Reviewer comment:
The authors mostly addressed my comments, even though they appear hesitant to put more work into the current study.

Dear Referee
Thank you for the thorough and in-depth look. I hope I can address your points

With kind regards,

Tim

A few major points:

1) If there is a study in prep, please make sure the current manuscript won't become entirely obsolete. The authors should avoid just writing a separate paper that actually includes some of the things I would have liked to see in this one.

I agree with the author that I would like to have included a through and satisfying description on the BRDF effects as well. But this has proven, with the current knowledge and data availability, simply not realistic.
One of the main difficulties is the validation of any BRDF corrections. For this, one requires a site with ground instruments, such as RailRoad Valley, or other sites in the RADCALNET network (although for SWIR wavelengths currently only RRV suffices due to the lack of wavelength coverage).
Our investigations into this effect have progressed significantly, but cannot be applied to this paper. Primarily due to the absence of validation of these effects at any of the Saharn sites.
The work on RRV currently is extensive and requires its own paper.

There were three conclusions so far that are worth presenting in this response:

- Over desert sites, the MODIS BRDF product was found not to be accurate enough. It often saturates over desert sites, in particular at longer wavelengths, providing a lack of data coverage. The uncertainties are also large.
- VIIRS datasets are not saturated, but appear to lack the so-called 'hot spot' in their derivation, at least in the current publicly released dataset. Together with GOSAT and OCO we see consistent differences (see Bruegge et al. for the hotspot). A VIIRS dataset using the correct MODIS model would be a potential solution, but such a dataset is not publicly available.
- Determining what/how the BRDF effects need to be corrected down a level below 5% in **absolute** terms has proven to be difficult. This is in line with the results as presented by previous work (Bruegge et al., 2019a/2019b, Kuze et al., 2014). 5% uncertainty is currently larger than the spread/variation we see. Note that we are aware we are comparing absolute (5% over RRV with atmospheric correction etc) and relative (3% spread) differences.

These results also directly show why the current paper will not be automatically be obsolete once the paper on RRV has been published. None of the sites given in this paper are instrumented. It thus cannot be validated that any BRDF effects derived for the RRV desert composition and applied to the Saharan, Namibian and Arabian deserts, is correct. This in particular is relevant at 2.3 micron where no other reliable instrumented site yet exists (at visible wavelengths one can compare RADCALNET sites).

Given the results of Bacour et al., the assumption that these effects are equal in RRV and any other desert is unlikely to hold down to the accuracies we have derived so far.

---

2) Can you be confident that BRDF is likely the main reason or can local heterogeneity (given the large footprint size) be a reason as well? As you haven't looked into MODIS yet, I doubt you can exclude the latter reason, right? Similarly for RR valley, Are there actually TROPOMI footprints that only covers the homogenous playa? PICS validation gets easier with finer spatial resolution, this should be clearly stated, esp. at the end when you make broad recommendations.

Local heterogeneity on scales of 20 and 100 km was investigated by Bacour et al., 2019 and is explicitly discussed there using the MODIS and Google Earth data. Heterogeneity has three possible effects, which can contribute:

- The effect would be seen by height heterogeneity (e.g. shadowing of dune formation), with the yearly variation. This has been mentioned. This is now expanded in the text as it is not fully explained.
- The pixel size of TROPOMI varies in size and would alter the BRDF by including different BRDF values at the edges. Any heterogeneity in the central part (~3 km^2) is not relevant as it is always included.
  If dominant, this would introduce a very regular pattern every ~16 days due to the repeatability of the S5p orbit cycle (which consists of 227 orbits to be more precise). The pattern would be regular but complex in form, as the size difference of the TROPOMI pixel is not a linear function of swath position due to the curvature of the Earth, in addition of the location of the central reference location varying at a subpixel scale. The pattern's shape would be site-dependent, but its period would not.

  One can approximate whether or not this heterogeneity is influencing the results by looking at the Standard deviation we derive and compare that to the homogeneity values.
  By taking the Bacour et al., SHOM 20 km (~max TROPOMI pixel size) values of their Table 2 (using Optimal locations), one would expect the spread to be largest for Algeria3, Algeria5, Algeria2, Niger3, Arabia2, Namibia_PICSAND1 and Sudan_PICSAND1. These have larger SHOM values above 0.7%. Note that due to the swath position argument above, this number is very difficult to relate to TROPOMI results.

  However, Comparison with the STDEV of Table 3 of our paper shows there is no clear correlation for these sites with higher STDEV.
  This conclusion is reinforced by the results of the nadir view. If heterogeneity was a larger effect, the standard deviation would have to be severely reduced when only viewing the nadir view, even with the much lower number of available data points. This due to the pixel size being nearly identical for this subset. Although for some it is slightly reduced, this is also not seen.
  This analysis is included in the paper.

---

3) You mention 5 impact factors: (detector position, instrument zenith, instrument azimuth, solar azimuth, solar zenith. This actually points to the lack of atmospheric correction being a primary reason, or stray light or other effects? on-board calibration also has some drawback as the light doesn't fully the exact same path as on orbit during

regular observations. It would be good to discuss these pros/cons and clearly state the lack of atmospheric correction here as well.

The referee is correct that the lack of atmospheric correction can indeed be an effect, primarily due to the presence of aerosols and its variation. This is included in the papers assumptions and discussion.
Note that I hope that in my definition of atmospheric correction, the actual reflection on the surface is not included in this nomenclature. This is now included.

For various reasons, the other effects are all far less likely and have been discussed in earlier work.
TROPOMI-SWIR Straylight has been carefully studied both on-gruond and in-flight (e.g. Tol et al., 2018, van Kempen et al., 2019, Ludewig et al., 2020) and has been shown to not be a factor for the SWIR detector. On-ground this was done over the identical optical path as the radiance (Tol et al., 2018). In-flight this is monitored by the diode lasers with variations seen less than tens of a percent (van Kempen et al., 2019, www.sron.nl/tropomi-swir-monitoring/swir-straylight). A similar argument can be made for the lack of variation in the ISRF inflight (van Hees et al., 2018).

Similarly, on-board calibration has shown the optics and detector to be stable down to tens of a percent or less (van Kempen et al., 2019). The only part of that is not actively monitored with either on-board lights or the irradiance, is the Earth-port entry/telescope itself (Figure 1 of van Kempen et al., 2019, green entrance at the lower right.). Long-term variation similar to what is seen over the Sahara deserts, is in itself monitored by reflectance monitoring over the entire Earth (monthly, see e.g. mps.tropomi.eu report, e.g.
http://mps.tropomi.eu/pdf/reporting?year=2021&number=07&freqchar=M&creationdatetime=20210801T121718, page 12, see also Ludewig et al., 2020). No variations are seen down to levels of less than 0.5%. Spikes are produced by averaging more/less cloudy areas).

We hope this explains the points raised by the referee.