Assessing the Feasibility of Using a Neural Network to Filter OCO-2 Retrievals at Northern High Latitudes

Referee comments are given in Blue.

Responses to comments are given in Black.

In the manuscript highlighted text is the added text and red crossed out text is deleted text.

Response to Referee 2

We thank the reviewer for the comments on our manuscript. Please see below for our responses.

Major Comment 1 - My only significant criticism on the method is the use of 0/1 quality parameters. Indeed (line 173) the target for the neural network is 0 when the XCO2 error is less than 2.5 ppm and 1 when it is larger than that. This means that a sounding with an error of 2.45 is considered as good as a sounding with an error of 0, while a sounding with 2.55 is as bad as that with an error of 7 ppm. I would have suggested to rather train the NN with a target that is a continuous function of the absolute error |TCCON-OCO2].

Similarly, I am surprise by the choice of the threshold at 01 (line 205) that is not justified. It would have been interesting to show the the standard deviation of the error as a function of the NN output (before the 0/1) classification. This would have provided arguments for the selection of the selection threshold (currently set at 0.1)

Training the NN with a continuous function of the absolute error between TCCON and OCO-2 would change the question from is the retrieval "good" or "bad" (in other words when does the retrieval algorithm work well vs when is it lacking) to a precision requirement. If the target is a precision requirement, the NN will filter out both "bad" retrievals as well as "good" retrievals where the signal is not good enough to achieve the required precision. This will lead to specific scene types (i.e snow) being filtered out which would lead to less soundings in the shoulder seasons. As long as this is understood this could be a viable path forward to augment the algorithm. The following was added on lines 333 - 339 to discuss this:

"The current implementation of the NN as a binary classifier was done to make the problem as simple as possible in order to filter out retrievals where the forward model of the retrieval algorithm is suboptimal, rather than scenes of high variance. A possible alteration to the algorithm would be to do the classification on a continuum where the output of the NN ($\hat{Y}$) would be related to the expected precision of the data. The downside to this configuration would be that the NN filter would be filtering out not only bad retrievals but also scenes with high variance. For example if you wanted a precision of better than 1 ppm, it is likely that you would not have any retrievals over snow getting through the NN, greatly reducing throughput in the winter and shoulder season at high latitudes."

The boundary set for the classification of the training data (2.5 ppm) does not translate to a hard boundary. As per the discussion from the comments (major comment 1 as well as minor comments 3 and 4) made by Referee 1, the way the NN works is by a majority determination from the training data. This means that after training the NN there could be retrievals that pass

the NN which have an absolute $XCO_2{}^{Diff} > 2.5$ ppm as long as the majority of retrievals with similar feature values have an absolute $XCO_2{}^{Diff} < 2.5$ ppm. Fig 3a shows that there are retrievals with absolute $XCO_2{}^{Diff} \sim 0$ ppm that don't pass the NN filter, which is attributed to no clear majority of retrievals with similar feature values that were classified as "good" in the training dataset.

The threshold of 0.1 was determined by not only looking at the standard deviation of the error as a function of the NN output (before manually setting it to 0 or 1), but also the number of major outliers and throughput at different threshold values (i.e. $\hat{Y}$). Fig 3a gives an indication of the standard deviation as a function of $\hat{Y}$ and also shows that as $\hat{Y}$ approaches 0.2 there are outliers as high as 20 ppm. Fig 3b also gives an indication of the standard deviation as a function of $\hat{Y}$ as well as the throughput for a given value of $\hat{Y}$. Setting a threshold value of 0.1 was found to give a good balance between increases in throughput, improvement in precision, and limiting the amount of major outliers getting through.

The following text on lines 211-213 was added:
"This threshold of 0.1 was determined by trying to balance throughput with degradation of precision as well as limiting the amount of individual retrievals with high absolute $XCO_2{}^{Diff} >$ 2.5 ppm passing the NN filter."

Minor Comment 1 – The abstract could mention the NN input features that seem to have the highest influence on the results

Extracting which features have the biggest influence on the entire coincident data set is not possible in the way that was done to show why the NN filtered out most of the data at Eureka. This is because when looking at the entire coincident data set there are a lot more features that become important because of different geophysical scenes provided at the different sites.

Minor Comment 2 – Some of the technical description of the NN approach (lines 150-160, lines 186-196) may not be needed in the paper as they are described in other documents

The authors believe that all essential details used to derive the results should be included in the manuscript so that anyone can reproduce the results given just the equations in the manuscript and supplementary material in combination with the OCO-2 and TCCON data sets. Also, this manuscript will serve as the main reference for future manuscripts that will deal with changes to the NN algorithm as well as using other training datasets.

Minor Comment 3 – Line 219: I do not think that "separated into two datasets" is appropriate as some elements of the original dataset are in none of the two while some others are in both

We see how this line can be misinterpreted. The following line "To validate the NN filtering, the validation data set was separated into two data sets. One data set was the OCO-2 bias-corrected XCO_2 values filtered using the NN filter and the other was filtered using the B10 qc_flag=0." was taken out and replaced with "To validate the NN filtering, the NN filter was applied to the validation data set and compared to the same validation data set but with the B10 qc_flag=0 applied to the soundings." on lines 236-238.

Minor Comment 4 – The paper shows the pass fraction as a function of 3 of the NN input parameters (features) (Figure 4). I assume the same has been done for the others. If not mentioned, I assume it means there is so significant variation. Please confirm

Yes, the same analysis and plots were done for all features. It was found that there was no significant variation for the rest of the features.

Minor Comment 5 – Figure 3a is impossible to read as they are two many datapoints. I strongy suggest to change the figure style, or to make a random sample of the datapoints before ploting

For fig 3, a red dashed line was added at $\hat{Y}$ =0.1 so that the reader can better compare the data that is considered "good" vs "bad". Other than that the figure was kept the same because the reader needs to be able to see how many major outlies (as well as their magnitude) gets through the NN filter as $\hat{Y}$ increases. This is important for the reader to see as this helps to illustrate the reason why the 0.1 threshold value was selected.

Minor Comment 6 – Figure 5-7 are difficult to read. I believe it would be better by showing two bars side my side rather than the plain/dash drawing

Figures 5-7 were plotted with the plain/dash bars side by side rather than over lapping (see below). The authors agreed that this way of plotting the figures did not enhance the clarity of the figures and find the overlapping bars easier to interpret. So figures 5-7 in the manuscript are unchanged.
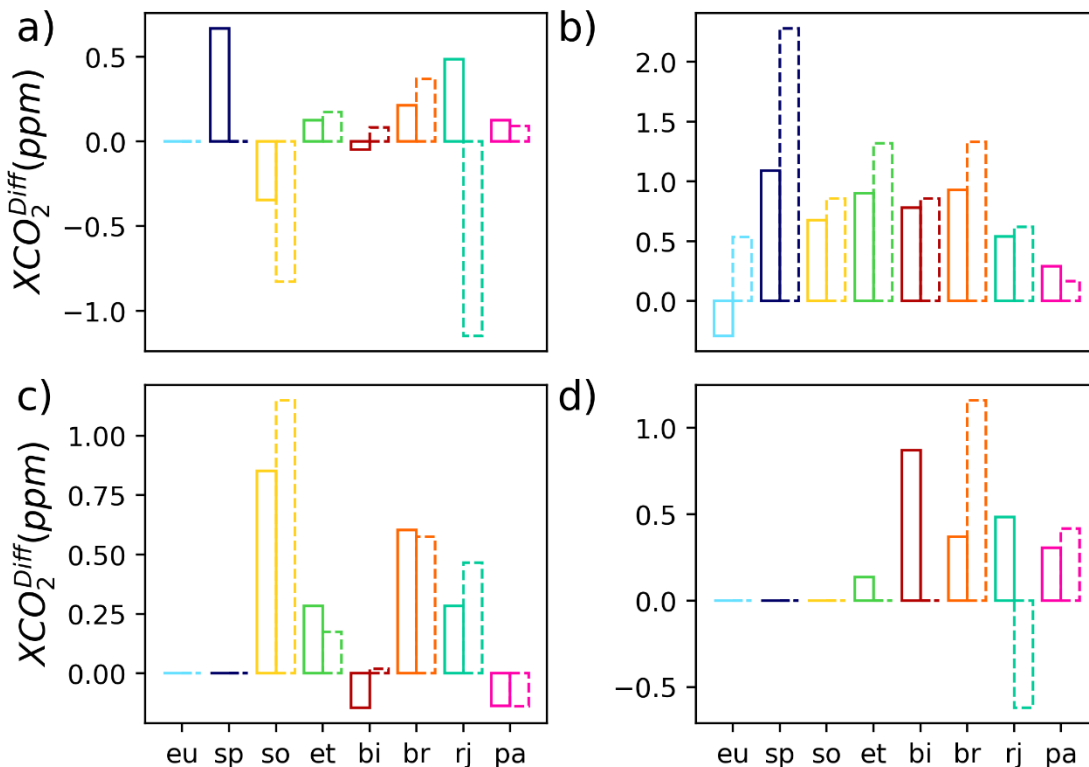
Figure 5 The bias at each site for the different seasons when the NN filter (bars with solid lines) and qc_flag (bars with dashed lines) is used to filter the OCO-2 retrievals in the validation data set. a) Spring (March, April, May). b) Summer (June, July, August). c) Fall (September, October, November). d) Winter (December, January, February). Note the different y-axis ranges for each plot. Note that bars that show a bias of zero are due to no data rather than a bias of zero.
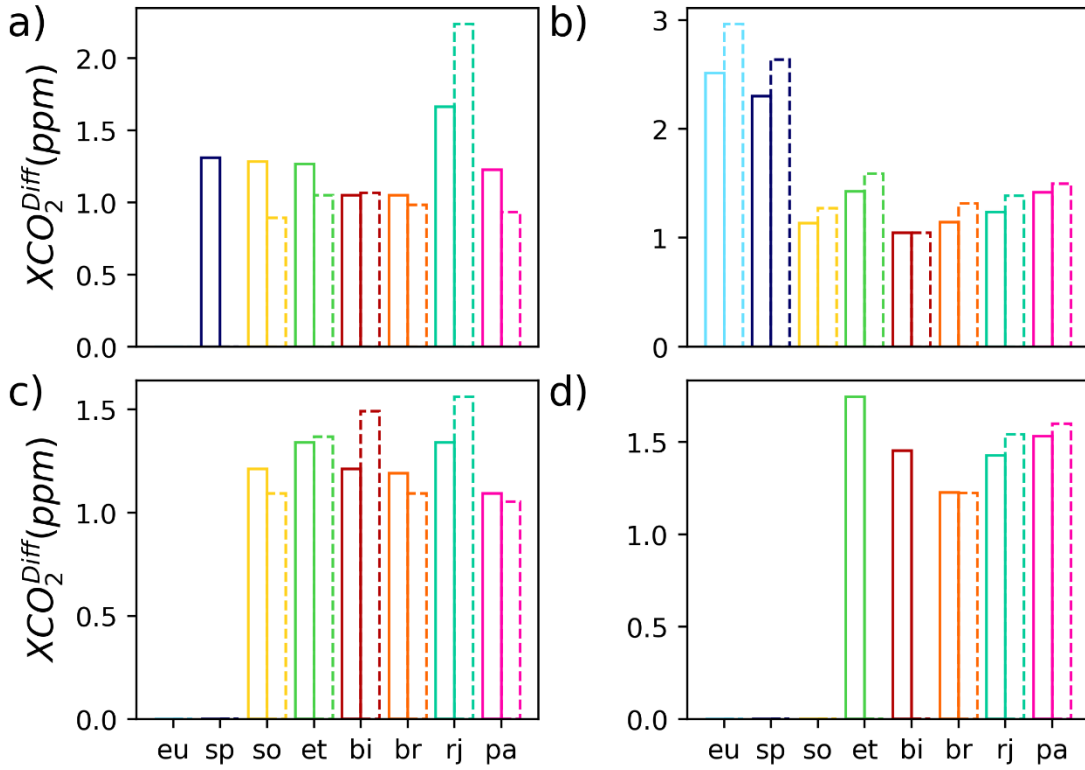

Figure 6 Same as Figure 5 but shows the precision at each site for each season. a) Spring (March, April, May). b) Summer (June, July, August). c) Fall (September, October, November). d) Winter (December, January, February). Solid bars indicate the NN filter, and dashed bars indicate the original B10 qc_filter.
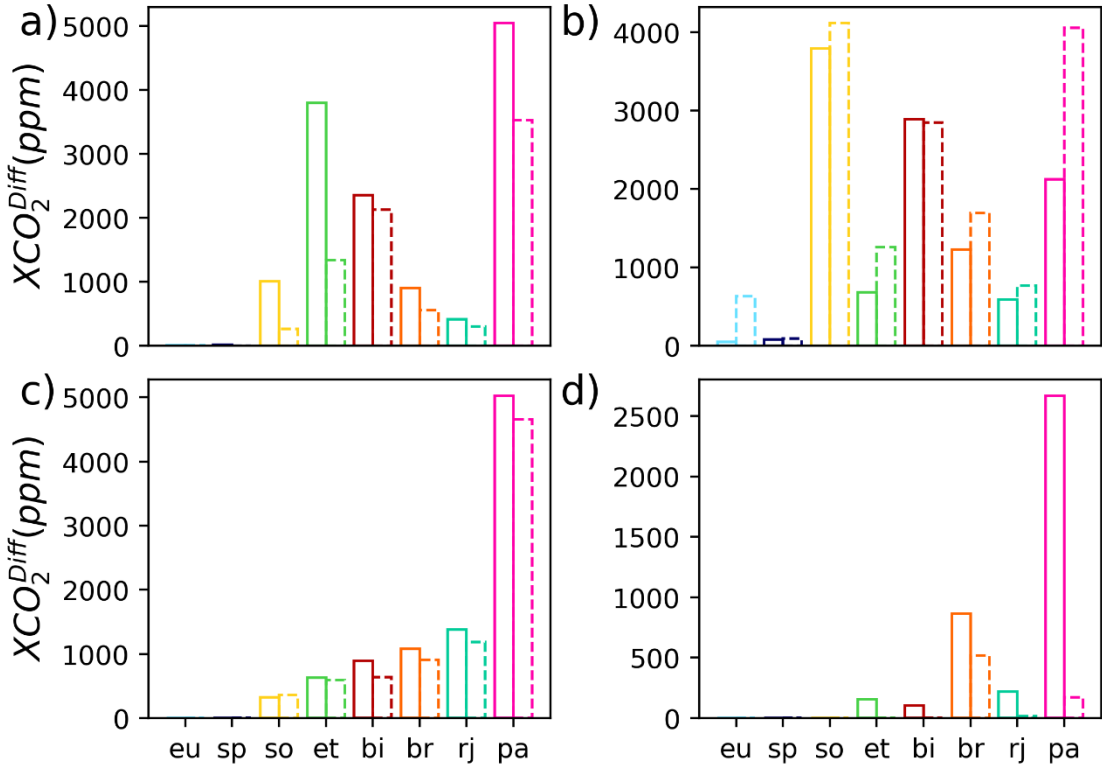
Figure 7 Same as Figure 5 but shows the number of soundings that pass each filter at each site for the different seasons. a) Spring (March, April, May). b) Summer (June, July, August). c) Fall (September, October, and November). d) Winter (December, January, February). Solid bars indicate the NN filter, and dashed bars indicate the original B10 qc_filter.