Overview:

Flood, et. al. has submitted a manuscript comparing ground-based mid-infrared FTIR measurements of tropospheric patrial column O3, CH4 and CO at five arctic sites to 11 model simulations. Daily and monthly averages are compared. Comparisons are conducted for the years 2008, 2009, 2014 and 2015 with the aim to test and comment on model validity, i.e., can the models reproduce the measurements, and if not, why?

The authors build upon prior University of Toronto research into Arctic atmospheric composition using FTIR data. The novelty of this study is that it is the first-time tropospheric patrial column O3, CH4 and CO measurements at the five arctic sites have been compared to this suite of 11 models. As mentioned in the manuscript, in situ and satellite data have been used in past studies to evaluate the performance of these 11 models, but not the FTIR datasets. The FTIR datasets provide an integrated partial column abundance that is quite different in footprint (spatial and temporal) and altitude sensitivity to the datasets used in previous studies hence bringing a new product to assist in model evaluation. This manuscript illustrates the benefits of using such partial column data in model evaluation and should be viewed as another standard dataset (along with in situ and satellite remote sensing) in future model comparison activities.

The manuscript is logically structured and well referenced. The writing is clear and ,in most instances, unambiguous. The analysis is robust and easily understood. The content is well within the scope of this journal. Information on data availability is given. The single conflict of interest is minor, stated up front and will be easily dealt with by the journal editors.

I recommend publication of the manuscript after some changes in the manuscript to mainly improve the clarity of content and the context of the investigation.

General comments:

G1/ The Introduction needs more detail to set the context of this research.

In the introduction, the AMAP SLCF assessment report (2021) was used as the basis for setting the context of this research into Arctic SCLFs and the importance of model validation. It is only in subsequent sections that the priori model validation work within Whaley, et al (2022 & 2023), Emmons et al. (2015) and POLARCAT/POLMIP were mentioned. Such past studies should be mentioned in the introduction to assist the reader in knowing where this current study fits in and what this study is to achieve that the past studies did not. It is only at line 89 where a single sentence states the aim of the study: “This study builds upon the model-measurement comparisons presented in the 2021 AMAP SLCF Assessment Report using an additional Arctic dataset that was not included in the original report.”. I view this current research as a natural extension of the work by Whaley, et al., 2022, but using a new dataset (FTIR site measurements) with a different temporal and spatial footprint to that of the in situ and satellite measurements.

The paragraph starting line 75 which introduces the FTIR measurement dataset should also be expanded to give examples of how such measurements from these 5 arctic sites have been used in past Arctic model validation studies. As it currently reads, it is unclear if this is the first time ever such measurements have been compared to models.

I think these changes will be easy to instigate and hopefully improvement context of this current research.

G2/ There is no mention of why column integrated measurements are used to validate/compared to the model simulations. This is one of the main novelties of this study.

At line 518 there is the statement: “NDACC FTIR spectrometers were selected for this project because of the wide range of species measured, high spectral resolution, multiple high-latitude sites, and publicly available data “, which seems the main justification of using the FTIR data (along with a brief contextual reference at line 72: “All of these factors lead to a scarcity of monitoring stations and a limited representation of atmospheric vertical information”).

I think these are secondary reasons, the main reason being a (partial) column integrated data product that has a spatial and temporal footprint which is more presentative of the tropospheric free atmosphere than in situ and satellite measurements.

I recommend adding a statement (in the Introduction) focusing on the benefits that validating models using partial column data (that FTIR can provide). The advantages and disadvantages of using column integrated data needs to be explained and how such data allows comparison to models in way in situ and satellite remotely sensed data cannot. It fills a gap.

G3/ It would be good to explicit state why CH_4, CO and O_3 were the selected species as both the models and measurements have other SLCFs products for which comparisons could be performed.

G4/ There is no mention why the four selected years (2008, 2009, 2014 and 2015) were chosen for the comparison activity. Reason/s why these years were selected need to be stated. Two other points should also be addressed: given this manuscript was submitted in 2023, why was the most recent year 2015? and why long-term trend comparison analysis , i.e., a continuous time series period, was not performed. I suspect model simulation temporal constraints, but this should be stated.

G5/ The partial column range used in comparisons is ground level to 7km. A prior study used 0-9km (Wespes, et al. 2012). Please state why the 0-7km range was selected.

G6/ There is no mention of the tropopause heights at the measurement sites. Even if the selected partial column upper boundary (7km) is less than the tropopause height, the averaging kernels might indicate 0-7km partial column measurement sensitivity to above the tropopause. How would it effect model measurement comparisons? Are stratospheric intrusions of major concern?

G7/ Reorganization of site-specific figures.

For CH_4: seasonal daily and monthly time series plots along with daily model measurement scatter plots are given for a single site, Eureka, i.e., figures 3,4 and 5 and the other four similar FTIR site data plots are in the appendix. Then all site data/metrics for CH_4 are displayed in figures 6 and 7. This is repeated for CO and O_3.

I found I was continually being referred to figures in the appendix, especially when it came to interpretation of the model measurement results at the end of each species section. I would like the authors to consider rearranging figures. I suggest that all the daily/individual measurement plots, e.g., figures 3,8, 13 be moved to the appendix. The monthly plot , e.g., figure 4 include all stations, in a 2x3 panel plot.

Also, figure 5 to include all stations. For CH_4 this would be a 3x5 panel plot.

For figures 10 and 15 I don’t think all stations scatter plots can be plotted is a reasonable way in a single figure , thus still relegated to the appendix. If the author can think of another way to concisely display the mentioned data in the main body of the manuscript it could be worth investigating.

G8/ Analysis interpretation of CH4.

Compared to CO and O3, the discussion and interpretation of the CH4 partial column measurement model comparison results are very short. Example: line 266 "satellite 265 instrument and finds that the models are biased low in the vicinity of the tropopause (300hPa) (Whaley et al., 2022)." What height is 300hPa? How much biased low? Is this expected? acceptable?

Please expand and include a greater discussion of the results in comparison to findings from Whaley, et. al., 2022, esp. in context of surface CH4 in situ measurements.

Specific comments:

S1/ Following on from G1, the first two sentences in the abstract could be improved. They currently do not add any specific information about this study. The abstract should also mention the novelty of this study, i.e., what has been done here that hasn't been done before.

S2/ line 51. "…causing most of the pollutants to remain predominantly localised" but throughout the manuscript there is multiple references indicating long range transport [of pollutants] (as at line 370) are a possible cause of measurement model differences. Can this disparity be rectified.

S3/ line 125. Degrees of freedom and the partial column averaging kernel (PC AVK): Could figure 1 altitude range be expanded to ~ 20km to see 'what happens above 8km'. If the PC AVK above 7km is ~ 1.0 this means retrieval information above 7 km is incorporated into the 0-7km PC. If so, please comment upon, and implications of.

S4/ line 124 and Table 2. Please expand commentary and implications for DOFs < 1.0. For CO and O3 the PC DOFs are ~1.0, but for CH4 the DOFs are < 1.0, and from figure 1, there is less sensitivity to near surface CH4. What are the implications of this for model comparisons?

S5/ Section 2.2. Relative to other manuscripts the section describing model simulations is brief, but I think it is justified as detailed model descriptions (and forcings) are given in Whaley, et. al., (2022). I see no need to repeat information that is already readily available.

Could the authors make sure that any model output that is used in this current study that differs from model output used in the study by Whaley, et al., (2022) be stated and the reasons for the change (e.g., an updated model or forcings) also be stated. This may seem a logical statement, but if the authors are going to heavily defer to Whaley, et al., (2022) to provide details then it is very important there are no changes or changes are identified.

S6/ The first two sentences in section 3 are not needed, as it is covered in the section 'Data Availability', or if the authors want to retain it in the manuscript, then relocate to section 2.

S7/ Figure 2. The flow chart alludes to that the 'nearest' model grid point (to a measurement site) is used. This should be mentioned in the text. To clarify, is there any spatial weighting of localised grid points? I.e., weighting/kriging of the closet model points/cells to the FITR location? Have tests been done concerning a geolocation weighted average model value? I gather any differences will be minimal but would be good to confirm, even if for a single site.

S8/ I think another paragraph is needed at the end section 3 concerning the type of analysis that is going to be performed using eqn. 1 and 2 as the quantification metrics. Are you going to investigate, diurnal, daily, monthly, or seasonal differences? Long term trends? Basically, what are you going to look at.

S8/ Best line fits: linear regression. Do the best line fits in all the analysis also take into account the uncertainty in the abscissa (measurements) as well as the ordinate (model)? If so, please state so, if not, then maybe prudent to perform a few tests to assess the effect on the linear fit. Since measurement and model uncertainties are of comparable magnitude, abscissa uncertainty could have a large effect on the calculated linear fit.

S9/ line 244:

"For all models, the R2 values for Ny Ålesund and Harestua are significantly smaller, while the overall mean percent difference is comparable to the other locations. The discrepancy is likely attributed to the smaller number of measurement points, causing outliers to have more weight in the linear regression, which is reflected in the elevated NRMSE for Ny Ålesund across all models."

I do not think a lack of lack of measurement points is a cause. Both Figs A9 and A12 show there are plenty of data points. Fig A12 clearly shows there are outlier measurements at Harestua. I would attribute this to either measurement/retrieval error that was not filtered out thus should be removed from the comparison datasets, or anomalous atmospheric events which if at fine temporal or spatial scale the models would be able to reproduce, thus this measurement period should also be omitted as the model would not be able to replicate it. Given that the anomalous measurements are both too high and too low I suspect measurement error. I recommend omitting such outliers (across all data sets, unless it can be accounted for) using a defined filtering method and perform analysis again.

This approach will not account for the low R^2 at Ny Ålesund and I cannot easily see why the R^2 is lower than at other sites.

S10/ line338: "Similar trends have been found in other Arctic model-measurement comparison studies." Please reference this statement, also do you mean trends or findings? As temporal trends are not investigated in this study. I think would also be helpful to quantitatively state the amount of underprediction in prior studies and then relative to this study (referring to table D1 would be a good idea when comparing the results from this study to prior studies).

S11/ line 355: "Further, the tracer investigation shows that OH differences account for more variability between the models than the transport mechanisms within the individual models."

Could this statement please be referenced.

S12/ line 366: "The results of the model-FTIR comparisons presented here support this reasoning, as the model with a positive bias (GEM-MACH) has a different emissions input, with possibly more complete emissions in the Arctic, as this was a high-resolution Arctic version."

This conjecture could be quite easily solved by looking at the model simulation parameters to see if this is true.

S13/ line 381: "In addition to atmospheric chemistry, its production is highly sensitive to meteorological conditions. Therefore, it is difficult for models to accurately simulate tropospheric O3." Ozone also can have a significant diurnal cycle due to photochemistry, complicating comparisons when measurements and model differ in time. Please include this cause as well.

S14/ line: 452. " To supplement the aircraft and satellite campaigns undertaken for the POLARCAT study, daily mean O3 measurements from the FTIR instruments at Eureka and Thule were compared to MOZART-4 simulations in Wespes, et al. (2012)".

Due to the daily diurnal cycle of ozone, comparisons of daily FTIR averaged ozone measurements would be biased high to model output (that uses daytime and nighttime values as I gather nighttime FTIR measurements are not taken). Can you confirm daily average MOZART ozone was used or matched to FTIR measurement times.

S15/ line 518:

"NDACC FTIR spectrometers were selected for this project because of the wide range of species measured, high spectral resolution, multiple high-latitude sites, and publicly available data."

As stated in G3, a better reason for using FTIR datasets should be given. This relates back to a general comment of the overall benefits of using column integrated measurements.

S16/ Defining FTIR uncertainty. This term (or variations of) is found within the text (e.g., lines 248, 323, 514) but not clearly defined. Is it the uncertainty of individual measurements as in table 2 , or the 1-sigma standard deviation of the daily/monthly measurement means?

S17/ The table 4 caption states:

"Summary of mean percent difference for each model and location by species. MMM is the multi-model mean. The colour scale indicates the mean percent difference relative to the FTIR measurements, from blue (-50%) to red (+50%). A square marker indicates that the mean difference is within the FTIR uncertainty. A triangle marker indicates that the mean difference is within the FTIR uncertainty combined with the standard deviation of the monthly mean percent difference."

It is difficult to understand what is being compared (and significance of the metric) when FTIR uncertainty is not clearly defined. Is FTIR uncertainty the monthly measurement 1-sigma S.D. or the uncertainties of a single measurement as given in table 2?

There is no explanation of why a double metric is used, could this be explained in the text. What does it mean if "the mean difference is within the FTIR uncertainty" but not "within the FTIR uncertainty combined with the standard deviation of the monthly mean percent difference".

S18/ Table D1: Is an important table. I recommend putting this in the main body of the manuscript and referred to in each species section.

Technical comments:

T1/ line 81. Arctic is not defined, are you implying >60N? Maybe define what 'Arctic' is.

T2/ Table 1 and Table 2 colour key columns are not needed.

T3/ Paragraph starting line 114 concerning technical details about the FTIR data and retrieval strategies. I think there is a need to mention the vertical grid spacing of the retrieval, i.e., how many layers, esp. in the troposphere, and from 0-7km.

T4/ Figure 1. The term 'mean column'. Do you mean total or partial (0-7km) column? Please make this clear in the label. If it is total column, then I recommend replotting as 0-7km partial column.

T5/ Figure 1. The abscissa axis (Partial? column AVK) needs units. [unitless] or [relative] would suffice if not [ppb/ppb].

T6/ line 166. (+/- 1.5 hours): I think it needs to be explicitly stated why this time frame was chosen (from previous model comparison studies?), just to make it clear why , say , +/-24h cannot be used. A tight time constraint is required for ozone due to diurnal photochemistry.

T7/ line 173. Partial column averaging kernel I gather? Maybe add ‘partial column’.

T8/ line 176. “ratio between the trace gas VMR and layer airmass (molec cm^-2)”. Best to add the term ‘layer airmass’ for clarity.

T9/ line 180. To clarify, is the station altitude is also used as the lower model partial column layer boundary in analysis? If so, then I think this needs to be stated.

T10/ line 206. Please replaced ‘important’ with a more specific descriptor. Important is too subjective (important in what context?).

T11/ line 213. ‘concentration’ should be replaced with ‘partial column’. The models are forced with concentrations (vmr), but the quantities under investigation are partial columns (molec. cm^-2).

T12/ Line 217. ‘Little variance’. Sorry, I found this unclear, do you mean between the models or intra-model (within a month or day)?

T13/ line 219. ‘uniformity’, of what?

T14/ Figure 3, 8 and 13: The measurement symbols are extremely hard to differentiate between. Can you make them easier to differentiate?

T15/ All figures: In all the figures, when model data is plotted, I gather it is modelled smoothed partial columns? If so, please add ‘smoothed’ to all ‘model data’ just to make it clear.

T16/ line 380. “It is a secondary pollutant” replace with “In the troposphere, ozone is a secondary…”

T17/ line 395. “However, the FTIR O3 seasonal cycle does not have a springtime minimum from surface ODEs, as one might expect from surface measurements”. Sorry, this does make sense. As it reads FTIR partial measurements are surface measurements? Can you please rewrite to make it clearer.

T18/ line 422. remove the word ‘difficult’.

T19/ line 425. remove ‘and as such recommended for future work’. I can understand what is trying to be conveyed, but nearly instance of a model measurement disagreement warrants future work.

T20/ line 497. remove the word ‘historical’.