Response to the reviewers on the manuscript "Methane retrieved from TROPOMI: improvement of the data product and validation of the first two years of measurements" by Alba Lorente et al.

The authors would like to thank the reviewers for their thoughtful and helpful comments and suggestions. Below are the comments by the reviewers in blue and replies in black. Any modification made to the text has been underlined. The line and page numbers correspond to the version of the manuscript available for online discussion.

Reviewer 2

Comment C 2.1 — Page 4, line 5. How do you determine the position of the 12 pressure layers? Are they fixed for every scene or are they calculated with respect to the surface pressure or tropopause height? If the location of the tropopause isn't accounted for in the construction of vertical layers, have you considered the uncertainty this could cause in calculating the total column XCH_4, compared to a method which aims to put a pressure layer boundary at the tropopause height?

Reply: The equidistant pressure layers are determined by the surface pressure and the top of atmosphere from the meteorological input, constrained by a value of 0.1 hPa. So the grid differs per retrieval, and during a single retrieval it remains fixed as the algorithm does not retrieve surface pressure. For the a priori vertical profile we use TM5 that varies with latitude, longitude and altitude also accounting for the effect of tropopause height variations.

Comment C 2.2 — Page 4, line 12. Could you please be more specific in which ECMWF data you are using. Is it ERA-5 or ERA-Interim for example.

Reply: The ECMWF data that we use is an operational analysis product; it corresponds to the first analysis performed after the forecast product, so it is not a reanalysis product as ERA-5 or ERA-Interim. Access to this ECMWF product is granted to us on behalf of the TROPOMI project. We have added "operational analysis product" to the text to make it clearer.

Comment C 2.3 — Page 4, line 21. How well do the results of the back-up filter compare to the VIIRS cloud filter data when you try to use it for scenes where you do have VIIRS to validate it? How many scenes in total for your two years of data use the VIIRS cloud clearing method, and how many use the back-up H2O retrieval method?

Reply: The cloud filtering is important in our processing, so we acknowledge the referee bringing it up, and we hope to clarify it as the Referee #1 also raised a question on this topic.

The quality of our XCH_4 retrieval relies on a very strict cloud filtering, for which we use VIIRS data that is able to identify small-scale cloud structures that could lead to errors in the retrieval if not filtered properly. VIIRS data used in the TROPOMI XCH_4 retrieval is processed operationally by the S5P-NPP cloud processor. If due to any circumstance the processing of the VIIRS data fails or it is delayed, we use the filtering based on a non-scattering retrieval as a back-up option. In this circumstance, the XCH_4 data is flagged (qa value downgraded to 0.4) as the data might be affected by cloud contamination, because the non-scattering retrieval is not as effective filter as VIIRS data, particularly for thick clouds. Originally, the filter based on the the non-scattering retrieval was optimized to filter cirrus over dark surfaces (Hasekamp et al., 2019), by applying a 6% and 22% threshold for the difference between the CH_4 and H_2O retrieved in the weak and strong absorption bands. This filter alone will effectively remove scenes with a cloud fraction higher than 15%, a fraction that is too high to keep the errors in the retrieved CH_4 below requirements. Together with the scattering filter (using the retrieved scattering parameters) scenes with a cloud fraction higher than 8% will be effectively filtered, but still far from the desired 1-2% for the CH_4 retrieval (Hasekamp et al., 2019). These numbers presented here correspond to the analysis made prior to launch, that need to be repeated using real data.

As VIIRS data is operationally processed, it is rarely missing or not available for its use in the XCH_4 retrieval. From all the orbits processed operationally since the beginning of the mission, for less than 1% the processing of VIIRS data was not nominal in the CH_4 retrieval. We added the following to stress this point: "In less than 1% of the cases when VIIRS data is not available, we use a back-up filter based on a non-scattering H_2O and CH_4 retrieval from the weak and strong absorption bands (Hu et al., 2016). These cases are flagged accordingly by the quality value indicator."

Comment C 2.4 — Page 4, lines 24-26. I am a little confused by how you cite a paper from 2019 (Hasekamp et al. 2019) to say that results of version 1.2.0 from June 2020 of your algorithm largely comply with mission requirements. Please could you elaborate on this.

Reply:

Hasekamp et al. (2019) is the reference to the ATBD for the operational algorithm version 1.2.0 mentioned in that sentence. We agree with the reviewer that this might be confusing for the reader, so we remove "as of June 2020". We wanted to specify the version of the operational algorithm when the manuscript was written/submitted (and that is why we added "as of June 2020"), having in mind that this version could have changed in the meantime. But we acknowledge that with the version number it should be sufficient to trace it back.

Comment C 2.5 — Page 4, line 27. You call this new version the beta version here, but do not refer to this again. However, on page 6, line 20 you say the updates to the algorithm correspond to v1.3.0. Is there a difference between this beta version and 1.3.0? If not then it might be clearer to call it v 1.3.0 here on page 4.

Reply: We acknowledge that the naming and version numbers might led to confusion when reading it, a remark also made by Referee #1 in comment C1.6. We try to make it clearer through the manuscript.

The reference to *beta version of the TROPOMI XCH_4 data product* is used for the data product that results from the scientific development activities within the L2 team at SRON. The next step for these developments is to be implemented in the operational processing whenever there is a a processor update.

We have removed the reference to version 1.3.0 that will eventually correspond to the future operational update, because this specific numbering is not certain as of now, and only causes confusion. Now page 6, line 20 reads: "The TROPOMI XCH_4 scientific data product from SRON retrieved with the updated algorithm will be suggested for use in the operational processing in the next processor update."

Comment C 2.6 — Page 5, line 8. Could you please comment on why you chose to use daily averaged TCCON instead of averaging only data which is within a shorter time frame. I understand that TROPOMI has 14 orbits in one day so I would assume it likely that more than one orbit may intersect the 600km diameter co-location criteria. Do you think there is merit in being stricter in your temporal co-location as a result so you are only matching TCCON at a similar time to an overpass?

Reply: We are glad that this was brought up as we do limit the TCCON measurements to ± 2 hours of the TROPOMI overpass, so the explanation on the manuscript is wrong, and we have changed it accordingly.

The mistake on the text is because we performed sensitivity tests by also using daily averages. Figure R1 shows the validation results with time constraint (left, same as Fig. 8a in the original manuscript) and without any time constraint (right). The overall validation results do not change significantly. The mean bias does not change significantly, and the station to station variability is only affected by 1 ppb. The number of collocation pairs did increase significantly (from 3203 to 8351). As an example for the validation over one of the stations, Fig. R2 shows the time series of the bias with time constraint (left) and without any time constraint (right).

Comment C 2.7 — Page 6, line 4. I think it's potentially misleading to say that the GOSAT swath is 790km with a 10.5km resolution without saying that its measurement method is different to TROPOMI's and that it usually only makes 3 of those 10.5km measurements across its swath. I think an additional sentence here on the sampling pattern of GOSAT would be helpful.

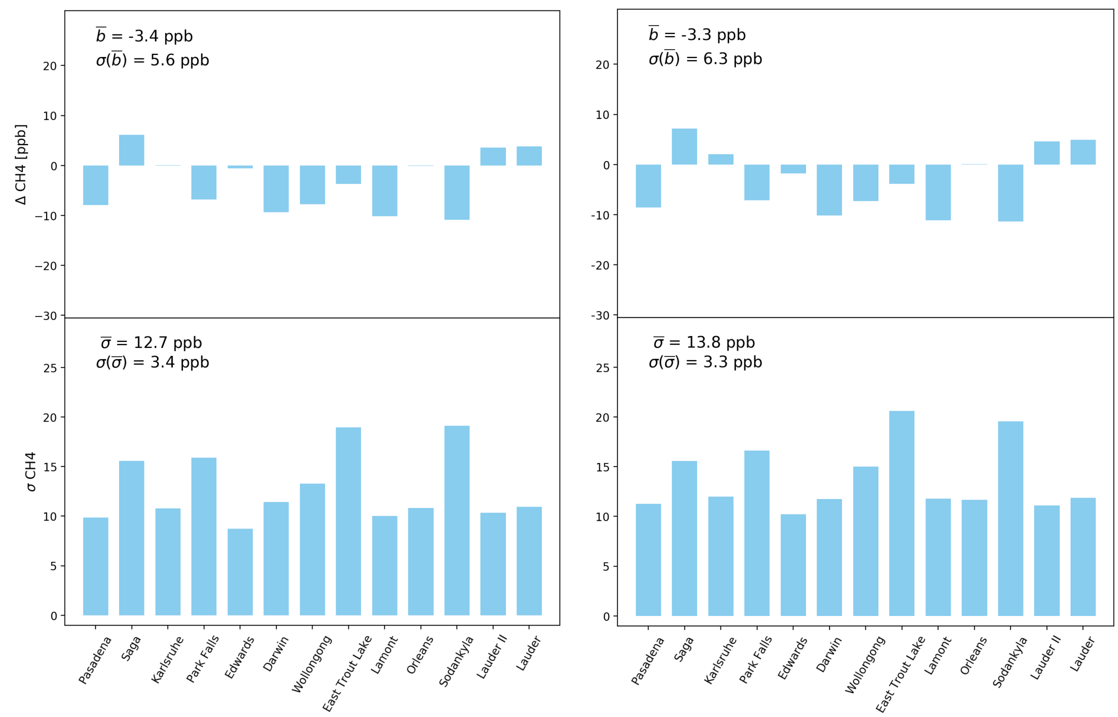

Figure R1: Mean differences between TROPOMI and TCCON XCH_4 (ΔXCH_4) and the standard deviation of the differences (σ_{XCH_4}) with (left) $\pm$ 2 hours of time constraint in TCCON (as in the manuscript) and (right) daily averages.

Reply: We have modified the sentence in page 6, line 4: "GOSAT was launched in 2009, and ~~with a~~ it performs three point observations in a cross-track swath of 790 km with 10.5 km resolution on the ground at nadir, which results in global coverage ~~is obtained~~ approximately every 3 days".

Comment C 2.8 — Page 6, line 16. You use a full-physics method for TROPOMI because the proxy method cannot be applied, and go on to say that the full-physics and proxy methods were found to perform similarly for GOSAT. What you don't explain in the paper is why you don't use the GOSAT full-physics data as this feels like a more natural comparison. Could you please comment on why you used gosat proxy over gosat full-physics?

Reply: The main reason to use the proxy product in this comparison is the fact that the data yield is higher. Furthermore, the comparison of TROPOMI XCH_4 and GOSAT with both approaches results in similar bias: mean bias of -10.3 ± 16.8 ppb and a Pearson's correlation coefficient of 0.85 with the proxy approach (as stated in the manuscript) and mean bias of -12.5 ± 14.9 ppb and

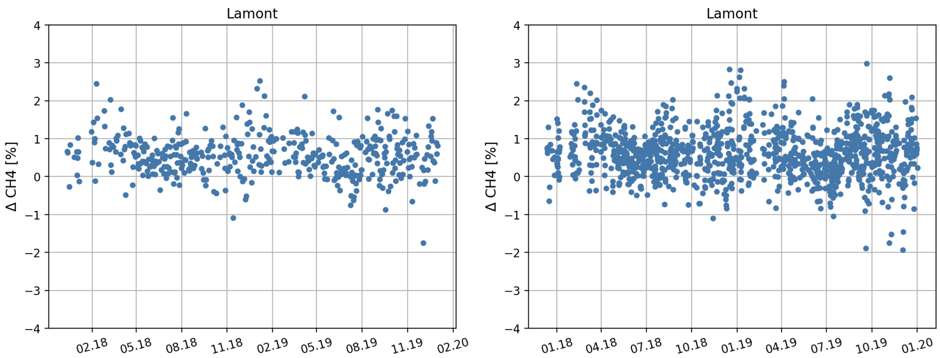

Figure R2: Time series of the bias between TROPOMI and TCCON XCH_4 over the Lamont station with time constraint (left) and without time constraint (right).

a Pearson's correlation coefficient of 0.86 with the full physics approach. The correlation plot of both comparisons is shown in Fig. R3, on the left for GOSAT proxy and on the right for GOSAT full-physics.

We have added this information to make clear the reason for the selection of the full-physics approach (page 6, line 15): "In the validation in Sect. 5 we found that there is no bias between the GOSAT proxy and full-physics products. However, we have selected for the comparison the GOSAT proxy product over the full-physics because of its higher data yield". And we also include the results for the full-physics in a sentence in page 17, line 7: "The overall comparison yields a mean bias of -12.5 ± 14.9 ppb if we use the GOSAT XCH_4 product retrieved with the full-physics approach".

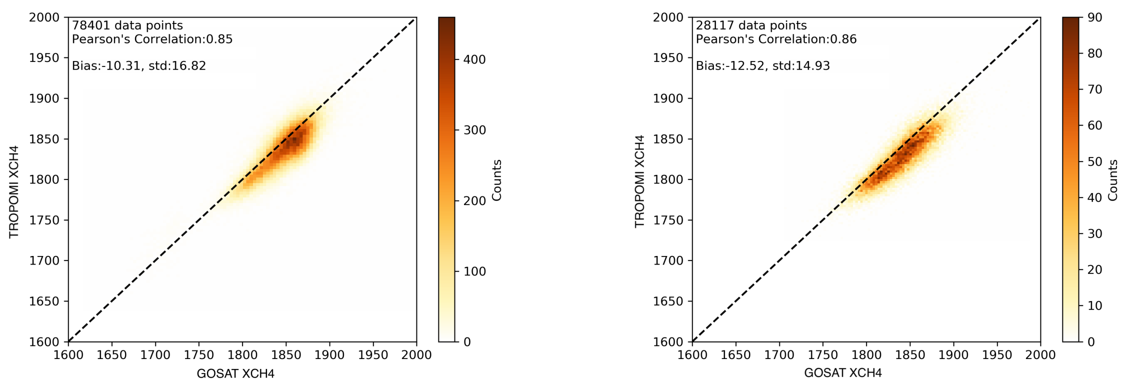

Figure R3: Correlation plot of TROPOMI XCH_4 and GOSAT XCH_4 retrieved with the proxy approach (left) and with the full physics (right). Daily collocations are averaged to a 2°x2° grid for the period 1 Jan 2018 – 31 Dec 2019.

Comment C 2.9 — On Section 3.1. It makes sense to me that using a constant gamma reduces the overall dispersion of the data, improving your results. But please can you comment on the theory behind why you think calculating gamma per iteration should result in a less accurate result than using an average value.

Reply: In theory calculating a gamma for each iteration should be actually superior than using an average gamma value. However, it is based on the idea of finding the minimum value of the elbow plots (χ^2 vs. regularization strength). Using real data, such a minimum does not exist in most cases and therefore can result in a more unstable inversion. We found using an average value for gamma results in a more stable retrieval and reduces the overall dispersion of the data.

Comment C 2.10 — Page 7, line 12. The reduction of 9% going from 19.7 ppb to 24.5 ppb doesn't make sense to me since it's becoming larger instead of reducing and the difference between these numbers is larger than 9%.

Reply: We appreciate the careful reading of the referee that led to spotting this typo. The reduction is from 21.5 ppb (and not 24.5) to 19.7 ppb, which corresponds to approximately 9%. We have corrected this.

Comment C 2.11 — Page 7, line 25. In your section on the TCCON validation you say that the overall bias with respect to HITRAN 2008 is +15.5 for HITRAN 2016. Table 2 shows that the difference between HITRAN 2008 and HITRAN 2016 is 20.3 ppb for TCCON.

Reply: We thank again for this careful check of the numbers. Indeed the bias from -2.4 ppb to 17.9 ppb is 20.3 and not 15.5 as it is written in the text (this is considering HITRAN 2008 bias as 2.4 ppb and not -2.4 ppb). We have corrected the text accordingly.

Comment C 2.12 — Page 7, line 31. Please could you give the global numbers as referred to here which show that SEOM-IAS has a significantly improved RMS and chi-squared over the other two.

Reply: We added these numbers to the text, page 7, line 31. ”Global mean χ^2 improves by 19% with SEOM-IAS cross-section and by 7% with HITRAN 2016 with respect to HITRAN 2008”.

Figure R4 shows the ratio of χ^2 of the retrieval with HITRAN 2008 and HITRAN 2016 (left) and HITRAN 2008 and SEOM-IAS (right), for one year of data averaged into daily 1° x 1° grid, which shows that SEOM-IAS cross section results in a significantly better χ^2 with respect to HITRAN 2008 and HITRAN 2016.

Comment C 2.13 — On section 3.3. I like the discussion on the differences of greater than 45m and 50m, but in figure 2 there are a lot of smaller systematic differences of 10-20m to be seen in

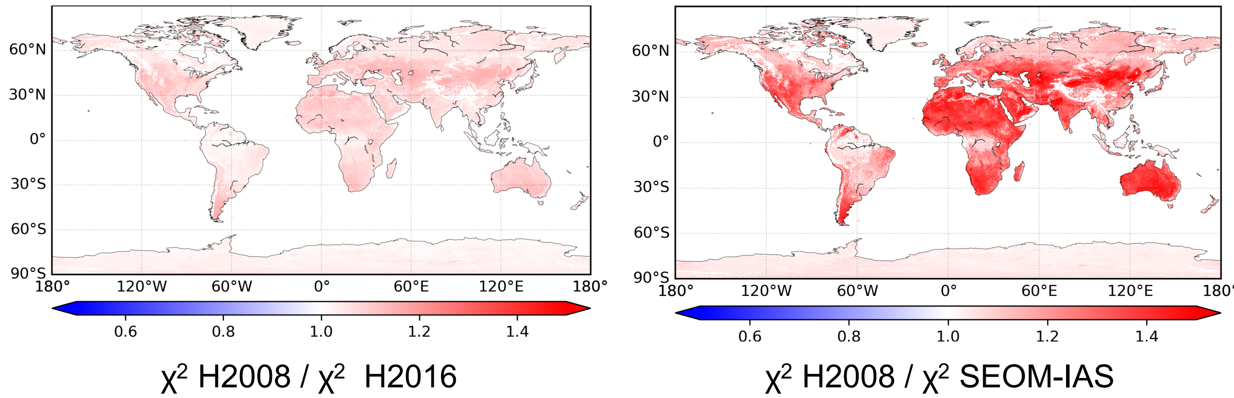

Figure R4: Ratio of χ^2 from the retrieval with HITRAN 2008 and HITRAN 2016 (left) and SEOM-IAS (right).

the Eastern US which lead to a net positive XCH4 difference over this half of the country. Firstly, please could you comment on why you think the higher resolution DEM would be on average higher elevation than the lower-res DEM over this region. And following on, could you please comment on the change to XCH4 overall as a result of any mean altitude difference between DEMs on a global scale (if one exists). I ask since you only focus on outliers between the DEMs in the paper and don't talk about any systematic differences.

Reply: The different East-West features over United States shown in Fig. 2 are only present in this region. As why on average the SRTM results higher in elevation over Eastern US we are not sure, but we assume that compared to the S5P-DEM, the SRTM database is a better representation on the terrain over the US as it is a database that uses national data and models. Globally, we see similar features as in the Western US in most of the mountain regions around the world, so there are not systematic differences. So overall XCH_4 changes are more pronounced over mountain regions, that is why we focused on the outliers in the discussion on Sect. 3.3.

Comment C 2.14 — On the Small Area Analysis. Could you please elaborate on how and why you chose the areas which you did. How dependent on your method is the choice of SAAs.

Reply: The reasoning for the choice of the specific areas used in the SAA analysis was mainly to have a representation of the challenging scenes for the XCH_4 retrieval, mainly low and high albedo. The areas also needed to include scenes with surface albedo around the reference value, and not include (as much as possible) big sources of methane, although this was less of a limiting factor because XCH_4 distribution is normalized for each region separately. Furthermore, we aimed at areas that had a relatively good coverage through all the different seasons, and we stayed away from big mountain regions.

For high albedo scenes it was straightforward to chose Sahara desert, and over this area we tested the choice of multiple regions and their size. The main challenge was to find regions that included the surface albedo reference value, and all the areas that we tested resulted in similar XCH_4 dependencies. Over Australia we made boxes of 5°x 5° and discarded those that had in the same box different modes in the XCH_4 distribution with respect to surface albedo. The most southern box was interesting because it includes low surface albedo values with strong XCH_4 underestimation, and the shape of this area is different to 5°x5° to avoid the location of strong XCH_4 sources as present in the EDGAR inventory (it is a region with an important oil and gas industry). Then areas over Canada were chosen because it represents the strong XCH_4 underestimation related to the low surface albedo values in these high latitudes, also present in Northern part of Europe and Russia. Adding areas of Russia did not change the dependence and the fit (Fig. 4) made to derive the correction.

Comment C 2.15 — On the bias correction method, page 10, line 1. You only apply a bias correction on the surface albedo and say that the other retrieved parameters show negligible dependence, showing surface albedo, AOD and SZA. For OCO-2 the parameter dP (the difference between the a priori and retrieved surface pressure) shows the largest dependence on the bias. Is this a parameter you have looked in to?

Reply:

In the TROPOMI XCH_4 retrieval we do not retrieve surface pressure. In Fig. 3 in the manuscript we show surface albedo, AOD and SZA as an example, but dependencies in other parameters such as χ^2, column of interfering absorbers H_2O and CO, retrieved aerosol parameters (aerosol size, altitude of aerosol distribution and aerosol column) were also investigated. Besides the fact that we also tried to use as few correction parameters as possible, all the other parameters showed negligible dependence compared to that on surface albedo. We have specified that SZA and AOD are examples on page 10, line 3 to avoid misunderstanding.

Comment C 2.16 — Page 16. Lines 5-6. Have you tried comparing with snow cover data to verify how reliable this method of detecting snow actually is?

Reply: We have not done that comparison ourselves, but we are in contact with colleagues from the Finish Meteorological Institute (FMI) to investigate the seasonality on the bias further. They have found a significant correlation between the seasonality of the bias and the presence of snow surface at Sodankylä, and as the blended albedo is as well correlated to this seasonality, it is suitable to use it to filter this complex scenes, but it is not aimed as an accurate method to actually detect snow. We specify this on page 17, line 1: "By applying it [...] a threshold value of 0.85 is optimal to remove these scenes <u>that cause the seasonality on the bias</u>". However, we do not apply this filter ourselves in an operational mode, as the source of these seasonality of the TROPOMI-TCCON bias is still unknown.

The correlation with snow surface found by FMI does not necessarily always imply the presence of vortex air which was our first hypothesis to explain the seasonality of the bias. We are also investigating the role of the prior profile in this specific retrieval scenarios, assuming that there might be cases with a strong depletion of XCH_4 in the upper troposphere (due mainly to vortex air) that impact our retrieval (and/or TCCON) if it is not captured properly by the prior. Furthermore, the different sensitivities between TCCON and TROPOMI might also play a role in the satellite and ground based comparison. All these effects we think need to be taken into account when making conclusions out of the validation results.

Comment C 2.17 — Page 5, Table 1. Caption missing versions for the instrument.

Reply: We assume that this refers to the typo as both instruments are referred as "ll" in the caption of Table 1. The instrument "ll" (Sherlock et al., 2017) was replaced by "lr" (Pollard et al., 2019). We have corrected this.

Comment C 2.18 — Page 10, line 7. Typo with full stop. I assume you wanted a capital T or a semicolon.

Reply: Corrected.

Comment C 2.19 — There are multiple instances throughout the paper where you've misspelt ppb as pbb. Such as Page 7 line 25, page 9 line 8, page 20 line 16 and twice on page 14 line 4.

Reply: We thank the referee for spotting this. We have changed it.