Reply to Referee Comments

Dear Editor, dear Referees,

thank you for your comments and remarks to the manuscript. We have considered the constructive criticism and feel that we have managed to improve the manuscript with your feedback.

We now devote a new section (now Sec. 5) to the comparison of the model and measurements. We compare the variability in the two datasets in order to justify the conclusion of the study, in which we apply the results from a model study to measurements. A newly added appendix (Sec. B) explains the method applied in Sec. 5.
The appendix contains two more sections (Sec. A1 and A2). One describes the influence of using a regional limit on the results, implemented as a limit in longitude corresponding to the Pacific Ocean (Sec. A1). In addition, the section on the influence of the pressure limit, which was previously part of the supplement, has been moved to the appendix (Sec. A2), as it is in structure very similar to Sec. A1.
We no longer include NO_2, as there are very few measurements.

Of course, the larger changes lead to numerous smaller changes in the manuscript, which are included in addition to those motivated directly by your comments. All changes are summarized here in order of appearance:
Abstract and introduction have been adapted due to the changes in the main body text, now also highlighting the new Sec. 5.
The description of the measurement frequency in Sec. 2.1 has been extended. In addition, the sentence describing the detrending of long-lived trace gases has been moved here, to make more clear that this method is applied to both, measurements and model data.
The description of the model run in Sec. 2.2 has been expanded, following the recommendations of both referees to include information on the boundary conditions and validation of the model.
In Sec. 3.1 (Representative for what parameter?), the limit in latitude of our analysis is now clearly defined at the beginning, as recommended by Referee 1. The appendices covering the limits in pressure and longitude are referenced.
The next Sec. 3.2 (Representative of which population?) describes the datasets used in the study. $MOD_{CARIBIC}$ has been renamed to $MOD^{regular}_{CARIBIC}$ to differentiate it from $MOD^{sampled}_{CARIBIC}$, thereby clearly marking the difference between the synthetic and flight path sampled data. $MOD^{sampled}_{CARIBIC}$ is now more often used, especially in the new Sec. 5 discussing model and measurement variability.

We now also include $MOD^3_{RANDPATH}$, which had so far been explained in all sections that use the dataset of reorganized random flights created from $MOD_{RANDPATH}$.

The paragraphs discussing the variability of $MOD_{CARIBIC}$ have been removed as Sec. 5 now covers this subject in a much more detailed manner.

As Sec. 5 discusses variability in detail, we also reconsidered Sec. 3.4, which discusses the variability used to sort results and test the statistical measures R_{rel} and R_{var}. Figure 2 now shows relative standard deviation σ_r (instead of its logarithm τ^*) of all datasets, not just of $MOD_{RANDPATH}$, as these datasets are also used in the new Sec. 5. This was also requested by Referee 2.

Sec. 4.1 mentions the other statistical tests which are now included in the study, as recommended by Referee 2. After the new Sec. 5, Sec. 6.1 describes the results of these additional tests.

The performance of the score using variability (R_{var}, results presented in Sec. 6.2) has been tested with $MOD^3_{RANDPATH}$, this is an extension of the previously submitted manuscript not directly motivated by the referees' comments. Both, Sec. 6.2 and 6.3 have been partly reworded, using the definition of $MOD^3_{RANDPATH}$ introduced in Sec. 3.2.

Changes in Sec. 6.4 (Representativeness uncertainty of the CARIBIC measurement data) come about by the changes in the names of datasets and the new material presented in Sec. 5. Referee 2 pointed out the importance of linking the thresholds for representativeness with the seasonality, which we now include here. This is highlighted especially in the discussion of separate climatologies displayed in Figure 10. Following a suggestion of Referee 2, the panels of Figure 10 are now labeled with letters from A to F.

Because $MOD^3_{RANDPATH}$ has been introduced above, Sec. 6.5 (Number of flights for representativeness) is now much shorter. In addition, we use a different figure that directly shows the relationship between the representativeness uncertainty and the number of flights for several species. It is therefore better suited to link a species' uncertainty to its seasonality, as suggested by Referee 2.

Due to the numerous changes of the manuscript, the conclusions have been adopted accordingly.

In the following, we present a point by point reply to all suggestions, followed by the marked up manuscript.

Best regards,

Johannes Eckstein

Referee 1

Major Comments

1. *However, the question arises if the model can be used as an appropriate tool for the question. I think this question has not been addressed sufficiently in the paper. How well can data from a course model resolution be representative of the state of the atmosphere as described here? The representation of the model climatology vs. flight track interpolation should depend on the models spatial and temporal resolution. If the grid or time span is too large (likely the case for global models), the model would not be able to represent the variability of the observations. A test would require to average the observations to the same model grid and then compare the variability.*

   We have now included a separate section (Sec. 5) that treats this question. We show the influence of the small scale variability on climatological mean values and discuss the differences between model and measurement variability on longer time scales. The section shows that the model reaches 50% to 100% of the variability of the measurement data, which have been smoothed to have the same small-scale variability as the model data. The ratio could be increased by a model run with a higher resolution, but is just as much influenced by the data used for binning the measurements, which has a much coarser resolution than the measurements themselves.

2. *Furthermore, I do not see any evaluation of the model. How well does the model represent the atmosphere? Especially water vapor is a gas that many models are not able to simulate appropriately, which is also the case for NOx and NOy. A discussion on how much this study depends on the performance of the model to represent chemical tracer should be added.*

   The new Sec. 5 also covers the differences between species. A detailed validation is beyond the scope of this study. Sec. 2.2, describing the model, has been expanded to include more references, e.g. to Hegglin et al. (JGR), 2010, who describe a validation of some aspects of the model, in addition to the validataion published by Jöckel et al. (GMD), 2016.

3. *Finally, little has been done to identify reasons for differences between the flight track comparison and the global comparison, based on the atmospheric character of different trace gases dependent on the region for instance. Depending on region, airmasses experience more pollution, convection, stratosphere/troposphere exchange. The Pacific experiences a lot of pollution from South East Asia in some seasons than the Atlantic. Since CARIBIC data do not cover the Pacific, what implication does that have of the representation of the data compared to a global average? I would suggest, plotting a lon/lat map for a certain altitude level, say 1 km below the tropopause. This may help explain why some tracers are representative and why others may be not. Certainty 35-70 degrees is a very large region that covers a lot of different airmasses reaching from the tropics to the polar regions.*

   We have included a section in the appendix that assesses the influence of a regional limit by excluding data taken at longitudes corresponding to the Pacific Ocean and comparing the results to the regular analysis (Sec. A1). The influence on climatological mean values is stronger for those species determined by source regions in Asia.

Minor Comments

1. *Page 3, Line 9. The assumption that species in the model show a similar variability has not been supported. A climatology of trace gases from the course model resolution is expected to show a much smaller variability than the observations. Wouldn't you expect a different result if you would run with a high model resolution spatial and temporal?*

   We now include a new section (Sec. 5) which treats the differences in model and measurement variability and the influence of the small scale processes on the climatological mean values, see also the answer to major comment 1 above.

2. *Page 4, line 15: Why is N2O5 counted twice, please explain.*

   N_2O_5 is measured by catalytic conversion to NO. One N_2O_5 molecule yields two NO molecules, this is why every N has to be counted. This is now explained in the manuscript in Sec. 2.2.

3. *Page 5: Line 6: is it +-4km (as stated above) or +- 4.25km?*

   This has been corrected here to +-4.25km. Heights are labeled with their centers, which corresponds to +-4km.

4. *Page 5, Line 17ff: Constraining the data to 35-75 degree N is not really removing different characteristics of tropical or polar airmasses and you would expect a larger variability. Earlier studies discussed differences in the characteristics of UTLS airmasses depending on the location with the jet stream and therefore with the height ofthe tropopause, which strongly varies with season. I think, constraining the comparison to 35-75 degrees N because of a good coverage of aircraft data would the better argument. There should be some discussion on the variability of the considered region.*

   We now discuss the questions of regional limits and coverage in more detail. The good coverage was also an argument for the limit in latitude and we now state so clearly in the text. The latitudinal limit is for sure not sufficient to exclude all influence of higher or lower latitudes, but is a first approximation. We do discuss data relative to the local tropopause, as all fields are presented with HrelTP (height relative to the tropopause) as vertical axis.

5. *Page 5, Line 23, if you define mid-latitude as 35-75deg, then please specify that here.*

   We now define mid-latitude more clearly, e.g. by specifying '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 °N < φ < 35 °N.' in the first paragraph of Sec. 3.

6. *Page 6: Line 6-7: The temperature comparison for the data is taken from meteorological analysis. Are those the same that were used to nudge the model? That would explain the high correlation coefficient. Please clarify.*

   The temperature data for the statement is taken from CARIBIC aircraft measurements. This is used for ERA-Interim via the AMDAR network, which is used for nudging the model. So the two are not independent, but the high Pearson correlation coefficient mentioned here serves to indicate the usefulness of the interpolation and is not meant as validation of the model.

7. *Page 7, Line 7-8: HrelTP does not look very similar to me. Distributions in the lower two rows in Figure 1 are more often above the TP than the flight track interpolated data. What implications will this have for the analysis?*

   Both, the distribution and differences in HrelTP and the different trace gas climatologies, are influenced by the sampling pattern. We have now clarified this relationship in the text.

8. *Page 7: Line 18. The text describes that the variability of the model data if interpolated to the flight track is only 40-70% of the actual observed data. Further, it is discussed that the variability in the model cannot capture the small scale variability of the data. Then the assumption is made that the variability of the model is similar for all species. I do not follow this conclusion. Why is this the case?*

   This paragraph has been completely revised and Sec. 5 now covers this subject.

9. *Page 9: Line 19: How does the model represent CO2, N2O and CH4? If those are prescribed as fixed boundary conditions, certainly the model would not identify the variability that exists in the real data.*

   Boundary conditions are not fixed. For CO_2, N_2O and CH_4, they are prescribed as latitude dependent monthly means. We have included a paragraph in the Sec. 2.2 on the boundary conditions of the chemical species in the model run.

10. *Page 13: I am not surprised about the different characteristics, since the different coverage of CARIBIC compared to the random distribution is very different, Figure 1 left column, the flight track sample more tropical air masses (being more concentrated in the south). Furthermore, the Pacific with different characteristic of tracers are not sampled by the CARIBIC data set. It would help to see for example a figure of CO at the altitude considered for example 1 km below the tropopause. A discussion on differences of the sampling location due to chemical characteristics that are different depending on sampling tropical or polar air masses, or characteristic longitudinal variability in different tracers would be helpful.*

    Whether the climatologies produced by the sampling pattern of CARIBIC are representative is just the question that we are investigating in this study. Regional differences are another, interesting subject, but are more difficult to investigate with CARIBIC data. As a first step, the influence of considering the Pacific ocean (or not) is now included as Sec. A2 of the appendix of the paper.

11. *Page 17: typo line 2 “while it is can be much”*

    The typo has been corrected by removing 'can be'.

12. *Page 17: Line 10: models usually have a poor representation of NO and NO2, especially in the UTLS it depends on lightning. Also convection is influencing NOx and can strongly vary with location, which is usually not well represented in models. Couldn’t this be the reason why there is a larger uncertainty?*

    NO_x production resulting from lightning activity is included in the model (Grewe et al., 2001). The geographical constraint of CARIBIC flight routes to flight corridors and thereby to the regions with high VMR of NO_x has the stronger influence on representativeness.

13. *Line 14: How is the model representing H2O in the stratosphere?*

    In the stratosphere and mesosphere the chemical H_2O tendencies (due to the methane oxidation) are calculated with the help of the chemical submodel MECCA (Sander et al., 2005). H_2O in the lower stratosphere is one of the compounds discussed by Hegglin et al, (JGR), 2010.

14. *Line 20; C2H6 and C3H8 are considered short-lived species with lifetimes of a few weeks or so.*

    We have changed the description from 'rather long-lived' to 'moderately long-lived'.

15. *Sec. 5.5 I think, the question should be changes for extended to: What would be a better regional coverage improve the statistic? This could be easily addressed within this paper, since one could extend the coverage over the pacific region, but keep the number of flights the same.*

    The influence of the Pacific region on our analysis is now covered in the appendix (Sec. A1). A more detailed study of the influence of different regions could be performed in the future.

16. *Conclusions: Page 21: Line 14: Sentence is unclear.*

    The sentence has been reworded: From 'R_{rel} is more applicable for answering the question, asking for the representativeness of for a climatology. It is therefore used for the analysis.' to 'R_{rel} (describing the representativeness of a climatology) is better suited for answering the question and is therefore used in the remaining analysis. '.

Referee 2

Major Comments

1. *Global scale chemistry transport model*
   There are two major concerns about using the EMAC model as a reference state of the atmosphere. First, the model description in the text is insufficient. It needs to be mentioned how the model was validated against other independent observations. For which species did the model perform well and for which not? Where is the model insufficient to reproduce variability on the scale given by the model resolution? This is especially important since one may suspect that the model will have difficulties reproducing vertical trace gas gradients in the UTLS region. Second, as shown in Figure 1 the model has only 3 levels in the UTLS region and output was only available every 12-hour. Therefore, the model misses large parts of the real variability (see also the CARIBIC comparison). How can it be justified that the model can still be assessed to analyse representativeness?

   We have now included a separate section (Sec. 5) that treats this question. We show the influence of the small scale variability on climatological mean values and discuss the differences between model and measurement variability on longer time scales. The section shows that the model reaches 50% to 100% of the measurement data that have been smoothed to have the same small-scale variability as the measurements. The ratio could be increased by a model run with a higher resolution, but is just as much influenced by the data used for binning the measurements, which has a much coarser resolution than the measurements themselves.

2. *Sampling strategy*
   Several choices seem to be arbitrary. I especially don't understand why the temporal domain is not sampled as a whole. Both sampling patterns RANDPATH and RANDLOC only sample 12 and 8 days per month, respectively. It would seem more appropriate to sample daily but on the other hand with a more realistic pattern that resembles that of the CARIBIC flights (i.e., on great arcs between major airports in the northern hemisphere, leaving out transpacific flights, since this region is never covered by CARIBIC). In that case the RANDPATH sampling could be viewed as the maximal achievable sampling pattern by commercial aircraft and RANDLOC could still be seen as sampling the northern hemispheric UTLS region as a whole.

   The alternative approach for creating RANDPATH proposed here may be more realistic, but the results would not be much different. The same is true for sampling the temporal domain. This is probable, as even RANDPATH and RANDLOC yield very similar climatologies, despite the differences in their sampling statistics.

3. *Selected statistical measures*
   Again there seem to be arbitrary choices concerning the statistical estimators and tests. If the Komogorov-Smirnov test turned out to be too strict because it requires similarity of the whole distribution, why did you not select other statistical tests that only evaluate one statistical parameter at a time (e.g., Mann-Whitby test for the mean and Levene's or Brown–Forsythe test for variance, all are non-parametric tests suited for atmospheric trace gas observations). Furthermore, the results need to be discussed together with observed seasonality of the trace species as is mentioned by the authors themselves on page 17, line 1, but than dropped without further reasoning 3 lines later. The relative difference does not contain much information in itself and as stated correctly depends on the lifetime of a species.

   We have checked all the tests proposed here and find that they do not provide more detailed answers than the Kolmogorov-Smirnov test used for the original manuscript. Short comments on this are included in the text. In discussing results, we now more explicitly state the need to consider the seasonality and do so when considering climatologies of individual trace gases in Figure 10, formerly Figure 8.

Minor Comments

1. *P1,L11: "formulated above". Not clear from the context where this was formulated*

   The sentence has been reformulated: From 'In contrast, the variability based scores pass the general requirements for representativeness formulated above. ' to 'In contrast, the two scores based on either variability or relative differences show the expected behaviour and thus appear applicable for investigating representativeness. '.

2. *P3,L28ff: Although no details on the measurement techniques are needed here, it would still be interesting to learn something about the overall uncertainties of the measurements and how these compare to the later discussion of representativeness.*

   It is difficult to give an overall uncertainty of the measurements, as these are taken by many different instruments. For example, the accuracy of acetone measurements is typically +-15%, which is mainly determined by the accuracy of the calibration gas standard and the reproducibility of the calibration. The relative precision becomes smaller for higher mixing ratios. At 1000 pptV, it is ~+- 3%, but it becomes +-25% at 200 pptV. For O_3, the precision is in the order of 0.3-1%. Since the instruments have such different characteristics, we have decided not to include too much detail on this in the manuscript.

3. *P4,L10: Model output every "eleven hours"? Did you mean 12 hours?*

   Model output for this model run was saved every eleven hours in order to be able to reproduce mean daily cycles.

4. *P4,L9ff: Additional information on emissions used in EMAC and vertical resolution in the UTLS region would be useful here.*

   The vertical resolution is displayed in Figure 1 and corresponds to about 1.5km in the UTLS.

5. *Sec. 3.1: It should be more prominently mentioned in the first paragraph of this Sec. that you restrict the analysis to the latitude region 35N to 75N. Details follow towards the end of the Sec. and can remain there, but it would be good to make this important detail clear from the beginning. It should also be stated in the abstract.*

   We now state these limits at the beginning of Sec, 3.1 as well as in the abstract.

6. *Table 1: For RANDPATH it is an adjusted Gaussian distribution, as mentioned in the text.*

   Wording has been adopted, changing 'gaussian' to 'adjusted gaussian' in the Table 1.

7. *Table 1 and elsewhere: "Uniform" or "rectangular" distribution should be used instead of "even".*

   Wording has been adopted, changing 'even' to 'uniform' in Table 1 and the description in Sec. 3.2.

8. *P6,L6f: The good correlation for temperature is not a big surprise, given the strong vertical stratification in the UTLS and the assumably large number of measurements. Since this is one of the few pieces of model validation mentioned, one could add a scatter plot to the supplement.*

   A validation of the model is not the focus of this study. It has been done elsewhere, e.g. Hegglin et al. (JGR), 2010 discuss a validation of some aspects of the model. The text now gives a reference to this study in Sec. 2.2.

9. *P6,L9f: It is not clear to me why the limited vertical model resolution is the reason you cannot compare CARIBIC directly to EMAC. The random sampling is still done using vertical interpolation to specific pressure levels. Would't the same argument apply to the random sampling strategy as well and could one not simply drop it and do the analysis of representativeness on discreet model levels instead?*

   Sampling on the discrete pressure levels would give rather poor results. The pressure would be limited to certain values only, which - in addition - are close to the limits CARIBIC ever reaches.

10. *P6,L20f: Why did you choose these cut-off values instead of simply using the standard deviation as a criterion (i.e., redistribute values outside +/- 2 sigma). I don't assume this would change much, but would seem statistically more sound. Alternatively, one could have sampled directly from the observed CARIBIC distribution.*

    We used these cut-off values as these correspond to the upper and lower limit of the CARIBIC measurements. The lower boundary was set to exclude ascents and descents of the aircraft.

11. *P7,L7ff: I don't agree with the statement that the distribution "is very similar for all datasets". There is a strong offset to higher HrelTP in both random sampling strategies. What is the actual mean HrelTP for all three samples?*

    The mean of HrelTP for the different datasets is now stated in Sec. 3.2. Both, the distribution and differences in HrelTP and the different trace gas climatologies, are influenced by the sampling pattern. We have now clarified this relationship in the text.

12. *p7,L11f: This requires some further justification (see major comment above). Without being aware of the details of Jöckel et al 2015, it seems a bit hard to believe that the model performs equally well for the very different set of species analysed here. There should be additional discussion of the species for which this may not be justified.*

    We have included a new, separate section (Sec. 5) that covers this subject, comparing model and measurement variability. The paragraphs you are referring to have therefore been removed.

13. *p9,l19: How was the mean tau* calculated? As the mean over all monthly tau* or as tau* of the mean mu and sigma?*

    This is now clearly stated in the text: As τ* (logarithm of the relative standard deviation) of the mean of μ and standard deviation σ using the whole time series.

14. *Figure2: It would be interesting to add CARIBIC observed tau* in the figure (where available).*

    Figure 2 now includes all τ* (logarithm of the relative standard deviation) of all the relevant datasets. In addition, we have modified the figure to show relative standard deviation, σ_r, which makes the figure easier to understand.

15. *Figure3 and others: The y-axis if often titled "variability". It would be useful to give a more concrete title, since the manuscript is dealing with all kinds of variability. This could reduce confusion. In this specific case I assume this is relative standard deviation?*

    The y-axis is now titled σ_r, the relative standard deviation.

16. *p12,l18: "The differences are small, mostly below an absolute value of 0.15." But this means that the absolute difference between both samples is 1.4 times larger than the value of the reference (or am I mistaken). I am not sure that I would call this small! In general using the log scaled relative difference seems a bit odd and only confuses. Why not use the relative difference as is?*

    The value 0.15 refers to the absolute values of which the logarithm has not been taken. 0.15 means there is 15% percent difference between the fields.

17. *p12,l29f: "A similar analysis has also been performed with data from a random number generator, leading to equivalent results." Are you referring to the RANDLOC sample here?*

    We are referring to the study of data created with a random number generator. It is documented in the supplement to the paper. The sentence has been reworded. It now states: '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 supplementary material to the article.'

18. *p13,l13: At least repeat the result of the sensitivity study here. The supplement should not be a paper on its own.*

    The reference in the following sentence has been made clearer by stating: 'This is also the result of a sensitivity study, which is discussed as supplementary material to this text. '

19. *p16,l5: Not clear which correlation is referred to here.*

    The reference to the correlation coefficient (with the number of samples) has been removed as it is not necessary for the argument.

20. *p16,l6: What is an "uncertainty error"? I think the use of representativeness uncertainty would in general work better.*

    The wording has been adopted, changing 'uncertainty error' to a plain 'uncertainty' here, then introducing the term 'representativeness uncertainty' in the sentence.

21. *p16,l8-13: This description is completely confusing. I don't understand what is done and why. Please improve the description.*

    The wording of this paragraph has been changed, please also refer to the highlighted differences at the end of this document. The paragraph now reads:
    'In order to asses the uncertainty for accepting CARIBIC measurement data to create a climatology, model data have to contain the same amount of data as $MEAS_{CARIBIC}$, which is why $MOD^{sampled}_{CARIBIC}$ (see Sec. 2) will be used in the following. In addition, $MOD_{RANDLOC}$ (see Table 1) was used as reference, as it has a random sampling pattern and represents the full model state, independent of the sampling pressure. The limits in pressure where again set to 180 hPa < p < 280 hPa. The resulting R_{rel} is shown in Figure 9. Using different wording, R_{rel} in this formulation can also be considered the sampling error of the measurements. '

22. *p17,l10ff: Since the discussion on NOx is along the EMAC results, it would be interesting to know how NOx sources in the UTLS are treated in the model. Does the model include a realistic representation of lightning NOx? Has this been analysed in previous studies?*

    NO_x production resulting from lightning activity is calculated with the help of the EMAC submodel LNOX (Grewe et al., 2001).

23. *p17,l33f: The representativeness uncertainty of 5 [% for long-] lived trace gases is huge considering their atmospheric abundance. It is much larger than their seasonal variability. This aspect needs to be considered in the analysis and discussed along with the results.*

    We now more clearly state the importance of the seasonal cycle. 'Less than 5%' has been reworded to be more meaningful for these substances by stating: '…all have representativeness uncertainties of less than 0.4 %, which is lower than their seasonal variability.' In addition, the color scale of the corresponding figure has been changed to highlight lower values more prominently.

24. *p18,l20ff: Finally there is some discussion using species specific thresholds, but again these thresholds are chosen without any justification. They should be related to seasonal variability.*

    The thresholds are now related to seasonal variability: 0.03% for CO_2 and 15% for O_3, one third and one fourth of the seasonal variability, respectively.

25. *p18,l32: Was it ever shown that R_rel "increases linearly"? May be an increasing relationship, but linearly?*

    A linear relationship has been shown to exist by the significant Pearson correlation coefficient in Sec. 6.3.

26. *Figures: It would be easier to follow the discussion of the figures if sub-panels would be labelled by letters (which is Copernicus style). For example discussion of Figure 8 on page 18.*

    We have included labels from A to F to the panels of Figure 10, formerly Figure 8.

27. *Figure 1 in supplement: Please explain black line in legend and add fit as additional line to the plot. Indicate which species are behind each point. Is the given fit applied to log(meas) and log(model)? Is it just my impression or does the model actually capture less of the variability for species that have a small relative variability? How could this be explained? I would have expected the opposite.*

    This part of the supplement has been removed as it is now covered in the new Sec. 5 on model and measurement variability.

Technical Comments

1. *P1,L2: It is "representative of" not "representative for".*

    This has been changed where 'representative for' stands for representative of a population.

2. *P5,L6 and elsewhere: "Data" is always plural. Change to "Data were*

    This has been changed everywhere in the manuscript.

Literature

- Grewe, V., Brunner, D., Dameris, M., Grenfell, J. L., Hein, R., Shindell, D., and Staehelin, J.: Origin and variability of upper tropospheric nitrogen oxides and ozone at northern mid-latitudes, Atmos. Environ., 35, 3421–3433, 10.1016/S1352-2310(01)00134-0, 2001
- 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, Journal of Geophysical Research: Atmospheres, 115, doi:10.1029/2010JD013884, 2010.
- Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51, Geoscientific Model Development, 9, 1153–1200, doi:10.5194/gmd-9-1153-2016, 2016.
- Sander, R., Kerkweg, A., Jöckel, P., and Lelieveld, J.: Technical note: The new comprehensive atmospheric chemistry module MECCA, Atmos. Chem. Phys., 5, 445–450, doi:10.5194/acp-5-445-2005, 2005

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

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

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

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

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

~~Generally, representativeness~~ Two requirements for any score describing representativeness are essential: Representativeness is expected to ~~decrease with increasing variability and to increase~~ increase (i) with the number of ~~available samples~~ samples and (ii) with decreasing variability of the species considered. Based on ~~this assumption~~these two requirements, we investigate the suitability of the different statistical measures for ~~our problem~~investigating representativeness. The Kolmogorov-Smirnov test ~~seems too~~ is very strict and does not identify any trace gas climatology as representative – not even ~~long lived well observed~~ of long lived trace gases. In contrast, the ~~variability based scores pass the general requirements for representativeness formulated above. In addition, even the simplest metric (relative differences) seems~~ two scores based on either variability or relative differences show the expected behaviour and thus appear applicable for investigating representativeness.

~~Using~~ For the final analysis of climatological representativeness, we use the relative differences score ~~we investigate the representativeness of a large number of different trace gases . For our final consideration we assume that the EMAC model is a reasonable representation of the real world and that representativeness in the model world can be translated to representativeness~~

~~for CARIBIC measurements. This assumption is justified by comparing the model variability to~~ 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 ~~variability of CARIBIC measurements~~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 ($\leq 0.4\,\%$), while shorter-lived species like O_3 have larger uncertainties (10-15 %). Finally, we ~~show how~~ translate the representativeness score ~~can be translated~~ 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 %.

1 Introduction

The UTLS (upper troposphere/lower stratosphere) is dynamically and chemically very complex and shows strong gradients in temperature, humidity and in many trace gases (Gettelman et al., 2011). As the ~~the~~ mid and upper troposphere have a strong influence on the atmospheric greenhouse effect, the UTLS plays an important role in our climate system (Riese et al., 2012). To characterize processes and evaluate the performance of chemistry-transport models in this area, ~~we require~~ spatially well resolved data collected on a global scale 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 ~~scope~~ 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 whether a flux tower station is representative ~~for~~ 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 analysed by means of a cluster analysis with backward trajectories (e.g. by Henne et al. (2008) or Balzani Lööv et al. (2008)). By this method, source regions for measured trace gases can be found and the type and origin of air masses contributing to an observed air mass determined, i.e. the airmass the data ~~is representative for~~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 analysed whether the dataset of the aircraft campaign SPURT (SPURenstofftransport in der Tropopausenregion - trace gas transport in the tropopause region, Engel et al. (2006)) is representative of the larger MOZAIC dataset (Measurements of OZone, water vapour, carbon monoxide and nitrogen oxides by in-service AIrbus airCraft, the precursor of IAGOS-core). Kunz et al. (2008) investigate distributions of two substances (O_3 and H_2O) in two atmospheric compartments (upper troposphere and lower stratosphere). They find that the smaller SPURT dataset is representative on every time scale of the larger MOZAIC set for O_3, while this is not the case for H_2O. While SPURT O_3 data can be used for climatological investigations, the variability of H_2O is too large to be fully captured by SPURT on the interseasonal time scales.

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

In Sec. 2, more details on the data from IAGOS-CARIBIC and the model run will be given. The general concept and definition of representativeness is discussed in Sec. 3. This section also gives details on sampling the model and on the variability,

which is used to group results by species. The statistical methods are then explained in Sec. 4, namely the Kolmogorov-Smirnov test, a variability analysis following the general idea of Kunz et al. (2008) and Rohrer and Berresheim (2006) and the relative difference of two climatologies. We then discuss the variability of the model data in comparison to that of the measurements in Sec. 5. The application of the methods to the different model samples is described in Sec. 6. After showing the result of each of the three methods seperately, Sec. 6.4 discusses the representativeness of the IAGOS-CARIBIC measurement data, while Sec. 6.5 answers the question how many flights are necessary to achieve representativeness. Sec. 7 summarizes and concludes.

2 Model and data

2.1 The observational IAGOS-CARIBIC dataset

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

During each CARIBIC flight, about 100 trace trace gas and aerosol parameters are measured. Some are measured continuously with a frequency between ~~5 Hz and 1/(5 min) and commonly available every~~ $5\,\mathrm{s}^{-1}$ and $0.2\,\mathrm{min}^{-1}$) and available from the database binned to $10\,\mathrm{s}$~~while others~~. Others (e.g. non-methane hydrocarbons) are taken from up to 32 air samples collected per flight. The substances considered in this study are $\mathrm{NO_y}$, $\mathrm{H_2O}$, $\mathrm{O_3}$, $\mathrm{CO_2}$, NO, ~~,~~$(\mathrm{CH_3})_2\mathrm{CO}$ (acetone), CO and $\mathrm{CH_4}$ from continuous measurements and $\mathrm{N_2O}$, $\mathrm{C_2H_6}$ and $\mathrm{C_3H_8}$ from air samples. $\mathrm{NO_y}$ is the sum of all reactive nitrogen species, measured by catalytic conversion to NO (Brenninkmeijer et al., 2007). Data of $\mathrm{N_2O}$, $\mathrm{CH_4}$ and $\mathrm{CO_2}$ 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 $\mathrm{MEAS_{CARIBIC}}$.

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

2.2 The chemistry-climate model EMAC

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

The model integration used in this study simulated the time between January 1994 and December 2013, with data output every eleven hours. Meteorology is nudged up to $1\,\mathrm{hPa}$ using divergence, vorticity, ground pressure and temperature from six-hourly ERA-Interim reanalysis. It includes the extensive EVAL-Chemistry using the kinetics for chemistry and photolysis

of Sander et al. (2011). This set of equations has been designed to simulate tropospheric and stratospheric chemistry equally well.

Boundary conditions for greenhouse gases (latitude dependent monthly means) are taken from Meinshausen et al. (2011) and continued to 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 first 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 ~~used~~ from the model used in this study are the same as those ~~used from measurements, summing up~~ 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 ~~was detrendedby subtracting the mean of each year from the values of that year and adding the overall mean~~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 ~~is~~ are the sample data to be seen as representative? 3. To which degree ~~is~~ 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 other' (Nappo et al., 1982). In this study, we investigate whether the CARIBIC data can be used to construct a climatology in the UTLS. We consider monthly binned data in the height of ~~±4 km~~ $\pm 4.25\,\mathrm{km}$ around the dynamical tropopause defined at the pressure at 3.5 PVU and in mid-latitudes with $75°\mathrm{N} < \varphi < 35°\mathrm{N}$.

In order to reference data to the tropopause, we use the geometric height in kilometers relative to the tropopause (HrelTP) at each datapoint. For the measurements, this height is provided by the meteorological support of CARIBIC by KNMI (Konin-

klijk Nederlands Meteorologisch Instituut) (http://www.knmi.nl/samenw/campaign_support/CARIBIC/), who use data from ECMWF (European Centre for Mendium-range Weather Forecast) for their calculation.

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

Even though all data of trace gases (be it from model or measurements) ~~is~~ 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 $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$. In addition, we explicitly limit pressure to this range in order to exclude data from ascents and descents of the aircraft. But since data ~~is~~ 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~~values~~. The properties of many trace substances are not only a function of their distance to the tropopause, but also of pressure. The limits in pressure inherent in the sample therefore also influence the climatology. They have to be considered and should be explicitly stated. This ~~efffect~~ effect is illustrated in ~~the supplementary material~~ Appendix A1 with the help of the methods developed in this study.

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

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 ~~for~~ of the tropopause region in mid-latitudes?

3.2 Representative ~~for~~ of which population?

When assessing the representativeness of the sample made up by all CARIBIC measurements (called $\mathbf{MEAS_{CARIBIC}}$, see Sec. 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 meaurements as regularly at the same spatial and temporal resolution. IAGOS-core and CONTRAIL sample with much higher frequency, but take measurements of only few

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

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

substances while satellites do not resolve the small scale-structures scale structures necessary to disentangle the dynamics around the tropopause. The population is therefore not accessible by the measurement platforms currently available.

This is the reason why the representativeness of the CARIBIC data is 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 Figure 1.

$\text{MOD}_{\text{CARIBIC}}^{\text{regular}}$: For the dataset $\text{MOD}_{\text{CARIBIC}}^{\text{regular}}$, 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 modelled to measured temperature can serve as an indication that this interpolation leads to reasonable results, despite the coarse coarser resolution in time and space of the model output.

$\text{MOD}_{\text{CARIBIC}}^{\text{sampled}}$: 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 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 $\text{MOD}_{\text{CARIBIC}}^{\text{regular}}$ 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 $\text{MEAS}_{\text{CARIBIC}}$ will be referred to as $\text{MOD}_{\text{CARIBIC}}^{\text{sampled}}$.

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

$\mathrm{MOD}_{\mathrm{RANDPATH}}$: 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 ~~$\mathrm{MOD}_{\mathrm{CARIBIC}}$~~ $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$. 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 $885.1\,\mathrm{km\,h^{-1}}$, the median of the speed of the true CARIBIC aircraft. The flights start at ~~0 : 00 UTC~~ 0:00 UTC and sample the model for ~~one day~~ 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 Figure 1.

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

$\mathrm{MOD}^{3}_{\mathrm{RANDPATH}}$: The dependecy 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 eight hour 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 ~~an even~~ a uniform distribution from 500 hPa to 10 hPa for each flight. Again, the datasets start at ~~0 : 00 UTC~~ 0:00 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 ~~$\mathrm{MOD}_{\mathrm{CARIBIC}}$ is representative for~~ $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ is representative of a climatology around the tropopause only within its pressure limits or also when expanding these limits.

As is visible in Figure 1, the distribution in HrelTP is very similar for ~~all datasets~~ $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ even though the pressure is ~~presribed~~ prescribed in very different ways ~~. This is an important prerequisite for the following investigation, as it shows that the relative~~ (mean of 0.79 km 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 amount of data ~~in each height bin is similar for all three datasets~~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.

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

~~The question whether this assumption is valid was also investigated with the available data. The relative standard deviation $\sigma_r = \sigma/\mu$ was calculated in each month of the climatlogies of CARIBIC measurements ($\mathrm{MEAS}_{\mathrm{CARIBIC}}$) and $\mathrm{MOD}_{\mathrm{CARIBIC}}$ (σ~~

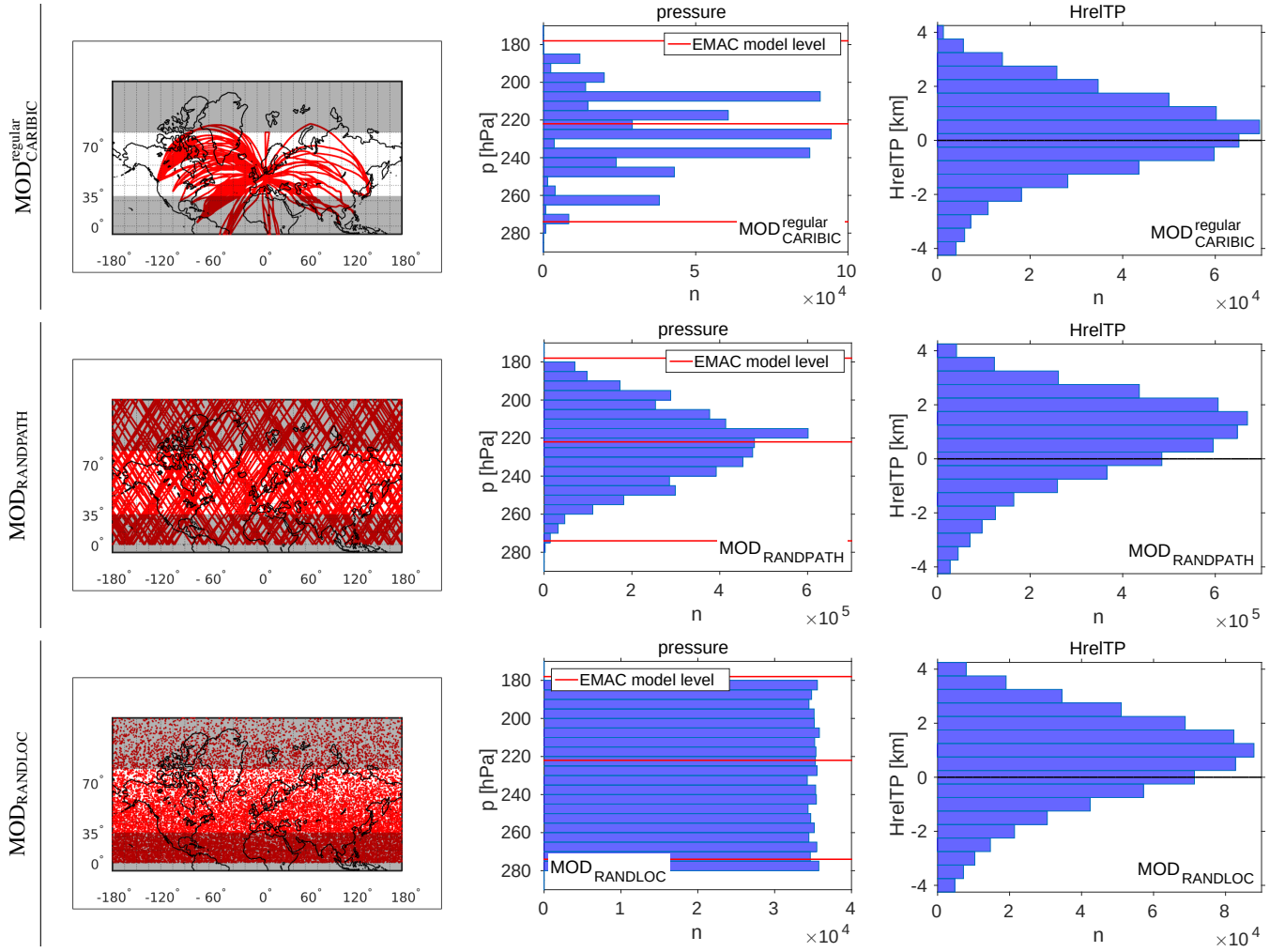

Figure 1. Flight ~~paths~~ path distribution (left), distribution ~~in~~ p ~~of~~ probed pressures (p, center) and ~~HrelTP~~ height relative to the dynamical tropopause (HrelTP, right) for the three datasets MOD~~CARIBIC~~ regular CARIBIC (top), $MOD_{RANDPATH}$ (center) and $MOD_{RANDLOC}$ (bottom).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

being the standard deviation, μ the mean) as a measure for the variability. By taking the mean over all months of the fraction of $\sigma_r^{\mathrm{MOD_{CARIBIC}}}$ and $\sigma_r^{\mathrm{MEAS_{CARIBIC}}}$, the fraction of variability of $\mathrm{MEAS_{CARIBIC}}$ reached by $\mathrm{MOD_{CARIBIC}}$ can be evaluated.

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

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

3.3 Confidence limits of ~~the~~ representativeness

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

Three definitions for representativeness are discussed and applied in this study: The Kolmogorov-Smirnov test, the ~~variabiltiy~~ variability analysis following Kunz et al. (2008) and the relative difference of two climatologies. The first method gives a yes-no answer within a chosen statistical confidence level. The other two approaches are formulated in such a way as to return a score. By (arbitrarily) setting a value for the score, the representative cases can be discriminated from the non-representative cases (see Sec. 4 and Sec. 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 implicitely also made by Kunz et al. (2008), as they investigate the representativeness of a smaller for a larger dataset and for two species of different variability. The measure for variability we use in this study is explained in the following section.

3.4 Defining a measure for variability

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

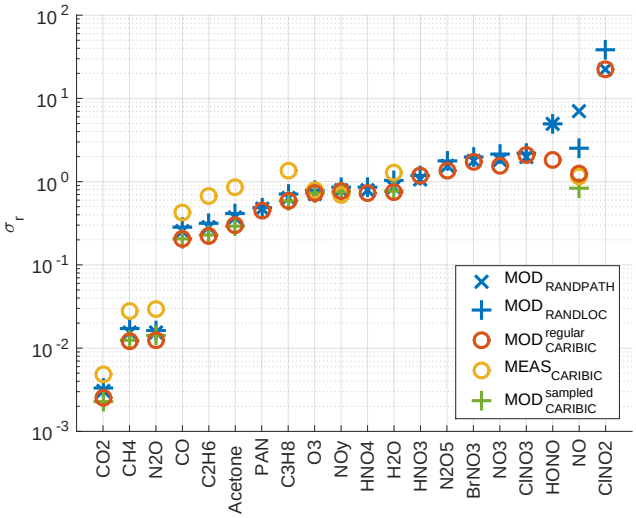

Figure 2. Variability ~~τ^*~~ σ_r calculated ~~from $\mathrm{MOD}_{\mathrm{RANDPATH}}$~~ for different datasets using Equation 1. The species are sorted ~~in τ^*~~ by σ_r, species with low variability ~~(high τ^*)~~ listed to the left, using the values from $\mathrm{MOD}_{\mathrm{RANDPATH}}$ for sorting.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

and standard deviation σ of each species.

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

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

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

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

It is important to note that the values determined from $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ 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 of the two values for $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$. $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ 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 Section 5. The model datasets are very similar, despite their different sampling patterns. They only differ for short-lived species (to the right in Figure 2), which have a strong daily cycle, e.g NO.

In Sec. 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 Equation 3 to test the relationship of representativeness and variability.

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

Sec. 4.2 will take a closer look at variability. It will be discussed how variability depends on the time scale for which it is calculated. The values shown in Figure 2 and used for the calculation of τ^* use the full time series, and thereby the overall variability. If shorter time scales had been considered, the values for σ_r in Figure 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 Section 3.1, we can specify more closely the question we answer in this study: For which species is a climatology compiled from CARIBIC data representative ~~for~~ 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~~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 $\text{MOD}_{\text{~~CARIBIC~~}}^{\text{regular CARIBIC}}$ and $\text{MOD}_{\text{RANDPATH}}$ in each height bin and month~~, see~~. 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 Sec. 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, Kunz et al. (2008) then applied it to aircraft data. A timeseries of data is subsequently divided into ever shorter time slices of increasing number and the variance is calculated for the data within

each time slice. By taking the mean over the whole number of slices and doing this for all divisions in time, a line is calculated, which is ~~characteteristic~~ characteristic for the development of variance in time.

Instead of considering variance in each time slice, we use the relative standard deviation $\frac{\sigma}{\mu}\sigma_x = \frac{\sigma}{\mu}$, which is the definition of ~~variabiltiy~~ variability following Junge (1974). It is calculated in each time slice and the mean gives the value for the corresponding time scale. In the following, time scale 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 $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}$ ~~and~~ $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$, $\mathrm{MOD}_{\mathrm{RANDPATH}}$ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$. The time scale changes from about 5 min to 5 a along the logarithmically spaced abscissa. As CO is a medium long-lived trace gas with an atmospheric lifetime of 2-3 months and a pronounced annual cycle, the mean variability increases up to time scales of 1 a. The variability of $\mathrm{MOD}_{\mathrm{RANDPATH}}$ ~~is larger~~ and $\mathrm{MOD}_{\mathrm{RANDLOC}}$ is larger than that of $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ on almost all time scales. For time scales of 30 d and more, however, the lines of all three datasets run in parallel, showing an increase up to 1 a, from when on the variability does not increase. This is consistent with the annual cycle of CO, which is also the cause for the relative decrease sharply at 0.5 a and 1.5 a. For time scales below 30 d, the distribution of flights in one month dominates the variability analysis. $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}$ $\mathrm{MOD}^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ includes only up to four flights on consecutive days, the mean variability does not decrease when going to time scales between 30 d and 4 d, while in $\mathrm{MOD}_{\mathrm{RANDPATH}}$, continuosly less data ~~is~~ are included in each time slice, leading to a continuous drop in the variability. For time scales of less than 1 d, the data ~~comes~~ 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 ~~in the~~ for time scales shorter than 30 d ~~(not shown)~~until time scales shorter than one day are reached, from when on 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 the representativeness of one dataset for another. A score $\mathrm{R}^{t,h}_{\mathrm{var}}$ describing the representativeness is defined from the difference of the values of the variability analysis, using the following equation:

$$\mathrm{R}^{t,h}_{\mathrm{var}} = \log_{10}\left(\left|\overline{\left[\frac{\sigma_1^{t,h}}{\mu_1^{t,h}}\right]} - \overline{\left[\frac{\sigma_2^{t,h}}{\mu_2^{t,h}}\right]}\right|\right) \tag{5}$$

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

Decreasing values of $\mathrm{R}^{t,h}_{\mathrm{var}}$ mean better representativeness, the value always being negative. Depending on t, the representativeness in different time scales can be evaluated. We used time scales of 30 d, 0.25 a, 0.5 a, 1 a, 2 a and 5 a to calculate $\mathrm{R}^{t,h}_{\mathrm{var}}$. When applying this method to all height bins, a profile in $\mathrm{R}^{t}_{\mathrm{var}}$ is calculated for each species. This is one possible definition for

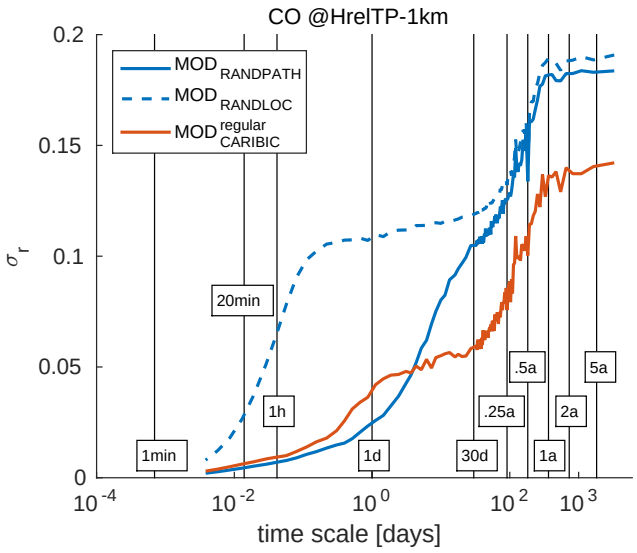

Figure 3. Variability analysis calculated for CO for $\mathrm{MOD_{RANDPATH}}$, $\mathrm{MOD_{RANDLOC}}$ and $\mathrm{MOD^{regular}_{CARIBIC}}$ at HrelTP $= -1\,\mathrm{km}$ (one kilometer below the tropopause)~~for $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDPATH}}$~~. The time scales used to calculate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

representativeness. Yet it has to pass the two requirements of being related to number of samples and variability outlined in Sec. 3.3. The results of testing this will be presented in Sec. 6.2.

4.3 Relative ~~difference~~differences

The third approach to assess representativeness is to analyze the relative differences between the climatologies from two differently large datasets. The procedure is summarized in Equation 6:

$$\mathrm{R^{\mathit{h}}_{rel}} = \log_{10}\left(\frac{1}{12}\sum_{m=1}^{12}\frac{|\mu_1^{m,h} - \mu_2^{m,h}|}{\mu_2^{m,h}}\right) \tag{6}$$

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

Figure 4 shows an example of relative differences between CO from $\mathrm{MOD_{~~CARIBIC~~}^{regular}}$ and the larger dataset $\mathrm{MOD_{RANDPATH}}$. The differences are small, mostly below an absolute value of 0.15. $\mathrm{R_{rel}}$ is defined (in Equation 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 Figure 4 due to less data there. Like for $\mathrm{R^{\mathit{t}}_{var}}$, decreasing values in $\mathrm{R_{rel}}$ mean better representativeness. And like $\mathrm{R^{\mathit{t}}_{var}}$, $\mathrm{R_{rel}}$ has to be tested for passing

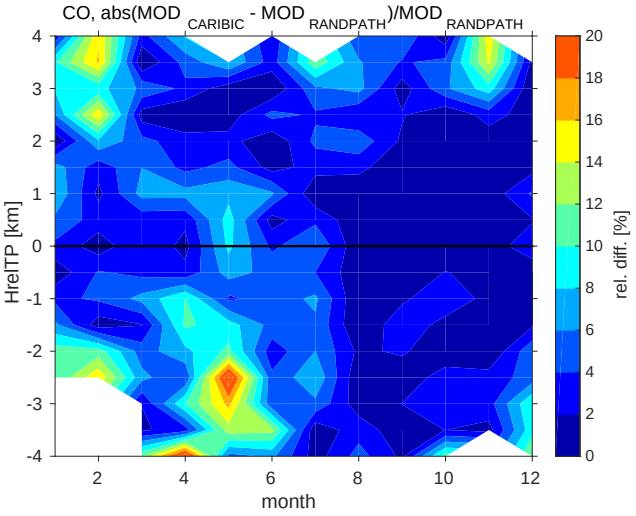

Figure 4. Relative differences of CO for MOD~~CARIBIC~~ regular CARIBIC and MOD$_{\mathrm{RANDPATH}}$. This is the basis used to calculate R_{rel}.

the requirements of being related to number of samples and variability (see Sec. 3.3) in order to be acceptable as a score for representativeness. The results of testing this will be discussed in Sec. 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 MOD~~CARIBIC~~ regular CARIBIC as a full model climatology. This is more obvious if taken to the power of 10, in which case the uncertainty will take values between 0 and 1. Use of this will be made in Section 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 representativeness of MEAS$_{\mathrm{CARIBIC}}$. MOD$^{\mathrm{regular}}_{\mathrm{CARIBIC}}$ and MOD$^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ are used as a placeholder for MEAS$_{\mathrm{CARIBIC}}$ and compared to other model datasets (MOD$_{\mathrm{RANDPATH}}$ and MOD$_{\mathrm{RANDLOC}}$) in the analysis. The results derived from these model datasets will be interpreted for MEAS$_{\mathrm{CARIBIC}}$ in Sec. 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 MOD$^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ will be used as an indicator of its complexity and compared to the variability of MEAS$_{\mathrm{CARIBIC}}$. Similar to Equation 1, we use the relative standard deviation $\sigma_r = \sigma/\mu$ as a measure for variability when comparing model and measurements. Variability of a certain time scale, e.g. 20 min, will be referred to as 20 min variability in the following, accordingly for other time scales.

5.1 Influence of short time scales on the climatological mean

All model datasets have been created from gridded datafiles with a certain resolution (2.8° or about 200 km, see Sec. 2.2). Considering the median airspeed of the CARIBIC aircraft of 885.1 km h^{-1}, this model resolution corresponds to a time scale of about 20 min. $MEAS_{CARIBIC}$ has a time resoltution 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 $MEAS_{CARIBIC}$ than in $MOD^{sampled}_{CARIBIC}$. To what extent this small scale variability influences the climatological values is investigated here.

By reducing the 20 min variability in $MEAS_{CARIBIC}$ to that of $MOD^{sampled}_{CARIBIC}$, it is possible to determine the influence of the small scale variability on the climatological mean values. The reduction in variability was done separately for each species and height to account for differences in terms of model complexity between the species. In order to reduce the variability in the time series, they were smoothed out, the method is presented in App. B. The smoothing number used in this method indicates how much variability has been removed. The 20 min variability of $MEAS_{CARIBIC}$ was then calculated for several smoothing numbers.

Figure 5 (left panel, solid lines) shows how the 20 min variability drops for all species if the data are smoothed progressively (increasing the smoothing number). The leftmost point for each species corresponds to the full 20 min variability, while this variability drops to zero if the time intervals considered in smoothing become much longer than 20 min. The dashed lines show the full model variability, which was not smoothed out. The crosspoints of the full and corresponding dashed line indicate the smoothing numbers for which $MEAS_{CARIBIC}$ has the same 20 min variability as $MOD^{sampled}_{CARIBIC}$. $MEAS_{CARIBIC}$ in which each species has been smoothed to this point will be referred to as $MEAS^{smoothed}_{CARIBIC}$.

Climatological mean values of $MEAS^{smoothed}_{CARIBIC}$ were then compared to mean values from $MEAS_{CARIBIC}$ with the full variability, thereby determining the influence of the reduced 20 min variability. A similar influence is expected by the coarse model resolution, which by definition has the same 20 min variability as $MEAS^{smoothed}_{CARIBIC}$.

The mean relative difference of the climatologies for different species between $MEAS^{smoothed}_{CARIBIC}$ and $MEAS_{CARIBIC}$ is displayed in Figure 5 (right panel). 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 % for the other species. 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. E.g. H_2O has very low stratospheric mixing ratios, that are reached in small-scale intrusions of stratospheric air encountered during flight. If these small-scale structures are smoothed out, the mean values become larger and the difference of $MEAS^{smoothed}_{CARIBIC}$ and $MEAS_{CARIBIC}$ is large and positive.

The relative differences show the influence of a lower variability that is equal to that of $MOD^{sampled}_{CARIBIC}$. This therefore shows that the coarse model resolution does in principle not lead to very large errors in climatological mean values. Nevertheless, the

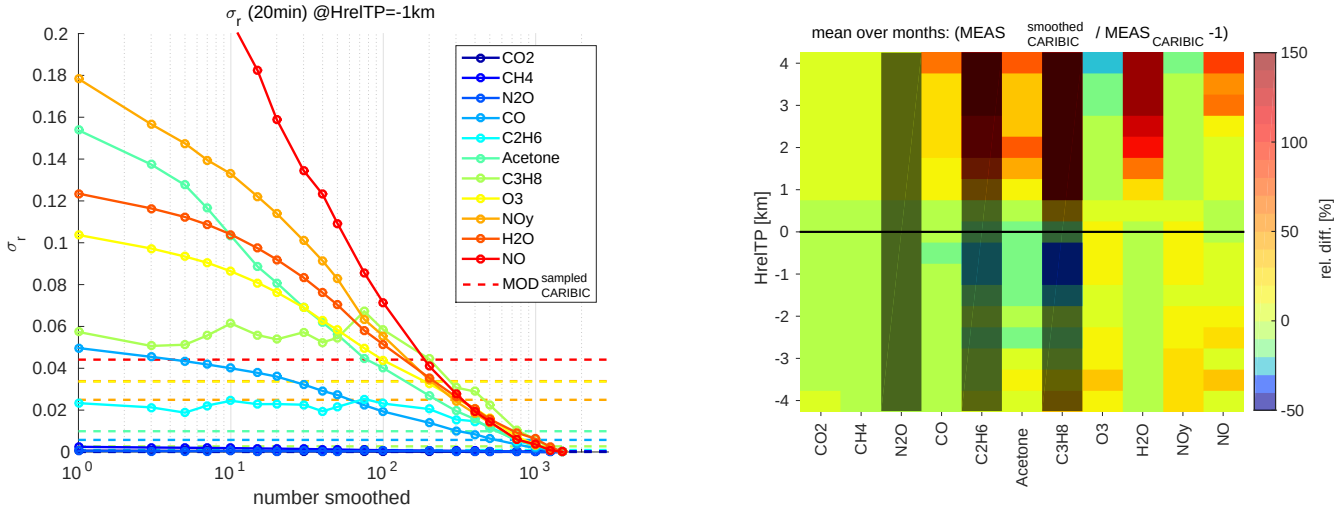

Figure 5. Left panel: 20 min variability of i) $\mathrm{MEAS}_{\mathrm{CARIBIC}}$, that has been smoothed out to an increasing degree, indicated by an increasing smoothing number (solid lines) and of ii) $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ (dashed lines), both for HrelTP $= -1$ km. The crosspoint of the dashed and corresponding full line indicate the smoothing number that is needed to reproduce the 20 min variability of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$. Right panel: Mean relative differences of $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}_{\mathrm{CARIBIC}}$. $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$ has been smoothed to have the same 20 min variability as $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$, using the smoothing number from the left hand panel. 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 therefore not considered here.

model could have other defiencies 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, $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ has been altered in such a way to reproduce the 20 min variability of $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$, see the preceeding 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 Sec. 4.2, variability depends on the time scale for which it is considered. In order to evaluate the model performance, we compare σ_r on time scales of 30 d and 1 a. 30 d variability includes data from typically 4 flights, so this is a measure for the atmospheric variabilty on the global, large scale dynamics. 1 a variability gives a good impression of the annual cycle, as it includes data from many flights and different years. Figure 6 shows $\sigma_r^{\mathrm{MOD}}/\sigma_r^{\mathrm{MEAS}}$ for time scales of 30 d (left) and 1 a (right), using the datasets $\mathrm{MOD}^{\mathrm{sampled}}_{\mathrm{CARIBIC}}$ and $\mathrm{MEAS}^{\mathrm{smoothed}}_{\mathrm{CARIBIC}}$

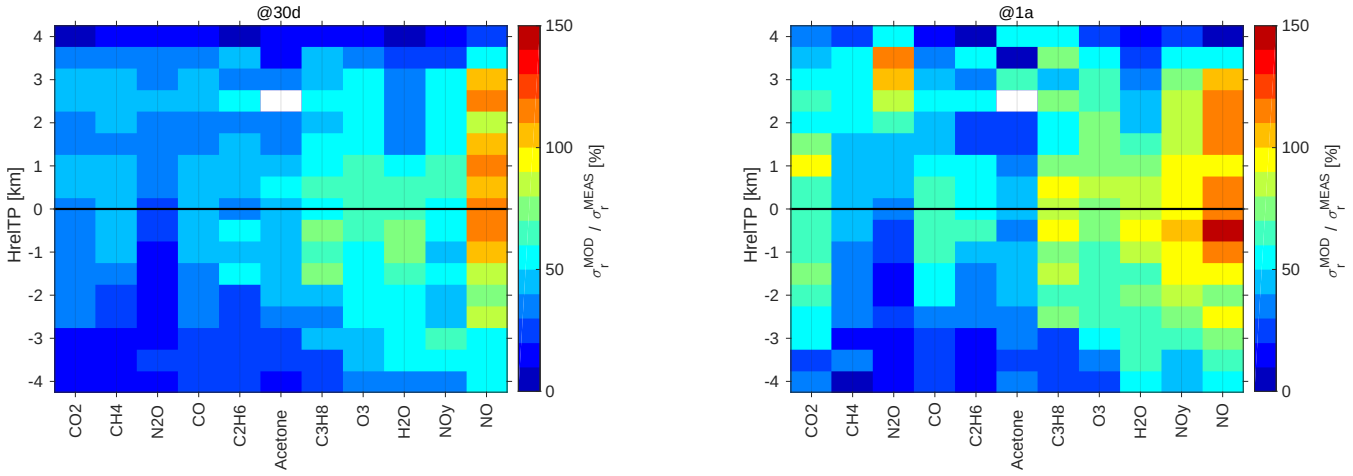

Figure 6. $\sigma_r^{MOD}/\sigma_r^{MEAS}$ given in percent for time scales of 30 d (left) and 1 a (right), where MOD stands for $MOD_{CARIBIC}^{sampled}$ and MEAS stands for $MEAS_{CARIBIC}^{smoothed}$. Values greater than 50 % indicate the high model complexity.

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 d variability shown in Figure 6 (left) 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 time scale of 1 a shows the variability that represents seasonality. The model does a better job for this time scale than for 30 d, 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 shorter lived species to the right. There are species that are not as well represented, while this also depends on the height considered (e.g. high values for stratospheric N_2O).

The model variability is influenced by many factors including the dynamics, 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 $MEAS_{CARIBIC}^{smoothed}$ is used as a reference for the comparison. If compared to the original $MEAS_{CARIBIC}$, 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 defiencies of the model and what leads to the differences between the species.

As is shown in Figure 6, the model reaches more than 50 % of the variability of the measurements, depending on the species and time scale. 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 (height relative to the tropopause) is 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° (≈ 110 km), while the measurement data have a much higher resolution (≈ 2.5 km, see Sec. 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. This also affects $MEAS_{CARIBIC}^{smoothed}$.

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 (Sec. 6.1), the variability analysis (Sec. 6.2) and the relative difference (Sec. 6.3) to $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}^{\mathrm{regular}}{}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$. All have to be related to the number of flights and the ~~the variabiltiy~~ variability of the species as discussed in Section 3.3. ~~A similar analysis has also been performed with data~~ These methods have also been applied to data not from an atmospheric model but from a random number generator, leading to equivalent results. ~~This study is~~ These are presented as supplementary material to the article. Sec. 6.4 interprets the results by species as a representativeness uncertainty. Finally, Sec. 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 $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}$ ~~and~~ $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}^{\mathrm{regular}}{}_{\mathrm{CARIBIC}}$ and $\mathrm{MOD}_{\mathrm{RANDPATH}}$ yields a first important result. Independent of the trace gas and ~~HrelTP~~ height considered, the result is always negative (not shown). This means that the data in each bin of $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}$ ~~is~~ $\mathrm{MOD}_{\mathrm{CARIBIC}}^{\mathrm{regular}}$ 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 ~~is~~ 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 ~~has~~ have already been reported by Kunz et al. (2008). On the one hand side this could mean that $\mathrm{MOD}_{\sout{\mathrm{CARIBIC}}}^{\mathrm{regular}}{}_{\mathrm{CARIBIC}}$ 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 ~~correct~~ 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 ~~further discussed with the help~~ also the result of a sensitivity study, ~~the results of which are presented~~ which is discussed as supplementary material to this text.

In addition to binning into twelve 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 nine 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. Especially, 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 postive 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.

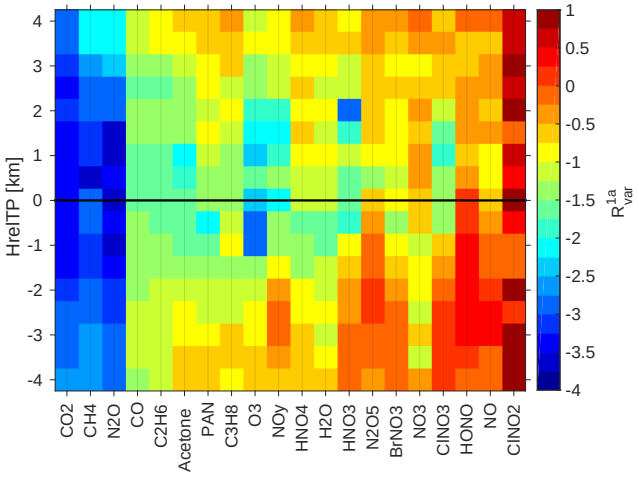

Figure 7. R_{var} calculated according to Equation 5 for a time scale of 1 a for all species in all height bins, using $MOD_{~~CARIBIC~~}^{regular\ CARIBIC}$ and $MOD_{RANDPATH}$. Low values indicate small differences in variability.

6.2 Applying the variability analysis

This section presents the results of the application of the variability analysis to $MOD_{~~CARIBIC~~}^{regular\ CARIBIC}$ and $MOD_{RANDPATH}$. Equation 5 was applied for different time scales (30 d, 0.25 a, 0.5 a, 1 a, 2 a and 5 a) to calculate R_{var}. The results are exemplarily discussed for a time scale of 1 a, shown in Figure 7, in which the results are sorted using the values of τ^* displayed in Figure 2.

R_{var} shows a strong ~~relationship with~~ dependancy on τ^*. This is visible from Figure 7, in which the results are sorted ~~using the~~ with decreasing values of τ^* ~~displayed in~~ (from Figure 2~~, that is~~), i.e. with increasingly higher atmospheric variabilty from left to right. The Pearson correlation coefficient ρ of R_{var} and τ^* is high, $|\rho| > 0.9$ in all height bins, independent of the time scale. R_{var} also shows a strong relationship to the number of samples: The amount of data in both $MOD_{~~CARIBIC~~}^{regular\ CARIBIC}$ 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 Sec. 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 time scale of 5 a, increasing continously when considering shorter time scales to $|\rho| \approx 0.95$ for the time scale of 30 d. Considering a logarithmic relationship inreases the goodness of fit for longer time scales, while it decreases that for shorter time scales ($|\rho| \approx 0.85$ for both 5 a and 30 d).

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

~~This is also supported by the study of random number data presented as supplementary material.~~

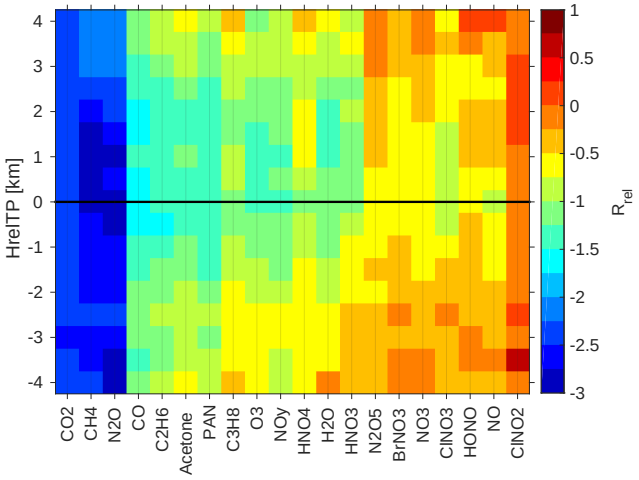

Figure 8. R_{rel} calculated ~~for~~ according to Equation 6 for all species in all height bins, using $MOD_{CARIBIC}$ ~~CARIBIC~~ regular CARIBIC and $MOD_{RANDPATH}$. Low values indicate small differences in climatological mean values.

This shows that by using the ~~relive~~ relative standard deviation (Equation 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 time scales of variability in a dataset and for this it remains a valid method. In order to infer representativeness, it is more appropriate to use the relative standard deviation in the analysis instead of the absolute variance.

6.3 Relative differences

R_{rel} was calculated for each species in each height bin according to Equation 6, ~~see~~ results are presented in Figure 8.

Figure 8 shows how low variability (decreasing to the left, values taken from Figure 2), is linked with good representativeness (low values in R_{rel}~~, or good representativeness, respectively (see Sec. 4.3)~~). R_{rel} decreases linearly with increasing variability τ^* with a high Pearson correlation coefficient greater than 0.95 for all height bins (not shown). ~~The relation of R_{rel} with the number of values is also visible~~ As visibile in Figure 8~~as the values decrease~~ , R_{rel} also decreases with the number of data points, ~~this number having its maximum~~ which maximizes just around the tropopause and ~~decreasing~~ decreases above and below it (see Figure 1).

~~This shows that R_{rel} passes the requirements of being related to number of samples and variability τ^* and can be used as a measure for representativeness.~~

This ~~relation with~~ dependance on the number of ~~values was tested in more detail: Each of the random paths of $MOD_{RANDPATH}$ was divided into three parts.Each part is then eight hours long, like a typical intercontinental flight with CARIBIC, and there are a total number of altogether 3888 shorter random flights. R_{rel} was then calculated for $MOD_{RANDPATH}$ and these subsamples,~~

~~increasing their size by including more of the 3888 shorter random flights~~data points was also tested by using $MOD^{3}_{RANDPATH}$ described in Sec. 3.3. The Pearson correlation coefficient ρ between the number of shorter random flights and R_{rel} was ~~larger than 0.9~~ $\rho \approx 0.95$ for all species in all heights~~(not shown)~~. 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 ~~using~~ Figure 7, Figure 8 can therefore be used to judge upon the representativeness of $MOD_{~~CARIBIC~~}^{regular}$ for $MOD_{RANDPATH}$. R_{rel} can be transformed into a relative difference in percent, by taking R_{rel} to the power of ten. A score of -2 stands for a mean relative difference of $1\,\%$.

The score that discriminates representative from the non-representative case has to be arbitrarily chosen (see Nappo et al. (1982) and Ramsey and Hewitt (2005)). This score gives the uncertainty within which the data ~~is~~ 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 seperated from those species that are not representative using the results from Figure 8. But the score of -2 is arbitrary. If it is reduced to -1.5 (roughly $3\,\%$ relative difference), $MOD_{~~CARIBIC~~}^{regular}$ 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 Equation 6) and R_{var} (see Equation 5) to be adequate scores to describe representativeness. After reconsidering the question we asked in the Section 3.1 (Is a climatology compiled from CARIBIC data representative ~~for~~ 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~~and shows the slightly higher correlation coefficient~~. A further discussion of R_{var} is beyond the scope of this paper. As noted in Sec. 4.3, R_{rel} is also comprehensible as an uncertainty ~~error~~ for using the smaller dataset to compile a climatology and will be called representativeness uncertainty correspondingly.

In order to asses the uncertainty for accepting CARIBIC measurement data to create a climatology, ~~all the gaps (e.g. due to instrument problems or measurement intervals $> 10\,s$) in measurements and HrelTP (calculated from ECMWF fields in the case of measurements) have to be mapped onto MOD~~model data have to contain the same amount of data as $MEAS_{CARIBIC}$ ~~of the corresponding species and HrelTP calculated from the model. This was done and R_{rel} - taken to the power of 10 - recalculated using~~, which is why $MOD^{sampled}_{CARIBIC}$ (see Sec. 2) will be used in the following. In addition, $MOD_{RANDLOC}$ ~~with an even distribution in pressure,~~ (see Table 1) was used as reference, as it has a random sampling pattern and represents the full model state, independent of the sampling pressure. The limits in pressure where again set to $180\,hPa < p < 280\,hPa$. The ~~result~~ resulting R_{rel} is shown in Figure 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 ~~different species are equally well represented in terms of the processes that act on them, as is the case here, see Section 3.2.Figure 9 therefore gives the representativeness uncertainty not only for a reduced set of $MOD_{CARIBIC}$, but also for the CARIBIC measurements. It can be used to answer~~ complexity of the model is sufficiently high for each species. This has been shown by comparing the variability of $MOD^{sampled}_{CARIBIC}$

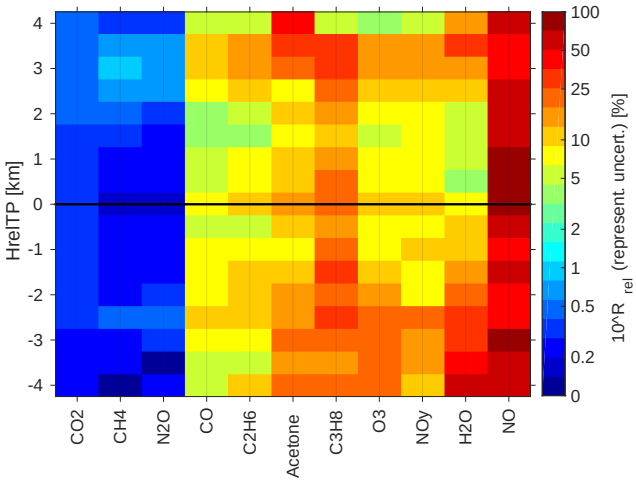

Figure 9. Representativeness uncertainty for using the CARIBIC data (that is 334 long-distance flights, see Table 1) to compile a climatology: $10^{R_{rel}}$ calculated from $MOD_{CARIBIC\ RANDLOC}$ and $MOD_{RANDLOC}^{sampled}{}_{CARIBIC}$. Low values indicate small representativeness uncertainties.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

and $MEAS_{CARIBIC}^{smoothed}$ for different time scales (see Sec. 5). The discussion of 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 Sec. 3.2: For which species is a climatology compiled from CARIBIC data representative ~~for~~ of the tropopause region in mid-latitudes?

~~The influence of the limit in pressure is shown in the supplement.~~

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

Many of the species that sum up to NO_y in the model are not actually measured by CARIBIC and therefore ~~get no value~~ are not displayed in Figure 9. In general, the representativeness uncertainty is lowest where there are most measurements, which is just around the tropopause (see Figure 1). This effect overlays the physical reasons for the different ~~values of the uncertainty for all species considered. If the limits in pressure are expanded in using $MOD_{RANDLOC}$, the uncertainty increases markedly, as is shown in supplementary material. The reasons for this have been discussed in Section 3.1~~uncertainties for the considered species.

~~and~~ NO ~~have~~ has the highest uncertainty of 90 %~~() and up to 100 % in the case of~~. We propose two possible reasons: On the one hand, there are many gaps in the observations. But ~~and~~ NO ~~are~~ 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 variabilty of many components is higher in the troposphere, where the uncertainty is 30 % at its maximum. Above, it is smaller than 10 % and the climatology therefore quite trustworthy.

It is interesting to note that C_2H_6 and C_3H_8, both collected in whole air samples still reach ~~uncertainty values~~ uncertainties comparable to those of other species in their range of τ^*. This is due to the fact that these are ~~rather~~ moderately long-lived species for which only a ~~moderate~~ 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 unnoticable compared to the vertical increase.

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

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

The long-lived trace gases CH_4, N_2O and CO_2 (all detrended as described in Sec. 2.1) all have representativeness uncertainties of less than ~~5 %~~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 example and summary, the representativeness uncertainty will be applied to climatologies of ~~,~~ CO, CO_2 and O_3, shown in Figure 10. CO is shown for $MOD_{\text{~~CARIBIC~~}}^{\text{sampled CARIBIC}}$ (top left, panel A), MOD_{RANDLOC} (top right, panel B) and CARIBIC measurements ($MEAS_{\text{CARIBIC}}$, center left, panel C). The white space in these figures has three possible reasons: the aircraft could have never flown in that bin, there could be measurement gaps in CO or a gap in HrelTP. The measurement gaps of CO and HrelTP from $MEAS_{\text{CARIBIC}}$ have been mapped onto $MOD_{\text{~~CARIBIC, the two upper left hand climatologies of Figure 10~~}}^{\text{sampled CARIBIC}}$, 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 ~~of MOD_{CARIBIC}~~ ($MOD_{\text{CARIBIC}}^{\text{sampled}}$ and MOD_{RANDLOC}, panels A and B) and found that R_{rel} is a good descriptor for the representativeness of one for the other. By ~~assuming~~ 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 $\mathrm{MEAS_{CARIBIC}}$.

By again defining $\mathrm{R_{rel}} = -1$ (10 % uncertainty, one third of the seasonal variation) as the limit for representativeness, the climatology of $\mathrm{MEAS_{CARIBIC}}$ ~~in~~ (Figure 10~~(center left~~, center left, panel C) was shaded in grey where it is not representative. The representativeness uncertainty shown in Figure 9 only serves as a first indication of the expected uncertainty when resolving monthwise. The center right panel (panel D) displays the standard deviation of CO from $\mathrm{MOD_{RANDLOC}}$. 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 Figure 10 (panels E and F), which shows climatologies of $\mathrm{CO_2}$ and O3. $\mathrm{CO_2}$ shows a small annual variation around a high background value. So 10 % uncertainty could be easily reached by a single measurement, which would certainly not be representative ~~for~~ of the whole year. The shading for $\mathrm{CO_2}$ in Figure 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 O3, 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 Figure 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 $\mathrm{SF_6}$, 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 $\mathrm{SF_6}$ 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 representative yet, how often does one have to fly in order to achieve a representative climatology?

~~As explained in Section 6.3, $\mathrm{R_{rel}}$ increases linearly with the number of flights considered, the Pearson correlation coefficient of this relationship exceeding 0.9 for all species . This was tested by cutting all paths of $\mathrm{MOD_{RANDPATH}}$ into three flight legs and testing~~ This question can be answered with the help of $\mathrm{MOD^3_{RANDPATH}}$. Figure 11 shows the representativeness uncertainty for some species and different numbers of ~~these against the whole dataset. For low numbers, the relationship of $\mathrm{R_{rel}}$ and the number of flights is better described by a logarithmic function. This is also motivated by the study using data from a random number generator, which is presented as supplementary material to this text. So here, $\mathrm{R_{rel}}$ was fit to the logarithm of the number of flights. The number of flights necessary to reach a specific representativeness uncertainty , can then be read from the regression~~

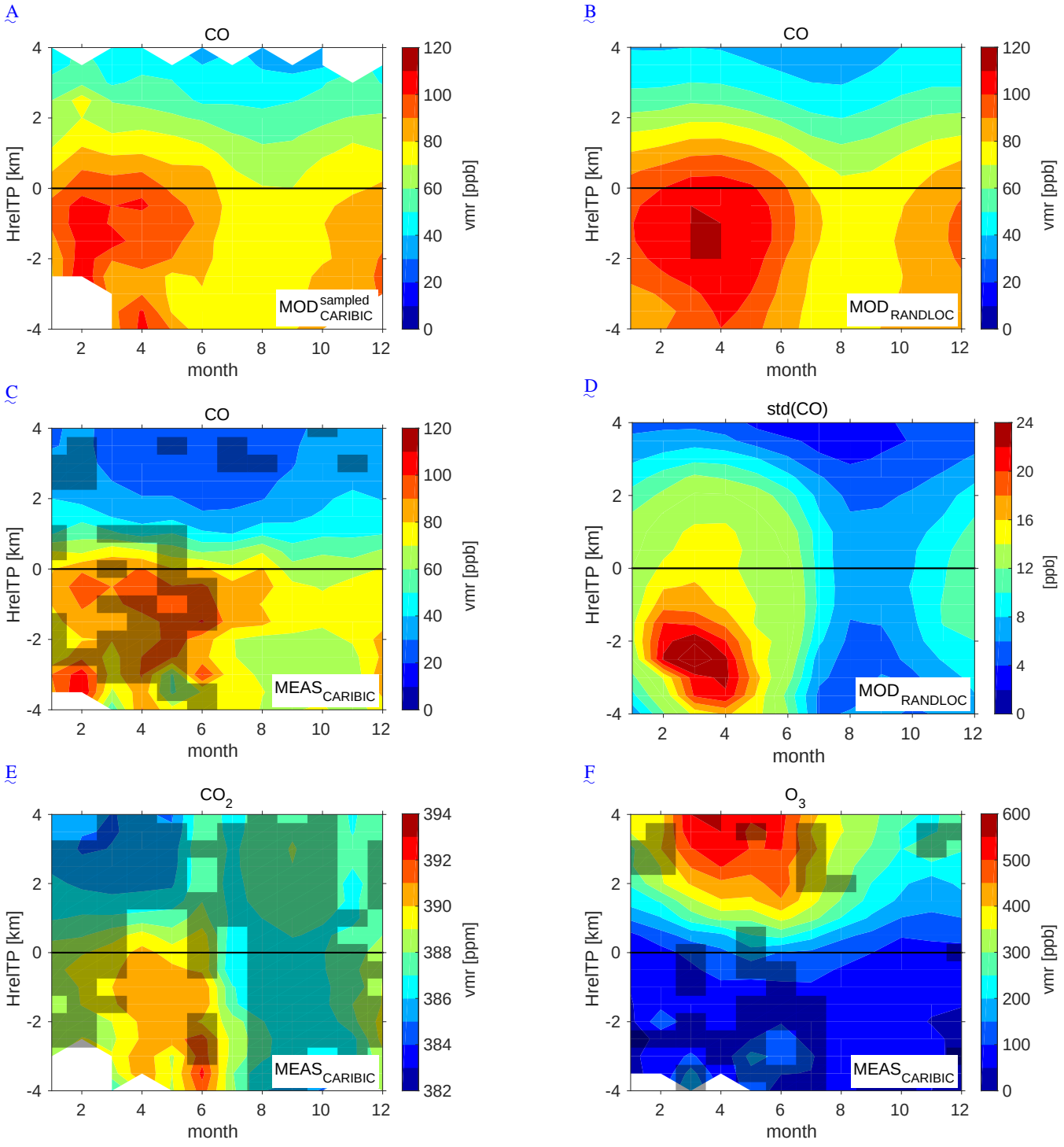

Figure 10. Climatology of CO, built from $MOD_{\sout{CARIBIC}}^{\text{sampled CARIBIC}}$ (~~top left, including the measurement gaps in $MEAS_{CARIBIC}$ due to instrument problems~~panel A), $MOD_{RANDLOC}$ (~~top right~~panel B) and the CARIBIC measurements ($MEAS_{CARIBIC}$, ~~center left~~panel C). Areas of $10\hat{}R_{rel} > 0.1$, calculated from the top row, were used to shade non-representative areas in the climatology of $MEAS_{CARIBIC}$ in grey. ~~The right center panel~~ Panel D displays the 1σ standard deviation of CO from $MOD_{RANDLOC}$. The bottom row (panels E and F) displays climatologies from $MEAS_{CARIBIC}$ of CO_2 (left) and O3, shaded with $10\hat{}R_{rel} > 0.003$ and $10\hat{}R_{rel} > 0.15$, respectively.

~~line calculated from R_{rel} and log(number of flights).The result for $R_{rel} = -1$, corresponding to a representativeness uncertainty of 10 %, is shown in Figure 11. It is in principle a translation of the value of R_{rel} from Figure 8 into a number of flights that are necessary to reach an uncertainty of 10 %. $R_{rel} = -1$, i.e. 10 % uncertainty are again set as a mean value, which may be too high for some species, depending on their annual cycle. Number of 8 h flights necessary to reach a representativeness uncertainty of 10 % ($R_{rel} = -1$). This result was calculated using $MOD_{RANDPATH}$, the method is explained in the text.~~

flights. As has been discussed in Section 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.

~~As is displayed~~ E.g., for (detrended) CO_2, the mean value of $MOD_{RANDLOC}$ 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 few flights, much less than those included in $MOD^{sampled}_{CARIBIC}$, indicated by the dashed line in Figure 11 ~~and goes in line with Sec. 6.4, CARIBIC with a total number of~~ at 334 flights~~from 2005-2013 is already representative for many long-lived species with low variability (high τ^*), to the left of the plot. For many of the nitrogen containing species with low τ^* (to the right), data representative of a climatology is probably impossible to collect within IAGOS-CARIBIC. The necessary number of flights reach up to more than 3000 in the tropospheric heights, corresponding to almost all data in $MOD_{RANDPATH}$. For those species in the center of the plot, the representativeness uncertainty may be further reduced by flying more often, especially for those with flight numbers below 1000 like~~ .

For O_3, ~~or . Due to their lower variability in the lower stratosphere, the climatological values of these species are already representative. In general, the uppermost and lowermost heights need more flights as they are less frequently probed by the aircraft.~~ 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 low already (< 5 % in this height), while the uncertainty is continuosly reduced if the number of flights increases.

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

Other species that are not included in Figure 11 can be deduced from their value of τ^* with the help of Figure 2. Those species measured in air samples need even more CARIBIC flights than indicated by the number in Figure **??**, 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 ~~our representativeness concept~~ the concept of representativeness, we apply general rules to investigate ~~the feasibility of compiling~~ whether climatologies from IAGOS-CARIBIC trace gas measurements can be seen as

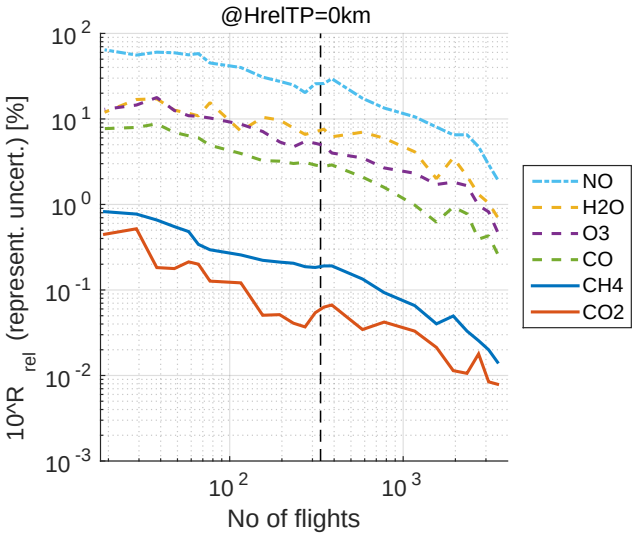

Figure 11. Representativeness uncertainty for different numbers of flights for some species. The number of flights in $MEAS_{CARIBIC}$ is indicated by the vertical dashed line. Other species can be deduced from their value of τ^* with the help of Figure 2.

representative. We answer the specific question: For which species is a climatology compiled from CARIBIC data representative ~~for~~ of the tropopause region in mid-latitudes?

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

~~Of these three datasets, $MOD_{CARIBIC}$ and $MOD_{RANDPATH}$ are used to develop methods describing representativeness , applying~~ 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 models 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 time scale, 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 ~~, a~~ (and the Mann-Whitney, Brown-Forsythe and Levene's test), (ii) variability analysis following Kunz et al. (2008) and ~~a relative differences~~

~~test~~ (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 ~~show~~ demonstrate that they pass ~~the two requirementswe defined as having to be met by any description of representativeness: Representativeness should increase with the number of measurements and decrease with the variability of the species. Variability was defined following~~ these two requirements, while the statistical tests are all too strict. R_{rel} ~~is more applicable for answering the question, asking for~~ (describing the representativeness of ~~for a climatology. It~~ a climatology) is better suited for answering the question and is therefore used ~~for the~~ in the remaining analysis.

~~A score of $R_{rel} = -1$ defines~~ The score R_{rel} is easily converted to a representativeness uncertainty ~~of 10 %. It is used to discriminate the representative from the non-representative compiled climatologies~~in percent and this measure is used in the discussion. The results ~~(using $MOD_{CARIBIC}$ and $MOD_{RANDLOC}$) show that the data of~~ show that CO_2, N_2O ~~,~~ and CH_4 ~~,~~ have very low uncertainties (below 0.4 %). CO, C_2H_6, and O_3 ~~can~~ reach higher values (5 % - 20 %), but can still be used to compile representative climatologies around the tropopause~~, while acetone, .~~ NO_y and H_2O are only usable in the ~~stratosphere. ,~~ 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 ~~results strongly depend on the accepted uncertainty of 10 % and would change if this limit is set to a different value.~~

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. ~~E.g. for , 1500 to 1000 flights are necessary for a representative climatology in the upper troposphere,~~ 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 ~~number strongly decreasing with height~~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 ~~data set~~ dataset should mirror the measurements, the other should be much larger, taking into account certain statistical properties of the measurement ~~data set~~dataset, so that the two ~~data sets~~ 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. ~~In following~~ By help of the methods presented here, representativeness is given a sound mathematical description, returning an uncertainty characterizing the specific dataset.

Appendix A: Limitations in longitude and pressure

$MEAS_{CARIBIC}$ is limited in longitude (the Pacific Ocean is never sampled) and pressure (as 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 $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$, 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 Equation 4) using $MOD_{RANDLOC}$ and $MOD^{regular}_{CARIBIC}$ within $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$. Figure 12 shows the results (right panel). For comparison, the left panel of Figure 12 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 on the right hand side ($10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$), except just above the tropopause, where $MOD^{regular}_{CARIBIC}$ 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.

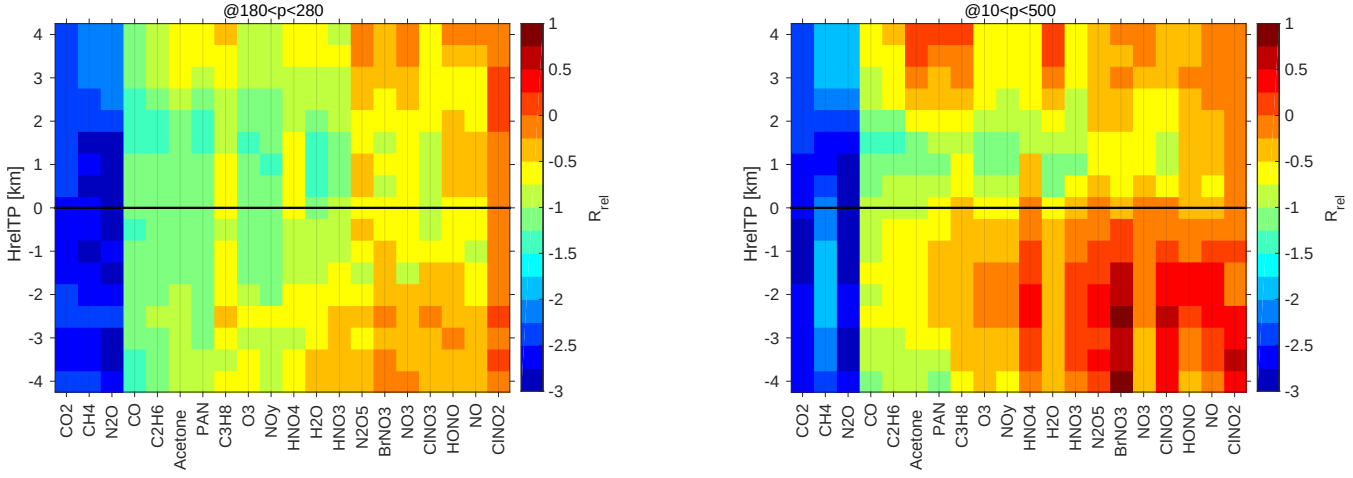

Figure 12. R_{rel} calculated from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$ with the range of p set to $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$ (left) and $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$ (right). Low values indicate small climatological differences. The difference between the two panels shows the influence of expanding the limits in p when calculating the climatological mean values with HrelTP used as a vertical coordinate.

The 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

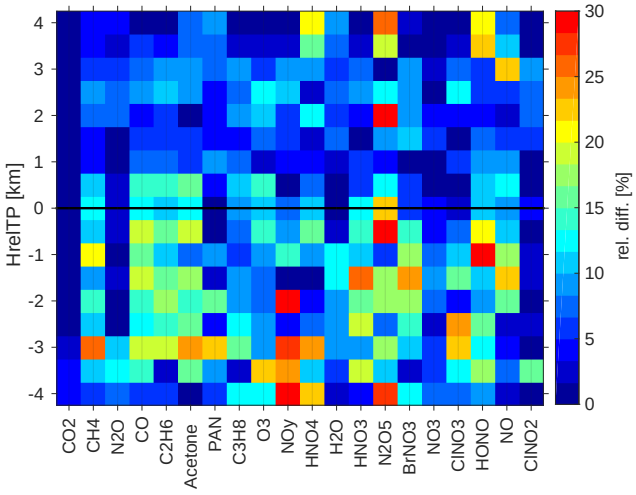

Figure 13. $|R^A_{rel}/R^B_{rel} - 1|$, given in percent. This is the fraction of the representativeness uncertainty introduced in R_{rel} calculated from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$ by including the Pacific ocean in $MOD_{RANDLOC}$, even though it is not sampled by $MOD^{regular}_{CARIBIC}$. Both, $textR^A_{rel}$ and $textR^B_{rel}$ have been calculed from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$, excluding the Pacific in $MOD_{RANDLOC}$ in the calculation of $textR^B_{rel}$.

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 $MOD_{RANDLOC}$ and $MOD_{RANDPATH}$ also cover the Pacific. The uncertainty introduced by taking the Pacific into account in $MOD_{RANDLOC}$ is investigated by calculating R_{rel} from $MOD^{regular}_{CARIBIC}$ and $MOD_{RANDLOC}$ in two different setups. R_{rel} is calculated from full $MOD_{RANDLOC}$ and $MOD^{regular}_{CARIBIC}$ (denoted by R^A_{rel}) and compared to R_{rel} calculated with $MOD_{RANDLOC}$ limited in longitude λ to $120°W < \lambda < 120°E$ (denoted by R^B_{rel}). The result is shown in Figure 13 as relative differences $|R^A_{rel}/R^B_{rel} - 1|$ between the two uncertainties. The relative differences show the share of the uncertainty inherent in $MOD^{regular}_{CARIBIC}$ because the Pacific is included in the reference dataset $MOD_{RANDLOC}$.

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 less 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 little (less than 3%) if the Pacific is included in the reference climatology of $MOD_{RANDLOC}$. 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 on lifetime, but on the source regions and the chemistry. Acetone, CO and C_2H_6 are air pollutants with strong

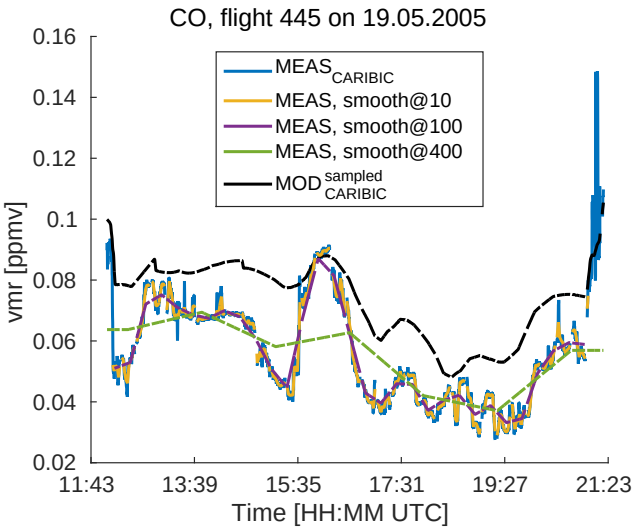

Figure 14. Timeseries of CO for flight 445 from Frankfurt to Tokyo. Shown is the time series of the interpolated model data and of the measurements. Measurements have been smoothed three times. The number indicates the length of the smoothing interval N.

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

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 datapoints 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 Figure 14 for three different lengths of the smoothing interval N.

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 manuscript 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.

References

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

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

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

Gettelman, A., Hoor, P., Pan, L., Randel, W., Hegglin, M., and Birner, T.: The extratropical upper troposphere and lower stratosphere, Reviews of Geophysics, 49, 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, Journal of Geophysical Research: Atmospheres, 115, doi:10.1029/2010JD013884, 2010.

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

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

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

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

Jöckel, P., Tost, H., Pozzer, A., Kunze, M., Kirner, O., Brenninkmeijer, C. A. M., Brinkop, S., Cai, D. S., Dyroff, C., Eckstein, J., Frank, F., Garny, H., Gottschaldt, K.-D., Graf, P., Grewe, V., Kerkweg, A., Kern, B., Matthes, S., Mertens, M., Meul, S., Neumaier, M., Nützel, M., Oberländer-Hayn, S., Ruhnke, R., Runde, T., Sander, R., Scharffe, D., and Zahn, A.: Earth System Chemistry integrated Mod-

elling (ESCiMo) with the Modular Earth Submodel System (MESSy) version 2.51, Geoscientific Model Development, 9, 1153–1200, doi:10.5194/gmd-9-1153-2016, http://www.geosci-model-dev.net/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., Heintzenberg, J., Schlager, H., Schuck, T., Slemr, F., Sprung, D., van Velthoven, P., Wiedensohler, A., et al.: Origin of aerosol particles in the mid-latitude and subtropical upper troposphere and lowermost stratosphere from cluster analysis of CARIBIC data, Atmospheric Chemistry and Physics, 9, 8413–8430, 2009.

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

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

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

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

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

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

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M., Lamarque, J., Matsumoto, K., Montzka, S., Raper, S., Riahi, K., et al.: 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., et al.: 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., et al.: Workshop on the representativeness of meteorological observations, June 1981, Boulder, Colo, Bull. Am. Meteorol. Soc., 63, 1982.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Supplement to:
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, Institute of Meteorology and Climate Research, Herrmann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[2] Karlsruhe Institute of Technology, Steinbuch Centre for Computing, Herrmann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany

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

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

~~Generally, representativeness~~ Two requirements for any score describing representativeness are essential: Representativeness is expected to ~~decrease with increasing variability and to increase~~ increase (i) with the number of ~~available samples~~ samples and (ii) with decreasing variability of the species considered. Based on ~~this assumption~~these two requirements, we investigate the suitability of the different statistical measures for ~~our problem~~investigating representativeness. The Kolmogorov-Smirnov test ~~seems too~~ is very strict and does not identify any trace gas climatology as representative – not even ~~long lived well observed~~ of long lived trace gases. In contrast, the ~~variability based scores pass the general requirements for representativeness formulated above. In addition, even the simplest metric (relative differences) seems~~ two scores based on either variability or relative differences show the expected behaviour and thus appear applicable for investigating representativeness.

~~Using~~ For the final analysis of climatological representativeness, we use the relative differences score ~~we investigate the representativeness of a large number of different trace gases . For our final consideration we assume that the EMAC model is a reasonable representation of the real world and that representativeness in the model world can be translated to representativeness~~

~~for CARIBIC measurements. This assumption is justified by comparing the model variability to~~ 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 ~~variability of CARIBIC measurements~~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 ($\leq 0.4\,\%$), while shorter-lived species like O_3 have larger uncertainties (10-15 %). Finally, we ~~show how~~ translate the representativeness score ~~can be translated~~ 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 %.

1 Introduction

This supplement discusses further results of the study of the representativeness of IAGOS-CARIBIC data using the chemistry-climate model EMAC. For abbreviations and methods, please refer to the main text. ~~Four~~ Two points are discussed here:

~~Section ?? briefly shows results of the comparison of model and measurement variability.~~ The methods to describe representativeness developed and tested with model data were also applied to data from a random number generator. This is described in Section 2. Section 3 discusses the sensitivity study of the Kolmogorov-Smirnov test using a subsample of $\mathrm{MOD}_{\mathrm{CARIBIC}}$. ~~Section ?? shows how the representativeness uncertainty of $\mathrm{MOD}_{\mathrm{CARIBIC}}$ decreases if the pressure range is increased to $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$, i.e. how the climatologies produced with data from IAGOS-CARIBIC are dependent on the pressure at which samples are taken.~~

2 ~~Comparing measurement and model variability~~

~~In order to compare model and measurement variability, the relative standard deviation $\sigma_r = \sigma/\mu$ (σ being the standard deviation, μ the mean) was calculated for $\mathrm{MEAS}_{\mathrm{CARIBIC}}$ (CARIBIC measurements) and $\mathrm{MOD}_{\mathrm{CARIBIC}}$ in each month. $\sigma_r^{\mathrm{MOD}_{\mathrm{CARIBIC}}}$ and $\sigma_r^{\mathrm{MEAS}_{\mathrm{CARIBIC}}}$ were calculated in each month. Figure ?? shows the correlation of $\sigma_r^{\mathrm{MOD}_{\mathrm{CARIBIC}}}$ and $\sigma_r^{\mathrm{MEAS}_{\mathrm{CARIBIC}}}$. Monthly variability σ_r of $\mathrm{MOD}_{\mathrm{CARIBIC}}$ over $\mathrm{MEAS}_{\mathrm{CARIBIC}}$. Colorcoding corresponds to the variability τ^* of each species. Data closer to the tropopause is plotted as larger circles.~~

~~As discussed in the main text, $\sigma_r^{\mathrm{MOD}_{\mathrm{CARIBIC}}}$ reaches 40 to 70 % of $\sigma_r^{\mathrm{MEAS}_{\mathrm{CARIBIC}}}$ for all species. The correlation coefficient of the two is 0.81. This shows that the model variability is similar for all species, justifying the use of results from the model datasets for CARIBIC measurements~~$^{\mathrm{regular}}_{\mathrm{CARIBIC}}$.

2 Calculating representativeness from random numbers

All three methods to investigate representativeness (Kolmogorov-Smirnov test, variability analysis and relative differences) have also been applied to data created with a random number generator. The results of this study are ~~discussed~~ presented here.

To produce the random numbers, 20 sets of 10^8 numbers were taken from a normal distribution. These 20 sets are referred to as species, well aware of the fact that they are purely ~~artifial~~artificial. From species to species, the standard deviation σ was set to vary from 10^{-3} to 10^3, values of the exponent ~~again~~ increasing linearly. 20 mean values μ (increasing from 10^4 to 10^8, with a linear increase in the exponent) where distributed randomly ~~onto~~ to the 20 species. This results in 20 species with different values for σ and μ. The statistics of each species will be indexed by the number 2. For short, this dataset will be called RAND.

3000 samples were taken from each of the 20 species. ~~The sample~~For each sample, 20 numbers were first randomly drawn from each species. These new numbers and all those that had been drawn before then make up this one sample. So the size increases by 20 for each sample~~, keeping the sample from before~~. This way, the relationship of the representativeness score with the sample size is directly accessible. ~~The statistics of each species will be denoted by the index 2, while samples~~ Samples are indexed by the number 1.

~~For short, this dataset will be named RAND.~~

The variability τ^* of each species ~~was~~ is defined as in Equation ~~5~~ 3 of the main text: ~~$\tau^* = \log_{10}(\mu_2/\sigma_2)$, where high values of τ^* stand for low variability~~$\tau^* = \log_{10}(\sigma_2/\mu_2)$. The two requirements set up in Section 3.3 for representativeness in general also have to hold here:

1. Representativeness has to increase with the number of samples.

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

With RAND defined in this way, it is possible to test representativeness using the variability analysis following Rohrer and Berresheim (2006) and Kunz et al. (2008) (see Section 4.2) and the relative differences (see Section 4.3). The Kolmogorov-Smirnov test was positive for very few samples (less than fifty numbers, independent of τ^*) and will not be further discussed. Its behaviour with aircraft data was subject of a sensitivity study, the results of which are shown in Sec. 3 of this supplement.

2.1 Variability analysis

The ~~variabiltiy~~ variability analysis (defined in Section 4.2 and Eq. 3) was applied in a simplified manner. As RAND is independent of time, R_{var} is reduced to just a single value containing the absolute difference of variability of each species of RAND and the sample taken thereof: $R_{var} = |\nu_1 - \nu_2|$, where ν is the mean variability. Figure 1 shows a result. The exact result is a matter of chance, as a random number generator is used. Similar to using $MOD_{\text{~~CARIBIC~~}}^{\text{regular CARIBIC}}$ and $MOD_{RANDPATH}$, a strong dependance on τ^* and a weak dependance on the number of samples is visible.

Similar to R_{var} when using $MOD_{\text{~~CARIBIC~~}}^{\text{regular CARIBIC}}$ and $MOD_{RANDPATH}$, the variability analysis using RAND meets the two requirements necessary for describing representativeness, which were described in Section 3.3 and above. This result supports the ~~findng~~ finding that R_{var} can be used as a statistic for describing representativeness.

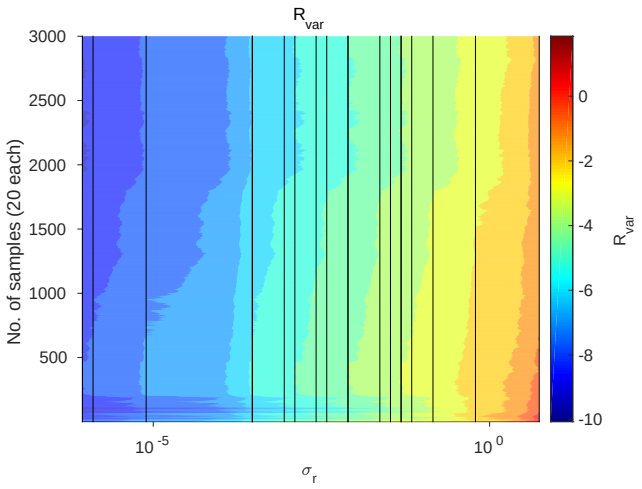

Figure 1. Representativeness score R_{var} applied to RAND. Vertical lines indicate the values of τ^* of each species.

2.2 Relative differences

Similar to R_{var}, R_{rel} is reduced to a simple relative difference when using RAND: $R_{rel} = |\mu_1 - \mu_2|/\mu_2$, where μ is the mean of the sample (index 1) and of the whole subset (index 2). Figure 2 shows ~~the~~ a result when applying R_{rel} to RAND. The dependance on τ^* is strong and linear. The result also depends on the number of samples, showing a slow increase with the number of samples. This dependance is sometimes disturbed by better values which are reached by chance when drawing from RAND.

Like for $MOD_{\text{~~CARIBIC~~ regular CARIBIC}}$ and $MOD_{RANDPATH}$, R_{rel} passes both conditions for a valid description of representativeness: it depends on variability τ^* and on the number of samples. The latter is also being influenced by chance and generally much weaker.

The fact that R_{rel} passes the two conditions for a description of representativeness can be understood with some theoretical considerations. The standard error of the mean is defined by

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \tag{1}$$

where $\sigma_{\overline{x}}$, the standard deviation of a sample, can be given by the following equation (N being the number of samples):

$$\sigma_{\overline{x}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{x}_i - \mu)^2} \tag{2}$$

For $N = 1$, this gives:

$$\sigma_{\overline{x}} = |\overline{x}_i - \mu| \tag{3}$$

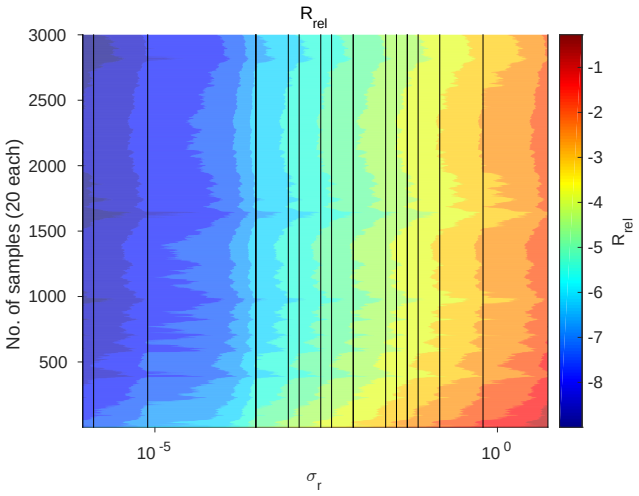

Figure 2. Like Figure 1, but for R_{rel}.

Plugging Eq. 3 into Eq. 1 gives:

$$\frac{|\overline{x}_i - \mu|}{\mu} = \frac{\sigma}{\mu\sqrt{n}} = \frac{10^{1/\tau^*}}{\sqrt{n}} \tag{4}$$

and therefore

$$R_{rel} = \log_{10}\left(\frac{|\overline{x} - \mu|}{\mu}\right) = -0.5\log_{10}(n) + \frac{1}{\tau^*} \tag{5}$$

So ideally, R_{rel} should depend inversely on τ^* and directly on the logarithm of the number of values. Figure 2 shows this is approximately true for RAND.

In the case of RAND, R_{rel} ~~can~~ and R_{var} can both be used to describe representativeness as ~~it passes~~ they pass the two conditions~~, while R_{var} does not~~. Theoretical considerations make the finding plausible for R_{rel}. RAND can be considered a theoretical abstraction of MOD. The finding here therefore strongly supports that of Sections 5.2 and 5.3, where R_{rel} and R_{var} have also been found to be good descriptors of representativeness when using $MOD_{~~CARIBIC~~}^{regular\ CARIBIC}$ and $MOD_{RANDPATH}$ or $MOD_{RANDLOC}$. In the main text, we use R_{rel} for final results, as it more suitable to answer the question of representativeness for a climatology.

3 Sensitivity study on the Kolmogorov-Smirnov test

When ~~using $MOD_{CARIBIC}$~~applying the Kolmogorov-Smirnov test to $MOD_{CARIBIC}^{regular}$, $MOD_{RANDPATH}$ or $MOD_{RANDLOC}$, ~~the Kolmogorov-Smirn~~ ~~test proved not usable, returning all~~ it returned almost only negative results. This indicates that $MOD_{~~CARIBIC~~}^{regular\ CARIBIC}$ is not representative of $MOD_{RANDPATH}$ in the definition of the Kolmogorov-Smirnov test. This behaviour was tested in a sensitivity study, the results of which are ~~described~~ discussed here.

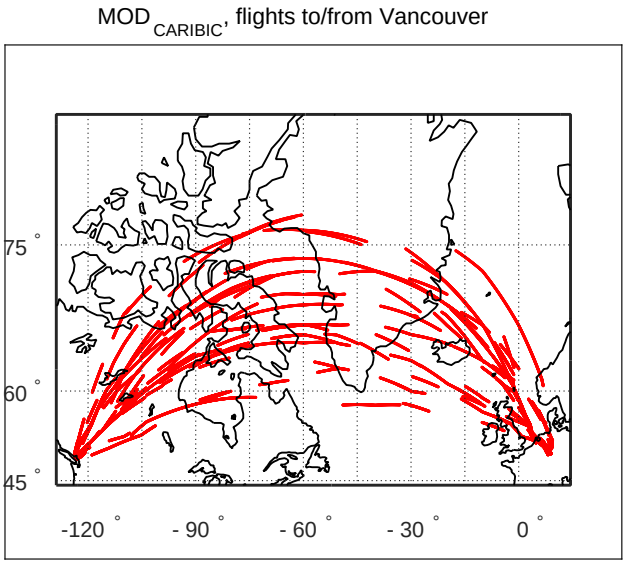

Figure 3. Flightroutes to Vancoucer, Canada, where each flight has been cut into 20 pieces and randomly chosen 30% of those pieces have been plotted. These are tested against the whole data from flights to Vancouver to give one point in Figure 4.

One of the most frequent destinations within the CARIBIC project is Vancouver, Canada (near 120 °W, 45 °N, see Figure 3), and only the subset of MOD~~CARIBIC~~ regular CARIBIC to this destination is considered in this example to minimize effects ~~stemming of~~ that may come from different flight routes. Parts of this reduced dataset were tested with the Kolmogorov-Smirnov test against the whole reduced dataset for all variables. ~~Data was not binned in months, including the whole distribution of datapoints in each height.~~ To produce these partial datasets, each flight was cut into an increasing number of pieces (corresponding to a certain time) and different percentages of these pieces were used in testing. Figure 3 ~~exemplifies this methodfor~~ shows an example of applying this method, by cutting each flight into 20 pieces and taking 30% of these by showing the corresponding flightpaths.

Data was not binned in months. When applying the Kolmogorov-Smirnov test without binning in months, the result is a profile in HrelTP for each variable. The result can then be diplayed in similar way to Figures ~~5 and 6.~~ 7 and 8. This matrix of height versus species was calculated for each combination of number of pieces and percent of pieces. In each combination, all the profiles of the different variables were averaged to end up with one value betwween 1 and 0 characterizing the result of the test for this combination of number of pieces and percent of pieces. The result can then give an impression of the strictness of the Kolmogorov-Smirnov test.

Figure 4 shows the result of the study. Independent of the number of pieces, the result is positive if all pieces are considered, as the definition of the test prescribes. But only when removing short pieces (shorter than 20 min) is the result also positive for less pieces, even though 70% percent of the data is still needed. When removing whole flights (at the top of the plot), more the 90% of the data has to be taken into account to achieve a positive result of the Kolmogorov-Smirnov test. This result is very

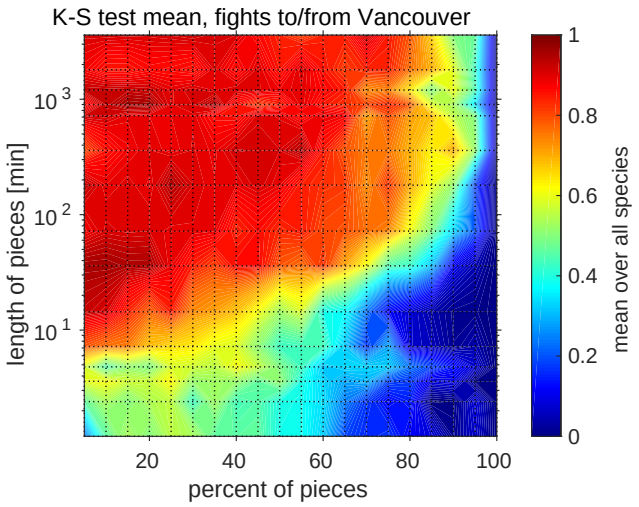

Figure 4. The Kolmogorov-Smirnov test applied to the flights to Vancouver, Canada, of $\mathrm{MOD}_{\mathrm{CARIBIC}}$ regular $\mathrm{CARIBIC}$ and subsets of these flights. Dotted lines indicate those lengths in time and those percentages that were tested. 0 stands for a passing the Kolmogorov-Smirnov test, 1 for not passing.

similar also for other error probabilities α, taking values of 0.001, 0.01, 0.05 (in the figure), 0.1 and 0.2. The area of failing increases only slightly with the error probability. This showcases the strictness of the test. The Kolmogorov-Smirnov test does not seem suitable to test a dataset measured with aircraft for representativeness of a larger dataset.

4 Aircraft tropopause pressure bias

By calculating $\mathrm{R_{rel}}$ using $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$, an important fact can be illustrated about data collected with instruments on civil aircraft. If data is resorted into heights relative to the tropopause (HrelTP), it still contains data taken at constant pressure altitudes in a limited range. Depending on the pressure at which the data was sampled, it contains information from different meteorological situations. The height of the tropopause relative to the sample pressure determines the range of values. The effect can be illustrated by calculating $\mathrm{R_{rel}}$ (see Equation 4) using $\mathrm{MOD_{RANDLOC}}$ and $\mathrm{MOD_{CARIBIC}}$ within $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$.

Figure **??** shows the results (right hand panel). For comparison, the left hand panel of Figure **??** shows $\mathrm{R_{rel}}$ of the same datasets when setting $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$, the range at which CARIBIC measures. On the right, the representativeness uncertainty increases strongly in all heights except just above the tropopause, where $\mathrm{MOD_{CARIBIC}}$ contains most data. Only the long lived species , and 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. $\mathrm{R_{rel}}$ calculated using $\mathrm{MOD_{CARIBIC}}$ and $\mathrm{MOD_{RANDLOC}}$ with the range of p set to $180\,\mathrm{hPa} < p < 280\,\mathrm{hPa}$ (left) and $10\,\mathrm{hPa} < p < 500\,\mathrm{hPa}$ (right).

The strong increase in representativeness uncertainty is due to the bias 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.

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 data analyzed in this paper can be accessed by signing the CARIBIC data protocol to be downloaded at .

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

References

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

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