Review of "Retrieving the atmospheric concentrations of carbon dioxide and methane from the European Copernicus CO2M satellite mission using artificial neural networks"

This paper describes a neural-network (NN) based direct retrieval of column mean concentrations of carbon dioxide (XCO2) and methane (XCH4), directly from spectra from a simulated CO2M satellite, using all of its instruments: its spectrometer (CO2I), multi-angle polarimeter (MAP), and it's cloud instrument as well. It uses realistic OSSEs that include aersols (but not clouds, it seems) to form a challenging dataset, though the authors are clear they do not include instrument artifacts in their simulations. Their predicted XCO2 and XCH4 values are excellent, and exceed the performance requirements of CO2M. Additionally, they provide a method to give both posterior uncertainties and averging kernels for their predicted XCO2 and XCH4 values. Notably, they also present a novel method to extend their training dataset forward in time, in order to avoid the problem of needing to retrain every couple of years as greenhouse gas concentrations continue to increase.

This paper is well-written, and does not suffer from some of the problems of previous machine-learning based XCO2/XCH4 retrieval papers. They have described their methods well, and the conclusions are indeed promising. However, they do have an implicit conclusion that the MAP instrument doesn't seem to be required, at least under their assumptions, as the error statistics are similar with and without MAP, and their claim that the lesser dependence of the dry-air column with MAP doesn't stand up to further scrutiny. They need to more fully examine this implicit conclusion. After they have sufficiently addressed this and my additional comments below, I recommend the manuscript for publication, as it will be an important addition to the literature.

All Comments

- 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. 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…
- 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.
- 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…". You could also add a sentence like "This method could easily be applied by future researchers training ML-based 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.

- 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.

- L43: BRDF → surface BRDF
- Fig1: For the love of god, please convince your CO2M colleagues to work in W m^-2 um^-1 sr^-1 units. We messed this up for OCO2/3. You can right this wrong.
- 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.
- 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.

  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.

- 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/)."

- 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?

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.

- 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.

- L502: short correlation length parts → or short correlation length parts

- 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?