General comments:

The manuscript by Flood et al. evaluates the accuracy of model estimates of partial atmospheric columns of CH4, CO, and O3 in Arctic regions by comparing against FTIR ground-based measurements from five northern high latitude sites in the Network for Detection of Atmospheric Composition Change (NDACC). The study considers 11 models and expands on previous analysis published in the Arctic Monitoring and Assessment Programme (AMAP) 2021 assessment report. That report included comparisons of 18 models to surface in situ, aircraft, and satellite-based measurements, but did not compare to ground-based FTIR measurements of partial column concentrations. Time-series of FTIR measurements and model estimates, correlations between FTIR measurements and model estimates, comparisons of seasonal cycles, and summarizing statistics of comparisons of model estimates and FTIR measurements are presented for each NDACC site. The site Eureka is the primary focus of most of the discussion in the paper, while comparisons at the remaining NDACC sites are mostly presented in the Appendices. There is also a balanced discussion of how the results of the comparison may or may not support previous analyses of this type.

Overall, the value of this analysis is clear. Comparing model estimates of greenhouse gas concentrations to FTIR retrievals is useful and provides a wealth of additional information that cannot be provided by comparing to surface-based measurements, and with much greater temporal frequency than aircraft-based measurements. Furthermore, this analysis is thorough and the interpretation of the results is reinforced by comparisons to previous literature. That being said, there are critical ways that the paper could and should be improved before publication. First, while references for each model considered are provided in the paper, the differences in model frameworks, assumptions, boundary conditions, and set up are not discussed much in the paper. The authors seem to expect the reader to conduct a high degree of background research or already have a very thorough understanding of all of these models. This is particularly relevant in the failure to state the spatial resolutions of the models being used and the failure to sight which models do or do not simulate the stratosphere (particularly in the context of the O3 discussion). There is certainly some discussion of factors that may be driving specific model biases, but these could be discussed more in the context of specific models. In addition, I think that some generalizations or conclusions are not as consistent with the plotted results as the authors suggest; however, many of the plots are difficult to read and I think reformatting some of the figures would really help.

Specific comments and suggestions:

In general, the paper would greatly benefit from a reorganization of how data is presented in the figures. Specifically, Figures 3, 4, 8, 9, 13, and 14 would be better presented as separate panelled subplots for each model, similar in format to Figures 5, 10, and 15. As they are, these figures are very difficult to read with many overlapping lines that make it hard to differentiate the behaviour of individual models.

Line 84, why are you only considering 11 of the 18 models used in the AMAP report?

In section 2, line 109, you say that SFIT4 is used to retrieve VMR profiles from NDACC FTIR measurements at all sites except Kiruna, which uses PROFFIT. There needs to be some discussion of how these two retrieval algorithms differ and how this may impact the resulting retrievals. Alternatively, you can cite a paper that compares the two. It seems unlikely that using different retrieval algorithms would not have some effect. Furthermore, in section 2, you should state what the typical temporal frequency of NDACC FTIR observations are. This becomes relevant in section 3 because it is useful to know approximately how many FTIR observations are included in the 3-hour averages that are compared to model estimates. Similarly, some information about the spatial resolutions of the models being considered and the distances between the NDACC sites and the referenced location of the model estimates is important. Do all the model data have the same spatial resolutions and do distances between site locations and model estimates vary among sites or among models? Could any of this variability explain differences seen in the model comparisons?

In Section 3, specifically the flow chart in Fig. 2, since the FTIR measurements are collected with greater temporal frequency than the model estimates, it would seem that the temporal matching would involve finding the FTIR measurements closest in time to model estimates, rather than the other way around. Also, please clarify whether averaging kernels and a priori values are retrieved for each FTIR observation or based on a general reference for the instrument. If they do vary with each observation, how is this handled when applying the corrections.

It seems like the second paragraph of section 4.1 belongs in methods. By extension, it would make sense to talk about how many models you are including in the comparison analysis for each gas earlier in the paper (in section 2 or 3), though the gases covered by each model are shown in Table 3 the number of models that estimate each gas should be summarized in the text as well.

Can you speculate on why GEOS-Chem has lower % differences (better accuracy), but also lower correlation coefficients (poor precision or more scatter) in the CH4 comparisons?

Line 262, you say that the FTIRs show good sensitivity to surface CH4, but Fig.1 shows that the instruments are more sensitive to higher altitudes in the partial column than they are to the surface. I also wonder if variations in the tropopause height or a poor representation of this in the models or in the FTIR retrievals could affect the accuracy of your CH4 partial columns in the comparisons.

In section 4.2, could the increased discrepancies between FTIR retrievals and model estimates of CO in spring also be at least partly explained by errors or biases in the FTIR observations due to low solar zenith angle or cloud cover?

Line 294, I think the seasonal shifts in bias for EMEP-MSC-W and WRF-Chem are more remarkable than for MATCH (at least at Eureka, which seems to be the site that the results discussion is primarily focused on), but these are not mentioned.

On page 17, when discussing Fig. 11, why not comment on the fact that WRF-Chem has correlation coefficients near zero and very high NRMSE relative to the other models?

Line 353-355, is this describing a global effect in which European emissions have a greater influence on surface CO everywhere, or were the studies conducted in Europe and the surface CO is more affected by local emissions?

In section 4.3, please clarify which months are included in "springtime". It seems that most models agree better relative to other models as well as FTIR in February and March than all other months except September. If these months are part of springtime that does not support your claim that springtime O3 concentrations are poorly characterized in the models. Furthermore, if April is part of springtime, WRF-Chem should be mentioned on Lines 400-401, along with UKESM1, GEM-MACH, and GEOS-Chem.

Line 401-402, please elaborate on the reasoning behind this conclusion ("may be attributed to a low bias in the models' lateral boundary condition, inaccuracies im model water vapour and/or a lack of O3 transported from mid-latitudes.").

Line 403-404, why is MRI-ESM2 not mentioned here? That model seems to track very well with the FTIR measurements in Figures 13 and 14.

Line 409-410, this claim does not seem to be as consistently relevant for O3 as it is with CO. There are a number of models, including EMEP-MSC-W, DEHM, and CESM, that appear to exhibit a high degree of scatter for lower O3 partial column concentrations (at least for Eureka).

Line 416, if the Wespes et al. study mentions stratospheric influence as a major driver in tropospheric concentrations of O3, why is there no further discussion of which models in the current study simulate the stratosphere and how this may or may not influence errors in the partial column model estimates when compared to FTIR measurements?

Line 441-444, it should also be mentioned that WRF-Chem and GEM-MACH do not have the same temporal coverage as the other models.

Line 446-447, the statement, "although again are largely underpredicted" needs more context. Do you mean to say, "models largely underpredict FTIR measurements"?

Line 457-458, drawing this conclusion in relation to the Wespes et al. study seems a bit tenuous because they only compared to one model and it was not one of the models considered in the current study.

Minor editing suggestions:

Do Figures 5, 10, and 15 need to have legends to indicate which line is 1:1 and which is the linear fit? I think Figures 7, 12, and 17 also need legends indicating FTIR data and model data points

Line 178, change "FITR," to "FTIR retrieval," or "FTIR retrieved partial column,"

Line 282, change "provided below" to "provided in Fig. 8-10"

Line 299, missing period between "2021)" and "Further". Change "comparison of" to "correlations between"

Line 300, change "1:1 comparison" to "1:1 correlation"

Line 326, change "FTIR comparisons" to "FTIR measurements" or "FTIR retrievals"

Line 330, suggest mentioning that this is eight pairs out of 36.

Line 366, change "emission fluxes" to "anthropogenic emissions" or "anthropogenic fluxes"

Line 388 and Line 391, change "show" to "shows" and "reduce" to "reduces", respectively. Go through the paper and make sure verb tenses when referring to a single figure are singular, the verb should fit the subject outside the parentheses.