Review of David et al., "XCO2 estimates from the OCO-2 measurements using a neural network approach.", by Chris O'Dell.

This work details a fast, artificial neural network (NN) approach to retrieving surface pressure and the column-mean dry air mole fraction of CO_2 (XCO_2) from high-spectral resolution measurements in the near-infrared from the Orbiting Carbon Observatory-2 (OCO-2). Tradiationally, the most accurate XCO2 retrievals have been from semi-physical ("Full-physics") retrievals. These are typically iterative, and typically include accurate calculations of multiple-scattering from thin layers of clouds and aerosols, which makes them exceedingly slow. They also tend to be subject to importance biases (of order 1 ppm) due to forward model errors (such as spectroscopic or instrument calibration). A neural-network approach is extremely appealing because it automatically solves the speed problem (NN's are very fast, likely tens of milliseconds per retrieval, vs. minutes for a typical FP approach), and may solve some of the bias problem as well, because they simply train on the "right answer", and do not have to know the details of spectroscopy or instrument models.

General Comments:

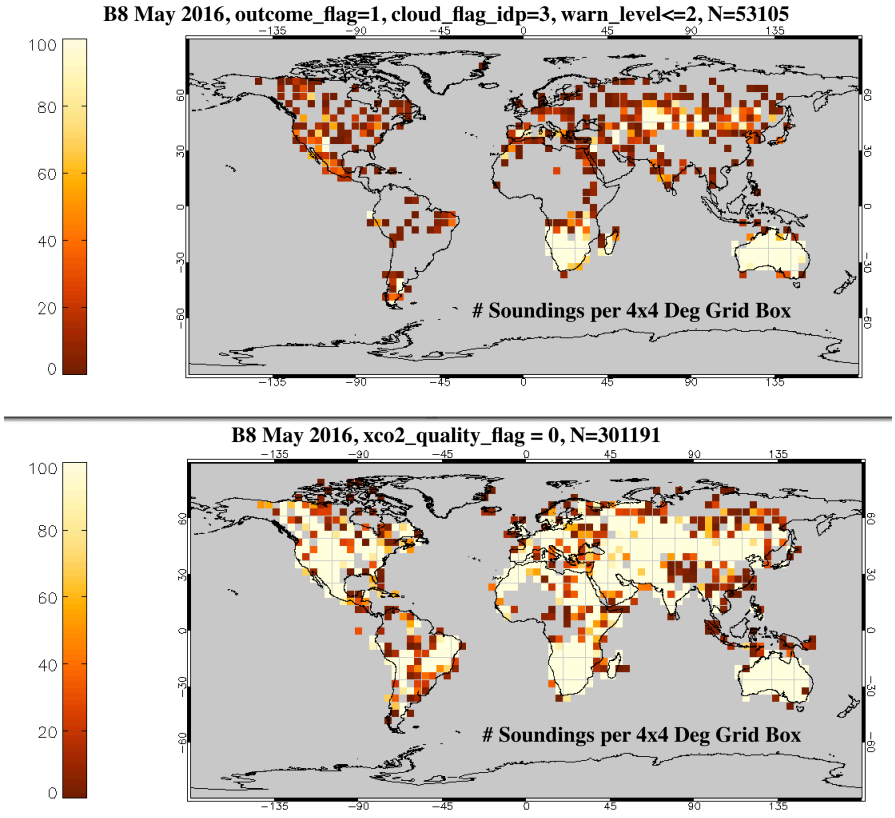
This work is the first serious effort to use NN's as applied to the XCO2 retrieval problem, and the author's mostly do a good job. However, there are a number of weaknesses and methodological problems in this work that need to be strengthened before I can recommend publication. I see this is a "major revision", but ultimately I believe this work can and should be published, as I believe (for the reasons above) that NN's hold great promise as applied to the XCO2-from-satellites problem.

Most significantly, as the first reviewer pointed out, it is difficult to ascertain from the manuscript alone exactly what the NN algorithm has learned. The authors trained it on a model (CAMS), and as a first validation, tested it against the same model. Using alternating months as training vs. testing is helpful, but the model certainly has deficiencies that are persistent longer than a month in certain regions, so testing well vs. that same model simply is not a validation. The only other real validation given is against TCCON, which seemed to perform well but because of TCCON's lack of good global coverage, again it is very hard to tell how well the model performs globally in any real sense. I am also worried about having to re-train the model every year to deal with the ~2 ppm secular increase, which over a mere 4 years is roughly equal to the entire global variability of XCO2. I believe it would make their argument much stronger if they ran their algorithm on a few select powerplant cases where the enhancement in the plume is reasonable well-constrained. This is possible for power plants with good bottom-up emissions esimates, and cases where the wind speed is reasonably well-known. See for instance Nassar et al. (2017) for some sample cases. If the NN doesn't see a plume at all, we know that it hasn't been properly trained; if it does, it will dramatically bolster the arguments in this work. Arguing that just because the NN doesn't have direct access to location or time information does not mean it cannot indirectly learn other relationships that allow it to appear to learn well. This is a hypothesis, not a proof that it has learned what you think it did.

Related to this, I'm somewhat concerned by training on CAMS and then considering to use the resulting XCO2 to correct CAMS. To show that your method works, you almost need to run a full (and fairly complicated) OSSE where you have a true world, a CAMS-like model world with some CO2 errors (spatially and temporally correlated), train on the latter, and then see if you can recover the former with the NN. I don't see how this is guaranteed to work, honestly. How do you know that you won't somehow reproduce systematic errors in CAMS by using the NN approach? You state in the text that you tacitly assume that CAMS errors are not correlated OCO-2 spectra in given areas and for given months. But because the CAMS errors (likely of order 1 ppm) are of a similar magnitude as the XCO2 signal, it is important to point out that this is merely an assumption, and more extensive validation (or a detailed OSSE study) is necessary to prove it.

Also, you claim to use the "ACOS cloud flag", which you say has values 0, 1, 2, or 3, as a way to define both your training and testing data sets. I think you mean "PreprocessingResults/cloud_flag_idp" in the L2Std file. If this is correct, please know that this flag is little used by the community. In fact, I've never heard of anyone using it, actually. It was defined about a decade ago for GOSAT and not really touched since then (I verified this with the author of the code that defines it). It has never been carefully validated and it appears to be extremely restrictive ("co2_ratio_idp" must be between 0.99 and 1.01 to pass, which is extremely restrictive and appears to cut out entire regions of the globe). Further, using outcome_flag=1 is also quite restrictive. Can you please comment on these flags, and why you didn't use the far simpler ACOS xco2_quality_flag, which is widely used by the community and is the generally adopted quality flag to use? In the plot below, I have attempted to show the differences between the two approaches for May 2016. I had to match the L2Std files (v8r) to the Lite files (v8), so there may be some differences to what you used in your work, but the general conclusion is that you miss a great deal of data with the highly restrictive data set you are working with. Thus, because it is so restrictive, it may be a far easier task than what ACOS tries to do, which is get the best error possible for the xco2_quality_flag = 0 dataset, which is roughly 6 times bigger.

Finally, in section 2 please give the sources of ACOS/OCO-2 data you used with more specificity. What specific versions and datasets of OCO-2 did you use? V8r, or just V8? Did you use L2Std files, L2Dia files, Lite files, etc?



What you train on is pretty critical. I think you should at the very least show a sounding density map of your training (and testing) set. Further, I think you should carefully explain your reasoning on how you choose the filtering. You must at least mention the xco2_quality_flag, and ideally you would retrain (or at least test) using this, if you aren't going to define your own quality flags. If you choose to train using very restrictive (clear-sky conservative) filters, please explain this is more detail.

Also, both outcome_flag and warn_level (which you use for filtering) come from the ACOS L2FP product (cloud_flag_idp comes from a fast, preprocessor code, the IMAP-DOAS Preprocessor, or IDP). It would be much better if you could avoid this entirely, because currently you are throwing away all the soundings that didn't converge or were skipped by the ACOS team, which relies on all the peculiarities of our specific algorithm. To make a useful NN algorithm, it ultimately must be independent of any full physics algorithm, unless you want to *train* on soundings that pass some smaller subset of data that includes L2FP quality flags, but *test* on a more complete set of soundings that doesn't use any L2FP quality flags. But you do not appear to do this.

Specific Comments:

L20, Abstract: I don't think TCCON is a "sunphotometer". That kind of implies more moderate resolution measurements. How about "reference ground-based spectrometer" or something similar?

L80: Please clarify “a limited set of spectral elements”. Make clear these are the solar (Fraunhofer) lines you’re talking about.

L85 (near there) : Do you try to mask deep solar lines as well (to remove their Doppler effects)? Please clarify, with a why or why not.

L90 (near there): Might you include the polarization angle directly to the NN, in addition to / instead of the relative azimuth? That might work even better.

L96: As the readers of this article are likely not NN experts, please discuss the pros and cons of # of hidden layers vs. # of neurons. Also please discuss how was the number 500 chosen or optimized.

L113: I think you mean “1 hPa”, not “1 ppm”.

L125: XCO2 is defined as weighted by the number of dry air molecules per square meter in each layer, not the pressure width. This can be shown to be roughly proportional to dP * (1-q) for a given layerm (e.g., O’Dell et al., 2012), where dP is the pressure width and q is the specific humidity in kg/kg. Please recalculate your model XCO2 using this more standard formulation, if possible, or defend your non-traditional XCO2 definition. The differences are generally small (tenths of a ppm), so it is defensible, but if you can be correct, it is best to do so.

L138/Figure 1: This is supposedly for the evaluation dataset, but includes N=381k soundings? In section 2, you say the evaluation dataset only includes 155k soundings. So something is wrong – please explain or fix.

L138-152: Based on Fig 5, there appears to be some problem in the surface pressure retrieval over mountains, specifically a high bias generally in these regions (visible over the Tibetan Plateau and the U.S. Rocky Mountains). Please discuss. I suggest that including the surface elevation in the NN may be a good idea, though technically it shouldn’t be necessary.

Regarding TCCON comparison: It would be useful to include the following statistics for ACOS, NN, and CAMS vs. TCCON: Overall Mean, Overall Std Dev, and Stddev of Station mean biases. These are useful to evaluate accurate vs. TCCON in simple statistics. See for example Fig 18b in O’Dell et al (2018). It shows a mean bias of Nadir Land observations vs. TCCON of 0.30 ppm, and a stddev of 1.04 ppm (it has not calculated the stddev of the station-level mean biases; some groups do this, others not). Finally, it doesn’t look like you’re applying the averaging kernel (AK) correction when comparing ACOS to TCCON. This typically makes the stddev about 0.1 ppm better. If you do not make this correction, please point this out in the text.

Fig4: Please include a horizontal dashed line so we can see the zero-level. Also, please be clear in the caption or the text if the ACOS and NN are sounding-matched. Typically, when we

compare ACOS to TCCON, which use all xco2_quality_flag=0 data. If you were to do this, it may change your results for ACOS vs. TCCON (though better vs. worse, I'm not sure).

Technical/Grammatical:
L92: "Although, the NN technique" → "Although the NN technique"
L124: "For each OCO-2 observation"
L129: "cosmic flux anomaly" → "cosmic ray flux anomaly"
L151: "lowest pressure" → "lowest pressures"
L181: please replace "classical" with "standard" or "traditional"
L250: "that are described in this paper."