Black: referee's comments red: authors' answers
First of all, we want to thank the two referees for the detailed analysis of our paper.
For the details, please look into the paper with keeping track of changes.

Referee #1

The manual is clearly written and well organized. It characterizes in detail and explains NDACC vs. TCCON retrieval performance discrepancies for N_2O (dry-air column averaged), along with providing insight into CTM model performance in the SH. Although the conclusions presented in the abstract and conclusions sections are supported by the data shown, there are a number of places where the manuscript lacks specificity. I recommend publication after the issues below are addressed.

Equation 1: SBF stands for? Alpha is a scaling factor, is beta too? Is the functional form of SBF necessary here?

SBF stands for a Source Brightness Fluctuation correction. Beta is a scaling factor too. This information is added now. We prefer to leave the functional form of SBF here.

Pg. 5, regularization: "OEM" vs "Tikhonov" is too general, especially since the OEM choice leads to 1 extra DOF in Table 3. Is the OEM prior covariance based on WACCM runs? Is Tikhonov diagonal or with a correlation length?

More information is added in the revised text.

The OEM prior covariance is based on WACCM runs, and each site made some smoothing to remove some non-physical relationship among layers.

Sa^{-1} is created by the Tikhonov L1 method $Sa^{-1} = \alpha L^{T}1\ T\ L1 \in R(n,n)$ (Tikhonov, 1963). The matrix T considers the thickness of each layer. The regularization strength α is the key parameter to control the strength of Sa^{-1}. For the test at Reunion, we choose the $\alpha = 1000$, which is corresponding to a DOFs to 3.0 (similar to what we get from the OEM retrieval).

P7L4: 0.06% is called "relatively small" here for differences in retrievals due to apriori profile, but 0.09% was called "negligible" on P5L24 in regard to differences caused by regularization schemes.

Thanks for pointing out these two adjs. Compare to the retrieval uncertainty, we think both are negligible.

P5L22: you're really testing for 4 things in Table 4: spectroscopy, regularization, window and apriori profile. Stating this early is less confusing.

Ok, this sentence is added now.

P7L9: "apart from spectroscopy causing a bias between different retrieval windows" is imprecise and also confusing w.r.t. to the previous statement that spectroscopy is the same (in a given microwindow). I suggest "apart from the different sensitivity of the forward model to the underlying true state in different microwindows, e.g., on account of spectroscopic differences,"

This sentence is revised now.

P8L5, regarding Fig3: "bias . . . increases with time" -> I cannot see this effect in Fig3

This sentence is removed now and we agree that it is difficult to see the difference increases with time (even I plot the linear trend for the difference, it is still difficult to see). In fact, the slight increasing in the difference between NDACC and TCCON is existed, which is related to the difference in XN2O trends from NDACC and TCCON.

Fig 4: The lowest TCCON N2O days occur for shades of orange corresponding roughly to April/May, when PV is low (maybe this is the later 2014 anomaly). What do we learn if we color the scatter plot by PV instead of month of year, or by SZA?

Thanks for your suggestion. You can find the scatter plot colored by PV here. The scatter colored with PV is more clear, and we added it in the revised version. The Figure 7 is also updated.

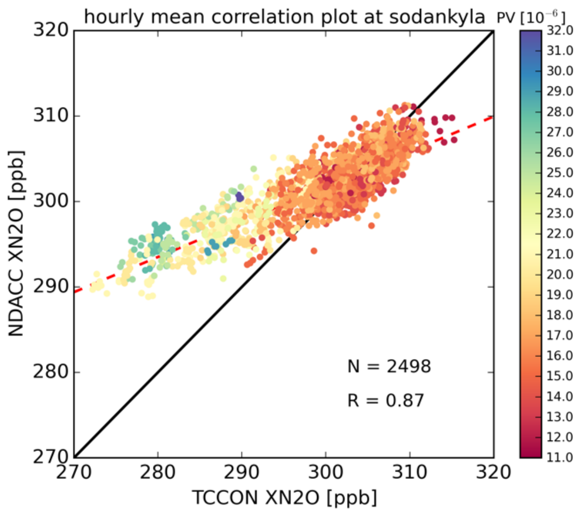

Figure 1. The scatter plot between the TCCON and NDACC XN2O data plot colored by PV value at Sodankyla.

We do not think coloring the scatter by SZA is a good idea, because the SZA is changing every day. For example, the SZA varies from 50 to 85 in spring while it varies from 30 to 85 in Summer, and thus it is difficult to distinguish the effect from the Polar Vortex.

Fig 5: Give n=3 and 40 for inside/outside measurements in caption. State the nature of the error bars. This plot is hard to read, consider splitting into two panels.

"n=3 and 40 for inside/outside measurements", " the error bars are the standard deviations of each profiles" are added in the caption. We do not separate Fig 5 into 2 panels, but we change the marker colors and error bar colors to make it more clear in Fig5.

Section 5: were the NyAlesund and Sodankyla TCCON data a priori-corrected in the model comparisons? If not, how would including this change Fig. 8/9/10?

The a priori-correction has not been applied for TCCON data at NyAlesund and Sodankyla when comparing to the GEOS-Chem model, mainly due to the following 3 reasons: 1) there is no co-located ACE-FTS measurement at Ny-Alesund, so that we can not apply the a priori correction at Ny-Alesund; 2) the purpose of the study is to understand the performances of the standard TCCON and NDACC data; 3) the Figure 7 already shows that the the XN2O trends and XN2O seasonality from TCCON (a priori corrected) and NDACC are very close to each other. Therefore, we prefer to keep the unchanged TCCON data in the text.

For testing, here we show the XN2O trends and seasonal cycles from TCCON (a priori corrction) and NDACC at Sodankyla in Table 1 and Figure 2.

Table 1. XN2O trend at Sodankyla in 2012-2017.

	TCCON	TCCON (a priori cor)	NDACC
XN2O trend	0.685±0.246 ppb/yr	0.812±0.282 ppb/yr	0.829±0.147 ppb/yr

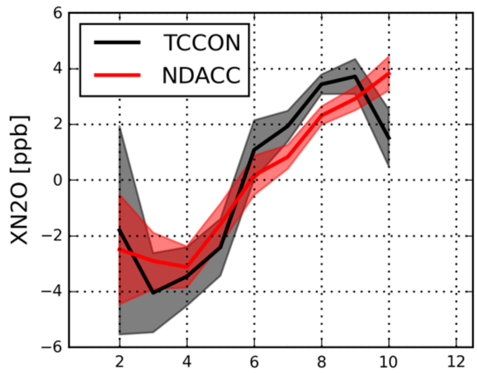

Figure 2. The seasonal cycles of XN2O from TCCON (a priori correction) and NDACC measurements at Sodankyla.

P14L3: "except at NyAlesund." -> and with very high variability at Sodankyla, which should be noted here as well.

OK, it is added now

P14L4: In this paragraph, the TCCON-derived trend is stated as "0.8-0.9 ppb/year" and described as "slightly smaller" than NDACC and flask sample trends of 0.9 – 1.0 ppb/year. I put a horizontal line to Fig 8 and find the highest TCCON trend to be ~0.85 (Wollongong), with all others being lower. Please calculate a single value of the TCCON, NDACC, flask, GEOS-priori and GEOS-posterior trends and be precise in comparing one to another.

The precise numbers are added in the revised version.

L9: updates to the TCCON profile with a 0.3%/yr growth rate: is that updated only at the surface or through a scaling of the whole profile? Give the precise change in the apriori structure and a precise trend value change.

It is the whole profile, which is added in the text now.

P14L18-19: "there are no full extent of the minimum . . . at NyAlesund in 2007, 2009, 2011" -> it's hard to see in that figure whether this is due to a lack of data or could it be because polar vortex intrusions were fewer in those years?

We agree with that it is difficult to distinguish, but both possibilities have been mentioned in the text.

P14L20: "comparatively more FTIR data" -> be specific: there is nearly 10X more data at Sodankyla. Also, since this data occurs in the shortest time series, this supports your previous argument about assuming constant growth rates and not worrying about different time series lengths at different TCCON/NDACC stations.

OK, they are added now.

Fig 9: It's hard to argue that there's a maximum in FTIR seasonal variations from Aug- Oct and a minimum from Feb-Apr at NyAlesund since there are no measurements in Feb, Sep, and Oct. This statement only holds true at Sodankyla and only for TCCON data (there's no clear maximum in autumn NDACC data). In this paragraph, the use of "slightly larger" (L6), "much larger" (L7), and "good agreement" (L10) is qualitative and debatable, especially in contrast with the high precision and high accuracy of the NDACC and TCCON data sets that was painstakingly laid out in Section 2. "good agreement" here appears to mean "better than factor of 2", though it is hard to see for Bremen and Izana. Please quantify. Also, I can't see an opposite pattern of seasonal variations at Wollongong as compared to the model; to me it appears rather flat, on average.

The description has been adapted in the text.

P16L4: "below 8 km" -> "from 0 to 8 km" will distinguish it from surface flask measure- ments better

OK.

P16L7: while Wollongong and Lauder may have comparable tropopause heights (check and quantify), they are separated by 10 degrees of latitude and exist in different climates, which should not be dismissed too quickly. For example, Fig 10 third row (8-17 km) clearly shows that if stratospheric processes are responsible for the discrepancy with GEOS-Chem, this is a stronger effect at the southern mid-latitude Lauder rather than the sub-tropical Wollongong.

We have double checked that the TP heights at Wollongong and Lauder are comparable (see Figure 3). The mean TP at Wollongong is about 0.3 km lower than that at Lauder.

Thanks for the suggestion, we agree that the difference in climate system might lead into the bias for Wollongong and Lauder. We added this information in the text now.

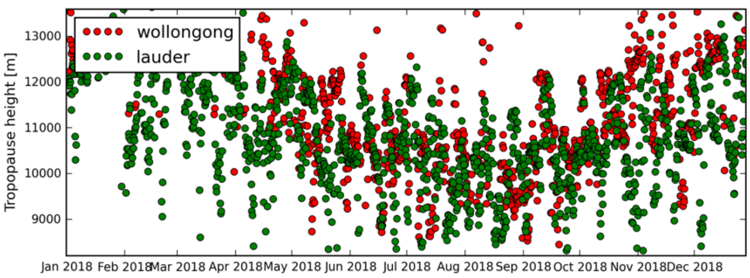

Figure 3. Tropopause height from ECMWF data in 2018 above Wollongong and Lauder.

P18L7: "slightly underestimated" -> please quantify

The precise values are added.

Minor comments:
P2L12: million -> billion
P3L11: "study is [that] to"

P3L29: “of [the] O2 in [the] dry air”

P5L15: “in these two spectroscopy” -> “in these two spectroscopic databases”

P6L4: “difference. Consequently,” -> “difference, consequently,”

P7L18: “are from” -> “range from”

P7L19: “with [the] standard deviations”

P8L10: “autumn” -> autumn

P9L5, L12, L18, L23: “polar vortex” -> “the polar vortex” (also in Fig. 6 and 7 caption and P16L28)

P9L8: “the isolation from mid-latitude refreshing” -> “dynamic confinement”?

P9L13: “[the] mole fractions”
P9L21: “criteria” -> “criterion”

P10L3: “[rapidly] decreases more rapidly”
P10L9: “[and] explaining why”
P14L16: “might be explained by [the lack of measurements;] the fact”
P17, Fig. 10 caption: “second to fourth panels” -> “second to fourth row panels”

Corrected