Referee 2

The manuscript Monitoring the TROPOMI-SWIR module instrument stability using desert sites by van Kempen et al outlines an approach for TROPOMI SWIR instrument stability monitoring using PICS site over deserts. While the manuscript is in general well written, I do have some high level questions/concerns regarding the ultimate utility of the study for TROPOMI validation apart from stating that it is in line with the onboard calibration routines. In the following, I will briefly describe some of these concerns, followed by a few detailed comments on typose, etc.

Dear referee, thank you for your constructive criticism. They proved useful. With the replies below, we hope to address your concerns.

with kind regards,

Tim

Major comments:

As far as I can see at the moment, the entire outcome of this investigation is a bulk characterization of the 2D detector response at one specific small wavelength band averaged over the entire spatial domain. However, I haven't seen any discussion on the implication of this limiting factor at all, which I consider rather large. First of all, the across-track pixels might all degrade to a different degree or have an overall different absolute calibration error.

We know from the monitoring of the TROPOMI-SWIR detector that these hypotheses are not applicable and that the detector is very stable.This is (in part) reported in van Kempen et al., 2019 for the first year of operations. Monitoring of the calibration can be found on the website of SRON: https://www.sron.nl/tropomi-swir-monitoring as well as derived from the weekly reports posted on http://mps.tropomi.eu/reporting corroborate these results for the full mission so far.
The referee is correct however that these points are far too implicitly assumed in the paper. We have added a sub-section in Section 3 to make these conclusions explicit.

However, their response is also impacted by potential BRDF effects as each across track element has its own viewing zenith angle. Given the data density of TROPOMI, I was somewhat surprised that the authors didn't try to at least disentangle some of the across-track element variations.

We started with the analysis of potential BRDF effects, but following results of e.g. Bruegge et al., 2019 and a collaboration with the JPL group using RailRoad Valley data, we quickly concluded that this is a much more complicated analysis with a significant amount of relevant parameters (BRDF of the surface, overpass time and angles of both sun and instrument). We are currently investigating these effects in much higher detail over RRV with the GOSAT and OCO teams. But to apply these correctly to the results in this work clearly was beyond the scope of this paper.

This 'problem' was only succinctly described in lines 100-105 in the preprint version, and I agree with the referee it was far too concise. This has been expanded to a new subsection and point to the future publication on this topic.

At the moment, I am somewhat uncertain what new information this manuscript reveals, especially as the scatter is surprisingly and the variations in slopes quite variable too. The authors would have to better explain the added value of this method (on top of the on-board calibration, which can characterize the entire FPA response).

The most relevant part of the information is the feasibility of this method for spectrometers in the SWIR region. I agree the scatter is surprising, and we are working on this to derive the methods to improve this (see above). The variations in slope are both an effect of the scatter and the surface properties themselves.

For TROPOMI itself, the added value on top of on-board calibration and a relatively poor validation, is indeed relatively marginal, although it would be the first L1b validation for the TROPOMI-SWIR. That in itself would be added value to the scientific community.

However, for future proposed small-sat missions where total weight is a major concern, the omission of on-board calibration will become more defendable given the existence and details of these methods. In particular, for SWIR missions this has not yet been demonstrated and is a major concern with some missions in development.

The origin of the scatter would have been an interesting feature to dig deeper into, but the authors chose not to, which is somewhat dissatisfying, as this could have been valuable to the community.

We agree with the referee, and are working on producing such a result in an upcoming publication (see above).

Minor issues:

Line 31: "calibrated column densities" These are retrieved products from calibrated spectra, not itself "calibrated" datasets (maybe validated and some post-hoc "calibration" like bias correction applied)

corrected

Line 48: allpart

corrected

Line 50++ Here you mention all kinds of impact factors but then chose to ignore all of them. Why not use TROPOMI and its large swath to actually see whether you can detect BRDF effects that can clearly be separated from detector effects across the spatial domain.

See comment above. We are carrying out a study to do this. This is now rephrased.

Line 57: "but also suffers from inaccuracies" Across the manuscript, statements like this are scattered. If you point out a weakness of an instrument, you will have to justify the statement with a citation or elaborate how you come to the conclusion. However, you can't just make a statement like this out of thin air in a peer-reviewed publication. Also, what does "most complete" mean in this context?

Rephrased

Line 57: "due to its very wide swath opening" --> just swath is fine, swath opening sounds awkward

Rephrased

Line 65 --> used for monitoring the stability of a large numberg of ...

corrected

Line 76: Why only every 5 days? The cloud cover should be low, so I dont understand the 5 day limit. Is it the overlap requirement? With MODIS data

being available, you should also be able to determine the impact of the exact spatial overlap across variable surfaces with slightly varying albedos. As far as I can see, no attempt has been made to compare against MODIS data (e.g. to look at sub-pixel variability, etc).

These refer to daily overpasses, and the occurrence, once every 5 days, of multiple overpasses within 1 day. Higher latitude sites (e.g. Gobi desert) would be covered even more often. This is unrelated to cloud cover, but originates with TROPOMI orbit parameters. In effect over 95% of the data is usable.

The referee is correct that no attempt was made to compare this to MODIS data. However, for a similar reason as the the BRDF, we quickly concluded that subpixel MODIS analysis rapidly increases the complexity of this comparison. In addition saturation effects of MODIS data due to the typical Solar zenith angles in combination with the high reflectance of desert surface limits this analysis severely.

Table 1 caption typos: cooerdinates... variatility...

corrected

Line 86++ Please better explain the cloud screening, at least give a proper citation. Has any screening for desert dust events been performed? Any other filters?

This section was re-ordered. Screening for desert dust events have not been performed.

Line 94: Why did you choose 50 degrees as cutoff even though this basically omits a non-significant fraction of TROPOMI's FPA? Have you checked whether adding the few additional degrees make any differences? Did you consider separating out the FPA (and thus VZA) dependence as mentioned above?

The choice of 50 is motivated by an estimate. We wanted as much data as possible without being heavily affected by the Lambertian assumption. Although the angle dependency is already apparent at smaller angles due to e.g. the hotspot (Bruegge et al., 2019), the data beyond 55 degrees suggests that BRDF effects of the SWIR even deviate from the angle-dependent BRDF effects as modelled by Bruegge et al., 2019. Without understanding both the range of BRDF effects themselves, as well as any deviations from their angle-dependency as observed compared to wavelengths, it was estimated as too impactful at large angles. Thus, the Lambertian assumption was made.

There isn’t a lot of difference up to an angle of 52, but significant higher scatter when including the full FPA or any angle from 55 degree or higher was seen.

Line 105: "are of insufficient quality to reliably improve the data" Please see my comment above. Without citation or justification with analysis, this statement is misplaced at beast and mean-spirited at worst. Any judgement statement like this requires corroboration.

Rephrased

Line 106: "A choice was made" What was the rationale of that choice? Did you consider the tradeoffs? Why not bin the analysis by viewing angles and see whether the scatter is reduced?

This binning scheme was calculated. The scatter was not reduced and appears to be dominated by BRDF effects depending on solar and instrument viewing angles. As such it was omitted.

Line 112: affect -> affects

Corrected

Line 116: "using standard mathematical rules" like what? Just gaussian error propagation? Please be specific if you can, esp. if it doesn't take up more space than "using standard mathematical rules", which is rather vague.

Corrected

Figure 2: Ths scatter is indeed large and clear outliers exists. Are these actually single measurements? If yes, can they be color-coded by the detector position or VZA (plus and minus)? Did you try to figure out why a few were low outliers by looking at the conditions during that time (or the specific detector position?)

We investigated this by color-coding, but found it to be a 5-variable problem (detector position, instrument zenith, instrument azimuth, solar azimuth, solar zenith). Although detector position and viewing zenith are related, small changes in viewing zenith show up as a few rows difference on the detector and thus give rise to larger changes.
The bulk of all outliers (not shown) were caused by cloud cover and correctly removed. None of the outliers are seen to be correlated to detector position, as these should have been detected. Due to the orbit of TROPOMI, the observations are cyclical in nature ~every 16 days, with the same detector pixel observing the

same location. other local phenomena (e.g. sand storms, moisture content) were not checked.

sine wave correction: This is rather vaguely described to be honest. It would be good to show such a fit. Does it look better if fitted against AMF or SZA? Is this something that is also seen in MODIS data? This is interesting and curious but again, the authors chose to not go the extra mile, which would have made this paper much more interesting. In general I have no problem with not diving deeper into all the issues but given that the overall relevance of this manuscript for TROPOMI validation or validation schemes using PICS in general is rather thin, I would have expected a somewhat deeper analysis into these small curiosities.

The text was improved to better motivate the usage of the sine wave. Example fits are now given. However, its dependency on AMF/SZA, as well as occurrences in the MODIS data is under investigation. It is clear by now that it is not a simple detail, as discussed above. The sine wave appears to be caused by the BRDF effects and the yearly 'dependency' of solar angle.

What we thought was indeed a small curiosity is in fact a more complex problem, as it is not also clearly site-dependent. Using a site such as RailRoad Valley, where instrumentation is available to verify results likely gives a far more in-depth answer to the questions posed by the referee than performing this over remote sites such as in the Sahara.

Line 131: "We attribute..." What is the basis of this, a hunch? You could actually look whether there is a VZA dependence! Why not do that, I really don't understand that choice.

As detailed, this is due to the result in ongoing research. It is clear that a VZA dependence is present, but it definitely also is clear that it is not just a VZA dependence that can easily be corrected for. We have made a reference to future work in which we will go in-depth on these dependencies.