We thank the reviewer for their insightful and helpful comments. Our responses are shown below in red, while changes to the text are shown in blue.

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.

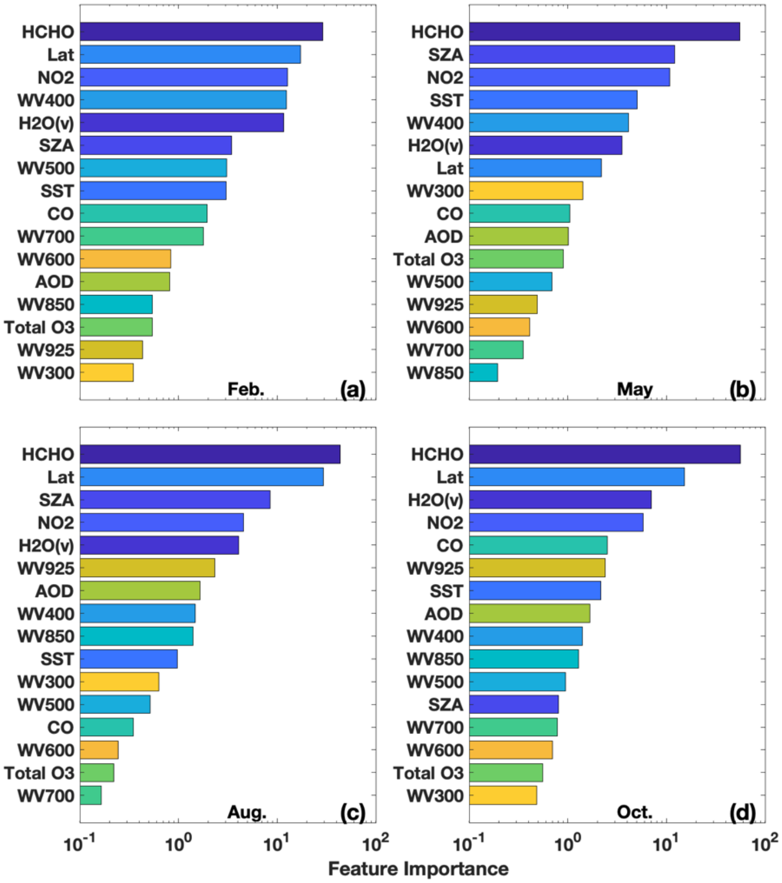


Figure S1: Feature importance, sorted by value, for the GBRT model for February (a), May (b), August (c), and October (d). Bars are colored so that variables have the same color in each panel.

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

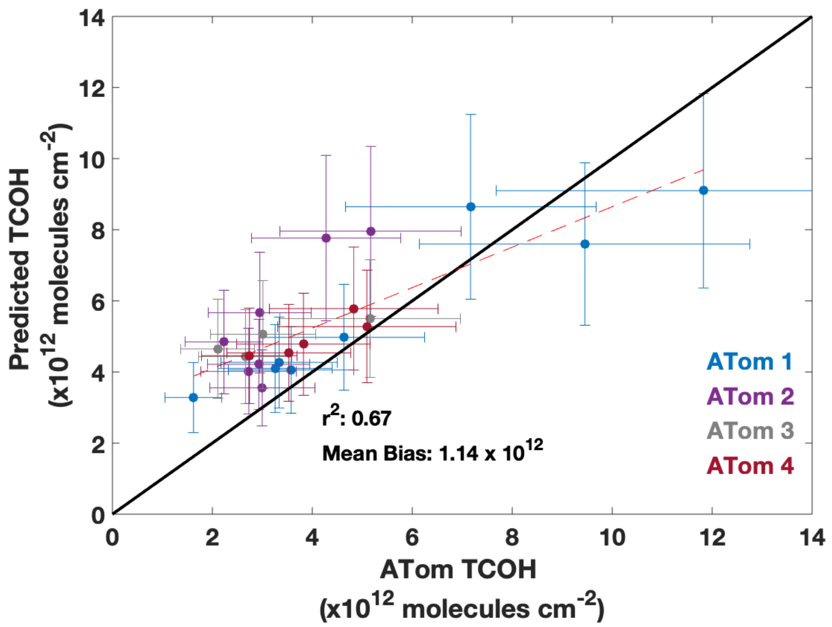
bias need to be further explored to clarify why the proposed framework still does not reproduce the observed OH values that are essential for chemical lifetime estimation.

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.

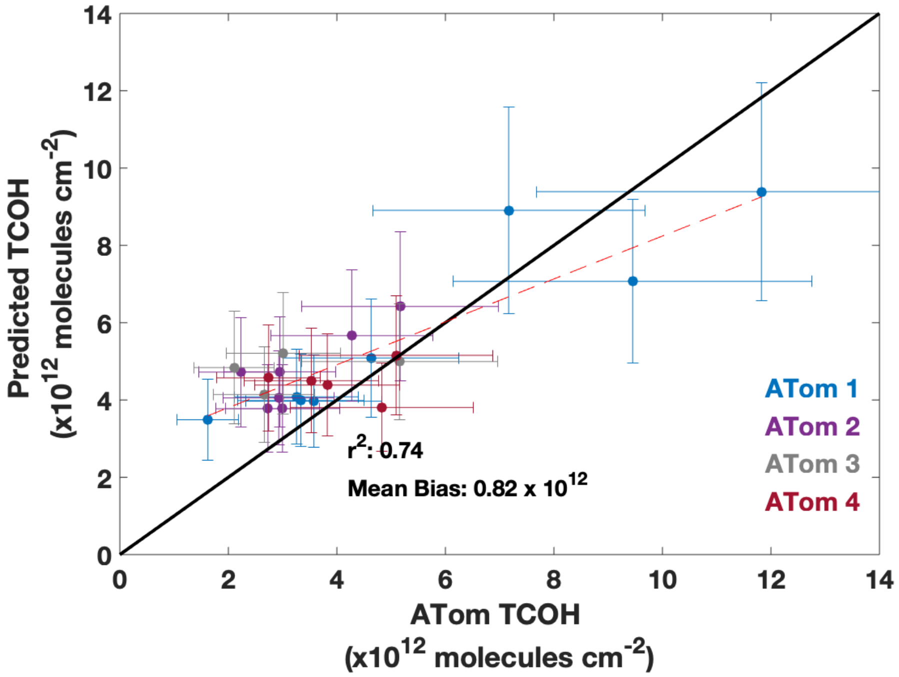


Figure S3: Same as Figure 5 except using a GBRT model that omits NO_2 as an input.

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.