Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016-179, 2016
Manuscript under review for journal Atmos. Chem. Phys.
Published: 4 April 2016

Atmospheric Chemistry and Physics
Discussions
Open Access

An assessment of the climatological representativeness of IAGOS-CARIBIC trace gas measurements using EMAC model simulations

Johannes Eckstein[1], Roland Ruhnke[1], Andreas Zahn[1], Marco Neumaier[1], Ole Kirner[2], and Peter Braesicke[1]

[1] Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research (IMK), Herrmann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[2] Karlsruhe Institute of Technology (KIT), Steinbuch Centre for Computing (SCC), Herrmann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

Correspondence to: Johannes Eckstein (johannes.eckstein@kit.edu)

Abstract. Measurement data from the long-term passenger aircraft project IAGOS-CARIBIC is often used to derive trace gas climatologies. We investigate to what extent such derived climatologies can be assumed to be representative for the true state of the atmosphere. Using the chemistry-climate model EMAC we sample the modelled trace gases along CARIBIC flight tracks. Different trace gases are considered and climatologies relative to the mid-latitude tropopause are calculated. Representativeness can now be assessed by comparing the CARIBIC sampled model data to the true climatological model state. Three statistical methods are applied for this purpose: the Kolomogorov-Smirnov test, and scores based on the variability and relative differences.

Generally, representativeness is expected to decrease with increasing variability and to increase with the number of available samples. Based on this assumption, we investigate the suitability of the different statistical measures for our problem. The Kolmogorov-Smirnov test seems too strict and does not identify any climatology as representative – not even long lived well observed trace gases. In contrast, the variability based scores pass the general requirements for representativeness formulated above. In addition, even the simplest metric (relative differences) seems applicable for investigating representativeness.

Using the relative differences score we investigate the representativeness of a large number of different trace gases. For our final consideration we assume that the EMAC model is a reasonable representation of the real world and that representativeness in the model world can be translated to representativeness for CARIBIC measurements. This assumption is justified by comparing the model variability to the variability of CARIBIC measurements. Finally, we show how the representativeness score can be translated into a number of flights necessary to achieve a certain degree of representativeness.

1 Introduction

The UTLS (upper troposphere/lower stratosphere) is dynamically and chemically very complex and shows strong gradients in temperature, humidity and in many trace gases (Gettelman et al., 2011). As the the mid and upper troposphere have a strong influence on the atmospheric greenhouse effect, the UTLS plays an important role in our climate system (Riese et al., 2012).

To characterize processes and evaluate the performance of chemistry-transport models in this area, we require spatially well resolved data collected on a global scale.

Aircraft are a suitable platform to carry out these measurements as they are able to probe in situ and at a high frequency. Measurements taken by commercial aircraft projects like IAGOS (In-Service Aircraft for a Global Observing System, Petzold et al. (2015)) and CONTRAIL (Comprehensive Observation Network for Trace gases by Airliner, Matsueda et al. (2008)) generate more continuous and regular datasets than research aircraft on sporadic campaigns and are therefore commonly given the attribute representative. But what is meant by this adjective?

Ramsey and Hewitt (2005) give a general introduction to representativeness, coming from soil sciences. As they state, the adjective representative has no meaning of its own, so a definition has to be given and 'it must be asked "representative of what?"'

In the scope of meteorology, Nappo et al. (1982) give the following definition: 'Representativeness is the extent to which a set of measurements taken in a space-time domain reflects the actual conditions in the same or different space-time domain taken on a scale appropriate for a specific application.' Representativeness in their understanding 'is an exact condition, i.e., an observation is or is not representative.' Only if 'a set of criteria for representativeness is established, analytical and statistical methods can be used to estimate how well the criteria are met.'

The mathematical definition given by Nappo et al. (1982) is mostly applied to data collected in the boundary layer, where it is used to answer the question whether a flux tower station is representative for the area in which it is positioned (e.g. by Schmid (1997), Laj et al. (2009) or Henne et al. (2010)). This can also be analysed by means of a cluster analysis with backward trajectories (e.g. by Henne et al. (2008) or Balzani Lööv et al. (2008)). By this method, source regions for measured trace gases can be found and the type and origin of air masses contributing to an observed air mass determined, i.e. the airmass the data is representative for. Köppe et al. (2009) apply this method to aircraft data from the project IAGOS-CARIBIC (Civil Aircraft for the Regular Investigation of the Atmosphere Based on an Instrument container, being part of IAGOS).

Lary (2004) and Stiller (2010) discuss the representativeness error in the field of data assimilation. Lary (2004) uses representativeness uncertainty as a synonym for variability within a grid cell, Stiller (2010) discusses the sampling error, which is considered to be part of the representativeness uncertainty. Larsen et al. (2014) study the representativeness of one dimensional measurements taken along the flight track of an aircraft to the three dimensional field that is being probed. But as they consider single flight tracks, their methods and definitions do not apply here.

The study of Schutgens et al. (2016) is more related to this study. They consider the sampling error on a global scale, comparing normal model means to means of model data collocated to satellite measurements. They find that this sampling error reaches $20-60\,\%$ of the model error (difference between observations and collocated model values).

We have been motivated by Kunz et al. (2008). They analysed whether the dataset of the aircraft campaign SPURT (SPURenstofftransport in der Tropopausenregion - trace gas transport in the tropopause region, Engel et al. (2006)) is representative of the larger MOZAIC dataset (Measurements of OZone, water vapour, carbon monoxide and nitrogen oxides by in-service AIrbus airCraft, the precursor of IAGOS-core). Kunz et al. (2008) investigate distributions of two substances (O_3 and H_2O) in two atmospheric compartments (upper troposphere and lower stratosphere). They find that the smaller SPURT dataset is representative on every time scale of the larger MOZAIC set for O_3, while this is not the case for H_2O. While SPURT O_3 data can be used for climatological investigations, the variability of H_2O is too large to be fully captured by SPURT on the interseasonal time scales.

This is similar to what is done in this study: We investigate the representativeness of data for different trace gases from IAGOS-CARIBIC (see Sec. 2.1) for a climatology in the UTLS. Possible mathematical definitions of the word representativeness are first discussed with the help of this data. Then, its representativeness following these definitions is investigated. By using data from the chemistry-climate model EMAC (see Sec. 2.2) along the flight tracks of IAGOS-CARIBIC and comparing this to a larger sample taken from the model, it becomes possible to investigate the representativeness of the smaller of the two model datasets. We assume that the different species are equally well represented in the model in terms of the processes acting on them and their variability. In this way, the representativeness of IAGOS-CARIBIC measurement data for a climatology in the UTLS can be quantified by using the two model datasets alone, using only the geolocation of the measurements. An exact reproduction of all measurements by the model is not necessary for this argument and is not investigated in this study.

In Sec. 2, more details on the data from IAGOS-CARIBIC and the model run will be given. The general concept and definition of representativeness is discussed in Sec. 3. This section also gives details on sampling the model and on the variability, which is used to group results by species. The statistical methods are then explained in Sec. 4, namely the Kolmogorov-Smirnov test, a variability analysis following the general idea of Kunz et al. (2008) and Rohrer and Berresheim (2006) and the relative difference of two climatologies. The application of the methods to the different model samples is described in Sec. 5. After showing the result of each of the three methods seperately, Sec. 5.4 discusses the representativeness of the IAGOS-CARIBIC measurement data, while Sec. 5.5 answers the question how many flights are necessary to achieve representativeness. Sec. 6 summarizes and concludes.

2 Model and data

2.1 The observational IAGOS-CARIBIC dataset

Within IAGOS-CARIBIC (CARIBIC for short), an instrumented container is mounted in the cargo bay of a Lufthansa passenger aircraft during typically four intercontinental flights per month, flying from Frankfurt, Germany (Munich, Germany, since August, 2014), see also Brenninkmeijer et al. (2007) and www.caribic-atmospheric.com.

During each CARIBIC flight, about 100 trace trace gas and aerosol parameters are measured. Some are measured continuously with a frequency between $5\,\mathrm{Hz}$ and $1/(5\,\mathrm{min})$ and commonly available every $10\,\mathrm{s}$ while others (e.g. non-methane hydrocarbons) are taken from up to 32 air samples collected per flight. The substances considered in this study are NO_y, H_2O, O_3, CO_2, NO, NO_2, $(CH_3)_2CO$ (acetone), CO and CH_4 from continuous measurements and N_2O, C_2H_6 and C_3H_8 from air samples. NO_y is the sum of all reactive nitrogen species, measured by catalytic conversion to NO (Brenninkmeijer et al., 2007).

The data of all flights from the year 2005 (beginning of the second phase of CARIBIC) to the end of December, 2013 (end of the model run) are considered in this study.

As this study investigates representativeness using model data, the geolocation of the CARIBIC measurements at $10\,\mathrm{s}$ resolution is used. In a second step, the gaps of the CARIBIC measurements and height information (due to technical problems etc.) are mapped onto their representation in the model data to infer the representativeness of the measurement data.

2.2 The chemistry-climate model EMAC

EMAC (ECHAM5/MESSy Atmospheric Chemistry model; Jöckel et al. (2006)) is a combination of the general circulation model ECHAM5 (Roeckner et al., 2006) and different submodels combined through the Modular Earth Submodel System (MESSy, Jöckel et al. (2005)). We use here a model configuration with 39 vertical levels reaching up to $80\,\mathrm{km}$ and a horizontal resolution of T42 (roughly $2.8°$ horizontal resolution).

The model integration used in this study simulated the time between January 1994 and December 2013, with data output every eleven hours. Meteorology is nudged up to $1\,\mathrm{hPa}$ using divergence, vorticity, ground pressure and temperature from six-hourly ERA-Interim reanalysis. It includes the extensive EVAL-Chemistry using the kinetics for chemistry and photolysis of Sander et al. (2011). This set of equations has been designed to simulate tropospheric and stratospheric chemistry equally well.

The substances used from the model are the same as those used from measurements, summing up $\mathrm{NO_y}$ from N, NO, $\mathrm{NO_2}$, $\mathrm{NO_3}$, $\mathrm{N_2O_5}$ (counted twice), $\mathrm{HNO_4}$, $\mathrm{HNO_3}$, HONO, HNO, PAN, $\mathrm{ClNO_2}$, $\mathrm{ClNO_3}$, $\mathrm{BrNO_2}$ and $\mathrm{BrNO_3}$. Data of $\mathrm{N_2O}$, $\mathrm{CH_4}$ and $\mathrm{CO_2}$ was detrended by subtracting the mean of each year from the values of that year and adding the overall mean.

3 Defining representativeness

As noted above and specified by Nappo et al. (1982) and Ramsey and Hewitt (2005), the word representative is meaningful only if accompanied by an object. Ramsey and Hewitt (2005) raise three questions to be answered in order to address representativeness: 1. For what parameter is the sample data to be seen as representative: e.g. the mean, a trend or an area? 2. Of which population is the sample data to be seen as representative? 3. To which degree is the data to be seen as representative? To assess the representativeness of CARIBIC data, these three questions have to be answered as well.

3.1 Representative for what parameter?

First, it is crucial to define what we anticipate the CARIBIC data to be representative of, since 'the same set of measurements may be deemed representative for some purpose but not other' (Nappo et al., 1982). In this study, we investigate whether the CARIBIC data can be used to construct a climatology in the UTLS. We consider monthly binned data in the height of $\pm 4\,\mathrm{km}$ around the dynamical tropopause defined at the pressure at $3.5\,\mathrm{PVU}$.

In order to reference data to the tropopause, we use the geometric height in kilometers relative to the tropopause (HrelTP) at each datapoint. For the measurements, this height is provided by the meteorological support of CARIBIC by KNMI (Koninklijk Nederlands Meteorologisch Instituut) (http://www.knmi.nl/samenw/campaign_support/CARIBIC/), who use data from ECMWF (European Centre for Mendium-range Weather Forecast) for their calculation.

From model output, the height relative to the tropopause (HrelTP) can be calculated, as the pressure value of the dynamical tropopause is known at each location, as well as the temperature and pressure profile. This HrelTP value calculated from the model data along the flight tracks of CARIBIC compares well with interpolated values from ECMWF provided by KNMI (Pearson correlation coefficient of $\rho = 0.97$), which is expected as the meteorology of the model is nudged using ERA-Interim data. The distribution of all values of HrelTP from the model is shown in Figure 1, showing a maximum right at the tropopause. Data was used within $\pm 4.25\,\mathrm{km}$ of the tropopause in steps of $0.5\,\mathrm{km}$, labelling the bins according to the central height at full and half kilometers.

Even though all data of trace gases (be it from model or measurements) is sorted into bins of HrelTP, it is important to keep in mind the limits in pressure. These are inherent in the CARIBIC dataset, as the aircraft flies on constant flight levels with $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$. In addition, we explicitly limit pressure to this range in order to exclude data from ascents and descents of the aircraft. But since data is considered relative to the tropopause, these limits are no longer visible directly from the resulting climatology, even though they can influence it strongly. The reason is that aircraft flying at constant pressure can measure far above (below) the tropopause only if the tropopause is located at high (low) pressure values. The properties of many trace substances are not only a function of their distance to the tropopause, but also of pressure. The limits in pressure inherent in the sample therefore also influence the climatology. They have to be considered and should be explicitly stated. This efffect is illustrated in the supplementary material with the help of the methods developed in this study.

In addition to limiting in HrelTP and p, it is necessary to apply a limit in latitude φ. Tropical data with $\varphi < 35°\mathrm{N}$ are excluded because of the considerably higher dynamical tropopause. Data with $\varphi > 75°\mathrm{N}$ are excluded because of the different chemistry in far northern latitudes, which leads to considerably different values for some some species that should not be combined with data from lower latitudes in one climatology. In addition, this latitudinal band is well covered by CARIBIC measurements. Other regions or latitudinal bands can be investigated using the same approach.

As a summary, we can specify more closely the question (Representative for what parameter?) asked in the beginning: Is a climatology compiled from CARIBIC data representative for the tropopause region in mid-latitudes?

3.2 Representative for which population?

When assessing the representativeness of the sample made up by all CARIBIC measurements, the population is the atmosphere around the tropopause and its composition. For many of the species measured by CARIBIC, there is no other project that takes such multi-tracer in-situ meaurements as regularly at the same spatial and temporal resolution. IAGOS-core and CONTRAIL sample with much higher frequency, but take measurements of only few substances while satellites do not resolve the small scale-structures necessary to disentangle the dynamics around the tropopause. The population is therefore not accessible by the measurement platforms currently available.

This is the reason why the representativeness of the CARIBIC data is investigated by comparing the model data along CARIBIC flight tracks to two larger samples taken from the model. These larger datasets are considered the population, in reference to which the representativeness of the smaller dataset (model along CARIBIC paths) is assessed. Three datasets were

Table 1. Summary of the specifications defining the three datasets MOD$_{\text{CARIBIC}}$, MOD$_{\text{RANDPATH}}$ and MOD$_{\text{RANDLOC}}$.

dataset	EMAC on	total sets	per month	duration	p distribution
MOD$_{\text{CARIBIC}}$	CARIBIC paths (2005-13)	334	up to 4 in 3 days	8-10h	flight levels show up, $\overline{p} = 223.42\,\text{hPa}$ $\sigma(p) = 18.94\,\text{hPa}$
MOD$_{\text{RANDPATH}}$	random paths	1296	12 in 28 days	24h	gaussian, $\overline{p} = 223.42\,\text{hPa}$ $\sigma(p) = 18.94\,\text{hPa}$
MOD$_{\text{RANDLOC}}$	random location	864	8 in 28 days	24h	even, $\min(p) = 10\,\text{hPa}$ $\max(p) = 500\,\text{hPa}$

created from the model output: the model along CARIBIC paths and two random model samples. All are presented in the following paragraphs, a summary being given in Table 1 and Figure 1.

MOD$_{\text{CARIBIC}}$: For the dataset MOD$_{\text{CARIBIC}}$, the model output was interpolated linearly in latitude, longitude, logarithm of pressure and time to the position of the CARIBIC aircraft, using the location at a resolution of $10\,\text{s}$ for all species. Figure 1 shows the flight paths considered in this study. Since CARIBIC also measures temperature, the high pearson correlation coefficient of $\rho = 0.97$ of modelled to measured temperature can serve as an indication that this interpolation leads to reasonable results, despite the coarse resolution in time and space of the model output.

As is visible in Fig. 1 (central column), only three of the model levels lay in the pressure range sampled by CARIBIC. This is why it is not feasible to compare MOD$_{\text{CARIBIC}}$ directly to the full model output, but two random model samples were created which are more similar in their statistical properties to MOD$_{\text{CARIBIC}}$.

MOD$_{\text{RANDPATH}}$: The dataset referred to as MOD$_{\text{RANDPATH}}$ is a larger set of flight paths used to sample the model. This set was mainly used to investigate the representativeness of MOD$_{\text{CARIBIC}}$. From the year 2005 to the end of 2013, 12 random flight paths were generated per month (1296 in total, evenly spaced in each month's first 28 days) and the model fields interpolated onto these paths. The starting point was randomly chosen in the northern hemisphere, as well as the direction taken by the aircraft. The speed was set to $885.1\,\text{km}\,\text{h}^{-1}$, the median of the speed of the true CARIBIC aircraft. The flights start at $0:00\,\text{UTC}$ and sample the model for one day in $10\,\text{s}$ intervals. They are reflected at the north pole and at the equator and reverse the sign of the increment in latitude direction once during flight. The first 100 of these paths are displayed in Figure 1.

The pressure was kept constant for each of the random flights, reproducing the statistics of the pressure distribution for CARIBIC as a whole. For this, a normal distribution centered around $223.42\,\text{hPa}$ with a standard deviation of $18.94\,\text{hPa}$ was used to choose the pressure value for each of the random flights. All pressure values of $p < 180\,\text{hPa}$ or $p > 280\,\text{hPa}$ were redistributed evenly between $200\,\text{hPa}$ and $250\,\text{hPa}$ to exclude unrealistically high or low values and sharpen the maximum.

MOD$_{\mathbf{RANDLOC}}$: For this sample, latitude and longitude were randomly drawn in the northern hemisphere (not aligned along a route) and the definition of the pressure distribution widened, drawing pressure from an even distribution from $500\,\mathrm{hPa}$ to $10\,\mathrm{hPa}$ for each flight. Again, the datasets start at $0:00\,\mathrm{UTC}$ and the separate points are $10\,\mathrm{s}$ apart, collecting 8640 samples on a sampling day. Eight of these sets are distributed evenly in each month, summing to a total of 864 sets of this type. This set was used to test whether MOD$_{\mathrm{CARIBIC}}$ is representative for a climatology around the tropopause only within its pressure limits or also when expanding these limits.

As is visible in Figure 1, the distribution in HrelTP is very similar for all datasets even though the pressure is presribed in very different ways. This is an important prerequisite for the following investigation, as it shows that the relative amount of data in each height bin is similar for all three datasets.

Representativeness was assessed using only model data in this study. In order to transfer the results from model data to measurements, we assume that different species are equally well represented in the model in terms of their variability. This inference is plausible, considering the equally good representation of the stratosphere and the troposphere in the model (Jöckel et al., 2015).

The question whether this assumption is valid was also investigated with the available data. The relative standard deviation $\sigma_r = \sigma/\mu$ was calculated in each month of the climatlogies of CARIBIC measurements (MEAS$_{\mathrm{CARIBIC}}$) and MOD$_{\mathrm{CARIBIC}}$ (σ being the standard deviation, μ the mean) as a measure for the variability. By taking the mean over all months of the fraction of $\sigma_r^{\mathrm{MOD_{CARIBIC}}}$ and $\sigma_r^{\mathrm{MEAS_{CARIBIC}}}$, the fraction of variability of MEAS$_{\mathrm{CARIBIC}}$ reached by MOD$_{\mathrm{CARIBIC}}$ can be evaluated.

The variability of MOD$_{\mathrm{CARIBIC}}$ is similar for all species, reaching between 40 and $70\,\%$ of that of MEAS$_{\mathrm{CARIBIC}}$. The Pearson correlation coefficient of $\sigma_r^{\mathrm{MOD_{CARIBIC}}}$ and $\sigma_r^{\mathrm{MEAS_{CARIBIC}}}$ is 0.81 (see supplementary material). These two facts show that the model represents all species equally well. On an absolute scale, the model cannot reach the variability of measurements due to its coarse resolution (see Section 2.2). The linear interpolation onto the location of the aircraft does not introduce the smaller scale variability present in the measurements. Also, the variability of MEAS$_{\mathrm{CARIBIC}}$ is not equal to the atmospheric variability, due to different characterisitics of the instruments for each species.

The assumption underlying this study is that the representativeness evaluated from the model data is also valid for the real atmosphere and the measurements taken by CARIBIC. This assumption is justified by the similar variability of the model for all species.

3.3 Confidence limits of the representativeness

When defining representativeness, one more question remains: What are the confidence limits of the representativeness?

Three definitions for representativeness are discussed and applied in this study: The Kolmogorov-Smirnov test, the variabiltiy analysis following Kunz et al. (2008) and the relative difference of two climatologies. The first method gives a yes-no answer within a chosen statistical confidence level. The other two approaches are formulated in such a way as to return a score. By (arbitrarily) setting a value for the score, the representative cases can be discriminated from the non-representative cases (see Sec. 4 and Sec. 5), the score corresponding to a confidence level.

There are two more requirements that we define as having to be met by representativeness in general:

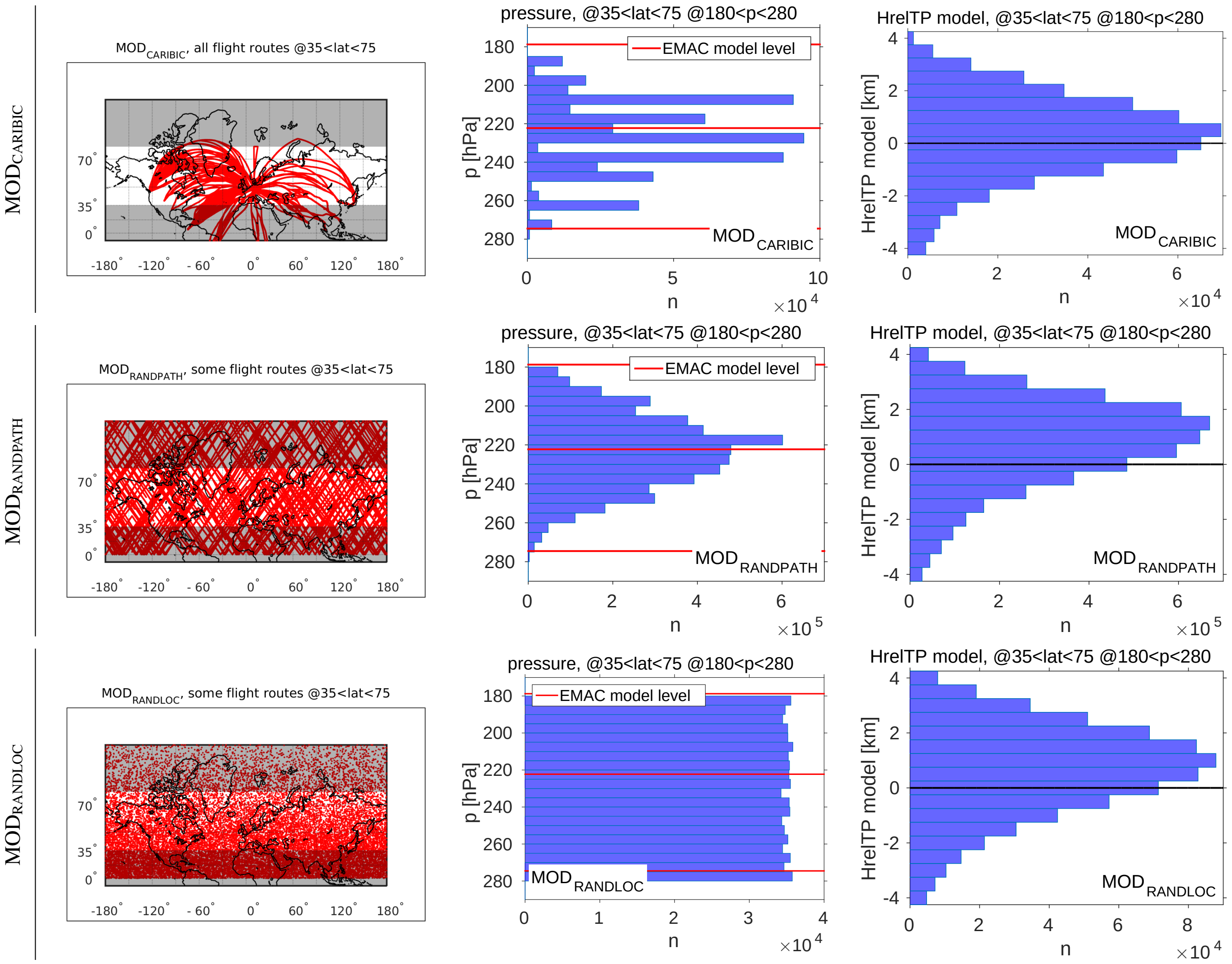

Figure 1. Flight paths (left), distribution in p (center) and HrelTP (right) for the three datasets $\mathrm{MOD_{CARIBIC}}$ (top), $\mathrm{MOD_{RANDPATH}}$ (center) and $\mathrm{MOD_{RANDLOC}}$ (bottom). Only parts of the paths of $\mathrm{MOD_{RANDPATH}}$ and $\mathrm{MOD_{RANDLOC}}$ are shown.

1. Representativeness has to increase with the number of samples (flights in the case of this study).

2. Representativeness has to decrease with increasing variability of the underlying distribution.

These two assumptions are implicitely also made by Kunz et al. (2008), as they investigate the representativeness of a smaller for a larger dataset and for two species of different variability. The measure for variability we use in this study is explained in the following section.

3.4 Defining a measure for variability

The representativeness is expected to differ for different species because of their atmospheric variability or atmospheric lifetime. This is part of the definition of representativeness given in Section 3.3. Kunz et al. (2008) also find that O_3 and H_2O are different in their representativeness and attribute this to the variability. It is therefore reasonable to consider results for representativeness relative to the variability of a species, which we denote by τ^*. It is calculated from $\mathrm{MOD}_{\mathrm{RANDPATH}}$ following Equation 1 using the mean μ and standard deviation σ of each species.

$$\tau^* = \log_{10}\left(\frac{\mu}{\sigma}\right) \tag{1}$$

Figure 2 shows the sorted values of τ^* for the species considered in this study. It is worthwile to note that in defining τ^* in this way, we closely follow Junge (1974), who showed that under certain constraints, the relationship

$$10^{-\tau^*} = \frac{\sigma}{\mu} = \sigma_r = a \cdot \tau^{-b} \tag{2}$$

holds, which links variability and lifetime τ using two species-dependent constants a and b. σ_r is the relative standard deviation used in Section 3.2 to compare model and measurement variability. This relationship has frequently been called Junge relationship in the past (e.g. by Stroebe et al. (2006) or MacLeod et al. (2013)). And indeed, as is visible in Figure 2, longer lived species like CO_2 or N_2O show lower variability (higher τ^*), while shorter lived species show higher variability (lower τ^*).

So including these thoughts on variability in the question formulated at the end of Section 3.1, we can specify more closely the question we answer in this study: For which species is a climatology compiled from CARIBIC data representative for the tropopause region in mid-latitudes?

4 Statistical methods

We use three different methods to evaluate representativeness: the Kolmogorov-Smirnov test, the variability analysis and relative differences.

4.1 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov two-sample test is a non-parametric statistical test that is used to examine whether two datasets have been taken from the same distribution (e.g. Sachs and Hedderich (2009)). It considers all types of differences in the sample

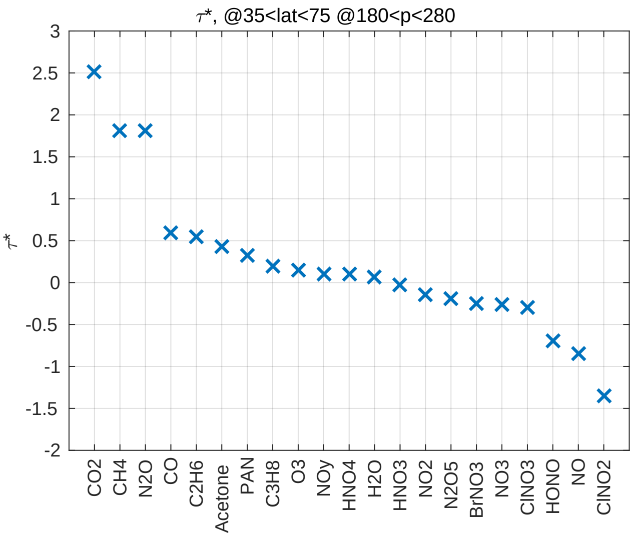

Figure 2. Variability τ^* calculated from MOD$_{\text{RANDPATH}}$ using Equation 1. The species are sorted in τ^*, species with low variability (high τ^*) listed to the left.

distributions that can be apparent in the mean, the standard deviation, the kurtosis, etc. The test statistic is the maximum absolute difference $\hat{D}$ in the cumulative empirical distribution functions $\hat{F}_x$ of the two samples x:

$$\hat{D} = \max|\hat{F}_1 - \hat{F}_2| \qquad (3)$$

The discriminating values D_α have been derived depending on the accepted confidence limit α. In this study, the two empirical distribution functions $\hat{F}_i$ were taken from MOD$_{\text{CARIBIC}}$ and MOD$_{\text{RANDPATH}}$ in each height bin and month, see Sec. 5.1.

4.2 Variability analysis

The variability analysis follows Rohrer and Berresheim (2006) and Kunz et al. (2008). Rohrer and Berresheim (2006) introduced a variance analysis for ground-based observations, Kunz et al. (2008) then applied it to aircraft data. A timeseries of data is subsequently divided into ever shorter time slices of increasing number and the variance is calculated for the data within each time slice. By taking the mean over the whole number of slices and doing this for all divisions in time, a line is calculated, which is characteteristic for the development of variance in time.

Instead of considering variance in each time slice, we use the relative standard deviation $\frac{\sigma}{\mu}$, which is the definition of variabiltiy following Junge (1974). It is calculated in each time slice and the mean gives the value for the corresponding time scale. By scaling the standard deviation σ with the mean μ, different species become comparable. Being a combination of variability as defined Junge (1974) and the variance analysis introduced by Rohrer and Berresheim (2006), this method is called variability analysis in the following paragraphs.

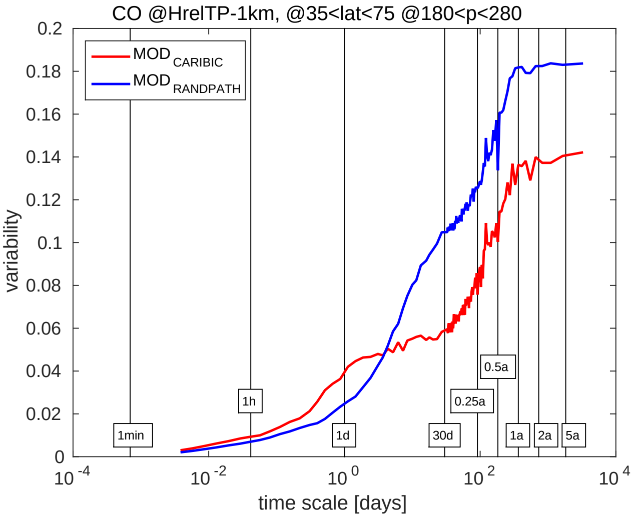

Figure 3. Variability analysis calculated for CO for HrelTP $= -1\,\mathrm{km}$ (one kilometer below the tropopause) for MOD$_{\mathrm{CARIBIC}}$ and MOD$_{\mathrm{RANDPATH}}$. The time scales used to calculate R$_{\mathrm{var}}$ using Equation 4 are indicated by vertical lines.

Figure 3 shows the variability analysis for CO just below the tropopause for MOD$_{\mathrm{CARIBIC}}$ and MOD$_{\mathrm{RANDPATH}}$. The time scale changes from about $5\,\mathrm{min}$ to $5\,\mathrm{a}$ along the logarithmically spaced abscissa. As CO is a medium long-lived trace gas with an atmospheric lifetime of 2-3 months and a pronounced annual cycle, the mean variability increases up to time scales of $1\,\mathrm{a}$. The variability of MOD$_{\mathrm{RANDPATH}}$ is larger on almost all time scales. For time scales of $30\,\mathrm{d}$ and more, however, the lines run in parallel, showing an increase up to $1\,\mathrm{a}$, from when on the variability does not increase. This is consistent with the annual cycle of CO, which is also the cause for the relative decrease sharply at $0.5\,\mathrm{a}$ and $1.5\,\mathrm{a}$. For time scales below $30\,\mathrm{d}$, the distribution of flights in one month dominates the variability analysis. MOD$_{\mathrm{CARIBIC}}$ includes only up to four flights on consecutive days, the mean variability does not decrease when going to time scales between $30\,\mathrm{d}$ and $4\,\mathrm{d}$, while in MOD$_{\mathrm{RANDPATH}}$, continuosly less data is included in each time slice, leading to a continuous drop in the variability. For time scales of less than $1\,\mathrm{d}$, the data comes from a single flight, showing another drop in variability that is linked to using data from geographic regions that are ever more close. Since the variability analysis is so closely linked to the distribution in time and space, the variability analysis of MOD$_{\mathrm{RANDLOC}}$ shows an almost constant value in the time scales shorter than $30\,\mathrm{d}$ (not shown).

Kunz et al. (2008) used the variance analysis to investigate whether the smaller SPURT dataset represents the variance present in MOZAIC dataset. Following this thinking, we consider the variability as one possible criterion to judge the representativeness of one dataset for another. A score $\mathrm{R}_{\mathrm{var}}^{t,h}$ describing the representativeness is defined from the difference of the values of the variability analysis, using the following equation:

$$\mathrm{R}_{\mathrm{var}}^{t,h} = \log_{10}\left(\left|\overline{\left[\frac{\sigma_1^{t,h}}{\mu_1^{t,h}}\right]} - \overline{\left[\frac{\sigma_2^{t,h}}{\mu_2^{t,h}}\right]}\right|\right) \qquad (4)$$

where $\sigma_x^{t,h}$ stands for the standard deviation and at $\mu_x^{t,h}$ for the mean in time scale t and height h of the datasets x. The overbar implies that the mean over all time slices corresponding to the time scale t of σ/μ are used. Considering Figure 3, the score can be interpreted as the absolute value of the difference of the two lines at certain time scales t.

Decreasing values of $\mathrm{R}_{\mathrm{var}}^{t,h}$ mean better representativeness, the value always being negative. Depending on t, the representativeness in different time scales can be evaluated. We used time scales of $0.25\,\mathrm{a}$, $0.5\,\mathrm{a}$, $1\,\mathrm{a}$, $2\,\mathrm{a}$ and $5\,\mathrm{a}$ to calculate $\mathrm{R}_{\mathrm{var}}^{t,h}$. When applying this method to all height bins, a profile in $\mathrm{R}_{\mathrm{var}}^{t}$ is calculated for each species. This is one possible definition for representativeness. Yet it has to pass the two requirements of being related to number of samples and variability outlined in Sec. 3.3. The results of testing this will be presented in Sec. 5.2.

4.3 Relative difference

The third approach to assess representativeness is to analyze the relative differences between the climatologies from two differently large datasets. The procedure is summarized in Equation 5:

$$\mathrm{R}_{\mathrm{rel}}^{h} = \log_{10}\left(\frac{1}{12}\sum_{m=1}^{12}\frac{|\mu_1^{m,h}-\mu_2^{m,h}|}{\mu_2^{m,h}}\right) \quad (5)$$

which was applied to each height bin h. $\mu_x^{m,h}$ stands for the mean of the data in the month m and in height bin h of the datasets x. The logarithm to the basis 10 was applied to the mean relative difference profile to end up with a profile in $\mathrm{R}_{\mathrm{rel}}$, similar to the score $\mathrm{R}_{\mathrm{var}}^{t}$ calculated from the variability analysis. Contrary to the Kolmogorov-Smirnov test or the variability analysis, this test statistic does not contain any information on the underlying distribution, because it uses only the mean in each bin.

Figure 4 shows an example of relative differences between CO from $\mathrm{MOD}_{\mathrm{CARIBIC}}$ and the larger dataset $\mathrm{MOD}_{\mathrm{RANDPATH}}$. The differences are small, mostly below an absolute value of 0.15. $\mathrm{R}_{\mathrm{rel}}$ is defined (in Equation 5) as the logarithm to the base 10 of the mean over all months (not shown). The score increases towards the top and bottom in Figure 4 due to less data there. Like for $\mathrm{R}_{\mathrm{var}}^{t}$, decreasing values in $\mathrm{R}_{\mathrm{rel}}$ mean better representativeness. And like $\mathrm{R}_{\mathrm{var}}^{t}$, $\mathrm{R}_{\mathrm{rel}}$ has to be tested for passing the requirements of being related to number of samples and variability (see Sec. 3.3) in order to be acceptable as a score for representativeness. The results of testing this will be discussed in Sec. 5.3.

Other than just as a score, the value of $\mathrm{R}_{\mathrm{rel}}$ can be understood as the average uncertainty for assuming the climatology of $\mathrm{MOD}_{\mathrm{CARIBIC}}$ as a full model climatology. This is more obvious if taken to the power of 10, in which case the uncertainty will take values between 0 and 1. Use of this will be made in Section 5.4.

5 Results

Here, we first present the results of the application of the Kolmogorov-Smirnov test (Sec. 5.1), the variability analysis (Sec. 5.2) and the relative difference (Sec. 5.3) to $\mathrm{MOD}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. All have to be related to the number of flights and the the variabiltiy of the species as discussed in Section 3.3. A similar analysis has also been performed with data from a random number generator, leading to equivalent results. This study is presented as supplementary material to the article. Sec. 5.4

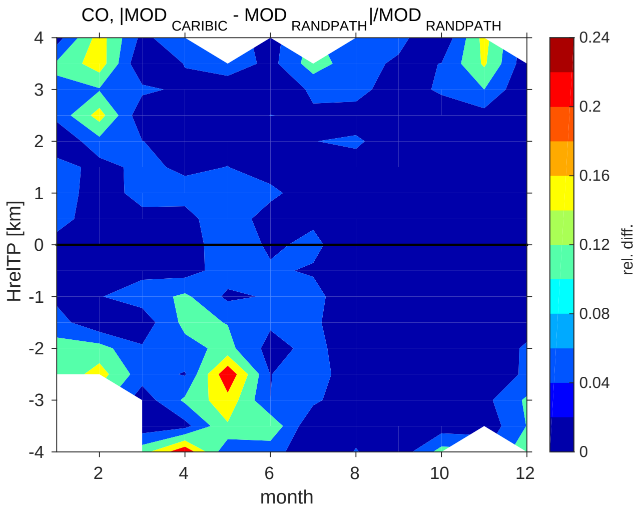

Figure 4. Relative differences of CO for $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDPATH}}$ used to calculate $\mathrm{R_{rel}}$.

interprets the results by species as a representativeness uncertainty. Finally, Sec. 5.5 answers the question of how many flights are necessary to achieve a certain degree of representativeness.

5.1 Applying the Kolmogorov-Smirnov test

The application of the Kolmogorov-Smirnov test to $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{CARIBIC}}$ yields a first important result. Independent of the trace gas and HrelTP considered, the result is always negative (not shown). This means that the data in each bin of $\mathrm{MOD_{CARIBIC}}$ is not representative of the corresponding bin in $\mathrm{MOD_{RANDPATH}}$ when defining representativeness by a positive result of the Kolmogorov-Smirnov test. This is also true if the data is not binned in months but only in HrelTP. The result also stays the same for all values of the confidence limit α (using values of 0.001, 0.01, 0.05, 0.1 and 0.2).

A similar finding for aircraft data has already been reported by Kunz et al. (2008). On the one hand side this could mean that $\mathrm{MOD_{CARIBIC}}$ is simply not representative of $\mathrm{MOD_{RANDPATH}}$. But if the other methods presented here are considered, the conclusion seems more appropriate that the Kolmogorov-Smirnov test is simply not the correct way to answer the question. It can be considered as too strict for the type of data and the question considered here. This is further discussed with the help of a sensitivity study, the results of which are presented as supplementary material to this text.

5.2 Applying the variability analysis

This section presents the results of the application of the variability analysis to $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDPATH}}$. Equation 4 was applied for different time scales (0.25 a, 0.5 a, 1 a, 2 a and 5 a) to calculate $\mathrm{R_{var}}$. The results are exemplarily discussed for a time scale of 1 a, shown in Figure 5, in which the results are sorted using the values of τ^* displayed in Figure 2.

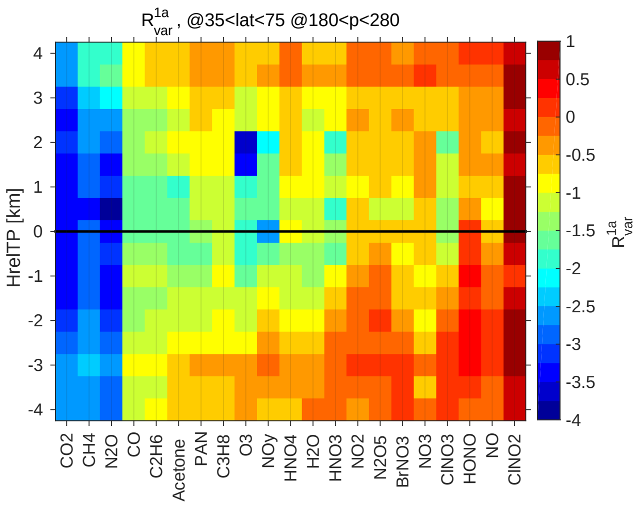

Figure 5. R_{var} calculated according to Equation 4 for a time scale of 1 a for all species in all height bins, using $MOD_{CARIBIC}$ and $MOD_{RANDPATH}$.

R_{var} shows a strong relationship with τ^*. This is visible from Figure 5, in which the results are sorted using the values of τ^* displayed in Figure 2, that is with increasingly higher atmospheric variabilty from left to right. The Pearson correlation coefficient ρ of R_{var} and τ^* is high, $|\rho| > 0.9$ in all height bins, independent of the time scale. R_{var} also shows a strong relationship to the number of samples: The amount of data in both $MOD_{CARIBIC}$ and $MOD_{RANDPATH}$ decreases below and above the tropopause, and R_{var} follows suit for practically all species. R_{var} therefore passes the requirements of being inversely related to τ^* and directly to the amount of used data points. Figure 5 can therefore be used to judge upon the representativeness of $MOD_{CARIBIC}$ for $MOD_{RANDPATH}$. This is also supported by the study of random number data presented as supplementary material.

This shows that by using the relive standard deviation (Equation 4) instead of the variance analysis applied by Kunz et al. (2008), the difference in variability can be used to infer representativeness. Rohrer and Berresheim (2006) originally introduced the variance analysis to investigate the sources and time scales of variability in a dataset and for this it remains a valid method. In order to infer representativeness, it is more appropriate to use the relative standard deviation in the analysis instead of the absolute variance.

5.3 Relative differences

R_{rel} was calculated for each species in each height bin according to Equation 5, see Figure 6.

Figure 6 shows how low variability (decreasing to the left, values taken from Figure 2), is linked with low values in R_{rel}, or good representativeness, respectively (see Sec. 4.3). R_{rel} decreases linearly with increasing variability τ^* with a high Pearson correlation coefficient greater than 0.95 for all height bins (not shown). The relation of R_{rel} with the number of values is also

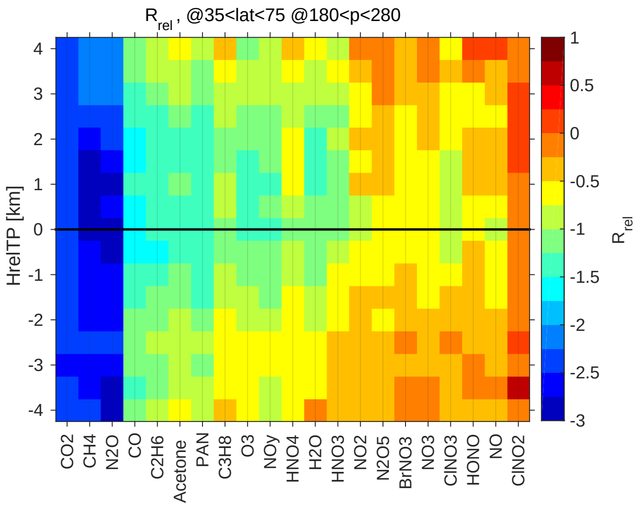

Figure 6. R_{rel} calculated for according to Equation 5 all species in all height bins, using $MOD_{CARIBIC}$ and $MOD_{RANDPATH}$.

visible in Figure 6 as the values decrease with the number of data points, this number having its maximum just around the tropopause and decreasing above and below it (see Figure 1). This shows that R_{rel} passes the requirements of being related to number of samples and variability τ^* and can be used as a measure for representativeness.

This relation with the number of values was tested in more detail: Each of the random paths of $MOD_{RANDPATH}$ was divided into three parts. Each part is then eight hours long, like a typical intercontinental flight with CARIBIC, and there are a total number of altogether 3888 shorter random flights. R_{rel} was then calculated for $MOD_{RANDPATH}$ and these subsamples, increasing their size by including more of the 3888 shorter random flights. The Pearson correlation coefficient between the number of shorter random flights and R_{rel} was larger than 0.9 for all species in all heights (not shown). This underlines how R_{rel} is well correlated with the number of measurements.

Using R_{rel} as a measure passes both conditions: It is directly proportional to the number of flights and indirectly to the variability. In addition to using Figure 5, Figure 6 can therefore be used to judge upon the representativeness of $MOD_{CARIBIC}$ for $MOD_{RANDPATH}$. R_{rel} can be transformed into a relative difference in percent, by taking R_{rel} to the power of ten. A score of -2 stands for a mean relative difference of $1\,\%$.

The score that discriminates representative from the non-representative case has to be arbitrarily chosen (see Nappo et al. (1982) and Ramsey and Hewitt (2005)). This score gives the uncertainty within which the data is considered representative. If a score of -2 is defined as representative (corresponding to $1\,\%$ mean relative difference), then representative species and heights can now be seperated from those species that are not representative using the results from Figure 6. But the score of -2 is arbitrary. If it is reduced to -1.5 (roughly $3\,\%$ relative difference), $MOD_{CARIBIC}$ can be seen as representative for many more species.

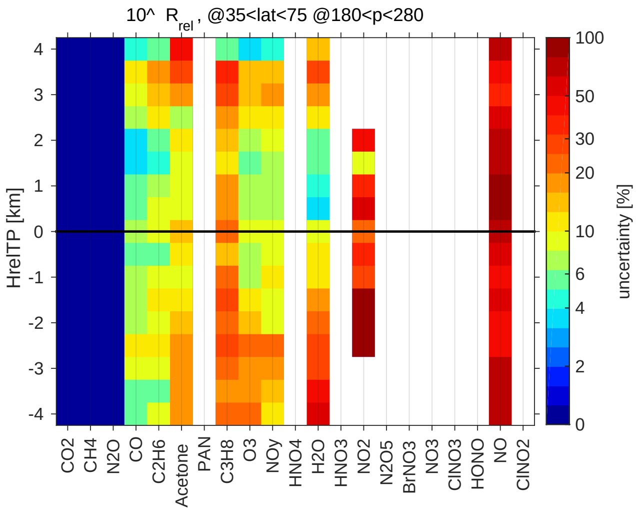

Figure 7. Representativeness uncertainty for using the CARIBIC data (that is 334 long-distance flights, see Table 1) to compile a climatology: $10^{R_{rel}}$ from $MOD_{CARIBIC}$ and $MOD_{RANDLOC}$.

5.4 Representativeness uncertainty of the CARIBIC measurement data

The last sections have shown R_{rel} (see Equation 5) and R_{var} (see Equation 4) to be adequate scores to describe representativeness. After reconsidering the question we asked in the Section 3.1 (Is a climatology compiled from CARIBIC data representative for the tropopause region in mid-latitudes?), we will use R_{rel} in the following. It is more intuitive (compared to R_{var}) as it describes the difference to a larger dataset, e.g. in percent and shows the slightly higher correlation coefficient. A further discussion of R_{var} is beyond the scope of this paper. As noted in Sec. 4.3, R_{rel} is also comprehensible as an uncertainty error for using the smaller dataset to compile a climatology and will be called representativeness uncertainty correspondingly.

In order to asses the uncertainty for accepting CARIBIC measurement data to create a climatology, all the gaps (e.g. due to instrument problems or measurement intervals $> 10\,\mathrm{s}$) in measurements and HrelTP (calculated from ECMWF fields in the case of measurements) have to be mapped onto $MOD_{CARIBIC}$ of the corresponding species and HrelTP calculated from the model. This was done and R_{rel} - taken to the power of 10 - recalculated using $MOD_{RANDLOC}$ with an even distribution in pressure, see Table 1. The limits in pressure where again set to $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$. The result is shown in Figure 7. Using different wording, R_{rel} in this formulation can also be considered the sampling error of the measurements.

This result - deduced from model data only - is also valid for the real world if the different species are equally well represented in terms of the processes that act on them, as is the case here, see Section 3.2. Figure 7 therefore gives the representativeness uncertainty not only for a reduced set of $MOD_{CARIBIC}$, but also for the CARIBIC measurements. It can be used to answer the question we asked in Sec. 3.2: For which species is a climatology compiled from CARIBIC data representative for the tropopause region in mid-latitudes? The influence of the limit in pressure is shown in the supplement.

When considering the representativeness uncertainty of a climatology, it is also important to consider the annual cycle of a species, e.g. 10 % can be much for a species that is more or less constant, while it is can be much for a species with a strong seasonality. Climatologies of CO_2, CO and O_3 are exemplarily discussed at the end of this section. The following paragraphs discuss representativeness by species, not explicitly considering the seasonal variations for each species.

Many of the species that sum up to NO_y in the model are not actually measured by CARIBIC and therefore get no value in Figure 7. In general, the representativeness uncertainty is lowest where there are most measurements, which is just around the tropopause (see Figure 1). This effect overlays the physical reasons for the different values of the uncertainty for all species considered. If the limits in pressure are expanded in using $MOD_{RANDLOC}$, the uncertainty increases markedly, as is shown in supplementary material. The reasons for this have been discussed in Section 3.1.

NO_2 and NO have the highest uncertainty of 90 % (NO) and up to 100 % in the case of NO_2. We propose two possible reasons: On the one hand, there are many gaps in the observations. But NO_2 and NO are also emitted by aircraft in the UTLS (Stevenson et al., 2004), and since CARIBIC flies in the flight corridors heavily frequented by commercial aircraft, it is unrealistic to assume a climatology of these species to be representative of the UTLS on a whole.

H_2O shows a strong gradient in its representativeness uncertainty, which is directly linked to the strong gradient in variability. The dry stratosphere can be described by relatively few measurements, which is why the uncertainty is low, only reaching 25 % at most. The humid and variable troposphere influenced by daily meteorology has a higher uncertainty, reaching more than 60 %.

NO_y, being a pseudo-species made up of many substances, is more difficult to disassemble. The variabilty of many components is higher in the troposphere, where the uncertainty is 30 % at its maximum. Above, it is smaller than 10 % and the climatology therefore quite trustworthy.

It is interesting to note that C_2H_6 and C_3H_8, both collected in whole air samples still reach uncertainty values comparable to those of other species in their range of τ^*. This is due to the fact that these are rather long-lived species for which only a moderate number of measurements are needed for a representative climatology. The climatology of C_3H_8 comes with an uncertainty of up to 25 %, while that of C_2H_6 is better with an uncertainty of less than 10 %.

The climatology of O_3 is very trustworthy, the uncertainty being smaller than 10 % for most height bins. The higher values in the tropospheric bins should not raise much concern, as O_3 increases strongly with height in the UTLS and an uncertainty of 15 % will be practically unnoticable compared to the vertical increase.

This is not true for acetone, where the gradient is just opposite to O_3. The climatology is trustable with an uncertainty only up to 10 % in upper levels, while it increases to 20 % in the lower heights, where the influence of spatially and temporally variable sources at the ground is stronger.

The climatology of CO is very good, the uncertainty in stratospheric height bins being less than 5 %. The troposphere, again stronger under the influence of sources, has a higher uncertainty reaching up to 10 %.

The long-lived trace gases CH_4, N_2O and CO_2 (all detrended as described in Sec. 2.1) all have representativeness uncertainties of less than 5 %. This is interesting especially for N_2O, which is measured only in the whole air samples.

As example and summary, the representativeness uncertainty will be applied to climatologies of CO_2, CO and O_3, shown in Figure 8. CO is shown for $MOD_{CARIBIC}$ (top left), $MOD_{RANDLOC}$ (top right) and CARIBIC measurements ($MEAS_{CARIBIC}$, center left). The white space in these figures has three possible reasons: the aircraft could have never flown in that bin, there could be measurement gaps in CO or a gap in HrelTP. The measurement gaps of CO and HrelTP from $MEAS_{CARIBIC}$ have been mapped onto $MOD_{CARIBIC}$, the two upper left hand climatologies of Figure 8. The representation of CO in the model, comparing top and center left figure, is similar (in the troposphere more so than in the stratosphere), but was not subject of this study. We compared the top row of $MOD_{CARIBIC}$ and $MOD_{RANDLOC}$ and found that R_{rel} is a good descriptor for the representativeness of one for the other. By assuming the result from the model to be valid also for measurements, we can now use the score calculated from the two model samples to determine the representativeness of $MEAS_{CARIBIC}$.

By again defining $R_{rel} = -1$ ($10\,\%$ uncertainty) as the limit for representativeness, the climatology of $MEAS_{CARIBIC}$ in Figure 8 (center left) was shaded in grey where it is not representative. The representativeness uncertainty shown in Figure 7 only serves as a first indication of the expected uncertainty when resolving monthwise. The center right panel displays the standard deviation of CO from $MOD_{RANDLOC}$. By comparing the center panels, it becomes evident that the variability specific to CO is one of the reasons for the higher representativeness uncertainty in spring, while it cannot explain all the features. The number of flights is a different reason, which explains the higher uncertainty in January, the month with the least flights (not shown).

The limit of $10\,\%$ should not be applied in general and has to be adapted to the species under consideration. This becomes evident by the bottom row in Figure 8, which shows climatologies of CO_2 and O3. CO_2 shows a small annual variation around a high background value. So $10\,\%$ uncertainty could be easily reached by a single measurement, which would certainly not be representative for the whole year. The shading for CO_2 in Figure 8 was set at a threshold of $0.3\,\%$. The high values in spring in the upper troposphere show an even lower uncertainty, the uncertainty of all data being less than $0.7\,\%$ (not shown). The opposite is true for O3, for which the threshold was set to $15\,\%$ uncertainty. Many tropospheric values in spring or at times of high gradients in the stratosphere at the beginning and end of spring have an uncertainty higher than these $15\,\%$.

As the results in Figure 7 are sorted by the variability of the species and this is linked to their lifetime in following Junge (1974), conclusions are possible for species even if they have not been explicitly considered in this study. This is true for SF_6, for example, which is measured in whole air samples by CARIBIC but was set to 0 in the model run and could therefore not be included in this study. As it is long-lived in both troposphere and stratosphere (Ravishankara et al., 1993), a climatology from CARIBIC SF_6 measurements can be considered to be representative even though it is measured only by whole air samples.

5.5 Number of flights for representativeness

One last question remains to be answered: For those substances not representative yet, how often does one have to fly in order to achieve a representative climatology?

As explained in Section 5.3, R_{rel} increases linearly with the number of flights considered, the Pearson correlation coefficient of this relationship exceeding 0.9 for all species. This was tested by cutting all paths of $MOD_{RANDPATH}$ into three flight legs and testing different numbers of these against the whole dataset. For low numbers, the relationship of R_{rel} and the number of

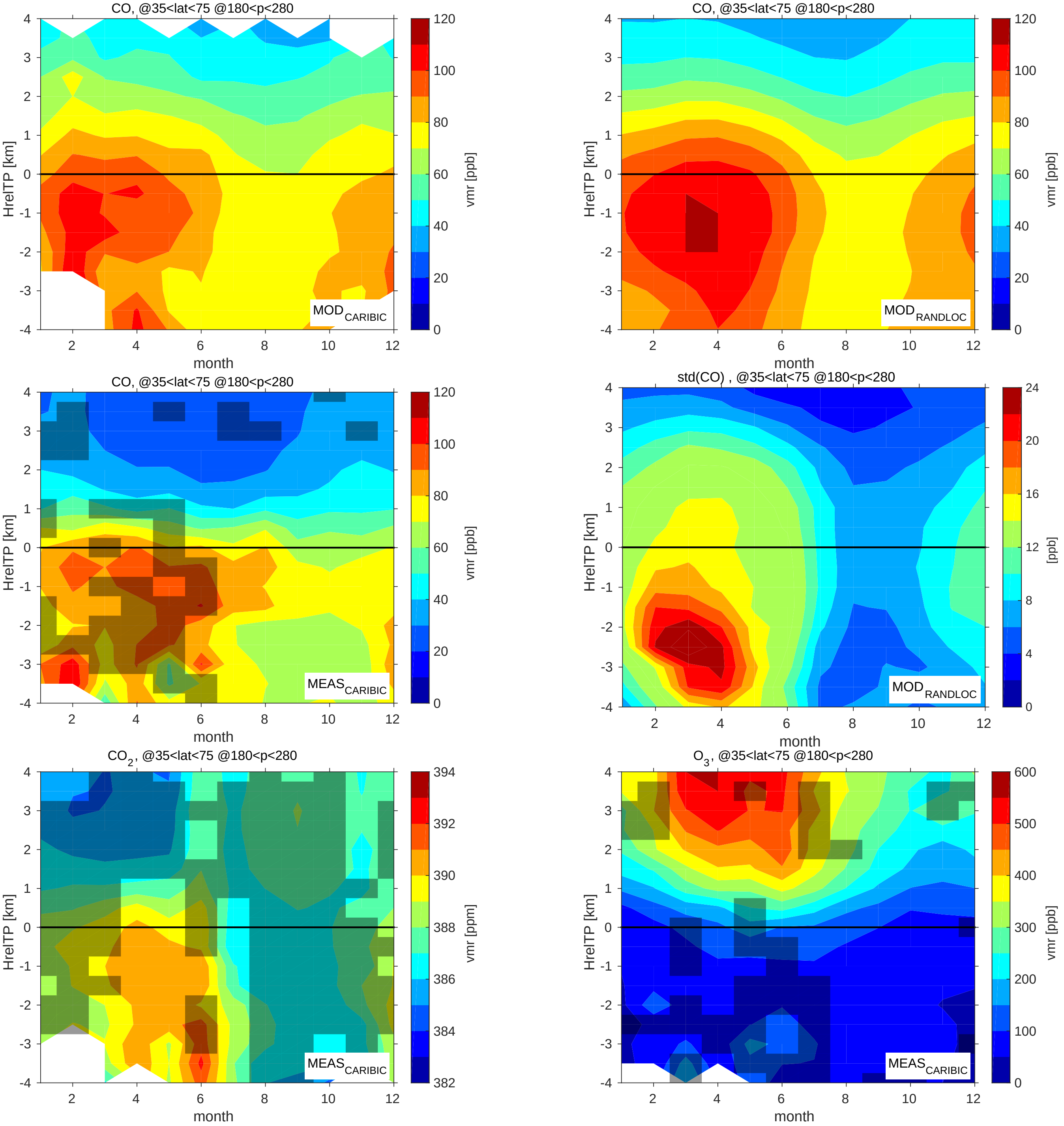

Figure 8. Climatology of CO, built from MOD$_{CARIBIC}$ (top left, including the measurement gaps in MEAS$_{CARIBIC}$ due to instrument problems), MOD$_{RANDLOC}$ (top right) and the CARIBIC measurements (MEAS$_{CARIBIC}$, center left). Areas of $10\hat{\ }R_{rel} > 0.1$, calculated from the top row, were used to shade non-representative areas in the climatology of MEAS$_{CARIBIC}$ in grey. The right center panel displays the 1σ standard deviation of CO from MOD$_{RANDLOC}$. The bottom row displays climatologies from MEAS$_{CARIBIC}$ of CO_2 (left) and O3, shaded with $10\hat{\ }R_{rel} > 0.003$ and $10\hat{\ }R_{rel} > 0.15$, respectively.

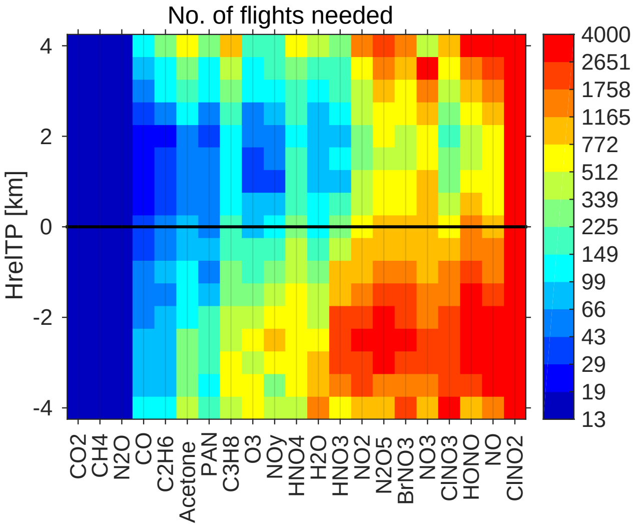

Figure 9. Number of 8 h flights necessary to reach a representativeness uncertainty of 10 % ($R_{rel} = -1$). This result was calculated using $MOD_{RANDPATH}$, the method is explained in the text.

flights is better described by a logarithmic function. This is also motivated by the study using data from a random number generator, which is presented as supplementary material to this text. So here, R_{rel} was fit to the logarithm of the number of flights. The number of flights necessary to reach a specific representativeness uncertainty, can then be read from the regression line calculated from R_{rel} and $\log(\text{number of flights})$. The result for $R_{rel} = -1$, corresponding to a representativeness uncertainty of $10\,\%$, is shown in Figure 9. It is in principle a translation of the value of R_{rel} from Figure 6 into a number of flights that are necessary to reach an uncertainty of $10\,\%$. $R_{rel} = -1$, i.e. $10\,\%$ uncertainty are again set as a mean value, which may be too high for some species, depending on their annual cycle.

As is displayed in Figure 9 and goes in line with Sec. 5.4, CARIBIC with a total number of 334 flights from 2005-2013 is already representative for many long-lived species with low variability (high τ^*), to the left of the plot. For many of the nitrogen containing species with low τ^* (to the right), data representative of a climatology is probably impossible to collect within IAGOS-CARIBIC. The necessary number of flights reach up to more than 3000 in the tropospheric heights, corresponding to almost all data in $MOD_{RANDPATH}$. For those species in the center of the plot, the representativeness uncertainty may be further reduced by flying more often, especially for those with flight numbers below 1000 like O_3, H_2O or C_3H_8. Due to their lower variability in the lower stratosphere, the climatological values of these species are already representative. In general, the uppermost and lowermost heights need more flights as they are less frequently probed by the aircraft.

6 Conclusions

After a general discussion of our representativeness concept, we apply general rules to investigate the feasibility of compiling climatologies from IAGOS-CARIBIC trace gas measurements. We answer the specific question: For which species is a climatology compiled from CARIBIC data representative for the tropopause region in mid-latitudes?

In order to answer this question, three datasets were created from a nudged model run of the chemistry-climate model EMAC: sampling the model at the geolocation of CARIBIC measurement data ($\mathrm{MOD_{CARIBIC}}$) and using the two different random samples $\mathrm{MOD_{RANDPATH}}$ (random flight tracks with similar properties as those of $\mathrm{MOD_{CARIBIC}}$) and $\mathrm{MOD_{RANDLOC}}$ (random locations).

Of these three data sets, $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDPATH}}$ are used to develop methods describing representativeness, applying the Kolmogorov-Smirnov test, a variability analysis following Kunz et al. (2008) and a relative differences test. By formulating the variability analysis and relative differences as scores ($\mathrm{R_{var}}$ and $\mathrm{R_{rel}}$ respectively), we show that they pass the two requirements we defined as having to be met by any description of representativeness: Representativeness should increase with the number of measurements and decrease with the variability of the species. Variability was defined following Junge (1974). $\mathrm{R_{rel}}$ is more applicable for answering the question, asking for the representativeness of for a climatology. It is therefore used for the analysis.

A score of $\mathrm{R_{rel}} = -1$ defines a representativeness uncertainty of $10\,\%$. It is used to discriminate the representative from the non-representative compiled climatologies. The results (using $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$) show that the data of $\mathrm{CO_2}$, $\mathrm{N_2O}$, $\mathrm{CH_4}$, CO, $\mathrm{C_2H_6}$, and $\mathrm{O_3}$ can be used to compile representative climatologies around the tropopause, while acetone, $\mathrm{NO_y}$ and $\mathrm{H_2O}$ are only usable in the stratosphere. $\mathrm{NO_2}$, NO and $\mathrm{C_3H_8}$ cannot be used for a representative climatology. Naturally, the results strongly depend on the accepted uncertainty of $10\,\%$ and would change if this limit is set to a different value. In addition, the uncertainty can be translated into a number of flights necessary to achieve representativeness. E.g. for $\mathrm{H_2O}$, 1500 to 1000 flights are necessary for a representative climatology in the upper troposphere, the number strongly decreasing with height.

The general concept of using two sets of model data to calculate the representativeness is easily applicable to other questions. One model data set should mirror the measurements, the other should be much larger, taking into account certain statistical properties of the measurement data set, so that the two data sets become comparable.

Questioning the representativeness of sampled data is important. Patterns might occur when sorting or averaging sparsely sampled data, but these patterns are not necessarily meaningful. We discuss and show a way to address this problem of representativeness by using model data. In following the methods presented here, representativeness is given a sound mathematical description, returning an uncertainty characterizing the specific dataset.

Acknowledgements. We thank all the members of the IAGOS-CARIBIC team, especially those who operate the CARIBIC container and Peter van Velthoven of KNMI who provides meteorological support. The collaboration with Lufthansa and Lufthansa Technik and the financial support from the German Ministry for Education and Science (grant 01LK1223C) are gratefully acknowledged. The CARIBIC measurement data analyzed in this paper can be accessed by signing the CARIBIC data protocol to be downloaded at http://www.caribic-atmospheric.com/.

This work was partially performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and Arts and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

References

Balzani Lööv, J., Henne, S., Legreid, G., Staehelin, J., Reimann, S., Prévôt, A., Steinbacher, M., and Vollmer, M.: Estimation of background concentrations of trace gases at the Swiss Alpine site Jungfraujoch (3580 m asl), Journal of Geophysical Research: Atmospheres (1984–2012), 113, n/a–n/a, 2008.

Brenninkmeijer, C. A. M., Crutzen, P., Boumard, F., Dauer, T., Dix, B., Ebinghaus, R., Filippi, D., Fischer, H., Franke, H., Frieß, U., Heintzenberg, J., Helleis, F., Hermann, M., Kock, H. H., Koeppel, C., Lelieveld, J., Leuenberger, M., Martinsson, B. G., Miemczyk, S., Moret, H. P., Nguyen, H. N., Nyfeler, P., Oram, D., O'Sullivan, D., Penkett, S., Platt, U., Pupek, M., Ramonet, M., Randa, B., Reichelt, M., Rhee, T. S., Rohwer, J., Rosenfeld, K., Scharffe, D., Schlager, H., Schumann, U., Slemr, F., Sprung, D., Stock, P., Thaler, R., Valentino, F., van Velthoven, P., Waibel, A., Wandel, A., Waschitschek, K., Wiedensohler, A., Xueref-Remy, I., Zahn, A., Zech, U., and Ziereis, H.: Civil Aircraft for the regular investigation of the atmosphere based on an instrumented container: The new CARIBIC system, Atmospheric Chemistry and Physics, 7, 4953–4976, doi:10.5194/acp-7-4953-2007, http://www.atmos-chem-phys.net/7/4953/2007/, 2007.

Engel, A., Bönisch, H., Brunner, D., Fischer, H., Franke, H., Günther, G., Gurk, C., Hegglin, M., Hoor, P., Königstedt, R., Krebsbach, M., Maser, R., Parchatka, U., Peter, T., Schell, D., Schiller, C., Schmidt, U., Spelten, N., Szabo, T., Weers, U., Wernli, H., Wetter, T., and Wirth, V.: Highly resolved observations of trace gases in the lowermost stratosphere and upper troposphere from the Spurt project: an overview, Atmospheric Chemistry and Physics, 6, 283–301, doi:10.5194/acp-6-283-2006, 2006.

Gettelman, A., Hoor, P., Pan, L., Randel, W., Hegglin, M., and Birner, T.: The extratropical upper troposphere and lower stratosphere, Reviews of Geophysics, 49, n/a–n/a, 2011.

Henne, S., Klausen, J., Junkermann, W., Kariuki, J., Aseyo, J., and Buchmann, B.: Representativeness and climatology of carbon monoxide and ozone at the global GAW station Mt. Kenya in equatorial Africa, Atmospheric Chemistry and Physics, 8, 3119–3139, 2008.

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J., and Buchmann, B.: Assessment of parameters describing representativeness of air quality in-situ measurement sites, Atmospheric Chemistry and Physics, 10, 3561–3581, 2010.

Jöckel, P., Sander, R., Kerkweg, A., Tost, H., and Lelieveld, J.: Technical Note: The Modular Earth Submodel System (MESSy)-a new approach towards Earth System Modeling, Atmospheric Chemistry and Physics, 5, 433–444, 2005.

Jöckel, P., Tost, H., Pozzer, A., Brühl, C., Buchholz, J., Ganzeveld, L., Hoor, P., Kerkweg, A., Lawrence, M., Sander, R., Steil, B., Stiller, G., Tanarhte, M., Taraborrelli, D., van Aardenne, J., and Lelieveld, J.: The atmospheric chemistry general circulation model ECHAM5/MESSy1: consistent simulation of ozone from the surface to the mesosphere, Atmospheric Chemistry and Physics, 6, 5067–5104, doi:10.5194/acp-6-5067-2006, 2006.

Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry Integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy, version 2.51), Geoscientific Model Development Discussions, 8, 8635–8750, doi:10.5194/gmdd-8-8635-2015, 2015.

Junge, C. E.: Residence time and variability of tropospheric trace gases, Tellus, 26, 477–488, 1974.

Köppe, M., Hermann, M., Brenninkmeijer, C., Heintzenberg, J., Schlager, H., Schuck, T., Slemr, F., Sprung, D., van Velthoven, P., Wiedensohler, A., et al.: Origin of aerosol particles in the mid-latitude and subtropical upper troposphere and lowermost stratosphere from cluster analysis of CARIBIC data, Atmospheric Chemistry and Physics, 9, 8413–8430, 2009.

Kunz, A., Schiller, C., Rohrer, F., Smit, H., Nedelec, P., and Spelten, N.: Statistical analysis of water vapour and ozone in the UT/LS observed during SPURT and MOZAIC, Atmospheric Chemistry and Physics, 8, 6603–6615, 2008.

Laj, P., Klausen, J., Bilde, M., Plass-Duelmer, C., Pappalardo, G., Clerbaux, C., Baltensperger, U., Hjorth, J., Simpson, D., Reimann, S., et al.: Measuring atmospheric composition change, Atmospheric Environment, 43, 5351–5414, 2009.

Larsen, M. L., Briner, C. A., and Boehner, P.: On the Recovery of 3D Spatial Statistics of Particles from 1D Measurements: Implications for Airborne Instruments, Journal of Atmospheric and Oceanic Technology, 31, 2078–2087, 2014.

Lary, D. J.: Representativeness uncertainty in chemical data assimilation highlight mixing barriers, Atmospheric Science Letters, 5, 35–41, 2004.

MacLeod, M., Kierkegaard, A., Genualdi, S., Harner, T., and Scheringer, M.: Junge relationships in measurement data for cyclic siloxanes in air, Chemosphere, 93, 830–834, 2013.

Matsueda, H., Machida, T., Sawa, Y., Nakagawa, Y., Hirotani, K., Ikeda, H., Kondo, N., and Goto, K.: Evaluation of atmospheric CO2 measurements from new flask air sampling of JAL airliner observations, Pap. Met. Geophys., 59, 1–17, doi:10.2467/mripapers.59.1, http://ci.nii.ac.jp/naid/130004484919/en/, 2008.

Nappo, C., Caneill, J., Furman, R., Gifford, F., Kaimal, J., Kramer, M., Lockhart, T., Pendergast, M., Pielke, R., Randerson, D., et al.: Workshop on the representativeness of meteorological observations, June 1981, Boulder, Colo, Bull. Am. Meteorol. Soc., 63, 1982.

Petzold, A., Thouret, V., Gerbig, C., Zahn, A., Brenninkmeijer, C., Gallagher, M., Hermann, M., Pontaud, M., Ziereis, H., Boulanger, D., Marshall, J., Nédélec, P., Smit, H., Friess, U., Flaud, J.-M., Wahner, A., Cammas, J.-P., and Volz-Thomas, A.: Global-scale atmosphere monitoring by in-service aircraft - current achievements and future prospects of the European Research Infrastructure IAGOS, Tellus B, 67, 2015.

Ramsey, C. A. and Hewitt, A. D.: A methodology for assessing sample representativeness, Environmental Forensics, 6, 71–75, 2005.

Ravishankara, A. R., Solomon, S., Turnipseed, A. A., and Warren, R. F.: Atmospheric Lifetimes of Long-Lived Halogenated Species, Science, 259, 194–199, doi:10.1126/science.259.5092.194, 1993.

Riese, M., Ploeger, F., Rap, A., Vogel, B., Konopka, P., Dameris, M., and Forster, P.: Impact of uncertainties in atmospheric mixing on simulated UTLS composition and related radiative effects, Journal of Geophysical Research: Atmospheres (1984–2012), 117, n/a–n/a, 2012.

Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U., and Schulzweida, U.: Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 atmosphere model, Journal of Climate, 19, 3771–3791, 2006.

Rohrer, F. and Berresheim, H.: Strong correlation between levels of tropospheric hydroxyl radicals and solar ultraviolet radiation, Nature, 442, 184–187, 2006.

Sachs, L. and Hedderich, J.: Angewandte Statistik : Methodensammlung mit R, Springer, Berlin, 13. edn., 2009.

Sander, S. P., Abbatt, J., Barker, J. R., Burkholder, J. B., Friedl, R. R., and Golden, D. M.: Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, Evaluation No. 17, 2011.

Schmid, H.: Experimental design for flux measurements: matching scales of observations and fluxes, Agricultural and Forest Meteorology, 87, 179–200, 1997.

Schutgens, N. A. J., Partridge, D. G., and Stier, P.: The importance of temporal collocation for the evaluation of aerosol models with observations, Atmospheric Chemistry and Physics, 16, 1065–1079, doi:10.5194/acp-16-1065-2016, 2016.

Stevenson, D. S., Doherty, R. M., Sanderson, M. G., Collins, W. J., Johnson, C. E., and Derwent, R. G.: Radiative forcing from aircraft NOx emissions: Mechanisms and seasonal dependence, Journal of Geophysical Research: Atmospheres, 109, doi:10.1029/2004JD004759, 2004.

Stiller, O.: A flow-dependent estimate for the sampling error, Journal of Geophysical Research: Atmospheres, 115, n/a–n/a, doi:10.1029/2010JD013934, 2010.

Stroebe, M., Scheringer, M., and Hungerbühler, K.: Effects of multi-media partitioning of chemicals on Junge's variability–lifetime relationship, Science of the total environment, 367, 888–898, 2006.