We are thankful to the reviewers for their positive and accurate feedback on our study, and for the improvements they allowed us to perform.
Before answering to the comments, we have to point out the changes made in Table 3, in Section 4.2, since the original manuscript version. They are due to a correction of an error in our method that was rightly spotted by Referee 1. More precisely, it concerns the additional mask based on the observed ozone values. Our mistake consisted in filtering out the individual measurement points if their corresponding ozone value was not consistent enough with the PV, instead of filtering out whole monthly grid cells. This correction results in a stronger loss of data and in the removal of several seasonal cycles from the figures, because they became incomplete. Nevertheless, it did not change the main conclusions.

Note also that since the submission of the GMDD version, we have decided to add the name of the methodology (Interpol-IAGOS) into the title of the revised version of the paper. There is also a change in the scatterplot in Fig. B3, at level 25 where the previous version showed two extreme outliers in observed CO (350 and 510 ppb) during summer, corresponding to two grid cells above Alaska. This removal has been explained in the figure legend too.

For the responses, the text from the reviewers is in blue. Our answers are in black, and the changes proposed for the revised manuscript are in italic (black for modified sentences, grey for unchanged sentences that have been pasted here in order to remind the context). In the revised manuscript, the main changes are shown in blue.

The paper is relatively well written, though I have noted a few instances where I had difficulty understanding the discussion. In general my comments are minor, detailed below, but I would like to raise one point that I found missing from the discussion of the results. A much better way to compare the models with the IAGOS observations would be to have highly time resolved instantaneous model outputs that are more directly comparable with the in-situ observations. This comparison would not be perfect because of errors in the model that arise from a whole host of different reasons that have been discussed at length in the literature. But, by and large, and particularly so for multi-model intercomparison projects, the large data volumes required to save high frequency output makes that type of comparison difficult and we must work with monthly average data. In addition to the usual list of reasons for model biases, now there is the added factor of having averaged the model fields in time. The authors have done a great job of exploring this effect with the observations by comparing IAGOS-HR with IAGOS-DM, but the discussion of the results does not include much mention of the effect of time averaging. It became a bit confusing when the discussion of the comparison of IAGOS-HR and IAGOS-DM zeroed in on the effect of mis-classification of points in either the UT or LS (see the comment on Page 19, Lines 4 – 6) and ignored the effect of time averaging. In the vicinity of the tropopause, over the course of a month, a particular model grid point will sometimes be in the stratosphere and sometime in the troposphere so the monthly average will reflect both of these influences. At least to my mind, this effect is an important part of the problem comparing monthly-average fields with in-situ observations but I do not find much discussion of this facet of the problem in the manuscript.

We fully agree that this explanation is clearer than the one we used in the paper. Actually, these explanations are closely linked (we would even say equivalent), but maybe our meaning was not clear. What we meant was that with a monthly average, for instance in the UT, we include grid points from undesirable stratospheric air masses. This leads to a mis-classification of a non-negligible part of the individual measurements. We changed the sentence, as below:

In other words, the effect of time averaging leads to a mis-classification of a non-negligible part of the individual measurements. For a given layer, it introduces a bias due to unexpected mixing with another layer.

And in the conclusion too:
The use of a monthly mean PV field and the ~800 m vertical resolution in the UTLS of MOCAGE onto which IAGOS observations are projected automatically result in an artificial increase of stratosphere-troposphere exchange. It is explained by the fact that the grid cells in the vicinity of the tropopause are crossed by both tospheric and stratospheric air masses in the course of a month. It results in a decreased vertical gradient between UT and LS. Nevertheless, the seasonal maxima and minima become less clear but remain visible in IAGOS-DM with respect to IAGOS-HR.

Minor comments:

Page 3, Line 28: When discussing the available frequency of model output to compare against the IAGOS data, the manuscript states 'the 3D outputs from the REF-C1SD simulations, which are monthly averages.' The fact that monthly average fields are the most commonly available output is not necessarily part of the REF-C1SD specification, but more of a result of asking for a large amount of data from a number of models participating in a multi-model intercomparison. The text here should be more general and refer to the fact that monthly average fields are the most commonly available type from multi-model intercomparisons.

Note that Reviewer 1 made the same remark. The sentence has been modified, as shown below:

However, these comparisons were led using frequent simulation outputs. Although the high frequency is necessary for their approach to separate accurately the air masses into different categories, it is not adapted to the assessment of monthly averaged fields used in multi-model intercomparisons.

Page 4, Lines 7 – 10: In discussing other available methods for comparing in situ measurements with model data, the authors state that interpolating neighbouring measurement points onto each grid point would be computationally expensive for the IAGOS data because it requires keeping track of a large number of measurement locations each month. But when the methodology is described in Section 3, it would seem to me that the proposed method also requires keeping track of the discrete locations of a large number of observations. I believe I see the

idea the authors are trying to convey– that a month of observation data at the original locations must be collected up before interpolating on to the grid point - but I would suggest rewording this point.

We agree with your comment. It was not clear. In contrast to our method, the method that we were citing here (notably in New et al., 2000) calculates linear combinations between measurement points, and thus require to keep track of a large amount of observations simultaneously, whereas for a given variable, our method requires to store only 2 quantities per grid cell: the sum of all the weights and the weighted sum of the current variable (ozone or CO). We brought precision to the text (as pasted below) to make it clearer:

Some of them consist in calculating ***a linear combination*** *from the neighbouring measurements points onto each gridpoint (e.g. New et al., 2000).* ***However, it requires to store simultaneously the information of all the measurement locations, and during a whole month.*** *It is thus convenient for measurements with regular locations as surface stations, whereas their use on the IAGOS database would be expensive computationally as well.*

From now on, the recommendations let without a response in this document have all been included in the manuscript.

Page 7, Line 4: The phrase 'measured a mixing ratio Cobs(X) for an X species' might be better as 'measured a mixing ratio Cobs(X) for species X'

Page 8, Line 3: Where it is written 'increasing linearly with the distance between the measurement point and the (i, j, k) grid point.', I think should be 'decreasing linearly with the distance'.

We really meant that the alpha, beta and gamma coefficients increased with the distance with the (i,j,k) gridpoint. In contrast, the corresponding weight defined by the product (1-alpha)*(1-beta)*(1-gamma) decreases with the distance. We have changed the wording to make it clearer:

[…] a normalized ***scalar*** *is then computed for each dimension (coefficients alpha, beta, gamma), increasing linearly with the distance between the measurement point and the (i, j, k) grid point.*

This change was also done in the legend of the corresponding figure.

Page 11, Line 1: I find the phrase 'before deriving monthly means in the two layers' a bit confusing because I am not familiar with the analysis performed in Cohen et al. (2018) and the earlier discussion of model layers. I assume the two layers are the UT and LS and the analysis is done separately for each region?

Exactly. This part of the text has been reworded. We added the word “separately” in the sentence below to make it explicit:
A second part of this assessment targets the behaviour of the model in the UT and the LS separately.

Page 11, Line 23: might be missing a word (such?) in ‘account, [such] as Western North America and Siberia.’

Page 14, Lines 13-14: ‘With respect to the 1:1 line, levels 25 and 26 are characterized by an overestimation of the lower part of the O3 distribution (< 120 ppb) and by an underestimation of the higher part.’ This sentence is referring to the pronounced underestimation of ozone between 150 ppb and 250 ppb shown in Figure 5 for P∼255 and P∼285 hPa, and from Figure 3 this looks to be related to an underestimation of ozone at high latitudes. Do the authors have any ideas why ozone in this one region seems to be underestimated to such a large degree? I will note that from the figures in the Appendix the underestimation does seem to be most extreme in the summer months.

These biases have been investigated, some reaction constants have been updated and for long CCMI-type simulations, in particular the CCMI phase-2, MOCAGE is now run in 60 levels (up to 0.1 hPa), giving a better representation of ozone at global scale.

Precision has been added into this paragraph.
[…] levels 25 and 26 are characterized by an overestimation of the lower part of the O3 distribution (< 120 ppb) and by an underestimation of the higher part, more pronounced during boreal summer according to Fig. B3 in Appendix.
A possible reason is that the summertime tropopause altitude in these regions can be overestimated by the model, or that the vertical stability is underestimated. These biases have been largely improved with the more recent version of MOCAGE used to run CCMI phase-2 simulations.

Page 18, Lines 17 – 18: In referring to the IAGOS-DM data separated into the UT and LS regions (‘In contrast, IAGOS-DM refers to the new product presented in this paper, i.e. the IAGOS data distributed on the model’s grid, directly comparable to the simulation.’) it might be helpful to the reader to remind them that the data is assigned into either the UT and LS based on the monthly average PV at each model grid point.

We replaced the expression “directly comparable to the simulation” by the more precise formulation you proposed.

Page 19, Lines 4 – 6: Here it is stated ‘In other words, by using the monthly mean PV from the simulation, some of the IAGOS measurement points may be attributed to the LS while being in the UT (or in the tropopause layer) and vice-versa.’. Here it

sounds as though for the IAGOS-DM dataset individual IAGOS observations are being assigned to either the UT or LS based on their position relative to the monthly average tropopause. I had thought the IAGOS-DM was constructed and then the individual monthly-average values on model grid points were assigned to either the UT or LS (Page 11, Lines 4 – 8)? In this case, a particular monthly average value at a particular model grid point may be affected by a mixture air from both the UT and LS. This is not necessarily a bad thing, as the model monthly averages that are being analyzed will have a similar problem.

We fully agree with this comment. As indicated in our response to Reviewer 1, a mistake has been corrected in our data processing. The clarification on the mixing between UT and LS air masses due to the PV monthly averages has already been added, in answer to a previous comment.

Page 19, Figure 7. There are a lot of lines on each panel of Figure 7 (likewise for Figure 8) and making a clear comparison between IAGOS-HR, IAGOS-DM and MOCAGE-M is not easy. I am sympathetic to the need to condense graphics to avoid figures with 20 panels, but I was wondering if the authors would consider adding figures to the appendix that directly compare the three datasets for each of the eight regions? For each region there would be one panel showing the annual cycle from IAGOS-HR, IAGOS-DM and MOCAGE-M for the LS and a second panel for the UT.

We agree with this suggestion. A whole set of new graphics has been added to the Appendix. They are mentioned in the text in Sect. 4.2, second paragraph:

The comparison between the two IAGOS products in matter of seasonal cycles is proposed in Figs. 7 and 8, respectively for O_3 and CO. They are shown with their corresponding interannual variability (IAV), defined as a year-to-year standard deviation. ***For complementary information, a more exhaustive representation is proposed in Figs. C1 and C2 in the Appendix showing the results with each region in a distinct panel.*** *In Fig. 7, both IAGOS versions show a summertime O_3 maximum […].*