First of all, we thank Christopher O'Dell (reviewer 1) for his effort in carefully reviewing our manuscript and his constructive comments.

Point-by-point answers to the comments of reviewer 1

General points

Authors: All points raised in the general part correspond to specific comments and are, therefore, discussed in the next section.

All comments

Reviewer 1: *Abstract: It would be helpful to say quickly in one sentence how the OSSE is set-up to make life "difficult" for the retrieval (includes plumes, realistic aerosols, etc) to make the results more meaningful.*
Authors: We added a sentence to the abstract that reads: "Since CO2M will not be launched until 2026, our study is based on simulated measurements over land surfaces from a comprehensive observing system simulation experiment (OSSE) that includes realistic meteorology, aerosols, surface BRDF (bidirectional reflectance distribution function), solar-induced chlorophyll fluorescence (SIF), and CO_2 and CH_4 concentrations."

Reviewer 1: *Also, has any other retrieval method demonstrated they can meet the accuracy and precision requirements of CO2M, or is this the first? If it is the first, it's important to say so. Though it looks like RemoTAP also does, based on Lu et al 2022, is that also your read? If so then I guess say nothing...*
Authors: Lu et al. (2022) and Noël et al. (2024) both conclude that their retrieval meet the requirements when applied to their simulated CO2M data.

Reviewer 1: *Abstract: I think it would be good to modify the abstract and conclusions to make it clear that you would have to re-train with real data once CO2M data are available, and that could change the storyline because of instrument artifacts, lack of sufficiently good training data (do you use TCCON, or a model, or... ?). So while this is a solid proof-of-concept, we can only really believe the amazing results once you apply it to real data somehow.*
Authors: We added the following sentence to the end of the abstract and a similar sentence to the conclusions: "While the presented results are a solid proof of concept, the actual achievable quality can only be determined once NRG-CO2M is trained on real data, where it is confronted, e.g., with unknown instrument effects and systematic errors in the training truth."

Reviewer 1: *Abstract: The sentence "We employ a hybrid learning*

approach that combines advantages of simulation-based and measurement-based training data to ensure coverage of a wide range of XCO2 and XCH4 values making the training data also representative of future concentrations." Is important! But it downplays the excellent work you've done here. Even if your NN approach didn't work, this one thing is great and could be utilized by any researcher trying to do direct ML-retrievals of GHGs. Maybe change to "We created a novel hybrid learning approach. . . ".
Authors: We rephrased one sentence of the abstract, which now reads: "We created a novel hybrid learning approach that combines advantages of ...".

Reviewer 1: *You could also add a sentence like "This method could easily be applied by future researchers training MLbased GHG retrievals, to avoid this common problem." Or something to that effect. I think it's just important to highlight this contribution to the literature, in addition to your actual ML model.*
Authors: We added the following paragraph to the conclusions: "It should be noted that the method could be applied to other instruments and applications. In addition to generating representative training data, spectra could also be modified, e.g., to study the ability of a machine learning model to predict changes in its target variable.".

Reviewer 1: *Abstract: I think you should also add a sentence to the effect of "Our ML model also provides accurate estimates of both the noise-driven uncertainties and the averaging kernels of XCO2 and XCH4 for each sounding." This is an important aspect of your model; not all ML models do this.*
Authors: We added to the abstract: "In addition, NRG-CO2M also provides estimates of both the noise-driven uncertainties and the averaging kernels of XCO2 and XCH4 for each sounding."

Reviewer 1: *L43: BRDF -> surface BRDF*
Authors: Done.

Reviewer 1: *Fig1: For the love of god, please convince your CO2M colleagues to work in W m-2 µm-1 sr-1 units. We messed this up for OCO2/3. You can right this wrong.*
Authors: I'm afraid it's too late for that. In the paper we aimed at consistency with the MRD. Personally, I also like the SI units W m-2 µm-1 sr-1 more, but I think the instrument scientists are into photons per second, probably because this has to be multiplied with the quantum yield of the detectors to calculate the signal. It could have been worse: the number of photons could be given in Mol and I assume that the imperial measurement system could still provide some really nice area and length units :)

Reviewer 1: *Page6: How are clouds modeled in the radiative transfer? Do they come from CAMS? From where does the effective radius for water and ice come? Clouds were excluded in Noel et al (2024) for the FOCAL tests. It*

seems like you are trying to include them here, so more details are welcome, since this is a specific difference to Noel et al.
Authors: On page 6, we included: “For the SCIATRAN RT simulations, we used pressure, temperature, specific humidity, cloud ice content, cloud water content, and cloud fraction from the ECMWF ERA5 reanalysis (Hersbach et al., 2020). Since we focus mainly on cloud-free conditions, we used static cloud microphysical properties for convenience, representing spherical water droplets with a gamma particle size distribution with an effective radius of 12µm and fractal ice particles with an effective radius of 50µm (Fig. 3 of Reuter et al. (2010) shows the corresponding volume scattering functions).” As discussed in Sec. 2.5.2, we include some of the cloudy scenes in the training data, especially, those with little cloud optical depth. This is intended to make the prediction less sensitive to residual cloud contamination and mimics imperfect cloud clearing of the training data set. However, as mentioned in Sec. 3, all quality analyses are performed only for cloud-free scenes (“Since the CO2M mission requirements are defined for cloud-free conditions, we filtered the evaluation data accordingly.”).

Reviewer 1: *Section 2.2. It’s not clear how these uncertainties in dry-air column, temperature, CO2 profile etc are used. Are you saying that you stochastically apply these terms to the truth training data before you simulate the spectra? Or that you stochastically supply them as input to the NN predictions, so the NN doesn’t have perfect knowledge of things like temperature profile, etc, when performing a retrieval on a given sounding? Please be clear. A flowchart might be helpful here. I think you ARE supplying these to the NN (you seem to say this in section 2.5) but please be explicit here. I think also saying WHY you need to supply this information is important.*
Authors: We added to Sect. 2.2: “’It should be noted that the input data for the RT simulations of the OSSE are free of noise. The main use of noise in our analyses is to generate realistically noisy training data. (Sect. 2.5).’ Additionally, we added to Sect. 2.5 the explanation: “It is important that the training data set contains noise, as all input and target features will of course be subject to inherent uncertainties during later training with real CO2M data. In addition, the noise supports generalized learning and suppresses overfitting.” Moreover, we added Fig. 1 of this document to the manuscript.

Reviewer 1: *Side note: I worry that you are telling your NN technique the answer by construction for each sounding, by supplying “truth data + gaussian noise” to it. It might be fine. But your “truth data + gaussian noise” for temperature, co2, surface pressure, etc, is not biased; there are no systematic errors. Instead, I would prefer that you had used a completely different model for your “prior information”. For instance, CarbonTracker for CO2, MERRA-2 for Temperature, humidity, surface pressure, etc. Your hypothesis would be that it doesn’t matter, but to me, that isn’t clear.*
Authors: The purpose of the training data set is, of course, to teach the network the correct data and their relationships. Systematic errors bear the

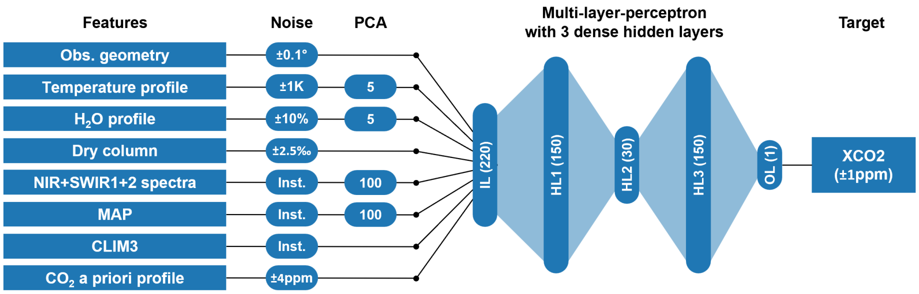

Figure 1: *Baseline* ANN training setup on the example of XCO2, including the amount of noise added to the training features and to the target variable (Sect. 2.2) and the PCA components used (Sect. 2.4). When training with actual measured data in the future, the addition of noise will be omitted. Inst=Noise of instrument model; IL=input layer; HL=hidden layer; OL=output layer.

risk that incorrect relationships are learned, which leads to a degradation in the prediction quality. This risk is particularly present if biases in the target truth correlate with input features (e.g. systematically too high CO2 concentrations at high latitudes, or over bright surfaces). However, reliable information on such biases and their covariance statistics do not exist which is why we have not considered them and assumed Gaussian noise for convenience. At least our results become better comparable to those of Noël et al. (2024) who also used an unbiased a prior and an unbiased training truth for their machine learning based post processing bias correction. In order to make the reader aware of this point, we discuss in the introduction: “Obviously, such errors would have the potential to reduce the accuracy of the prediction, but a realistic estimate of the to be expected error patterns of the training truth is difficult and beyond the scope of this study.”

Reviewer 1: *Near line 360. Feel free to add a contextual comment like: “For comparative purposes, the dry air column dependence for the operational OCO-2 XCO2 retrieval (v11.1) is roughly 85%, making it highly dependent on the accuracy of the prior meteorology, the prior surface elevation, and the instrument pointing (Jacobs et al., 2024, https://amt.copernicus.org/articles/17/1375/2024/).”*
Authors: Thanks, we added to section 3.3.1: “For comparison, the dry column dependence of the FOCAL CO2M XCO2 retrieval is 100% by design (Noël et al., 2024) and the dry column dependence of the operational OCO-2 XCO2 retrieval (v11.1) is approximately 85% (Jacobs et al., 2024).”

Reviewer 1: *Near line 420. I don’t get why removing the NIR band doesn’t increase the dependence on the dry air column to 100% ! Where is information on the dry column coming from? I guess from the fact that your prior co2 profiles are pretty good, so it can partially deduce the dry column from*

the co2 bands alone?
Authors: It cannot come from a too good a priori XCO2 because this would result in a larger dependency to the a priori. However, you probably meant the a priori profile shape. We agree that the CO2 profile shape has to be somewhat constraint in order to get dry column information from the CO2 bands. The a priori profile shape will contribute to this, but for the ANN, it would be sufficient that the CO2 profile shapes of the training data set vary not arbitrarily.

Reviewer 1: *Also, regarding the increase in the dry column dependence when you remove MAP, from 6% to 16%. Typical surface pressure uncertainties are on order 1-2 hPa (or often even smaller). +- 2 hPa is 2/1000 roughly, and 10% of this is 2/10000. For a typical XCO2 of 400 ppm, this would induce an uncertainty of 0.08 ppm. This implies that removing MAP from CO2M which add an additional +- 0.08 ppm uncertainty to XCO2, due to errors in the prior surface pressure, relative to the with-MAP case. Which basically means that, according to your analysis, MAP really is not necessary. That's a pretty big conclusion that you are currently glossing over. Please address this directly in the manuscript. Presumably its due to some assumption you've made? FYI this also affects your interpretation in the conclusions (near 520), where you are implying that this is an important difference for the no-MAP case. It's really not, honestly. OCO-2/3 would kill to only have a 15% dependence on the dry air column, which leads to nearly negligible errors in the target gases.*
Authors: Within the conclusions, we modified the corresponding paragraph which now reads: "This had an apparently small effect on accuracy and precision, which is not consistent with the results of Lu et al. (2022), whose retrieval method became significantly less accurate under these conditions. We can only speculate about possible reasons for this. i) We use a different aerosol microphysical model, which is consistent with the MACC aerosol model, but is less complex than the one used by Lu et al. (2022). ii) Their CO2I-only retrieval method is fundamentally different from ours and also from FOCAL, which may result in different sensitivities to aerosol-induced biases. In this context, it should be noted that our CO2I-only results are in good agreement with those of Noël et al. (2024), suggesting that it may be possible to meet the CO2M mission requirements without using MAP. iii) The statistics computed by Lu et al. (2022) to quantify the systematic and stochastic errors differ from those computed by us. However, we observe that the dependence of the XCO2 prediction on the dry column increases when MAP is not used, which may introduce systematic errors of the order of 0.1 ppm in reality when perfect knowledge of the dry column cannot be expected."

Reviewer 1: *Near Line 470, and Figures 10+11. Can't you plot the AK-corrected Truth minus Prediction, instead of straight truth – prediction? You should! I *always* do this in my OSSE experiments, it is important. It would also show if your hypothesis is correct on the source of this hotspot in the difference plot of figure 11. In fact a comparison of these two plots (with and without AK-correction) would be very illuminating. Your statement on using* *the true profiles as prior comes close to accomplishing this, but is not nearly as powerful. Plus, you are expecting modelers to make the AK correction; therefore I think It's important to set a good example and do the same, and show the effect when you don't.*

Authors: When AKs are taken into account, the difference between modeled and true XCO2 is

$$\Delta X = \hat{X} - \sum w_i[c_i^{apr} + A_i(c_i^{mod} - c_i^{apr})] \tag{1}$$

where $\hat{X}$ is the retrieved XCO2, w is the weighting of layer i, c^{apr} is the a priori profile, and c^{mod} is the model profile. Most of our analyses have been done with an a priori equal to the truth, i.e. MACC. In this case, c^{apr} becomes c^{mod}, so that the difference between retrieved and true XCO2 becomes

$$\Delta X = \hat{X} - X^{mod}. \tag{2}$$

This is the quantity we analyze to assess the systematic errors, as shown in Fig. 6 and 7. This means using the truth as a prior has the advantage that all deviations from the truth can be directly attributed to retrieval deficiencies without explicitly accounting for the AKs. However, it has the disadvantage that it rewards retrievals that put little weight on the measurement and much weight on the a priori. In order to demonstrate, that this is not the case here, we performed the anaylses of the Berlin scene on purpose with a constant a priori so that it is clear that the retrieved XCO2 variability only comes from the measurement but not the a priori.

If we understand the comment correctly, you are suggesting to also show results for the Berlin scene with AKs applied. In this case, ΔXCO2 would become

$$\Delta X = X^{con} - \sum w_i[c_i^{con} + A_i(c_i^{mod} - c_i^{con})] \tag{3}$$

where X^{con} is the retrieved XCO2 using the constant a priori and c^{con} is the constant a priori profile. Using the AKs, we can compute X^{con} from the retrieval result $\hat{X}$ obtained using the model as a priori:

$$X^{con} = \hat{X} - \sum w_i(1 - A_i)(c_i^{con} - c_i^{mod}) \tag{4}$$

so that

$$\Delta X = \hat{X} - X^{mod}. \tag{5}$$

This equation is the same as Eq. 1 which means, that the difference between the prediction using the truth as a priori and the model equals the difference between the prediction using the constant a priori and the model with AKs applied. In other words, instead of applying the AKs to the model, we can also use the truth as a priori (as, e.g., in Fig. 6 and 7). We added the corresponding figures to the appendix of the manuscript (see Fig. 2 and 3 of this document).

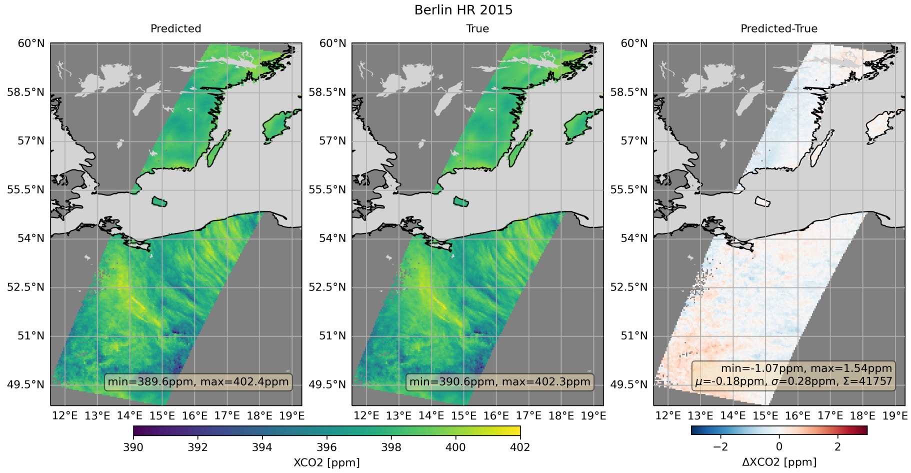
Berlin HR 2015
Predicted
True
Predicted-True
60°N
58.5°N
57°N
55.5°N
54°N
52.5°N
51°N
49.5°N
12°E 13°E 14°E 15°E 16°E 17°E 18°E 19°E
min=389.6ppm, max=402.4ppm
min=390.6ppm, max=402.3ppm
min=-1.07ppm, max=1.54ppm
μ=-0.18ppm, σ=0.28ppm, Σ=41757
390 392 394 396 398 400 402
XCO2 [ppm]
−2 0 2
ΔXCO2 [ppm]

Figure 2: As Fig. 11, but using the true CO_2 concentration profiles as a prior instead of their scene-wide average.

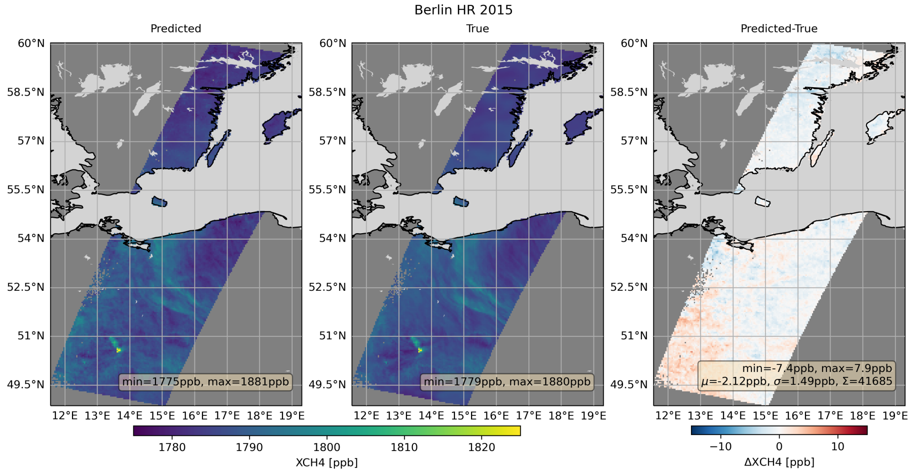

Figure 3: As Fig. 12, but using the true CH_4 concentration profiles as a prior instead of their scene-wide average.

Reviewer 1: *L502: short correlation length parts -> or short correlation length parts*
Authors: Done.

Reviewer 1: *I think the conclusions section really needs a paragraph on what it would take to "operationalize" this algorithm for real satellite data. Presumably you would train it on observed spectra, along with your method to extend it to larger truth values of XCH4 and XCO2? What would you use for the training truth: TCCON, Models, something else? Would your methods to get at the AK and posterior Xgas uncertainties still work? Would you have any reason to expect worse performance?*
Authors: We replaced the last paragraphs of the conclusions, which now reads:

"In order to use NRG-CO2M to retrieve XCO2 and XCH4 as well as the associated uncertainties and averaging kernels from real CO2M radiance measurements, once available, the PCAs and the training of the ANNs would have to be repeated with real data. In this case, the training truth could, e.g., consist of model data confirmed by an ensemble of models as done for NASA's OCO-2 XCO2 bias correction (O'Dell et al., 2018) or by corresponding TCCON measurements as done for FOCAL's GOSAT and GOSAT-2 XCO2 bias corrections (Noël et al., 2021). We expect that at least one full year should be used for training, although the modification of the training spectra makes them representative of a wider range of atmospheric conditions.

In the analysis of real data, several effects, the detailed investigation of which is beyond the scope of this paper, may lead to somewhat degraded retrieval quality. These include unknown systematic errors in the training truth, a priori, and met profiles, non-ideal sampling of the training data set, and potential instrument or RT features that are not well approximated by our spectrum modification method. Therefore, the actual retrieval quality achievable can only be determined after NRG-CO2M has been trained on and applied to real data.

However, due to the quality achieved in the analysis of synthetic CO2M data, the proposed retrieval algorithm NRG-CO2M can be considered as a promising candidate to meet the high accuracy and precision mission requirements of CO2M while providing high data yield and negligible computational requirements, making it a valuable addition to the ensemble of conventional algorithms."

References

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Jacobs, N., O'Dell, C. W., Taylor, T. E., Logan, T. L., Byrne, B., Kiel, M., Kivi, R., Heikkinen, P., Merrelli, A., Payne, V. H., and Chatterjee, A.: The importance of digital elevation model accuracy in XCO_2 retrievals: improving the Orbiting Carbon Observatory 2 Atmospheric Carbon Observations from Space version 11 retrieval product, Atmospheric Measurement Techniques, 17, 1375–1401, https://doi.org/10.5194/amt-17-1375-2024, 2024.

Lu, S., Landgraf, J., Fu, G., van Diedenhoven, B., Wu, L., Rusli, S. P., and Hasekamp, O. P.: Simultaneous Retrieval of Trace Gases, Aerosols, and Cirrus Using RemoTAP—The Global Orbit Ensemble Study for the CO2M Mission, Frontiers in Remote Sensing, 3, https://doi.org/10.3389/frsen.2022.914378, 2022.

Noël, S., Reuter, M., Buchwitz, M., Borchardt, J., Hilker, M., Bovensmann, H., Burrows, J. P., Di Noia, A., Suto, H., Yoshida, Y., Buschmann, M., Deutscher, N. M., Feist, D. G., Griffith, D. W. T., Hase, F., Kivi, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Podolske, J. R., Pollard, D. F., Sha, M. K., Shiomi, K., Sussmann, R., Té, Y., Velazco, V. A., and Warneke, T.: XCO_2 retrieval for GOSAT and GOSAT-2 based on the FOCAL algorithm, Atmospheric Measurement Techniques, 14, 3837–3869, https://doi.org/10.5194/amt-14-3837-2021, 2021.

Noël, S., Buchwitz, M., Hilker, M., Reuter, M., Weimer, M., Bovensmann, H., Burrows, J. P., Bösch, H., and Lang, R.: Greenhouse gas retrievals for the CO2M mission using the FOCAL method: first performance estimates, Atmospheric Measurement Techniques, 17, 2317–2334, https://doi.org/10.5194/amt-17-2317-2024, 2024.

O'Dell, C. W., Eldering, A., Wennberg, P. O., Crisp, D., Gunson, M. R., Fisher, B., Frankenberg, C., Kiel, M., Lindqvist, H., Mandrake, L., Merrelli, A., Natraj, V., Nelson, R. R., Osterman, G. B., Payne, V. H., Taylor, T. E., Wunch, D., Drouin, B. J., Oyafuso, F., Chang, A., McDuffie, J., Smyth, M., Baker, D. F., Basu, S., Chevallier, F., Crowell, S. M. R., Feng, L., Palmer, P. I., Dubey, M., García, O. E., Griffith, D. W. T., Hase, F., Iraci, L. T., Kivi, R., Morino, I., Notholt, J., Ohyama, H., Petri, C., Roehl, C. M., Sha, M. K., Strong, K., Sussmann, R., Te, Y., Uchino, O., and Velazco, V. A.: Improved retrievals of carbon dioxide from Orbiting Carbon Observatory-2 with the version 8 ACOS algorithm, Atmospheric Measurement Techniques, 11, 6539–6576, https://doi.org/10.5194/amt-11-6539-2018, 2018.

Reuter, M., Buchwitz, M., Schneising, O., Heymann, J., Bovensmann, H., and Burrows, J. P.: A method for improved SCIAMACHY CO_2 retrieval in the presence of optically thin clouds, Atmospheric Measurement Techniques, 3, 209–232, https://doi.org/10.5194/amt-3-209-2010, 2010.