Flood et al., Evaluating modelled tropospheric columns of CH_4, CO and O_3 in the Arctic using ground-based FTIR measurements, https://doi.org/10.5194/egusphere-2023-1161

Response to Reviewers

Review 2

We thank the reviewer for their comments, which have helped us improve the manuscript. Our author responses are given in a blue font, while the italicized text in the indented bullet points has been added in the manuscript.

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.

After considering the feedback from both reviewers, the figures in the text have been modified and moved to help with flow, but retain clarity. Figures 3, 8, and 13 have been moved to the appendix, and since we don't discuss the difference between each of the four years, we have removed the different symbols for them. We feel that Figures 4, 9, and 14 show a better view of the models in relation to each other when they are presented on the same axis and prefer to keep them in one. These have been moved to the appendix and a subpanel plot with all of the monthly mean results at all locations has been added in the text.

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

Added the following text to Section 2.2:

- *While more models participated in the AMAP SLCF Assessment (18 total) and other species were simulated, these were not included in the current study because either the models did not have 3-hourly outputs or the FTIR retrievals had insufficient tropospheric sensitivity (e.g., NO_2).*

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?

Added the following text to Section 2.1:

- *All sites included in this paper use SFIT4, except Kiruna, which uses a comparable retrieval code called PROFFIT, which has been shown to agree well with SFIT (Hase et. al, 2004).*

Table 1 lists the operational season of the FTIR instruments at each site and Table 2 indicates how many measurements were taken in the years being investigated. The temporal frequency of measurements would depend on operations of each location, including instrument downtime and weather. There are very few instances (<10 across sites) of multiple FTIR measurements being averaged from falling closest to the same model interval.

The NDACC locations are listed within Table 1.

The model spatial resolution has been added to Table 3.

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.

Although multiple FTIR measurements can occur within a 3-hour period, there were very few instances where this occurred. When it did (as stated in the text) the partial columns were averaged. Given that there are far more modelled results, the FTIR measurements were matched with the model output that was closest in time to the measurement.

This statement in Section 3 indicates that each FTIR measurement has its own averaging kernel and that the smoothing applied to the model profile uses the averaging kernel of the relevant FTIR measurement:

> *"Then, the model VMR profile is smoothed using the respective FTIR measurement's averaging kernel and a priori profile."*

This statement in Section 2.1 indicates that each station has a single a priori profile, but that the pressure and temperature profiles correspond to the conditions near the time of each measurement:

> *"The a priori information for the modelled spectra is provided by 40-year-average profiles from the Whole Atmosphere Community Climate Model (WACCM) (Marsh et al., 2013), with spectroscopic absorption parameters from the HITRAN 2008 line-list (Rothman et al., 2009) and daily pressure and temperature profiles from the U.S. National Centers for Environmental Prediction (NCEP) (Kalnay et al., 1996)."*

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.

Added the following text to the manuscript in Section 2.2:

- *Note that not every model has provided all three gases; there are three which have CH_4, nine with CO, and 11 with O_3 (see Table 3).*

Moved text in Section 4.1 regarding CH_4 prescribed in models to Section 2.2, as suggested.

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?

Added the following text to Section 4.1:

- *GEOS-Chem does simulate a north-south gradient, which is reflected in the smaller overall model-measurement percent difference, compared to other models, in all locations (note Fig. 6 in Whaley et al., 2022). However, the R^2 of GEOS-Chem vs. FTIR is smaller than that for the other models at some locations (Eureka and Kiruna), which can be attributed to the increase in variability the gradient introduces – including some instances of overestimation.*

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.

Added /modified the text in Section 4.1:

- *The FTIR retrievals show good sensitivity to tropospheric CH_4 (sensitivity >0.5), however, as these column measurements average out CH_4 biases over the tropospheric column, they are not expected to exactly match the surface measurement comparisons. Furthermore, due to the sharp decrease in CH_4 above the tropopause (Whaley et al., 2022), a poor representation of the tropopause height may contribute to the low bias in the 0-7 km partial columns, as shown from O_3 data in Whaley et al. (2023).*

We have also included the partial column averaging kernels for 0-7 km and 7-20 km to show the difference between the altitude ranges in the partial columns.

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?

We do not believe these factors account for the discrepancies as the other Arctic modelling papers discussed found similar results for CO comparisons using in situ and satellite measurements (e.g., Whaley et al., 2022).

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.

Added the following text to Section 4.2:

- *WRF-Chem is biased low in the spring and summer, but agrees better with the observations from August onwards, in contrast to EMEP-MSC-W, which tends to diverge from the measurements in the mid- to late summer.*

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?

The following text has been appended to the statements which were already included in the text on this topic:

- *WRF-Chem shows better agreement with the FTIR measurements from Eureka, where the NRMSE is comparable to CESM, CMAM and GEOS-Chem. This is likely a result of the increased density of measurement points in August and September, when WRF-Chem exhibits a minimum bias compared to the FTIR data, and because the comparison only includes data points from 2014 and 2015. The large negative biases earlier in the year lead to low R^2 and high NRMSE at all sites. This appears to be linked to negative biases in modelled surface CO over mid-latitude source regions, and in the free troposphere compared to MOPITT data, as reported by Whaley et al. (2022).*

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?

The results discussed in relation to the POLMIP study are regarding the Arctic. This is noted in the preceding sentence: *"Using an idealized tracer, POLMIP examined anthropogenic and biomass burning influences in Arctic regions, demonstrating a seasonal dependence of transport efficiency"*.

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.

We have added (late February - May) to the text on first mention of "spring" to define it. We agree that the MMM has little to no bias in the springtime O_3 at Eureka, however, there is a large spread in springtime O_3 values across models. While we discussed each of the models' behavior in the springtime, we did not state that overall it is poorly characterized, just that it is quite variable.

Added WRF-Chem to the statement, as suggested.

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 in model water vapour and/or a lack of O3 transported from mid-latitudes.").

Added/modified text in Section 4.3:

- *The discrepancies may arise from inaccuracies in model water vapor leading to an increase in O_3 destruction and/or a lack of O_3 transported from mid-latitudes, which is a substantial source of tropospheric O_3 in the Arctic (Hirdman et al., 2010; Whaley et al., 2023). In the case of the regional GEM-MACH model, low biases in O_3 or precursor species at the lateral boundary conditions may also be contributing.*

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.

Added as suggested.

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).

Reworded in Section 4.3 to better describe the results shown in the figures:

- *The general underprediction towards the largest values could be related to the underestimation in precursor species (such as CO or NO_x), a lack of long-range transport, an underestimation of ozone production in air masses during long-range transport to the Arctic, or a combination thereof.*

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?

We have added a column to Table 3 to indicate the level of stratospheric chemistry for each model.

Further text was added in Section 4.3 to support this:

- *The model-FTIR comparisons reveal that the spatial resolution and inclusion of stratospheric chemistry in the models does not necessarily improve results (refer to Table 3 for horizontal resolution and stratospheric chemistry). For example, WRF-Chem, EMEP MSC-W, and GEM-MACH show a low R^2 and higher NRMSE (varying between sites and models), although contributing to this for WRF-Chem and GEM-MACH could be the limited number of analysis years (two and one, respectively). These air-quality focused models have detailed chemistry and were run at higher spatial resolutions, whereas for example CMAM, a climate-focused model, has a coarser resolution with simplified tropospheric chemistry and demonstrates larger R^2 and smaller mean percent differences (Fig. 13). However, when considering the stratosphere, CMAM, which* *includes comprehensive stratospheric chemistry, has comparable metrics in Fig. 13 to DEHM, which uses prescribed climatologies for the stratosphere. Similarly, Whaley et al. (2022) stated that the degree of stratospheric chemistry in the models did not reveal a consistent benefit or handicap when comparing the models with surface 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.

Added the following text to Section 4.3 to reiterate:

- *For example, WRF-Chem, EMEP MSC-W, and GEM-MACH show a low R^2 and higher NRMSE (varying between sites and models), although contributing to this for WRF-Chem and GEM-MACH could be the limited number of analysis years (two and one, respectively).*

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

Changed text to suggestion.

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.

Modified text in Section 4.3:

- *Results here are similar to those presented in Wespes et al. (2012), where across all the locations and models, 24 of the 55 model-measurement mean percent differences were within ±15% (see Table 4).*

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

The 1:1 line and linear fit are described in the figure caption. Previous iterations had more detailed legends, however these were reduced to allow for larger text of the current legend, and avoid additional clutter on the plots.

A legend has been added to the MMM plots.

Line 178, change "FITR," to "FTIR retrieval," or "FTIR retrieved partial column," Line 282, change "provided below" to "provided in Fig. 8-10"

Changed as per suggestion.

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

Changed as per suggestion.

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

Changed as per suggestion.

Line 326, change "FTIR comparisons" to "FTIR measurements" or "FTIR retrievals" Line 330, suggest mentioning that this is eight pairs out of 36.

Added as per suggestion.

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

Changed as per suggestion.

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.

Edits made.