We thank the reviewer for their comments. Responses to comments are shown below in red, while quotations from the revised manuscript are indicated in blue. Line numbers for the reviewer comments refer to the original manuscript while line numbers in the author responses refer to the revised manuscript unless otherwise indicated.

Reviewer 1

With the advance of satellite instruments and machine learning techniques, atmospheric chemistry research is developing fast. This paper does an attempt to build a machine learning method for total (tropospheric) column OH, based on satellite observations. To this end, they train a GBRT model on results of a chemistry transport model and apply this model to satellite observations from mainly MOPITT, OMI, AIRS, but also location (i.e. solar intensity). The basic idea is that OH in the remote atmosphere in the tropics is driven mainly by the abundance of O3, H2O, NOx, CO, and hydrocarbons. In that respect, this provides an original contribution and shows some promises for the future.

The main problem I have with the paper is that is insufficiently credits and discusses other developments in the field. Reading the paper, I was wondering if the authors are aware of these developments at all? In a fast-advancing field, reading, referencing, and discussing the work of others is of utmost importance. And the paper fails to do this. References to own work dominate. Below I outline how the paper should improve to become acceptable for publication.

General methodology

In studying OH in the remote atmosphere, we have to rely on knowledge on atmospheric chemistry. In this paper, the authors use results of atmospheric chemistry simulations to train a machine learning algorithm. No criticism here. In the discussion, however, they "come up" with the idea to reduce the uncertainties between the 3D model results and satellite observations (without any references). Long-standing efforts have been made to "merge" satellite information and models in a process called data assimilation. First, there is the idea of chemical data assimilation, performed in e.g. the EU Copernicus services (e.g. https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/3rd-jointtraining/ACT2021_AInness.pdf).

The authors of this manuscript work in the NASA Atmospheric Chemistry and Dynamics Laboratory, whose researchers have contributed to advances in atmospheric data assimilation for decades, so we have substantial familiarity with data assimilation. However, the methodology we describe here is fundamentally different from data assimilation in that we are only using the 3D model as a training dataset for a machine learning model. We are making no efforts to ingest satellite data into a 3D model to improve representation of any species within that model. At no point are we running our own 3D model. As such, we did not feel that a discussion of data assimilation methods in the introduction was warranted. We have, however, added the following (Line 84):

> Miyazaki et al. (2020) created a data assimilation framework that ingested satellite observations of CO, NO_2, O_3, and HNO_3 (nitric acid) into multiple CTMs. The data assimilation reduced the spread in average OH among the models and brought the interhemispheric ratio closer to unity, in line with values suggested by MCF observations (e.g. Patra et al., 2014). These results suggest that the incorporation of satellite observations into a modeling framework can improve the representation of OH.

For the discussion in Section 6, we acknowledge that there are similar applications between data assimilation methods and what we propose here, although again, they are fundamentally different in that we are not proposing to re-run a 3D model. We have added the following to the manuscript (Line 754):

> "This would serve as a computationally efficient complement to other methodologies constraining models with observations (e.g. Miyazaki et al., 2020, Miyazaki et al., 2021) to identify the impacts of these errors on the atmospheric oxidation capacity."

Second, some authors worked their whole life on the subject of OH, satellites, and models, and do not receive even a citation in the manuscript (e.g. https://acp.copernicus.org/articles/20/931/2020/). I am not claiming that the work in this paper is useless. What I am saying is that the added value could be much more when proper credit and discussion is dedicated to related studies.

We have added the following references to the paper, in addition to the Miyazaki et al (2020) and Boersma et al (2018) papers suggested by the reviewer, so that the cited works are more comprehensive:

Stevenson et al. (2020) (Line 52): "…with modeled trends disagreeing with those derived from observationally constrained methods (Stevenson et al., 2020)."

Wild et al. (2020) (Line 55): "Using Gaussian emulation, Wild et al. (2020) found that the relative importance of drivers of OH variability differed widely among three CTMs."

Spivakovsky et al. (1990) and Lelieveld et al. (2016) (Line 64): "This spatial heterogeneity is further caused by the large variation in the relative importance of drivers of OH loss and production in different regions of the atmosphere (e.g. Spivakovsky et al., 1990;Lelieveld et al., 2016)."

Naus et al. (2021) (Line 76) : "…although there has been recent success when using three dimensional inversion techniques (Naus et al., 2021)."

Patra et al. (2014) (Line 88): "The data assimilation reduced the spread in average OH among the models and brought the interhemispheric ratio closer to unity, in line with values suggested by MCF observations (e.g. Patra et al., 2014)."

Miyazaki et al. (2021)

We would also point out that OH modeling and chemistry are vast topics with decades of literature behind each, so we have focused on providing the background that we feel is necessary to lay the groundwork for our study while also keeping the paper to a manageable length.

In the revised manuscript, the authors should catch up with existing work and should discuss that in the introduction and discussion. This should replace the current self-centered manuscript with restricted references to work of other groups.

In the original manuscript, 12% of the cited works have one of the co-authors of this paper as the primary author, so we do not believe that our manuscript focuses too heavily on our work at the

expense of other researchers. When we do cite ourselves, it is mostly because we are building on previous work that is highly relevant to this study.

NO2 satellite data

More or less along the same lines. NO2 abundance in the remote tropics appears to be very important in determining TCOH. In section 5, the authors attempt to use alternative satellite products. In their evaluation of the results, they systematically refer to differences with their product as 'biases'. Although they evaluated to some extend their TCOH product against Atom data, with relative OK result, this does not imply that their product is OK in May 2018 (the analyzed month) and that all the other results are biased.

First, the evaluation with the ATom data is an evaluation of the methodology itself, not of the satellite product TCOH. We use observations from the ATom campaign as inputs to the model to determine whether we can reproduce observed TCOH. This is different than evaluating the satellite product determined from the OMI/MOPITT/AIRS observations. Nowhere in the paper do we make assertions on the absolute accuracy of the OMI/MOPITT/AIRS TCOH product, primarily, because as we state in the conclusions, there are insufficient observations of TCOH to fully evaluate the product.

Second, it was not our intention to judge the absolute accuracy of any satellite retrieval or imply that any retrieval was better than another, as we clearly stated in the original manuscript (Line 628 of the original manuscript). Indeed, one of the conclusions of the work is that we need more observations in the remote atmosphere to determine which retrievals, if any, are accurate in this area. To make this point more explicitly clear, we have removed the word bias from the paper and use "difference" or analogous words. For example, the paragraph comparing OMI/MOPITT/AIRS retrievals to TROPOMI now reads (Line 664):

> In general, observations from TROPOMI agree with those from the satellites in Table 1, with the exception of NO_2 and HCHO. Ozone, $H_2O_{(v)}$, and CO from TROPOMI are highly correlated (r^2 of 0.85 or higher) and agree within 10% on average with their respective retrievals from OMI, MOPITT, and AIRS. On the other hand, TROPOMI KNMI-NO_2 is systematically higher (145% on average), and TROPOMI HCHO is 20% lower than their corresponding OMI retrievals. The higher TCOH from the TROPOMI product is consistent with the increase in NO_2, which would lead to higher secondary production of OH. Further, while TROPOMI KNMI-NO_2 is modestly correlated with OMI NO_2 (r^2 = 0.61), TROPOMI and OMI HCHO are not correlated (r^2 = 0.23), highlighting the difficulty of the HCHO retrieval. Note that we are not seeking to determine which retrieval, if any, is more accurate. We are highlighting the differences to emphasize the impact that systematic differences in retrieval magnitudes of GBRT model inputs can have on the resultant TCOH.

On top of that, they fail to refer to an extensive (EU-funded) program QA4eCV in which the NO2 products (e.g. of OMI) have been evaluated (e.g. https://amt.copernicus.org/articles/11/6651/2018/). This effort is so central to the discussion, that it really shameful that relevant literature is not cited.

We have added a reference to this paper (Line 776):

> Recent efforts, such as the QA4ECV (Quality Assurance for the Essential Climate Variables, to improve NO_2 retrieval algorithms have reduced uncertainty, particularly over land (Boersma et al., 2018), although it is unclear how the accuracy of these retrievals translates to the remote tropics as validation data are still extremely limited.

As you point out, our work is centered in the remote tropics, a region where there is very little published literature with NO_2 retrieval validation. The QA4eCV paper only shows validation of the OMI product against observations at one site in China, a region outside of our study domain. Because uncertainties over the remote ocean are generally higher than in more polluted regions, it is unlikely that these results are applicable to the research region. In the original manuscript, we do cite the work of Wang et al (2020) who evaluated the TROPOMI retrieval used in this study over the remote Pacific. We now also include a reference to Verhoelst et al, 2021 which compares TROPOMI retrievals over Reunion to ground-based observations (Line 626).

> Wang et al. (2020) found that this retrieval was biased high when compared to ship-based observations from a MAX-DOAS instrument over the remote oceans, while Verhoelst et al. (2021) found good agreement between the retrieval and ground-based observations in Reunion.

I found the paper a pleasant read, presenting an interesting view for future exploration. In that respect, publication is possible, but the paper should discuss and give credit to internationally well-established efforts, which would require a major overhaul of the introduction and discussion.

Specific Comments:

Line 144 – 145: Nothing said yet about satellites… I assume OMI?
That is correct. We have added the following text for clarification (Line 155):

> We use instantaneous OH output from MERRA2 GMI at 14:00 local time for each day of a given month across the years 2005 to 2019, a timeframe that maximizes overlap between the operational lifetime of the satellites listed in Table 1 and the period of the MERRA2 GMI simulation.

Line 148: I can imagine that cloud fractions < 0.3 still could be included in machine learning?
As stated in the manuscript, we omit grid boxes with cloud fractions greater than 0.3 in the training dataset. While those values could be included, we wish to make the training dataset as close as possible to the satellite data that will be used as inputs. We've clarified this in the text (Line 158):

> For a given month and year, we calculate daily tropospheric column values across the grid, filtering out columns where the maximum cloud fraction in that column was greater than 30% in order to align the training targets more closely with satellite data, where retrievals of some species are often filtered for cloud cover.

Table 1: CO overpass is not at 14:00.
We mention the MOPITT overpass time in Lines 192 and again in Lines 271. For clarification, we have also added the following text to the Table 1 caption:

> Overpass times are ~13:30 LST for all satellites except MOPITT, which has a ~10:30 LST overpass.

Line 174: Why not 14:00?

The time used for the SZA calculation is arbitrary. Because we're looking at monthly averages and SZA is independent of longitude when calculated for a given LST, SZA in the GBRT model primarily provides additional information about the distance from the latitude with maximum insolation. Because this is not a process-based model, the actual SZA value is not relevant to the calculation, only the relative difference of the SZA from one location to another.

Line 232: Mention if you use tropospheric sub-column (stratospheric correction…)
Good point. We mention that we use the tropospheric column NO_2 in Table 1 but failed to highlight that here. We have made the correction.

Line 247: Above, 10 LST
While the MOPITT overpass is ~10:30 LST, output from MERRA2 GMI is at 10:00 LST. We have clarified this point in the text (Line 144):

> Output is available at daily- and monthly-averaged resolution, as well as instantaneous values at 10:00 and 14:00 LST. These times are within approximately 30 minutes of the overpass times of the satellites described in Section 2.2.

Line 420: Rather strange. What differences. I could imagine that you could perform an analysis by which the satellite data are one-by-one replaced by the model counterpart, to understand what drives the lower TCOH in the multi-satellite product compared to the model. Why would this be beyond the scope?
This type of analysis is certainly possible, as demonstrated between our comparisons of the OMI and TROPOMI NO_2. Our goal with this manuscript, however, is not to understand the differences between the MERRA2 GMI simulation and a satellite-constrained OH product. As stated in the last paragraph of the introduction, our goals for this manuscript are "assessing the feasibility of our methodology, identifying potential limitations, and suggesting areas of improvement in the current observational network." While understanding these differences are interesting questions and are something we will most likely look at in future work, adding a discussion here would be a distraction from our intended objectives.

Line 505: now I am confused. I though Atom TCOH was derived from measured OH? How this suggests a box model is employed to derive Atom OH.
In response to the second reviewer, we have removed this discussion.

Line 543: This is a strange addition, because you seem to propagate errors through the GBRT model.
We've reworded the text to clarify our point (Line 577):

> "In contrast to NO_2, uncertainties in TCOH resulting from HCHO maximize in regions with higher HCHO columns (Fig. 6). The magnitude of that uncertainty is likely an overestimate as the actual retrieval uncertainty for HCHO in these regions is significantly lower than the value assumed for the error analysis."

Line 553: Some indication of these scales is needed. Yearly? 1x1 degree?
We specify the temporal and spatial scales as monthly and 1 x 1 degree in the second paragraph of the section.

Line 583: I thought this was only for the NO2 product?

The text now reads “Horizontal resolution for the month examined here (May 2018) is as high as 7km x 3.5 km at nadir”.

Line 607: this now becomes very messy. Above you say that TROPOMI has a water vapour product. So, why not include that in the training (I guess the model has to be sampled slightly different)
TROPOMI has a water vapor column product, but, as stated in the text, there is not an analogous water vapor layers product to that provided by AIRS. We have added the word column to clarify further that we are using the TROPOMI column product. As the text stands, it clearly states that we created a new model with all inputs except for the water vapor layers, implying that we include the water vapor column.

Line 612: This makes more sense in section 5.3.1, which should then be renamed.
We have renamed Section 5.3.1 and moved the first paragraph of Section 5.3.2 to the end of Section 5.3.1.

Line 616: AIRS does not make sense here, because AIRS is not used
AIRS is used. As discussed above and in the text, we only removed the water vapor layers from the product, not the water vapor column.

Line 622: I do not agree with the wording here. Why would this estimate be overestimated. The validation of the other product with Atom was not that convincing. So, I propose not to qualify one product better than another...
The text now reads (Line 659):

> While there is modest correlation between the two (r^2 = 0.63), the TROPOMI product is 27.6% larger than the OMI/MOPITT/AIRS product, with higher values across almost the entire domain.

Line 627: again, wording suggests that OMI/MOPITT/AIRS is the truth.
As we explicitly state in the text, we are not judging the accuracy of any particular retrieval. We are not saying that TROPOMI is wrong and OMI/MOPITT/AIRS retrievals are correct. We reported the difference in TROPOMI retrievals with respect to the other retrievals because those are the retrievals that underlie our baseline TCOH product. We have changed the paragraph to the following to make this point even more explicitly clear (Line 664):

> In general, observations from TROPOMI agree with those from the satellites in Table 1, with the exception of NO_2 and HCHO. Ozone, $H_2O_{(v)}$, and CO from TROPOMI are highly correlated (r^2 of 0.85 or higher) and agree within 10% on average with their respective retrievals from OMI, MOPITT, and AIRS. On the other hand, TROPOMI KNMI-NO_2 is systematically higher (145% on average), and TROPOMI HCHO is 20% lower than their corresponding OMI retrievals. The higher TCOH from the TROPOMI product is consistent with the increase in NO_2, which would lead to higher secondary production of OH. Further, while TROPOMI KNMI-NO_2 is modestly correlated with OMI NO_2 (r^2 = 0.61), TROPOMI and OMI HCHO are not correlated (r^2 = 0.23), highlighting the difficulty of the HCHO retrieval. Note that we are not seeking to determine which retrieval, if any, is more accurate. We are highlighting the differences to emphasize the impact that systematic differences in retrieval magnitudes of GBRT model inputs can have on the resultant TCOH.

Line 645: ??? This overvalues the capabilities of the machine learning model...I agree it would somehow represent the non-lineair nature, but here things are mixed up....
It's unclear what you mean by this statement. Swapping in variables to understand their relative importance on a target variable, even for non-linear systems, has been done before (see, for example, Nicely et al. 2016).

Line 693: get more confused: is this against observation of a (potentially biased low?) OMI retrieval
We have changed this sentence to read:

When compared to OMI, the MINDS NO_2 retrieval is 58% higher, as compared to 145% higher for the KNMI retrieval.

Line 674: what I understand is that this information is available, but not analysed in this manuscript? Why not?
As we state in the introduction, the point of this manuscript is not a detailed analysis of OH using the TCOH product, rather it is to demonstrate the viability of the methodology and to understand its strengths and weaknesses. An analysis of OH temporal and spatial variability is a topic unto itself and will be examined in future work.

Line 679: VOC chemistry is also present in the tropics. So this is more a land-ocean aspect.
Yes, we agree that there is VOC chemistry in the tropics. To make it more explicitly clear we now say "Expansion of the product over land will likely require…".

Line 683: Very USA centric. You fail to refer to S4 (Copernicus) and GEMS (Asia, already launched.)....
The text now reads (Line 723):

> For example, current and upcoming geostationary air quality satellites such as Sentinel 4, TEMPO (Tropospheric Emissions: Monitoring of Pollution), and GEMS (Geostationary Environment Monitoring Spectrometer) could provide most of the necessary inputs to the machine learning model…

Line 707: See my main point: this methodology is followed by other groups, but you fail to address these methods in the introduction....
We have added the following (Line 753):

> "This would serve as a computationally efficient compliment to other methodologies constraining models with observations (e.g. Miyazaki et al., 2020) to identify the impacts of these errors on the atmospheric oxidation capacity."

Line 732: I think the issue is NOT the instrument design, but simply there is not enough effort to perform calibration and validation of satellite products. There has been a huge effort in Europe (QA4ECV) including NO2. Seems the authors are unaware if this, which is kind of frightening in light of this paper.

We are aware of this effort and similar efforts at NASA (e.g. the MINDS project mentioned in the paper). The point of the final paragraph of the paper is to highlight the need for further validation of satellite

retrievals in the remote tropics, the region relevant to this work. While there have been several validation efforts of the various OMI and TROMPOMI retrievals in polluted regions, to our knowledge, there is little published literature in the remote atmosphere. The QA4ECV paper you cite does show excellent agreement between ground-based observations from one site in China and the OMI retrieval, but this is not located in our target region. And as NO_2 retrievals have evolved with time, their uncertainties have definitely improved, particularly in polluted regions, but, based on our understanding of the literature (e.g. Lamsal et al, 2021), in the remote atmosphere, uncertainties still remain high because of the methodology used to separate the tropospheric portion from the total column.

The final paragraph now reads (Line 772):

> Finally, accuracy of the TCOH product is dependent on the accuracy of the satellite retrievals input into the machine learning model, with the NO_2 retrieval having the largest effect. To reduce the uncertainty of the TCOH product, more information about the accuracy of individual NO_2 retrievals is required. Currently, there is little validation of OMI and TROPOMI NO_2 retrievals in the remote, tropical atmosphere, so it is difficult to assess, which retrievals, if any, are correct. Recent efforts, such as the QA4ECV (Quality Assurance for the Essential Climate Variables, to improve NO_2 retrieval algorithms have reduced uncertainty, particularly over land (Boersma et al., 2018), although it is unclear how the accuracy of these retrievals translates to the remote tropics as validation data are still extremely limited. Even retrievals of TROPOMI and OMI made with the same algorithm show differences, suggesting that instrumental differences could also affect the results. Future satellite missions should focus on trying to reduce the uncertainty in NO_2 retrievals, particularly in the remote atmosphere, both through improvements in instrument design and algorithm development.

References:

Boersma, K. F., Eskes, H. J., Richter, A., De Smedt, I., Lorente, A., Beirle, S., van Geffen, J. H. G. M., Zara, M., Peters, E., Van Roozendael, M., Wagner, T., Maasakkers, J. D., van der A, R. J., Nightingale, J., De Rudder, A., Irie, H., Pinardi, G., Lambert, J. C., and Compernolle, S. C.: Improving algorithms and uncertainty estimates for satellite NO2 retrievals: results from the quality assurance for the essential climate variables (QA4ECV) project, Atmos. Meas. Tech., 11, 6651-6678, 10.5194/amt-11-6651-2018, 2018.

Lelieveld, J., Gromov, S., Pozzer, A., and Taraborrelli, D.: Global tropospheric hydroxyl distribution, budget and reactivity, Atmos. Chem. Phys., 16, 12477-12493, 10.5194/acp-16-12477-2016, 2016.

Miyazaki, K., Bowman, K. W., Yumimoto, K., Walker, T., and Sudo, K.: Evaluation of a multi-model, multi-constituent assimilation framework for tropospheric chemical reanalysis, Atmos. Chem. Phys., 20, 931-967, 10.5194/acp-20-931-2020, 2020.

Miyazaki, K., Bowman, K., Sekiya, T., Takigawa, M., Neu, J. L., Sudo, K., Osterman, G., and Eskes, H.: Global tropospheric ozone responses to reduced NOx emissions linked to the COVID-19 worldwide lockdowns, Science Advances, 7, eabf7460, 10.1126/sciadv.abf7460, 2021.

Naus, S., Montzka, S. A., Patra, P. K., and Krol, M. C.: A three-dimensional-model inversion of methyl chloroform to constrain the atmospheric oxidative capacity, Atmospheric Chemistry and Physics, 21, 4809-4824, 10.5194/acp-21-4809-2021, 2021.

Patra, P. K., Krol, M. C., Montzka, S. A., Arnold, T., Atlas, E. L., Lintner, B. R., Stephens, B. B., Xiang, B., Elkins, J. W., Fraser, P. J., Ghosh, A., Hintsa, E. J., Hurst, D. F., Ishijima, K., Krummel, P. B., Miller, B. R., Miyazaki, K., Moore, F. L., Muhle, J., O'Doherty, S., Prinn, R. G., Steele, L. P., Takigawa, M., Wang, H. J., Weiss, R. F., Wofsy, S. C., and Young, D.: Observational evidence for interhemispheric hydroxyl-radical parity, Nature, 513, 219-223, 10.1038/nature13721, 2014.

Spivakovsky, C. M., Yevich, R., Logan, J. A., Wofsy, S. C., McElroy, M. B., and Prather, M. J.: Tropospheric OH in a three-dimensional chemical tracer model: An assessment based on observations of CH3CCl3, Journal of Geophysical Research: Atmospheres, 95, 18441-18471, 10.1029/JD095iD11p18441, 1990.

Stevenson, D. S., Zhao, A., Naik, V., amp, apos, Connor, F. M., Tilmes, S., Zeng, G., Murray, L. T., Collins, W. J., Griffiths, P., Shim, S., Horowitz, L. W., Sentman, L., and Emmons, L.: Trends in global tropospheric hydroxyl radical and methane lifetime since 1850 from AerChemMIP, Atmos. Chem. Phys., 10.5194/acp-2019-1219, 2020.

Verhoelst, T., Compernolle, S., Pinardi, G., Lambert, J. C., Eskes, H. J., Eichmann, K. U., Fjæraa, A. M., Granville, J., Niemeijer, S., Cede, A., Tiefengraber, M., Hendrick, F., Pazmiño, A., Bais, A., Bazureau, A., Boersma, K. F., Bognar, K., Dehn, A., Donner, S., Elokhov, A., Gebetsberger, M., Goutail, F., Grutter de la Mora, M., Gruzdev, A., Gratsea, M., Hansen, G. H., Irie, H., Jepsen, N., Kanaya, Y., Karagkiozidis, D., Kivi, R., Kreher, K., Levelt, P. F., Liu, C., Müller, M., Navarro Comas, M., Piters, A. J. M., Pommereau, J. P., Portafaix, T., Prados-Roman, C., Puentedura, O., Querel, R., Remmers, J., Richter, A., Rimmer, J., Rivera Cárdenas, C., Saavedra de Miguel, L., Sinyakov, V. P., Stremme, W., Strong, K., Van Roozendael, M., Veefkind, J. P., Wagner, T., Wittrock, F., Yela González, M., and Zehner, C.: Ground-based validation of the Copernicus Sentinel-5P TROPOMI NO2 measurements with the NDACC ZSL-DOAS, MAX-DOAS and Pandonia global networks, Atmos. Meas. Tech., 14, 481-510, 10.5194/amt-14-481-2021, 2021.

Wild, O., Voulgarakis, A., amp, apos, Connor, F., Lamarque, J.-F., Ryan, E. M., and Lee, L.: Global sensitivity analysis of chemistry-climate model budgets of tropospheric ozone and OH: exploring model diversity, Atmospheric Chemistry and Physics, 20, 4047-4058, 10.5194/acp-20-4047-2020, 2020.

Reviewer 2

This study introduces a new framework to infer tropospheric column OH, TCOH, over the tropical remote oceans based on a random forest regression and gradient boosted regression trees (GBRT) techniques using outputs from MERRA2 and several satellite observation data sets, including trace gaces and H2O. Satellite-based TCOH estimates were compared to the OH field of the original MERRA2 data, and an independent validation was then conducted against the ATom aircraft measurements. This methodology appears to be unique and innovative. The discussion section also contains interesting implications. Nevertheless, I have some concerns as described below. If the authors address them, this paper can be published. Note that another reviewer already posted many constructive comments to improve the manuscript, especially with regard to the relevance to other studies and the value of other satellite data. I have similar concerns and agree with most of the comments, so I do not duplicate the concerns in my comments.

1. Because of the localized non-linear chemistry, the use of aggregated satellite information at the monthly scale and 1x1 degree resolution would not correctly capture the OH distributions needed to predict detailed chemical mechanisms and then OH mean states and variability. Daily

L2 data would clearly be a better input for better use of satellite information. Combining the L2 retrieval uncertainty information (and averaging kernels) provided for each pixel may also allow for improved use of satellite data products in ML. Because the current framework does not yet fully take the advantage of satellite products, the implications for future satellite requirements obtained may be limited or biased. In particular, the demonstrated large biases indicate the need for further refinement. Consistent with my concerns, Nicely et al. (2020) clearly stated that "Much future work is needed, though; observations must be incorporated to introduce a ground truth element to this analysis in a manner that either adjusts for or avoids disconnects between coarse versus local/instantaneous spatiotemporal scales and appropriately accounts for measurement uncertainty; an analysis of model output with much higher temporal frequency is needed to identify exactly where model differences in chemical mechanisms lie". I see no reason to continue to use the aggregated L3/L4 data in this study. The increased computational costs should be manageable with a computationally efficient ML approach.

We agree that higher temporal and spatial resolution would aid in both understanding OH variability and providing information on understanding the relative importance of OH drivers in different portions of the atmosphere. We have, however, focused on a 1° x 1° horizontal resolution and monthly-averaged data in this work for two reasons.

First, the satellite retrievals used as the basis for this study require relatively coarse spatial and temporal scales. We added the following to the text (Line 240):

> We use these resolutions because, in the study domain, individual pixel retrievals, particularly of NO_2 and HCHO, are frequently at or below detection limits (Gonzalez Abad, 2015; Lamsal, 2021), necessitating averaging to relatively coarse temporal and spatial scales. Missing data due to cloud cover and the OMI row anomaly further increase the need for monthly-scale averaging. While other satellites, such as OMPS and TROPOMI, provide retrievals with increased signal to noise ratios and more complete data coverage, the satellites used here cover a far longer time period. The 1.0° × 1.0° and monthly resolutions, in combination with the long data record, are sufficient to understand regional trends in TCOH and some aspects of TCOH temporal and spatial variability.

Second, the goal of this particular work is to demonstrate the validity of our approach to constraining TCOH. Doing this work at a relatively coarse resolution allows us to explore the feasibility of the methodology without the potential complications of the noisier data from finer temporal and spatial resolutions. We also note that, even at 1° x 1°, the product we present here provides observational constraints on OH at far finer scales than existing methods (e.g. MCF inversions).

2. With regard to the uncertainty discussion and the predicted positive bias against the Atom measurements, it is essential to understand the relative importance of each satellite measurement used in GBRT, in order to provide an optimal framework for inferring TCOH and suggesting future satellite measurements meaningfully. This can be simply done by applying each satellite data set separately in GBRT, validated against Atom measurements, similar to observing system impact analysis widely used in data assimilation. Furthermore, whether removing the OMI NO2 from the calculation reduces the positive bias against Atom is a relevant question here (see also my comments below). Meanwhile, the random forest feature importance would provide

additional information on the relative values. The current uncertainty section includes a related discussion and provides suggestions on the relative contribution of each measure, but is more indirect and limited. The suggested additional effort should also help to better understand the comparisons of OMI/AIRS/MOPITT and TROPOMI results. While the purpose of this paper is to present a methodology, understanding the role of each measurement cannot be ignored to ensure that the proposed methodology that combines multiple satellite data works properly and synergistically.

As discussed in response to your next point, a GBRT model that excludes NO_2 as an input shows a slight increase in agreement between the model and observations, suggesting that errors in the ATom NO_2 might be contributing to measurement/model disagreement. In addition, we note that when the model excluding NO_2 is applied to the hold-out set, the NRMSE increases by about 50%, suggesting a strong degradation in performance which could result in additional errors that result in a high bias. See the response to your next point for further discussion and changes to the text in regard to this point.

The uncertainty analysis discussed in Section 5.2 suggests that the model responds strongly to changes in NO_2 and to a lesser extent HCHO. Likewise, TROPOMI analysis in section 5.3, in which we systematically replace TROPOMI observations into the OMI/AIRS/MOPITT product, demonstrates a large response (29%) to changes in NO_2 and a much more muted response (3%) to changes in HCHO, despite the poor agreement between OMI and TROPOMI HCHO retrievals. Both these examples suggest that NO_2 values are more important to the final TCOH than HCHO, whereas the feature importance has HCHO as the most important variable. We have added the following to the text, along with Figure S11, which we also reproduce below (Line 582):

> This uncertainty analysis is in general agreement with the model feature importance (Supplementary Fig. 11), a measure of the relative importance of GBRT model inputs, where HCHO and NO_2 consistently have the largest values of the satellite inputs.

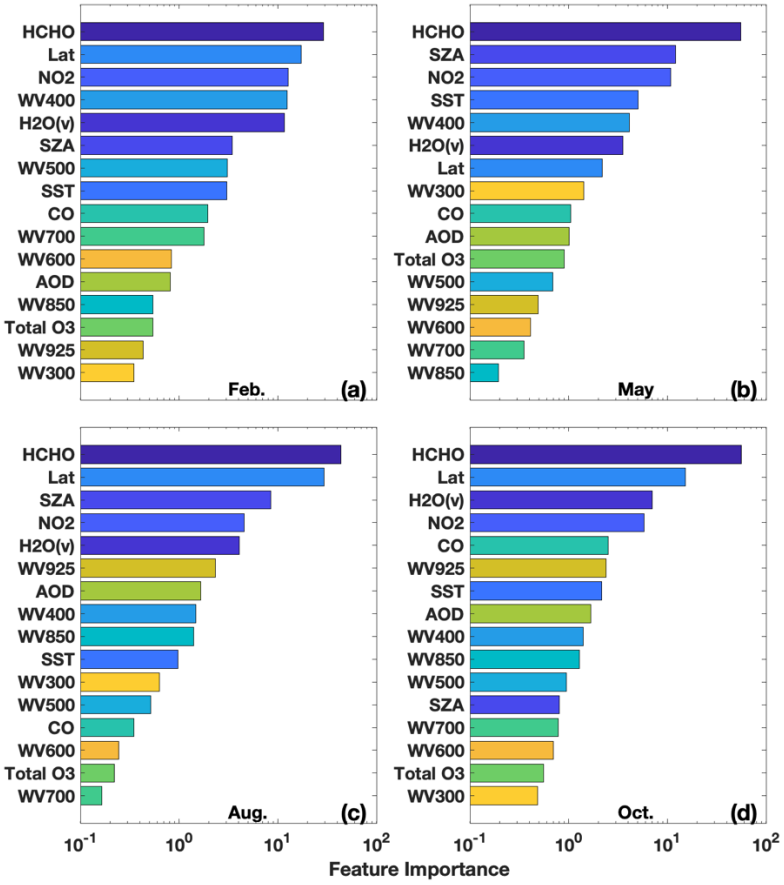


Figure S1: Feature importance, sorted by value, for the GBRT model for February (a), May (b), August (c), and October (d). Bars are colored so that variables have the same color in each panel.

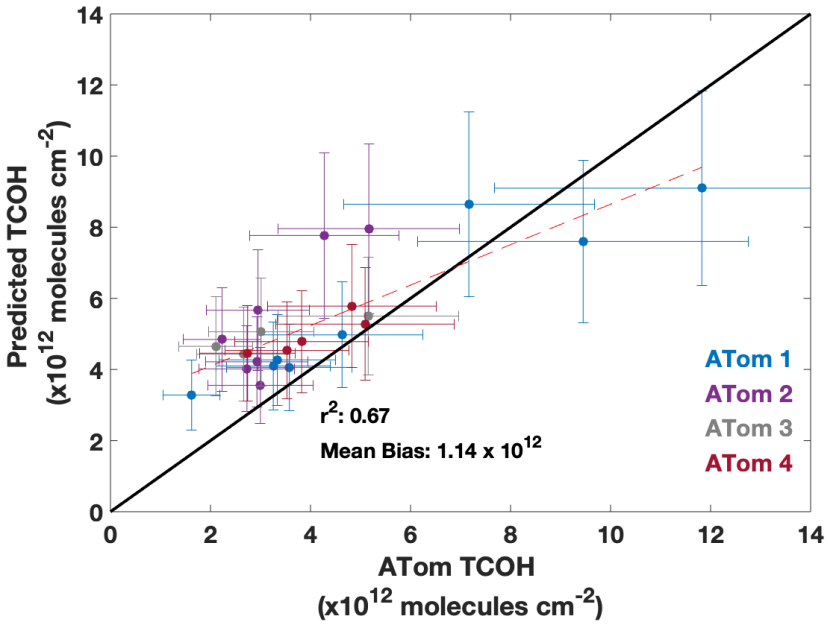
3. Although the variability of OH is relatively well reproduced, the large positive bias against AToM measurements remains a serious concern. Knowing the realistic OH magnitude can even be more important than variability for some important applications, such as chemical lifetime estimation. The highly biased estimates can have limited impacts on future applications. Several potential error sources are discussed in the manuscript, but they are not very convincing. The first point, "spanned 300 – 400 km in latitude" might not be the main reason, as the authors also discussed. The second point, "if a large fraction of the tropospheric column of one input was outside the range of the ATom profile, this would likely cause large errors in calculated TCOH." can be verified by comparing the entire tropospheric column with that based on ATom sampling, using the OH field in MERRA2. This requires assuming that MERRA-2 provides realistic vertical profiles, while multi-model simulation data can provide that uncertainty information. As for the third point, "Recalculating the TCOH from ATom with NO2 from a box model constrained with NO observations", its approach and validity are unclear in the manuscript. While the purpose of this paper is to present the methodology, the reasons for the large positive bias need to be further explored to clarify why the proposed framework still does not reproduce the observed OH values that are essential for chemical lifetime estimation.

In order to make a more direct comparison between the ATom columns and the GBRT output as well as to address the concern about inputs being outside of the range of the ATom profile, we have changed the way we evaluate the GBRT model using ATom data. We describe this in the text as follows (Line 499):

> Because ATom profiles did not span the entire tropospheric column, we trained a separate GBRT model where OH and all tropospheric column input variables were substituted for columns spanning 990 – 250 hPa, the median range of ATom profiles. This allows for a more direct comparison between observed and modeled TCOH.

Limiting the training dataset to the range of the ATom profiles greatly improves the comparison between observed TCOH and that predicted by the GBRT model, with the near uniform high bias of 2.87×10^{12} molecules/cm^2 being reduced by more than a factor of two to 1.14×10^{12} molecules/cm^2, where the model somewhat underpredicts TCOH at higher values, although still within the observational uncertainty. The r^2 also increases slightly from 0.61 to 0.67. The portion of the tropospheric OH column in MERRA2 GMI outside the vertical extent covered by sampling during ATom was 22% for February 2005, on average. Consistent with this value, predicted columns from the modified GBRT model were ~30% lower than for the model trained on the full tropospheric column. The high bias presented in the initial version of the manuscript was therefore an artifact resulting from the difference in the range of the ATom columns from the training dataset. We have omitted references to the previous comparison in the text, and only use the new comparison here. The discussion of the comparison in the text now reads (Line 509):

> The GBRT model captures the variability of the observed TCOH. While there is a modest overall high bias, the median normalized absolute error of 26% is within observational uncertainty. When applied to all ATom deployments, predicted TCOH is correlated with the observations with an r^2 of 0.67 and a mean bias of 1.14×10^{12} molecules/cm^2 (Fig.5). Many of the data points agree within the combined modeled and observational uncertainty. The r^2 values for individual deployments are 0.88 for ATom 1, 0.73 for ATom2, and 0.78 for ATom 3 and 4. The level of agreement between observed and predicted OH is comparable or better than that of other methods to infer OH from space. For example, Pimlott et al. (2022) found an r of 0.78 (r^2 = 0.61) when estimating ATom OH using a steady state approach, with r values ranging from 0.51 to 0.85 (r^2 of 0.26 to 0.72) for the different deployments. The level of agreement we show here therefore demonstrates the validity of the machine learning method to capture the variability of OH.



Updated Figure 2: Regression of TCOH observed from the ATom deployments against that predicted from the GBRT model. Error bars represent the 2σ observational uncertainty as reported in Brune et al. (2020) and the GBRT uncertainty described in Section 5.2. The r^2 of a linear least squares fit and the mean bias are also shown.

There is still disagreement between observations and the predicted OH, however, although the median normalized absolute error is 26%, within observational uncertainty. NO_2 is still a likely contributor to at least some of this disagreement, and we include the following discussion in the text (Line 521):

> The source of the model/measurement disagreement, with over- and underprediction at low and high column content respectively, is unclear, although there are multiple potential error sources. For example, a typical profile taken during ATom spanned 300 – 400 km in latitude, disconnecting the top and bottom of the profile in space. This is in contrast to the data used to train the model, which were vertical columns over one location. This could lead to a degradation in model performance when applied to ATom, since the columns are not directly analogous to the training dataset. These effects are likely limited because ATom observations are in the remote atmosphere, where the spatial distribution of relevant species is likely to be more homogeneous than over land.
>
> Further, there is a known interference with the ATom NO_2 observations, suggesting another possible contributor to disagreement between measured and modeled OH. Because of thermal degradation of NO_2 reservoir species, such as organic nitrates and peroxyacetyl nitrate, in the instrument inlet, ATom NO_2 observations are likely biased high (Silvern et al., 2018;Shah et al., 2023;Nault et al., 2015). To test the potential impact of NO_2 on the predicted OH columns, we applied the ATom observations to a model that omits NO_2 as an input. Removing NO_2 increases the r^2 to 0.74, decreases the mean bias to 0.82×10^{12} molecules/cm^2, and decreases the median normalized absolute error slightly to 25.7% (Fig. S8). These improvements in performance suggest that errors in NO_2 could be contributing to the measure/model differences. Omitting NO_2 does, however, likely introduce additional errors as NO_X compounds are essential to OH production in some regions of the atmosphere. When we apply the hold out set from MERRA2 GMI to

this model, for example, the NRMSE increases by approximately 50%, highlighting the importance of keeping NO_2 as an input variable.

For more certain evaluation of the GBRT model with observations, greater certainty in the in situ NO_2 observations is needed. Although the in situ observations are insufficient to evaluate the absolute accuracy of the product, the results presented here demonstrate that a machine learning model trained on data from a CTM simulation can capture TCOH variability in the actual atmosphere and suggest that predicted OH columns agree with observations within instrumental uncertainty.

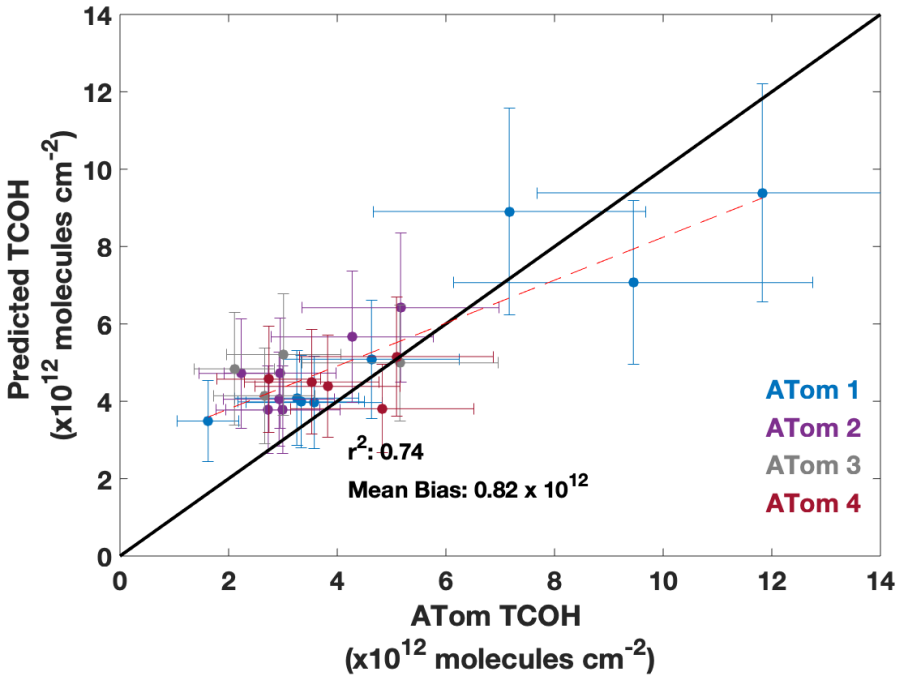


Figure S3: Same as Figure 5 except using a GBRT model that omits NO_2 as an input.

4. The large discrepancy between the MERRA2 and satellite HCHO remains a concern. This could lead to significant degradation in OH predictions. This can be demonstrated based on the ML framework with and without HCHO data.

We have trained a model that omits HCHO as an input variable, finding little difference from the satellite constrained OH product that includes all inputs. We now mention that here (Line 339):

> Similarly, the satellite-constrained TCOH product discussed in Section 4.2 differs by only 3% on average for one determined with a GBRT model that excludes HCHO as an input, suggesting the limited impact of potential errors in the MERRA2 GMI HCHO distribution on model performance.

These results are also consistent with the relative small change in TCOH when using TROPOMI vs OMI HCHO discussed in Section 5.3. While this analysis suggests that we could exclude HCHO from the model, we continue to include it because of its importance as a proxy for other VOCs and their role as an OH sink.

5. Future discussion is needed on satellite data products. In particular, satellite column measurements should have different vertical information due to different vertical sensitivities and profiles among measurements and variables. Meanwhile, OH variability can be largely independent between the lower and upper troposphere. This would complicate the prediction and interpretation of TCOH.

As we discuss in Section 2.1, we did not find a significant difference in the satellite constrained TCOH product when applying averaging kernel/shape factor information for CO and HCHO. This is likely because the model seems relatively insensitive to HCHO and applying MOPITT CO averaging kernel information does not markedly change the MERRA2 GMI CO. We do not apply OMI NO_2 averaging kernel/shape factor information as a GEOS simulation with a similar setup to MERRA2 GMI is used for the shape factors. This would not be the case for TROPOMI KNMI-NO_2, however, so this could be one source of error. We have added the following (Line 699):

> In addition, the training dataset does not take TROPOMI averaging kernels and shape factors into account, which could also contribute to the observed differences.

We agree that using tropospheric column OH could obscure variability/trends in different levels of the atmosphere. For instance, in MERRA2 GMI, ENSO-related OH variability in the UT is controlled by changes in NO_X while, near the surface, O_3-related variability drives OH. We do show in Anderson et al, 2021, however, that there is still an ENSO-related signal in the tropospheric column, so some information still exists. Further, tropospheric columns would still be a significant advance over the current observational constraints on tropical OH. As we reference in Section 6, we are investigating trying to use a similar methodology to constrain OH at different layers in the atmosphere, although that work is not significantly enough advanced to discuss here. We mention the utility of understanding OH drivers in different levels of the atmosphere here (Line 732):

> Vertically-resolved OH could also help understand differences in OH drivers in the upper and lower troposphere {Spivakovsky, 1990; Lelieveld, 2016}, which can often be decoupled from the column.

Reviewer 3

The authors present a gradient-boosted regression tree (GBRT) machine-learning (ML) model for tropospheric OH columns trained on synthetic satellite observations from the NASA Global Modeling Initiative (GMI) chemical transport model driven by the MERRA-2 meteorological reanalysis. They evaluate the ML model with available in situ observations from the ATom field campaign and then apply the model to actual satellite observation inputs for the recent past. The paper is well-written and clear, and a rare example of a manuscript I have reviewed that I think needs no modifications to be suitable for publication. That being said, I do think the other two reviewers have provided some helpful and constructive comments that would add to the discussion in useful ways. However, the scientific guts of the paper are sound and interesting, and I recommend publication.

We thank the reviewer for their time and response.