Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-145-RC1, 2021

Comment on amt-2021-145

Anonymous Referee #1

Referee comment on "Assessing the feasibility of using a neural network to filter Orbiting Carbon Observatory 2 (OCO-2) retrievals at northern high latitudes" by Joseph Mendonca et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-145-RC1, 2021

Review of "Assessing the Feasibility of Using a Neural Network to Filter OCO-2 Retrievals at Northern High Latitudes" submitted for publication in *Atmospheric Measurement Techniques*

This manuscript presents a machine learning approach to QC filtering satellite-retrieved XCO_2, using collocated measurements from OCO-2 and from ground-based TCCON stations as an independent data source. A neural network (NN) is trained on the offsets between the two measurements to classify a retrieval as "good" or "bad" based on inputs of satellite retrieval variables such as albedo, solar and sensor zenith angle, surface elevation standard deviation, etc. NN-filtered retrieval bias, precision, and throughput are assessed by season and location, and compared to the standard product QC flag. Bias and precision were improved for most seasons/locations while increasing throughput in all seasons except summer.

The work presented here is a novel, well constrained application of machine learning to a tractable problem. The co-dependencies of retrieval quality on a large number of variables are difficult to discern and tease apart by traditional methods, making this an application well-suited to machine learning. The focus on one region (~high northern latitudes) with a very simplified question – is a retrieval of good quality (1) or not (0) – is likely a safe problem to tackle, while supplying plenty of data points on which to train the model. However, I do question whether it is appropriate to reduce this problem down to such a binary determination and wonder whether some lack of clarity in the results (e.g., numerous good retrievals being filtered out by the NN) may result from this choice. Overall, though, this is a well-written manuscript, with clearly presented figures and a well-described methodology. The topic is of interest to the readers of *Atmospheric Measurement Techniques* and could represent either a sound method for fine-tuning traditional quality control algorithms or a step towards using machine learning for quality control in the future. Below I describe further my major hesitation along with some more minor concerns to be addressed.

Major comments

It seems a bit of a mismatch to me that NN's would be chosen for an application that is reduced to such a binary distinction; did the authors try or consider other ML techniques such as self-organizing maps or regression trees? Alternatively, the problem could be posed such that the output values are on a continuum, such that values indicate some measure of confidence, e.g., Y=0 if OCO-2=TCCON, Y=0.5 if OCO-2 within +/- 2.5 ppm of TCCON, Y=1 if OCO-2 exceeds +/- 5 ppm, or similar. As the methods are described, it seems as if the training "target" values are strictly 1's and 0's, making the eventual NN outputs in the middle somewhat ambiguous, as referenced below.

I suggest the authors explain their reasoning for the choice of NNs with the binary result and/or discuss alternate setups that might be attempted for future directions.

Minor comments

L147: Was there any attempt to optimize the features being fed into the algorithm?

L205: How is the threshold of 0.1 determined?

L214: I do worry a bit that so many small values of XCO_2^{Diff} have values of Y^ >0.1. It may be true that the greatest density of points with low XCO_2^{Diff} exists at Y^ <0.1, but if you integrate along the XCO_2^{Diff} =0 line from Y^ =0.2-1.0 (in Fig. 3b), you get a non-insignificant portion of the total samples that should be considered "accurate," according to the TCCON data. Could you discuss the implications? If this means that the NN filter is being overly restrictive, I just wonder if this justifies a re-framing of the problem, re:the first comment, to use a different ML algorithm or change to a non-binary determination.

Fig. 3: Should XCO_2^{Diff} have units of ppm? It would be helpful if this were indicated.

Also, if 2.5 ppm is the threshold for deeming a value of XCO_2^{Diff} as "good" or "bad," it would make sense to indicate those values (+/- 2.5) as horizontal lines in both panels.

Likewise, since Y^ values are considered good if <0.1, bad if >0.1, indication of that threshold as a vertical line would also be helpful.

L216: The influence of the proportion of "good" vs. "bad" training data could be tested, granted there are enough data available. Subsample the good retrievals so as to still be representative of various conditions (perhaps sample across percentile bins of each input feature) and allow roughly equal numbers of good points as bad points. It would be interesting to see how the results change; this is likely not be the best way to set up the NN training but could be illustrative.

L290: Could the authors discuss the considerations that go into the OCO-2 team's determination of the B10 qc_flag?

Along these lines, it seems as though, from this discussion and from Fig. 8, that the qc_flag may be too dependent on the presence of snow, when in fact there are other complicating factors, e.g., during summer, that should be more heavily weighted when checking the quality of the OCO-2 data. Meanwhile, some of the over-snow retrievals may contain better data than previously acknowledged. This could be a valuable contribution to the field if the authors agree this conclusion is supported by their analysis. If the lack of independent observations precludes confidence in this supposition, then please disregard.

L325: Another potential future direction I would offer is performing bias correction on the retrieved XCO_2. Recent modeling studies have moved in this direction, with forecasts of surface air quality being adjusted on a site-specific basis using machine learning and observations (e.g., https://acp.copernicus.org/articles/20/8063/2020/acp-20-8063-2020.html and https://acp.copernicus.org/articles/21/3555/2021/acp-21-3555-2021.html). It seems this approach may be applicable to satellite-retrieved species with available independent measurements.

Technical corrections:

L299: Instead of "topography," perhaps "variable topography" would be more clear?

Table 1, 4^{th}- and 3^{rd}-to-last rows: "form" should be "from"

Fig. 3 caption: “the all three” should be “all three”