Reviewer-2

Xie et al. present an ML approach for performing XCO_2 retrievals based on OCO-2 measurements. Their work follows the publications of David and Bréon who have before shown the general success of this type of method, applied to the same instrument. This work of Xie et al. implements a similar technique. The added novelty then seems to be the training on simulated data (via the ACOS forward model) that is covering a wider range of XCO_2, in order to mitigate the issue of the MLP not being able to follow the growth of atmospheric CO_2. There are some more general issues that the authors did not mention, such as quality assessment - how does a science data user know an ML-based retrieval is considered “good” and should be used in a study (or what the uncertainty on the estimate is)? Should people use the ACOS averaging kernels, since the ML-based method does not provide any?

Thank you for raising this important point. Estimating uncertainty is important for ensuring the reliability of machine learning models for satellite-based XCO_2 retrieval. While our current MLP-XCO_2 model does not provide direct uncertainty estimates, we recognize this as an area for future improvement.

A number of techniques exist for quantifying uncertainty in deep neural networks [1, 2]. However, generating robust uncertainty estimates requires a substantial volume of matched input and target data. The ideal dataset would contain numerous examples of OCO-2 observations paired with corresponding TCCON XCO_2 measurements across diverse conditions. This would enable sampling to assess the variance in our model’s predictions across the full distribution. Unfortunately, the limited availability of matched OCO-2 and TCCON data, especially for TCCON sites in “nadir” mode, means the current sample size is insufficient for comprehensive uncertainty quantification in the presented study.

Our primary contribution in the current study is demonstrating that simulated training data can complement real OCO-2 data to enable stable and accurate XCO_2 retrieval on future observations. While we do not currently provide uncertainty bounds, estimating prediction intervals is an important next step.

These have been clarified in the updated manuscript.

The manuscript is very well written and contains useful figures for the most part. In some places, minor re-wording or additional explanation would be appropriate and helpful, I have listed those below. My major comments would be the following:

[1] In Bréon et al. 2022, it is revealed that their ML approach inadvertently resulted in the NN "using" the weak CO_2 band as a proxy for geographical location and time. They thus removed the weak CO_2 band from the training process. However, the authors of this manuscript do indeed use the weak CO_2 band and have not explained as to how they overcome this issue. That would be important information, especially since they are following the general layout of Bréon et al. 2022. Maybe the issue does not manifest itself due to the much smaller region of interest, but the authors must show that.

[2] Related to (1), the "glitch" discovered in Bréon et al. 2022 was only found after they investigated specific features present in the original ACOS OCO-2 retrievals which were missing in the ML-based retrievals (specifically, strong plumes). The authors here only look at broad bulk-type statistics by comparing to TCCON and to ACOS OCO-2 via simple scatter plots. The strengths and weaknesses of the proposed ML-approach should be investigated more thoroughly by analyzing the results more carefully. Do the same biases appear in the derived retrievals, compared to the training data? Are global-scale features retained just like regional and small-scale ones? Are new biases introduced? There is possibly more to learn from the data than is shown in Figures 10 and 11. While the approach is promising, the authors should attempt to show an assessment of the quality of the ML-retrievals beyond the simple scatter plots.

My suggestion to the authors would be to (1) demonstrate that their approach, while using the weak CO_2 band, does not result in a loss of local features, such as plumes (analogous to Bréon et al. 2022, Figure 4). Further (2) they should demonstrate that their ML-based retrievals retains other characteristics of the training set (regional-scale, or local-scale; observe differences on maps etc.)

Thank you for your valuable suggestions and concerns. After incorporating feedback from you and other reviewers, we conducted an analysis using Explainable AI (XAI) on our original MLP-XCO_2 model and optimized the input parameters for the new MLP-XCO_2 model. While we did include the weak CO_2 band as an input in our model, our training results with simulated data did not show a significant impact of including or excluding the WCO_2 band on the detection of plume features. Instead, we found that the use of retrieved pressure was a crucial determinant in our training process. Thus, in addition to prior surface pressure input, we implemented a compact MLP-P model. Trained on historical OCO-2 data, this model infers surface pressure from spectral and angular information. Unlike MLP-XCO_2, MLP-P is expected to maintain relative stability and broad reliability in its surface pressure retrieval results in the future, according to our tests.

One possible explanation for this finding could be the idealized nature of the spectral data in our simulations, which differs from real-world spectra and is not affected by noise. This idealization might enable the ML model to more accurately capture the underlying relationship between spectral absorption and XCO_2. Conversely, using only OCO-2 product data, potential noise may mask weak absorption band characteristics, leading the ML model to focus on training better plume detection capabilities in strong absorption bands, as explained in Bréon et al. 2022.

Regarding the updates made:

- Plume Detection: Following your suggestions, we included tests for plume detection at potential high emissions sites, such as thermal power plants, using the updated MLP-XCO_2. Figs 12 and 13 in the revised manuscript demonstrate that our updated model detects sudden increases in XCO_2 in areas with plume emissions from OCO-2 spectra, providing substantial evidence of genuine atmospheric CO_2 retrieval from spectral data.

- Error Analysis: In the updated manuscript, we conducted an in-depth analysis of error metrics across different subregions within East Asia, as shown in Table 4 in the updated manuscript. These subregions, including Northeast, Northwest, Southwest, and Southeast, exhibit distinct geographical characteristics. Results show consistent predictive performance across these subregions, indicating uniform inversion effectiveness without introducing regional biases or instability. This indicates the model's robustness and its ability to handle diverse geographic and environmental conditions without sacrificing accuracy or performance consistency, despite variations in regional representation within the training dataset.

Minor suggestions:

[1] Line 4: ”low retrieval efficiency” and ”insufficient retrieval accuracy” are somewhat diffuse terms; I would simply mention challenges regarding computational efficiency.

Thanks for your suggestions. The sentence has been rephrased to “However, the next generation of greenhouse gas monitoring satellites is expected to face challenges, particularly in terms of computational efficiency in atmospheric CO_2 retrieving and analysis.”

[2] Line 38: ”not limited spatially or temporally” is not quite true - space-based platforms have observational coverage in space and time as a result of their orbital characteristics and other instrument parameters.

You are correct and my original description was indeed not accurate and has been changed to “satellite remote sensing offers broader spatial coverage and more flexible temporal observation.”

[3] Line 62: The interpolation of absorption coefficients for the calculation of optical property inputs for RT calculations are generally quite fast and can be done in less than a second typically, if the code is optimized enough (amongst other things). The computational effort is mostly driven by the RT calculations.

Yes, it is true that the majority of the computational cost in the overall forward model is devoted to solving the radiative transfer equation. This has been changed to “However, executing these complex optimizations requires computationally expensive interpolation of high-spectral-resolution gas absorption reference data and solving the radiative transfer equations in each iteration.”

[4] Figure 5: It is not fully clear to me what these represent. Did the authors take the outputs of the ACOS retrieval L2STD products and use them as input in their ReFRACtor-driven set-up? Please clarify.

Yes, you’ve got it right, and we apologize for any confusion arising from our unclear explanation. In Figs 5 and 6, we’re comparing the simulated spectra generated by the modified ReFRACtor model with the actual observed spectra from the OCO-2 satellite.

The information required to set up the ReFRACtor model is taken from the OCO-2 L2std v10r product. This has been explained more clearly in the updated manuscript.

[5] Line 338: When discussing the computational effort, it is mentioned that the forward model takes 12.16s to process two bands; but in forward-model “mode”, Jacobians are presumably not calculated, so the actual retrieval set-up would be even slower than the mentioned 36.48s. It would also be very interesting to learn how long the training process took!

Thank you for your concerns. Calculating Jacobians layer-by-layer does require significant computational cost in traditional retrieval models. The mentioned processing time of 12.16 seconds for two CO2 bands includes rapid Jacobian calculations within the ReFRACtor model. In a test with ”max_iteration” set to 1, the total time for computing all three bands and completing one fully optimized iteration was 23.13 seconds on an Intel 13700K CPU. The breakdown is: 8.85 seconds for Band 1 (O_2-A), 5.49 seconds for Band 2 (WCO_2), 6.54 seconds for Band 3 (SCO_2), totaling 20.88 seconds for the full radiative transfer (RT) computation.

The training process for machine learning models, which includes both generating simulated data and tuning the hyperparameters of the network, can be very time-consuming, requiring up to hundreds of CPU hours. The most computationally expensive parts - generating training data and training the model - are done beforehand. So while the upfront costs of developing a machine learning model can be high, once the model is trained, making predictions is extremely fast.

References

[1] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in neural information processing systems 30.

[2] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, M. Bennamoun, Hands-on Bayesian neural networks—A tutorial for deep learning users, IEEE Computational Intelligence Magazine 17 (2) (2022) 29–48.