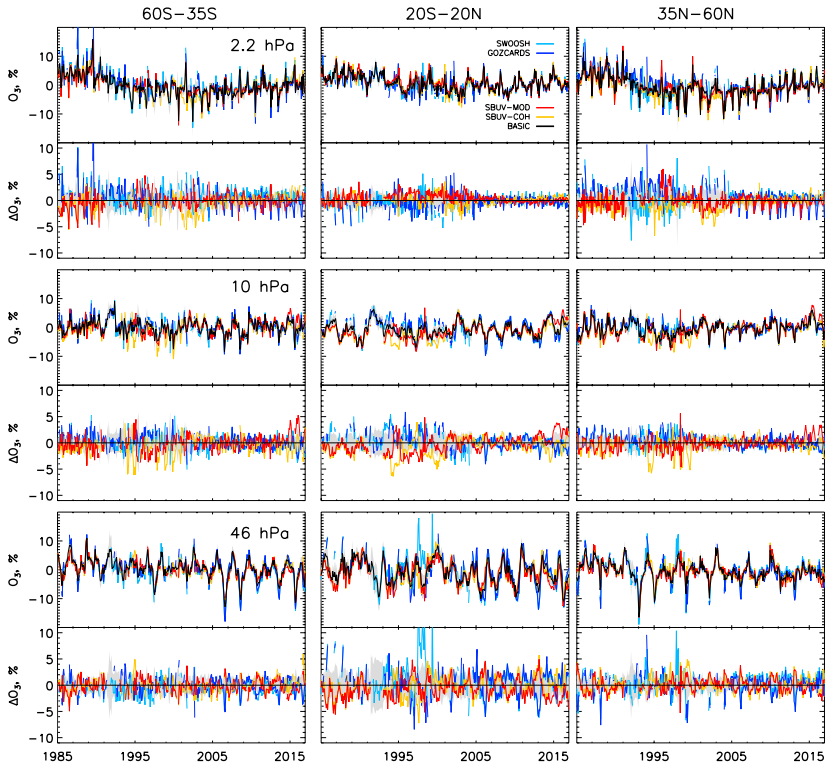

Figure 3.12: *Selected pressure levels in three latitude bands for the four VMR composites and the BASIC$_{VMR}$ composite. Each pair of plots show the relative (%) deseasonalised time series bias-shifted to agree with SWOOSH for the July 2005 to December 2013 period (upper half) and anomalies relative to the BASIC$_{VMR}$ composite (lower half). The 2-sigma uncertainty on the BASIC$_{VMR}$ is shown with grey shading.*

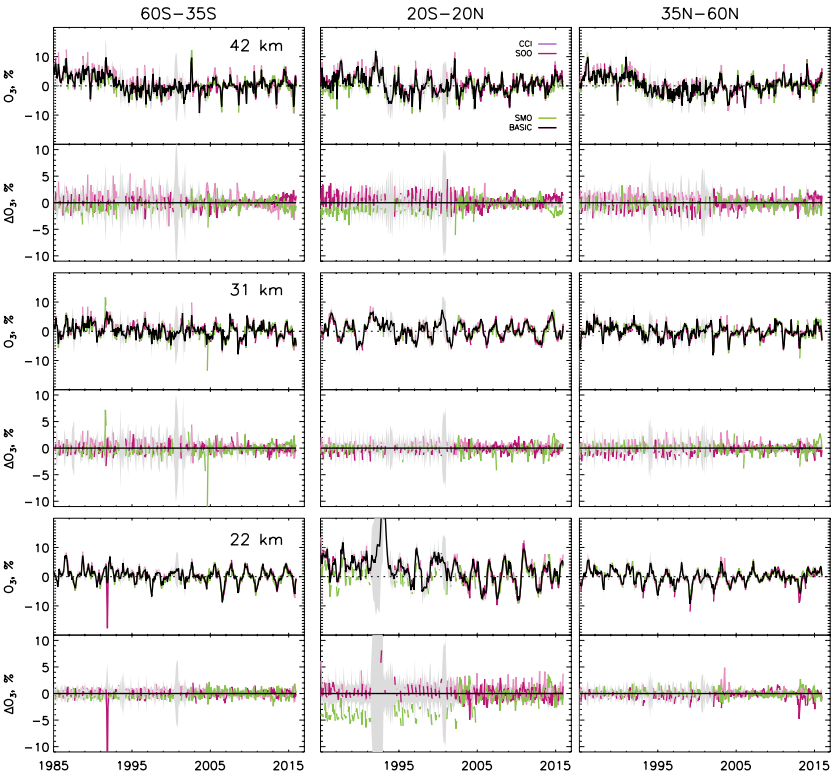

Figure 3.13: *As for **Figure 3.12** but for the number density composites and BASIC$_{NDEN}$ derived from these.*

Examples of the SVD-estimated uncertainties are shown in **Figure 3.11**. These uncertainties are typically at least two times smaller than those provided with composites (where available), which might suggest that provided uncertainties are conservative given that the SVD-uncertainties represent a deviation of the estimated mean uncertainty from the group. Here we have separately estimated uncertainties in four VMR-based composites and three number density-based composites; the uncertainties are not comparable between the sets of composites, only within. Nevertheless, it is interesting to compare all seven in the same sub-figures. For example, in the VMR-based composites it is clear that the uncertainties in the 20°S–20°N region at 10 hPa and 46 hPa are larger between 1994 and 2001; at 10 hPa this is mainly due to the SBUV MOD and SBUV COH composites and reflects a period of known drift in the SBUV-based composites (see *Section 3.1.4*). The number-density uncertainties are typically lower than the VMR-composites, but this reflects the fact that the number density composites are based on very similar underlying data with similar vertical resolutions. The VMR-based uncertainties integrate information from the lower resolution SBUV-based composites, with those at a similar resolution to the number density composites and, as such, the uncertainties are larger to accommodate the higher uncertainty in the absolute level. Nevertheless, the uncertainties in BASIC$_{VMR}$ (based on the four VMR-composites: GOZCARDS, SWOOSH, SBUV MOD, and SBUV COH; solid, black line) and BASIC$_{NDEN}$ (based on the three number density composites: SAGE-CCI-OMPS, SAGE-OSIRIS-OMPS, and SAGE-MIPAS-OMPS; dotted, black) in **Figure 3.11** are similar. Note that sometimes the BASIC$_{NDEN}$ uncertainty becomes temporarily large, which occurs because of missing data in all the underlying composites; this can also be seen during the Mt. Pinatubo eruption, particularly over the Equator at 46 hPa. What appears to be common to all regions presented is that uncertainties prior to 2005 are larger than after this time, especially between 1995 and 2000, which may have a significant effect on the estimated decadal

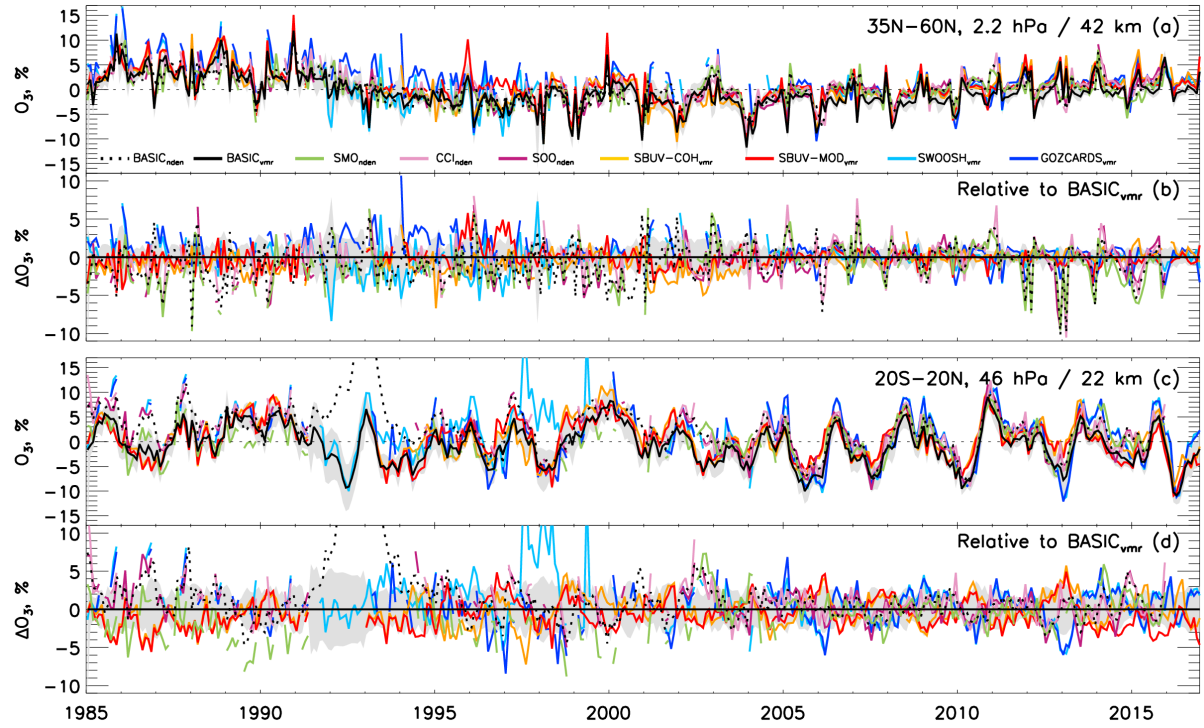

Figure 3.14: *Two levels from* ***Figures 3.12*** *and* ***3.13*** *overlaying the VMR and number density time series for comparative purposes. Each pair of plots show the relative (%) change compared to the July 2005 – December 2012 mean (a & c) and the change relative to BASIC$_{VMR}$ (b & d). Note that while number density and VMR time series shown correspond to approximately the same region in the atmosphere, they are not exact and should be considered only indicative.*

trend (see below) especially since this larger uncertainty results directly from different offsets in the composites. Another interesting feature, apparent in many regions, is that the uncertainties begin to increase again after 2014, which suggests composites are diverging again. This divergence may lead to an inflation in uncertainties in trend estimates even though more data are being accumulated. Understanding why data are diverging again and which are most likely correct will be an important issue to follow up in future work.

Figure 3.12 presents the four VMR-based composites; (upper) relative to the mean of July 2005 to December 2012 and (lower) relative to BASIC$_{VMR}$ for the same regions shown in **Figure 3.11**. Similarly, **Figure 3.13** provides the same for the three number-density composites and BASIC$_{NDEN}$. Panels in **Figures 3.12** and **3.13** represent approximately the same pressure/altitude region in the stratosphere, and the similar short- and long-term variability reflects this. As discussed above, the relative differences between composites (lower panels) are typically larger for the VMR composites than the number density and are larger at higher pressures (lower altitudes), which reflects the decreasing vertical resolution in SBUV. The number density composites contain mostly similar data, so while there are clearly periods of drift and rapid divergence (**Figure 3.12**) between the SWOOSH/GOZCARDS composites (constructed using similar instruments) and the SBUV-composites (especially for 1994–2004 and prior to 1990), the offsets between composites are clearer in the number density composites (lower panels of **Figure 3.13**) because the underlying data are usually the same (SAGE II until 2005, OMPS from 2011). So, if drifts exist in these number density-based data they will not become apparent, and differences will mainly reflect the differences in composite merging or data-screening prior to the merge. This is especially apparent in the SAGE-MIPAS-OMPS data set where, at 22 km in the equatorial region, this data set is offset by 4 % from the other two number density composites, and the BASIC$_{NDEN}$ composite rejects this as an unlikely level of ozone prior to 2004. This is a good example of how the BASIC method goes beyond a simple composite average. Offsets are seen in other panels mainly associated with the SAGE-MIPAS-OMPS composite, and it is these offsets that likely contribute to the larger positive post-2000 decadal trends in SAGE II-MIPAS-OMPS presented in the tropical lower stratosphere presented in *Steinbrecht et al.* (2017) and in *Section 5.1* of this Report (**Figure 5.2**). Once again, the divergence between composites after 2014 appears in several panels in **Figures 3.12** and **3.13**.

We note that the BASIC composites can only estimate ozone as accurately as the information available within the considered composites can permit, that is if all the data are wrong in the same way, at the same time, BASIC cannot estimate the true state of ozone at that time. Therefore, in principle, the more composites that can be incorporated in the analysis, the more useful and robust the result should be. However, the BASIC$_{VMR}$ and BASIC$_{NDEN}$ composites show different variability on monthly and approximately two-yearly timescales because they use different data-types (sources, resolution, vertical grid, units, *etc.*).

In **Figure 3.14**, we make this clear by showing all seven composites and both BASIC composites from two of the panels presented in the preceding figures (35°N–60°N at 2.2 hPa/42 km and 20°S–20°N at 46 hPa/22 km); the lower panels are all relative to the $BASIC_{VMR}$ composite, and the difference between $BASIC_{VMR}$ and $BASIC_{NDEN}$ is clear (solid and dotted black lines, respectively). There is some sensitivity to the exact altitude/pressure level compared in these plots, but choosing an altitude above or below the ones presented leads to broadly similar results. It is important to note that prior to 2004, SAGE II data are in all composites except SBUV COH and SBUV MOD. Therefore, applying the BASIC method to all seven composites treated independently would lead to a bias in the composites containing SAGE II, a concern raised by *Harris et al.* (2015). It would also require a transfer function through a model (*e.g.*, European Center for Medium-Range Weather Forecast Re-Analysis (ERA-Interim)) to put all the composites on the same coordinate system and therefore applying such a coordinate change introduces additional uncertainties (*e.g.*, due to unknown uncertainties in the parameters of the reanalysis models) not considered explicitly in the uncertainty estimate. As such, because $BASIC_{NDEN}$ is based only on the SAGE II composites, and the $BASIC_{VMR}$ equally between those with and without SAGE II data, it is unsurprising that the two BASIC composites show differences prior to 2004, with $BASIC_{NDEN}$ lining up closer to the GOZCARDS and SWOOSH composites that contain SAGE II.

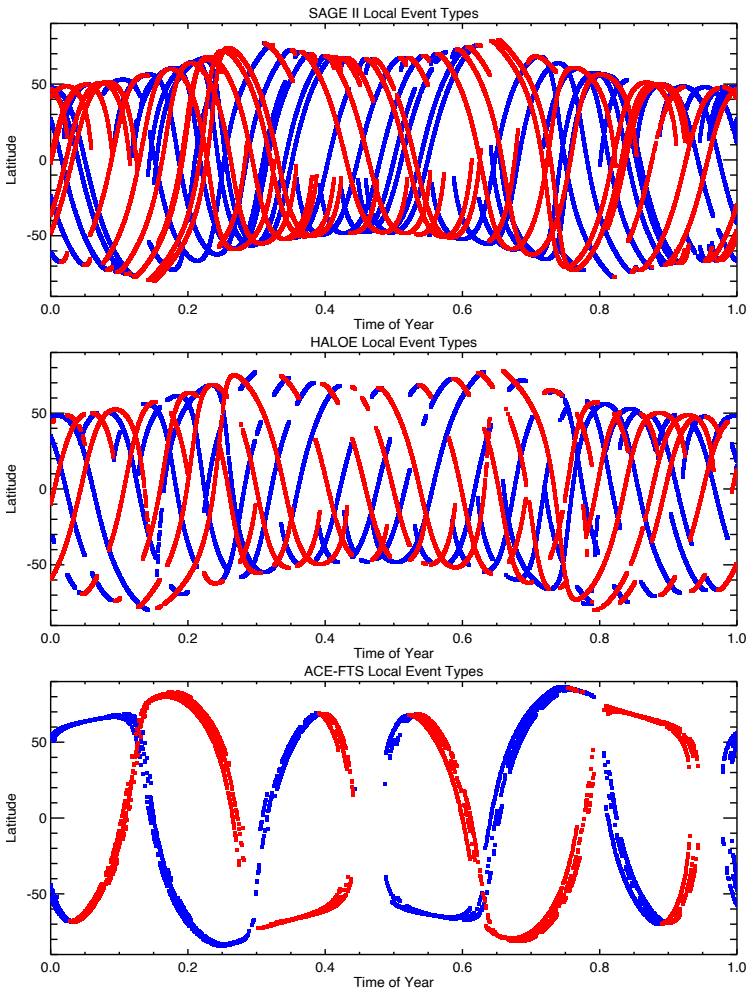

Figure 3.15: *Latitude and time of year of all events for SAGE II, HALOE, and ACE-FTS separated by local event type (blue for sunrise and red for sunset) plotted every 3 years (to reduce clutter) illustrating the drifting sampling patterns over time. Sampling patterns can systematically shift several weeks over a few years for instruments like SAGE II (in its later years) or HALOE (continuously) while ACE-FTS is essentially constant. Time of year is expressed as the modulus of the year fraction.*

A more thorough and detailed analysis of the $BASIC_{VMR}$ composite presented in *Ball et al.* (2017, 2018), along with a trend analysis, reveals that after accounting for many of the artefacts and drifts in the data, differences in post-1997 ozone change profile shapes that have been presented in previous studies (*e.g.*, *Tummon et al.*, 2015) disappear, and the trends in both hemispheres look similar, suggesting that artefacts are indeed important in biasing trend estimates. The enhanced uncertainty between 1995 and 2005, also presented here, shows that trends may also be sensitive to the inflection date used in piecewise linear trend (PWLT) multiple linear regression (MLR, see *Section 4.3.4*). Additionally, *Ball et al.* (2017) found that the use of MLR led to a large, post-1997 negative trend in the tropics (20°S–20°N) peaking at 7 hPa (also reported elsewhere), which disappeared in BASIC when applying dynamical linear modelling (DLM) to estimate the change.

The use of the BASIC method and composites, and comparisons similar to the limited set presented here, should aid composite teams in understanding artefacts in the composites and improving the merging procedure. Revealing artefacts allows for evidence of the reliability of composites to inform our understanding of why composites show diverging decadal trend estimates and to make informed decisions about how these artefacts might be addressed in the future. Ultimately, a more advanced approach is to merge the instrument data underlying each composite only once, using a methodology adapted from that developed in BASIC, which will lead to a single composite that provides the best estimate of ozone given all the satellite data available.

3.2 Sampling bias and uncertainty correction characterisation

WMO (2014) identified three factors that were not accounted for in trend analyses with a potential major impact on resulting trends: Diurnal variability of ozone, biases between data sets, and long-term drifts between data sets. However, there is an additional complication that is intricately tied to these three factors in trend analyses, namely the non-uniform temporal, spatial, and diurnal sampling

of the different instruments used for those analyses. This non-uniform sampling can have a detrimental impact not only on the regression techniques used to derive long-term trends in ozone but also on other analyses performed to determine diurnal variability or the magnitude of potential biases and drifts between data sets.

In order to perform regression analyses to determine long-term ozone trends, data sets are first typically reduced to monthly zonal mean (MZM) values that are utilised as though they are representative of the centre of the month and the centre of the latitude bin. While this assumption is reasonable for highly sampled data sets (*e.g.*, nadir and limb scatter measurements) it generally breaks down when applied to sparsely sampled data sets (*e.g.*, ground or occultation measurements), though even highly sampled data sets are susceptible to changes in the local solar time of observations that can be problematic in the presence of diurnal variability (*Bhartia et al.*, 2013, *Frith et al.*, 2017). This is not a new concept; T*oohey et al.* (2013) and *Sofieva et al.* (2014) both investigated non-uniform temporal sampling as an added source of noise and uncertainty that could be characterised and included in trend analyses. However, orbital drift can lead to a systematic drift in sampling patterns over time, making the standard practice of using deseasonalised anomalies for trend analysis insufficient to remove potential sampling biases. *Millán et al.* (2016) investigated the impacts of non-uniform sampling biases on resulting trends from different instruments by repeating a "representative year" of sampling for each data set and running a model through it over ~30 years to analyse the effect on trends. While illustrative, this did not account for the actual sampling bias as it changed from year to year for those instruments. As such, it is still necessary to consider the non-uniform sampling of different satellite data sets and how representative the derived MZM ozone values are of the actual month and zonal band. Data from ground-based observations can exhibit similar problems and have the additional complication of making measurements from only a single latitude and longitude such that one must also consider their representativeness of the zonal band itself.

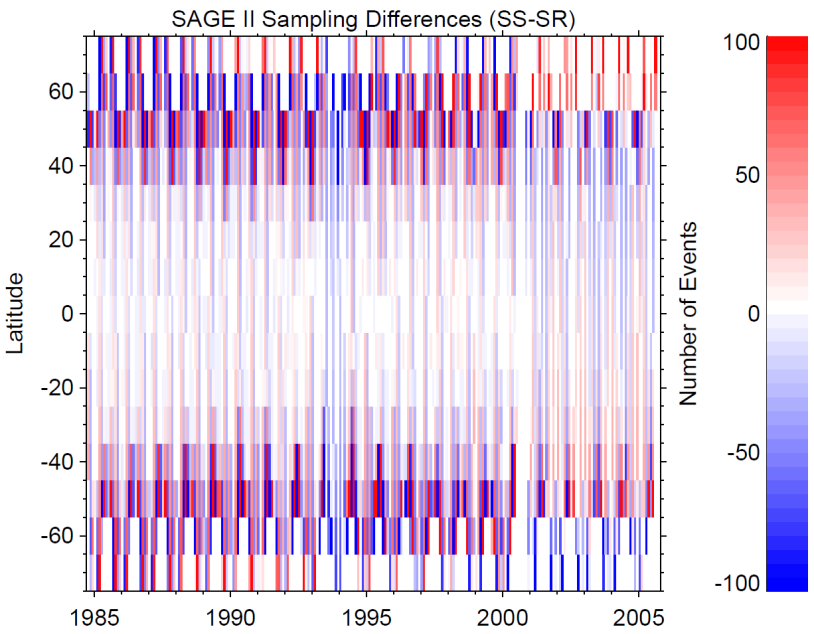

Figure 3.16: *The difference in the total number of sunset (SS) and sunrise (SR) events in each month and 10 degree latitude bin from SAGE II. In addition to the rapid oscillation between SR and SS dominated months, instrument anomalies resulted in large periods and locations of SR/SS dominated sampling (bottom panel of* ***Figure 8*** *of Damadeo et al., 2018).*

3.2.1 Sampling bias for occultation instruments estimated using simultaneous temporal and spatial (STS) analysis

Occultation instruments are classic examples of non-uniform temporal sampling drift. **Figure 3.15** illustrates the drifting sampling patterns of three occultation instruments (*i.e.*, SAGE II, HALOE and ACE-FTS) over their mission lifetimes. SAGE II and HALOE exhibit a systematic drift in sampling towards earlier times of year at every latitude over their mission lifetimes. ACE-FTS has a much slower drift but has a non-uniform distribution of sunrise and sunset measurements as a function of time of year for every latitude, making the separation of the seasonal cycle and diurnal variability impossible when computing monthly zonal means. In addition to non-uniform seasonal sampling, occultation instruments can also exhibit non-uniform diurnal sampling over the mission lifetime. SAGE II is an extreme case of this, where instrument anomalies caused long periods of sunrise or sunset dominated sampling (**Figure 3.16**). In the presence of diurnal variability (*e.g.*, in the upper stratosphere where trends are of the largest magnitude) these diurnal sampling biases can detrimentally impact trend analyses.

Damadeo et al. (2018) discusses the non-uniform temporal, spatial, and diurnal sampling of occultation instruments in great detail and how the use of MZM values can create sampling-induced biases that alias into long-duration variability (*i.e.*, solar cycle and/or long-term trends). Ultimately where (*i.e.*, at what latitudes and altitudes) the sampling biases alias into trend and/or solar cycle results is somewhat "random" as it is dependent upon the chance combination of drifting sampling patterns, spatially-varying seasonal gradients, and frequency of interannual variability. That work also utilises a simultaneous temporal and spatial (STS) regression technique that properly accounts for the non-uniform sampling patterns of occultation instruments (*Damadeo et al.*, 2014) and applies it to the SAGE II, HALOE, and ACE-FTS data sets simultaneously to derive trend results unaffected by sampling biases.

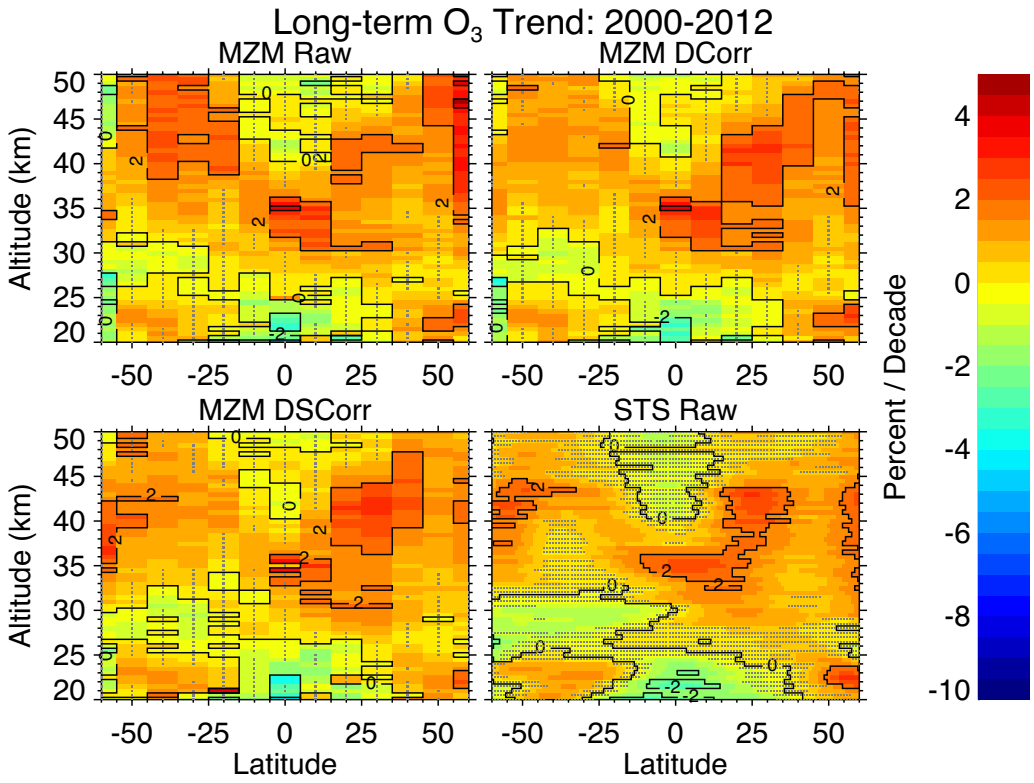

Figure 3.17: *Long-term trends derived from both the MZM and the STS regressions during the potential recovery period. Results are also shown when using the STS regression results to create a diurnally corrected (DCorr) and a diurnally & seasonally corrected (DSCorr) data set for use with the MZM regression. The diurnal correction has the greatest influence on the upper stratosphere while the seasonal correction has the greatest influence at higher latitudes. Stippling denotes areas where the trend results are not significant at the 2σ level. Contour lines are plotted at 2 % intervals. (**Figure 11** from Damadeo et al., 2018).*

Lastly, in an effort to quantify the impact of non-uniform sampling on derived trends when using MZM methodology (*i.e.*, regressing to MZM values separately for each latitude bin), *Damadeo et al.* (2018) uses the results of the STS analysis to create diurnally as well as seasonally corrected versions of these data sets for use with MZM analysis. Each version (*i.e.*, raw, diurnally corrected, and diurnally plus seasonally corrected) is then run through an MZM regression model to derive long-term ozone trends. **Figure 3.17** illustrates the difference in trend results derived between the different "corrected" data sets. The diurnal correction exhibits the largest influence, showing differences in trend results of about 1–2 % per decade in the upper stratosphere at mid-latitudes (*i.e.*, where typical positive trends are largest). The seasonal correction has the largest influence at high latitudes and at the tropical middle stratosphere although at a reduced magnitude of about 0.5–1 % per decade.

Since typically derived "recovery" trends are only about 2–3 % per decade, the influence of non-uniform sampling patterns on derived trends can be significant and is strongly dependent upon what data sets are used and how they are incorporated into the analysis.

3.2.2 Station means versus zonal means

This section focuses on the question of whether monthly averaged ozone partial columns in the middle and upper stratosphere at single lidar stations are representative of the monthly zonal mean. The motivation for this dedicated analysis arises from the wide use of zonal means in calculating trends and in studies related to the interannual ozone variability in cross sections of the middle and upper stratosphere.

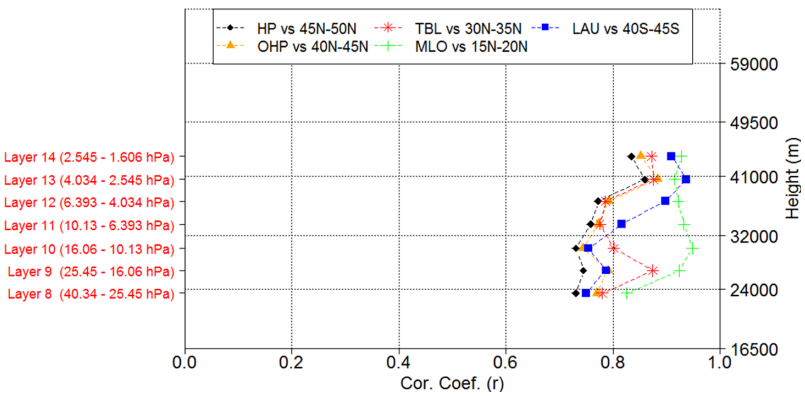

Figure 3.18: *Correlation between monthly mean SBUV overpass data at five lidar stations versus the corresponding 5° monthly zonal mean SBUV data of each site.*

We performed a comparison of lidar station overpass SBUV MOD data and zonal mean SBUV MOD data. A comparison of zonal mean SBUV MOD data to the station mean data by the lidar instrument itself would introduce uncertainty due to the use of different instruments. Five lidar stations with long-term ozone profile data records were chosen: Hohenpeissenberg (47.8°N, 11.0°E), Haute Provence (43.9°N, 5.7°E), and Table Mountain (34.4°N, 117.7°W) in the northern mid-latitudes; MLO (19.5°N, 155.6°W) in the tropics; and Lauder (45.0°S, 169.7°E) in the southern mid-latitudes. Furthermore, the analysis is confined to SBUV layers 8 (40–25 hPa) up to 14 (2.5–1.6 hPa), because the accuracy of the lidar data is limited in the upper stratosphere and that of the SBUV data is limited below the 30 hPa level.

Figure 3.18 shows that SBUV overpass data at the five selected lidar locations are highly correlated with the respective SBUV zonal mean data. Natural oscillations (seasonal, QBO, *etc.*) were removed prior to computing the correlation but not when computing the long-term trends. Correlation coefficients increase with altitude from about 0.75 to 0.9 for all sites and these values are statistically significant. This finding is of particular importance, especially when it comes to the calculation of long-term trends. It suggests that the variability at a single point in the middle and upper stratosphere is comparable to that found in the 5-degree zonal mean data encompassing the lidar station location. This implies that higher frequency spatial variability has little impact at these altitudes, making the derived trends from station data and satellite zonal mean data more directly comparable.

Although the level of agreement between ozone variability at single stations and from zonal means encompassing the stations has yet to be quantified (*WMO,* 2014; *Frith et al.,* 2017; *Zerefos et al.,* 2018), **Figure 3.19** shows an example of the spatial distribution of the correlation coefficients between SBUV overpass data at Hohenpeissenberg and at 633 station locations around the globe. The SBUV MOD data at station locations were downloaded from **https://acd-ext.gsfc.nasa.gov/anonftp/toms/sbuv/MERGED**. The "zonality representativeness" is obvious in the chromatic scale of **Figure 3.19** as well as the finding that as we move higher in altitude, higher correlations are found even at distances exceeding 1000 km. The results are similar when the calculations are repeated between overpasses over the other four available lidar stations and all 633 locations of SBUV overpasses (not shown here, see *Zerefos et al,* 2018). Overall, stations correlate well and are representative over a fairly wide range of longitudes and latitudes. These findings are also true for MLO but, as mentioned in *Section 5.4*, MLO Umkehr cannot represent the tropical belt between 20°S and 20°N. Instead, according to these findings, MLO represents the northern zone well between 15–20°N.

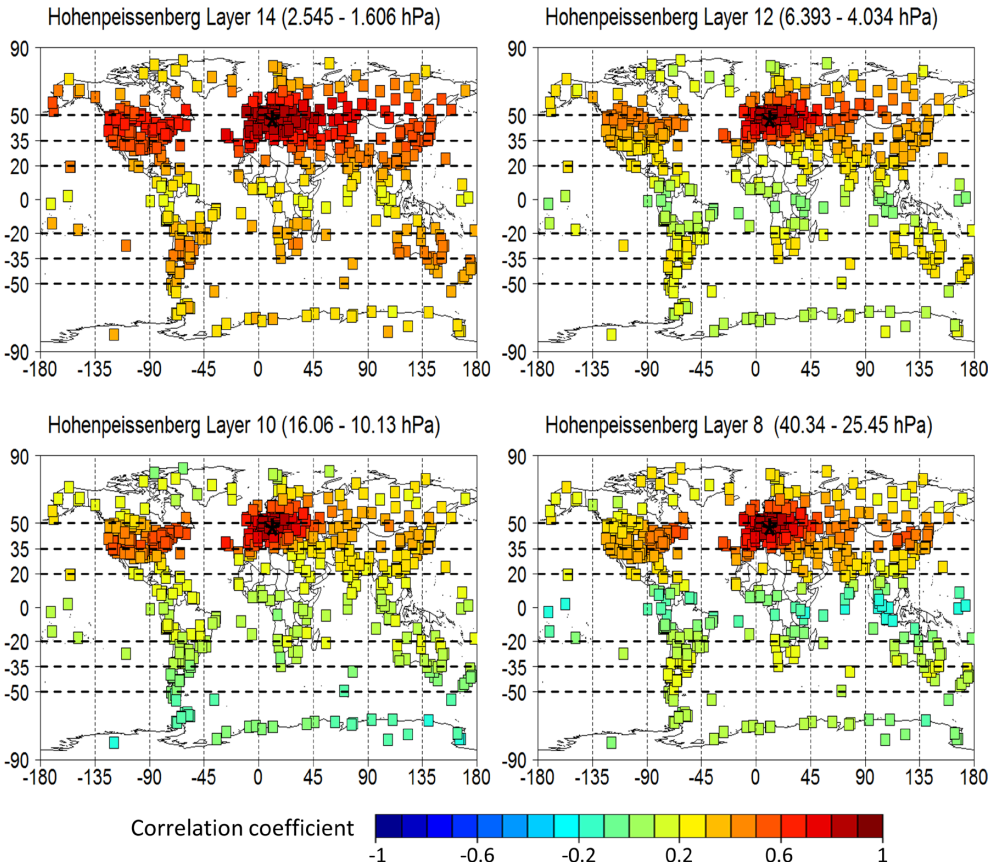

Figure 3.19: *Correlation between the time series, previously deseasonalised and known variability removed, of layered ozone monthly SBUV MOD overpasses at the Hohenpeissenberg station and the SBUV MOD overpasses at various other locations around the globe. Four layers are shown in the panels. The black star indicates the location of Hohenpeissenberg.*

3.3 Summary

Any measurement process unavoidably brings about uncertainties, which ultimately propagate into ozone profile trend uncertainties. Some sources of uncertainty can be directly estimated by the regression algorithm from the time series, others have to be quantified by independent means. For instance, a constant drift in ozone levels over time can be, to a large degree, collinear with the trend proxy term in the regression. It will therefore be absorbed in the trend estimate but not in the trend uncertainty estimate.

In this particular case, as in others, there is a clear benefit of having several complementary contemporary data records since none of the individual satellite or ground-based records provide superior stability over the entire spatio-temporal domain of interest. Intercomparisons of ozone time series of various kinds (single profile measurements, local and monthly zonal means or monthly deseasonalised anomalies, single-sensor or merged records) have revealed measurement-related artefacts, such as drifts, discontinuities, and spikes. For some artefacts the evidence was comprehensive enough to exclude (part of) the data record from further analyses. Other issues were not, or could not be, removed, but they have been taken into consideration in the interpretation of the trend results in *Chapter 5*. These include the drift in a few satellite data records in part of the stratosphere, most notably for OSIRIS and OMPS-LP. Improvements are required, especially for the OMPS-LP data record as it drifts by 5–10 % per decade, most likely as a result of unstable altitude registration. Most ground-based station records exhibit anomalous behaviour during some periods in time. Although the anomalies are broadly consistent with reported systematic errors of 5–10 %, they are episodic rather than systematic in nature. Despite these residual artefacts, the agreement between observational records has generally been improved when compared to the consistency found for earlier data versions used by previous assessments (*e.g.*, *WMO,* 2014; *Tummon et al.,* 2015; *Harris et al.,* 2015; and references therein).

Complementary analysis methods and tools are an asset as well. Comprehensive approaches that intercompare not one or two but many data sets in a coherent way are key in attributing issues to a particular data record. The Bayesian algorithm BASIC proves more robust against outliers than traditional methods to infer the underlying ozone time series from a set of (imperfect) data records. This recent development has shown clear potential in providing insights in more subtle uncertainty patterns relevant for trend studies. MC simulations have proven useful in estimating the additional trend uncertainty related to remaining potential artefacts that cannot be cleanly identified and removed as well as how the merging process deals with these artefacts. For example, seemingly statistically significant discrepancies between trends derived from two SBUV-based records are found to overlap within uncertainty estimates when those estimates include the uncertainty of the individual SBUV data records propagated through the merging process using MC simulations.

The impact of sampling uncertainty on trends is now much better understood. This source of uncertainty is unrelated to the performance of the instrument and becomes only important if the data are analysed at an aggregate level sufficiently far away from that of the original individual profile measurements. Studies using SBUV data showed high correlations between time series at individual sites and those averaged in corresponding 5° latitude belts. The impact of sampling uncertainty is a more important issue for the analysis of monthly zonal mean ozone values by the sparsely-sampled occultation sounders. The interplay of changes in the measurement pattern and diurnal and seasonal gradients lead to systematic changes in derived trends by up to 1–2 % per decade in parts of the stratosphere.

The results described in this chapter are further considered in the interpretation of the trend results in *Chapter 5.*