We are thankful to the modifications proposed by Reviewer 2. We addressed the comments in the same way as in the first authors' response.

Review of revised version of Cohen et al. (2021), Interpol-IAGOS: a new method...

The author response to the comment from Anonymous Referee #1 regarding the use of Specified Dynamics (SD) simulations ('... the problem with the SD-simulations could be that certain nudging methodologies may lead to introducing noise (and hence too much mixing) especially visible around the tropopause (see discussion in Orbe et al. 2020).') misses the mark a bit. I agree that MOCAGE, being a CTM, would not suffer from problems due to nudging of dynamical variables, however the problems introduced by nudging in a SD simulation should be seen as a caveat to the results found comparing SD simulations and not as a possible application. As argued by Orbe et al (2020), the dynamical noise and resulting loss of consistency between tracers and dynamical variables is a problem with SD simulations and should be pointed out here as a possible aggravating factor. The text inserted at Page 27, Lines 9 – 10 does not adequately bring this forward. It is an important point since the use of SD simulations (or a CTM using reanalysis) is an important component of the current comparison with IAGOS observations as one would expect any significant biases in the position of the tropopause in the model would be reduced making the comparison with IAGOS-DM more straightforward.

We corrected this formulation, as follows:
To a greater extent, it can be used on a wide range of long-term simulations including both CCMs nudged and free runs in order to perform climatological comparisons. Precaution must be taken while extending this work to the specified-dynamics simulations from CCMs, regarding the loss of consistency between chemical and dynamical variables that is introduced by nudging, as highlighted in Orbe et al. (2020). Notably, inconsistencies between ozone and potential vorticity are likely to introduce noise in the simulated upper-tropospheric and the lower-stratospheric behaviours.

Page 4, Lines 21 – 23: There is revised text here that reads 'To compare the REF-C1SD simulations against IAGOS data, interpolating the simulation outputs onto the high-resolution observations would be expensive computationally, and not required because our study is not focused on processes but on climatologies.' and was added to address the comment of Referee #1 on the original text at P4L3, that the interpolation of the model outputs on to the IAGOS observations is 'not possible'. There is new text inserted at Page 4, Lines 16 – 18 that much better addresses the substance of the original comment by Reviewer #1. And even with the focus on climatologies ('... our study is not focused on processes but on climatologies.') would it not be advantageous to have high frequency model outputs that could be interpolated on to the IAGOS observations to construct a climatology that is more directly comparable with the climatology derived from IAGOS? It is difficult to argue against the view that the correct way to perform the comparison would be to have high frequency model outputs and, while computationally expensive, it is nowhere near as computationally expensive as the original model simulations. But for now we do not have high frequency model outputs from multi-model intercomparisons such as CCMI. I would suggest the authors revise the text at Page 4, Lines 21 – 23.

Thanks for this suggestion. We substituted Referee 2's argument to the ones we were using. The new sentence is:
To compare the REF-C1SD simulations against IAGOS data, interpolating the simulation outputs onto the high-resolution observations would be the most accurate way, but high-frequency outputs from multi-model intercomparisons such as CCMI are not available yet.

Page 10, Lines 6 – 12: Here the authors have added a brief description of the comparison of sampling MOCAGE daily output with the monthly sampling in response to the second half of the comment of Referee #1 to P4L3 that begins ‘In fact, it would be important to prove for a methodology paper as you have presented here that your claim of the gridded IAGOS data being representative of the monthly mean is true.’ In the response to the comment, the authors did perform a new analysis that is described in the response to Reviewer #1 using MOCAGE-M-day. There are some statistics from this comparison in the response to Reviewer #1, but the authors do not include any quantitative results from the comparison to MOCAGE-M-day in the article – as it is, the authors state ‘MOCAGE-M monthly means could be considered as representative of the month’ with no supporting information. In particular, I am not too worried about the sampling in locations with a high frequency of aircraft sampling (the North Atlantic flight corridor, for example) but some mention of the magnitude of the larger differences would be instructive.

We agree with this comment, the mentioned sentence in the paper should rely on some values describing the distribution of the relative biases between MOCAGE-M-day and MOCAGE-M monthly means. Below is a table showing three percentiles for each species and season, characterizing the relative bias, in absolute values (the same as involved in the FGE equation). We chose to show the percentiles 90, 95 and 99 to the reviewers in order to give more details, but for a better clarity, the manuscript does not include the percentile 95.

Percentiles		90	95	99
O3	DJF	0.104	0.136	0.228
	MAM	0.098	0.133	0.222
	JJA	0.064	0.087	0.147
	SON	0.081	0.105	0.167
	ANN	0.060	0.080	0.131
CO	DJF	0.079	0.109	0.181
	MAM	0.064	0.088	0.158
	JJA	0.049	0.067	0.130
	SON	0.063	0.085	0.144
	ANN	0.041	0.055	0.103

The 90 percentile of the bias spreads from 4.1 to 10.4 %. The 95 percentile is generally lesser than 10 % also, except during boreal winter and spring when the ozone bias reaches until 13.6 %. Concerning the extremely high values, the annual 99 percentile equals 13.1 % for ozone and 10.3 % for CO. For the seasonal biases, the 99 percentile spreads from 14.7 up to 22.8 % for ozone, and from 13.0 up to 18.1 % for CO.

We also calculated PDFs for seasonal cycles, as in the four histograms shown below:

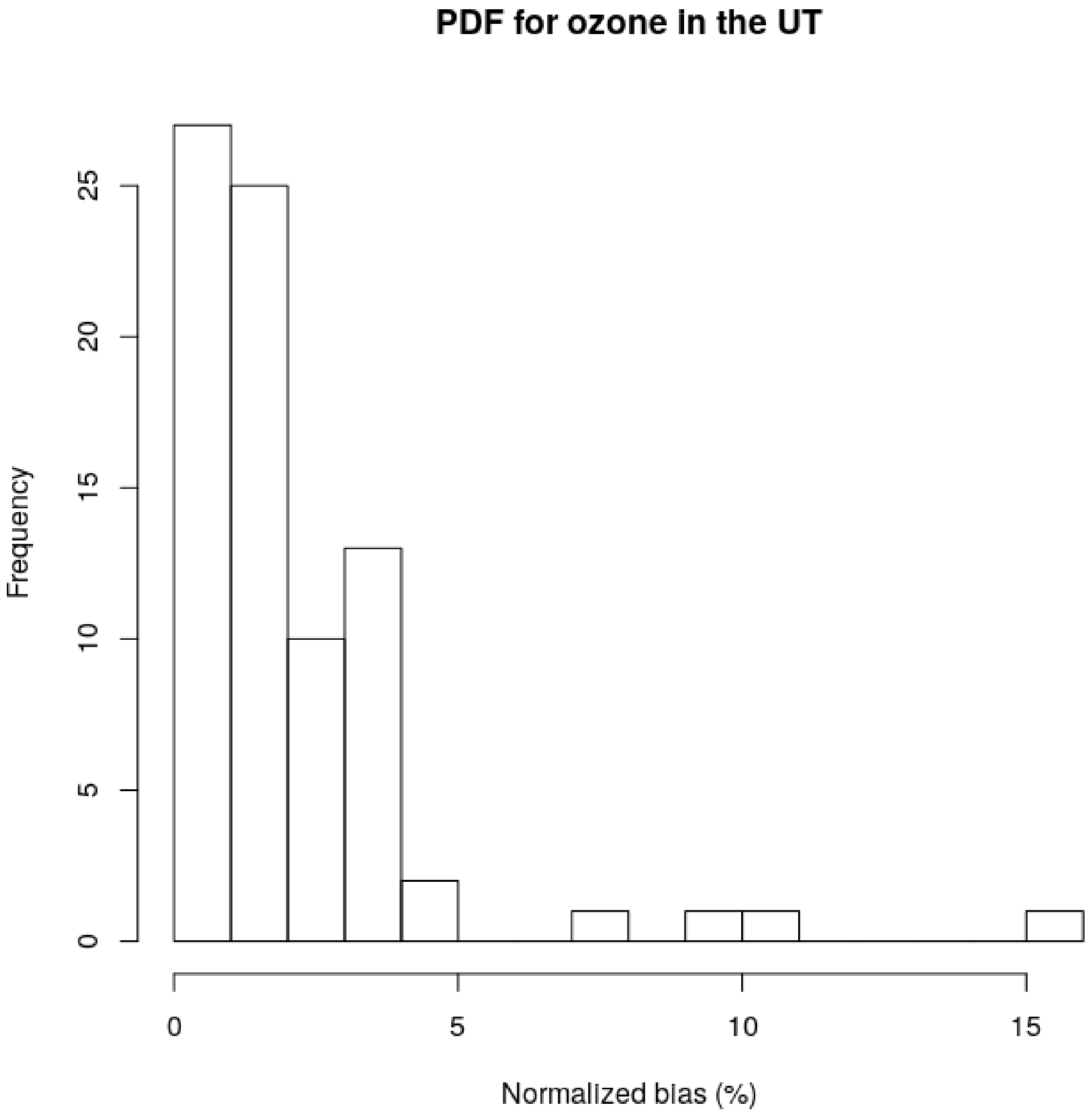

Figure AR1: Histogram showing the PDF of the normalized biases for ozone in the UT, the sample being the relative biases between the seasonal cycles from MOCAGE-M-day and MOCAGE-M during 2003 - 2007, concatenated across the regions. The sample is made of 81 values.

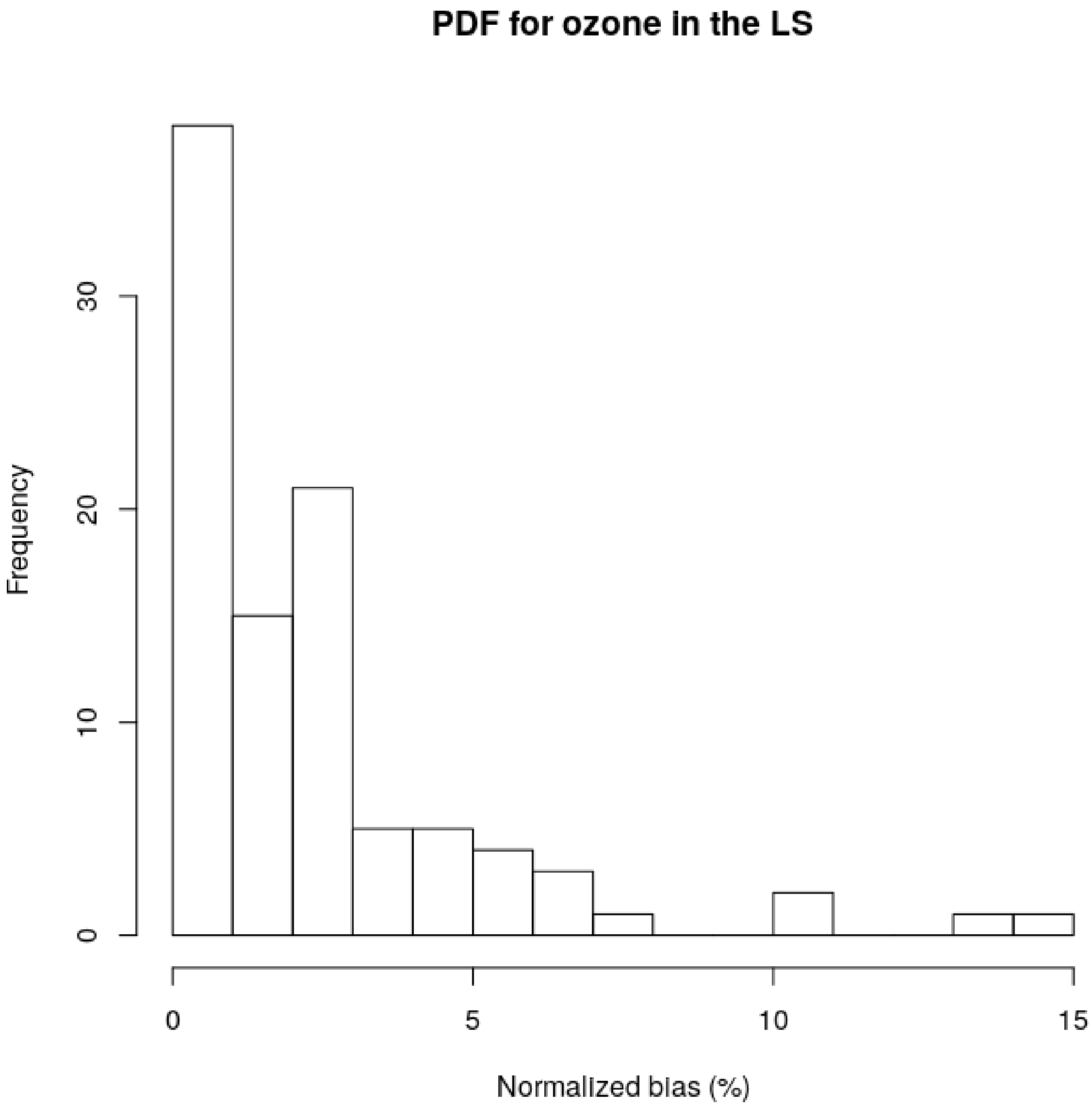

Figure AR2: Same as Fig. AR1 for ozone in the LS. The sample is made of 96 values.

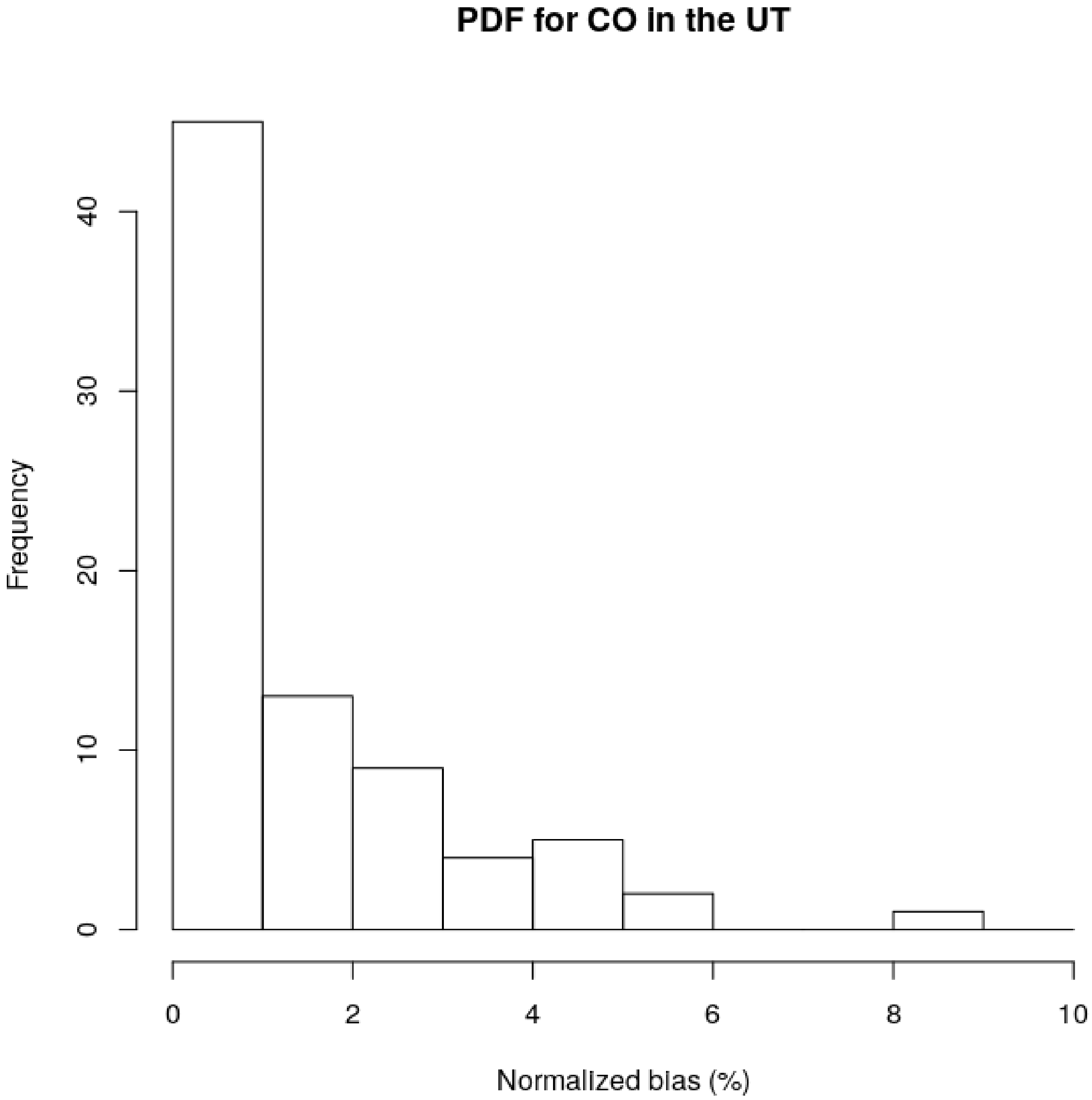

Figure AR3: Same as Fig. AR1 for CO in the UT. The sample is made of 79 values.

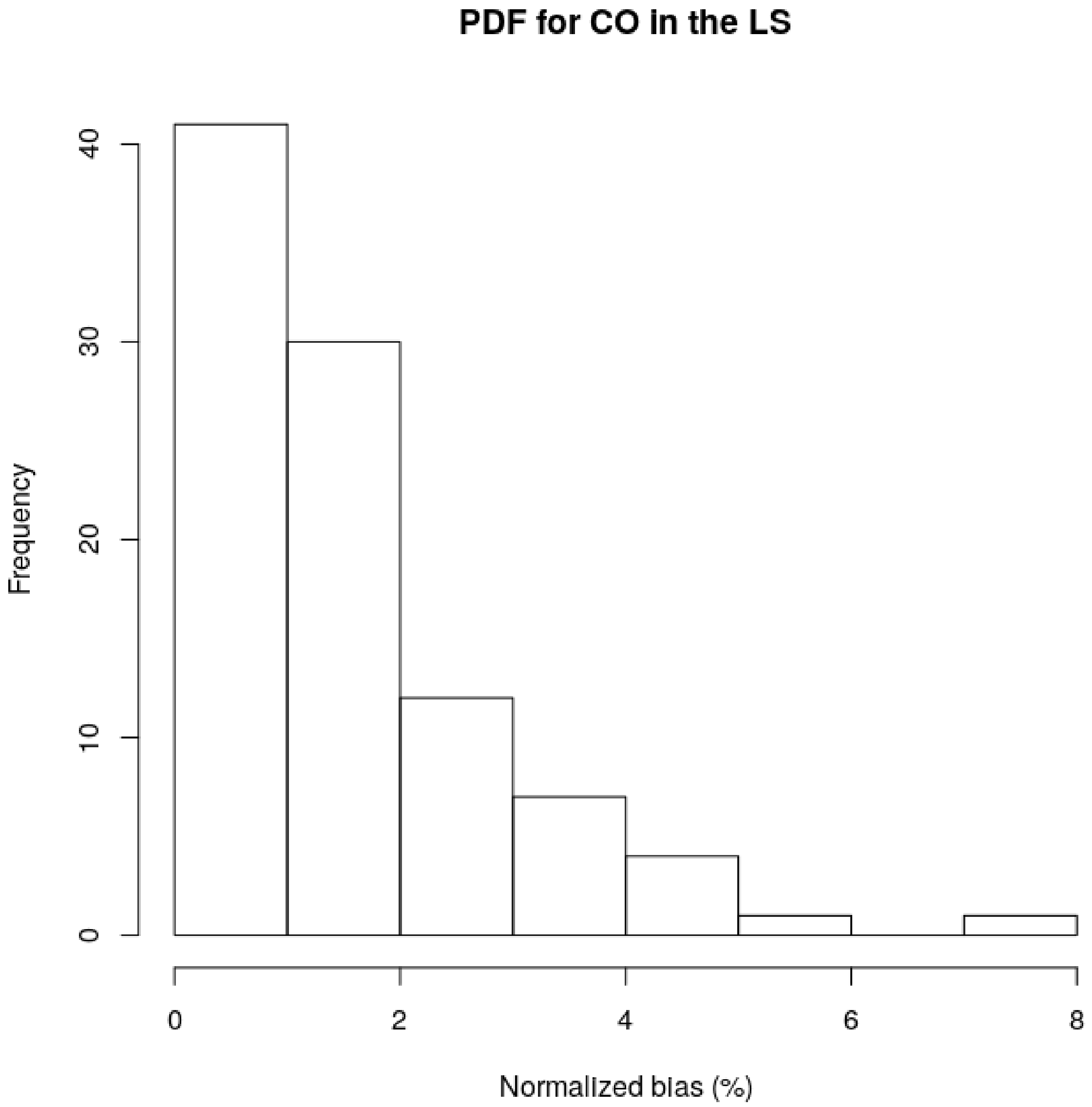

Figure AR4: Same as Fig. AR1 for CO in the LS. The sample is made of 96 values.

These histograms show that the quasi-totality of the biases are below 5 %. Ozone biases outreach 10 % twice in the UT (-10.7 % in Northwest America, in October; 15.2 % in Middle East, in February) and four times in the LS (-10.9 % in Northwest America, in October; -13.2 % in North Atlantic, in November; -14.1 and 10.4 % in the West Mediterranean basin, in January and September respectively), the maximum value being ~ 15 %.

The changes in the paper are as follows:

- At the end of Section 3.2:

The latter point has been tested (not shown) on a 5-year subsample using the simulation daily outputs instead of monthly outputs. It consisted of comparing MOCAGE-M to a test product derived by calculating monthly averages from the daily outputs and applying a mask based on the IAGOS daily sampling. The results from this test are briefly presented in Sect. 4.

- In the results section, a short subsection has been added:

4.1 Monthly representativeness

A first step in the assessment of the methodology consists of testing the monthly representativeness of the IAGOS-DM mean values, in order to evaluate the temporal consistency between IAGOS-DM and MOCAGE-M. For this purpose, as mentioned at the end of Section 3.2, we compared MOCAGE-M to a test product derived by calculating monthly averages from the simulation daily outputs, after applying a mask based on the IAGOS daily sampling. For this test, the chosen period spreads from 2003 until 2007 included, an uninterrupted measurement period for both ozone and CO. Concerning the mean 3D distributions, a mean normalized difference between the two products has been found below 1.7 % for each season and each species. In absolute values, 10 % of the yearly mean biases are greater than 6.0 % (4.1 %) for ozone (CO), and 1 % greater than 13.1 % (10.3 %). Seasonal mean biases are characterized by a 90 percentile generally lower than 10 %, and a 99 percentile from 14.7 up to 22.8 % for ozone and from 13.0 up to 18.1 % for CO. The maximum values correspond to winter and spring. Concerning the seasonal cycles, the relative difference between the two MOCAGE products was found to be almost systematically below 5 %, and amongst all the regions, its ozone values seldom outreach 10 %, with a maximum value at 15.2 %. In conclusion of this comparison, the similar results obtained between MOCAGE-M and the test product suggested that in most cases, the IAGOS-DM monthly means could be considered as representative of the month.

On the reply of the authors to the comment 'On a similar note, it would also be good to see what the benefit of the weighted gridding versus a gridding of the observations without the weighting function would be...' To estimate the benefit I would think you would have the weighted and unweighted gridding and compare these two products to some estimate of the truth. Here, we have the original aircraft IAGOS observations, a weighted gridding of the IAGOS observations, an unweighted gridding of the IAGOS observations, and the model output. One could show differences between the weighted and unweighted gridding, but how does one show the benefits of weighted gridding? To put it another way, because the model differences are smaller for weighted or unweighted gridding, does that mean one is more correct? I suggest the authors be careful about stating they have analyzed the benefits of weighting. In particular, I am not convinced the authors have shown that 'using a weighting function is a necessary step for a more accurate assessment' as stated in the abstract of the revised version at Page 2, Lines 8 – 11. Given the difficulty of showing a benefit to weighting, I would suggest the authors revise the text to be clear they are showing 'differences' and not 'benefits'.

It is true that we did not choose our words correctly for this test (MOCAGE-M vs IAGOS-DM against MOCAGE-M-noW vs IAGOS-DM-noW), since we cannot show that the results are "better" although the scores are enhanced.

In the abstract:
Along this model evaluation, we also assess the differences caused by the use of a weighting function in the method when projecting the IAGOS data onto the model grid compared to the scores derived in a simplified way. We conclude that the data projection onto the model's grid allows to filter out biases arising from either spatial or temporal resolution, and the use of a weighting function yields different results, here by enhancing the assessment scores. Beyond the MOCAGE REF-C1SD evaluation presented in this paper, the method could be used by CCMI models for individual assessments in the UTLS and for model intercomparisons with respect to the IAGOS data set.

And in the last sentence in Section 4.1:

The general improvement of normalised biases, normalised errors and spatial correlations, compared to a simplified gridding method, suggests that the use of a weighting function in our methodology can significantly enhance the model assessment.

Page 20, Lines 24 – 27: The text at this point was added in response to a comment from Reviewer 2 about the lack of directly addressing the effects of time averaging for the comparison with IAGOS data around the tropopause ('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.'). The text in this section still discusses the inability to correctly classify air masses using monthly average PV, but the problem is much more fundamental than. There is no clean separation of the stratosphere and troposphere left in the IAGOS-DM data after it is monthly averaged on constant pressure surfaces. By using monthly averages, of either the IAGOS data on a particular pressure level (IAGOS-DM) or model data, the sharp separation of tropospheric and stratospheric air around the tropopause, which would be preserved in the IAGOS-HR data, is lost. This will be a fundamental problem with analysing monthly average data in the vicinity of the tropopause and is illustrated to some extent by the differences between IAGOS-HR and IAGOS-DM. This is the point that is still missing in the text and, I think, it is an important one because it illustrates that treating the IAGOS-DM data in the way it is treated does make it more like the monthly average model data so the comparison of IAGOS-DM and the model is more valid. But because it is monthly averaged data the sharpness of the separation between tropospheric and stratospheric air masses is lost to some extent.

As written in the first authors' response, we agree with this explanation. We thought our previous answer to this comment was sufficient, as we consider that the misclassification of individual measurement points and the loss of sharpness in the tropopause are the same direct consequence of the loss of resolution caused by time averaging, because both generate a mixing between the UT and the LS. In other words, we thought we were telling the same thing, although with a different view. In order to make it explicit, now we have modified the text as follows:
In other words, the effect of time averaging leads to a loss of tropopause sharpness, thus resulting in 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.

A couple of minor comments:

Page 6, Line 12: 'expanding from 1980 to 2010' should be 'extending from 1980 to 2010'
Thanks for this correction, the change has been made.

Page 10, Lines 27 – 28: The new text 'less measurements are needed to characterize the climatologies' in reference to the CO observations implies that the CO climatology is somehow equally well characterized as the ozone climatology using less data. It would seem that the reason 60 CO observations are judged sufficient is that the sampling period for CO is shorter and to use the same Nthres=100 as for ozone means throwing out too much data. It is not that less measurements are needed to characterize the the CO climatology, so the wording should be modified here.

We agree with this comment. We reworded the text, as quoted below:
Accounting for the shorter CO measurement period compared to O_3 (~60 % of the O_3 period), the same threshold applied to the CO climatologies would result in a greater proportion of filtered-out grid cells. Thus, the corresponding N_{thres} threshold for this species is derived by applying a factor 0.6, leading to 60.