Responses to Reviewer #1

We thank the reviewer for their comments and questions. Our responses are formatted as follows:

The reviewer's comment/question (numbered) is written in black italic text.

> Our responses are written in normal black text (indented).

> The revised text as it appears in the manuscript is written in normal blue text (indented), with relevant changes underlined.

Line numbers refer to the edited manuscript. We have also provided a tracked-changes document, but that has different line numbers.

1. There is no mention of instrument noise in the paper, for example equation (2) (and the related equations) rely on a fractional absorption model to infer column enhancements from reflectance changes in the measured radiance. The measured radiances will include instrument noise, with R and R' not showing the same noise due to capture under difference conditions. The fractional absorption model does not seem to account for instrument noise, meaning that the minimisation step between the R and m values may be underestimated. It is possible that the instrument noise is accounted for in the forward model calculations as identified in Jervis et al. (2020), if this is so, it should be explicitly stated, and the noise model identified.

Thank you for raising this point. An instrument noise model does not need to be specified to perform the methane retrieval. However, for the purpose of assessing retrieval precision, we assume that the instrument noise is normally distributed and uncorrelated. We now state this in the text on L. 203-206.

We estimate single-pixel column retrieval precision as the standard deviation of non-plume methane enhancements across the scene, assuming uncorrelated and normally-distributed instrument noise. We obtain precisions of 0.32 mol m^{-2} (roughly 49% of background) for the SBMP retrieval and 0.31 mol m^{-2} (roughly 48% of background) for the MBSP retrieval.

2. Further to this, do Sentinels-2A and B exhibit different behaviour due to differing instruments, and different noise levels?

We now explain that the fractional signal change is different between the two instruments due to their different spectral ranges (L. 153-154).

The difference between these two values is due to the instruments' slightly different spectral ranges (Figure 1).

3. The paper would benefit from some more validation, understandably this is challenging due to minimal measurements at this resolution (both spatially and temporally), and while there are some comparisons with GHGSat-D in this study (which show good results), I would have liked to have seen some TROPOMI retrievals over the test cases. These comparisons would distinctly show how TROPOMI cannot identify these plumes (or maybe they can), while Sentinel-2 can.

We explain in the introduction that TROPOMI spatial resolution is too coarse to resolve individual point sources except in special cases (e.g., Pandey et al., 2019), and have demonstrated this in a previous publication (Varon et al., 2019). We now direct the reader to this previous work in our discussion of TROPOMI point-source detection capabilities (L. 37-38).

The TROPOspheric Monitoring Instrument (TROPOMI) aboard the Sentinel-5 Precursor satellite provides daily global methane measurements at up to 5.5×7 km^2 pixel resolution (Hu et al., 2018; Schneising et al., 2019), sufficient to detect major accidental blowouts at oil/gas facilities (Pandey et al., 2019), but generally too coarse to resolve point sources

(Varon et al., 2019), which are often spatially clustered and typically produce plumes <1 km in scale (Frankenberg et al., 2016; Duren et al., 2019).

4. The authors briefly state at the beginning of page 10 that only unusually strong methane sources are detectable using the methods identified in this paper (~>3 t /hr). However, there is very little detail in the paper about how common such sources are and what they might be. The statement that most large sources fall between 1-10 t/hr is difficult to conceptualise without context. Therefore the overall utility of this method with Sentinel-2 remains unclear, and this should be improved. I recommend including a section placing the sources in the case studies in this paper in context with other global sources e.g. biomass burning events, or coal mining emissions, both significant methane sources (Saunois et al., 2020), therefore giving readers an idea of how the work in this paper could be applied globally.

Thank you for raising this important point. We elaborate on the significance of methane point sources of this magnitude (L. 266-272).

The largest methane point sources under normal operating conditions (landfills, wastewater treatment plants, and the vents of underground coal mines) can emit 1-10 t h^{-1} as annual means (Jacob et al., 2016; Varon et al., 2020; Scarpelli et al., 2020a). Our detection-limit estimates for Sentinel-2 thus restrict application to unusually strong sources, but such sources are routinely detected in abnormal operations (Pandey et al., 2019; Varon et al., 2019; Cusworth et al., 2020). Duren et al. (2019) found that sources stronger than 2.5 t h^{-1} accounted for <5% of point source emissions in California, but Frankenberg et al. (2016) found that they were responsible for more than 25% in the Four Corners oil/gas production region of New Mexico.

5. P2 L 36 - (Schneising et al., 2020) should be identified here as well.

Schneising et al. (2020) investigate methane emissions from oil/gas basins using TROPOMI observations. This is not directly relevant to the passage in question, which is about point sources, but we realize that the work of Schneising et al. (2019) on TROPOMI methane retrievals should be acknowledged here (in addition to the Hu et al. study). Thank you for bringing this to our attention. We have added the Schneising et al. (2019) citation to L. 36 and to the reference list.

The TROPOspheric Monitoring Instrument (TROPOMI) aboard the Sentinel-5 Precursor satellite provides daily global methane measurements at up to 5.5×7 km^2 pixel resolution (Hu et al., 2018; Schneising et al., 2019)…

6. P2 L45 – It'd be useful to identify the main aims of Sentinel-2 in this section i.e. identifying land changes etc. This would help differentiate the scope of this study from the main aims of the Sentinel-2 mission.

Thanks for this suggestion. We now briefly discuss the original aims of the Sentinel-2 mission in the introduction (L. 47-49).

Here we demonstrate the capability of the current Sentinel-2 twin satellites to detect and quantify strong methane point sources globally with both fine pixel resolution and frequent revisits. Sentinel-2 was originally designed to provide operational data products for environmental risk management, land cover classification, land change detection, and terrestrial mapping, as a complement to the Landsat and SPOT satellite missions. It comprises two satellites positioned 180° out of phase in the same sun-synchronous orbit, with an equator-crossing time of 10:30 (local solar time) at the descending node. Sentinel-2A (S2A) was launched in June 2015 and Sentinel-2B (S2B) in March 2017.

7. P2 L53 – While 2300 nm does include significantly more methane spectral lines than 1600 nm, solar irradiance is several times higher at 1600 nm, and surface reflectance is typically higher at 1600 nm. Therefore, one has to be careful about interpreting the next statement '2300 nm being considerably more sensitive to methane than 1600 nm'. I recommend this section be qualified with some statement about solar irradiance.

Thank you for raising this point. Differences in solar irradiance between bands are not a concern because Sentinel-2 bands 11 and 12 have identical SNR of 100 at their respective reference radiances. Therefore we are only concerned with reflectance differences between the bands. If the reflectance in a given band is systematically higher than in another, this is accounted for by our factor 'c'. We clarify in the text on L. 57-59.

Band 12, overlapping with the stronger and broader 2300-nm feature, is considerably more sensitive to methane than band 11. This is despite the comparatively lower solar irradiance in band 12, because the two bands have equivalent signal-to-noise ratios (SNR) at their respective reference radiances (Drusch et al., 2012).

8. P2 L54 – "Band 11 can therefore be used as a proxy for the continuum". It is not clear what is meant here? Surface reflectance values can still be significantly different between Band 11 and 12, especially in high albedo scenes. This combined with higher solar irradiance in band 11, indicate that there are significant differences between the bands.

Please see our response to comment #7.

9. P3 L67 – Are there any upcoming instruments that this method could be applied to?

We add that our methods can be applied to the Sentinel-3 SLSTR instrument as well.

Our techniques are developed with a focus on Sentinel-2 satellite observations, owing to the exceptional spatial and temporal resolution of MSI data, but can easily be extended to observations from other multispectral surface imagers with similar spectral bands, such as the Landsat 7 and Landsat 8 instruments with 30-m pixel resolution and a combined 8-day revisit rate, or the Sentinel-3 SLSTR instrument with 500-m pixel resolution and daily revisits.

10. P4 L93 – The risk of artefacts due to water vapour lines is identified here, but artefacts due to spectroscopic database uncertainties have not been identified. Since 2300 nm is very complex

spectrally, it is likely retrieval artefacts will exist, especially over high reflectance environments where the lines can saturate. Therefore, I recommend to include a short discussion here (or elsewhere as appropriate) on the potential impact of spectroscopic parameter uncertainty.

Thank you for this suggestion. We now note the possibility of quantification errors from database uncertainty (L. 104-105).

Additional quantification errors may arise from spectroscopic database uncertainties.

11. P4 L100 – Here the concept of residual radiance analysis is introduced, the core of the analysis of this paper. The sentence reads a little bit as though this is a new technique, which it is not. I recommend that a discussion or identification of past uses of methane detection with residual radiance be included here e.g. (Leifer et al., 2006; Roberts et al., 2010).

Thank you for making this important point. We now mention these previous studies and another one by Innocenti et al. (2017) that used similar techniques (L. 115-117).

But one can also in principle retrieve methane column concentrations together with surface albedo from just two spectral measurements, one featuring methane absorption and one not. This could be done for a single spectral band by comparing observations of the same scene with and without a methane plume. It could also be done for a single scene with two adjacent spectral bands that are sufficiently close to have similar surface and aerosol reflectance properties, but differ in their methane absorption properties. Similar techniques have previously been used to retrieve methane column concentrations from ground-based (Innocenti et al., 2017) and airborne (Leifer et al., 2006; Roberts et al., 2010) remote sensing instruments. We demonstrate this here using Sentinel-2 bands 11 and 12.

12. P4 L105 – Here a radiative transfer model is briefly described, can access to this model be provided? Fundamentally, the results shown in the paper need to be reproducible, and the RTM is key in this regard.

We will make the radiative transfer model code available upon request and have added a Code Availability section stating this (L. 469-470).

13. Further, no mention of the model used for surface reflectance is identified which is surprising given the importance of reflectance in this study. Assuming the same model used in Jervis et al. (2020) is employed, is a Lambertian model sufficient at such high spatial resolution? Would not a BRDF model yield improved results, possibly dealing with some of the heterogeneous scenes?

We agree with the reviewer that the surface reflectance model should be explicitly mentioned. This information has been added to L. 125. The scaling factor c will automatically account for any angle-dependent reflectance effects in the SBMP retrieval method. For the MBSP method, the reflectance measurements of different bands are at

the same observation and thus scattering angle. For these reasons, we don't believe that incorporating a BRDF reflectance model will improve results.

The incident solar irradiance is from Clough et al. (2005) and we assume the surface reflectance to be Lambertian.

14. P4 L110 – Presumably the HITRAN2016 database?

Fixed, thanks.

15. P5 L115 – The assumption about aerosols is not well justified here, especially given the location of the case studies in this paper are in desert regions, well known to be affected by aerosols.

Aerosols in desert scenes do not contradict the justifications we provide, but to clarify we have rephrased and expanded on this justification (L. 128-131).

Aerosol effects are ignored because methane sources generally do not co-emit aerosols, and because background aerosol such as from dust can be assumed uniform across a given scene (like water vapour and CO_2). Neglecting aerosol scattering may produce methane retrieval errors of a few percent (Huang et al., 2020), but as we show below this is much smaller than typical Sentinel-2 methane retrieval errors.

16. P5 Equation 2 – It is unclear as to how instrument noise in band 12 is accounted for in this calculation. The radiances used in equation 1 will include instrument noise, so there should be some accounting for this in equation 2?

Please see our response to comment #1.

17. P5 L136 – Are the differences between Sentinel-2A and 2B purely due to the spectral range differences?

Please see our response to comment #2.

18. P5 L141-142 – It is very unclear what the authors have undertaken here with assessing the impact of water vapour. It is stated that variations in background water vapour have a minor impact of 6% (although how does a 6% variation affect the precision). But what background is being used in this test? Are the water vapour variations the max that could be seen in the US standard atmosphere, or based off the max in a tropical scene?

The range of water vapor column abundances used reflects the range observed by Nelson et al. (2016) over land using OCO-2. We now state this explicitly (L. 158-160).

Water vapour columns over land may vary from 1 to 40 kg m^{-2}, as observed by the Orbiting Carbon Observatory 2 (OCO-2; Nelson et al., 2016), but under nominal observing conditions m_{SBMP} varies by only 6% over this range (i.e., by roughly $\pm$ 0.002).

19. P6 L149 – I'm not convinced that relying on similar radiance levels between the spectral bands works as suggested in this section. Even if surface reflectances are similar (which they may not be), solar irradiance and instrument noise will likely mean different SNRs between band-12 and band-11, yielding large magnitude differences. Can this all be accounted for in the 'c' factor in equation 3?

We discuss in Section 3.5 the consequences of surface reflectance differences between bands; if the differences are too large, the retrieval becomes too noisy to resolve all but the most anomalous plumes. We address the question of solar irradiance and SNR differences between bands in our response to comment #7. Our retrieval demonstration in Section 3.4 and our documentation of precision levels in Figures 3, 5, and 7 show that the 'c' factor can indeed account for broad differences between spectral bands and satellite passes.

20. P8 L197 – It is stated that 'c' is used to account for scene-averaged differences in reflectances between satellite passes or bands. However earlier in the text (P6 L150), 'c' is identified as being used to account for calibration differences, implying minor variations. This line should be moved up to p6 to give more detail about the use of 'c'.

The parameter 'c' can be interpreted differently in the SBMP and MBSP methods. In the former, it accounts for scene-wide differences between passes; in the latter, scene-wide differences between bands. We explain on L. 170-172 that in the MBSP method 'c' addresses both calibration differences *and* spectral dependences of the albedo. For clarity we now also elaborate on the meaning of 'c' in the SBMP method (L. 143-144; see our response to Reviewer #2 comment #4). Our statement that 'c' accounts "for scene-averaged differences in reflectance between satellite passes or bands" (L. 219-221) covers both the SBMP and MBSP methods.

21. P9 L218 – The term 'plume detection limit' is not explicitly identified, please explain this term and how it is calculated. This term is also used in Figure 3, with no explanation as to what this is or how it is calculated.

We explain how these values are obtained in Section 3.5 (L. 261-266) and in the caption of Figure 3. We have rephrased our explanation for clarity.

The single-pixel column retrieval precisions of Figure 3 can be related to empirical plume detection limits—the smallest source rate Sentinel-2 can detect in a given scene—by using the long-term observation records for the Hassi Messaoud and Korpezhe scenes shown in Section 4. For Hassi Messaoud with 27% retrieval precision, the minimum detected source is 2.6 t h^{-1}. For Korpezhe with 33% retrieval precision but more temporally variable surface conditions, the minimum detected source is 3.5 t h^{-1}. The plume detection limit is expected to be proportional to precision (Jacob et al., 2016) and is shown in Figure 3 for the savannah, urban, and farmland scenes by scaling of the Hassi Messaoud MBMP value.

22. P9 L227 – This MBMP precision is still significantly worse than the homogeneous scenes, is this still due to reflectance errors?

> Yes, and also because this surface is darker than the Algerian and Turkmen scenes. For example, comparing to the Hassi Messaoud scene, the farmland scene is 4 times darker in band 11 and 9 times darker in band 12 (mean reflectances on L. 239-241).

23. P10 L221 – This section should go in the conclusions.

> We assume the reviewer is referring to the following passage, beginning around L. 241: "Our detection-limit estimates for Sentinel-2 thus restrict application to unusually strong sources, but such sources are frequently detected in abnormal operations (Pandey et al., 2019; Varon et al., 2019; Cusworth et al., 2020)." We now reiterate this point in the conclusions (L. 441-443).

> For 27% precision we estimated a plume detection limit of 2.6 t h^{-1}. This is at the high end of methane point sources in normal operating conditions, but can detect anomalously large point sources that contribute disproportionately to emissions from a given region or sector and have been routinely observed by satellite (Pandey et al., 2019; Varon et al., 2019; Cusworth et al., 2020).

24. P10 L253 – With regards to TROPOMI, it'd be useful to contrast the results of TROPOMI with those found by the methods shown in this paper. Therefore providing direct proof of the utility of this method with Sentinel 2, if such data is available at this time.

> Please see our response to comment #3.

25. P2 L35 – TROPOMI now operates at a spatial resolution of 3.5x5.5 km2.

> That is true for e.g., the NO_2 retrieval, but not for methane. The nominal resolution of the TROPOMI methane product is 5.5x7 km^2.

26. P4 L88 – HITRAN -> HITRAN2016

> Fixed, thanks.