We kindly thank the referees for their relevant comments, suggestions, and corrections.

The responses are organized as follows: the reviewer's comment is in blue, our answers are in black, and the changes proposed for the revised manuscript are in italics (black for modified sentences, grey for unchanged sentences that have been copy-pasted here to remind the context). The numbering of the pages and lines corresponds to the preprint, not to the modified document. Last, we answer "Done." to all the comments that suggest a modification, that we completely agree with, and that do not require any clarification in this document.

Before addressing the comments, please note that we found an inconsistency in the text, and corrected it as follows:

P31, L456: *"Another limitation of this approach is that the mean altitude of the measurements changes with the latitude: as the tropopause altitude decreases with the latitude, the subtropics are more sampled in the UT than in the LS, and reversely, the high latitudes are more sampled in the LS than the UT"*

This has been replaced by:
*"Another limitation of this approach is that **the tropopause altitude decreases with the latitude whereas the cruise altitude does not depend on latitude. Consequently,** the subtropics are more sampled in the UT than in the LS, and reversely, the high latitudes are more sampled in the LS than the UT."*

Comments from Referee 2

Cohen et al. processed long-term IAGOS measurements to generate climatologies of CO, O3, H2O, and NOy in the UTLS region, and then utilized these observations to evaluate five different chemical transport/chemistry-climate models. They also highlighted dynamic features observed in the IAGOS data, such as the seasonal shifts of the ITCZ above Africa. The manuscript is quite comprehensive, perhaps containing more results than can be effectively presented in a single paper. I hope the authors could address the following comments before being published in ACP.

Major comments: The manuscript is challenging to follow from beginning to end, as each section feels somewhat disconnected. For example, while the introduction emphasizes assessing the impact of aviation emissions, the paper itself does not present results specifically related to aviation emissions. Additionally, Sections 3.1 and 3.2 both discuss biases due to cross-tropopause transport, yet they are presented separately. Figures 10 and 11-13 seem redundant. The manuscript could be significantly improved by simplifying the introduction and focusing on presenting the most important results.

We kindly thank Referee 2 for the relevant suggestions, helpful for the clarity of the paper.

Figures 11–13 have been moved into the Appendix, but we keep their analysis in the manuscript. Their additional information is relevant, as they provide an understanding of the metrics shown in Figure 10, and meridional information as well.

According to Referee 1's comments too, several modifications were brought to the introduction. We agree that emphasizing the impact of aircraft NOx emissions can bring confusion as the scope is only to assess the models using the IAGOS database, but mentioning the ACACIA project and the estimation of the impact of aircraft NOx emissions remains important as it is the main motivation for this study. Thus we tried to make the introduction clearer about the aim of this study by merging the paragraph presenting ACACIA into another paragraph. Now the introduction is organized as follows:

- Importance of ozone, water vapour, CO, and NO_x in the UTLS for the climate system
- Why NO_x emissions are a factor controlling UTLS composition when they are emitted in the free troposphere
- Why simulating UTLS chemical composition is important to understand the impact of these high-altitude NOx emissions
- In this context, assessing the models in the UTLS is crucial
- For this purpose, the IAGOS database is well-suited
- For this use, the Interpol-IAGOS software is well-suited
- This study aims to assess 4 CCMs/CTM in the UTLS, for the ACACIA EU project

Other comments:

Line 40: How do you differentiate the impact of aviation NOx emissions from lightning NO emissions?

We generalized as *"free-tropospheric NOx emissions"* instead, thus including lightning, in order to avoid some confusion. Aviation emissions are still mentioned later in the introduction, but we hope we managed to rephrase it such that it does not seem to be the topic of this study anymore.

In the companion papers (but not in this study), the impact of aviation NOx emissions is investigated with a perturbation approach, i. e. using a couple of runs. One run includes aviation emissions, and the other one is made without it. Then we analyze the difference between the two runs.

Line 54-67: This paragraph describes the impact of individual processes on measured species but does not explain how the IAGOS dataset can assess the sensitivity of model responses to aircraft emissions.

We apologize for the confusion, we hope that our explanations (above) and clarification in the manuscript are sufficient to address this question.

Line 75: Do most measurements during the cruise occur in the LS? Are measurements in the UT primarily taken during departure and landing, potentially limiting the spatial representation of observed climatology in the UT?

In the extra-tropics, there are more measurements in the LS, especially in winter. However, the cruise measurements remain the first source of measurements in the UT. Precision has been added in the text below:

"The monitoring began in 1994 for ozone and H_2O, 1997 for NO_y, and 2001 for CO, with an abundant sampling in most of the northern extratropics (above and below the tropopause) and several tropical transects."

Table 1: It would be helpful to indicate whether these models simulate chemistry in the stratosphere.

Yes, it is a required condition for this paper. We added this information in the general comment P6, L121.

Line 175: What is meant by "online model" here? Does it imply no interaction between chemistry and meteorology?

We guess that Reviewer 2 meant "offline model". It does imply no interaction between chemistry and meteorology, or more exactly, no impact from chemistry on meteorology.

Line 283-293: How do model simulated tropopause heights compare to layers defined by PV fields? Could differences in tropopause height affect the comparison between IAGOS data and model results?

The PV fields are taken from the models' output. The model simulated tropopause and the layers are thus defined by the same data set. In order to avoid confusion, we clarified the text as follows:

"For each model, the tropopause is defined dynamically as the isosurface of 2 PVU (potential vorticity units) derived from the model output. The UT spreads from 400 hPa up to the tropopause level but excludes the top grid cell in order to avoid the strongest mixing zone, directly impacted by both layers (e.g. Thouret et al. (2006); Cohen et al. (2018)). The LS corresponds to all the sampled grid cells above the 3 PVU isosurface."

To the second question, we investigated the mean pressure difference between observations and modelled tropopause (2 PVU isosurface), represented in the new figures in the supplementary material (Figs. 1–5).

For a given layer, the pressure differences are relatively low across the models. Generally, it is not higher than 5 hPa, except in winter and spring in the LS, for NO_y and ozone, but it does not reach 10 hPa (~ 200 m). It could still be problematic for water vapour in the LMS as the vertical gradient from the tropopause is particularly high, including two orders of magnitude (Zahn et al., 2014), but there is no apparent link between Figs. 5–8 and Figs. S2–S5. For instance, ozone (Fig. 5) is higher in the UT with EMAC and lower with LMDZ-INCA, but the sampled grid cells are closer to the model tropopause (Fig. S2) for LMDZ-INCA. In the LS, the sampled grid cells are closer to the tropopause for EMAC and UKESM, and yet, they do not minimize ozone.

The comparison with Figs. S2–S5 is added to the manuscript as follows:

“This issue might be partly addressed by our methodology, notably the definition for the layers that enhances the isolation between them, and the exclusion of the grid cells with an inconsistent PV value regarding ozone observations. The mean pressure differences between observations and the model 2-PVU tropopause shown in Supplementary Material (Figs. S1--S5) do not exhibit noticeable differences between the models: mostly, they are less than 5 hPa (except in winter and spring for ozone and NO_y), and are always less than 10 hPa. For each species and layer, the distance between the sampled grid cells and the tropopause does not vary enough across the models to play a significant role in the inter-model discrepancies, as there is no apparent correlation with the chemical tracers.”

Line 327: In Section 3.1, the O3/CO ratio is used to indicate biases in cross-tropopause exchanges, while Section 3.2 attributes H2O variation to cross-tropopause mixing as well. These two sections could either be combined or require additional explanation to clarify the differences.

It can effectively seem redundant as it concerns the same thematic (cross-tropopause exchanges). The main difference is that the explanation suggested in Section 3.1 concerns biases in cross-tropopause exchanges, whereas the diagnostic for water vapour in Section 3.2 does not mention biases, only H_2O differences between the UT and the LS. It only explains the lower-stratospheric behaviour of water vapour with cross-tropopause exchanges. To clarify this distinction, we added this precision:

“The lower stratosphere shows a similar pattern, though the contrast between the summertime water vapour maximum and the rest of the year is more pronounced ***than in the UT****. This feature is consistent with the increased impact from the troposphere during this season, and the extremely steep vertical gradient in water vapour.”*

This way makes it clearer that it does not concern the comparison between observations and models.

Figure 11-13: where is the figure for the comparison of CO?

This figure has been restored. As the scatterplots have moved into the Appendix, it is now labeled as Fig. B4.