Review of "On the potential of a neural network-based approach for estimating XCO2 from OCO-2 measurements" by Breon, David, Chatelanaz, and Chevallier.

Overview

This paper follows on from a previous work, David et al. (2021, AMT). That paper used a neural network approach (NN) to retrieve column mean carbon dioxide concentrations (XCO2) from the Orbiting Carbon Observatory-2 (OCO-2) satellite, directly from the hyperspectral near-infrared radiances recorded by the satellite. This method is very much faster, and could possibly have smaller biases, than traditional Bayesian "full-physics" retrievals, which in particular require costly, multiple-scattering radiative transfer calculations to be performed for each retrieval. This new work shows that the previous NN could actually estimate the latitude and date of each observation with surprising accuracy, even though direct information about those quantities was not provided to the NN. Therefore, it implies that the previous NN was of little value. This new work in particular showed that plume features could not be recovered from the first NN, a necessary (but not sufficient) step to demonstate that the NN "works". The authors find that by removing the middle of OCO-2's three spectral bands, the NN can no longer easily reproduce the date (but still can reproduce the latitude with surprising accuracy), and can now reproduce plume features. This suggests the new NN is indeed properly inferring CO2 from the depth of the spectral absorption features contained within the OCO-2 radiances. The new NN is shown to compare equally well as a standard full physics approach (the ACOS retrieval) to ground-truth measurements from the TCCON network, and has slightly better precision than ACOS as well.

General Comments

The findings of the deficiencies of the previous NN (described in David et al., 2021, hereafter D21) are very illuminating and a welcome addition to the literature. In addition, in this new work they use all 8 OCO-2 cross-track footprints, instead of just a single one. In fact, they imply (but do not directly ever state) that they use a single NN for all 8 footprints, which would be a significant achievement. The accuracy and precision of the new NN against TCCON is impressive, as is the finding that the new NN seems to accurately identify and characterize plume features for local fossil fuel sources, such as power plants or urban areas. However, this paper has a number of shortcomings which must be addressed prior to publication.

Most importantly, like its predecessor paper D21, there are some hypotheses in this paper that are presented as solid truths but in fact may not be so. In my review of D21, I strongly argued that the presentation by the authors that the NN had learned how to independently estimate XCO2 from the spectra was simply a hypothesis, and suggested that they check well-known plume features (such as from large, isolated power plants) before publication. They did not do so, and argued against me. They were proven incorrect, and thankfully state that clearly in this publication. However, that does not mean that all the stated hypotheses in this new paper, again presented as truths, are indeed so. Primarily, that *now* the NN retrieval with the weak CO2 band removed really does accurately retrieve XCO2 in the way the authors think it does. Again, this is merely a hypothesis. Granted, it is supported by the fact that the model can reproduce plume features, but it is by no means proved. I believe it is certain that the NN is indeed taking some

information from the spectra features directly related to CO2 concentration. However, this does *not* preclude the possibility that there still may be other features in the spectra that the NN could be using to reproduce features of the CAMS model, features themselves which indeed may be incorrect in reality.

For instance, the CAMS model differs from other models in certain areas of the globe. It is possible that CAMS is more accurate than the others, but the reverse is also possible. Therefore, why risk using the CAMS model to train the NN over times & places where it significantly disagrees with other models? We are explicitly trying to figure out which model is more correct by using satellite data. If the NN is somehow replicating CAMS biases, we would have no easy way of knowing. And therefore, the NN results will always be suspect. Others (like me!) may suspect that the NN-derived OCO-2 values agree better with CAMS simply because it was trained on CAMS, not because CAMS is actually correct. I understand that the authors' goals are merely to show the potential of a NN approach. But this is also linked, I believe, to extremely careful training. If we have learned anything from D21, it is that the NN can learn ways to predict things in very different ways than you think it does. It tricked the authors in D21. It can do so again, unless the authors are extraordinarily careful and do many supporting checks to ensure that this is not the case. I'm not convinced that this is sufficiently done in the current manuscript.

One simple test is to retrain the data on relatively uninteresting soundings in places where there is not a lot of disagreement among models. For instance, OCO-2 results over the Amazon region and the Sahel region of Africa, as well as eastern China, are all areas of some disagreement and debate (see for instance Peiro et al., 2021, ACP, https://acp.copernicus.org/preprints/acp-2021-373/). Why not remove these areas from the training? Surely the rest of the globe has adequate ranges of surface albedo, viewing geometry, aerosols, etc., that XCO2 in other regions should be sufficient to teach the NN how to retrieve XCO2 in these regions? I would like to see more tests like this to strengthen the findings. If the authors insist that such tests are "beyond the scope of this work", then statements about success of this NN *must* be toned, or given appropriate caveats, prior to publication.

The other main critique is that the authors heavily rely on ACOS quality filtering to select soundings on which to retrieve. This is a major difficulty faced by all satellite XCO2 retrievals. ACOS uses a number of variables on which to screen data for retrieval, as described in detail by O'Dell et al. (2018, AMT). It is not at all clear how the NN could address this. The authors suggest that by using the difference from the NN-retrieved to the prior surface pressure, it would be "easy" to accomplish that goal. But they do not show this to be the case, and in my experience many other variables besides that one will come into play. Without other variables to help indicate quality (such as goodness of fit statistics, albedo mean and slope retrieval discrepancies, retrieved aerosol, etc), it is unclear if it is indeed possible at all to accomplish this with a standalone NN. This should be stated clearly in the discussion section, that this is an unsolved problem.

Beyond these critiques, there are additional questions & suggestions given below which must be adequately addressed prior to publication.

Specific Comments

L83: “The uncertainties

L79: “Our hypothesis was that the CAMS … model constrained by surface air-sample measurements provides a fairly accurate estimate of the atmospheric CO2 concentration, including the growth rate over multiple years.” Please provide evidence for this statement. (e.g., https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-GHG/CAMS73_2018SC2_D73.1.4.1-2020-v5_202109_v1.pdf, Chevallier et al., 2019, https://acp.copernicus.org/articles/19/14233/2019/).

L83: “The uncertainties on the modeling are small with respect to the range of XCO2 samples that is available in the multi-year dataset.” Please defend this statement quantitatively. How big are each?

L105: Was a single NN used for all 8 footprints, or did you train 8 different NN’s? Please state clearly in the main text. It’s relevant, because the line features move around due to the slightly different wavelength calibration of each footprint. Ie, channel 500 of footprint 1 is not at the same wavelength as channel 500 of the other footprints.

Section3 / Figure A1: Since one of the main points of this paper is to discuss the failure of the first NN and how it was improved, showing the failure in the main text is critical. Therefore, the failure of the first NN to find plumes should be figure 1 rather than A1. Also, because there can be “false plumes” in the OCO-2 data associated with dust or other aerosol features, it is important that you know that the plume seen by ACOS is *real*. How do we know that the multiple plumes in fig A1 are not some source of ACOS-induced bias? Therefore, this figure requires you to use a documented case caused by a known urban or power plant emission source. Many examples abound, for example Nassar et al. (2021, *RSE*, https://doi.org/10.1016/j.rse.2021.112579) and Reuter et al. (2019, *ACP*, https://doi.org/10.5194/acp-19-9371-2019).

Figure 3: Please list the fossil fuel sources of the plumes. If you cannot, please use other examples where the source is documented, again so that we know that these plumes are real (see previous comment). If possible, cite supporting sources.

L184: “Standard deviation of the latitude estimate”. I think the authors mean “Standard devation of the latitude error”. Please correct. Similarly for the statements about the longitude and date errors.

L183-215: Regarding the estimate of date & latitude. Can you please state whether the accuracy on these variables was independent of footprint or not? Ie, was it different for footprints 1-8 at all? Often, calibration artifacts such as bad pixels affect the different footprints a little differently, so if it is dependent on footprint, that would tell you if it was more likely to be some calibration artifact that the NN is keying off of for its estimates.

L223: Please repeat this analysis for the O2+sCO2 NN results (sigma_lat = 8.9 deg, sigma_lon=57 deg, sigma_date = 195 days), and state the resulting XCO2 accuracy, to show that the inherent accuracy from latitude and date alone is relatively poor for that band combination, further justifying the second version of the NN.

L248: You may also wish to state that the use of the NWP surface pressure as input to your NN is further justified considering the fact that the ACOS algorithm also explicitly uses it in its posterior bias correction, and in fact it is the most important term in the bias correction (O'Dell et al., 2018).

L262: "there is no satellite data input to CAMS". The informed reader will know that this is not true for all versions of CAMS. FT20r3, for example, assimilates OCO2 rather than surface/in-situ data. As you report the standard deviation of your result vs. CAMS (0.85 ppm), it may also be interesting to report the same but for the CAMS version which assimilates OCO2. If your hypothesis is true, that standard deviation should be lower.

L291: Please define and justify the statement "significantly correlated". The R-value for land nadir is merely 0.39 as shown in your figure 7; which seesm to imply that only 15% of the innovation difference variance is common to the two datasets. Some of this may be due to instrument noise, which you could reduce by averaging up the data (say to all soundings that fall in a given 10-second block, as is commonly done by modelers, see for example Peiro et al., 2021, ACP, https://acp.copernicus.org/preprints/acp-2021-373/). Further, the best-fit line appears to fall significantly away from the 1-1 line. However, that could be to due more "noise" in the ACOS fit.

L322: I think you mean that the comparison to TCCON does not *suggest* favoring one satellite product of the other. It would allow it if there were any obvious difference, it just doesn't suggest it with this analysis.

L326: Your statement on the value of the satellite data relative to the CAMS model makes little sense. There are many models in addition to CAMS, and they disagree about many, many things of importance to the carbon cycle. The TCCON data seem to have limited value in resolving most of these questions, especially in the tropics where the TCCON data are incredibly sparse. In addition, the in-situ-driven CAMS results typically run 12 months behind real-time, while satellite data are available within 1 month of data collection. Indeed, this was the motivation behind the CAMS "FastTrack" (FT) product, which assimilates OCO-2 rather than in-situ data, and has been shown to compare equally well with independent aircraft data (Chevallier et al., 2019, ACP, https://doi.org/10.5194/acp-19-14233-2019). Please modify or remove this statement.

L335: Regarding your statement on the "agreement with CAMS": You seem to imply that the better agreement with CAMS for the NN implies that the NN product is "better" than ACOS. You simply cannot draw this conclusion when the NN was trained to agree with CAMS. If it didn't agree better with CAMS than ACOS, something would be wrong. The agreement with CAMS tells you literally nothing about the quality of the NN beyond the fact that it has been properly trained. Please rephrase this statement to reflect this fact.

Discussion section: Please also mention / highlight the fact that (if I've interpreted your paper correctly), the same NN training was applied to all 8 OCO-2 footprints. That's quite amazing. If so, it's necessary to perform a brief analysis on the quality of the XCO2 analysis from the 8 different footprints. Are they all comparable? If so, this is a remarkable result given that the NN does not "know" a-priori the wavelength grid of each footprint. If not, it is important to know if each footprint is required to be treated with a separate NN training. This is important for future sensors such as CO2M and GeoCarb, which may have 100s to 1000 different cross-track footprints (and thus training 1000 different NN's may be challenging).

L400: Using the surface pressure difference to the met forecast *might* provide such a quality flag. It might not. It's a hypothesis that would need to be tested. ACOS uses many variables, both pre- and post-retrieval, as indicators of quality, of which surface pressure error is just one.

Technical Comments

L155: as input, the training → as input, and the training
L157: as → in that
L159: worrisome however. → worrisome, however.
L160: well documented → well-documented
L160-1: local enhancement → local enhancements; plume → plumes
L162 : South African → South Africa
L195 : a combination of O2 band with either CO2 bands → a combination of the O2 band with either CO2 band
L212: provides an indirect information → provides indirect information
L240: shown on Figure 3 → shown in Figure 3
L253: leads to a slightly better → leads to slightly better
L295: remotely sensed → remotely-sensed
L321: agreements are → agreement is
L344: contrarily → contrary
L365: provides → provided (to keep with the same verb tense as this earlier finding provided motivation for the present study)