Answers to the referee's comments

We would like to thank the referee for her/his review of our paper and for giving us the opportunity to improve it.

The referee's question is copied in italic and the answer is written in normal font.

We have added 'review' in the legend of the figures shown in this document to distinguish from the figures of the manuscript.

We copy the modified part of the paper here in the response. The added modifications are written in red.

We mean by 'old version' the previous revision of the manuscript and 'new version' the revised paper.

General comments:

1. *I previously inquired about the final condition number of the R matrix and in response learned that it is around 8E6. I understand that the authors took care to ensure that the inverse of R was properly computed. However, with such a large condition number I worry that the IASI observations could be strongly down-weighted, and I wonder how the cost function in RfullExp compares to that in RdiagExp. Is it possible to include a plot of the cost function versus iterations for a cycle of RfullExp and RdiagExp, either in the response or in the revised article?*

Indeed, the condition number is large. To compare the behavior of the cost function in RfullExp to that in RdiagExp, we present in Figure 1_review three assimilation cycles picked arbitrarily after 15 days of assimilation for both experiments: 15/07/2010 12-13h UTC, 20/07/2010 05-06h UTC, and 27/07/2010 09-10h UTC.

The six cycles (three for each experiment) are shown on separate plots because the absolute values of the cost function are not comparable among cycles and experiments. We selected three dates and different times to show that the behavior of the minimizer's iterations is somehow systematic and not scene-dependent. The plots are shown only here and not reported in the revised paper for conciseness. However, the revised article was modified to partly include the following discussion.

Figure 1_review and Figure 2_review show the cost function versus the number of iterations for the 3 cycles, for RdiagExp and for RfullExp respectively.

In the RfullExp case, the minimizer converges after almost 90 iterations (89 iterations in average over the entire month), whereas it exceeds the maximum threshold (150 iterations) in the case of RdiagExp. The two convergence criteria used in the LBFGS minimizer are based on the reduction of the cost function and of the norm of the gradient to values below typically small thresholds (1.e-9 for the accuracy of the reduction of the cost function between to iterations and 1.e-3 for the gradient). We remind that the limit of 150 iterations was set to save computational time. Hence, within the RdiagExp the minimization does not achieve a full convergence. However, the further reduction of the cost function during the final iterations is quite small compared to the overall reduction. As a consequence, letting the minimizer reach the full convergence (after about 200 iterations) does not affect the O3 analysis significantly (not shown). For the RfullExp, the convergence is achieved due to the stationarity of the cost function (1st criterion). The fact that the observations are downweighed in RfullExp is likely the reason for the faster convergence.

<u>This part L13 to L15 P13 of the old version of the paper:</u>

"In fact, the introduction of the estimated R reduces the number of iterations from 150 (a fixed value to stop iterations if the convergence criteria were not achieved to save computational time) to 90 iterations in average. This means that the CPU time is reduced by more than 150% for each assimilation cycle."

<u>is replaced by:</u>

"In fact, the introduction of the estimated R reduces the number of iterations from 150 (a fixed value to stop iterations if the convergence criteria were not attained to save computational time) to 89 iterations in average. This means that the CPU time is reduced by more than 150% for each assimilation cycle. The convergence criteria of the LBFGS algorithm is based on either the reduction of the cost function or the norm of its gradient below some given small thresholds. For the RfullExp, the convergence is achieved due to the stationarity of the cost function (1st criterion). The widespread correlations (high condition number) and larger variance of the estimated R matrix conduct to a downweight of the observations and are likely the reason for the improved convergence in RfullExp." <u>in the new version of the paper L1 to L5 P14.</u>

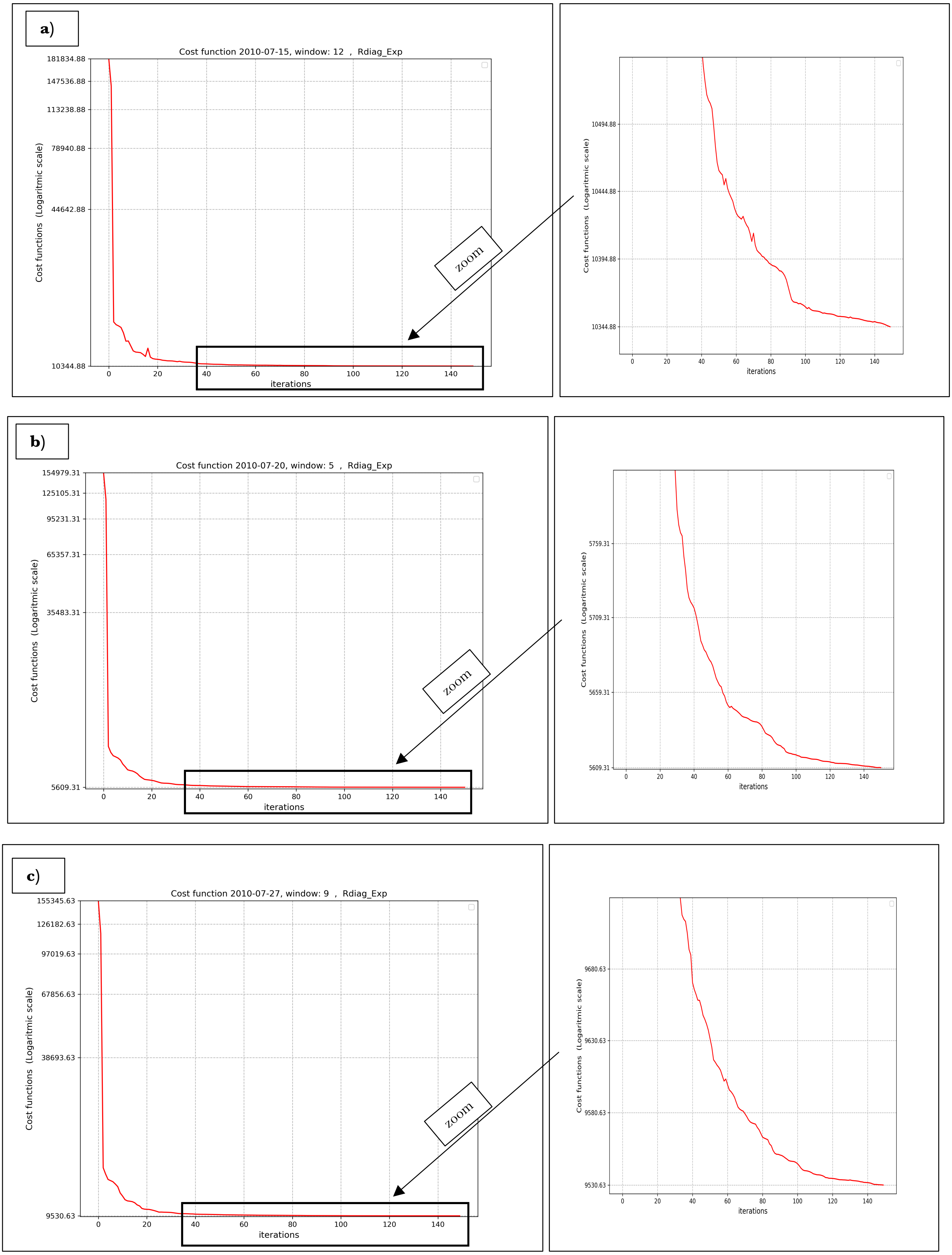

Figure 1_review: the cost function versus iterations for 15/07/2010 12-13h UTC (a), 20/07/2010 05-06h UTC (b), and 27/07/2010 09-10h UTC (c) for RdiagExp.

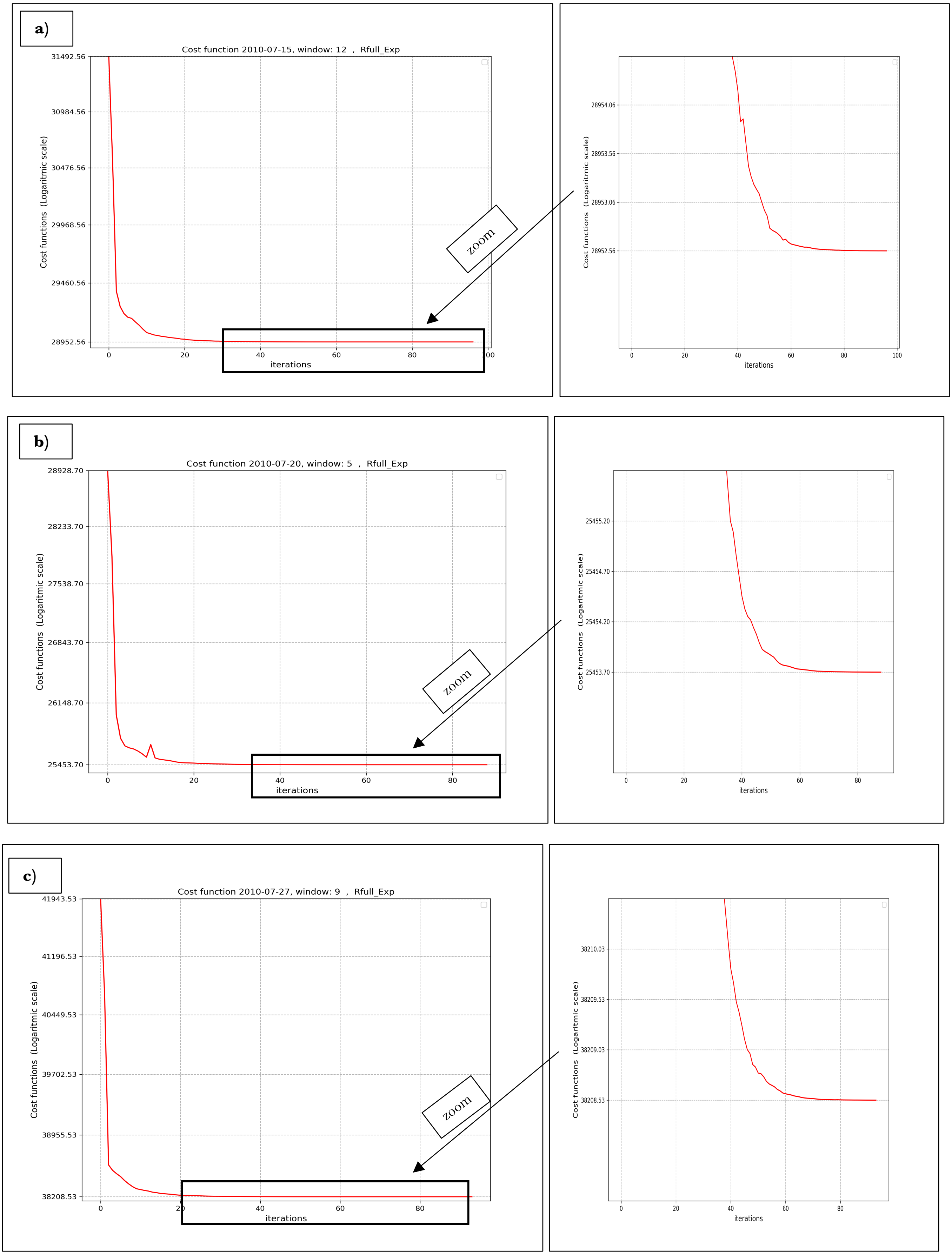

Figure 2_review: the cost function versus iterations for 15/07/2010 12-13h UTC (a), 20/07/2010 05-06h UTC (b), and 27/07/2010 09-10h UTC (c) for RfullExp.

2. *I do not feel that an adequate response was given. To claim that results are "significant" requires a statistical analysis, for example, computing a confidence interval around the difference between the zonal averages of two experiments.*

Indeed, we used the word 'significant' in the paper as well as in the previous revision. With 'significant' we meant that accounting for a more realistic observation-error estimation brought a 'remarkable' improvement in terms of results. Obtaining remarkable improvement in the validation against three independents observation networks (OMI, MLS and Ozonesondes) made us conclude that the results were 'significant'. However, the referee is right. To claim the differences between the two experiments are significant needs a statistical analysis. We present below (Figure 3_review) a t-test to evaluate the statistical significance of the differences between the two experiments in terms of the zonal averages reported in Figure 3 of the paper.

Figure 3_review reports results of Student's t-test comparing the zonal averages of the two experiments (RfullExp and RdiagExp). In fact, the zonal averages (whose differences are shown in Figure 3 of the paper) are obtained by averaging the analysis over the month of the study and over longitudes (allowing us to have a sample of size of 24(hours)x30(days)x180(longitudes)). We have used the standard deviation computed for each average to perform our test. Regions with green color reports the null hypothesis H_0 (the two experiments are not significantly different in terms of zonal averages) and red color report the alternative hypothesis H_1 (the results are significantly different in terms of zonal averages). The results shown below are obtained at 0.05 level of significance. We notice that the majority of regions report significant differences. Moreover, the regions where the differences are large in Figure 3 of the paper (between 300 hPa and 10 hPa) are statistically significant as it is shown in Figure 3_review.

Another test of the significance of the differences of the analyses with respect to the MLS and ozonesoundings measurements is reported in the question 5 of the specific comments.

We would like also to remind that our main objective was to assess the update of observation-error covariances on the assimilation results. For this, we kept the same period (one month) and system configuration already discussed in the literature (Emili et, al. 2019). Nevertheless, accounting for a long period is important to assess the robustness of the approach for a potential operational implementation. Emili et, al. 2020 have used an estimated R-matrix (as in our paper) to assess the impact of IASI measurements on global ozone reanalyses for a duration of one year (manuscript already submitted to Geoscientific Model Development). The results were similar in terms of the covariance estimation (strong correlations) and on the impact on the assimilation results (improvement of the reanalysis over the considered year). This suggests that the presented results are robust and can be extrapolated to other periods.

This discussion was added to the paper.

<u>This part L32 P11 to L1 P12 of the old version of the paper:</u>

'On the other hand, an important reduction of ozone is observed in the tropics at 20 hPa (more than 600 ppbv). To better…'

<u>is replaced by:</u>

'On the other hand, a large reduction of ozone is observed in the tropics at 20 hPa (more than 600 ppbv). We have performed a t-test to evaluate the significance of these differences between the two experiments in terms of zonal averages. These were obtained by averaging the analysis over the month of the study and over longitudes. We have used the standard deviation computed for each average to perform our test. We have noticed that in the majority of regions, especially where the differences are large (between 300 hPa and 10 hPa), the differences are statistically significant (not shown). To better understand the impact of the estimated…'

in the new version of the paper L33 P11 to L4 P12.

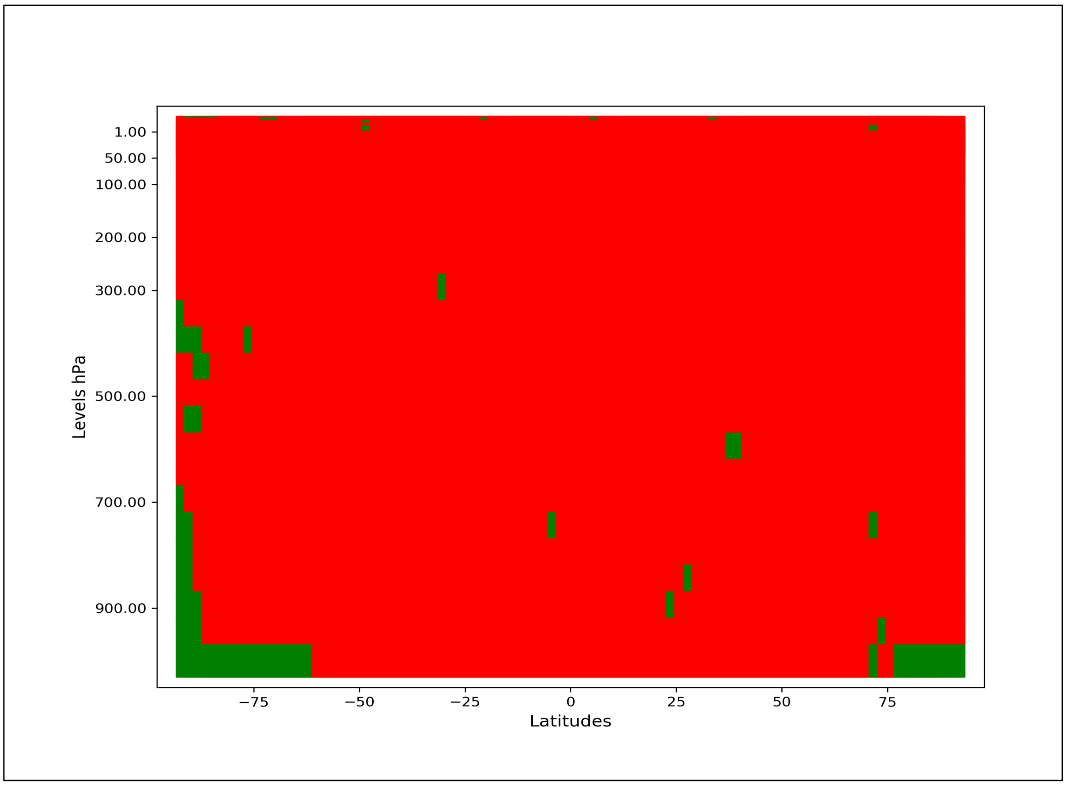

Figure 3_review: T-test of the zonal averages of the two experiments (RfullExp and RdiagExp); the green color reports the null hypothesis H_0 (the two experiments are not significantly different in terms of zonal averages) and red color reports the alternative hypothesis H_1 (The results are significantly different in terms of zonal averages).

Specific comments:

1. *Page 2 line 24 and elsewhere: There is another recent work that can be cited here, Bathmann and Collard 2020. It might be worthwhile and relevant to include this reference as the authors also examined IASI error correlation matrices over land and sea, and assimilated IASI ozone channels.*

   Indeed, it is a relevant reference to be cited here. we have added it to the manuscript

2. *Page 9 last paragraph and first paragraph of page 10: Where do you remark that the estimated standard deviation is proportional to the radiance values? I think the larger standard deviations in the SST channels (compared to ozone channels and in general) can probably be attributed to greater sensitivity to emissivity and cloud detection error, as well as greater representivity error.*

We show in Figure 4_review the **R** standard deviation, the average of observations, and the average of the background in the observation space ($H(x_b)$). At first glance, we notice that the estimated standard deviation has a very similar shape to that of the observed radiances or the equivalent of the background in the observation space. The ratio of the estimated standard deviation over the observation is about 5 % for SST channels and 2 % for ozone channels. We have suggested in the paper that the larger absolute error in the SST band compared to the ozone channels might be explained by the large values of the observation and the background for the SST channels in comparison with respect to the ozone channels. The (Desroziers) statistics are computed by multiplying the observation minus background values times the observation minus analysis. Since larger absolute values of **y** correspond generally to larger deviations, the observed spectral behavior of the errors seems natural. As the referee has suggested, the slightly larger relative error in the SST band could also be attributed to greater sensitivity to emissivity and representivity error. However, it is difficult to draw such conclusions from the sole Desroziers estimation procedure and we consider that a more detailed analysis of the individual sources of errors is required to better assess the effect of the different errors.

The paper was modified to include this comment.

This part of the old version of the paper L34 P9 to L2 P10:

'We remarked that the estimated standard … for the entire spectral window (not shown)'

Was replaced by:

'We have plotted the R standard deviation, the average of observations, and the average of the background in the observation space on the same figure (not shown). We have noticed that the estimated standard deviation has a very similar shape to that of the observed radiances or the equivalent of the background in the observation space. This may suggest that the larger absolute error in the SST band compared to the ozone channels might be explained by the large values of the observation and the background for the SST channels in comparison with respect to the ozone channels. It could also be attributed to greater sensitivity to emissivity and representivity error.'

This part was added to the new version of the paper L8 to L13 P10:

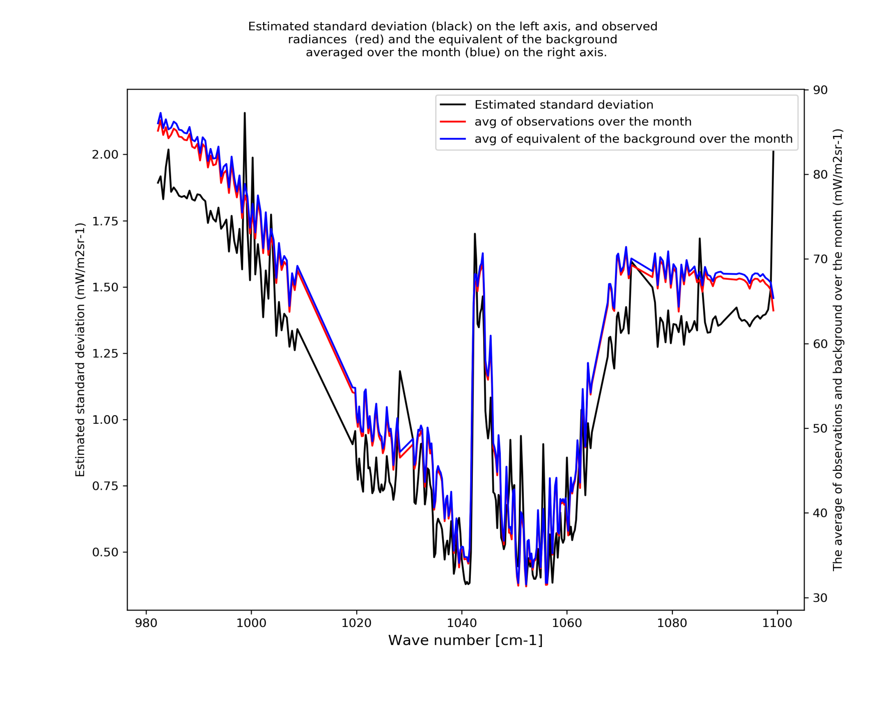

Figure 4_review: Estimated standard deviation (black) on the left axis, and observed radiances (red) and the equivalent of the background averaged over the month (blue) on the right axis.

3. *Section 4.3 The conclusion that I can draw from this section is that larger error variances increase the convergence rate of the minimization algorithm. It is mentioned that the diagonal matrix pulls the analysis solution closer to the observations, and I think the discussion can refer back to Fig 1. The errors are generally larger in RfullExp, so these observations are being downweighed in RfullExp.*

Indeed, in the case of RfullExp the errors are larger and observations are downweighted as a result. We added a reference to the Figure 1 in the discussion.

This part of the paper was modified (see comment 1 of 'general comments')

Section 4.3 The discussion about the number of iterations that are necessary for the minimization to converge is a little confusing and I wonder if the results are robust. In the first paragraph, it is stated that it converges in 90 iterations if a non-diagonal R is used. Then in the second paragraph, it is stated that it takes more than 100, 60 and 70 iterations to converge with the 1st, 2nd and 3rd estimates of R. Where did 90 come from? Also, are these numbers of iterations averaged over multiple assimilation cycles, or are they just from one cycle?

Indeed, the way we have presented the number of iterations was not done properly and may create confusion. In fact, we have picked arbitrarily the number of iterations (90) from an assimilation cycle we presented in the first paragraph. In the second paragraph, we have changed the assimilation cycle considered. To correct this, we have averaged the total of iterations over all cycles.

We found that the first estimation needs an average of 149 iterations to converge whereas the second estimation requires only an average of 89 iterations.

The paper was modified to include this discussion.

This part of the new version of the paper was corrected L14 P14:

'The minimizer takes 149 iterations in average to converge. (average computed for all the assimilation cycles of the month). We used the analysis given by the 1st experiment to estimate another R-matrix. We have used this estimation to run another assimilation cycle (2nd experiment). We have noticed that the minimizer needs about 89 iterations in average to converge. We have modified the R-matrix of the 1st experiment by keeping its correlations and replacing its standard deviation with that of R used in the 2nd experiment. The resulting matrix was used to run a 3rd assimilation experiment. The minimizer needs less than 90 iterations to converge. The results of the 3rd experiment seem to suggest that updating the variance has a larger impact on the convergence speed.'

4. *Page 16 line 8: How many ozonesondes observations are available in the high latitudes? Are there enough to quantify the significance of these results?*

4.1. *How many ozonesondes observations are available in the high latitudes?*

We have used 219 radiosoundings whose geographical distribution is presented in Figure 5_review.

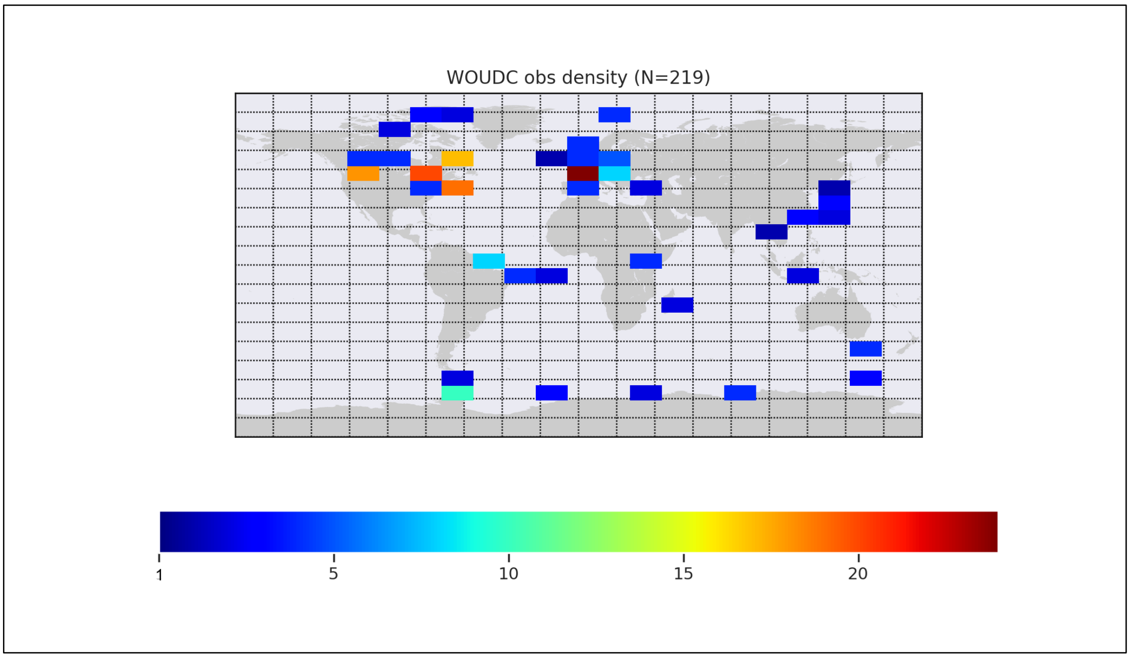

Figure 5_review: Geographical distribution of used ozone soundings.

4.2. *Are they enough to quantify the significance of the results?*

As we have mentioned in the comment 2 of general comments, we meant by ‘significant’ that accounting for a more realistic observation-error estimation brought a ’remarkable’ improvement in terms of results. However, quantifying the statistical significance requires statistical analysis.

To discuss this question, we have applied the t-test to the differences between analyses of the two experiments and observations (ozonesondes then MLS). For each observation type (MLS and ozonesondes) and for both experiments, we have computed H(x_a)-observation. We have averaged the differences over the number of available observations (ozonesondes and MLS separately). The significance of these differences between the two experiments is performed with 0.05 level confidence.

In other words, we have performed the t-test for the averages (over the number of observations) of these two quantities:

$$\varepsilon^{RdiagExp} = H(x_{a\_RdiagExp}) - Y. \qquad \text{and} \qquad \varepsilon^{RfullExp} = H(x_{a\_RfullExp}) - Y$$

with $x_{a\_RdiagExp}$ is the analysis from the RdiagExp, $x_{a\_RfullExp}$ the analysis from the RfullExp, H the observation operator and Y stands for observations (ozonesondes or MLS).

We present below the results and the number of used observations for ozonesoundings (Figure 6_review) and for MLS (Figure 7_reveiw). H0 stands for the null hypothesis, the averages of the analysis minus observations are not significantly different between RdiagExp and RfullExp. It is set to 'False' (red points) when the differences are statistically significant and 'True' (green points) for the inverse. The levels where the observations are not available are shown in blue points.

We notice, in Figure 6_review, that the significance of the differences between the two experiments' analyses and the ozonesoundings differs over the levels. The question raised by the referee is relevant. In fact, the conclusion we have made 'significant results' has to be discussed in detail (as function of levels). The reduction of the error between 20 and 50 hPa, and between 300 and 400 hPa reported in Figure 6 of the paper (all) is significant. For the low troposphere the differences are not statistically significant.

To complete this discussion, we present the MLS results in Figure 7_review. Unlike the ozonesoundings results, the differences with respect to the MLS are statistically significant for the all levels discussed in the paper (between 10 and 170 hPa, showed in Figure 6 of the paper (all)).

In conclusion, we admit that ozonesoundings as a single source of information is not sufficient to quantify the significance of the differences noted when accounting for an updated observation error. However, accounting for other sources of information (OMI and MLS) that report the improvement of the results encountered in the radiosoundings validation combined with the significance of MLS over the considered levels suggests that our results are significant.

We have modified the revised paper to include this discussion.

<u>This part added to the new version L12 P16:</u>

'To evaluate the significance of the differences between the analyses of the two experiments with respect to MLS and ozonesoundings measurements, we have performed the t-test of the differences between analyses and observations (ozonesondes then MLS). We have noticed that for the ozonesoundings, the significance differs among vertical levels. The reduction of the error between 20 and 50 hPa, and between 300 and 400 hPa reported in Figure 6 is statistically significant. For the low troposphere the differences are not significant. Unlike the ozonesoundings results, the differences with respect to the MLS measurements are statistically significant for all levels discussed in MLS validation.'

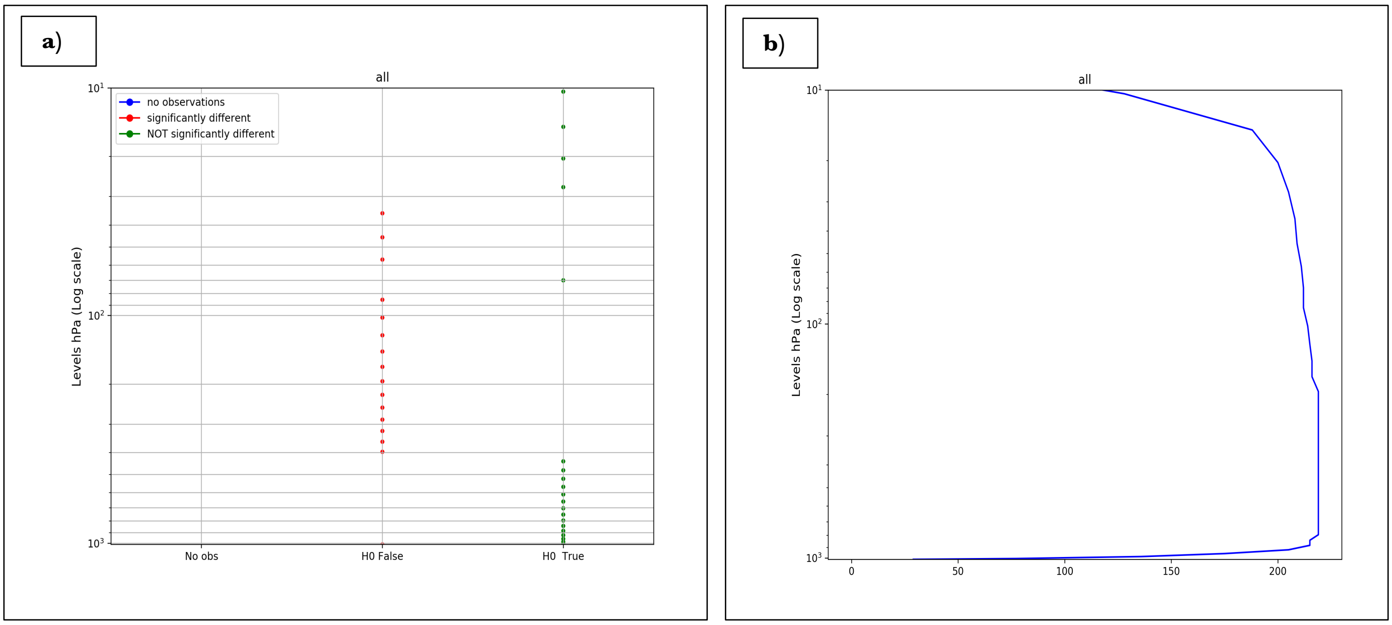

Figure 6_review: (a) T-test of the averages of the analyses minus observations (given by the ozonesoundings) of the two experiments (RfullExp and RdiagExp); the green points report the null hypothesis H_0 (the averages are not significantly different) and red points report the alternative hypothesis H_1 (the averages are significantly different), and the blue show the levels where observations are not available. The number of used observations is shown on (b).

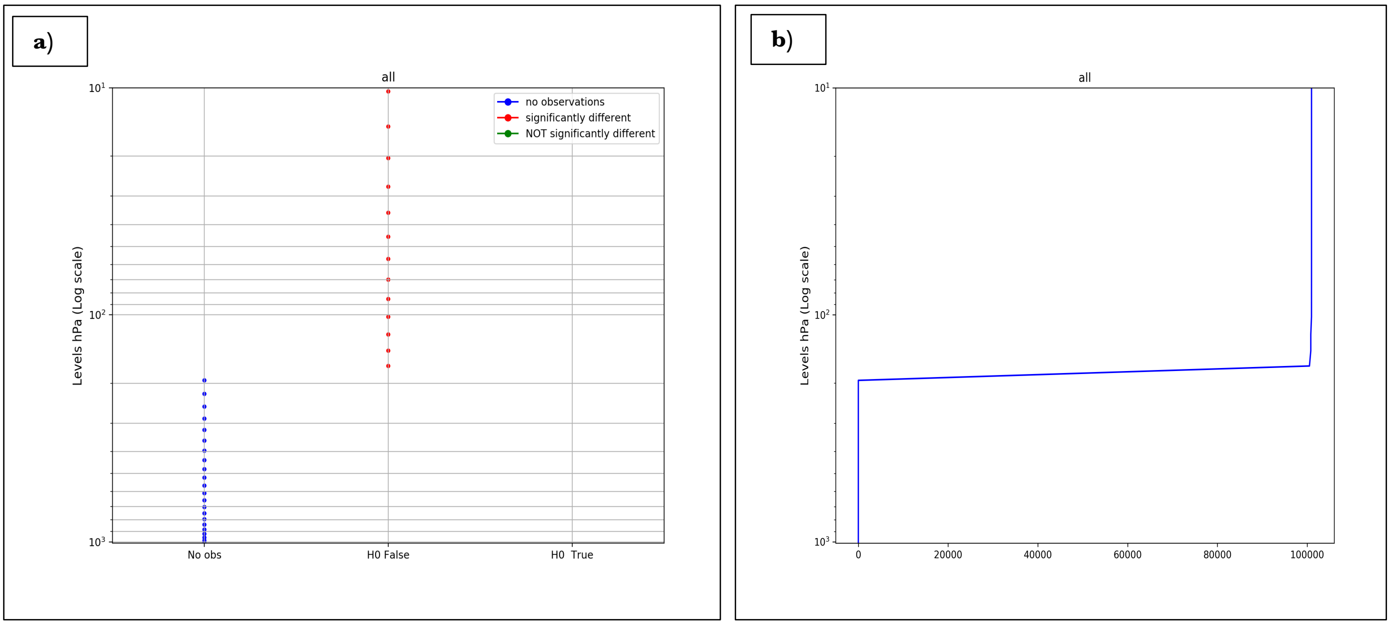

Figure 7_review: (a) T-test of the averages of the analyses minus observations (given by MLS) of the two experiments (RfullExp and RdiagExp); the green points report the null hypothesis H_0 (the averages are not significantly different) and red points report the alternative hypothesis H_1 (the averages are significantly different), and the blue show the levels where observations are not available. The number of used observations is shown on (b).

3. Technical comments:

Page 1 line 5: "adopted in some" should be "adopted in many":

Corrected

Page 1 line 16: "and in the climate" should be "and in climate"

Corrected

Page 1 line 23: "component of the observation's network" should be "component of an observational network"

Corrected

Page 2 line 4: after the colon, this sentence is not grammatically correct. Furthermore, parameters and climate change are not applications. Estimation of parameters and climate change studies are applications.

Corrected

Page 2 line 5 and elsewhere: change MetopA to Metop-A

Corrected

Page 2 line 7: can "stratosphere layer" be changed to "stratosphere"?

Corrected

Page 2 line 9" change "construct more accurate" to "construct a more accurate"

Corrected

Page 2 line 12 and elsewhere: change "chemistry transport model (CTM)" to just "CTM". CTM was introduce at line 8.

Corrected

Page 2 line 23: change "some Numerical Weather Prediction (NWP) systems" to "many Numerical Weather Prediction (NWP) centers."

Corrected

Page 3 line 5: this should say "using the Desroziers method"

Corrected

Page 3 line 6: abbreviate CTM

Corrected

Page 4 line 6: change "the TOVS instrument" to "TOVS instruments"

Corrected

Page 4 line 10: "The radiative transfer..." multiple verb tenses are used in this sentence. It should probably only be in past tense.

Corrected

Page 4 line 28: change "the Skin Surface Temperature (SST) and the ozone" to "Skin Surface Temperature (SST) and ozone"

Corrected

Page 5, line 14 remove the last access statement

Corrected

Page 6 line 1: change "transmitting continuously" to "continuously transmitting"

Corrected

Page 6 line 11: change "section of the results" to "the results section"

Corrected

Page 6, line 22: change "examine here exclusively" to "exclusively examine" and "as already reminded in the introduction and in the conclusion" to "as mentioned in the introduction"

Corrected

Page 6, line 23 change "and correlation" to "and a correlation"

Corrected

Page 7 lines 8-10. "The systematic error..." This sentence is redundant with the one that comes after it.

Corrected

Page 8 line 5: change "have used the channel" to "used IASI channel"

Corrected

Page 8 line 7: change "for the long" to "for a long"

Corrected

Page 8 line 8: measurements should be singular

Corrected

Page 8 line 12: change "we were used" to "we used"

Corrected

Page 8, line 15 delete "also"

Corrected

Page 8 lines 16-17: IASI is not an ozone instrument

Corrected

Page 8 line 18: change "analyses" to "analysis"

Corrected

Page 8 line 22: change "accounted by" to "accounted for by"

Corrected

Page 8 line 26: change "statistics of error" to "error statistics"

Corrected

Page 9 line 1: "which may not always be the case" In practice it is almost never the case. Page 9 line 15: change "from 3D-Var experiment that uses" to "from a 3D-Var experiment that used" and "1st" to "the 1st"

Corrected

Page 9 line 22-23: "The differences…" there are mixed verb tenses in this sentence

Corrected

Page 10 line 4-5 You can delete the date on this personal communication.

Corrected