The authors would like to thank the editor and the reviewers for their precious time and invaluable comments. The corresponding changes and refinements are highlighted in yellow in the revised paper and are also summarized in our responses below. Authors' responses are in blue. Reviewers' comments are in black. When the manuscript is cited, it is shown in italics.

Referee #1

This study builds a 0.1° × 0.1° daily XCH_4 product covering 2020–2023. The authors do this by bias-correcting GOSAT-2 with TCCON as reference, then bias-correcting GOSAT and TROPOMI with the TCCON-informed GOSAT-2 product as reference, and then finally filling each daily 0.1° × 0.1° grid cell with the bias-corrected data from the three sensors, giving priority to GOSAT-2, then TROPOMI, then GOSAT. This fused product performs well against withheld TCCON data and provides additional coverage past that of TROPOMI in challenging retrieval environments. The presentation quality is excellent, but I encourage the authors to consider the following comments at their discretion to improve the scientific quality and significance.

Major Comments:

1. While TCCON is an excellent reference dataset, there are downsides to using it as the basis of your bias correction (e.g., limited independent validation data remains). Do the GOSAT-2 and TCCON co-locations represent diverse enough conditions (e.g., in surface albedo) to justify its use as reference? Figure 6 shows a small range of surface albedos relative to the global distribution (cf. Figure 6 in Balasus et al., 2023). In the case of TROPOMI, a small-area approximation can be used to generate data with more diverse retrieval conditions for reference purposes, either training or validation (Lorente et al., 2021).

   ➔ Thank you for this insightful comment. We agree with the reviewer's opinion. Although TCCON provides a high-precision reference, its spatial distribution is limited and therefore cannot fully represent the diverse retrieval conditions encountered globally. In particular, most TCCON sites used in this study are located in areas with SWIR surface albedo below approximately 0.4 (Fig. 1). We therefore clarified this limitation of TCCON-based bias correction in the revised manuscript and performed additional analyses to assess the dependency on surface albedo.

   Using the TCCON co-location dataset (Fig. R1), the standard TROPOMI product showed a strong negative dependency on surface albedo, with R = -0.49 (Fig. R1(a)). The operational bias-corrected product, which applies the a posteriori albedo-bias correction described by Lorente et al. (2021), reduced this dependency, but a residual dependency remained, with R = -0.17 (Fig. R1(b)). In contrast, machine learning (ML)-based bias correction reduced this dependency to R = -0.02. For GOSAT and GOSAT-2, the standard products already showed weaker surface albedo dependencies than TROPOMI, and the dependencies remained low after ML-based correction, with R = -0.03 and R = -0.04, respectively (Fig. R1(c)–(f)). These results indicate that, within the albedo range represented by the TCCON sites, the ML-based bias correction effectively reduced surface-albedo-related bias.

To further evaluate the stability of the final target (i.e. TCCON-based bias-corrected GOSAT-2), under surface albedo conditions broader than those sampled by TCCON sites, we performed an additional satellite match-up analysis between GOSAT bias-corrected product and TCCON-based bias-corrected GOSAT-2 (Fig. R2). Because the GOSAT retrievals are relatively less sensitive to surface albedo and has been used as a reference or validation dataset in previous studies (Lorente et al., 2021; Balasus et al., 2023; Li et al., 2024; Fan et al., 2024), we used it as a complementary reference to evaluate the stability of the final target under broader albedo conditions. This analysis covered a surface albedo range of approximately 0–0.8. The TCCON-based bias-corrected GOSAT-2 showed a weaker dependency than the standard GOSAT-2 product, with R decreasing from -0.20 to -0.12.

We also evaluated the surface albedo dependency of TROPOMI harmonized to this target scale (Fig. R3). The dependencies observed in the standard and operational bias-corrected TROPOMI products were almost removed by applying ML-based harmonization. This result indicates that harmonization to the TCCON-based bias-corrected GOSAT-2 target scale effectively mitigated the surface-albedo-related bias in TROPOMI.

Therefore, although these additional analyses do not eliminate the intrinsic limitation of the TCCON-based approach, the site cross-validation within the TCCON albedo range and the satellite match-up analysis over a broader surface albedo range together provide complementary evidence that the TCCON-based bias-corrected GOSAT-2 using ML can serve as a reasonable common reference scale for multi-sensor harmonization.

Lines 309-312: "*This inter-sensor harmonization further assessed the major retrieval-parameter dependencies identified in the TCCON-based analysis over broader parameter ranges (Fig. 8). After ML-based harmonization, the SWIR surface albedo dependency of TROPOMI and the ΔPs dependency of GOSAT were nearly removed, indicating improved consistency with the ML-based bias-corrected GOSAT-2 scale.*"

Lines 454-459: *"First, the ML-based bias correction relies on TCCON as the primary reference dataset. Although TCCON provides high-precision ground-based XCH_4 observations, its spatial distribution is limited and does not fully represent the diverse retrieval conditions encountered globally, particularly high-surface-albedo conditions. We partly addressed this limitation through LOSOCV strategy and additional satellite match-up analyses over broader surface albedo ranges, but independent validation under underrepresented retrieval conditions remains necessary."*

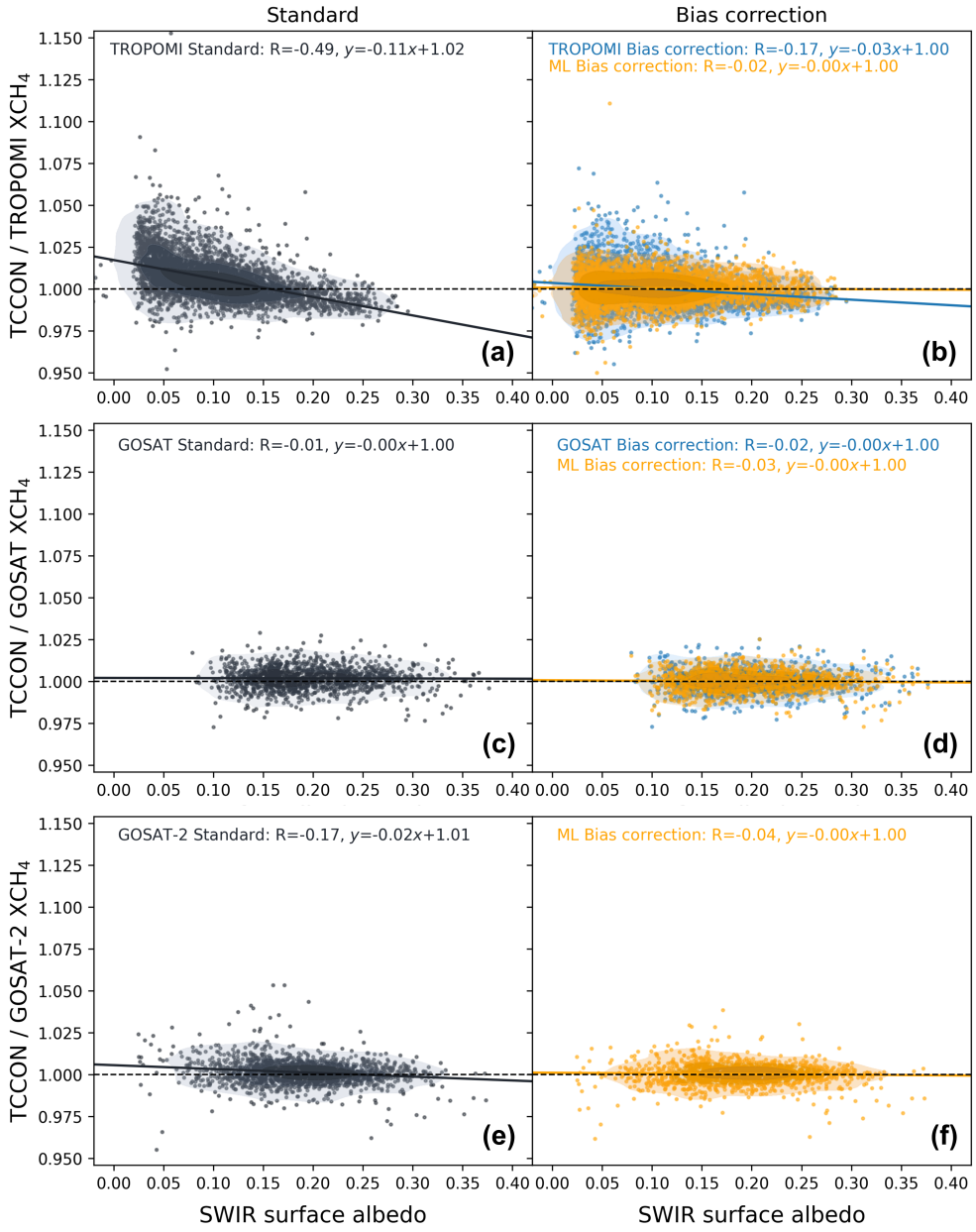

Figure R1. SWIR surface albedo dependency of satellite XCH_4 products relative to TCCON XCH_4. Scatter density plots show the relationship between SWIR surface albedo and the ratio of satellite XCH_4 (TROPOMI, GOSAT, and GOSAT-2) to TCCON XCH_4. For each sensor, the left panel (a, c, e) shows the standard product, and the right panel (b, d, f) shows the bias-corrected product, including the operational bias-corrected product and the machine learning (ML)-based bias-corrected results for all sensors. The ML-based bias-corrected results are based on leave-one-site-out cross-validation, in which the validation site was excluded from model training.

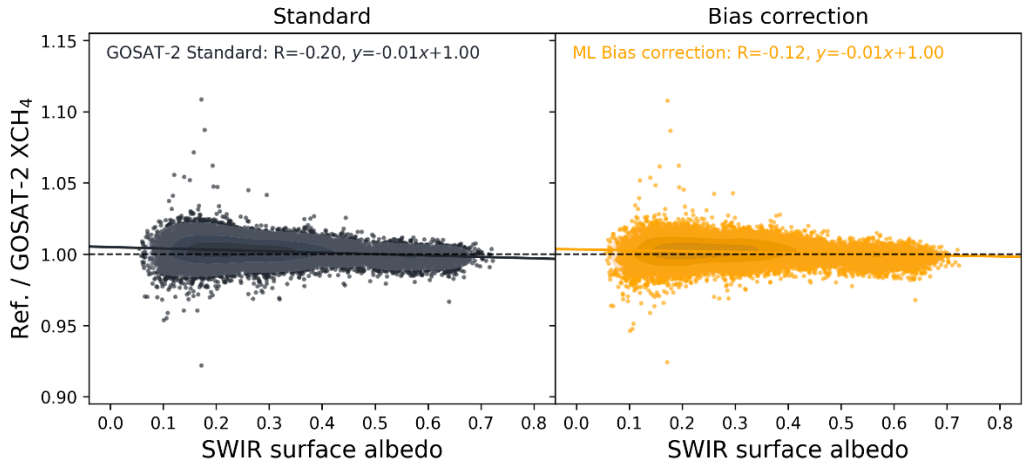

Figure R2. SWIR surface albedo dependency of GOSAT-2 XCH_4 relative to GOSAT bias-corrected XCH_4 (Reference). Scatter density plots show the ratio of GOSAT-2 XCH_4 to GOSAT bias-corrected XCH_4 as a function of SWIR surface albedo using GOSAT–GOSAT-2 satellite co-location samples.

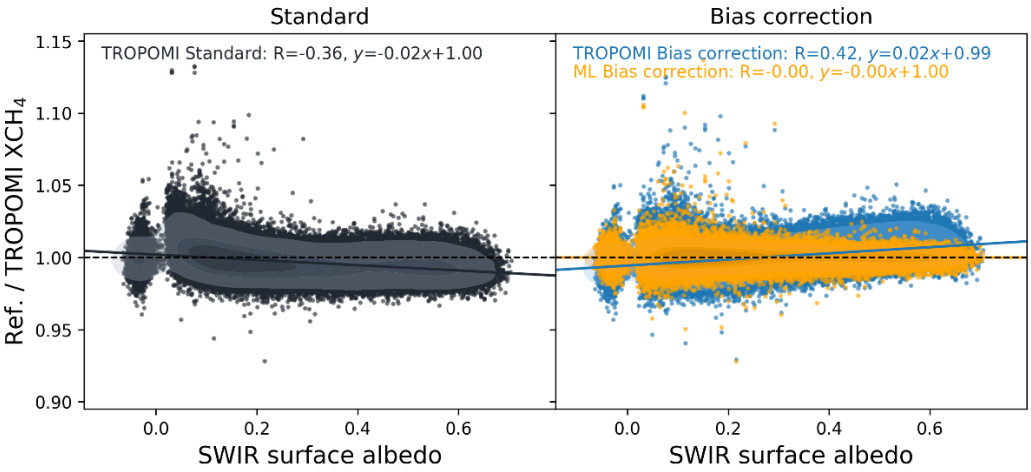

Figure R3. SWIR surface albedo dependency of TROPOMI XCH_4 relative to the final harmonization target. Scatter density plots show the ratio of TROPOMI XCH_4 to the final target, defined as the ML-based bias-corrected GOSAT-2 XCH_4, as a function of TROPOMI SWIR surface albedo. The comparison among the standard, operational bias-corrected, and ML-based harmonized TROPOMI products.

References:

Lorente, A., Borsdorff, T., Martinez-Velarte, M. C., & Landgraf, J. (2023). Accounting for surface reflectance spectral features in TROPOMI methane retrievals. *Atmospheric Measurement Techniques*, *16*(6), 1597-1608.

Balasus, N., Jacob, D. J., Lorente, A., Maasakkers, J. D., Parker, R. J., Boesch, H., ... & Varon, D. J. (2023). A blended TROPOMI+ GOSAT satellite data product for atmospheric methane using machine learning to correct retrieval biases. *Atmospheric Measurement Techniques*, *16*(16), 3787-3807.

Li, K., Bai, K., Jiao, P., Chen, H., He, H., Shao, L., ... & Chang, N. B. (2024). Developing unbiased estimation of atmospheric methane via machine learning and multiobjective programming based on TROPOMI and GOSAT data. *Remote Sensing of Environment*, *304*,

114039.

Fan, L., Wan, Y., & Dai, Y. (2024). Development of a multi-source satellite fusion method for XCH4 product generation in oil and gas production areas. *Applied Sciences*, *14*(23), 11100.

2. The authors have focused mostly on the delivery of the 0.1° × 0.1° product as the outcome of the study (as opposed to, for example, what their bias corrections tell us about shortcomings in any of the retrievals). Could the authors suggest some studies that could benefit from their data? In many cases, XCH_4 data alone is insufficient. In inverse modeling studies using chemical transport models, users would also need e.g. averaging kernel, prior profile, and pressure grid information (and might want data from all available sensors, not just the best one, in order to reduce random errors), though inverse modeling studies are not the only application of XCH_4 data.

➔ Thank you for this important comment. We agree that providing only the final gridded XCH_4 values may be limited for some downstream applications. In response, we expanded the publicly available dataset beyond the final daily 0.1° fused XCH_4 product to include the sensor-specific products used in the fusion. Specifically, the released dataset includes the TCCON-based bias-corrected GOSAT-2 product, the harmonized GOSAT-2-like TROPOMI and GOSAT products, and the source sensor identifier for the final fused product. These additional data allow users to trace which sensor contributed to each fused grid cell and to interpret sensor-specific differences and the fusion process more transparently.
Sections 4.1 and 4.2 also include analyses showing how the bias correction and harmonization results diagnose and mitigate residual bias structures in the individual satellite retrievals. For example, TROPOMI showed a strong dependency on SWIR surface albedo, while GOSAT showed a dependency on ΔPs. These retrieval-parameter-dependent biases were substantially reduced after ML-based correction and harmonization. Thus, the study does not only deliver a fused product but also provides bias-corrected and harmonized XCH_4 fields that address limitations of the operational satellite products.
We further added examples of downstream analyses that could benefit from the dataset. The TCCON-based bias-corrected GOSAT-2 product can serve as a reference-scale benchmark for future XCH_4 bias-correction or harmonization studies. The daily 0.1° fused XCH_4 product can support gap-filling studies, regional methane assessment, and hotspot identification. In addition, the corrected and harmonized sensor-specific products can be used to develop and evaluate more advanced fusion strategies. We clarified these dataset components and potential applications in the revised manuscript.

Lines 479-482: "*The bias-correction and harmonization analyses also revealed and mitigated sensor-specific residual bias structures, including the SWIR surface albedo dependency in TROPOMI and the ΔPs dependency in GOSAT. These results indicate that the proposed framework not only improves the final fused product but also provides diagnostic insight into retrieval-condition-dependent limitations of the individual satellite products.*"

Lines 489-492: "*The dataset can support downstream applications such as XCH_4 gap*

filling, multi-sensor intercomparison and harmonization, regional methane assessment, hotspot identification, and the development of advanced fusion strategies. The sensor-specific corrected and harmonized products can also serve as reference-scale benchmarks for future XCH_4 bias-correction studies"

Lines 496–500: *"The publicly available dataset includes the daily globally harmonized fused XCH4 product at 0.1° spatial resolution for 2020–2023, individual products used in the fusion process (TCCON-based bias-corrected GOSAT-2, GOSAT-2-like harmonized TROPOMI and GOSAT), and source sensor identifiers indicating which satellite selected each valid observation in the fused product. All products are provided in HDF5 format and are available on Zenodo at* https://doi.org/10.5281/zenodo.20304047*."*

3. Have the authors found it necessary to account for differences in prior profiles and vertical sensitivities of the different instruments (TCCON, TROPOMI, GOSAT, GOSAT-2) when making comparisons?

➔ We appreciate this important methodological point. The final objective of this study is to generate a daily 0.1° fused XCH_4 product by integrating GOSAT, GOSAT-2, and TROPOMI. To achieve this, we first needed to evaluate the relative performance of the three satellite products against a common external reference, select a reference scale for fusion, and then harmonize the other sensors to that scale.

In this study, TCCON was used as a common reference for bias correction and reference-scale selection. We did not explicitly apply averaging-kernel smoothing in Step 1–2. Applying satellite-specific averaging kernels can make each satellite–TCCON comparison more profile-aware, but it also smooths the TCCON reference differently according to the prior profile and vertical sensitivity of each satellite. This can make the effective reference comparison sensor-specific, which is not fully consistent with our objective of comparing the relative performance of the three sensors within a common TCCON reference framework.

Nevertheless, as the reviewer noted, differences in prior profiles and vertical sensitivities can affect satellite–TCCON comparisons. We therefore performed an additional sensitivity test to quantify their potential impact. In this analysis, we applied the TCCON pressure-weighting and averaging-kernel smoothing approach described in Appendix A of Balasus et al. (2023) and the TCCON data-comparison guidance. Before comparison, the TCCON profile was adjusted to the satellite-specific prior profile and averaging-kernel sensitivity, and the resulting validation statistics were compared with those from the original comparison.

The results show that averaging-kernel smoothing produced only limited changes in the validation statistics across all sensors. Across all TCCON stations, RMSE changes remained within ±0.6 ppb and ΔR^2 within ±0.02 for all datasets (Fig. R4). Similar results were obtained for the independent test stations used in this study, with RMSE differences within ±0.2 ppb and ΔR^2 within ±0.01 across all sensors (Table R1).

Therefore, although prior profiles and vertical sensitivities clearly differ among TCCON, TROPOMI, GOSAT, and GOSAT-2, our sensitivity test indicates that their impact on the validation statistics is relatively small compared with the systematic retrieval biases addressed by the ML-based bias correction and harmonization framework.

Reference:

Balasus, N., Jacob, D. J., Lorente, A., Maasakkers, J. D., Parker, R. J., Boesch, H., ... & Varon, D. J. (2023). A blended TROPOMI+ GOSAT satellite data product for atmospheric methane using machine learning to correct retrieval biases. *Atmospheric Measurement Techniques*, *16*(16), 3787-3807.

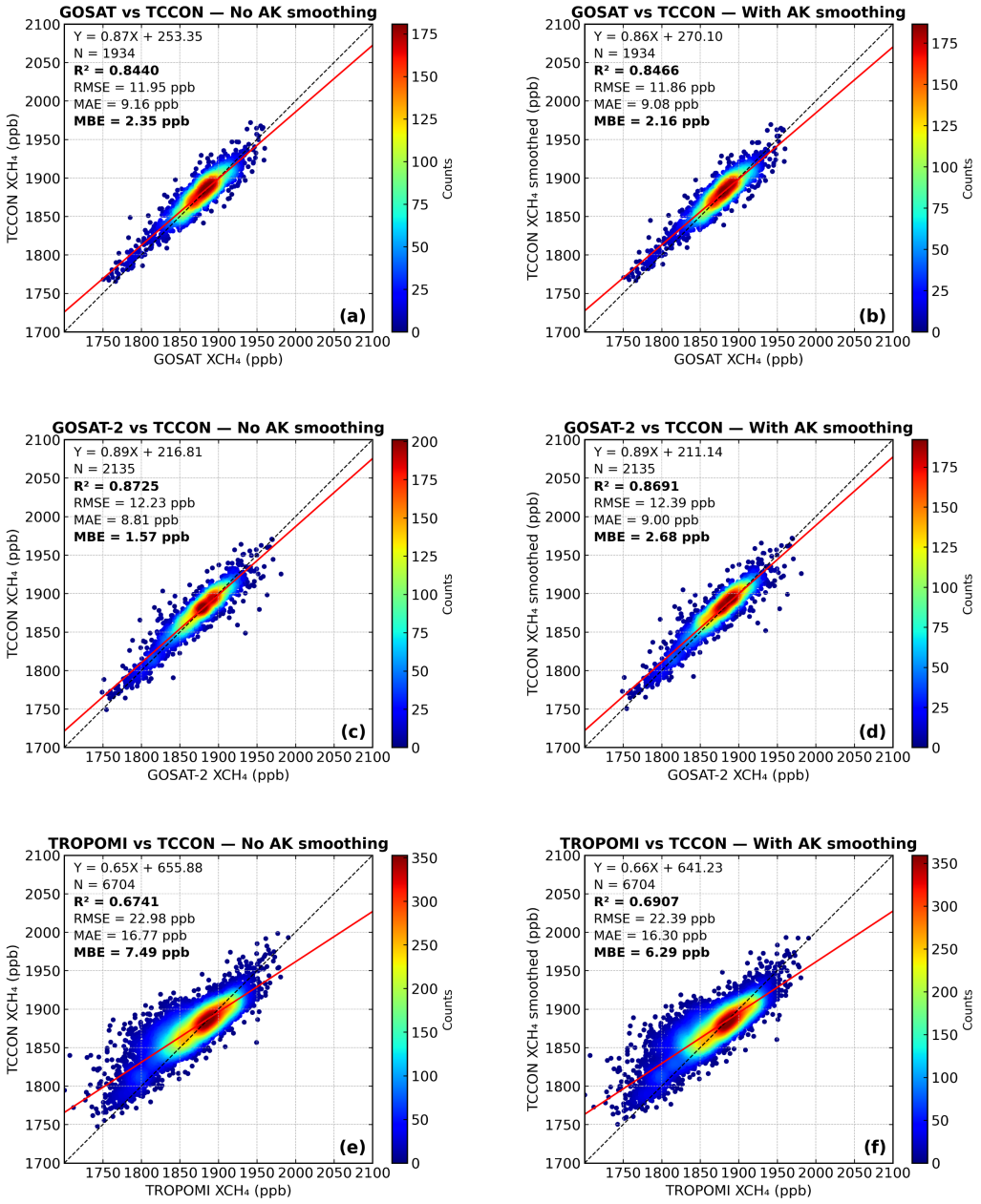

Figure R4. Comparison of satellite XCH_4 products against TCCON observations with and without averaging kernel (AK) smoothing. Panels (a), (c), and (e) show direct comparisons without AK smoothing,

while panels (b), (d), and (f) show comparisons after applying AK smoothing to TCCON observations. Results are shown for GOSAT, GOSAT-2, and TROPOMI, respectively. Each point represents an individual collocation pair and is colored by point density. The red line denotes the least-squares linear regression fit and the dashed black line indicates the 1:1 reference. Validation statistics including R^2, RMSE, MAE, and MBE are shown in each panel.

Table R1. Validation statistics (RMSE and MAE) for standard satellite XCH_4 products against TCCON at the three independent test stations (Edwards01, Garmisch01, Xianghe01), comparing direct comparison (no average kernel (AK) smoothing) and AK-smoothed TCCON results

	RMSE(Direct) ppb	RMSE(AK) ppb	MAE(Direct) ppb	MAE(AK) ppb	N
GOSAT	11.20	11.15	8.53	8.46	629
GOSAT-2	11.05	11.20	7.95	8.17	567
TROPOMI	17.48	17.43	13.48	13.49	1598

Minor Comments:

1. Line 39: misspelled reference?

   ➔ We thank the reviewer for pointing this out. The reference spelling has been corrected to "Saunois et al. (2025)" in line 36.

   Lines 35-36: "*Recent observations show that CH_4 levels have been rising at an unprecedented rate, with 2020–2022 marking the fastest growth since systematic monitoring began (Saunois et al., 2025).*"

2. Line 46: consider defining XCH_4

   ➔ We thank the reviewer for this suggestion. XCH_4 has been defined at its first occurrence in the revised manuscript (Lines 43–44) as 'the column-averaged dry-air mole fraction of atmospheric methane (XCH_4)

   Lines 40-41: "*Accurate quantification of the column-averaged dry-air mole fraction of atmospheric methane (XCH_4) is therefore essential for identifying emission sources and evaluating progress toward climate goals*."

3. Line 61: Lorente et al. (2021) uses a small-area approximation, not TCCON

   ➔ We thank the reviewer for this correction. The original text incorrectly implied that Lorente et al. (2021) used TCCON as a reference for bias correction. The revised text now accurately distinguishes between TCCON-based statistical regression approaches and the small-area approximation method.

   Lines 57-59: "*Previous bias correction studies have relied on statistical regression approaches that account for factors such as surface albedo, aerosol loading, and viewing*

geometry (Inoue et al., 2016), as well as small-area approximation methods that derive surface-albedo-related biases directly from satellite observations (Lorente et al., 2021)"

4. Line 99: please check the 30 times number (cf. Table 1 in Jacob et al., 2022)

➔ We thank the reviewer for this insightful reference. Upon checking Table 1 in Jacob et al. (2022) and Noël et al. (2021), we confirmed that the spectral resolution is 0.06 nm for GOSAT and GOSAT-2 and 0.25 nm for TROPOMI. We found that the "~30 times" statement was incorrectly calculated. We therefore revised the text to state approximately fourfold difference and explicitly added the spectral resolution of each sensor to Table 2.

Lines 105-108: "*GOSAT and GOSAT-2 provide high-precision measurements with spectral resolution approximately four times finer than that of TROPOMI (Kuze et al., 2009, 2016; Suto et al., 2021; Jacob et al., 2022; Noël et al., 2021), whereas TROPOMI offers daily global mapping, which increases data density despite higher susceptibility to atmospheric interference (Hu et al., 2018).*"

Reference:

Jacob, D. J., Varon, D. J., Cusworth, D. H., Dennison, P. E., Frankenberg, C., Gautam, R., ... & Duren, R. M. (2022). Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. *Atmospheric Chemistry and Physics*, *22*(14), 9617-9646.

Noël, S., Reuter, M., Buchwitz, M., Borchardt, J., Hilker, M., Bovensmann, H., ... & Warneke, T. (2021). XCO 2 retrieval for GOSAT and GOSAT-2 based on the FOCAL algorithm. *Atmospheric Measurement Techniques*, *14*(5), 3837-3869.

5. Line 224: extra "and" at the end of the sentence

➔ Thank you for pointing this out. We have removed the extra "and" at the end of the sentence.

Lines 231-232: "*Model performance was evaluated using three standard metrics: coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) (Equations S1–S3).*"

6. Line 253: specify northern-hemisphere/boreal summer and autumn

➔ Thank you for the suggestion. To avoid ambiguity and ensure consistency with the figure, we revised the seasonal terms to the corresponding month ranges.

Lines 261-263: "*Across sensors, the standard products showed negative biases and broad interquartile ranges (IQRs) throughout the year, with strong negative shifts from June to September (Fig. 5). The seasonal aggregation confirmed that this negative bias was most pronounced in June-July-August (JJA).*"

7. Figure 4: metrics (e.g. RMSE in subplots c,f,h) look much more optimistic than the LOSOCV column of Table S6

➔ Thank you for pointing this out. Figure 4 and Table S5 are both based on the leave-one-site-out cross-validation (LOSOCV) results, but the metrics were calculated at different aggregation levels and were intended to evaluate different aspects of performance. The LOSOCV metrics in Table S5 were calculated using individual satellite–TCCON co-location samples, and therefore reflect sample-level accuracy, including the effects of observation-level errors and variability, as well as uneven sample numbers across sites, as shown in the Figure 4 legend.

In contrast, Figure 4 was constructed after aggregating the LOSOCV predictions and TCCON observations at each TCCON site. Each point represents the site-mean XCH_4, and the error bars indicate the standard deviation of the co-location samples within each site, representing within-site variability. Thus, Figure 4 was intended to evaluate how effectively the bias correction reduces mean satellite–TCCON differences at the site level and how consistently this improvement is achieved across different concentration ranges and sites, rather than to assess sample-level prediction accuracy.

Because random sample-level errors and short-term variability are averaged during site-level aggregation, the performance metrics in Figure 4 can appear more optimistic than the sample-level LOSOCV metrics in Table S6. Therefore, the two results are not contradictory but represent different aspects of performance. We have clarified this distinction in the revised manuscript text.

Lines 236-237: "*In Figure 4, LOSOCV results and TCCON observations were aggregated at each TCCON site to examine site-level mean bias, within-site variability, and station-to-station consistency.*"

Line 243: "*In contrast, the ML-based correction tightened site-level agreement for all sensors, reducing both RMSE and MAE.*"

8. Figure 5: why not leave one season out in your cross validation if you are going to plot like this?

➔ Thank you for pointing this out. The validation was performed using leave-one-month-out cross-validation (LOMOCV), but the previous version of Figure 5 presented only seasonally aggregated results, which may have obscured the distinction between the validation unit and the visualization unit. We revised Fig. 5 to first show the monthly bias distributions from the LOMOCV results and then the seasonally aggregated bias distributions from the same LOMOCV outputs. This revision clarifies the monthly validation basis while also allowing the seasonal bias patterns to be interpreted more intuitively. The corresponding text and figure caption were revised accordingly.

Lines 261-269: "*Across sensors, the standard products showed negative biases and broad interquartile ranges (IQRs) throughout the year, with strong negative shifts from June to September. The seasonal aggregation confirmed that this negative bias was most pronounced in June-July-August (JJA). This seasonal pattern is consistent with known*

sensitivities of SWIR-based XCH_4 retrievals to surface albedo and atmospheric scattering by aerosols and cirrus (Inoue et al., 2016; Hu et al., 2016; Lorente et al., 2021; Oshio et al., 2020). These factors can modify the effective light path and vary seasonally with vegetation, humidity, cloud conditions, and solar geometry. Operational bias correction shifted the median bias toward zero but still exhibited larger seasonal dispersion than the ML-based correction. The ML-based correction maintained seasonal medians near zero and reduced both the IQRs and non-outlier spread, with most IQRs remaining within ±8 ppb, indicating more stable performance under seasonal variability."

References:

Inoue, M., Morino, I., Uchino, O., Nakatsuru, T., Yoshida, Y., Yokota, T., ... & Tanaka, T. (2016). Bias corrections of GOSAT SWIR XCO 2 and XCH 4 with TCCON data and their evaluation using aircraft measurement data. *Atmospheric Measurement Techniques*, *9*(8), 3491-3512.

Hu, H., Hasekamp, O., Butz, A., Galli, A., Landgraf, J., Aan de Brugh, J., ... & Aben, I. (2016). The operational methane retrieval algorithm for TROPOMI. *Atmospheric Measurement Techniques*, *9*(11), 5423-5440.

Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., Schneider, A., Wu, L., ... & Landgraf, J. (2021). Methane retrieved from TROPOMI: improvement of the data product and validation of the first 2 years of measurements. *Atmospheric Measurement Techniques*, *14*(1), 665-684.

Oshio, H., Yoshida, Y., Matsunaga, T., Deutscher, N. M., Dubey, M., Griffith, D. W., ... & Wunch, D. (2020). Bias correction of the ratio of total column ch4 to co2 retrieved from gosat spectra. *Remote Sensing*, *12*(19), 3155.

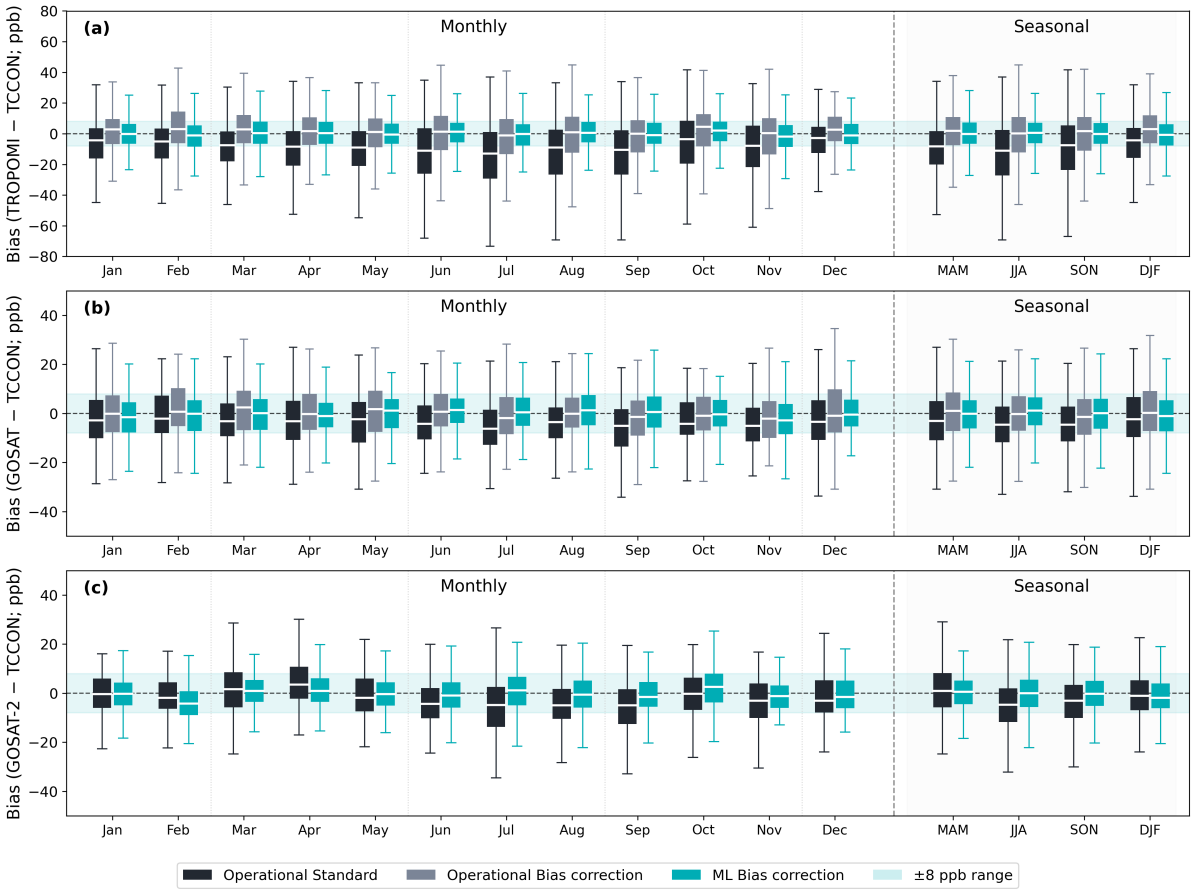

Figure 5. Monthly and seasonally aggregated XCH_4 bias distributions under leave-one-month-out cross-validation (LOMOCV). Box-and-whisker plots show bias distributions (Satellite − TCCON) for each calendar month and their seasonal aggregations for (a) TROPOMI, (b) GOSAT, and (c) GOSAT-2. The compared datasets include the uncorrected operational standard products, operational bias-corrected products, and ML-based bias-corrected results. Boxes indicate the interquartile range (IQR) with the median shown as a horizontal white line. The shaded band marks the ±8 ppb range around zero bias.

9. SI: double-check (e.g., no bolds in Tables S6 and S7, no Table S8, etc.)

➔ Thank you for pointing this out. We checked the Supplementary Information, added the missing bold formatting in Tables S5 and S6, and corrected the table numbering from Table S6 to Table S7.

10. Line 525: Jacob et al. (2022) no longer is in Discussion

➔ Thank you for checking this. Jacob et al. (2022) is still cited in the Introduction and Datasets sections. We have checked the citation locations accordingly.

Referee #2

First of all, I want to thank the authors for this well-structured and clear and concise article. In terms of grammar, I have nothing to add. However, I do have three important issues that need to be addressed.

Major Comments:

1. The first is the issue of proper accreditation. Each of the individual TCCON datasets used, should have their data reference added to the references list (See https://tccon-wiki.caltech.edu/Main/DataLicense for TCCON citation guidelines). These individual per station citations can be found on https://tccondata.org/.

   ➔ Thank you for pointing this out. Following the TCCON data citation guidelines, we added the individual data references for all TCCON station datasets used in this study to the reference list. We also moved the TCCON site table to the main manuscript to clearly present the stations, site information, and corresponding citations. These revisions ensure proper accreditation of the TCCON datasets used in this study.

   Lines 87-88: "*Stations used in this study are shown in Fig. 1 and listed in Table 1.*"

2. The second point pertains to the selection of GOSAT-2 as the standard to which the other satellite products are bias-corrected (step 2 in the overall process). This selection of GOSAT-2 is based on the results listed in Table-3. However, it is not clearly stated if the common data sample on which Table 3 is based, is the training sample or instead comes from the LOSOCV approach. If the first, I consider this to be a weak basis for selection, if the latter it is a stronger one as we are fundamentally interested in the performance of the bias corrected products outside the scope of the training dataset. That said, we also need to take the global distribution of the TCCON network into account, which under-samples large swaths of the globe. In that view it is hard to state, with confidence, based on the analysis performed here, that GOSAT-2 should be taken as the definitive reference. I would very much prefer it if the authors performed 3 different step2 analysis wherein in turn, GOSAT, GOSAT-2 and TROPOMI are taken as a reference. This will allow the authors to perform an intercomparison, assess the impact of this choice on a global scale, identifying regions where things converge and diverge, and make a more thorough determination on whether all 3 of these end products turn out to be valid candidates or that one is superior.

   ➔ We thank the reviewer for this important methodological question. We confirm that Table 4 is based entirely on leave-one-site-out cross-validation (LOSOCV) results, not the training sample, ensuring that the evaluation is independent of the training data and provides a cross-validated basis for selecting GOSAT-2 as the harmonization reference. We have clarified this in the revised manuscript text and table caption.

   In addition, we agree that the spatial distribution of the TCCON network is uneven and does not fully represent the diverse retrieval environments encountered globally. Therefore, we acknowledge that the TCCON-based bias-corrected GOSAT-2 product should not be

regarded as an absolute global truth standard. Given this limitation, we used the product with the most stable performance in the LOSOCV evaluation against TCCON as a practical reference scale for multi-sensor harmonization. Because the validation site conditions are excluded from training in LOSOCV, this evaluation provides an independent basis for reference-scale selection within the constraints of the TCCON network. This limitation has been clarified in the revised manuscript and added to the limitation discussion in Section 4.5.

We agree that comparing Step 2 harmonization results using GOSAT, GOSAT-2, and TROPOMI as alternative references would be scientifically valuable. However, TROPOMI showed lower and more variable performance than the GOSAT series across different cross-validation strategies in Table S5, with LOSOCV $R^2 = 0.79$ for TROPOMI compared with 0.87–0.88 for the GOSAT series. Furthermore, TROPOMI XCH_4 retrievals are known to carry inherent surface-reflectance-dependent biases from the 2.3 μm spectral band (Lorente et al., 2021; Lorente et al., 2023; Somkuti et al., 2025), and previous studies have specifically used GOSAT as a harmonization anchor because of its higher spectral precision and retrieval stability (Balasus et al., 2023; Li et al., 2024; Fan et al., 2024). Figure R6 confirms that TROPOMI shows strong albedo dependence in the standard product. Although ML-based correction reduces this within the TCCON-sampled range, the training data are limited to a narrow albedo range of approximately 0.0–0.30. The GOSAT series, in contrast, shows stable retrieval characteristics even within this limited range, suggesting comparably stable performance beyond it as well. Therefore, we considered TROPOMI a lower-priority candidate as the harmonization reference for the reference-scale selection objective of this study. Instead, we performed an additional analysis comparing GOSAT-anchored and GOSAT-2-anchored harmonization within the GOSAT series, which we considered the more plausible candidate set.

Figure R7 showed that GOSAT-2 provided more inter-satellite co-location samples with TROPOMI across all years, months, and latitudinal zones, which is advantageous for training a more stable and geographically representative harmonization model. Cross-validation (CV) results further showed that the GOSAT-2-anchored harmonization produced lower errors than the GOSAT-anchored harmonization across different CV strategies (Table R2), latitudinal zones, and months (Fig. R8). Both reference frameworks substantially reduced surface albedo and AOT dependencies in the harmonized TROPOMI product (Fig. R9), but the GOSAT-2-anchored framework showed more stable overall performance.

Taken together, these results do not establish GOSAT-2 as an absolute global reference. However, considering the LOSOCV-based TCCON evaluation, the reference-scenario sensitivity analysis within the GOSAT series, and the larger co-location sample with TROPOMI, the TCCON-based bias-corrected GOSAT-2 product is the most reasonable choice as a practical and well-validated reference scale for the multi-sensor harmonization framework in this study.

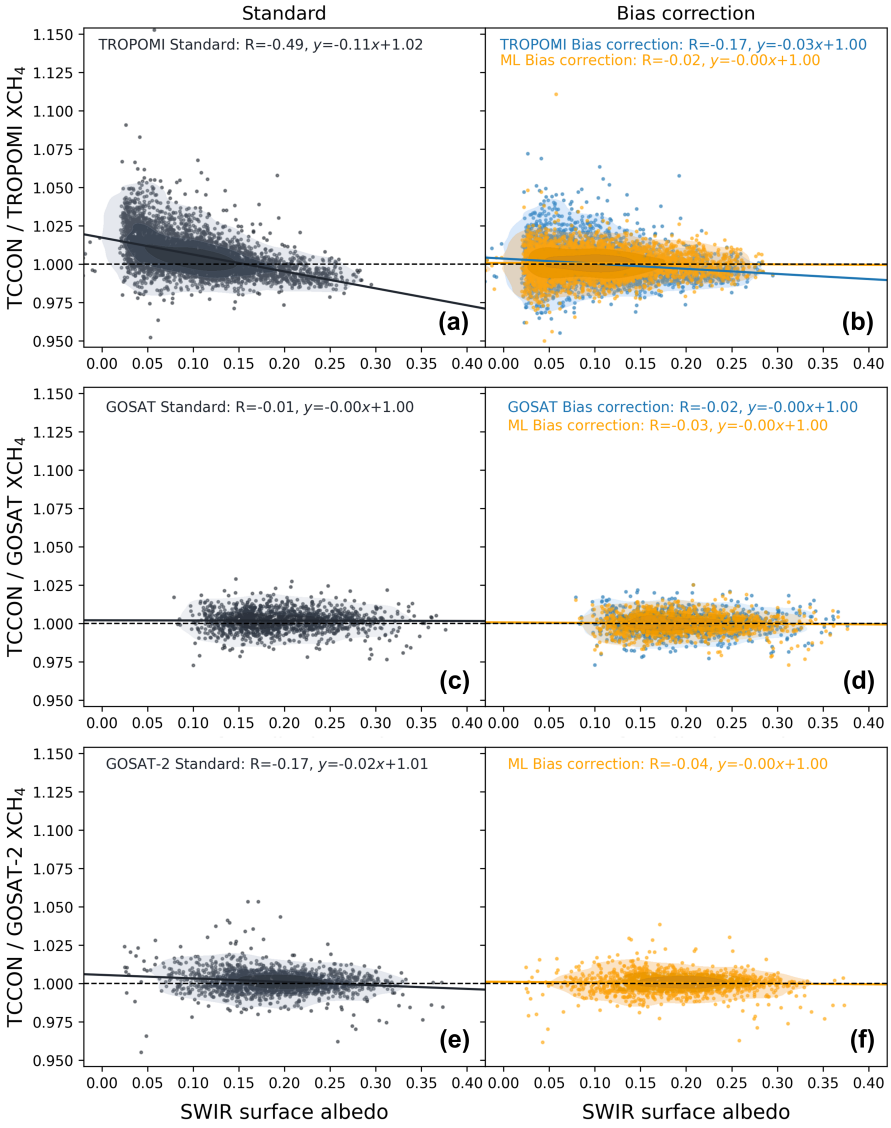

Figure R6. SWIR surface albedo dependence of satellite XCH_4 products relative to TCCON under LOSOCV, across an albedo range of 0.0–0.40. Left column shows standard products and right column shows bias-corrected products for TROPOMI (a, b), GOSAT (c, d), and GOSAT-2 (e, f). Orange points indicate ML-based bias-corrected results and blue points indicate operational bias-corrected results. Solid lines denote least-squares regression fits and the dashed horizontal line indicates the unbiased reference level (ratio = 1). Pearson correlation coefficients (R) and regression equations are reported in each panel.

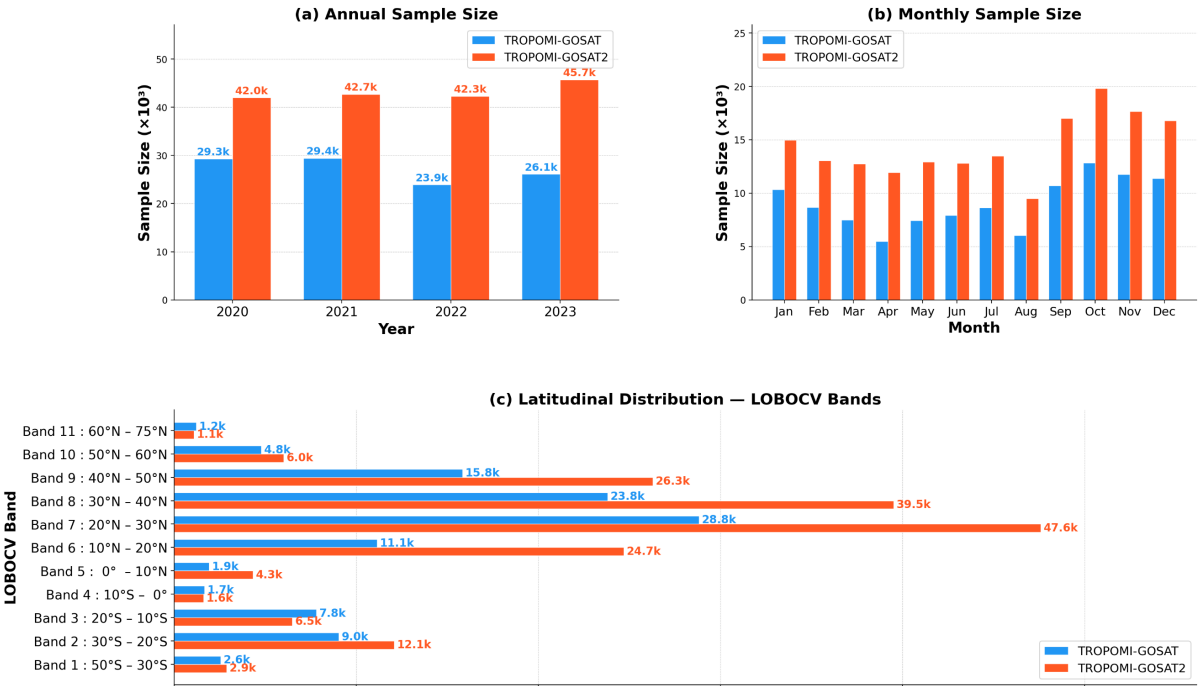

Figure R7. Comparison of inter-satellite colocation sample sizes for TROPOMI-GOSAT (blue) and TROPOMI-GOSAT-2 (orange) pairs used in Step 2 harmonization for 2020–2023. (a) Annual sample size, (b) monthly sample size, and (c) latitudinal distribution across LOBOCV bands.

Table R2. Cross-validation performance comparison of TROPOMI harmonization using GOSAT and GOSAT-2 as alternative harmonization references. Boldface indicates the better-performing value for each metric

		LOBOCV[a]	LOMOCV[b]	LOYOCV[c]
GOSAT as reference	N		112,341	
	R^2	0.86	0.88	0.87
	MAE (ppb)	8.62	8.06	8.39
	RMSE (ppb)	11.32	10.61	11.02
GOSAT-2 as reference	N		183,550	
	R^2	**0.91**	**0.92**	**0.92**
	MAE (ppb)	**6.78**	**6.32**	**6.59**
	RMSE (ppb)	**9.07**	**8.49**	**8.82**

[a] LOBOCV: Leave-One-Band-Out Cross-Validation
[b] LOMOCV: Leave-One-Month-Out Cross-Validation
[c] LOYOCV: Leave-One-Year-Out Cross-Validation

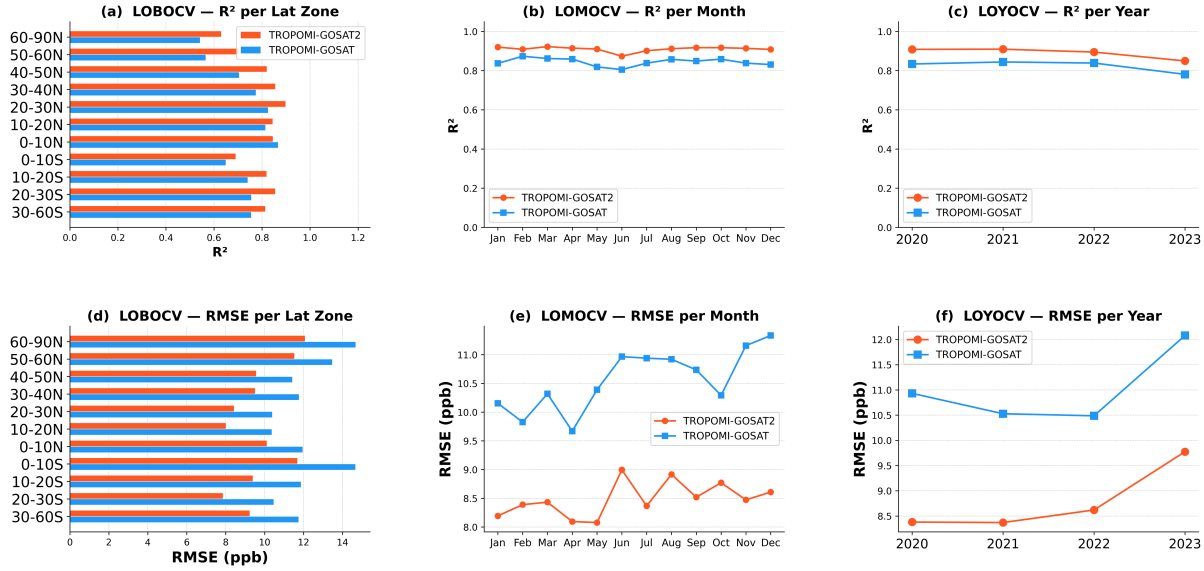

Figure R8. Cross-validation performance comparison of TROPOMI harmonization using GOSAT (blue) and GOSAT-2 (orange) as alternative harmonization references. Panels (a–c) show R^2 results for LOBOCV by latitudinal zone, LOMOCV by month, and LOYOCV by year, respectively. Panels (d–f) show the corresponding RMSE results for LOBOCV by latitudinal zone, LOMOCV by month, and LOYOCV by year.

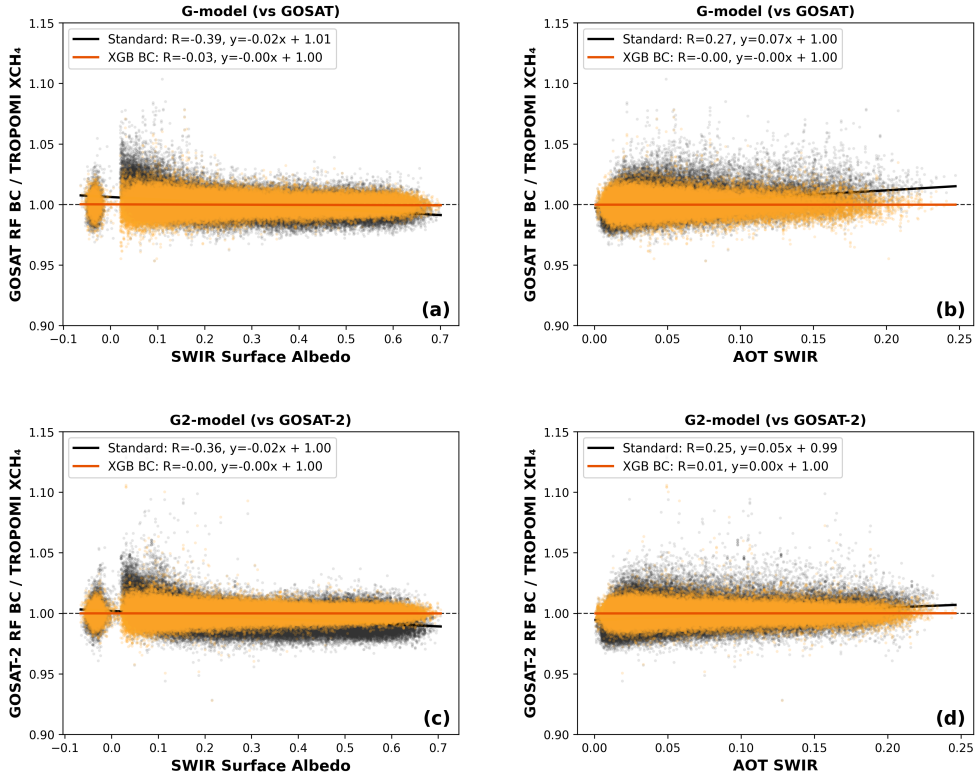

Figure R9. Dependence of the harmonized TROPOMI XCH_4 ratio on SWIR surface albedo (left column) and aerosol optical thickness (AOT SWIR; right column) before (black) and after (orange) XGBoost-based harmonization. Results are shown for TROPOMI harmonized to GOSAT (G-model; panels a, b) and to GOSAT-2 (G2-model; panels c, d). Solid lines denote least-squares regression fits and the dashed horizontal line indicates the unbiased reference level (ratio = 1).

Lines 296-298: *"Table 4 compares the ML-based bias-corrected products from all three sensors evaluated against TCCON using common collocated samples under LOSOCV, ensuring that the reference selection is based on independent out-of-site performance."*

Lines 454-459: "*First, the ML-based bias correction relies on TCCON as the primary reference dataset. Although TCCON provides high-precision ground-based XCH_4 observations, its spatial distribution is limited and does not fully represent the diverse retrieval conditions encountered globally, particularly high-surface-albedo conditions. We partly addressed this limitation through leave-one-station-out cross-validation and additional satellite match-up analyses over broader surface albedo ranges, but independent validation under underrepresented retrieval conditions remains necessary.*"

References:

Lorente, A., Borsdorff, T., Butz, A., Hasekamp, O., aan de Brugh, J., Schneider, A., Wu, L., Hase, F., Kivi, R., Wunch, D., Pollard, D. F., Shiomi, K., Deutscher, N. M., Velazco, V. A., Roehl, C. M., Wennberg, P. O., Warneke, T., and Landgraf, J.: Methane retrieved from TROPOMI: improvement of the data product and validation of the first 2 years of measurements, Atmospheric Measurement Techniques, 14, 665–684, https://doi.org/10.5194/amt-14-665-2021, 2021.

Lorente, A., Borsdorff, T., Martinez-Velarte, M. C., and Landgraf, J.: Accounting for surface reflectance spectral features in TROPOMI methane retrievals, Atmospheric Measurement Techniques, 16, 1597–1608, https://doi.org/10.5194/amt-16-1597-2023, 2023.

Somkuti, P., McGarragh, G., O'Dell, C., Di Noia, A., Vogel, L., Crowell, S., Ott, L. E., and Bösch, H.: Surface reflectance biases in XCH_4 retrievals from the 2.3 µm band are enhanced in the presence of aerosols, Atmospheric Measurement Techniques, 18, 4647–4663, https://doi.org/10.5194/amt-18-4647-2025, 2025.

Balasus, N., Jacob, D. J., Lorente, A., Maasakkers, J. D., Parker, R. J., Boesch, H., Chen, Z., Kelp, M. M., Nesser, H., and Varon, D. J.: A blended TROPOMI+GOSAT satellite data product for atmospheric methane using machine learning to correct retrieval biases, Atmospheric Measurement Techniques, 16, 3787–3807, https://doi.org/10.5194/amt-16-3787-2023, 2023.

Li, K., Bai, K., Jiao, P., Chen, H., He, H., Shao, L., Sun, Y., Zheng, Z., Li, R., and Chang, N.-B.: Developing unbiased estimation of atmospheric methane via machine learning and multiobjective programming based on TROPOMI and GOSAT data, Remote Sensing of Environment, 304, 114039, https://doi.org/10.1016/j.rse.2024.114039, 2024.

Fan, L., Wan, Y., and Dai, Y.: Development of a Multi-Source Satellite Fusion Method for XCH4 Product Generation in Oil and Gas Production Areas, Applied Sciences, 14, 11100, https://doi.org/10.3390/app142311100, 2024.

3. The third point addresses the rank order merging method used in step three, where each 0.1° daily grid is filled with GOSAT-2 if available, then TROPOMI if available, then GOSAT. If the ML harmonization between the satellite products performed in step 2 is successful, I see no reason why this method is superior to simply taking the median of all products in the daily grid cell. If the ML harmonization is unsuccessful, then clearly the rank order creation of the merged dataset isn't the solution either.

   ➔ Thank you for this important comment. We agree that the merging strategy after

harmonization should be clearly justified. To address this, we additionally quantified the overlap frequency of available satellite observations in the valid daily 0.1° grid cells of the final fused XCH_4 product.

Figure R10 shows the overall, daily, and spatial sensor-overlap frequencies for 2020. More than 99% of valid grid cells contained only one satellite observation, whereas cases with two or more simultaneous satellite observations accounted for less than 1% (Figs. R10a, b). Spatially, the frequency of three-sensor overlap was mostly below 0.1% of valid days, and the frequency of two-sensor overlap was generally below 10% (Figs. R10c, d). Under this sparse-overlap structure, the practical advantage of median-based merging as a robust estimator is limited. When only one satellite observation is available, median, mean, and priority-selection methods produce the same value, and the advantage of the median becomes meaningful only when multiple observations are frequently available within the same grid cell.

Therefore, the main purpose of fusion in this study is not to statistically average multiple sensor values within the same grid cell, but to use complementary observations from the GOSAT series in regions or conditions where TROPOMI retrievals are limited. As discussed in Section 4.3, the overall increase in global coverage from fusing the three satellites was modest, and most coverage was contributed by TROPOMI. However, the GOSAT series can provide additional XCH_4 information under some conditions and regions where TROPOMI observations are unavailable; thus, the benefit of fusion lies more in complementary sampling than in a large increase in coverage percentage.

For this reason, we considered the priority-selection strategy to be the most appropriate. ML-based harmonization reduces systematic inter-sensor discrepancies by converting GOSAT and TROPOMI to the bias-corrected GOSAT-2 scale, and Figure 7 shows that the inter-sensor bias distributions became more aligned with reduced spread after harmonization. Nevertheless, residual inter-sensor differences can remain even after harmonization; therefore, a simple mean or median is not necessarily optimal when multiple values are available in the same grid cell. We therefore prioritized the ML-based bias-corrected GOSAT-2 product, which defines the reference scale, followed by TROPOMI and GOSAT based on their agreement with reference and sampling density.

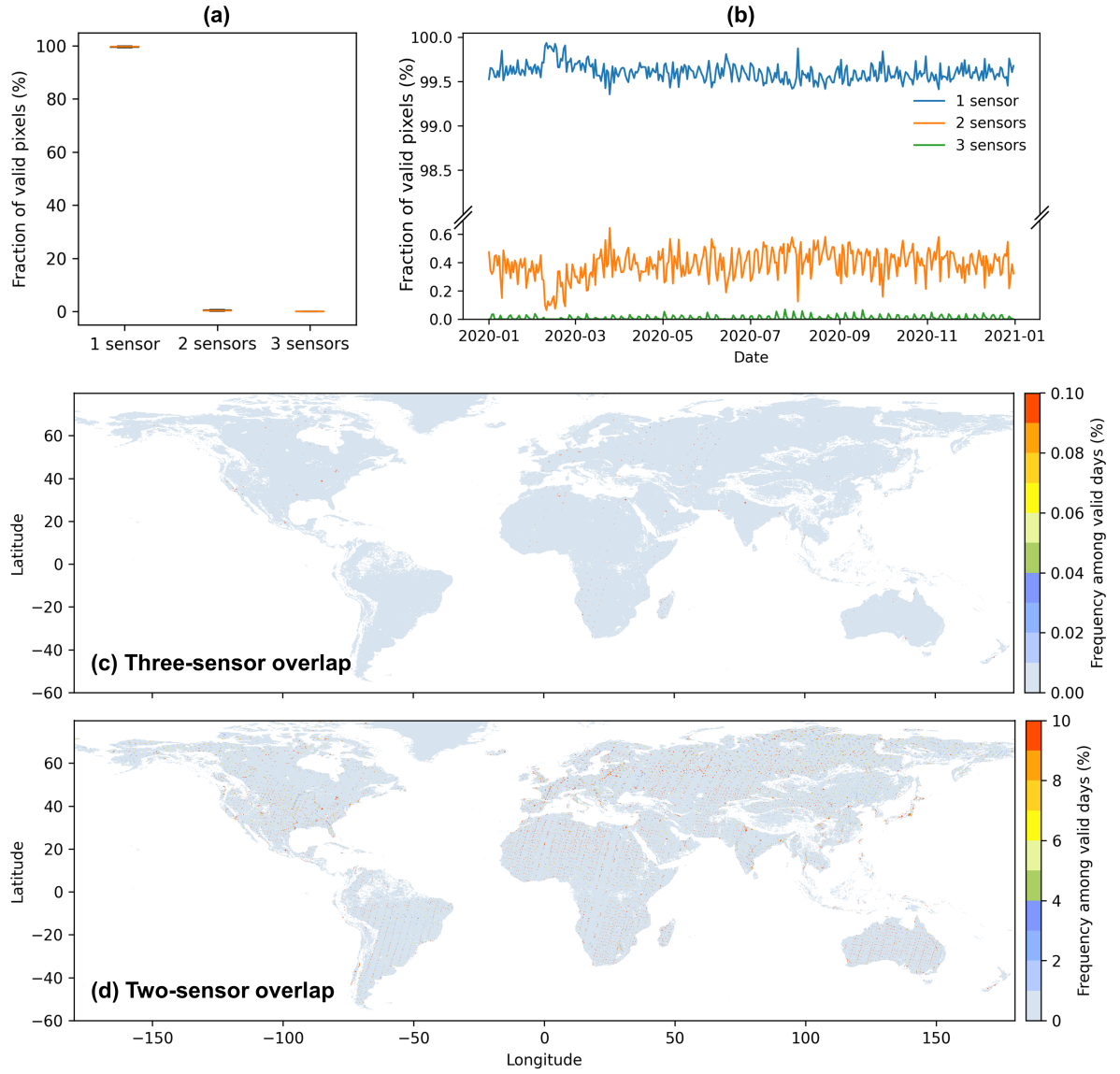

Figure R10. Sensor-overlap frequency in valid daily 0.1° grid cells. Distribution of the number of available satellite observations used for the fused XCH_4 product in 2020. (a) Overall fraction of valid grid cells containing one, two, or three available sensors. (b) Daily variation in the fraction of valid grid cells by the number of available sensors. (c, d) Spatial frequency of three-sensor and two-sensor overlaps, expressed as the percentage of valid days at each grid cell.

Minor Comments:

1. Line 164: Sha et al. 2021 is used as a source for the used validation colocation criteria. Note however that Sha et al. consider the line of sight of the FTIR instrument. To quote the paper:"An effective location of the FTIR measurement on the line of sight (i.e. at a 5 km altitude) is used to do the co-location". This should be acknowledged.

   Sha, M. K., Langerock, B., Blavier, J.-F. L., Blumenstock, T., Borsdorff, T., Buschmann, M., Dehn, A., De Mazière, M., Deutscher, N. M., Feist, D. G., García, O. E., Griffith, D. W. T., Grutter, M., Hannigan, J. W., Hase, F., Heikkinen, P., Hermans, C., Iraci, L. T., Jeseck,

P., Jones, N., Kivi, R., Kumps, N., Landgraf, J., Lorente, A., Mahieu, E., Makarova, M. V., Mellqvist, J., Metzger, J.-M., Morino, I., Nagahama, T., Notholt, J., Ohyama, H., Ortega, I., Palm, M., Petri, C., Pollard, D. F., Rettinger, M., Robinson, J., Roche, S., Roehl, C. M., Röhling, A. N., Rousogenous, C., Schneider, M., Shiomi, K., Smale, D., Stremme, W., Strong, K., Sussmann, R., Té, Y., Uchino, O., Velazco, V. A., Vigouroux, C., Vrekoussis, M., Wang, P., Warneke, T., Wizenberg, T., Wunch, D., Yamanouchi, S., Yang, Y., and Zhou, M.: Validation of methane and carbon monoxide from Sentinel-5 Precursor using TCCON and NDACC-IRWG stations, Atmos. Meas. Tech., 14, 6249–6304, https://doi.org/10.5194/amt-14-6249-2021, 2021.

➔ Thank you for pointing this out. We agree that Sha et al. (2021) used a line-of-sight-based effective FTIR location for co-location, whereas our TROPOMI–TCCON co-location followed the spatial-temporal criteria described by Balasus et al. (2023). We therefore revised the sentence and replaced the citation accordingly.

Lines 171-173: "*We therefore constructed collocated training pairs using satellite-specific spatiotemporal and elevation constraints, adopting criteria from previous studies and official validation strategies (Balasus et al., 2023; Yoshida et al., 2023).*"

Reference:

Balasus, N., Jacob, D. J., Lorente, A., Maasakkers, J. D., Parker, R. J., Boesch, H., ... & Varon, D. J. (2023). A blended TROPOMI+ GOSAT satellite data product for atmospheric methane using machine learning to correct retrieval biases. *Atmospheric Measurement Techniques*, *16*(16), 3787-3807.

2. Line 241: a space is missing between "conditions." and "Given".

➔ Thank you for pointing this out. We have corrected the missing space between "conditions." and "Given."

3. Line 295: please repeat that we are building a 0.1° daily product here.

➔ Thank you for the suggestion. We added the requested clarification that the final output is a daily 0.1° fused XCH_4 product.

Lines 316-317: "*Finally, we generated a daily 0.1° fused XCH_4 product by integrating the harmonized observations from the three satellites.*"

4. Line 308: I would not describe Xianghe as a rural site (it sits within 100 km of Beijing Centre) in a heavily industrialized and urbanized region. There is probably a mix-up with Edwards (described as urban), which 100 km radius touches the outskirts of Los Angeles but in and of itself is situated in the desert.

➔ Thank you for catching this. We have revised the site descriptions to better reflect each station's actual environment. Xianghe is located approximately 50 km east-southeast of

Beijing, within a densely populated and economically active region of China, and is surrounded by urban areas, croplands, and strong anthropogenic emission sources. Edwards is located in the arid high desert of California and is characterized by high surface albedo associated with nearby bright dry lakebeds. Accordingly, Xianghe is now described as a peri-urban site influenced by strong anthropogenic emissions, while Edwards is described as an arid high-desert site with high surface albedo. The descriptions have been corrected in the revised manuscript.

Lines 333-336: *"The three stations represent distinct environments: a peri-urban site influenced by strong anthropogenic emissions (Xianghe, China; mean = 1903.16 ppb), an arid high-desert site with high surface albedo (Edwards, USA; mean = 1880.37 ppb), and a mountainous site (Garmisch, Germany; mean = 1872.98 ppb)."*

5. Paragraph 336 onwards: Here we see an improvement of the fusion product compared to TROPOMI when comparing to GOSAT-2, but all components within the fusion product are ML bias corrected towards GOSAT-2. It is thus basically telling us the same as Figure 7.

➔ We thank the reviewer for this observation and agree that the previous Fig. 10b partially overlapped with the harmonization evaluation already presented in Fig. 7. To avoid redundancy and potential ambiguity in interpretation, we removed the regional discrepancy comparison figure from the revised manuscript. Section 4.3 now focuses on the primary objective of the fusion framework in retrieval-challenged environments: improving observational coverage and spatial complementarity among sensors. We have clarified this focus in the revised manuscript.

6. Paragraph 363 onwards: How exactly was the growth rate calculated. Also, no uncertainty values on the growth rate are given.

➔ We thank the reviewer for this helpful comment. The mean annual growth rates shown in Fig. 11b were estimated using linear regression of the monthly global mean XCH_4 time series over 2020–2023, with uncertainties reported as the standard errors of the regression slopes. The regional interannual growth rates in Fig. 11c were calculated as year-over-year differences (i.e., each month minus the corresponding month of the previous year). Annual markers represent the mean of the monthly growth rates within each year, while error bars represent the standard deviation (σ, $n-1$) of the monthly growth rates within each year. All linear trends were statistically significant ($p < 0.001$). Figure 11 and the corresponding discussion have been revised accordingly.

Lines 387-397: *"Mean annual growth rates were calculated from the monthly time series over 2020–2023, while uncertainties represent the variability of monthly growth rates within each year. The fused product captures a sustained increase from approximately 1850 ppb in early 2020 to nearly 1900 ppb by late 2023, corresponding to a mean global growth rate of* $12.28 \pm 1.00 ppb\ yr^{-1}$ *(Fig. 10b). This agrees well with TCCON* $(13.56 \pm 1.25 ppb\ yr^{-1})$ *and NOAA* $(14.77 \pm 0.64 ppb\ yr^{-1})$*, supporting the ability of the fused dataset to represent large-scale atmospheric* CH_4 *variability and interannual trends.*

Interannual growth rates exhibited strong regional variability (Fig. 10c). The NH (0–80°N), SH (60°S–0°), and the high-emission zone (0–40°N) all peaked in 2021 at 14.3 ± 2.2, 15.4 ± 2.7, *and* 15.5 ± 2.1 *ppb* yr^{-1}, *respectively, consistent with record global* XCH_4 *increases reported for that period (Saunois et al., 2025). Following this peak, growth rates generally declined during 2022–2023, reaching approximately* 6.0 ± 1.8 *ppb* yr^{-1} *in the NH,* 11.4 ± 3.9 *ppb* yr^{-1} *in the SH, and* 6.4 ± 3.0 *ppb* yr^{-1} *in the high-emission zone by 2023, consistent with recent inverse modeling estimates (Pendergrass et al., 2025)."*

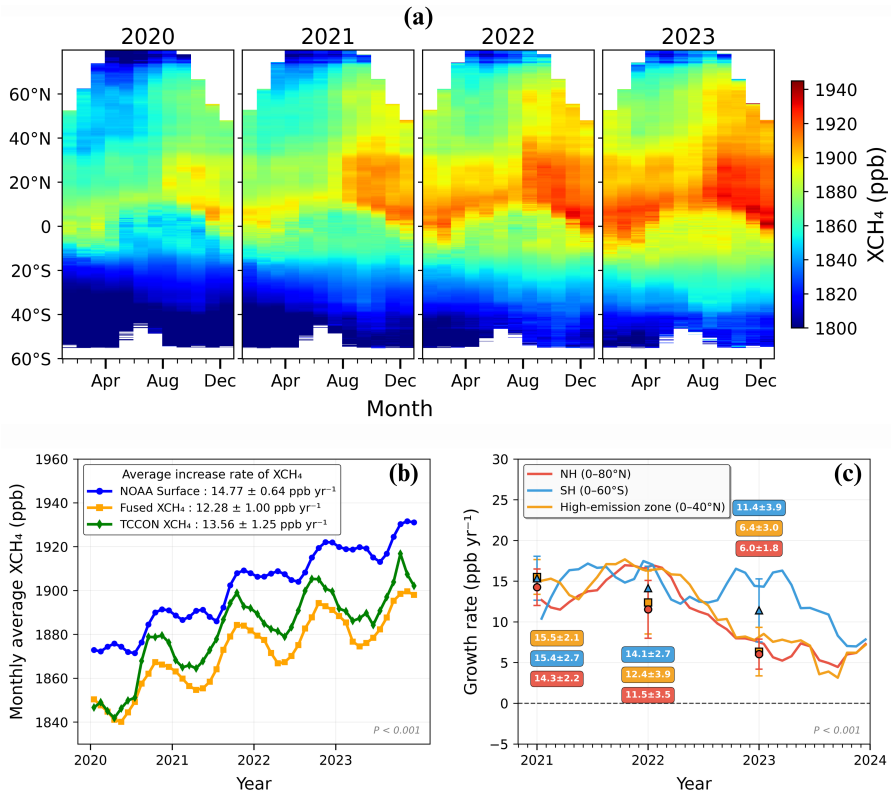

Figure 11. Spatiotemporal characteristics and growth of fused XCH_4 ***during 2020–2023. (a) Monthly latitudinal distribution of fused*** XCH_4 ***concentrations at 0.1° spatial resolution from 2020 to 2023. (b) Comparison of monthly global mean methane time series from the fused*** XCH_4 ***product (orange), NOAA marine surface*** CH_4 ***measurements (blue), and TCCON ground-based*** XCH_4 ***observations (green). Values in the legend indicate mean annual growth rates over 2020–2023 with associated uncertainties. (c) Monthly hemispheric year-over-year differences (solid lines) and annual mean growth rates (markers) for the Northern Hemisphere (NH; 0–80°N, red), Southern Hemisphere (SH; 0–60°S, blue), and high-emission zone (0–40°N, orange). Error bars represent the standard deviation of monthly growth rates within each year. All trends shown in panels (b) and (c) were statistically significant (p < 0.001).***

7. Line: 580: This is a AMTD reference, replace by its non-discussions final paper

➔ Thank you for pointing this out. We have replaced the AMTD reference with the appropriate final published reference and no longer cite the AMTD discussion paper in the revised manuscript.