Reviewer comments are italicized

Author responses are indented and unitalicized

Changes to the manuscript text are in blue

The authors would like to thank the anonymous reviewers for their time and attention to this manuscript. Your comments and suggestions have been greatly appreciated and have significantly improved the paper.

Response to anonymous referee 3

G1: I find that the biases, which the authors are trying to correct for, appear mostly linear (see Figure 2). I am not convinced that using a machine learning correction is necessary and I fear that the introduction of such highly non-linear models will lead to overfitting. How have you ensured that you are fitting signal and not noise?

A Machine learning model does not try to impose non-linearity upon the fit, so if a linear fit is most appropriate according to the model training, then the resulting model will appear more linear. Using the machine learning model could be computational overkill, but the results would not be negatively affected.

Overfitting can be monitored for in the training process of the ML models and we did so to ensure we were not overfitting. The results of which are available in the Supplementary Information figure S3, which describe the training and validation loss throughout the training process for each monthly model. Overfitting would be identified in these plots as a significant increase in validation loss while training loss remains low. The consistent gap between the training and validation loss at the end of the training period actually indicates we are underfitting with our models. Unfortunately, the solution to underfitting is introducing more data to the model, which we are in limited supply of due to the low amount of overlap between GOSAT and TROPOMI in this region. Underfitting in this case means that our model could still be improved by incorporating more data, but the output values of the model are valid for our purposes.

To clarify this for readers, we have added the following to the methods section:

"To monitor against overfitting, training and validation loss for the training period of each model were calculated and are presented in **Figure S3**."

And we have added more text to explain the SI figure further to ensure readers understand there has been no overfitting

"To ensure the monthly models are not over- or under-fitting the data, training and validation loss plots were constructed for the training process for each monthly model. The number of epochs for these training processes are generally quite low, but the lack of loss value fluctuation later in the training process demonstrates that more training is unnecessary. Models that are underfit will show significant gaps between the training and validation loss, and one could argue that the models presented here show some signs of underfitting. However, the solution to underfitting is generally to incorporate more data into the model, and the co-located GOSAt and TROPOMI points used for training were already stretched to their limit. An overfit model will appear in these plots as a divergence, generally later in the training process, of the training and validation losses. These models do not demonstrate any patterns of overfitting, suggesting that the models were trained accurately and successfully."

G2: I am very surprised to see such strong dependence of the XCH4 bias correction on land cover type, which is much more typically observed (and intuitively understood) in trace gas retrievals from coarse-spectral-resolution sensors like AVIRIS-NG. Please comment in the text why such hyper-local corrections might be physically plausible.

The XCH4 bias correction is strongly dependent on very specific land cover types. The referee will notice that "Non-Agricultural" land cover types include a multitude of cover types and the variability within this group is quite small. Where we observe strong dependence is between water-intensive and drought-resistant crops and we believe this is due to the soil or vegetation water content and its effect on the measurement light frequencies. This is already described in the manuscript:

"…water sources like these two eventual tributaries to the Missouri River dictate where larger water-intensive agricultural operations exist. As such, larger densities of water-intensive crop farms are co-located with these rivers, bringing their albedo-influencing crops and plant-life, and thus requiring an albedo correction which is not necessarily reflected in magnitude by the surrounding scrubland. It has been shown that water intensive crops, like corn, sugarbeets, and alfalfa; and drought resistant crops, like winter wheat, millet, and dry beans; reflect SWIR light differently, allowing for identification of crops from space with the SWIR reflectance variable along with other variables (Chen et al., 2005). This effect is possibly due to water content or leaf size of the vegetable matter."

G3: The Pearson correlation coefficient, which has been chosen as a metric for the performance of different albedo corrections here, is hard to interpret (correlation between albedo and Delta_XCH4 (TROPOMI-GOSAT) is almost the same for the uncorrected data and the proposed correction in the month of December). Please explain your reasoning in more detail and provide additional metrics to measure how your new correction performs in comparison to the existing corrections by both Lorente et al. and Balasus et al..

Theoretically there should be no correlation between the surface albedo and the methane mixing ratio. The fact that we observe a correlation means that the instruments and measurement techniques are introducing a bias, such as in the uncorrected, Lorente et al. and Balasus et al. lines in Figure 3. Our correction remains between -0.1 and 0.1 Pearson correlation value for all months, indicating that our correction is the least biased with respect to surface albedo. The uncorrected data being similar to our correction in December is not an issue, there is simply less albedo bias in the winter months, we believe this is related to agricultural land-use in the area not affecting winter months as much as this is not a growing season at this latitude.

We prefer not to devise a new metric, as this is the one used in Lorente et al. 2021 and is the most relevant metric to the topic of the manuscript. We have added more clarifying text to the manuscript to make interpretation of the Pearson correlation more clear.

"To evaluate against the other models, Pearson correlations were calculated and presented in Fig 3a where different constructions of Pearson values have been unified according to Tab. S2, where the Pearson correlations have been calculated the same ways as in Fig. 2 with the correlation between GOSAT/TROPOMI and surface albedo. To reiterate, a Pearson correlation of 0 is

the preferred value, as the difference between the two data sets does not depend on surface albedo. The surface albedo is the SWIR albedo as retrieved by TROPOMI."

M1: Line 15-17: the 5-6 ppb correction occurs only at some times during the year. Be more specific here and in the text, for instance in Line 291-293.

Amended, the second instance was removed entirely.

"requiring a correction of 5-6 ppb larger than areas covered in water-intensive crops in the summer."

M2: Line 44-46: Rephrase this sentence to make sure that readers understand that a) the performance of proxy retrievals is not generally unaffected by albedo and b) GOSAT XCH4 can also exhibit some (weak) albedo bias, but less than TROPOMI XCH4 due to a number of reasons (imaging grating vs FTS, spectral resolution and band, etc.).

Amended, this is also repeated in line 116 of the original manuscript

"There have been several recent updates to the dataset to mitigate this albedo effect using TROPOMI retrieval data over areas without emissions, and also by comparison with proxy retrievals from GOSAT, which are much less affected by surface albedo"

"This dataset has been used extensively before as a measurement that is less affected by changing surface albedo"

M3: Line 85-86: "Satellite methods for…are going to be biased…" -> "Satellite methods for…can be biased…"

Amended

M4: Line 86-88: Rephrase, because CH4 emissions from oil and gas operations can be strongly time-dependent.

We prefer to keep this sentence as is, it is not uncommon to assume that emissions from O&G operations do not fluctuate considerably with the seasons, see:

Wilson, C.; Chipperfield, M. P.; Gloor, M.; Chevallier, F. Development of a Variational Flux Inversion System (INVICAT v1.0) Using the TOMCAT Chemical Transport Model. *Geosci. Model Dev.* **2014**, *7* (5), 2485–2500. https://doi.org/10.5194/gmd-7-2485-2014.

Shen, L.; Gautam, R.; Omara, M.; Zavala-Araiza, D.; Maasakkers, J. D.; Scarpelli, T. R.; Lorente, A.; Lyon, D.; Sheng, J.; Varon, D. J.; Nesser, H.; Qu, Z.; Lu, X.; Sulprizio, M. P.; Hamburg, S. P.; Jacob, D. J. Satellite Quantification of Oil and Natural Gas Methane Emissions in the US and Canada Including Contributions from Individual Basins. *Atmospheric Chemistry and Physics* **2022**, *22* (17), 11203–11215. https://doi.org/10.5194/acp-22-11203-2022.

M5: Line 98-100: move this sentence to section 2.2 ?

Amended

M6: Figs. 2,3/ Lines 175-181: Fig. 2 indicates that a small bias exists only in winter ($|R| > 0.1$). So based on your threshold in Pearson's correlation coefficient, no correction would be needed in

summer? This appears to be supported by Fig. 3 which seems to indicate that the Lorente et al. correction works well in 6 out of 12 months.

The reviewer is correct, the Lorente et al. correction falls within the ideal Pearson value range for about half the year. This speaks further to the "seasonality" required for a more accurate correction. Indeed, the Balasus et al. correction falls within this range much of the time that the Lorente correction does not, so someone could put together a combined Balasus-Lorente correction and it may be mostly correct most of the time.

M7: Line 198-201: Please provide more reasoning with respect to the choice of the threshold in |R| and possibly replace the reference Kuckartz et al. 2013, since this German reference may be a challenging resource for many readers (if you don't replace it, double-check the page number).

The choice of Pearson R value is largely arbitrary, but was selected with the idea that essentially nobody would argue that a Pearson R value <0.1 was anything other than negligible. We have altered the language in the manuscript to reflect this point and have updated our references to more accessible and appropriate sources.

"When all data are used (Fig 2a) the Pearson correlation is indeed calculated to be low, i.e. below a threshold of 0.1, which we chose here as a target value for minimal correlation between SWIR surface albedo and the albedo corrected methane retrieval. Though the significance of Pearson coefficients is up to interpretation, most would agree that a value of <0.1 is negligible (Akoglu, 2018; Schober et al., 2018)."

Akoglu, H.: User's guide to correlation coefficients, Turk J Emerg Med, 18, 91–93, https://doi.org/10.1016/j.tjem.2018.08.001, 2018.

Schober, P., Boer, C., and Schwarte, L. A.: Correlation Coefficients: Appropriate Use and Interpretation, Anesthesia & Analgesia, 126, 1763, https://doi.org/10.1213/ANE.0000000000002864, 2018.

M8: Fig. 4: Can you comment why albedo in the NIR appears to be more important than the albedo in the SWIR? This appears counter intuitive.

We cannot say with 100% certainty why the NIR albedo appears to be more important than the SWIR albedo due to the translucent-box nature of the machine learning algorithm. As described in the manuscript, we define our correction as a SWIR albedo correction due to our choice of correction validation. Over the course of this work we trained dozens of sets of models using various hyperparameters and other changes, but the models we chose to use were the ones where the correlation between surface albedo SWIR and GOSAT/TROPOMI was minimized. This does not mean that the SWIR albedo is necessarily the most or even an important factor in the actual model correction. The higher importance of the NIR albedo suggests that the NIR albedo variable has a larger impact on the final output value than the SWIR albedo. While this does not make intuitive sense knowing that the SWIR albedo is more likely to affect the actual measurement of the XCH4, the model does not have this knowledge to bias its decision making. We have clarified this further in the manuscript.

"Figure 4 shows that other features may be more important than the surface albedo SWIR in the actual model calculation. "importance" in a ML model is the magnitude of effect that variable has on the final output value of the model. The variables that appear higher on the y-axis than "surface albedo SWIR" tended to be more important and should be analyzed as well. Some of these

variables have clear reasonings as to why they are more important: XCH4 apriori, XCH4 corrected, and XCH4 are all the measurements of methane mixing ratio that were either priors for the TROPOMI measurement (XCH4 apriori) or direct measurements of the methane mixing ratio by TROPOMI (XCH4 and XCH4 corrected). XCH4 and "XCH4 corrected" directly measure methane mixing ratios via TROPOMI, serving as primary data sources for our predictive models. The reasoning for other important variables is not so clear: "surface albedo SWIR precision" and "chi square SWIR". The precision of the surface albedo SWIR measurement being important was not expected, but may be the result of a well-trained model successfully making the association between the SWIR albedo measurement and its precision. A less precise measurement would be less heavily relied upon for the model's predictions, so the importance may come from the association between the precision measurement and how much a particular measurement affected the model during training."

Technical comments:

T1: Line 55: "Kansas. (Petron et al...)." -> "Kansas (Petron et al ...)."

Amended

T2 : Line 61 : (CAFO)s -> (CAFOs)

Amended

T3 : Line 64 : Collocation -> Colocation

Amended

T4: Line 119 : Balasus et. al. (Balasus et al., 2023) -> Balasus et al. (2023)

Amended