Response to Reviewer 2:

We would like to thank Dr. Lang for performing a thorough review and for the many helpful suggestions to improve the paper.

Below, we respond to each of the review comments. For the sake of clarity, the review comments are given in blue italics and our response is printed in normal font. Changes to the manuscript are printed in green.

The paper by Tilstra et al. on the In-orbit Earth reflectance validation of TROPOMI on board the Sentinel-5 Precursor satellite is an important contribution to the meanwhile significant history of knowledge in characterising this class of instruments, since the launch of GOME-1 on ERS-2. The paper is well written and organized and it shows robust and convincing results on the assessment of the TropOMI instruments radiometric in-flight calibration accuracy using independent radiative transfer (RT) forward modelling of the expected Top-Of-Atmosphere (TOA) radiances in the region between 328 and 2314 nm.

The fact that the results are convincing and robust with respect to the provided statistical analysis of the reflectance correlations (between simulated and measured results) is not a given, since previous attempts to use RT model results in evaluating the in-flight calibration accuracy and performance of this class of sensors were limited predominantly by uncertainties in model inputs.

In this respect the statistical significance of the presented results, and the fact that they may be used as a robust evaluation of the sensor performance for most of the large spectral regions investigated here, makes it a unique contribution, generally in the field of post-launch high-spectral resolution sensor calibration in the UV to the near infra-red (NIR).

In this respect, and in view of future missions in development (Sentinel-5), or planned mission like the proposed High-Priority Copernicus Candidate CO2 monitoring mission (HPCC CO2M), it is unfortunate that the short-wave infra-red (SWIR) band evaluation had to be excluded from the results, probably because of systematic error contributions in the RT model surface reflectance input. In view of the significantly stronger focus of S5 and CO2M on the SWIR, improving the analysis and associated methods in this spectral region would be very welcome (and as suggested by the authors they consider this possible provided a better surface albedo input becomes available).

It is indeed very unfortunate that no strong conclusions could be drawn from the SWIR results. It is still not understood what exactly causes the unexpected and suspicious behaviour for the 2314-nm wavelength band. Because we cannot explain the behaviour for this wavelength band we are forced to dismiss the end result for this wavelength band. We are at the same time, however, not excluding the possibility that the results that were found for the 2314-nm wavelength band are "real". Further studies will have to investigate this.

The authors address the two key error-sources for the overall approach in quite some detail. One is the selection of real clear-sky Rayleigh scattering scenes, with no or negligible residual cloud or aerosol scattering contribution. The other is the accuracy of the used surface reflection data, and in particular their angular and spectral variation (spectral BRDF), which is a large contributor - in addition to the accurate knowledge of the surface albedo itself - to the total budget of the forward modelling RT error and has significantly hampered previous attempts for carrying out such studies.

Cloud screening is predominantly done using co-located VIIRS cloud detection, and aerosols are detected by deriving an Aerosol Absorbing Index (AAI) from the measurements themselves. Here the authors state that they avoid a vicious circular problem by "correcting" the reflectances. However it seems not quite clear if the way these valuse are corrected is then actually implying an overall iterative approach, i.e. going through the full procedure, coming up with an radiometric offset (or residual between measurement and RT result), then use these results to correct the radiances, and finally repeat the procedure until some convergence criterion is fulfilled. This aspect should be, from my point of view, better described or clarified in the paper.

The corrections on the reflectances that are used to calculate the AAI are not completely based on the validation study, but partly derived independently.

To explain, we independently analysed the reflectances at 340/380 nm which are stored in the TROPOMI AAI product. From this, a (time dependent) correction for degradation was derived for the reflectances following the method explained in Tilstra et al. [2012]. This correction for degradation was applied to the TROPOMI 340/380-nm reflectances that go into our calculation of the AAI. This part of the correction does not depend on the outcome of the study, because it is derived independently.

A day-zero correction was also applied to the 340/380-nm reflectances, by focusing on one day only ("day zero", the first day studied after the end of the commissioning phase, 1 May 2018, indicated by the blue dots in Figure 8). For this particular day, the reflectance correction factors were found from the reflectance validation study and these were then used as day-zero corrections for the reflectances that go into the AAI calculation. This part of the correction *does* depend on the validation study itself.

However, the correction factors resulting from the validation study are not changing significantly if a second round is performed to calculate the day-zero correction. Note that the AAI threshold value was set to a relatively high value of 2.0 index points. The impact of the filtering on AAI is therefore rather low to begin with.

In conclusion, there is no real need for an interative scheme.

Note that the primary reason for calculation the AAI ourselves is the fact that it is not allowed to average AAI values, but it is allowed to average the reflectances from which the AAI is derived, as explained in Section 4.3.

We have added the following sentences to Section 4.3 of the paper, to better explain the approach that was followed to correct the reflectances that go in to the calculation of the AAI:

" The corrections consist of a time-dependent multiplicative correction to account for the linear drift observed in the 340/380-nm reflectances that are stored in the S5P AAI product, and of a day-zero multiplicative reflectance correction. The linear drift was determined using the method presented in Tilstra et al. [2012]. The day-zero correction factors were found from applying the reflectance validation study on 1 May 2018 ("day zero") and determining the correction factors found for 340 and 380 nm. In principle this step could be repeated by applying the correction factors, calculating newer correction factors from these, until convergence is reached. An iterative scheme for this was not pursued, however, because a secondary iteration turned out to already provide almost the same correction factors as the first one. "

The accuracy of the used surface reflectance is for sure the most critical aspect for the performance of the RT forward model results in the described validation approach. The author's present 4 methods to validate

the Lambertian Equivalent Reflectance (LER) database derived from OMI and SCIAMACHY, the latter being probably used in order to cover the SWIR. The missing angular variation is considered as one of the largest errors in the existing OMI and SCIA LER databases, especially over vegetation, and this is why the chosen filtering on low quality LER values makes use of BRDF and albedo information as provided by Modis the Terra or Aqua satellite platforms. This however raises two questions:

1) Since the problem in the missing angular information in the LER has already been described for the existing LER databases derived from GOME-2, and therefore an angular dependent version of the GOME-2 LER database has been developed in the context of the Atmospheric Composition Satellite Application Facility (ACSAF) activities for Metop, the question arises why this database has not been used here. Especially since, in the end, the SCIA LER database in the SWIR shows deficiencies, and the analysis of this region had to be excluded anyway.

The *directional* GOME-2 surface albedo database can unfortunately not be used to provide the missing angular information, because the directional information stored in this database is representative for polar orbits with a 09:30 LT equator crossing time. The GOME-2 instruments are all kept in 09:30 LT orbits, while TROPOMI is in a 13:45 LT orbit. It is only possible to use the GOME-2 directional information for data observed from polar orbits with equator crossing times close to 09:30 LT. This includes data from GOME-1, SCIAMACHY, GOME-2 itself, the future Sentinel-5/UVNS, but not data from OMI or TROPOMI.

It is possible to use the *non-directional* GOME-2 surface LER information which is also stored in this product. In fact, this is what we did during the initial phase of the study. Unfortunately, the GOME-2 surface LER has some issues for wavelengths below 400 nm. These issues have been reported in the past and one could easily take this into account via corrections. However, this would not improve the readability of the paper and it would also not improve the results of this study because the SCIAMACHY surface LER database does not have a lower quality than the GOME-2 surface LER database. SCIAMACHY does have the advantage that its equator crossing time (10:00 LT) is a bit closer to the 13:45 LT of TROPOMI.

SCIAMACHY also provides surface albedo for the SWIR channel. Although we indeed had to exclude the results for the 2314-nm wavelength band from the manuscript, we were at least able to report that linearity problems and offset issues do not seem to be present for the SWIR channel. We would not have been able to draw these conclusions using the GOME-2 surface LER database.

In the revised manuscript we now mention in Section 3.4 that the GOME-2 surface LER database could also have been used, but that we decided to use the SCIAMACHY surface LER database primarily because of the additional information in the SWIR wavelength range:

" As an alternative to the SCIAMACHY surface LER database we could also use the very similar GOME-2 surface LER database [Tilstra et al., 2017]. However, the GOME-2 surface LER database does not provide information for the SWIR wavelength range, which is the primary reason for adopting the SCIAMACHY surface LER database instead. Alternatively, we can make use of MODIS surface BRDF . . . "

2) In the context of TOA test-data simulations of this kind of spectrometers, the MODIS surface albedo and BRDF is frequently used in conjunction with MODIS surface type characterisation and the ADAMs (A surface reflectance Database for ESA's earth observation Missions - https://nebula.esa.int/sites/default/files/

The possibility of using the ADAM database was considered and discussed, but in the end we did not go for this option. The principal component method used in the ADAM database & toolkit seems to be a responsible way of handling the interpolation between the MODIS wavelength bands. Certainly, the wavelength interpolation provided this way is addressing the problem that we mentioned in the manuscript (in Section 3.4) of the – for our specific purpose – inconvenient location of the MODIS wavelength bands.

However, we do have our doubts that the interpolation can be done with a high enough accuracy, because the seven MODIS bands are quite far apart. For example, the shortest MODIS wavelength band is centred at 469 nm, which would involve extrapolation (not interpolation) to get towards 328 nm. A lower quality is to be expected here, as reported in, for instance, Bacour et al. [2020]. For the longest MODIS wavelength band centred at 2130 nm we would need extrapolation to get to 2314 nm. For the wavelength region at about 700 nm, interpolation has to been done between 645 and 858 nm with no support from space-based observations. Especially here, around the red edge, interpolation of any kind seems to be challenging. The OMI and SCIAMACHY surface LER databases provide surface LER for all wavelength bands studied in our validation. The (monochromatic) wavelength bands of these databases were defined in exactly the same way as the wavelength bands defined in the manuscript.

In the recent paper by Bacour et al. [2020] the ADAM database and its toolkit are introduced extensively. The paper mentions that the ADAM database *overestimates* the BRDF near the "hot spot" by a factor of about 1.5 for barren surfaces. For vegetated and other surfaces, the ADAM database *underestimates* the BRDF near the hot spot by a factor of 1.5–2.0. This is quite a large underestimation/overestimation, and it is not clear to us whether these errors are caused by the ADAM interpolation scheme or by the underlying MODIS BRDF database. It would be an alarming fact if these errors would be present in the MODIS surface BRDF information. In any case, the reported underestimation/overestimation by a factor of 1.5–2.0 is a thing to worry about, because with the close to noon (13:45 LT) equator crossing time time of TROPOMI and the relatively wide orbit swath the backscatter condition is found in a rather large part of the TROPOMI orbit.

The water BRDF in the ADAM database is not derived from MODIS, but modelled. Although modelling surface reflection for water surfaces is perfectly possible, it could lead to a difference in quality between land and water pixels. Such differences are not found in our analyses, because we treat water and land pixels in the same way. If there are deficiencies in our approach because of issues in the OMI or SCIAMACHY surface LER used, then at least the error is made in a consistent way for both land and water surfaces.

Another thing to consider is that using ADAM surface albedo input would link the validation results to the radiometric calibration of MODIS. Right now, they are linked to OMI and SCIAMACHY. One could argue that the validation results are also linked to MODIS, because of the filtering which is partly based on MODIS.

But for the most part, the link is with OMI and SCIAMACHY. This seemed to us to be the proper thing to do, especially considering the heritage of the TROPOMI instrument and of its retrieval algorithms.

In the end, for the reasons mentioned above, we decided to rely on the OMI and SCIAMACHY databases.

In the revised manuscript we now mention the ADAM database, that this approach would address the problem of the position of the MODIS wavelength bands, and explain why we decided not to use it. The text now reads:

" Another possible source of surface albedo is the ADAM database [Bacour et al., 2020]. This database is based on MODIS BRDF, but provides interpolation to any wavelength in the wavelength range 240–4000 nm. This would address the issue related to the position of the seven MODIS wavelength bands. However, interpolation does lead to a lower quality and reliability, as explained in Bacour et al. [2020]. Next to this, the ADAM database shows both underestimation and overestimation by a factors of 1.5–2 in backscatter directions [Bacour et al., 2020, p. 18]. For water surfaces, the surface reflection is modelled, and bears no relationship with the MODIS BRDF data. This means that consistency in calibration between land and ocean is broken. For the above reasons we decided to use the OMI/SCIAMACHY approach. "

Minor and editorial issues:

p.2, l. 13: The reference to GOME-2 on Metops could be associated to the relevant paper by Munro et al.

Yes, it should indeed. We have added the reference to Munro et al. [2016] to the revision of the paper.

We have also added the reference to TROPOMI, which was also missing.

General, the exclusion of band 8 should probably be motivated more towards the beginning of the paper.

We agree and now do this by referring to Section 4.1 already in the Introduction of the manuscript:

" The analyses cannot be performed for TROPOMI bands 1, 2, and 8 for reasons explained in Sect. 4.1. "

Section 3.4. The temporal aspects on using a database derived from SCIAMACHY and its application to a recent missions, should probably be mentioned and/or addressed, especially for vegetation and crop surfaces.

We agree and have added the following lines to Section 3.4:

" It should also be noted that the OMI and SCIAMACHY surface LER databases are mostly representative for the time periods from which they were derived (OMI: 2005–2009; SCIAMACHY 2002–2012). Systematic changes in surface reflectivity occurring after these time periods, for instance due to changes in land use, are not covered by the databases and will result in errors. "

p. 8, l.3. "From Fig. 1 it can be inferred. . .". I find it actually quite difficult to infer it from the Figure if the differences are small. From the Figure 1 one can only for sure infer that they are in the right overall magnitude and spectral relation.

We agree and have changed the sentence:

" However, the impact is relatively small and of little concern. "

The term "inhomogeneity" is meant as inhomogeneity due to the occurrence of both land and water inside the box. It is based on the surface type information that is available in the TROPOMI L2 products.

Inhomogeneity in terms of the standard deviation of the TROPOMI reflectances was calculated, but used to track down outliers which were not detected by the TROPOMI flagging. This is sometimes an issue for the SWIR channel. It was not used for filtering out scenes/boxes that were considered too inhomogeneous.

The idea behind the one-by-one degree boxes is to have the OMI and SCIAMACHY surface albedo information line up as best as possible with the TROPOMI observations.

In the revised manuscript, we now clarify how the inhomogeneity of a box is determined:

" ... sun glint or solar eclipse events. Inhomogeneity and snow/ice presence are determined based on the surface type and surface condition indicators inside the TROPOMI L2 products. A box is only considered inhomogeneous if it contains both land and water. The number of boxes ... "

Figure 3 shows the location and number of measurements over the defined period and number of days for clear sky scenes. However, it is not clear if this is the final statistics for all 56 days applying method 4 for cloud screening. The distribution of the locations of the latter, which actually go into the results, would here be of highest interest.

It is the final statistics for all 56 days, so Figure 3 illustrates the distribution and frequency of the clear-sky one-by-one degree boxes that were collected over all 56 days. These are all boxes which were not filtered away for reasons explained in the text, such as presence of snow or absolute latitudes above 60 degrees. Filtering such as performed in Method 4 was not applied for this image, because the other three methods are also discussed in the paper, and because Method 4 has not been defined at this point in the manuscript.

To clarify the meaning of Figure 3 we have changed the caption of Figure 3 in the following way:

" Figure 3. Location and frequency of the $1° \times 1°$ cloud-free latitude/longitude boxes that were selected by the algorithm every two weeks for the period from May 2018 until February 2020. A total of 56 days contributed to the image. The selection and filtering approach is described in Sect. 4.4. "

Changes to the manuscript:

During the review phase of this manuscript the TROPOMI instrument kept generating new data and we took the opportunity to extend the time range that was studied originally (May 2018–February 2020) by three months. The studied time period is now May 2018–May 2020, covering two years. Some of the numbers reported in the paper have changed slightly, but not significantly. Figure 8 and Table 3 have also been updated. The changes w.r.t. the previous version are in all cases insignificant, well within the reported accuracies. The extension of the time range that was studied has led to an increased accuracy of the results. Figure 3 was also updated.

References:

Bacour, C., Bréon, F.-M., Gonzalez, L., Price, I., Muller, J. P., Prunet, P., Straume, A. G.: Simulating multi-directional narrowband reflectance of the Earth's surface using ADAM (a surface reflectance database for ESA's Earth observation missions), Remote Sens., 12, 1679, doi:10.3390/rs12101679, 2020.

Munro, R., Lang, R., Klaes, D., Poli, G., Retscher, C., Lindstrot, R., Huckle, R., Lacan, A., Grzegorski, M., Holdak, A., Kokhanovsky, A., Livschitz, J., and Eisinger, M.: The GOME-2 instrument on the Metop series of satellites: instrument design, calibration, and level 1 data processing – an overview, Atmos. Meas. Tech., 9, 1279–1301, doi:10.5194/amt-9-1279-2016, 2016.

Tilstra, L. G., de Graaf, M., Aben, I., and Stammes, P.: In-flight degradation correction of SCIAMACHY UV reflectances and Absorbing Aerosol Index, J. Geophys. Res., 117, D06209, doi:10.1029/2011JD016957, 2012.

Tilstra, L. G., Tuinder, O. N. E., Wang, P., and Stammes, P.: Surface reflectivity climatologies from UV to NIR determined from Earth observations by GOME-2 and SCIAMACHY, J. Geophys. Res.-Atmos., 122, 4084–4111, doi:10.1002/2016JD025940, 2017.