Atmos. Chem. Phys., 17, 2775–2794, 2017
www.atmos-chem-phys.net/17/2775/2017/
doi:10.5194/acp-17-2775-2017

Atmospheric Chemistry and Physics Open Access EGU

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)

Received: 29 February 2016 – Discussion started: 4 April 2016
Revised: 11 January 2017 – Accepted: 7 February 2017 – Published: 23 February 2017

Abstract. Measurement data from the long-term passenger aircraft project IAGOS-CARIBIC are often used to derive climatologies of trace gases in the upper troposphere and lower stratosphere (UTLS). We investigate to what extent such climatologies are representative of the true state of the atmosphere. Climatologies are considered relative to the tropopause in mid-latitudes (35 to 75° N) for trace gases with different atmospheric lifetimes. Using the chemistry–climate model EMAC, we sample the modeled trace gases along CARIBIC flight tracks. Representativeness is then assessed by comparing the CARIBIC sampled model data to the full climatological model state. Three statistical methods are applied for the investigation of representativeness: the Kolmogorov–Smirnov test and two scores based on the variability and relative differences.

Two requirements for any score describing representativeness are essential: representativeness is expected to increase (i) with the number of samples and (ii) with decreasing variability of the species considered. Based on these two requirements, we investigate the suitability of the different statistical measures for investigating representativeness. The Kolmogorov–Smirnov test is very strict and does not identify any trace-gas climatology as representative – not even of long-lived trace gases. In contrast, the two scores based on either variability or relative differences show the expected behavior and thus appear applicable for investigating representativeness. For the final analysis of climatological representativeness, we use the relative difference score and calculate a representativeness uncertainty for each trace gas in percent.

In order to justify the transfer of conclusions about representativeness of individual trace gases from the model to measurements, we compare the trace gas variability between model and measurements. We find that the model reaches 50–100 % of the measurement variability. The tendency of the model to underestimate the variability is caused by the relatively coarse spatial and temporal model resolution.

In conclusion, we provide representativeness uncertainties for several species for tropopause-referenced climatologies. Long-lived species like CO_2 have low uncertainties (≤ 0.4 %), while shorter-lived species like O_3 have larger uncertainties (10–15 %). Finally, we translate the representativeness score into a number of flights that are necessary to achieve a certain degree of representativeness. For example, increasing the number of flights from 334 to 1000 would reduce the uncertainty in CO to a mere 1 %, while the uncertainty for shorter-lived species like NO would drop from 80 to 10 %.

Published by Copernicus Publications on behalf of the European Geosciences Union.

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 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, spatially well-resolved data collected on a global scale are required.

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 area 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 of whether a flux tower station is representative of 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 analyzed 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 can be determined, i.e., the air mass the data are representative of. 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 analyzed 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 Vapor, 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 timescale 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 inter-seasonal timescales.

This is similar to what is done in this study: we investigate the representativeness of data for different trace gases from IAGOS-CARIBIC (see Sect. 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 Sect. 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 also assess whether the complexity of the model is similar to that portrayed by the measurements, using the variability as a measure for the complexity. We find that the variability of the model is high enough and therefore quantify the representativeness of IAGOS-CARIBIC measurement data for a climatology in the UTLS by using the two model datasets alone.

In Sect. 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 Sect. 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 Sect. 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 clima-

tologies. We then discuss the variability of the model data in comparison to that of the measurements in Sect. 5. The application of the methods to the different model samples is described in Sect. 6. After showing the result of each of the three methods separately, Sect. 6.4 discusses the representativeness of the IAGOS-CARIBIC measurement data, while Sect. 6.5 answers the question how many flights are necessary to achieve representativeness. Section 7 summarizes and concludes the paper.

2 Model and data

2.1 The observational IAGOS-CARIBIC dataset

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

During each CARIBIC flight, about 100 trace gas and aerosol parameters are measured. Some are measured continuously with a frequency between $5\,s^{-1}$ and $0.2\,min^{-1}$ and are available from the database binned to 10 s. 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, $(CH_3)_2CO$ (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). Data of N_2O, CH_4 and CO_2 were detrended by subtracting the mean of each year from the values of that year and adding the overall mean.

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. This dataset will be referred to as $MEAS_{CARIBIC}$.

As this study investigates representativeness using model data, the geolocation of the CARIBIC measurements at 10 s resolution is used. In a second step, the gaps in 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 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 11 h. Meteorology is nudged up to 1 hPa using divergence, vorticity, ground pressure and temperature from 6-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.

Boundary conditions for greenhouse gases (latitude-dependent monthly means) are taken from Meinshausen et al. (2011) and continued until 2013 from the RCP 6.0 scenario (Moss et al., 2010). Boundary conditions for ozone-depleting substances (CFCs and halons) are from the WMO-A1 scenario (WMO, 2010). Emissions for NO_x, CO, and non-methane volatile organic compounds are taken from the EDGAR data base (http://edgar.jrc.ec.europa.eu/index.php).

The setup of the model in this study is similar to that made for the run RC1SD-base-08 of the Earth System Chemistry integrated Modelling (ESCiMo) initiative, presented by Jöckel et al. (2016). It differs in vertical resolution (47 versus 39 levels), but horizontal resolution, nudging and the chemistry are the same. The study by Jöckel et al. (2016) gives a detailed description and presents validation results.

Hegglin et al. (2010) performed an extensive inter-model comparison including EMAC with the same horizontal resolution as the setup for this study. Dynamical as well as chemical metrics have been used in this study, focussing on the UTLS. Overall, they find EMAC performs well within the range of the models that were tested. The reader is referred to the study for further details.

The substances from the model used in this study are the same as those from measurements. NO_y, which is simulated in its components, is summed up from N, NO, NO_2, NO_3, N_2O_5 (counted twice because measurements of NO_y are taken by catalytic conversion), HNO_4, HNO_3, HONO, HNO, PAN, $ClNO_2$, $ClNO_3$, $BrNO_2$ and $BrNO_3$. Data of N_2O, CH_4 and CO_2 were detrended, using the same method applied to the measurements.

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 are the sample data to be seen as representative? (3) To what degree are 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 others" (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 ±4.25 km around the dynamical tropopause defined at the pressure at 3.5 PVU and in mid-latitudes with $75°\text{N} < \varphi < 35°\text{N}$.

In order to reference data to the tropopause, we use the geometric height in kilometers relative to the tropopause (HrelTP) at each data point. 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 Medium-range Weather Forecasts) 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 Fig. 1, showing a maximum right at the tropopause. Data were used within ±4.25 km around the tropopause in steps of 0.5 km.

Even though all data of trace gases (be it from model or measurements) are 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 $180\,\text{hPa} < p < 280\,\text{hPa}$. In addition, we explicitly limit pressure to this range in order to exclude data from ascents and descents of the aircraft. But since data are 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. 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 effect is illustrated in Appendix A1 with the help of the methods developed in this study.

In addition to limiting HrelTP and p, it is necessary to apply a limit to latitude φ. We limit the data by including only mid-latitudes with $75°\text{N} < \varphi < 35°\text{N}$. Tropical data with $\varphi < 35°\text{N}$ are excluded because of the considerably higher dynamical tropopause. Data with $\varphi > 75°\text{N}$ are excluded because of the different chemistry in far northern latitudes, which leads to considerably different mixing ratios for 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.

Like the limit in pressure, CARIBIC data are also limited in longitude, as the Pacific Ocean is never probed. The effect of this limit on the climatology is discussed in Appendix A2.

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 of the tropopause region in mid-latitudes?

3.2 Representative of which population?

When assessing the representativeness of the sample made up by all CARIBIC measurements (called $\text{MEAS}_{\text{CARIBIC}}$, see Sect. 2.1), 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 measurements 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 are 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 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 Fig. 1.

$\text{MOD}^{\text{regular}}_{\text{CARIBIC}}$: for the dataset $\text{MOD}^{\text{regular}}_{\text{CARIBIC}}$, 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 s for all species, independent of the time resolution in $\text{MEAS}_{\text{CARIBIC}}$. Figure 1 shows the flight paths considered in this study. Since CARIBIC also measures temperature (at 10 s resolution), the high Pearson correlation coefficient of $\rho = 0.97$ of modeled to measured temperature can serve as an indication that this interpolation leads to reasonable results, despite the coarser resolution in time and space of the model output.

$\text{MOD}^{\text{sampled}}_{\text{CARIBIC}}$: the measurement frequency for some species in $\text{MEAS}_{\text{CARIBIC}}$ is lower (e.g., those taken by whole air samples), all species contain gaps because of instrument

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

Dataset	EMAC on	Total sets	Per month	Duration	p distribution
$\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$	CARIBIC paths (2005–13)	334	up to 4 in 3 days	8–10 h	flight levels show up, $\overline{p} = 223.42$ hPa $\sigma(p) = 18.94$ hPa
$\mathrm{MOD}_{\mathrm{RANDPATH}}$	random paths	1296	12 in 28 days	24 h	adjusted gaussian, $\overline{p} = 223.42$ hPa $\sigma(p) = 18.94$ hPa
$\mathrm{MOD}_{\mathrm{RANDLOC}}$	random location	864	8 in 28 days	24 h	uniform, $\min(p) = 10$ hPa $\max(p) = 500$ hPa

problems at some point and some of the species considered by the model datasets are not measured at all. Sometimes, it is interesting to consider $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ reduced to the exact number of measurement points, i.e., reduced by all these measurement gaps. The model dataset along CARIBIC paths that has the same gaps as $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ will be referred to as $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$.

As is visible in Fig. 1 (central column), only three of the model levels lay in the pressure range sampled by CARIBIC. To have comparable statistics, $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ was compared to two random model samples.

$\mathrm{MOD}_{\mathrm{RANDPATH}}$: the dataset referred to as $\mathrm{MOD}_{\mathrm{RANDPATH}}$ is a larger set of flight paths used to sample the model. This set was mainly used to investigate the representativeness of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$. 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 km h^{-1}, the median of the speed of the true CARIBIC aircraft. The flights start at 00:00 UTC and sample the model for 24 h in 10 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 Fig. 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 hPa with a standard deviation of 18.94 hPa was used to choose the pressure value for each of the random flights. All pressure values of $p < 180$ hPa or $p > 280$ hPa were redistributed evenly between 200 and 250 hPa to exclude unrealistically high or low values and sharpen the maximum.

$\mathrm{MOD}^{3}_{\mathrm{RANDPATH}}$: the dependency of representativeness on the number of flights is an important part of this study. Each of the random paths was divided into three parts, resulting in 3888 8-h flights, the duration of a typical intercontinental flight with CARIBIC. Representativeness was then calculated with the different methods for $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and these subsamples, increasing their size by including more of the 3888 shorter random flights. This dataset of randomized shorter flights will be referred to as $\mathrm{MOD}^{3}_{\mathrm{RANDPATH}}$.

$\mathrm{MOD}_{\mathrm{RANDLOC}}$: 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 a uniform distribution from 500 to 10 hPa for each flight. Again, the datasets start at 00:00 UTC and the separate points are 10 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 $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ is representative of a climatology around the tropopause only within its pressure limits or also when expanding these limits.

As is visible in Fig. 1, the distribution in HrelTP is very similar for $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ even though the pressure is prescribed in very different ways (mean of 0.79 and 0.64 km respectively). The distribution of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ is different (mean of 0.26 km), which is due to the larger number of data from southern latitudes (not shown). The different regional sampling is one of the reasons why climatologies from $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$ differ, and this difference also affects the distribution in HrelTP.

3.3 Confidence limits of 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 variability analysis following Kunz et al. (2008) and the relative difference of two climatologies. The first method gives a yes or no answer within a chosen statistical confidence level. The

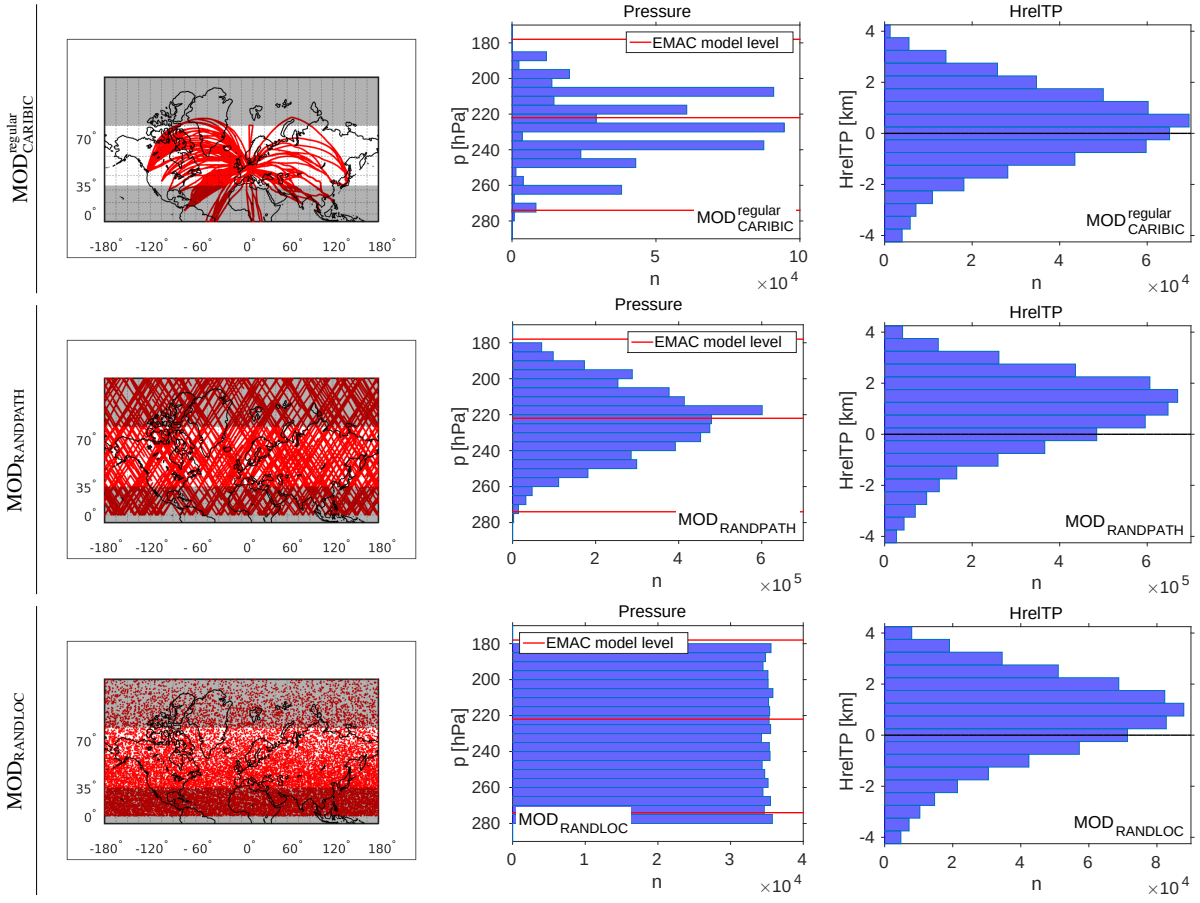

Figure 1. Flight path distribution (left), distribution of probed pressures (p, center) and height relative to the dynamical tropopause (HrelTP, right) for the three datasets $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ (top), $\mathrm{MOD}_{\mathrm{RANDPATH}}$ (center) and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ (bottom). Only parts of the paths of $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ are shown in the left column.

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 Sects. 4 and 6), the score corresponding to a confidence level.

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

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 implicitly also made by Kunz et al. (2008), as they investigate the representativeness of a smaller dataset 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

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 Sect. 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. In this study, we use the relative standard deviation σ_r as a measure for variability. It is calculated following Eq. (1) using the mean μ and standard deviation σ of each species.

$$\sigma_\mathrm{r} = \frac{\sigma}{\mu} \tag{1}$$

Figure 2 shows the sorted values of σ_r for the species considered in this study, using the full time series to calculate σ_r. It is worthwhile to note that in defining variability in this way, we closely follow Junge (1974), who showed that under

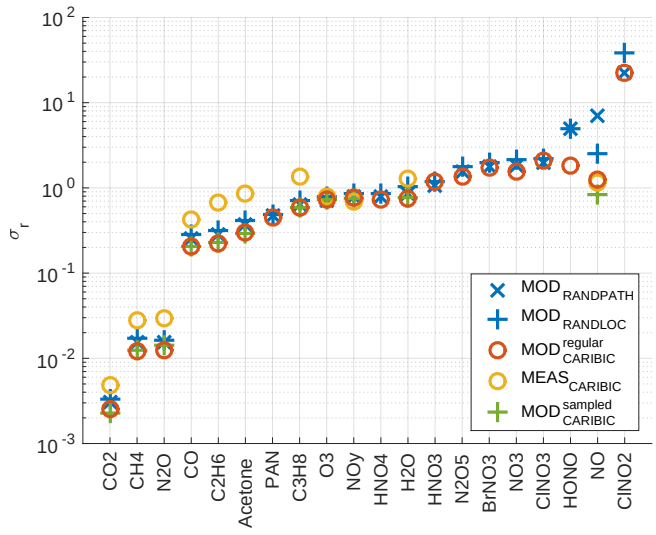

Figure 2. Variability σ_r calculated for different datasets using Eq. (1). The species are sorted by σ_r and species with low variability are listed to the left, using the values from $MOD_{RANDPATH}$ for sorting. Note that $\log_{10}(\sigma_r) = \tau^*$, see Eq. (3).

certain constraints, the relationship

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

holds, which links variability and lifetime τ using two species-dependent constants a and b. This relationship has frequently been called the Junge relationship in the past (e.g., by Stroebe et al., 2006 or MacLeod et al., 2013). And indeed, as visible in Fig. 2, longer-lived species like CO_2 or N_2O show lower variability, while shorter-lived species show higher variability.

It is important to note that the values determined from $MEAS_{CARIBIC}$ are affected by the measurement frequency in case of data sampled by whole air samples (N_2O, C_2H_6 and C_3H_8) and by gaps due to instrument problems. But the influence of these gaps is small, as can be seen by the small differences between the two values for $MOD^{regular}_{CARIBIC}$ and $MOD^{sampled}_{CARIBIC}$. $MEAS_{CARIBIC}$ has a slightly higher variability than the model datasets for most species. The relationship of model and measurement variability is discussed in more detail in Sect. 5. The model datasets are very similar, despite their different sampling patterns. They only differ for short-lived species (to the right in Fig. 2), which have a strong daily cycle, e.g., NO.

In Sect. 3.3, we defined representativeness as having to decrease with increasing variability. Because we want to emphasize the relationship of σ_r with τ and in order to differentiate this variability (calculated from the complete time series) clearly from other similar terms, we use τ^* defined in Eq. (3) to test the relationship of representativeness and variability.

$$\tau^* = \log_{10}(\sigma_r) = \log_{10}(a) - b \cdot \log_{10}(\tau) \tag{3}$$

Section 4.2 will take a closer look at variability. It will be discussed how variability depends on the timescale for which it is calculated. The values shown in Fig. 2 and used for the calculation of τ^* use the full time series, and thereby the overall variability. If shorter timescales had been considered, the values for σ_r in Fig. 2 would change, but not the order of the species that follows from the values.

So including these thoughts on variability in the question formulated at the end of Sect. 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 of 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 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:

$$\hat{D} = \max|\hat{F}_1 - \hat{F}_2| \tag{4}$$

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 $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDPATH}$ in each height bin and month. In addition to the Kolmogorov–Smirnov test, we also applied the Mann–Whitney test for the mean and Levene's and the Brown–Forsythe test for variance (see again Sachs and Hedderich, 2009). All results of applying these tests are presented in Sect. 6.1.

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, and Kunz et al. (2008) then applied it to aircraft data. A time series 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 characteristic for the development of variance in time.

Instead of considering variance in each time slice, we use the relative standard deviation $\sigma_r = \frac{\sigma}{\mu}$, which is the definition

of variability following Junge (1974). It is calculated in each time slice and the mean gives the value for the corresponding timescale. In the following, timescale therefore refers to the length of the interval in time in which the variability is calculated. By scaling the standard deviation σ with the mean μ, different species become comparable. Being a combination of variability as defined by Junge (1974) and the variance analysis introduced by Rohrer and Berresheim (2006), this method is called variability analysis in the following paragraphs.

Figure 3 shows the variability analysis for CO just below the tropopause for $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$, $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$. The timescale changes from about 5 min to 5 years 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 timescales of 1 year. The variability of $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ is larger than that of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ on almost all timescales. For timescales of 30 days and more, however, the lines of all three datasets run in parallel, showing an increase up to 1 year, after which 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 and 1.5 years. For timescales below 30 days, the distribution of flights in 1 month dominates the variability analysis. $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ includes only up to four flights on consecutive days, and the mean variability does not decrease when going to timescales between 30 and 4 days, while in $\mathrm{MOD}_{\mathrm{RANDPATH}}$, continuously fewer data are included in each time slice, leading to a continuous drop in the variability. For timescales of less than 1 day, the data come from a single flight, showing another drop in variability that is linked to using data from geographic regions that are ever more close in the case of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. Since the variability analysis is so closely linked to the distribution in time and space, the variability analysis of $\mathrm{MOD}_{\mathrm{RANDLOC}}$ shows an almost constant value for timescales shorter than 30 days until timescales shorter than 1 day are reached, after which the variability also drops.

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 how representative one dataset is of another. A score $R^{t,h}_{\mathrm{var}}$ describing the representativeness is defined from the difference of the values of the variability analysis, using the following equation:

$$R^{t,h}_{\mathrm{var}} = \log_{10}\left(\left|\overline{\left[\frac{\sigma^{t,h}_1}{\mu^{t,h}_1}\right]} - \overline{\left[\frac{\sigma^{t,h}_2}{\mu^{t,h}_2}\right]}\right|\right), \tag{5}$$

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

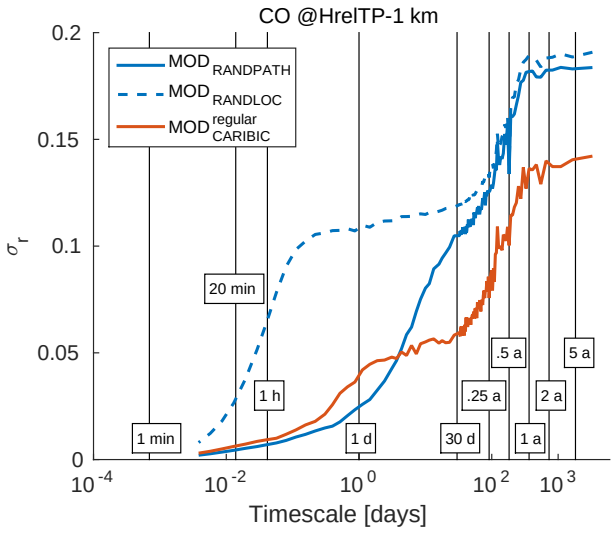

Figure 3. Variability analysis calculated for CO for $\mathrm{MOD}_{\mathrm{RANDPATH}}$, $\mathrm{MOD}_{\mathrm{RANDLOC}}$ and $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ at HrelTP $= -1$ km (1 km below the tropopause). The timescales used to calculate R_{var} using Eq. (5) are indicated by vertical lines.

Decreasing values of $R^{t,h}_{\mathrm{var}}$ mean better representativeness, the value always being negative. Depending on t, the representativeness in different timescales can be evaluated. We used timescales of 30 days, 0.25, 0.5, 1, 2 and 5 years to calculate $R^{t,h}_{\mathrm{var}}$. When applying this method to all height bins, a profile in R^{t}_{var} 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 Sect. 3.3. The results of testing this will be presented in Sect. 6.2.

4.3 Relative differences

The third approach in assessing representativeness is to analyze the relative differences between the climatologies from two differently large datasets. The procedure is summarized in Eq. (6):

$$R^{h}_{\mathrm{rel}} = \log_{10}\left(\frac{1}{12}\sum_{m=1}^{12}\frac{|\mu^{m,h}_1 - \mu^{m,h}_2|}{\mu^{m,h}_2}\right), \tag{6}$$

which was applied to each height bin h. $\mu^{m,h}_x$ 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 R_{rel}, similar to the score R^{t}_{var} calculated from the variability analysis. Contrary to the Kolmogorov–Smirnov test or the variability analysis, this test statistic does not contain any in-

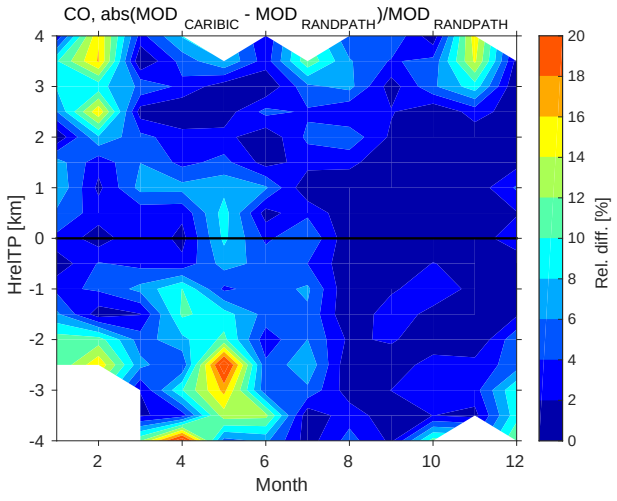

Figure 4. Relative differences of CO for $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. This is the basis used to calculate R_{rel}.

formation on the underlying distribution, because it uses only the mean in each bin.

Figure 4 shows an example of relative differences between CO from $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and the larger dataset $\mathrm{MOD}_{\mathrm{RANDPATH}}$. The differences are small, mostly below an absolute value of 0.15. R_{rel} is defined (in Eq. 6) as the logarithm to the base 10 of the mean over all months (not shown). The score increases towards the top and bottom in Fig. 4 due to fewer data there. Like for R^t_{var}, decreasing values in R_{rel} mean better representativeness. And like R^t_{var}, R_{rel} has to be tested for passing the requirements of being related to the number of samples and variability (see Sect. 3.3) in order to be acceptable as a score for representativeness. The results of testing this will be discussed in Sect. 6.3.

Other than just as a score, the value of R_{rel} can be understood as the average uncertainty for assuming the climatology of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ 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 Sect. 6.4.

5 Model and measurement variability

Representativeness was assessed using only model data in this study, yet the final goal was to investigate the climatological representativeness in mid-latitudes of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$. $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ are used as a placeholder for $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ and compared to other model datasets ($\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$) in the analysis. The results derived from these model datasets will be interpreted for $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ in Sect. 6. This means that conclusions drawn from model data alone will be applied to measurements.

To justify this reasoning, it is important to investigate the differences between the model and the real atmosphere. It is not crucial that the model reproduces the exact values of the measurements, but rather that the complexity for each species in the model is similar to the real complexity. This will be investigated in the following two sections. The variability of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ will be used as an indicator of its complexity and compared to the variability of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$. Similarly to Eq. (1), we use the relative standard deviation $\sigma_{\mathrm{r}} = \sigma/\mu$ as a measure for variability when comparing model and measurements. Variability of a certain timescale, e.g., 20 min, will be referred to as 20 min variability in the following, and accordingly for other timescales.

5.1 Influence of short timescales on the climatological mean

All model datasets have been created from gridded data files with a certain resolution (2.8° or about 200 km, see Sect. 2.2). Considering the median airspeed of the CARIBIC aircraft of 885.1 km h^{-1}, this model resolution corresponds to a timescale of about 20 min. $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ has a time resolution of up to 10 s, depending on the instrument. Model data has been linearly interpolated to this high 10 s resolution, but this does not introduce the variability that is present in the measurements. The 20 min variability is therefore always larger in $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ than in $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$. To what extent this small-scale variability influences the climatological values is investigated here.

The data of each species in $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ was smoothed by interpolating between the 20 min mean values. These smoothed measurements then resemble time series taken from model output with a resolution of about 200 km. With this smoothed dataset, it is possible to determine the influence of the small-scale variability on the climatological mean values. The exact method of smoothing is presented in Appendix B. The smoothed dataset will be referred to as $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ in the following.

Climatological mean values of $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ were compared to mean values from $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ with the full variability, thereby determining the influence of the reduced 20 min variability. A similar influence is expected by the coarse model resolution. The mean relative difference of the climatologies for different species between $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ is displayed in Fig. 5. The differences depend strongly on the species. Those species that are measured by air samples (N_2O, C_2H_6 and C_3H_8) have been shaded in grey, since they contain very little data far above and below the tropopause and are therefore not considered in this section.

The mean relative differences are smaller than 1 % for the long-lived species to the left and reach 10–20 % at most for the other species. The largest values appear where the mixing ratios of the species are small and vertical gradients are strong, i.e., in stratospheric CO, acetone or H_2O and tropospheric O_3. For example, H_2O has very low stratospheric mixing ratios that are reached in small-scale in-

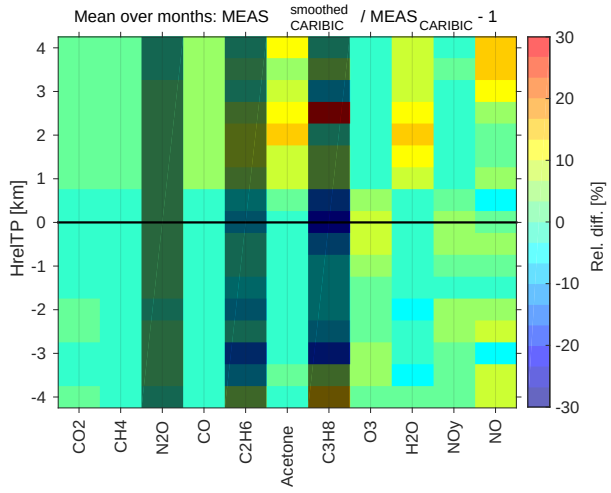

Figure 5. Mean relative differences of $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}_{\mathrm{CARIBIC}}$. $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ has been smoothed by interpolating between the 20 min mean values, the exact method being presented in Appendix B. The relative differences correspond to the error in the climatologies of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ due to the coarse model resolution. N_2O, C_2H_6 and C_3H_8 are measured by air samples with a low measurement frequency and are therefore not considered here.

trusions of stratospheric air encountered during flight. If these small-scale structures are smoothed out, the mean values become larger and the difference of $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ is large and positive.

The relative differences show the small influence of a lower 20 min variability on climatological mean values. This therefore shows that the coarse model resolution does not in principle lead to very large errors in climatological mean values. Nevertheless, the model could have other deficiencies in the description of the different species. These are made visible in the following section by comparing model and measurement variability directly.

5.2 Comparing model and measurement variability

In this section, the variability of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ is compared directly to that of $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$. For this dataset, the 20 min variability of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ has been reduced; see the preceding section. As this study argues completely within the model world, it is important that the model has similar values for the variability, which is used as an indicator of the underlying complexity. If the model cannot reproduce the measurement variability at all, it is not plausible why conclusions on representativeness drawn from model data should also be true for the real atmosphere.

As has been discussed in Sect. 4.2, variability depends on the timescale for which it is considered. In order to evaluate the model performance, we compare σ_{r} on timescales of 30 days and 1 year. Variability calculated in a timescale of 30 days typically includes data from 4 flights, so this

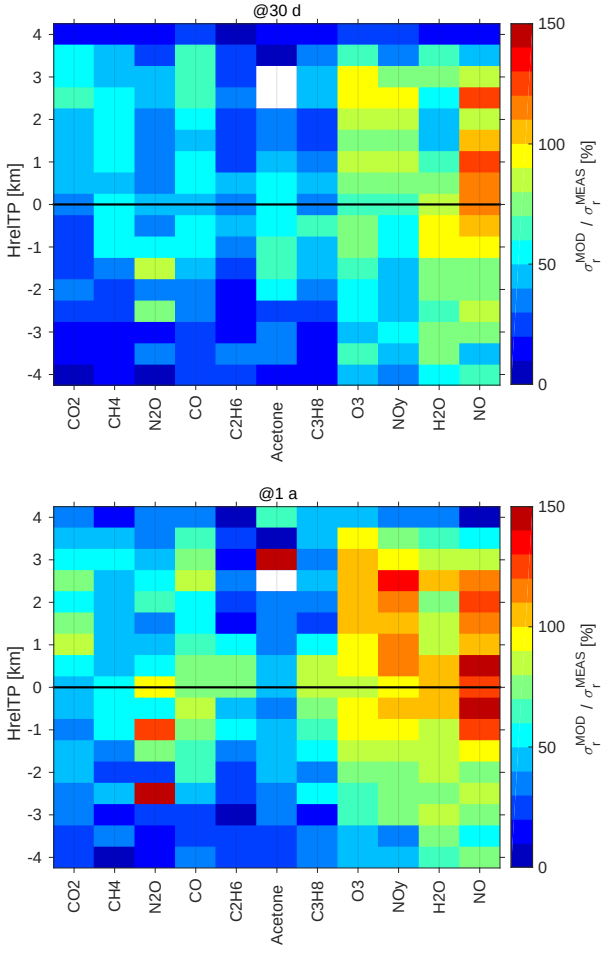

Figure 6. $\sigma_{\mathrm{r}}^{\mathrm{MOD}}/\sigma_{\mathrm{r}}^{\mathrm{MEAS}}$ given in percent for timescales of 30 days (top) and 1 year (bottom), where MOD stands for $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ and MEAS stands for $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$. Values greater than 50 % indicate the high model complexity.

is a measure for the atmospheric variability on the global, large-scale dynamics. Variability calculated in a timescale of 1 year gives a good impression of the annual cycle, as it includes data from many flights and different years. Figure 6 shows $\sigma_{\mathrm{r}}^{\mathrm{MOD}}/\sigma_{\mathrm{r}}^{\mathrm{MEAS}}$ for timescales of 30 days (top) and 1 year (bottom), using the datasets $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$.

Figure 6 shows that the variability in the measurements reached by the model differs between species. In general, the variability reached for shorter-lived species better fits that of the measurements. Short-lived species also undergo a more complex chemistry in the model, which adds variability. The 30 days variability shown in Fig. 6 (top) reveals to what extent the model is able to capture variability related to the large-scale dynamics. Most species reach 40–80 %. NO is very short lived and strongly determined by its daily cycle,

which is the reason why the variability in the model reaches higher values.

The timescale of 1 year shows the variability that represents seasonality. The model does a better job for this timescale than for 30 days, with short-lived species and CO_2 reaching well over 60 % of the variability, approaching 100 % for some species. Here again, the model chemistry increases the variability for the shorter-lived species to the right. There are species that are not as well represented, while this also depends on the height considered. N_2O, C_2H_6 and C_3H_8 are also affected by the lower measurement frequency, as they are only measured in air samples.

The model variability is influenced by many factors including the dynamics and the representation of the chemistry and of the sources included in the model. The limited horizontal and vertical resolution also plays a role, even though $\mathrm{MEAS}_{\mathrm{CARIBIC}}^{\mathrm{smoothed}}$ is used as a reference for the comparison. If compared to the original $\mathrm{MEAS}_{\mathrm{CARIBIC}}$, the percentages of variability reached by the model drop by 10–20 % (not shown). It is beyond the scope of this paper to further disentangle what causes the deficiencies of the model and what leads to the differences between the species.

As is shown in Fig. 6, the model reaches more than 50 % of the variability of the measurements, depending on the species and timescale. In general, the model variability can be increased by using a run with a higher resolution, because a decrease in spatial resolution requires a decrease in the time step of the integration. The variability of the measurements in each bin of HrelTP is also influenced by the choice of reference for HrelTP. For this study, HrelTP has been derived from model output fields from ECMWF at a resolution of 1° ($\approx$ 110 km), while the measurement data have a much higher resolution ($\approx$ 2.5 km, see Sect. 2.1). The highly variable measurements are then sorted into bins of coarsely resolved HrelTP, artificially increasing the variability of the measurements in each bin of HrelTP. To a lesser extent, this also affects $\mathrm{MEAS}_{\mathrm{CARIBIC}}^{\mathrm{smoothed}}$. Considering these complementing thoughts on the model and measurement variability, the fraction of variability reached by the model (more than 50 %) justifies the application of the representativeness evaluated from the model to $\mathrm{MEAS}_{\mathrm{CARIBIC}}$.

6 Results

Here, we first present the results of the application of the Kolmogorov–Smirnov test (Sect. 6.1), the variability analysis (Sect. 6.2) and the relative difference (Sect. 6.3) to $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. All have to be related to the number of flights and the variability of the species as discussed in Sect. 3.3. These methods have also been applied to data not from an atmospheric model but from a random number generator, leading to equivalent results. These are presented as a Supplement to the article. Section 6.4 interprets the results by species as a representativeness uncertainty. Finally, Sect. 6.5 answers the question of how many flights are necessary to achieve a certain degree of representativeness. In addition, Appendix A discusses the influence of the limitations in longitude and in pressure which are inherent in the CARIBIC dataset.

6.1 Applying the Kolmogorov–Smirnov test

The application of the Kolmogorov–Smirnov test to $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$ yields a first important result. Independent of the trace gas and height considered, the result is always negative (not shown). This means that the data in each bin of $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ are not representative of the corresponding bin in $\mathrm{MOD}_{\mathrm{RANDPATH}}$ when defining representativeness by a positive result of the Kolmogorov–Smirnov test. This is also true if the data are 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 have already been reported by Kunz et al. (2008). On the one hand side this could mean that $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ is simply not representative of $\mathrm{MOD}_{\mathrm{RANDPATH}}$. But if the other methods presented here are considered, the conclusion seems more appropriate that the Kolmogorov–Smirnov test is simply not the appropriate way to answer the question. It can be considered as too strict for the type of data and the question considered here. This is also the result of a sensitivity study, which is discussed as a Supplement to this text.

In addition to binning into 12 months (January to December), we have also tested $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$ when first binning into separate months (108 months in 9 years) and then using this monthly mean data to compile a climatology. For this monthly mean data, the Kolmogorov–Smirnov test does give a positive result in some heights and months. But no meaningful pattern could be determined from the results. In particular, the result does not depend on τ^* (not shown). The same is true for the Mann–Whitney test for the mean and Levene's and the Brown–Forsythe test for variance. They give no positive result for data binned directly into months. The result is positive for some months and heights if data are first binned into separate months the monthly mean data used for testing. The positive results seem randomly distributed and no relationship to τ^* could be found. These tests therefore also seem not to be suitable for answering the question of representativeness.

6.2 Applying the variability analysis

This section presents the results of the application of the variability analysis to $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. Equation (5) was applied for different timescales (30 days, 0.25, 0.5, 1, 2 and 5 years) to calculate R_{var}. The results are exemplarily discussed for a timescale of 1 year, shown in Fig. 7, in

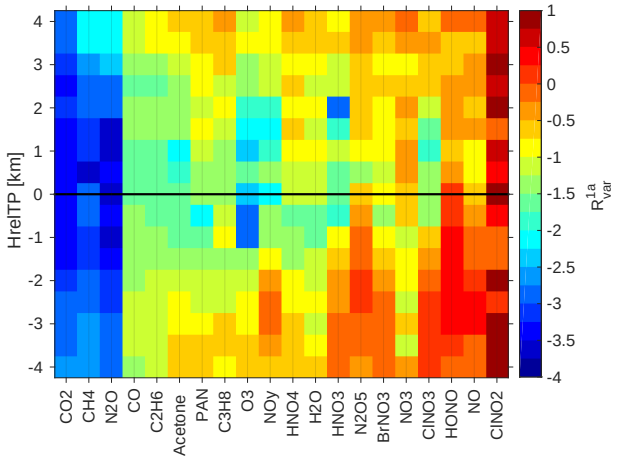

Figure 7. R_{var} calculated according to Eq. (5) for a timescale of 1 year for all species in all height bins, using $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDPATH}$. Low values indicate small differences in variability.

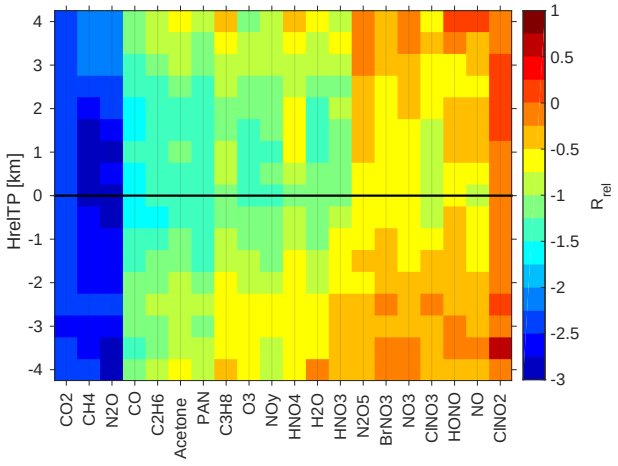

Figure 8. R_{rel} calculated according to Eq. (6) for all species in all height bins, using $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDPATH}$. Low values indicate small differences in climatological mean values.

which the results are sorted using the values of τ^* displayed in Fig. 2.

R_{var} shows a strong dependency on τ^*. This is visible from Fig. 7, in which the results are sorted with decreasing values of τ^* (from Fig. 2), i.e., with increasingly higher atmospheric variability 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 timescale. R_{var} also shows a strong relationship to the number of samples: the number of data in both $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDPATH}$ decreases below and above the tropopause, and R_{var} follows suit for practically all species.

The relation of R_{var} and the number of flights was also tested by using $MOD^{3}_{RANDPATH}$, defined in Sect. 3.3. R_{var} was correlated with the number of flights for each species and height. When investigating a linear relationship, the Pearson correlation coefficient was approximately $|\rho| \approx 0.75$ for the timescale of 5 years, increasing continuously when considering shorter timescales to $|\rho| \approx 0.95$ for the timescale of 30 days. Considering a logarithmic relationship increases the goodness of fit for longer timescales, while it decreases that for shorter timescales ($|\rho| \approx 0.85$ for both 5 years and 30 days).

R_{var} therefore passes the requirements of being inversely related to τ^* and directly to the number of included data points and flights. Figure 7 can therefore be used to judge the representativeness of $MOD^{regular}_{CARIBIC}$ for $MOD_{RANDPATH}$.

This shows that by using the relative standard deviation (Eq. 5) 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 timescales 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.

6.3 Relative differences

R_{rel} was calculated for each species in each height bin according to Eq. (6); results are presented in Fig. 8.

Figure 8 shows how low variability (decreasing to the left, values taken from Fig. 2), is linked with good representativeness (low values in R_{rel}). R_{rel} decreases linearly with increasing variability τ^* with a high Pearson correlation coefficient greater than 0.95 for all height bins (not shown). As visible in Fig. 8, R_{rel} also decreases with the number of data points, which maximizes just around the tropopause and decreases above and below it (see Fig. 1).

This dependance on the number of data points was also tested by using $MOD^{3}_{RANDPATH}$, described in Sect. 3.3. The Pearson correlation coefficient ρ between the number of shorter random flights and R_{rel} was $\rho \approx 0.95$ for all species in all heights. Less variable species like CO_2 show a better relationship with the logarithm of the number of flights. 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 Fig. 7, Fig. 8 can therefore be used to judge the representativeness of $MOD^{regular}_{CARIBIC}$ for $MOD_{RANDPATH}$. R_{rel} can be transformed into a relative difference in percent, by taking R_{rel} to the power of 10. A score of -2 stands for a mean relative difference of 1 %.

The score that discriminates the 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 are 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 separated from those species that are not representative using the results from Fig. 8. But the score of −2 is arbitrary. If it is reduced to −1.5 (roughly 3 % relative difference), $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ can be seen as representative for many more species.

6.4 Representativeness uncertainty of the CARIBIC measurement data

The last sections have shown R_{rel} (see Eq. 6) and R_{var} (see Eq. 5) to be adequate scores to describe representativeness. After reconsidering the question we asked in the Sect. 3.1 (is a climatology compiled from CARIBIC data representative of 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. A further discussion of R_{var} is beyond the scope of this paper. As noted in Sect. 4.3, R_{rel} is also comprehensible as an uncertainty for using the smaller dataset to compile a climatology and will be called representativeness uncertainty correspondingly.

In order to assess the uncertainty for accepting CARIBIC measurement data to create a climatology, model data have to contain the same number of data as $\mathrm{MEAS}_{\mathrm{CARIBIC}}$, which is why $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ (see Sect. 2) will be used in the following. In addition, $\mathrm{MOD}_{\mathrm{RANDLOC}}$ (see Table 1) was used for reference, as it has a random sampling pattern and represents the full model state, independent of the sampling pressure. The limits in pressure were again set to $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$. The resulting R_{rel} is shown in Fig. 9. 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 complexity of the model is sufficiently high for each species. This has been shown by comparing the variability of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ for different timescales (see Sect. 5). The discussion in the following paragraphs is therefore also valid for the real atmosphere, even though results have been derived from model data alone. Figure 9 answers the question we asked in Sect. 3.2: for which species is a climatology compiled from CARIBIC data representative of the tropopause region in mid-latitudes?

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 a lot for a species that is more or less constant, while it is a lot for a species with a strong seasonality. The following paragraphs discuss representativeness by species, not explicitly considering the seasonal variations for each species. The monthly resolved climatologies of CO, CO_2 and O_3 will be discussed exemplarily at the end of this section.

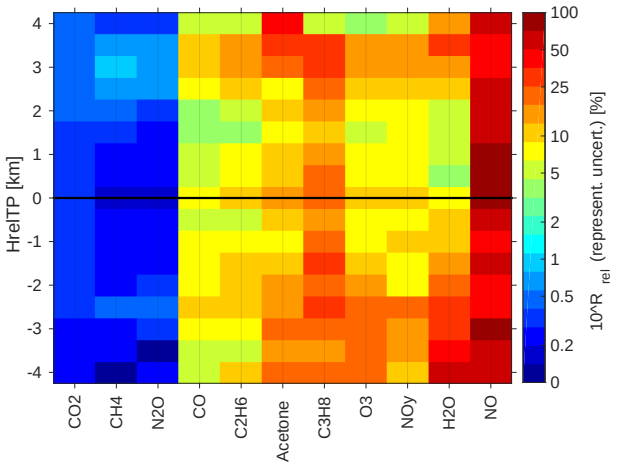

Figure 9. Representativeness uncertainty for using the CARIBIC data (that is 334 long-distance flights, see Table 1) to compile a climatology: $10^{R_{\mathrm{rel}}}$ calculated from $\mathrm{MOD}_{\mathrm{RANDLOC}}$ and $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$. Low values indicate small representativeness uncertainties. N_2O, C_2H_6 and C_3H_8 are measured from air samples, which increases the uncertainty, especially for C_3H_8.

Many of the species that sum up to NO_y in the model are not actually measured by CARIBIC and therefore are not displayed in Fig. 9. In general, the representativeness uncertainty is lowest where there are most measurements, which is just around the tropopause (see Fig. 1). This effect overlays the physical reasons for the different uncertainties for the considered species.

NO has the highest uncertainty of 90 %. We propose two possible reasons: on the one hand, there are many gaps in the observations. On the other hand, NO is 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 variability 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 is therefore quite trustworthy.

It is interesting to note that C_2H_6 and C_3H_8, both collected in whole air samples, still reach uncertainties comparable to those of other species in their range of τ^*. This is due to the fact that these are moderately long-lived species for which only a smaller 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 unnoticeable 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 Sect. 2.1) all have representativeness uncertainties of less than 0.4 %, which is lower than their seasonal variability. This is interesting especially for N_2O, which is measured only in the whole air samples.

As an example and summary, the representativeness uncertainty will be applied to climatologies of CO, CO_2 and O_3, shown in Fig. 10. CO is shown for $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ (top left, panel a), $\mathrm{MOD}_{\mathrm{RANDLOC}}$ (top right, panel b) and CARIBIC measurements ($\mathrm{MEAS}_{\mathrm{CARIBIC}}$, center left, panel c). The white space in these figures is there for three possible reasons: the aircraft could have never flown in that bin, there could be measurement gaps in CO or there could be a gap in HrelTP. The measurement gaps of CO and HrelTP from $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ have been mapped onto $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$, but HrelTP differs slightly and therefore also the white space. The representation of CO in the model, comparing top and center left figure (panels a and c), is similar to measurements (in the troposphere more so than in the stratosphere), but was not subject of this study. We compared the top row ($\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$, panels a and b) and found that R_{rel} is a good descriptor for the representativeness of one for the other. By accepting 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 uncertainty of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$.

By again defining $R_{\mathrm{rel}} = -1$ (10 % uncertainty, one-third of the seasonal variation) as the limit for representativeness, the climatology of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ (Fig. 10, center left, panel c) was shaded in grey where it is not representative. The representativeness uncertainty shown in Fig. 9 only serves as a first indication of the expected uncertainty when resolving month-wise. The center right panel (panel d) displays the standard deviation of CO from $\mathrm{MOD}_{\mathrm{RANDLOC}}$. By comparing the center panels (c and d), 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 Fig. 10 (panels e and f), which shows climatologies of CO_2 and O_3. 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 of the whole year. The shading for CO_2 in Fig. 10 was set at a threshold of 0.3 %, again just above one-third of the seasonal variation. 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 O_3, for which the threshold was set to 15 % uncertainty (around one-fourth of the seasonal variation). 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 Fig. 9 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.

Two limitations are inherent in the CARIBIC data: the Pacific Ocean is never sampled and the pressure is limited to flight levels. The influence of both these limitations is discussed in Appendix A.

6.5 Number of flights for representativeness

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

This question can be answered with the help of $\mathrm{MOD}^{3}_{\mathrm{RANDPATH}}$. Figure 11 shows the representativeness uncertainty for some species and different numbers of flights. As has been discussed in Sect. 6.4, the yearly variation of a species is one of the factors that determines the threshold of the uncertainty with which the species can be considered to be representative.

For example, for (detrended) CO_2, the mean value of $\mathrm{MOD}_{\mathrm{RANDLOC}}$ is 385.7 ppmv with a yearly variation of 2.5 to 3.5 ppmv. A representativeness uncertainty of at least 0.5 % has therefore to be set as the minimum threshold for CO_2. This can be reached with only a few flights, much less than those included in $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$, indicated by the dashed line in Fig. 11 at 334 flights.

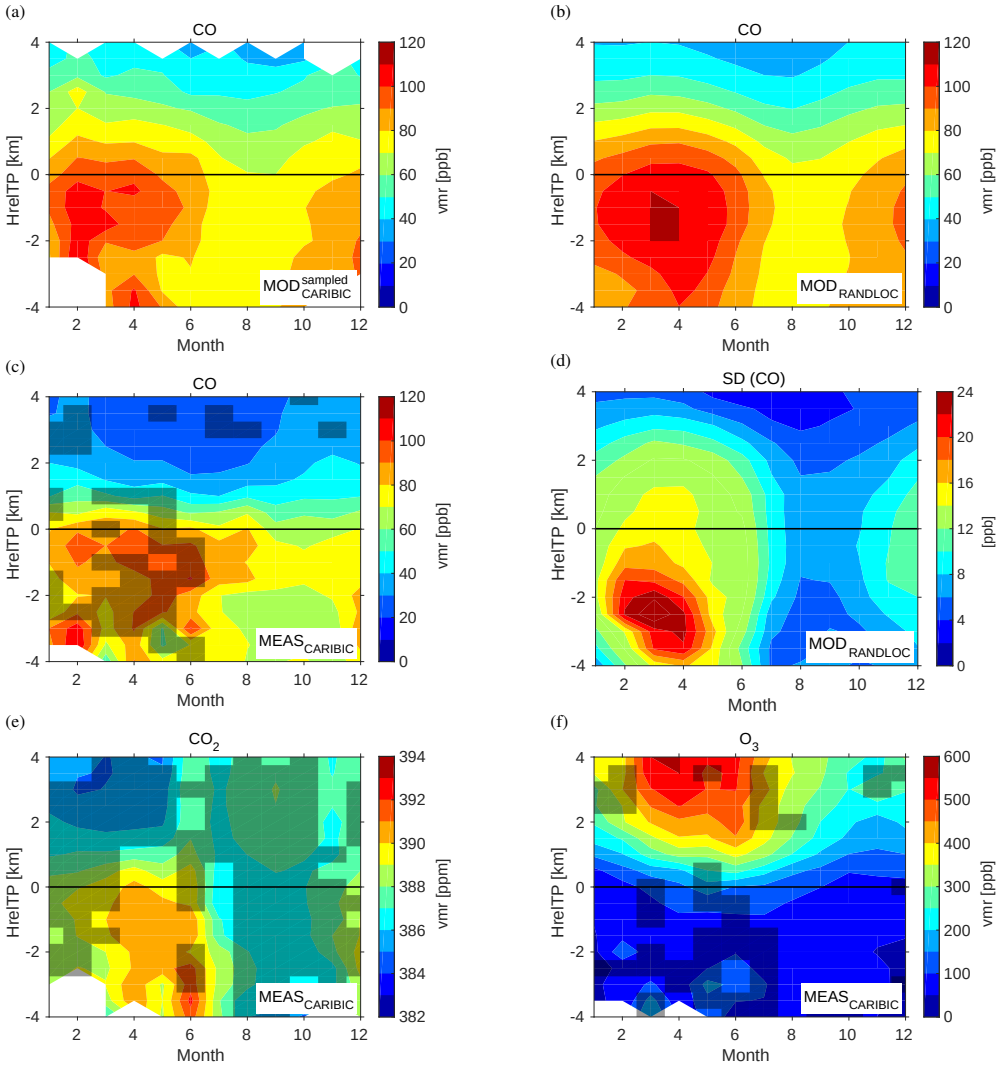

Figure 10. Climatology of CO, built from $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ **(a)**, $\mathrm{MOD}_{\mathrm{RANDLOC}}$ **(b)** and the CARIBIC measurements ($\mathrm{MEAS}_{\mathrm{CARIBIC}}$, **c**). Areas of $10^{R_{\mathrm{rel}}} > 0.1$, calculated from the top row, were used to shade non-representative areas in the climatology of $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ in grey. Panel **(d)** displays the 1σ standard deviation of CO from $\mathrm{MOD}_{\mathrm{RANDLOC}}$. The bottom row **(e, f)** displays climatologies from $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ of CO_2 (left) and O_3, shaded with $10^{R_{\mathrm{rel}}} > 0.003$ and $10^{R_{\mathrm{rel}}} > 0.15$, respectively.

For O_3, on the other hand, the yearly cycle proposes an uncertainty of 50 % or more. While this is the minimum value to reproduce the yearly cycle at all, it may still not be sufficient for the application. With the number of CARIBIC flights, the uncertainty in O_3 is already low (< 5 % in this height), while the uncertainty is continuously reduced if the number of flights increases.

As is indicated by Fig. 11, highly variable species like NO need many flights in order for their climatologies to reach low uncertainties. Even 1000 flights, approximately 10 more years of flying the CARIBIC observatory, will not reduce the uncertainty below 10 %.

Other species that are not included in Fig. 11 can be deduced from their value of τ^* with the help of Fig. 2. Those species measured in air samples need even more CARIBIC flights than indicated by the number in Fig. 11, as the measurement frequency is much lower.

7 Conclusions

We describe and assess the degree of climatological representativeness of data from the passenger aircraft project IAGOS-CARIBIC. After a general discussion of the concept of representativeness, we apply general rules to investigate whether climatologies from IAGOS-CARIBIC trace gas measurements can be seen as representative. We answer the specific question: for which species is a climatology com-

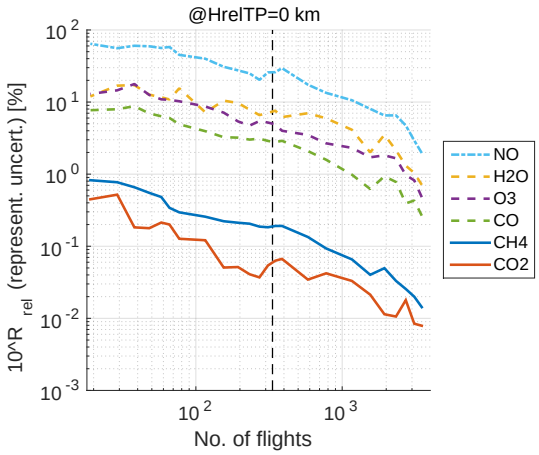

Figure 11. Representativeness uncertainty for different numbers of flights for some species. The number of flights in $\mathrm{MEAS_{CARIBIC}}$ is indicated by the vertical dashed line. Other species can be deduced from their value of τ^* with the help of Fig. 2.

piled from CARIBIC data representative of the tropopause region in mid-latitudes?

In order to answer this question, four datasets were created from a nudged model run of the chemistry–climate model EMAC. Two datasets sample the model at the geolocation of CARIBIC measurement data ($\mathrm{MOD^{regular}_{CARIBIC}}$ and $\mathrm{MOD^{sampled}_{CARIBIC}}$). These datasets are contrasted to the much larger datasets $\mathrm{MOD_{RANDPATH}}$ (random flight tracks with similar properties as those of $\mathrm{MOD^{regular}_{CARIBIC}}$) and $\mathrm{MOD_{RANDLOC}}$ (random locations).

As a first step, we demonstrate that these model datasets are appropriate to answer our question, which asks for the representativeness of CARIBIC measurement data. In order to justify the validity of the conclusions drawn from model data to the measurements, we compare model and measurement variability, using the variability as an indication of the model's ability to reproduce changes in space and time. To compare like with like, variability on scales smaller than the model resolution is removed from the measurements. With this prerequisite the model reproduces 50–100 % of the variability of the measurements, depending on timescale, height relative to the tropopause and species. This is sufficient to transfer our results from the model world to the real atmosphere considering the coarse resolution of the model and of the data used for binning the measurements into height relative to the tropopause.

Three methods to describe representativeness are developed and applied: (i) the Kolmogorov–Smirnov test (and the Mann-Whitney, Brown–Forsythe and Levene's test), (ii) variability analysis following Kunz et al. (2008) and (iii) a test interpreting the relative difference between two datasets. Two fundamental requirements are essential for representativeness: its increase (i) with the number of measurements and (ii) with decreasing atmospheric variability of the species, which is related to atmospheric lifetime following Junge (1974). By formulating the variability analysis and relative differences as scores (R_{var} and R_{rel} respectively), we demonstrate that they pass these two requirements, while the statistical tests are all too strict. R_{rel} (describing the representativeness of a climatology) is better suited for answering the question and is therefore used in the remaining analysis.

The score R_{rel} is easily converted to a representativeness uncertainty in percent and this measure is used in the discussion. The results show that CO_2, N_2O and CH_4 have very low uncertainties (below 0.4 %). CO, C_2H_6, and O_3 reach higher values (5–20 %), but can still be used to compile representative climatologies around the tropopause. NO_y and H_2O are only usable in the lower stratosphere (uncertainties of 5 to 8 % there, higher elsewhere), while NO and C_3H_8 cannot be used for a representative climatology (uncertainties of 25 % and more). Naturally, the interpretation of results strongly depends on the chosen threshold uncertainty and should depend on the seasonal variability of the species under consideration. This is demonstrated by setting different limits for climatologies of CO_2, CO and O_3.

In addition, the uncertainty can be translated into a number of flights necessary to achieve representativeness. This is demonstrated for some species by showing the relationship of the number of flights and the representativeness uncertainty. For long-lived species like CO_2 and CH_4, the 334 IAGOS-CARIBIC flights used in this study already provide enough data, while short-lived species like NO need around 1000 flights to reduce the uncertainty to 10 %, sufficient to reproduce the strong annual cycle.

The general concept of using two sets of model data to calculate the representativeness is easily applicable to other questions. One model dataset should mirror the measurements, the other should be much larger, taking into account certain statistical properties of the measurement dataset, so that the two datasets 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. With the help of the methods presented here, representativeness is given a sound mathematical description, returning an uncertainty characterizing the specific dataset.

8 Data availability

Measurement data from IAGOS-CARIBIC can be obtained by signing the CARIBIC Data Protocol available from www.caribic-atmospheric.com. The data of the model run that has been used in this study is not publicly available, but can be obtained by contacting Ole Kirner (ole.kiner@kit.edu).

Appendix A: Limitations in longitude and pressure

$\mathrm{MEAS_{CARIBIC}}$ is limited in longitude (the Pacific Ocean is never sampled) and pressure (as with all civil aircraft, CARIBIC flies at a certain pressure level). Both limitations influence the climatologies calculated from the dataset. They are discussed in the following sections.

A1 Limitation in pressure: aircraft tropopause pressure bias

By calculating R_{rel} using $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$, an important fact can be illustrated about data collected with instruments on civil aircraft. As the aircraft flies at constant pressure levels, data are also taken at these pressure altitudes only. If data are then resorted into heights relative to the tropopause (HrelTP), this limit in pressure is no longer visible. Nevertheless, it influences the results as the volume mixing rations of many trace substances are not only a function of their distance to the tropopause, but also of pressure.

The effect on the climatological values can be illustrated by calculating R_{rel} (see Eq. 4) using $\mathrm{MOD_{RANDLOC}}$ and $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ within $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$. Figure A1 shows the results (bottom panel). For comparison, the top panel of Fig. A1 shows R_{rel} of the same datasets when setting $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$, the range at which CARIBIC measures. The representativeness uncertainty is much higher in almost all heights in the bottom panel ($10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$), except just above the tropopause, where $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ contains most data. Only the long-lived species CO_2, N_2O and CH_4 retain their low uncertainties. For the more variable species to the right of the figure, the representativeness uncertainty increases strongly, especially in the troposphere, where the variability increases if data taken at higher pressure are included.

The strong increase in representativeness uncertainty is always present in measurement data from commercial aircraft, which can only collect data high above the tropopause when the tropopause is at high pressure and far below when it is at low pressure values. This bias is naturally contained in all data measured at constant pressure and then sorted relative to the tropopause, and should be kept in mind when examining climatologies from corresponding platforms.

A2 Limitation in longitude: the influence of the Pacific Ocean

As visible in Fig. 1, there are no CARIBIC measurements over the Pacific Ocean, while $\mathrm{MOD_{RANDLOC}}$ and $\mathrm{MOD_{RANDPATH}}$ also cover the Pacific. The uncertainty introduced by taking the Pacific into account in $\mathrm{MOD_{RANDLOC}}$ is investigated by calculating R_{rel} from $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$ in two different setups. R_{rel} is calculated from full $\mathrm{MOD_{RANDLOC}}$ and $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ (denoted by $R^{\mathrm{A}}_{\mathrm{rel}}$) and compared to R_{rel} calculated with $\mathrm{MOD_{RANDLOC}}$ limited in longitude λ to $120°\mathrm{W} < \lambda < 120°\mathrm{E}$ (denoted by $R^{\mathrm{B}}_{\mathrm{rel}}$). The result is shown in Fig. A2 as relative differences $|R^{\mathrm{A}}_{\mathrm{rel}}/R^{\mathrm{B}}_{\mathrm{rel}} - 1|$ between the two uncertainties. The relative differences show the share of the uncertainty inherent in $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ because the Pacific is included in the reference dataset $\mathrm{MOD_{RANDLOC}}$.

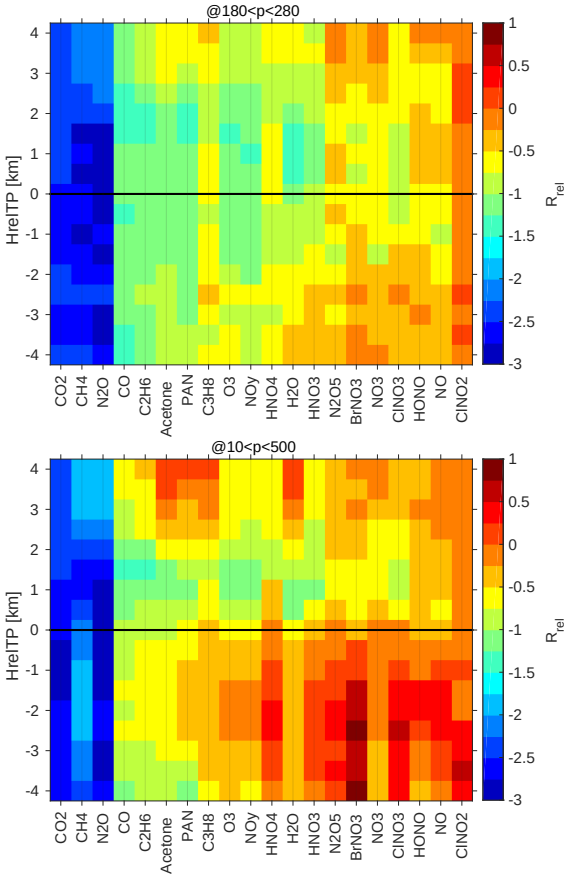

Figure A1. R_{rel} calculated from $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$ with the range of p set to $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$ (top) and $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$ (bottom). Low values indicate small climatological differences. The difference between the two panels shows the influence of expanding the limits in p when calculating the climatological mean values with HrelTP used as a vertical coordinate.

The importance of the Pacific depends on the species under consideration and whether the stratosphere or troposphere are considered. The influence on stratospheric values is very small for all species. In addition, those heights with fewer data (top and bottom) are most strongly influenced if the Pacific is not considered. For the long-lived species CO_2 and N_2O, the uncertainty increases only a little (less than 3 %) if the Pacific is included in the reference climatology of $\mathrm{MOD_{RANDLOC}}$. But tropospheric CH_4 is more influenced by surface values. Interestingly, $ClNO_2$ is also not affected, which clearly shows that the effect does not depend

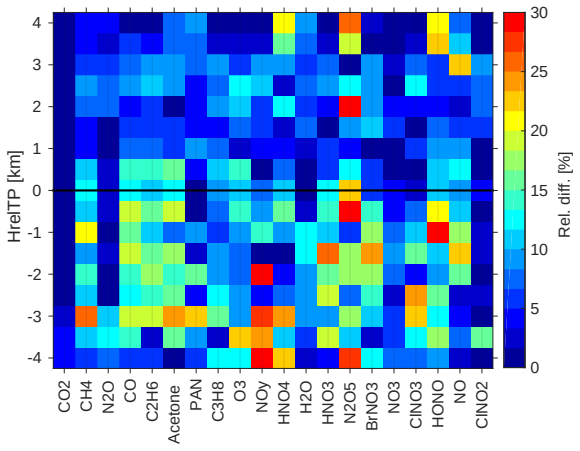

Figure A2. $|R^{A}_{rel}/R^{B}_{rel} - 1|$, given in percent. This is the fraction of the representativeness uncertainty introduced in R_{rel} calculated from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$ by including the Pacific ocean in $MOD_{RANDLOC}$, even though it is not sampled by $MOD^{regular}_{CARIBIC}$. Both R^{A}_{rel} and R^{B}_{rel} have been calculated from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$, excluding the Pacific in $MOD_{RANDLOC}$ in the calculation of R^{B}_{rel}.

on lifetime, but on the source regions and the chemistry. Acetone, CO and C_2H_6 are air pollutants with strong sources in Asia. Parts of these sources are excluded if the Pacific is not considered, which is why the inclusion of the Pacific in $MOD_{RANDLOC}$ is responsible for 15–20 % of the total uncertainty. The situation is similar for HNO_3, N_2O_5, $BrNO_3$ and HONO. For the other species, the uncertainty introduced by the Pacific is smaller.

Appendix B: Method of smoothing

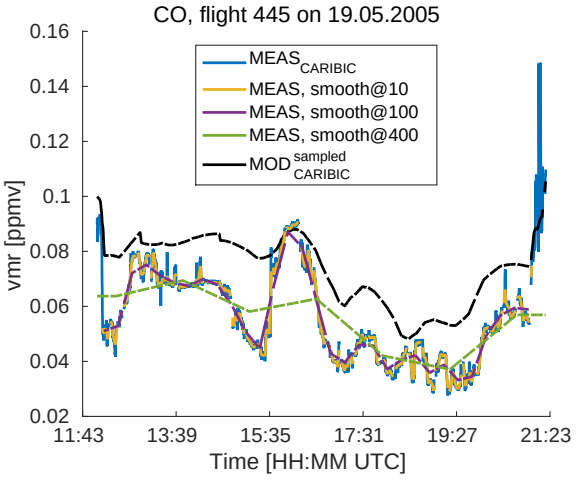

Figure B1. Time series of CO for flight 445 from Frankfurt to Tokyo. Shown is the time series of the interpolated model data and of the measurements. Measurements have been smoothed three times. The number indicates the length of the smoothing interval N.

This section shortly describes the method of smoothing used for creating the dataset $MEAS^{smoothed}_{CARIBIC}$.

Each species and each flight is considered separately. For smoothing a certain interval of the time series (consisting of a certain number of data points N), the time series is first cut into the corresponding number of pieces and the mean value of the N data points calculated within each piece. In a second step, these mean values are associated with the center of each piece of the time series. Then, a linear interpolation is performed between the central points. The corresponding mean value is applied directly from the beginning of the flight to the center of the first interval and from the center of the last interval to the end of the flight. Finally, the gaps in the original time series are mapped onto the smoothed data. The original and the resulting smoothed time series are shown in Fig. B1 for three different lengths of the smoothing interval N.

The Supplement related to this article is available online at doi:10.5194/acp-17-2775-2017-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank Andreas Engel for his work as editor and two anonymous referees, whose comments and the discussions they spawned improved the paper substantially. We would also like to thank Markus Hermann for his ongoing interest and support.

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.

The article processing charges for this open-access publication were covered by a Research Centre of the Helmholtz Association.

Edited by: A. Engel
Reviewed by: two anonymous referees

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), J. Geophys. Res.-Atmos., 113, D22305, doi:10.1029/2007JD009751, 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, Atmos. Chem. Phys., 7, 4953–4976, doi:10.5194/acp-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, Atmos. Chem. Phys., 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, Rev. Geophys., 49, RG3003, doi:10.1029/2011RG000355, 2011.

Hegglin, M. I., Gettelman, A., Hoor, P., Krichevsky, R., Manney, G. L., Pan, L. L., Son, S.-W., Stiller, G., Tilmes, S., Walker, K. A., Eyring, V., Shepherd, T. G., Waugh, D., Akiyoshi, H., Añel, J. A., Austin, J., Baumgaertner, A., Bekki, S., Braesicke, P., Brühl, C., Butchart, N., Chipperfield, M., Dameris, M., Dhomse, S., Frith, S., Garny, H., Hardiman, S. C., Jöckel, P., Kinnison, D. E., Lamarque, J. F., Mancini, E., Michou, M., Morgenstern, O., Nakamura, T., Olivié, D., Pawson, S., Pitari, G., Plummer, D. A., Pyle, J. A., Rozanov, E., Scinocca, J. F., Shibata, K., Smale, D., Teyssèdre, H., Tian, W., and Yamashita, Y.: Multimodel assessment of the upper troposphere and lower stratosphere: Extratropics, J. Geophys. Res.-Atmos., 115, D00M09, doi:10.1029/2010JD013884, 2010.

Henne, S., Klausen, J., Junkermann, W., Kariuki, J. M., Aseyo, J. O., and Buchmann, B.: Representativeness and climatology of carbon monoxide and ozone at the global GAW station Mt. Kenya in equatorial Africa, Atmos. Chem. Phys., 8, 3119–3139, doi:10.5194/acp-8-3119-2008, 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, Atmos. Chem. Phys., 10, 3561–3581, doi:10.5194/acp-10-3561-2010, 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, Atmos. Chem. Phys., 5, 433–444, doi:10.5194/acp-5-433-2005, 2005.

Jöckel, P., Tost, H., Pozzer, A., Brühl, C., Buchholz, J., Ganzeveld, L., Hoor, P., Kerkweg, A., Lawrence, M. G., 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, Atmos. Chem. Phys., 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, Geosci. Model Dev., 9, 1153–1200, doi:10.5194/gmd-9-1153-2016, 2016.

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

Köppe, M., Hermann, M., Brenninkmeijer, C. A. M., Heintzenberg, J., Schlager, H., Schuck, T., Slemr, F., Sprung, D., van Velthoven, P. F. J., Wiedensohler, A., Zahn, A., and Ziereis,

H.: Origin of aerosol particles in the mid-latitude and subtropical upper troposphere and lowermost stratosphere from cluster analysis of CARIBIC data, Atmos. Chem. Phys., 9, 8413–8430, doi:10.5194/acp-9-8413-2009, 2009.

Kunz, A., Schiller, C., Rohrer, F., Smit, H. G. J., Nedelec, P., and Spelten, N.: Statistical analysis of water vapour and ozone in the UT/LS observed during SPURT and MOZAIC, Atmos. Chem. Phys., 8, 6603–6615, doi:10.5194/acp-8-6603-2008, 2008.

Laj, P., Klausen, J., Bilde, M., Plass-Duelmer, C., Pappalardo, G., Clerbaux, C., Baltensperger, U., Hjorth, J., Simpson, D., Reimann, S., Coheur, P.-F., Richter, A., De Mazière, M., Rudich, Y., McFiggans, G., Torseth, K., Wiedensohler, A., Morin, S., Schulz, M., Allan, J. D., Attié, J.-L., Barnes, I., Birmili, W., Cammas, J. P., Dommen, J., Dorn, H.-P., Fowler, D., Fuzzi, S., Glasius, M., Granier, C., Hermann, M., Isaksen, I. S. A., Kinne, S., Koren, I., Madonna, F., Maione, M., Massling, A., Moehler, O., Mona, L., Monks, P. S., Müller, D., Müller, T., Orphal, J., Peuch, V.-H., Stratmann, F., Tanré, D., Tyndall, G., Abo Riziq, A., Van Roozendael, M., Villani, P., Wehner, B., Wex, H., and Zardini, A. A.: Measuring atmospheric composition change, Atmos. Environ., 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, J. Atmos. Ocean. Tech., 31, 2078–2087, 2014.

Lary, D. J.: Representativeness uncertainty in chemical data assimilation highlight mixing barriers, Atmos. Sci. Lett., 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 CO_2 measurements from new flask air sampling of JAL airliner observations, Pap. Met. Geophys., 59, 1–17, doi:10.2467/mripapers.59.1, 2008.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M., Lamarque, J., Matsumoto, K., Montzka, S., Raper, S., Riahi, K., Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and van Vuuren, D. P. P.: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300, Climatic change, 109, 213–241, 2011.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., Van Vuuren, D. P., Carter, T. R., Emori, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thomson, A. M., Weyant, J. P., and Wilbanks, T. J.: The next generation of scenarios for climate change research and assessment, Nature, 463, 747–756, 2010.

Nappo, C., Caneill, J., Furman, R., Gifford, F., Kaimal, J., Kramer, M., Lockhart, T., Pendergast, M., Pielke, R., Randerson, D., and Shreffler, J. H.: 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, 28452, doi:10.3402/tellusb.v67.28452, 2015.

Ramsey, C. A. and Hewitt, A. D.: A methodology for assessing sample representativeness, Environ. 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, J. Geophys. Res.-Atmos., 117, D16305, doi:10.1029/2012JD017751, 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, J. 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, Agr. Forest Meteorol., 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, Atmos. Chem. Phys., 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 NO_x emissions: Mechanisms and seasonal dependence, J. Geophys. Res.-Atmos., 109, D17307, doi:10.1029/2004JD004759, 2004.

Stiller, O.: A flow-dependent estimate for the sampling error, J. Geophys. Res.-Atmos., 115, D22206, 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, Sci. Total Environ., 367, 888–898, 2006.

WMO: Scientific Assessment of Ozone Depletion: 2010, Global Ozone Research and Monitoring Project-Report No. 52, World Meteorological Organization, 2010.