remote sensing

MDPI

Supplementary Materials

- Predicted CVS split by samples after random shuffling is presented in Table S1
- Predicted CVS of cases with above-cloud aerosols by our model is presented in Figure S1.
- A cloud profile showing the tops and bottoms of predicted CVS along the OCO-2 track is presented in Figure S2.
- Predicted CVS of multi-layer cloud cases by our model is presented in Figure S3.

Supplementary Material S1: Predicted CVS Split by Samples

To compare with the longitude-based split, we finished an 8-fold cross-validation, with 12.5% of total samples reserved for test dataset in each experiment, and 6% of the training dataset (5.25% of total samples) used for validation dataset (close to 12.8% and 5.5% in the paper). Results of the samples-based CV are listed in the Table S1 below.

The prediction based on random sample split have a better performance in p_top accuracy. Although the improvement can be attributed to the consistence of training and test dataset, we guess there may be overfitting in training and samples leakage when splitting the samples.

Besides, although p_tops of longitude-split test dataset have larger bias, we think it's acceptable. Compared with physical based model, machine learning model rely on statistical result, so shifts in distribution may occur when the algorithm applied to predictions in other years, which may lead to biases of test dataset. So we think split by longitude may have an impact on verifying the accuracy of our algorithm, but it's also a try for our model to test on a slightly shifted dataset versus training dataset.

What's more, the main scope of the study is to develop an algorithm using only OCO-2 oxygen A-band hyperspectral measurements to simultaneously predict COD, p_top and CPT of single-layer liquid clouds. We successfully proved the feasibility of our algorithm, and the method of split won't influence the main scope of our study.

Table S1. Models in the sample-based cross-validation and their prediction accuracy.

Cross-Validation Setup (Fold Number)	RMSE			Bias		
	COD/-	p_top/hPa	CPT/hPa	COD/-	p_top/hPa	CPT/hPa
1	7.38	32.80	26.93	-0.08	-2.18	0.33
2	7.21	31.43	26.84	0.00	-0.88	0.65
3	7.07	30.47	26.60	-0.14	-0.20	0.64
4	7.15	31.12	26.80	-0.12	-0.63	0.88
5	7.18	31.69	26.72	-0.03	0.93	0.41
6	7.13	31.00	26.66	-0.16	-0.69	0.33
7	7.13	30.86	26.67	0.05	-1.47	0.81
8	7.20	31.70	26.74	-0.10	1.17	0.32
Average result	7.18	31.38	26.75	-0.07	-0.49	0.55
Split by longitude	7.31	35.06	26.66	0.32	-8.25	3.49

Supplementary Material S2: Predicted CVS of Cases with Above-Cloud Aerosols

Aerosol-over-cloud scenes could influence the prediction. And it will have a negative impact for p_top and CPT. For the whole dataset of 2016, 350244 cases with above-cloud aerosol are tested based on Model I, with RMSEs of 6.88, 44.79hPa and 27.16hPa for COD, p_top and CPT, and biases of 0.22, -21.00hPa and 9.27hPa. Above-cloud aerosols don't have obvious impact on COD, no matter for RMSE or bias. But above-cloud aerosols cause larger error for p_top retrieval. RMSE increases to 44.79hPa and bias increases to -21hPa.

Because above-cloud aerosols cause absorption in oxygen A-band, then increase retrieval error of p_top. Similarly, RMSE and bias of CPT also increases.

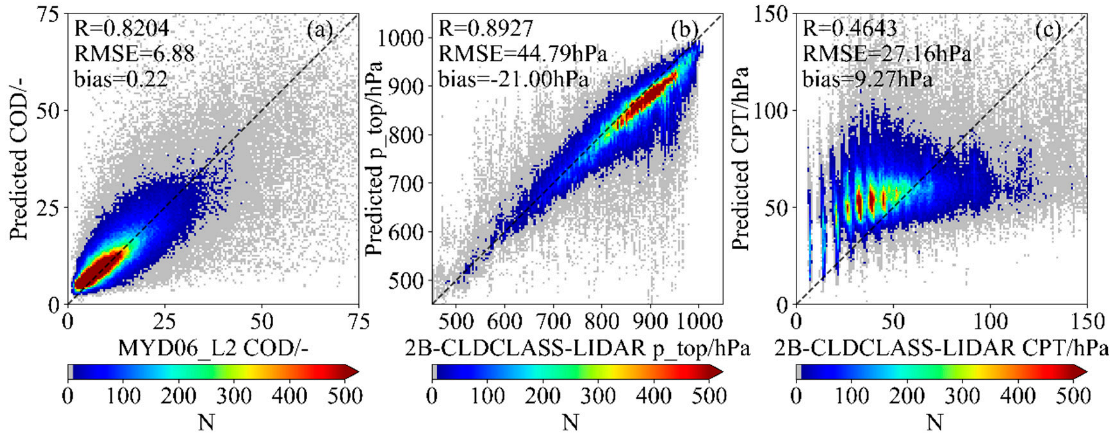

Figure S1. Prediction result based on the liquid clouds with above-cloud aerosols. (a) COD, bin size = 0.5 × 0.5, (b) p_top, bin size = 4 hPa × 4 hPa, (c) CPT, bin size = 1 hPa × 1 hPa. Bins with counts less than 10 are shown in gray.

Supplementary Material S3: A Cloud Profile Showing the Tops and Bottoms of Predicted CVS along the OCO-2 Track

A view along the OCO-2 track (orbit number: 12586) between 52°S and 2°S in the Figure S2 below (Model I). Predicted p_top and cloud base pressure are close to 2B-CLDCLASS-LIDAR, but there are some biases for p_top and cloud base pressure (like overestimations near 22°S).

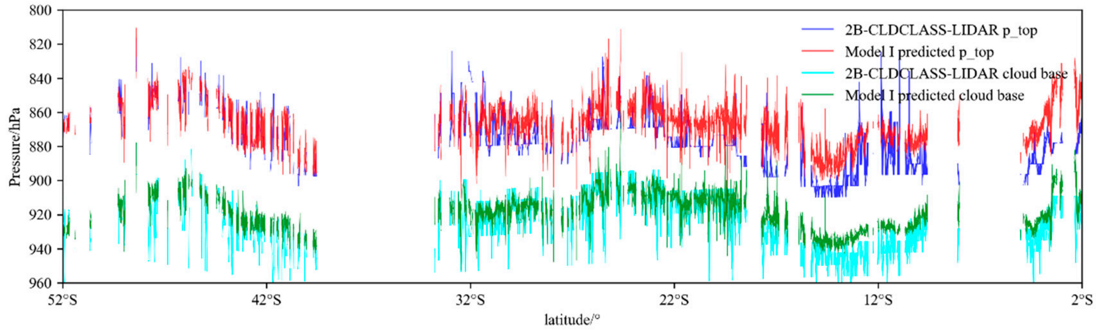

Figure S2. A view of predicted CVS of Model I along OCO-2 track (orbit number: 12586). Blue and cyan lines represent p_top and cloud base pressure (CBP) from 2B-CLDCLASS-LIDAR. Red and green lines represents predicted p_top and CBT.

Supplementary Material S4: Predicted CVS of Multi-Layer Cloud Cases

When Model I is trained, only single-layer liquid cloud cases are used, so Model I can only have three outputs (cloud optical depth, cloud top pressure and cloud pressure thickness). But actually multi-layer clouds have several cloud top pressure and several geometrical thickness. Because of the limitation of Model I (Model I can't retrieve multiple cloud top heights or multiple cloud geometrical thickness), we have to define multi-layer cloud before the test on multi-layer clouds. For a multi-layer cloud, we define the top pressure of the top cloud as cloud top pressure (p_top) reference and the base pressure of the bottom cloud as cloud base pressure reference. Then the pressure difference between them is calculated as the cloud pressure thickness reference.

Total of 5,163,487 multi-layer cloud cases without above-cloud aerosols are collocated. Our model can only retrieve liquid clouds, so we just test the cases with p_top>650hPa to ensure most cases are probably liquid clouds, with 2,432,787 cases tested.

In Figure S3, all three variables have much worse performance. COD still have a RMSE of 14.70, while p_top and CPT have RMSEs of 108.26hPa and 60.55hPa and biases of -42.97hPa and -14.95hPa. Our algorithm needs further improvement for multi-layer clouds prediction.

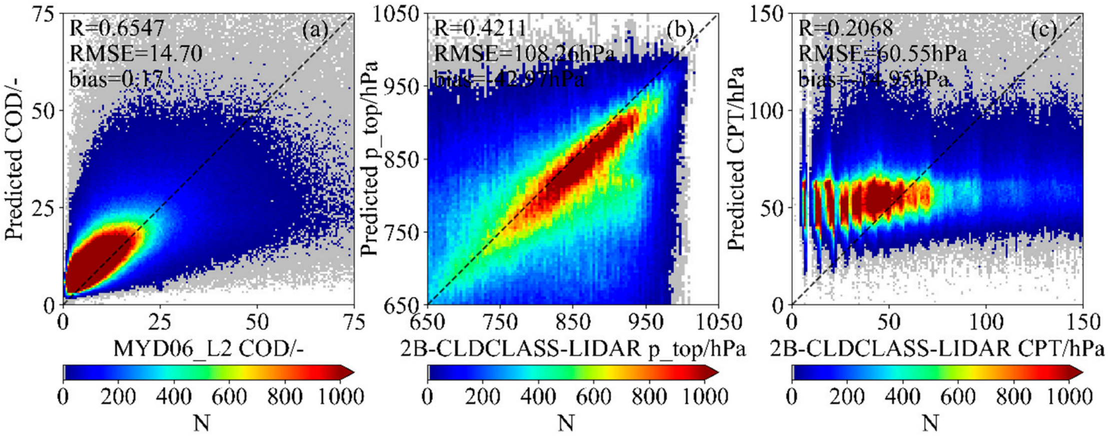

Figure S3. Prediction result based on the multi-layer clouds with p_top > 650 hPa. (a) COD, bin size = 0.5 × 0.5, (b) p_top, bin size = 4 hPa × 4 hPa, (c) CPT, bin size = 1 hPa × 1 hPa. Bins with counts less than 10 are shown in gray.