1 **A Machine Learning Methodology for the Generation of a Parameterization of the Hydroxyl Radical: a**
2 **Tool to Improve Computational-Efficiency in Chemistry Climate Models**
3

4 Daniel C. Anderson[1,2], Melanie B. Follette-Cook[2,3], Sarah A. Strode[2,3], Julie M. Nicely[2,4], Junhua Liu[2,3],
5 Peter D. Ivatt[2,4], Bryan N. Duncan[2]
6
7 [1] GESTAR II, University of Maryland Baltimore County, Baltimore, MD, USA
8 [2] Atmospheric Chemistry and Dynamics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD,
9 USA
10 [3] GESTAR II, Morgan State University, Baltimore, MD, USA
11 [4] Earth System Science Interdisciplinary Center, University of Maryland, College Park, MD, USA
12
13 *Correspondence to*: Daniel C. Anderson (daniel.c.anderson@nasa.gov)

14 **Abstract**

15   We present a methodology that uses gradient boosted regression trees (a machine learning
16 technique) and a full-chemistry simulation (i.e., training dataset) from a chemistry climate model (CCM)
17 to efficiently generate a parameterization of tropospheric hydroxyl radical (OH) that is a function of
18 chemical, dynamical, and solar irradiance variables. This surrogate model of OH is designed to allow for
19 computationally-efficient simulation of nonlinear feedbacks between OH and tropospheric constituents
20 that have loss by reaction with OH as their primary sinks (e.g., carbon monoxide (CO), methane (CH_4),
21 volatile organic compounds (VOCs)). Such a model framework is advantageous for studies that require
22 multi-decadal simulations of CH_4 or multi-year sensitivity simulations to understand the causes of trends
23 and variations of CO and CH_4. The methodology that we present provides for the relatively easy creation
24 of a new parameterization in response to, for example, changes in the underlying CCM chemistry and/or
25 dynamics schemes. We show that a parameterization of OH generated from a CCM simulation is able to
26 reproduce OH concentrations with a normalized root mean square error of approximately 5%, as well as
27 capturing the global mean methane lifetime within approximately 1%. The accuracy of the
28 parameterization is dependent on inputs being within the bounds of the training dataset. However, we
29 show that the parameterization predicts large deviations in OH for an El Niño event that was not part of
30 the training dataset, and that the spatial distribution and strength of these deviations are consistent
31 with the event. This result gives confidence in the fidelity of the parameterization to simulate the spatial
32 and temporal responses of OH to perturbations from large variations in the chemical, dynamical and
33 solar irradiance drivers of OH. In addition, we discuss how two machine learning metrics, Gain feature
34 importance and SHAP values, indicate that the behavior of the parameterization of OH generally
35 comports with our understanding of OH chemistry, even though there are no physics- or chemistry-
36 based constraints on the parameterization.
37

38 **1.0 Introduction**
39   The hydroxyl radical (OH) is the dominant tropospheric oxidant. It removes numerous species from
40 the atmosphere, including carbon monoxide (CO) and methane (CH_4), which are the largest OH sinks on
41 a global scale (Spivakovsky et al., 2000;Spivakovsky et al., 1990). Recent trends in CH_4, the second most
42 important anthropogenic greenhouse gas, can potentially be explained by changes in OH abundance
43 (Rigby et al., 2017), although changes in emissions are also a likely contributor (Turner et al., 2017).
44 Likewise, the large increase in CH_4 during 2020 has been attributed to decreases in OH resulting from
45 COVID-19 related changes in NO_x (NO_x = NO + NO_2) abundance (Laughner et al., 2021). Understanding
46 the non-linear chemistry of the drivers of OH and feedbacks among these species is therefore important

47 for characterizing past and present changes in the atmosphere as well as in projecting future climate
48 scenarios.
49
50   Chemistry-climate models (CCMs) with detailed chemical mechanisms are used to investigate the
51 complex, non-linear chemistry between these species and their impacts on the atmosphere (e.g.,Fiore et
52 al., 2006;Voulgarakis et al., 2015;Gaubert et al., 2017;Holmes, 2018). The utility of CCMs for this
53 purpose is limited, however, by the large computational expense of a CCM with a full representation of
54 $O_3 - NO_x - VOC$ (Ozone, NO_x, Volatile Organic Compound) chemistry combined with the need to model
55 over decadal time scales in order to let CH_4 perturbations fully evolve (Prather, 1996). Because of this
56 computational expense, simulations are necessarily limited to a short time frame, performed at coarse
57 horizontal resolutions, and/or limited in the number of sensitivity runs that can be performed (e.g., Fiore
58 et al., 2006;Holmes, 2018;Voulgarakis et al., 2015).
59
60   There are several alternatives (i.e., surrogate models) to running a full chemical mechanism that
61 capture some of the relationship between OH and trace gases, such as CO and CH_4, and are less
62 computationally expensive. Prescribed OH fields, either static or annually-varying, from a full chemistry
63 simulation or a climatology have been used for decades to simulate and understand trends in CO and
64 CH_4 in a computationally-efficient way (e.g., Saito et al., 2013;Wang et al., 2004). However, this method
65 linearizes CO and CH_4 chemistry with OH, preventing the simulation of nonlinear feedbacks in changes in
66 CO and CH_4 on OH, and thus could bias, for instance, interannual CH_4 changes (Chen and Prinn, 2006).
67
68   Spivakovsky et al. (1990) developed a parameterization of OH, later updated by Duncan et al. (2000),
69 that captures many of the nonlinear feedbacks between OH and tropospheric constituents (e.g., CO,
70 CH_4, VOCs) that have loss by reaction with OH as their primary sinks. The method to generate the
71 parameterization uses higher order polynomials with various chemical species, meteorological variables,
72 and variables related to solar irradiance as inputs. The degree of the nonlinear impacts simulated by the
73 parameterization of OH depends on the method used to populate these inputs. For instance, many of
74 the meteorological and solar irradiance variables may be provided by the model at run time. Though the
75 chemical variables that are not all simulated explicitly in the surrogate model may be provided as
76 climatological or monthly means from a full chemistry simulation. Whatever the approach, users can
77 strategically select input variables to meet the specific goals of an individual study, such as
78 understanding the causes of trends and variations in OH, CO, and CH_4 around the globe.
79
80   These parameterizations have been incorporated into CCMs to simulate the nonlinear feedbacks
81 between atmospheric CO and OH (Duncan et al., 2007a;Duncan and Logan, 2008;Strode et al., 2015),
82 methylchloroform and OH (Spivakovsky et al., 1990b), and the CO-CH_4-OH cycle (Elshorbany et al.,
83 2016). Through numerous sensitivity simulations, Duncan et al. (2007a) and Duncan and Logan (2008)
84 used a parameterization of OH in an atmospheric model of CO to elucidate the causes of trends and
85 interannual variations in CO from 1988-1997 on regional to global scales as well as those observed by
86 individual *in situ* monitors around the world. Strode et al. (2015) used the Efficient CH_4-CO-OH (ECCOH)
87 chemistry module (Elshorbany et al., 2016) implemented in the NASA Goddard Earth Observing System
88 (GEOS) general circulation model (GCM) to investigate the effects of different model biases in GEOS on
89 simulated OH. To do this, they performed multiple sensitivity simulations, adjusting tropospheric water
90 vapor, ozone, and NO_x to match satellite observations, to understand the impacts on OH and CH_4
91 lifetime. Similarly, Elshorbany et al. (2016) investigated the impacts of varying CH_4 and CO emissions on
92 the growth rate of atmospheric methane concentrations through multiple sensitivity runs.
93

94   A variety of machine learning techniques, such as neural networks (Nicely et al., 2017;Nicely et al.,
95   2020;Kelp et al., 2020), random forest regression (Keller and Evans, 2019), and gradient boosted
96   regression trees (GBRTs) (Ivatt and Evans, 2020;Stirnberg et al., 2020) show promise for use in new,
97   efficient methods to generate parameterizations of OH as these techniques have been successfully used
98   in atmospheric chemistry applications. In particular, GBRT models (Elith et al., 2008;Chen and Guestrin,
99   2016) use an ensemble of decision trees to predict the value of a target based on multiple inputs.
100   Decision trees are created sequentially, with each new tree minimizing a cost function based on the
101   results of the previous tree (Elith et al., 2008;Stirnberg et al., 2020). Unlike other machine learning
102   algorithms, regression tree methods have easily interpretable metrics that help highlight the influence
103   of the different input variables (Yan et al., 2016). These metrics can help further understanding of the
104   model behavior in ways other machine learning techniques cannot.
105
106   We present a new method for the efficient generation of a parameterization of OH using GBRTs and
107   a full chemistry simulation from a CCM, which serves as the training dataset. Our methodology allows
108   for the parameterization to be easily and rapidly regenerated in response to changes in, for instance, the
109   underlying model chemical mechanism (e.g., updates to the chemical rate constants or absorption cross-
110   sections) or model dynamics, which affect many of the variables that influence OH (e.g., Anderson et al.,
111   2021). Likewise, the parameterization can be modified to include new input variables. In Section 2, we
112   describe the new parameterization and the method used to develop it. In Section 3, we evaluate the
113   performance of the parameterization, and, in Section 4, we summarize the results and discuss applying
114   the parameterization to scientific research.
115
116   **2.0 Description of the Methodology to Generate the Parameterization of OH**
117   In this section, we describe the creation of the training dataset (Sect. 2.1) and the methodology used
118   to create the parameterization of OH (Sect. 2.2).
119
120   **2.1 Creation of the Training Dataset for the Parameterization**
121   We created the training dataset using output from the MERRA2 GMI simulation (https://acd-
122   ext.gsfc.nasa.gov/Projects/GEOSCCM/MERRA2GMI/). MERRA2 GMI is a 40-year (1980 – 2019)
123   GEOSCCM simulation run in replay mode (Orbe et al., 2017) with MERRA2 (Modern Era Retrospective
124   analysis for Research and Applications) meteorology (Gelaro et al., 2017) and the Global Modeling
125   Initiative (GMI) chemical mechanism (Duncan et al., 2007b;Strahan et al., 2007). Aerosols were
126   calculated with the Goddard Chemistry Aerosol Radiation and Transport (GOCART) module (Chin et al.,
127   2002;Colarco et al., 2010). The model was run at a resolution of c180 on the cubed sphere
128   (approximately 0.625° longitude by 0.5° latitude) with 72 vertical layers. In this analysis, we use only
129   tropospheric output at daily and monthly resolutions. The GMI chemical mechanism includes
130   approximately 120 species and 400 reactions, characterizing the photochemistry of the troposphere and
131   stratosphere. Further simulation details, including a description of the emissions, are available
132   elsewhere (Anderson et al., 2021;Strode et al., 2019).
133
134   We created a dataset of training targets, representing the full range of simulated OH values, for
135   each month using daily-averaged output from the MERRA2 GMI simulation. We sample on a monthly
136   basis instead of on a yearly basis to increase computational efficiency of the generation of the
137   parameterization. The spatiotemporal variability in the abundance and emissions of OH drivers on the
138   yearly scale would necessitate a far larger dataset and more complicated sampling procedures to ensure
139   representativeness of both OH and the input variables. As demonstrated in Section 3.0, the adopted
140   monthly approach accurately captures OH while limiting the size of the training dataset.
141

Geoscientific
Model Development
Discussions

142   For each day of a month, we divided all simulated tropospheric OH concentrations from the 40-year
143   simulation into 20 equally-sized percentile bins (i.e., $0 - 5^{th}$ percentile, $5^{th} - 10^{th}$ percentile, etc.). Then,
144   we randomly selected 200,000 values from each bin, resulting in 4,000,000 training targets for each day
145   of the month. We also included the maximum and minimum OH values of the entire dataset to
146   represent the full range of values. We then combined training targets to form one large dataset with
147   120,000,000 values (for a 30-day month), encompassing the full range of OH concentrations from each
148   day of the month. To limit the size of the training dataset, we then subsampled these targets, again
149   randomly selecting 200,000 values from equally-sized percentile bins of OH concentration. The
150   procedure resulted in a dataset with 4,000,000 training targets that span all days within a given month.
151   A schematic of this process is shown in Figure S1. We omitted data from 5 years (1985, 1995, 2005,
152   2015, 2016) from the training dataset for model evaluation and a case study as discussed below. We
153   also created a training dataset for monthly-averaged output, discussed in Sect. 4.0, using an analogous
154   process.

155
156   Then, we extracted training dataset inputs for the regression tree parameterization corresponding
157   to the OH targets from the MERRA2 GMI output. We list parameterization inputs in Table 1. The
158   parameterizations of Spivakovsky et al. (2000), Duncan et al. (2007a) and Elshorbany et al. (2016), along
159   with expert knowledge of OH chemistry, informed our choice of inputs. As discussed in the next section,
160   NO_2 serves as a sufficient proxy for the impact of NO_x and NO_y on OH, so NO_2 is the only reactive
161   nitrogen species included as an input parameter. For both ice and water cloud as well as aerosol optical
162   depths, we include the optical depth above and below each datapoint as inputs. We use aerosol optical
163   depth (AOD) at 550 nm, calculated from the GOCART aerosol module. We took all 27 inputs from the
164   MERRA2 GMI simulation except UV albedo, which we took from the Ozone Monitoring Instrument (OMI)
165   climatology of Kleipool et al. (2008).

166
167   *Table 1:* Inputs to the machine learning parameterization of OH

Chemical Inputs			Meteorological/Radiative Inputs		
NO_2	Formaldehyde (HCHO)	Isoprene (C_5H_8)	Temperature	Pressure	Water Vapor
CO	Hydrogen peroxide (H_2O_2)	Propene (C_3H_6)	Cloud Fraction	Stratospheric O_3 Column	Latitude
CH_4	Methyl hydroperoxide (CH_3OOH; MHP)	Propane (C_3H_8)	Ice Cloud Optical Depth (above & below)	Aerosol Optical Depth (above & below)	Water Cloud Optical Depth (above & below)
		C4 & C5 Alkanes			
O_3	Acetone (CH_3COCH_3)	Isoprene (C_5H_8)	UV Albedo	Solar Zenith Angle	

168
169   **2.2 Creation of the GBRT Parameterization**
170   We used the XGBoost package (Chen and Guestrin, 2016) in Python to create 12
171   parameterizations of OH (one for each month), training the parameterizations on the MERRA2 GMI
172   datasets described in Sect. 2.1. We refer to the GBRT models as parameterizations to prevent confusion
173   when referring to 3-dimensional models. For each month, we used 80% of the dataset (3.2 million
174   datapoints) for model training and the balance for model validation. In addition, as outlined in-depth in
175   Sections 2.1 and 3.0, the model was also evaluated on 5 years of data not included in the model training.
176   Increasing the size of the training dataset did not improve model performance but did increase model
177   training time, so the training set was restricted to a size that represented the full ranges of OH values.

178
179   To maximize parameterization performance while also balancing the potential of overfitting, we
180   tuned hyperparameters, including the learning rate, the maximum tree depth, and the number of trees.

181 We chose hyperparameter values that minimized the parameterization normalized root mean square
182 error (NRMSE) (Eq. 1.) of the training dataset. In Eq. 1, N is the number of samples, y is the MERRA2 GMI
183 OH, $\hat{y}$ is the parameterized OH, and IQR is the interquartile range of the dataset. We set the learning
184 rate, which controls the magnitude of change when adding a new tree, to 0.1, while we varied the
185 maximum tree depth and number of trees from 6 to 22 and from 10 to 150, respectively. For both
186 maximum tree depth and number of trees, NRMSE initially dropped significantly with increasing value,
187 representing sharp improvement in parameterization performance. NRMSE values eventually
188 plateaued, increasing parameterization runtime without noticeably improving performance. A
189 combination of a maximum tree depth of 18 and 100 trees balanced performance with model training
190 and run time.

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}}{IQR} \qquad (1)$$

191

192 We also evaluated inputs into the parameterization to ensure that each did not lead to decreased
193 performance, finding that no single variable dominates model performance and no variable reduces
194 performance. We retrained the parameterization 27 times for July, removing each input successively, to
195 determine its impact on the NRMSE. When we applied the resultant models to the July 2005 validation
196 dataset, the percentage change in the NRMSE generally increased by less than 1%. The small differences
197 in NRMSE indicate that there are likely variables that provide duplicate information to the
198 parameterization. As will be discussed in Sect. 3.2, however, the relative importance of inputs varies by
199 month, and some variables, though not important on average, have a large influence in specific chemical
200 environments. Because of these factors and a desire to use a consistent set of input variables across all
201 months, we did not remove any inputs from the parameterization as a result of this analysis.

202

203 Finally, we omit NO_X and NO_y as parameterization inputs because we find that NO_2 is sufficient as an
204 input to capture the impact of reactive nitrogen on OH in the parameterization. Because of the
205 importance of NO_X in OH production (Spivakovsky et al., 2000;Anderson et al., 2021), we tested
206 performance by substituting different reactive nitrogen species for NO_2 as inputs to the
207 parameterization. We trained three additional parameterizations, including ones with NO_X, NO_y (NO_y =
208 $NO + NO_2 + PAN + 2N_2O_5 + HNO_3$ + alkyl nitrates), and the individual NO_y species. Parameterization
209 performance did not improve noticeably with the inclusion of NO_X or the individual NO_y species.
210 Including NO_y as a group actually decreased performance.

211

212 **3.0 Evaluation of the machine learning parameterization of OH**
213 We now evaluate the performance of the parameterization of OH. In Section 3.1, we compare the
214 OH calculated with the parameterization to that from the MERRA2 GMI simulation, showing agreement
215 in both local OH concentrations as well as in global metrics, such as CH_4 lifetime (τ_{CH4}). In Section 3.2, we
216 investigate the parameterization Gain feature importance and SHapley Additive Explanations (SHAP)
217 values to understand the relative contributions of inputs to parameterization performance and to
218 demonstrate that, even though there are no physics- or chemistry-based constraints, parameterization
219 behavior comports with our understanding of OH chemistry. We explore the ability of the
220 parameterization to predict OH in response to strong deviations in its drivers from the climatological
221 mean in Section 3.3, by examining two El Niño events.

222

223 **3.1   Ability of the parameterization to reproduce modeled OH and global OH metrics**
224 The parameterization is able to reproduce both the spatial distribution and concentration of
225 daily-averaged OH, although with noticeable errors at high latitudes in the winter hemisphere, which is
226 unimportant as OH is seasonally low. Figure 1a shows the fractional difference between OH calculated
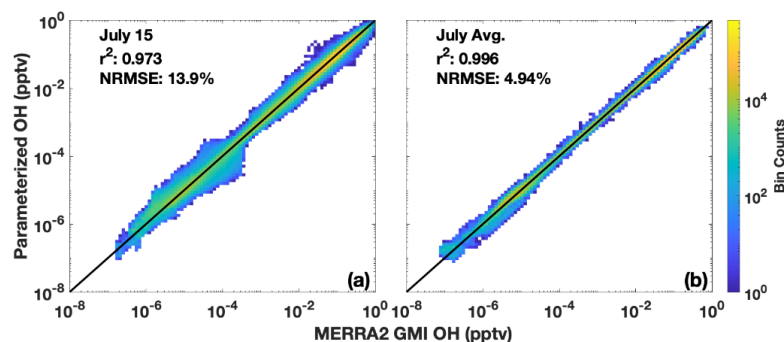
227 with the parameterization and OH from the MERRA2 GMI simulation for July 15, 2005, a date omitted
228 from the training dataset. The parameterized and MERRA2 GMI OH fields are shown in Figure S2. The
229 OH in Figure 1 has been averaged over the lower free troposphere (LFT), defined as pressures between
230 the top of the planetary boundary layer (PBL) and 500 hPa. Agreement is similar throughout the
231 troposphere, but we highlight this region because of its importance for CH_4 and CO loss (Spivakovsky et
232 al., 2000). For July 15, there are notable regions of bias, particularly poleward of 30° S, that seem to
233 reflect synoptic scale meteorological patterns, and where OH is low (Fig. S2). This results in a NRMSE for
234 the entire troposphere of 13.9% (Fig. 2a). At higher concentrations, the correlation between the
235 MERRA2 GMI simulation and the parameterized OH is much tighter than at lower concentrations,
236 although the highest density at all concentrations is centered around the 1:1 line. Because the CO and
237 CH_4 lifetimes are much longer than one day, the accuracy of the parameterization on monthly
238 timescales is more relevant to the applications of the parameterization than an individual day.
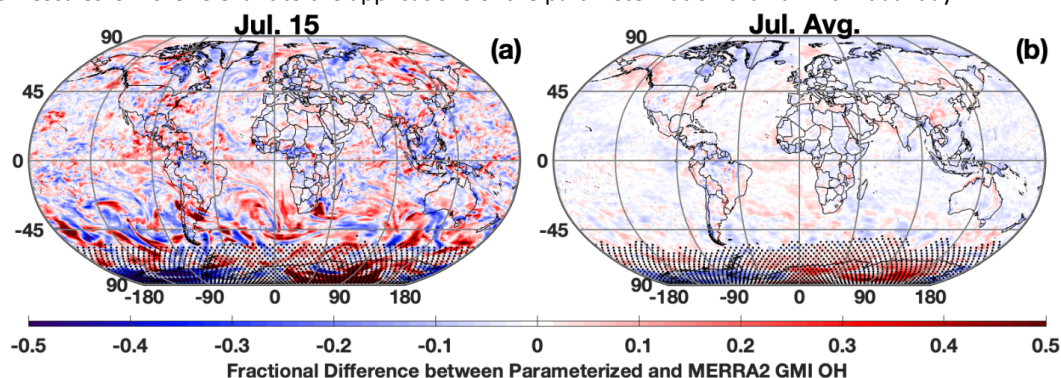
239
240 **Figure** 1: Fractional difference between the parameterized and MERRA2 GMI OH averaged over the LFT for July 15,
241 2005 (a) and averaged across all days for July 2005 (b). Regions with low OH, defined as a mixing ratio of less than
242 0.005 pptv, are indicated with stippling.

243   Averaged on a monthly basis, the parameterization can reproduce the global OH distribution
244 with little error (Fig. 1-2). For July 2005, the percentage difference between the parameterized OH and
245 output from the MERRA2 GMI simulation in the LFT (Fig. 1b) and throughout the troposphere (Fig. S3) is
246 generally within 10%, outside of the Southern Hemispheric high latitudes, where it is polar night and OH
247 concentrations are negligible. The random errors evident in the daily data in Figure 1a average out on
248 the monthly timescale, resulting in a tight correlation (r^2 = 0.996) and a NRMSE of 4.94% for all
249 tropospheric values (Fig. 2b). Similar results are found for the July model when applied to other years
250 (Table S1) and for parameterizations developed for other months (Fig. S3 and S4). Averaging the daily
251 output over the monthly period yields a better NRMSE than for a parameterization trained exclusively
252 on monthly-averaged data (NRMSE = 7%) and over climatology (NRMSE = 11%), defined as the mean OH
253 from the MERRA2 GMI simulation averaged over 1980 to 2019.
254

Geoscientific
Model Development
Discussions

Figure 2: Scatter density plot of tropospheric OH from the MERRA2 GMI simulation plotted against OH calculated by the parameterization for July 15, 2005 (a) and averaged across all July days (b) for 2005. Colors indicate the number of data points in each bin. The r^2 of a linear least squares regression and the NRMSE are also indicated.

In regions where global CO and CH_4 loss are most important, parameterization biases and errors are low. For CO and CH_4, tropospheric loss to OH maximizes in the LFT in the 0 - 30° latitude band of the summer hemisphere with near negligible loss in the winter hemisphere polar region (Fig. 3). The comparatively large over- and underestimates over Antarctica evident in Figure 1 are irrelevant to the OH/CO/CH_4 cycle because of the low loss rate in this region. In contrast, in regions where CO and CH_4 loss maximize, the parameterization is biased low by only -0.3 to -1.4%. The normalized absolute error varies between 2.2% and 4.6% in the tropics and Northern Hemispheric mid-latitudes for all tropospheric layers (MFT: pressures between 500 and 300 hPa, UFT: pressures between 300 hPa and the tropopause). Results are similar for other months.
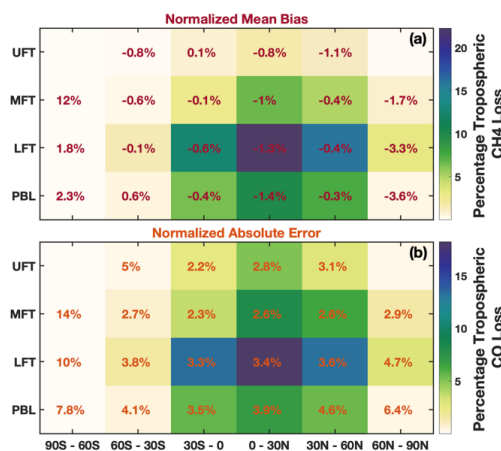

Figure 3: (a) Percentage of total tropospheric CH_4 lost to reaction with OH for July 2005 averaged over 30° zonal mean bins and 4 atmospheric layers is shown by the background colors. The percentage loss values account for the mass of each region relative to the total atmospheric mass. Numbers in red indicate the normalized mean bias of the parameterization with respect to the MERRA2 GMI simulation. Statistics for the polar UFT are omitted because low tropopause heights limit the data amount in these regions. (b) Same as (a) except for tropospheric CO loss and the normalized absolute error indicated in orange.

The parameterization is also able to reproduce global mean metrics of OH, such as τ_{CH4}, within 1.3% on average. For each month of 2005, we calculated the global, mean mass-weighted tropospheric

279   OH as described in Lawrence et al. (2001) and the mean tropospheric τ_{CH4} with respect to OH as
280   described in Nicely et al. (2020) for the MERRA2 GMI simulation, the parameterization, and the
281   climatological mean, defined as the average value from the MERRA2 GMI simulation between 1980 and
282   2019. Results for τ_{CH4} are shown in Figure 4, and for mass-weighted OH in Figure S5. The
283   parameterization captures the seasonality of the τ_{CH4}, with a minimum in boreal summer and a
284   maximum in boreal winter. Agreement varies slightly by month, differing by only 0.8% in January and up
285   to 2.5% in August. These values are reasonable and much smaller than the inter-model variability often
286   seen in model intercomparison projects (e.g., Nicely et al., 2020;Voulgarakis et al., 2013). Similar results
287   are found for the global, mean mass-weighted OH. The Northern Hemispheric/Southern Hemispheric
288   OH ratio (Fig. S5) also generally agrees within 0.5% for all months, again with the exception of August.
289   The comparatively weaker performance for August is limited to 2005, however, as performance of the
290   August parameterization in the other validation years (1985, 1995, and 2015) is closer to the 1%
291   difference shown by the parameterizations for the other months. The parameterizations present a
292   significant improvement over the climatological mean, which for 2005, consistently underestimates τ_{CH4}
293   for all months and by up to 6% in March.
294
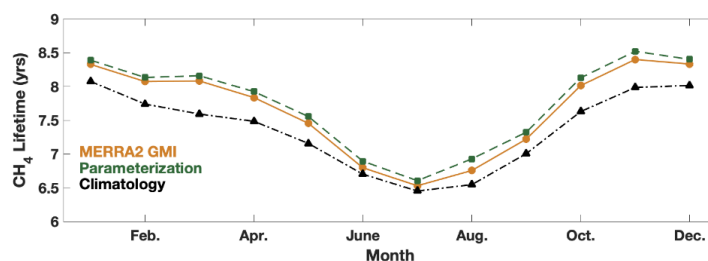
295
296   *Figure* 4: Global mean methane lifetime with respect to tropospheric OH from the parameterization (green squares)
297   and MERRA2 GMI (orange circles) for 2005 and the climatological average (black triangles) calculated from MERRA2
298   GMI for 1980 - 2019.

3.2   Understanding the relative importance of input parameters

299
300   While we have demonstrated that the parameterization is able to reproduce OH accurately, it is
301   also instructive to understand the relative importance the parameterization places on each of the
302   inputs. Although this parameterization is not process-based, understanding how the parameterization
303   responds to different inputs can help determine if the regression tree is responding in a way consistent
304   with current understanding of OH chemistry, although there are limitations to the information that can
305   be gleaned from these metrics. We evaluate the regression tree parameterization using two metrics,
306   the Gain feature importance as output by the XGBoost package, and SHAP values.
307
3.2.1   Investigating the Gain feature importance
308
309   The Gain feature importance (Chen and Guestrin, 2016) is a measure of the improvement in model
310   accuracy achieved from adding branches in the model corresponding to a specific input variable. The
311   Gain value therefore indicates the relative importance of each input for the model as a whole but not
312   for individual datapoints. The Gain values for each input for the January and July models are shown in
313   Figure 5. While there are differences between the two months, several features are similar. Variables
314   that indicate geographic location (e.g., SZA, latitude, and pressure) and chemical species that have
315   previously shown to be dominant drivers of OH variability (e.g., NO_2, O_3, CO) and/or OH proxies (e.g.,
316   HCHO) (Wolfe et al., 2019;Murray et al., 2021) have some of the highest Gain values. As we show
317   below, caution should be used in extrapolating results from the Gain values to a process-based
318   understanding of OH without prior knowledge of its response to chemical and dynamical drivers.

319
320   The relative importance of variables that indicate location is consistent with OH chemistry and
321   previous parameterization studies. Primary OH production is driven by the photolysis of O_3 followed by
322   the subsequent reaction of the O^1D radical, produced from that photolysis, with water vapor (e.g.,
323   Spivakovsky et al., 2000). Thus, the OH distribution is strongly dependent on SZA, latitude, and
324   pressure. This is consistent with the parameterization, where SZA and latitude have the highest Gain
325   values for both months examined here, as well as with the results of Duncan et al. (2000), who
326   highlighted the importance of latitude in their parameterization.
327
328   Similarly, the chemical species that are most important for controlling OH distribution on the global
329   scale also tend to have higher Gain values. As discussed above, O_3 and NO_x chemistry is instrumental in
330   controlling primary and secondary OH production on global scales (e.g., Spivakovsky et al.,
331   2000;Anderson et al., 2021), consistent with their comparatively high Gain values. HCHO, an oxidation
332   product of the reaction of OH with many VOCs, has been found to be a suitable proxy for OH in the
333   remote atmosphere (Wolfe et al., 2019), consistent with its relative importance in both the July and
334   January models.
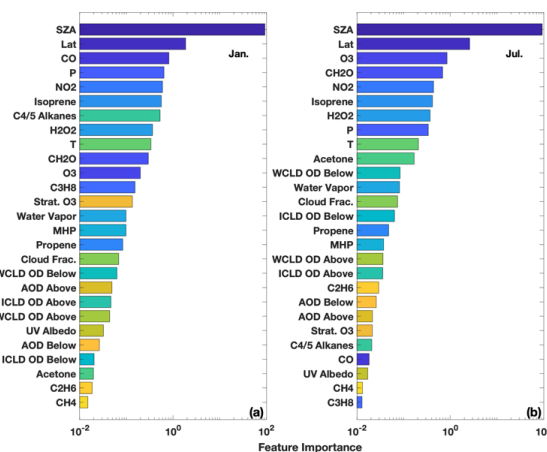
335
336   **Figure 5:** The feature importance (gains) of the January (a) and July (b) parameterizations as calculated by XGBoost.
337   Variables are sorted by their relative importance. WCLD = Water cloud; ICLD = Ice Cloud; OD = Optical Depth.
338   "Above" and "below" for the optical depth variables indicate the optical depth above and below a particular model
339   grid box.
340
341   The relative importance of global OH sinks in the parameterization, however, demonstrates the
342   limitations of using the Gains values to interpret the regression tree model in a process-based way. CO,
343   the dominant OH sink on a global scale (Spivakovsky et al., 2000), is the most important chemical input
344   for the January parameterization, although it is relatively unimportant in the parameterizations for all
345   other months. While tropical CO variability in MERRA2 GMI and biomass burning emissions (Duncan,
346   2003b) are larger in boreal winter than July, there is no process-based explanation for why CO should be
347   different in January from December or February. Differences in the relative importance of CO between
348   the two months does not imply that OH sensitivity to CO in MERRA2 GMI or the atmosphere varies in
349   the same manner. Instead, the differences simply indicate that the parameterization algorithm finds CO
350   to be more useful in predicting OH in January than July. Similarly, CH_4, the second most important OH
351   sink on the global scale, has low Gain values, suggesting it has little impact on model performance. This
352   is likely because, in the MERRA2 GMI simulation, CH_4 concentrations vary little within a given latitude

353   band due to CH$_4$ surface concentrations being set as a boundary condition. The methane distribution
354   therefore provides little additional information beyond that already contained in the variables that
355   indicate location.
356
357   **3.2.2. Investigating parameterization SHAP values**
358   While the Gain values indicate the relative importance of species in the parameterization and can
359   provide some information as to whether the parameterization behaves in a manner consistent with our
360   understanding of OH chemistry, the metric only provides information about the dataset as a whole.
361   Gain values can therefore obscure the importance of variables that only strongly impact the
362   parameterization for a small subset of the data. To better understand the relative importance of
363   variables as well as the spatial variability in that importance, we also calculate the SHAP values
364   (Lundberg and Lee, 2017), which provide information on the relative importance of each datapoint input
365   into the model.
366
367   In the context of machine learning, Shapley values, an idea first developed for game theory
368   (Shapley, 1953), indicate the average contribution of an individual model input to all possible
369   combinations of inputs. For example, to calculate the Shapley value of the variable X for a hypothetical
370   machine learning model with three input variables X, Y, and Z, first a model would be trained with all
371   three variables. A new model would then be retrained, omitting X, and the difference between the two
372   models would be calculated to determine the contribution of X. This process would then be repeated
373   with different permutations of input variables (e.g., X and Y, X and Z) to determine the contribution of X
374   in those instances. The final Shapley value is the average of the contribution from these different
375   models. SHAP values use the same concept but in a manner that reduces the computation time
376   (Lundberg and Lee, 2017), as this process could become prohibitive for a model, such as the
377   parameterization of OH, that contains 27 inputs.
378
379   We calculate SHAP values using the TreeExplainer API of the SHAP package available for Python.
380   One limitation of the algorithm used to calculate SHAP values is that it is too computationally expensive
381   to calculate the SHAP values for the tuned regression tree model. Computational time to calculate SHAP
382   values for the troposphere at the native model resolution for one day is several months. Maximizing
383   computational speed by degrading the model resolution and running the SHAP package with GPUs,
384   would take approximately 4 days for one model day. Calculating SHAP values for a model with default
385   model hyperparameters, however, takes minutes. This is due to the large reduction in the number of
386   trees (100 to 10) and the maximum tree depth (18 to 6) in the parameterization, which significantly
387   speeds up the creation of new regression trees needed in the SHAP value calculation.
388
389   We first evaluate the feasibility of using the SHAP values for the untuned model to explain the
390   parameterization behavior. To test this, we created a subset of 5000 OH values from the
391   parameterization training dataset that spanned the full range of OH concentrations. We then calculated
392   the SHAP values for the July parameterization with tuned hyperparameters as well as for a July
393   parameterization using the default XGBoost hyperparameters. For the variables found to be most
394   important for the parameterization (e.g., SZA, NO$_2$, O$_3$, isoprene, HCHO, latitude), there are strong
395   correlations (r^2 of 0.97 or higher) for the SHAP values between the tuned and untuned model, resulting
396   in similar spatial distributions, although there are differences in the magnitude. For other variables,
397   correlation is much weaker, although the relative importance of variables is similar for the tuned and
398   untuned parameterizations. We therefore restrict our analysis primarily to variables that have highly
399   correlated SHAP values between the tuned and untuned models and discussion to the relative
400   importance of the different variables.

401
402   The distribution of SHAP values for the training dataset for July demonstrates the importance of
403   including each of the variables as inputs to the parameterization as well as the large variability in their
404   relative importance. Figure 6 shows the distribution of the SHAP values for each input parameter of the
405   approximately 3.2 million datapoints used to train the July parameterization. The median SHAP values
406   (Fig. 6) show similar ordering as the Gains feature importance (Fig. 5), with variables that indicate
407   location as well as O_3 and NO_2 being the most important in both cases. When looking at the distribution
408   of the SHAP values, however, it becomes apparent that variables that appear to be unimportant for
409   parameterization performance in the mean (e.g., propene and CH_4) can have large importance for
410   individual datapoints. For example, although propene can be locally important for OH chemistry, due to
411   its reactivity, concentrations in the remote atmosphere are low, making the species seem unimportant
412   in the aggregate. Similar results are found for the January parameterization (Fig. S6). As discussed in
413   Section 2.2, the SHAP values provide a rationale for including each of these species in the
414   parameterization.
415
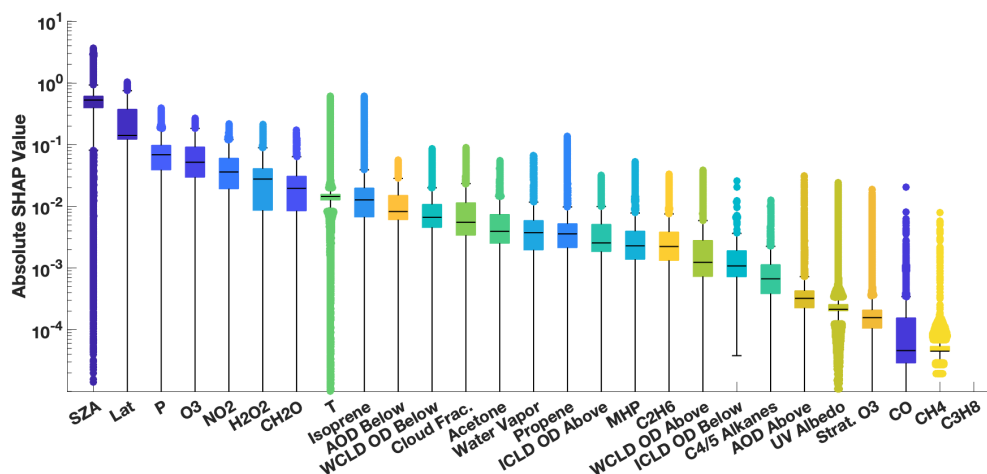
416
417   *Figure 6:* Distribution of the absolute SHAP value for each parameterization input for July from an untuned version
418   of the parameterization of OH. Input parameters are sorted by order of relative importance. The median is indicated
419   with the black line, edges of the box represent the interquartile range, and whiskers represent the 5th and 95th
420   percentile. Values outside this range are indicated with circles. Note that the SHAP value for propane is zero,
421   indicating that it is not used by the untuned parameterization.

422
423   The SHAP values also demonstrate the spatial distribution of the relative importance of the
424   different chemical OH drivers. Figure 7 shows the relative importance of NO_2, as determined by the
425   SHAP values for the untuned parameterization, for both the zonal mean and the LFT. Note that the
426   untuned parameterization has large errors for low OH concentrations, so SHAP values poleward of 45 °S
427   should be viewed as more uncertain than those elsewhere. In both the horizontal and vertical, the SHAP
428   values demonstrate that the parameterization captures the spatial pattern of the relative importance of
429   NO_X for OH production. The spatial pattern in Figure7a, for example, has the highest contribution of
430   NO_2 to the total SHAP value in the tropical UFT and in the northern hemisphere midlatitudes. This is
431   nearly identical to the spatial pattern of the relative contribution of the $NO + HO_2$ reaction to overall OH
432   production in the MERRA2 GMI simulation (Anderson et al., 2021). Likewise, in the LFT, the contribution
433   from NO_2 maximizes over continental regions with high emission and minimizes over the remote oceans.

434 The spatial pattern of SHAP values of isoprene also agree with OH chemistry, maximizing in regions of
435 strong biogenic emissions and minimizing over oceans (Fig. S7). These SHAP values demonstrate that,
436 although the parameterization is not process-based, its behavior at least partially comports with our
437 understanding of OH chemistry.
438
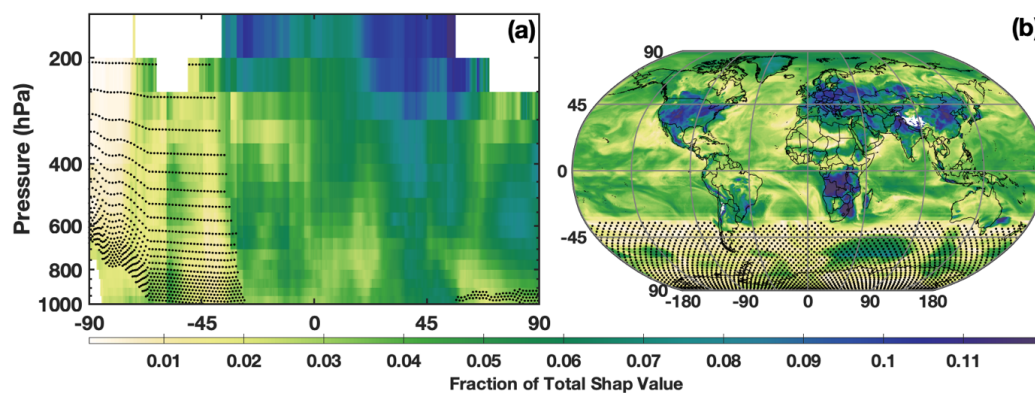
439
440 **Figure 7:** The fraction of the contribution of the NO_2 SHAP value to the sum of the absolute SHAP value of all inputs
441 in July is shown for the zonal mean (a) and the LFT (b). Regions where mean OH mixing ratios are below 0.03 pptv,
442 the point below which the untuned parameterization is unable to reasonably predict OH, are indicated by the
443 *stippling*.
444

3.3 Case Study: Testing the parameterization response to the 2016 El Niño Event

446 While we have demonstrated that the parameterization can satisfactorily reproduce OH during all
447 months of 2005, we now investigate its ability to capture OH accurately during the 2016 El Niño event,
448 which we excluded from the training dataset. Evaluating how the parameterization responds to
449 deviations from the climatological mean of the inputs during a large-scale event, such as El Niño, is a
450 strong test of its ability to predict extremes in OH as well as to respond to deviations from the
451 climatological mean of the parameterization inputs. The response of the parameterized OH to these
452 extremes in inputs will also provide a further test of the ability of the parameterization to behave in a
453 process-based way.
454
455 El Niño events lead to dramatic changes in the concentrations and distributions of many OH
456 drivers, including O_3 (Oman et al., 2011;Oman et al., 2013), CO (Duncan, 2003a;Rowlinson et al., 2019),
457 NO_X (Murray et al., 2013;Murray et al., 2014) and water vapor (Shi et al.;Anderson et al., 2021). As such,
458 the El Niño Southern Oscillation (ENSO) is the dominant mode of OH variability throughout much of the
459 troposphere and can result in localized changes in OH on the order of 40 – 50% from the climatological
460 mean (Anderson et al., 2021;Turner et al., 2018). Changes in secondary production from NO_X in the UFT
461 and changes in primary production from O_3 in the PBL and LFT drive the ENSO related variability of OH
462 (Anderson et al., 2021). Methane emissions also vary strongly with the ENSO phase (Zhang et al.,
463 2018;Worden et al., 2013). In order to capture the $OH/CH_4/CO$ system correctly, the parameterization
464 must be able to capture ENSO-related OH variability.
465
466 Here, we investigate the ability of the parameterization to capture OH during the El Niño events of
467 1997/98 and 2015/16, two of the largest such events during the period of the MERRA2 GMI simulation
468 according to the Multivariate ENSO Index (Wolter, 2011). The 1997/98 event, which was included in the
469 training dataset, was a prototypical example of an Eastern Pacific (EP) El Niño, characterized by sea

470 surface temperature (SST) anomalies extending to the coast of South America. In contrast, the 2015/16
471 event was a blend of an EP and a Central Pacific (CP) El Niño, also known as El Niño Modoki, where SST
472 anomalies extend only to the international dateline (Paek et al., 2017).   These different "flavors" of El
473 Niño affect atmospheric distributions of OH drivers, such as water vapor (Du et al., 2021), in different
474 ways, suggesting different impacts on OH. While we did include other blended El Niños (1986/87,
475 1987/88, and 1991/92) (Kug et al., 2009) in the training dataset, each had responses in the atmospheric
476 distribution of OH and its drivers distinct from the 2015/16 event. We focus our investigation on
477 January and the MFT, the time and location of the strongest correlation between ENSO and OH
478 (Anderson et al., 2021) in the MERRA2 GMI simulation. We also restrict the analysis to the OH
479 precursors, NO_2, CO, and O_3, as they have both a strong influence in the variability of ENSO-related OH
480 production/loss and have comparatively large feature importance and SHAP values in the January
481 parameterization.
482
483   For both the 1997/98 and 2015/16 El Niño events, each OH driver examined deviates strongly
484 from the climatological mean, defined as the average value from the MERRA2 GMI simulation over all
485 Januarys from 1980 – 2019. Both O_3 and NO_2 have pronounced positive anomalies over the western
486 Pacific and maritime continent and negative anomalies over the eastern Pacific (Fig. 8) that extend
487 throughout much of the free troposphere (Fig. S8), likely associated with changes in the Walker
488 Circulation as described in Oman et al. (2011).   The positive anomalies over Indonesia show a distinct
489 westward shift during the 1997/98 event as compared to 2015/16, highlighting the variability in the
490 effects of ENSO on emissions and transport. CO has a large positive anomaly over much of the globe,
491 attributable to the increases in biomass burning during El Niño events (e.g., Duncan, 2003a). As with O_3
492 and NO_2, there are large differences in the spatial pattern of the CO anomalies between the two events,
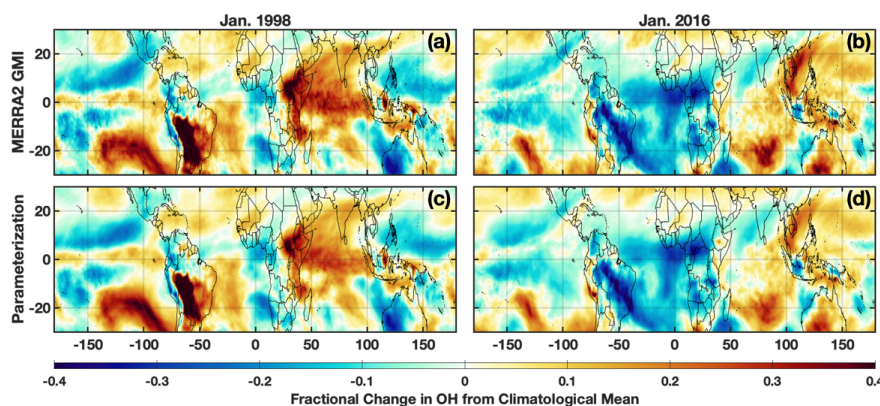493 particularly over the Indian Ocean, central Africa, and South America.
494
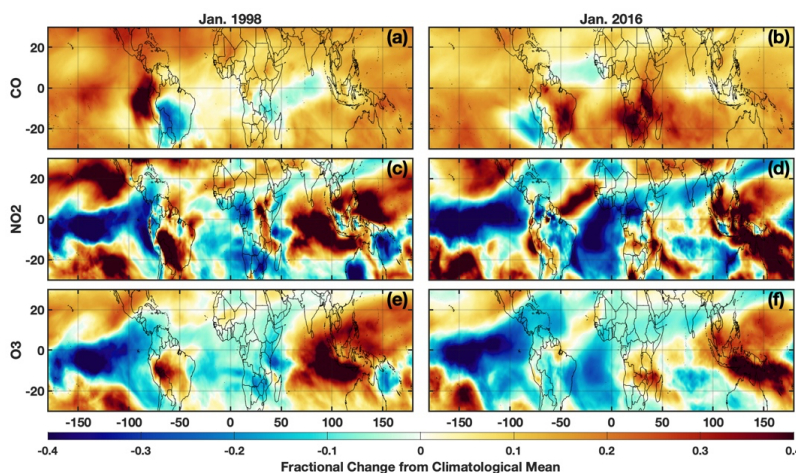
495
496 Figure 8: Fractional difference of the indicated variable between January 1998 (left) and the climatological mean
497 (1980 – 2019) calculated from the MERRA2 GMI simulation for the MFT. The same values but for January 2016 are
498 indicated on the right. Species shown are CO (a,b), NO_2 (c,d) and O_3 (e,f).
499
500   The differences in anomalies of the OH drivers between the 1997/98 and 2015/16 El Niño events
501 lead to distinct anomaly patterns in OH itself. During the 1997/98 event, in the MFT, there are
502 noticeable positive anomalies in OH over much of the Indian Ocean basin, the southeastern Pacific,
503 South America, and the western Atlantic Ocean (Fig. 9). During 2015/16, the positive anomalies were

504   more limited and most noticeable in the tropical western Pacific Ocean and southern Indian Ocean.
505   Along the equator, the positive anomalies extended throughout a larger portion of the troposphere
506   during January 1998 than 2016. Both the parameterization inputs and the OH itself respond strongly
507   and in different ways to each El Niño event, providing a strong test to determine the robustness of the
508   parameterization.
509   The parameterization reproduces the ENSO-related OH anomalies for both El Niño events with
510   remarkable fidelity. We ran the parameterization for all Januarys from 1980 to 2016 to calculate a
511   climatology and calculated the deviations for 1998 and 2016 from that value. For both events, the
512   parameterization accurately captures the location and magnitude, as well as the spatial pattern, of the
513   OH anomalies with a few minor exceptions in the horizontal and vertical (Figs. 9 and S8). Correlation
514   between the MERRA2 GMI and parameterized anomalies plotted in Figure 9 has an r^2 of 0.93 or higher
515   for both years. The parameterization is therefore capable of reproducing both the climatological mean
516   of OH as well as large deviations in the mean in response to strong climatological deviations in the
517   model inputs, even for years excluded from the training dataset.
518

Figure 9: Fractional difference of the indicated variable between January 1998 (left) and the climatological mean (1980 – 2019) calculated from the MERRA2 GMI simulation averaged over the MFT. The same values but for January 2016 are indicated on the right. Species shown are OH from the MERRA2 GMI simulation (a,b) and OH calculated by the parameterization (c,d).

Section 4.0: Discussion and Summary: The parameterization of OH as a tool for scientific research

525   In this manuscript, we present a new methodology to generate a parameterization of OH that, as
526   compared to previous methods (e.g., Spivakovsky et al., 1990; Duncan et al., 2000), is efficient and easy
527   to use, allowing the user to rapidly update the parameterization of OH as necessary. The new method
528   uses GBRTs and a full-chemistry simulation from a CCM as the training data to generate the
529   parameterization of OH with a high degree of accuracy relative to the full-chemistry simulation.
530   As discussed in Section 1, previous incarnations of the parameterization of OH have been an
531   important tool in CCMs for decades to efficiently simulate OH, CO, and CH_4 and the feedbacks between
532   them. Researchers have successfully used this framework to perform sensitivity simulations through
533   manipulation of the input parameters (i.e., chemical, meteorological, and solar irradiance variables) to
534   the parameterization of OH, as well as emissions and dynamics, to deconvolve the causes of local to
535   global trends and variations in OH, CO, and CH_4. In addition, sensitivity simulations have been valuable
536   for the selection of computationally-expensive, full-chemistry simulations to target for further analysis
537   in a given research study.

538   Despite the potential of a parameterization of OH, one must take care to ensure that a particular
539 research application of this tool is within its intended scope. In the next sections, we discuss some of
540 the strengths and limitations of the parameterization of OH as a tool for scientific research.
541
542 **Section 4.1: Excursions from training data**
543   The parameterization of OH accurately reproduces OH from the full-chemistry simulation on
544 which it was trained, but it may not produce the desired accuracy for a given time period or scenario
545 outside the range represented in the training data. Of course, the degree of degradation in accuracy
546 depends on how far inputs exceed the ranges of the training dataset. In addition, the parameterization
547 of OH generated using inputs from one model may not be portable to another model or even a different
548 configuration of the same model as shown below. The simulated relationships between OH and the
549 input parameters may differ because of inter-model variations in the chemical, dynamical, and radiative
550 schemes. Ultimately, it is up to the user to determine an acceptable level of degradation for a specific
551 research topic. In this section, we give an example of the degree of degradation in accuracy for a
552 parameterization of OH that uses 1) a different time period for the same model and 2) input parameters
553 from a different model.
554
555 **Section 4.1.1: Input parameters from a different time period for the same model setup**
556   Analysis of a separate model simulation, the Chemistry Climate Model Initiative (CCMI) GEOS
557 simulation (Morgenstern et al., 2017), highlights possible limitations in extending the parameterization
558 to years outside of those on which it was trained, particularly if there is a strong trend in one of the
559 inputs. The GEOS CCMI simulation has unconstrained meteorology, spans 1960 – 2100, and has a
560 horizontal resolution of 2.0° latitude × 2.5° longitude. Emissions for precursors of tropospheric O_3 and
561 aerosols are from the RCP6.0 scenario. We trained two new parameterizations on the CCMI dataset,
562 denoted CCMI2019 and CCMI2060, using data from 1980 – 2019 and 1980 – 2061, respectively. We
563 used the same methodology to create the training datasets as for the MERRA2 GMI parameterization.
564 CCMI output are only available at monthly resolution, so we trained the CCMI parameterizations on
565 monthly, instead of daily, averaged values. Every 10th year, staring in 2000, was omitted from the
566 training dataset for validation.
567   While the CCMI2019 parameterization performed similarly to the MERRA2 GMI
568 parameterization for years included in the training dataset, performance degraded significantly for years
569 beyond 2019. The CCMI2019 parameterization captured the τ_{CH4} for 2000 and 2010 within 1% (Fig. 10,
570 red line) and the NRMSE within 5% (not shown). When we applied the parameterization to years
571 outside of the training window, however, performance degraded quickly and, by 2060, underestimated
572 τ_{CH4} by about 4%. The CCMI2060 parameterization, on the other hand, captures the τ_{CH4} lifetime within
573 0.5% for all validation years.
574   The reason for this performance degradation is likely due to input parameters that extend
575 beyond the range used in the training dataset. For example, there is a strong positive trend in the
576 stratospheric O_3 column (Fig. 10), which results in chemical environments in 2060 that did not exist in
577 the 1980 – 2019 training dataset. Other variables with strong trends, such as CH_4 and temperature, as
578 well as different large-scale dynamical patterns, could also decrease parameterization performance.
579 These results strongly suggest caution when applying the parameterization to future scenarios outside
580 of the training window. For example, experiments investigating relative changes in a species (e.g.,
581 understanding how a 10% decrease in OH would affect tropospheric CH_4 abundance) could likely be
582 successfully implemented, while those trying to understand absolute differences in CH_4 would require
583 significant care to ensure valid results.
584
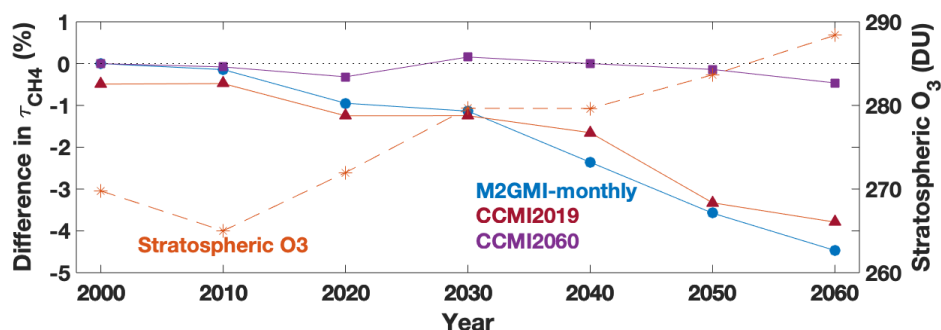
585
586 **Figure 10:** *Time series showing the percent difference between τ_{CH4} calculated from the CCMI simulation and from*
587 *three separate parameterizations: one trained on MERRA2 GMI output from 1980 – 2019 (blue circle), one trained*
588 *on CCMI output spanning 1980 – 2019 (red triangle), and one trained on CCMI output spanning 1980 – 2060 (purple*
589 *square). The stratospheric O_3 column (orange star) from the CCMI simulation averaged over 30 S to 60 N, the region*
590 *where tropospheric CH_4 loss to OH maximizes (Fig. 3), is also shown. All data are for July.*

591 **Section 4.1.2: Input parameters from a different model setup**
592   Similar to applying the parameterization outside of the timeframe on which it was trained,
593 applying the parameterization to a different model setup also warrants caution, as the differences can
594 result in new chemical environments on which the parameterization was not trained. We now discuss
595 parameterization performance when output from the CCMI simulation discussed in Section 4.1.1 is input
596 into the MERRA2 GMI-trained parameterization. Despite both simulations being from the GEOS
597 framework, the CCMI simulation setup differs from the MERRA2 GMI simulation in emissions, time
598 frame, resolution, and meteorology (unconstrained vs specified dynamics), among others. Because
599 CCMI output is only available at a monthly resolution, we created a separate parameterization,
600 hereafter referred to as "M2GMI-monthly", using MERRA2 GMI output with identical parameterization
601 inputs as the daily parameterization but using monthly-averaged values. Performance is similar to that
602 of the parameterization trained on daily data and averaged over monthly timescales, with the NRMSE
603 for the troposphere on the order of 6 -7% depending on the year.
604
605   When output from the CCMI simulation is used as inputs to the M2GMI-monthly
606 parameterization, performance degrades significantly. For July 2000, for example, there are distinct
607 regions of both positive and negative biases (Fig. 11a) in parameterized OH, resulting in a NRMSE of
608 13%, on par with using climatology as an OH estimate. Because the largest discrepancies are centered
609 outside of the tropics and/or in regions with low concentrations, τ_{CH4} for year 2000 is identical between
610 the CCMI and parameterized OH. When applied to 2060 (Fig. 11c), which is far outside the training
611 period of the M2GMI-monthly parameterization, there is a near universal high bias in parameterized OH,
612 resulting in a NRMSE of 16% and a τ_{CH4} biased low by 4.5%. This overestimate results in a negative trend
613 in τ_{CH4} from parameterized OH from 2000 to 2060 (Fig. 10, blue line), despite the trend in the CCMI
614 simulation being positive. Applying the MERRA2 GMI parameterization to a study using the CCMI setup
615 would therefore misrepresent the OH/CH_4 cycle.
616
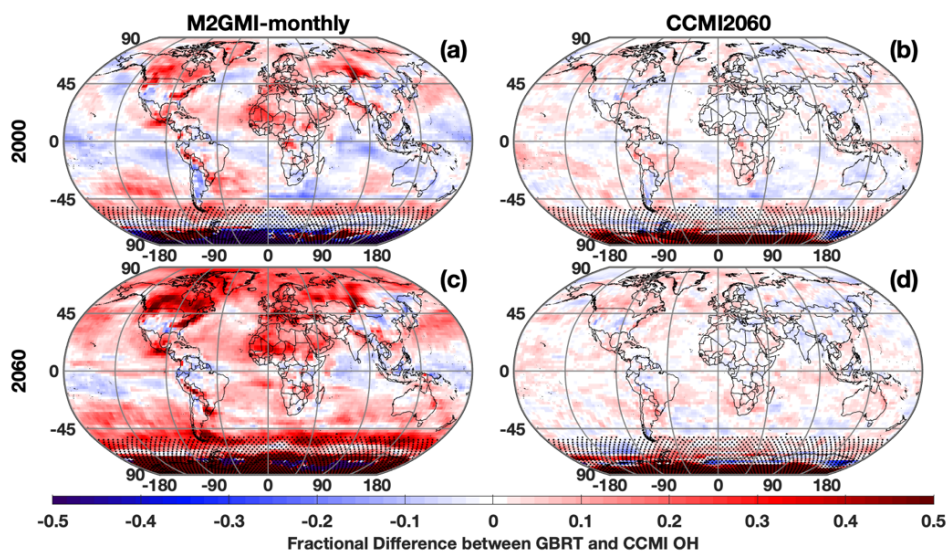
Figure 11: *Fractional difference between OH calculated by the M2GMI-monthly (left) and the CCMI2060 (right) parameterizations and OH output from the CCMI simulation. Data are averaged between 850 and 500 hPa for July 2000 (top) and July 2060 (bottom). Regions with low OH, defined as a mixing ratio of less than 0.005 pptv, are indicated with stippling.*

Through an analysis of the ranges of input parameters from both simulations, we found that the differences in parameterization performance for inputs from MERRA2 GMI and CCMI are likely largely attributable to differences in the stratospheric O_3 distributions between the two simulations. In 2060, for example, CCMI stratospheric O_3 has a much higher frequency of values above 300 DU than the M2GMI-monthly training dataset (Fig. 12). A smaller, but still noticeable, shift in the distribution is also found for the year 2000 (Fig. S9). Likewise, the accuracy of the M2GMI-monthly parameterization decreases from 2000 – 2060 as the stratospheric O_3 burden increases (Fig. 10, red line). Mechanistically, higher stratospheric ozone in CCMI should result in lower tropospheric OH because the reduction in incoming ultraviolet radiation limits tropospheric O_3 photolysis. This could also lead to a higher CO burden, due to the smaller OH sink. Comparisons between the OH and CO distributions from the two simulations are consistent with this hypothesis. Even though the M2GMI-monthly training dataset spanned the full range of stratospheric O_3 values from the CCMI simulation, the frequency of stratospheric O_3 values at higher concentration likely creates chemical environments in the CCMI simulation distinct from those in MERRA2 GMI, forcing the parameterization to extrapolate to a chemical space on which it was not trained. This highlights the need to compare the distribution of any parameterization inputs to that of the training dataset.
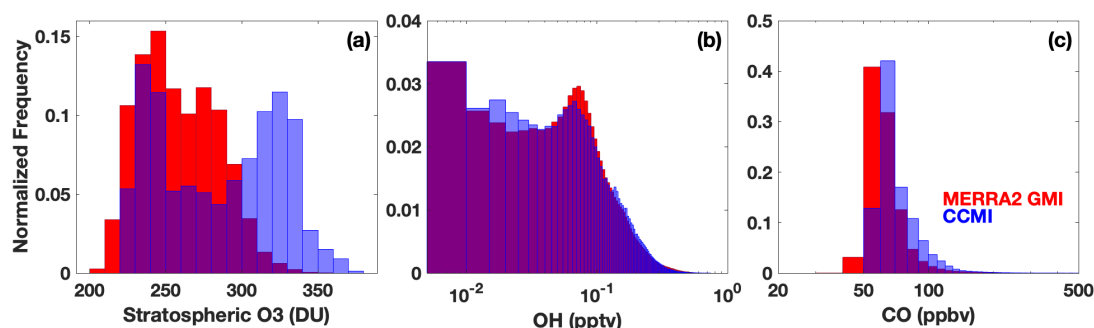
Figure 12: *Histograms showing the distribution of stratospheric column O₃ (a), tropospheric OH (b) and tropospheric CO (c) from the M2GMI-monthly parameterization training dataset (red) and from the CCMI simulation for 2060 (blue). Purple indicates areas of overlap between the two distributions.*

Again, performance improves significantly when we apply output from the CCMI simulation to the CCMI2060 parameterization. The regions of consistent high bias notable when CCMI output was applied to the M2GMI-monthly parameterization are absent for both 2000 and 2060, and NRMSE shows a factor of three improvement over the previously discussed scenario. Likewise, as discussed in Section 4.2.1, for all validation years, the parameterized OH resulted in a τ_{CH4} that agreed with the CCMI simulation between 0 and -0.46% (purple line, Fig. 10).

We conclude that for best performance, a separate parameterization should be created for each new modeling framework to capture OH variability accurately. This will not create an undue computational expense on an experiment. Because a full chemistry simulation is necessary to create the parameterization inputs of chemical species that are not calculated online, the necessary data to create a training dataset will already be available. The only additional time will be that required to format the regression tree inputs and to train the model, which takes approximately 2 – 3 hours for each month. This process can be performed using previously created python scripts with minimal changes. The flexibility that this modeling framework permits will facilitate its use even if there are major changes to the underlying model dynamics or chemistry.

5.0 Code Availability
The scripts used to create the training datasets and a sample script to create a parameterization have been archived by Zenodo and are available on github at https://doi.org/10.5281/zenodo.6046037 (Anderson, 2022).

6.0 Data Availability
Output from the MERRA2 GMI simulation are publicly available at https://acd-ext.gsfc.nasa.gov/Projects/GEOSCCM/MERRA2GMI/. Output from the GEOSCCM simulation for CCMI is available at the Centre for Environmental Data Analysis (CED), the Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation, at http://data.ceda.ac.uk/badc/wcrp-ccmi/data/CCMI-1/output.

7.0 Author Contributions
DCA wrote the manuscript, performed the data analysis, and created the parameterizations. BND and MBFC developed the idea for the parameterization. SAS performed three-dimensional modeling for the work. JMN and PDI provided advice on machine learning. All authors helped develop ideas for the analysis and contributed to the manuscript,

677
678 **8.0 Competing Interests**
679 The authors declare that they have no conflict of interest.
680
681 **9.0 Acknowledgements**

684
685
686 **10. References**

687 Anderson, D. C., Duncan, B. N., Fiore, A. M., Baublitz, C. B., Follette-Cook, M. B., Nicely, J. M., and Wolfe,
688 G. M.: Spatial and temporal variability in the hydroxyl (OH) radical: understanding the role of large-scale
689 climate features and their influence on OH through its dynamical and photochemical drivers,
690 Atmospheric Chemistry and Physics, 21, 6481-6508, 10.5194/acp-21-6481-2021, 2021.

691 Anderson, D.C., Follette-Cook, M.B., Strode, S.A., Nicely, J.M., Liu, J., Ivatt, P.D., Duncan, B.N.: Code for
692 the development of.a parameterization of OH for CCMs, zenodo[code],
693 https://doi.org/10.5281/zenodo.6046037.

694 Chen, T., and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, KDD '16: Proceedings of the 22nd
695 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785-794.

696 Chen, Y.-H., and Prinn, R. G.: Estimation of atmospheric methane emissions between 1996 and 2001
697 using a three-dimensional global chemical transport model, Journal of Geophysical Research:
698 Atmospheres, 111, https://doi.org/10.1029/2005JD006058, 2006.

699 Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., Martin, R. V., Logan, J. A.,
700 Higurashi, A., and Nakajima, T.: Tropospheric Aerosol Optical Thickness from the GOCART Model and
701 Comparisons with Satellite and Sun Photometer Measurements, Journal of the Atmospheric Sciences,
702 59, 461-483, 10.1175/1520-0469(2002)059<0461:TAOTFT>2.0.CO;2, 2002.

703 Colarco, P., da Silva, A., Chin, M., and Diehl, T.: Online simulations of global aerosol distributions in the
704 NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth, Journal of
705 Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012820, 2010.

706 Du, M., Huang, K., Zhang, S., Huang, C., Gong, Y., and Yi, F.: Water vapor anomaly over the tropical
707 western Pacific in El Niño winters from radiosonde and satellite observations and ERA5 reanalysis data,
708 Atmos. Chem. Phys., 21, 13553-13569, 10.5194/acp-21-13553-2021, 2021.

709 Duncan, B., Portman, D., Bey, I., and Spivakovsky, C.: Parameterization of OH for efficient computation in
710 chemical tracer models, Journal of Geophysical Research: Atmospheres, 105, 12259-12262,
711 10.1029/1999JD901141, 2000.

712 Duncan, B. N.: Indonesian wildfires of 1997: Impact on tropospheric chemistry, Journal of Geophysical
713 Research, 108, 10.1029/2002jd003195, 2003a.

714 Duncan, B. N.: Interannual and seasonal variability of biomass burning emissions constrained by satellite
715 observations, Journal of Geophysical Research, 108, 10.1029/2002jd002378, 2003b.

716 Duncan, B. N., Logan, J. A., Bey, I., Megretskaia, I. A., Yantosca, R. M., Novelli, P. C., Jones, N. B., and
717 Rinsland, C. P.: Global budget of CO, 1988–1997: Source estimates and validation with a global model,
718 Journal of Geophysical Research, 112, 10.1029/2007jd008459, 2007a.

719 Duncan, B. N., Strahan, S. E., Yoshida, Y., Steenrod, S. D., and Livesey, N.: Model study of the cross-
720 tropopause transport of biomass burning pollution, Atmos. Chem. Phys., 7, 3713-3736, 10.5194/acp-7-
721 3713-2007, 2007b.

722 Duncan, B. N., and Logan, J. A.: Model analysis of the factors regulating the trends and variability of
723 carbon monoxide between 1988 and 1997, Atmospheric Chemistry and Physics, 8, 2008.

724 Elith, J., Leathwick, J. R., and Hastie, T.: A working guide to boosted regression trees, J Anim Ecol, 77,
725 802-813, 10.1111/j.1365-2656.2008.01390.x, 2008.

726 Elshorbany, Y. F., Duncan, B. N., Strode, S. A., Wang, J. S., and Kouatchou, J.: The description and
727 validation of the computationally Efficient CH4–CO–OH (ECCOHv1.01) chemistry module for 3-D model
728 applications, Geoscientific Model Development, 9, 799-822, 10.5194/gmd-9-799-2016, 2016.

729 Fiore, A. M., Horowitz, L. W., Dlugokencky, E. J., and West, J. J.: Impact of meteorology and emissions on
730 methane trends, 1990–2004, Geophysical Research Letters, 33, 10.1029/2006gl026199, 2006.

731 Gaubert, B., Worden, H. M., Arellano, A. F. J., Emmons, L. K., Tilmes, S., Barré, J., Martinez Alonso, S.,
732 Vitt, F., Anderson, J. L., Alkemade, F., Houweling, S., and Edwards, D. P.: Chemical Feedback From
733 Decreasing Carbon Monoxide Emissions, Geophysical Research Letters, 44, 9985-9995,
734 https://doi.org/10.1002/2017GL074987, 2017.

735 Gelaro, R., McCarty, W., Suarez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C., Darmenov, A.,
736 Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty,
737 A., da Silva, A., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G.,
738 Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era
739 Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), J Clim, Volume 30, 5419-
740 5454, 10.1175/JCLI-D-16-0758.1, 2017.

741 Holmes, C. D.: Methane Feedback on Atmospheric Chemistry: Methods, Models, and Mechanisms,
742 Journal of Advances in Modeling Earth Systems, 10, 1087-1099,
743 https://doi.org/10.1002/2017MS001196, 2018.

744 Ivatt, P. D., and Evans, M. J.: Improving the prediction of an atmospheric chemistry transport model
745 using gradient-boosted regression trees, Atmospheric Chemistry and Physics, 20, 8063-8082,
746 10.5194/acp-20-8063-2020, 2020.

747 Keller, C. A., and Evans, M. J.: Application of random forest regression to the calculation of gas-phase
748 chemistry within the GEOS-Chem chemistry model v10, Geoscientific Model Development, 12, 1209-
749 1225, 10.5194/gmd-12-1209-2019, 2019.

750 Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D., and Tessum, C. W.: Toward Stable, General Machine-
751 Learned Models of the Atmospheric Chemical System, Journal of Geophysical Research: Atmospheres,
752 125, e2020JD032759, https://doi.org/10.1029/2020JD032759, 2020.

753   Kleipool, Q. L., Dobber, M. R., de Haan, J. F., and Levelt, P. F.: Earth surface reflectance climatology from
754   3 years of OMI data, Journal of Geophysical Research: Atmospheres, 113,
755   https://doi.org/10.1029/2008JD010290, 2008.

756   Kug, J.-S., Jin, F.-F., and An, S.-I.: Two Types of El Niño Events: Cold Tongue El Niño and Warm Pool El
757   Niño, Journal of Climate, 22, 1499-1515, 10.1175/2008JCLI2624.1, 2009.

758   Laughner, J. L., Neu, J. L., Schimel, D., Wennberg, P. O., Barsanti, K., Bowman, K. W., Chatterjee, A.,
759   Croes, B. E., Fitzmaurice, H. L., Henze, D. K., Kim, J., Kort, E. A., Liu, Z., Miyazaki, K., Turner, A. J.,
760   Anenberg, S., Avise, J., Cao, H., Crisp, D., de Gouw, J., Eldering, A., Fyfe, J. C., Goldberg, D. L., Gurney, K.
761   R., Hasheminassab, S., Hopkins, F., Ivey, C. E., Jones, D. B. A., Liu, J., Lovenduski, N. S., Martin, R. V.,
762   McKinley, G. A., Ott, L., Poulter, B., Ru, M., Sander, S. P., Swart, N., Yung, Y. L., and Zeng, Z. C.: Societal
763   shifts due to COVID-19 reveal large-scale complexities and feedbacks between atmospheric chemistry
764   and climate change, Proc Natl Acad Sci U S A, 118, 10.1073/pnas.2109481118, 2021.

765   Lawrence, M. G., Jöckel, P., and von Kuhlmann, R.: What does the global mean OH concentration tell
766   us?, Atmos. Chem. Phys., 1, 37-49, 10.5194/acp-1-37-2001, 2001.

767   Lundberg, S. M., and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, 31st Conference
768   on Neural Information Processing Systems, Long Beach, CA, USA, 2017.

769   Morgenstern, O., Hegglin, M. I., Rozanov, E., Connor, F. M., Abraham, N. L., Akiyoshi, H., Archibald, A. T.,
770   Bekki, S., Butchart, N., Chipperfield, M. P., Deushi, M., Dhomse, S. S., Garcia, R. R., Hardiman, S. C.,
771   Horowitz, L. W., Jöckel, P., Josse, B., Kinnison, D., Lin, M., Mancini, E., Manyin, M. E., Marchand, M.,
772   Marécal, V., Michou, M., Oman, L. D., Pitari, G., Plummer, D. A., Revell, L. E., Saint-Martin, D., Schofield,
773   R., Stenke, A., Stone, K., Sudo, K., Tanaka, T. Y., Tilmes, S., Yamashita, Y., Yoshida, K., and Zeng, G.:
774   Review of the global models used within phase 1 of the Chemistry–Climate Model Initiative (CCMI),
775   Geoscientific Model Development, 10, 639-671, 10.5194/gmd-10-639-2017, 2017.

776   Murray, L. T., Logan, J. A., and Jacob, D. J.: Interannual variability in tropical tropospheric ozone and OH:
777   The role of lightning, Journal of Geophysical Research: Atmospheres, 118, 11,468-411,480,
778   10.1002/jgrd.50857, 2013.

779   Murray, L. T., Mickley, L. J., Kaplan, J. O., Sofen, E. D., Pfeiffer, M., and Alexander, B.: Factors controlling
780   variability in the oxidative capacity of the troposphere since the Last Glacial Maximum, Atmospheric
781   Chemistry and Physics, 14, 3589-3622, 10.5194/acp-14-3589-2014, 2014.

782   Murray, L. T., Fiore, A. M., Shindell, D. T., Naik, V., and Horowitz, L. W.: Large uncertainties in global
783   hydroxyl projections tied to fate of reactive nitrogen and carbon, Proceedings of the National Academy
784   of Sciences, 118, e2115204118, 10.1073/pnas.2115204118, 2021.

785   Nicely, J. M., Salawitch, R. J., Canty, T., Anderson, D. C., Arnold, S. R., Chipperfield, M. P., Emmons, L. K.,
786   Flemming, J., Huijnen, V., Kinnison, D. E., Lamarque, J.-F., Mao, J., Monks, S. A., Steenrod, S. D., Tilmes,
787   S., and Turquety, S.: Quantifying the causes of differences in tropospheric OH within global models,
788   Journal of Geophysical Research: Atmospheres, n/a-n/a, 10.1002/2016JD026239, 2017.

789   Nicely, J. M., Duncan, B. N., Hanisco, T. F., Wolfe, G. M., Salawitch, R. J., Deushi, M., Haslerud, A. S.,
790   Jöckel, P., Josse, B., Kinnison, D. E., Klekociuk, A., Manyin, M. E., Marécal, V., Morgenstern, O., Murray, L.
791   T., Myhre, G., Oman, L. D., Pitari, G., Pozzer, A., Quaglia, I., Revell, L. E., Rozanov, E., Stenke, A., Stone, K.,

Strahan, S., Tilmes, S., Tost, H., Westervelt, D. M., and Zeng, G.: A machine learning examination of hydroxyl radical differences among model simulations for CCMI-1, Atmospheric Chemistry and Physics, 20, 1341-1361, 10.5194/acp-20-1341-2020, 2020.

Oman, L. D., Ziemke, J. R., Douglass, A. R., Waugh, D. W., Lang, C., Rodriguez, J. M., and Nielsen, J. E.: The response of tropical tropospheric ozone to ENSO, Geophysical Research Letters, 38, n/a-n/a, 10.1029/2011gl047865, 2011.

Oman, L. D., Douglass, A. R., Ziemke, J. R., Rodriguez, J. M., Waugh, D. W., and Nielsen, J. E.: The ozone response to ENSO in Aura satellite measurements and a chemistry-climate simulation, Journal of Geophysical Research-Atmospheres, 118, 965-976, 10.1029/2012jd018546, 2013.

Orbe, C., Oman, L. D., Strahan, S. E., Waugh, D. W., Pawson, S., Takacs, L. L., and Molod, A. M.: Large-Scale Atmospheric Transport in GEOS Replay Simulations, Journal of Advances in Modeling Earth Systems, 9, 2545-2560, 10.1002/2017ms001053, 2017.

Paek, H., Yu, J.-Y., and Qian, C.: Why were the 2015/2016 and 1997/1998 extreme El Niños different?, Geophysical Research Letters, 44, 1848-1856, https://doi.org/10.1002/2016GL071515, 2017.

Prather, M. J.: Time scales in atmospheric chemistry: Theory, GWPs for CH4 and CO, and runaway growth, Geophysical Research Letters, 23, 2597-2600, https://doi.org/10.1029/96GL02371, 1996.

Rigby, M., Montzka, S. A., Prinn, R. G., White, J. W. C., Young, D., O'Doherty, S., Lunt, M. F., Ganesan, A. L., Manning, A. J., Simmonds, P. G., Salameh, P. K., Harth, C. M., Muhle, J., Weiss, R. F., Fraser, P. J., Steele, L. P., Krummel, P. B., McCulloch, A., and Park, S.: Role of atmospheric oxidation in recent methane growth, Proc Natl Acad Sci U S A, 114, 5373-5377, 10.1073/pnas.1616426114, 2017.

Rowlinson, M. J., Rap, A., Arnold, S. R., Pope, R. J., Chipperfield, M. P., McNorton, J., Forster, P., Gordon, H., Pringle, K. J., Feng, W., Kerridge, B. J., Latter, B. L., and Siddans, R.: Impact of El Niño–Southern Oscillation on the interannual variability of methane and tropospheric ozone, Atmospheric Chemistry and Physics, 19, 8669-8686, 10.5194/acp-19-8669-2019, 2019.

Saito, R., Patra, P. K., Sweeney, C., Machida, T., Krol, M., Houweling, S., Bousquet, P., Agusti-Panareda, A., Belikov, D., Bergmann, D., Bian, H., Cameron-Smith, P., Chipperfield, M. P., Fortems-Cheiney, A., Fraser, A., Gatti, L. V., Gloor, E., Hess, P., Kawa, S. R., Law, R. M., Locatelli, R., Loh, Z., Maksyutov, S., Meng, L., Miller, J. B., Palmer, P. I., Prinn, R. G., Rigby, M., and Wilson, C.: TransCom model simulations of methane: Comparison of vertical profiles with aircraft measurements, Journal of Geophysical Research: Atmospheres, 118, 3891-3904, https://doi.org/10.1002/jgrd.50380, 2013.

Shapley, L. S.: A Value for N-Person Games, in: Contributions to the Theory of Games, edited by: Kuhn, H. W., and Tucker, A. W., Ann. Math. Studies, 28, Princeton University Press, 307-317, 1953.

Shi, L., Schreck, C., and Schröder, M.: Assessing the Pattern Differences between Satellite-Observed Upper Tropospheric Humidity and Total Column Water Vapor during Major El Niño Events, Remote Sensing, 10, 10.3390/rs10081188, 2018.

Spivakovsky, C. M., Wofsy, S. C., and Prather, M. J.: A numerical method for parameterization of atmospheric chemistry: Computation of tropospheric OH, Journal of Geophysical Research: Atmospheres, 95, 18433-18439, https://doi.org/10.1029/JD095iD11p18433, 1990.

Geoscientific
Model Development
Discussions

830 Spivakovsky, C. M., Logan, J. A., Montzka, S. A., Balkanski, Y. J., Foreman-Fowler, M., Jones, D. B. A.,
831 Horowitz, L. W., Fusco, A. C., Brenninkmeijer, C. A. M., Prather, M. J., Wofsy, S. C., and McElroy, M. B.:
832 Three-dimensional climatological distribution of tropospheric OH: Update and evaluation, Journal of
833 Geophysical Research: Atmospheres, 105, 8931-8980, 10.1029/1999jd901006, 2000.

834 Stirnberg, R., Cermak, J., Fuchs, J., and Andersen, H.: Mapping and Understanding Patterns of Air Quality
835 Using Satellite Data and Machine Learning, Journal of Geophysical Research: Atmospheres, 125,
836 e2019JD031380, 10.1029/2019JD031380, 2020.

837 Strahan, S. E., Duncan, B. N., and Hoor, P.: Observationally derived transport diagnostics for the
838 lowermost stratosphere and their application to the GMI chemistry and transport model, Atmos. Chem.
839 Phys., 7, 2435-2445, 10.5194/acp-7-2435-2007, 2007.

840 Strode, S. A., Duncan, B. N., Yegorova, E. A., Kouatchou, J., Ziemke, J. R., and Douglass, A. R.: Implications
841 of carbon monoxide bias for methane lifetime and atmospheric composition in chemistry climate
842 models, Atmospheric Chemistry and Physics, 15, 11789-11805, 10.5194/acp-15-11789-2015, 2015.

843 Strode, S. A., Ziemke, J. R., Oman, L. D., Lamsal, L. N., Olsen, M. A., and Liu, J.: Global changes in the
844 diurnal cycle of surface ozone, Atmospheric Environment, 199, 323-333,
845 10.1016/j.atmosenv.2018.11.028, 2019.

846 Turner, A. J., Frankenberg, C., Wennberg, P. O., and Jacob, D. J.: Ambiguity in the causes for decadal
847 trends in atmospheric methane and hydroxyl, Proceedings of the National Academy of Sciences, 114,
848 5367, 10.1073/pnas.1616020114, 2017.

849 Turner, A. J., Fung, I., Naik, V., Horowitz, L. W., and Cohen, R. C.: Modulation of hydroxyl variability by
850 ENSO in the absence of external forcing, Proc Natl Acad Sci U S A, 115, 8931-8936,
851 10.1073/pnas.1807532115, 2018.

852 Voulgarakis, A., Naik, V., Lamarque, J. F., Shindell, D. T., Young, P. J., Prather, M. J., Wild, O., Field, R. D.,
853 Bergmann, D., Cameron-Smith, P., Cionni, I., Collins, W. J., Dalsøren, S. B., Doherty, R. M., Eyring, V.,
854 Faluvegi, G., Folberth, G. A., Horowitz, L. W., Josse, B., MacKenzie, I. A., Nagashima, T., Plummer, D. A.,
855 Righi, M., Rumbold, S. T., Stevenson, D. S., Strode, S. A., Sudo, K., Szopa, S., and Zeng, G.: Analysis of
856 present day and future OH and methane lifetime in the ACCMIP simulations, Atmospheric Chemistry and
857 Physics, 13, 2563-2587, 10.5194/acp-13-2563-2013, 2013.

858 Voulgarakis, A., Marlier, M. E., Faluvegi, G., Shindell, D. T., Tsigaridis, K., and Mangeon, S.: Interannual
859 variability of tropospheric trace gases and aerosols: The role of biomass burning emissions, Journal of
860 Geophysical Research: Atmospheres, 120, 7157-7173, 10.1002/2014jd022926, 2015.

861 Wang, J. S., Logan, J. A., McElroy, M. B., Duncan, B. N., Megretskaia, I. A., and Yantosca, R. M.: A 3-D
862 model analysis of the slowdown and interannual variability in the methane growth rate from 1988 to
863 1997, Global Biogeochemical Cycles, 18, https://doi.org/10.1029/2003GB002180, 2004.

864 Wolfe, G. M., Nicely, J. M., St Clair, J. M., Hanisco, T. F., Liao, J., Oman, L. D., Brune, W. B., Miller, D.,
865 Thames, A., Gonzalez Abad, G., Ryerson, T. B., Thompson, C. R., Peischl, J., McCain, K., Sweeney, C.,
866 Wennberg, P. O., Kim, M., Crounse, J. D., Hall, S. R., Ullmann, K., Diskin, G., Bui, P., Chang, C., and Dean-
867 Day, J.: Mapping hydroxyl variability throughout the global remote troposphere via synthesis of airborne

868   and satellite formaldehyde observations, Proc Natl Acad Sci U S A, 116, 11171-11180,
869   10.1073/pnas.1821661116, 2019.

870   Worden, J., Jiang, Z., Jones, D. B. A., Alvarado, M., Bowman, K., Frankenberg, C., Kort, E. A., Kulawik, S.
871   S., Lee, M., Liu, J., Payne, V., Wecht, K., and Worden, H.: El Niño, the 2006 Indonesian peat fires, and the
872   distribution of atmospheric methane, Geophysical Research Letters, 40, 4938-4943,
873   https://doi.org/10.1002/grl.50937, 2013.

874   Yan, R., Ma, Z., Zhao, Y., and Kokogiannakis, G.: A decision tree based data-driven diagnostic strategy for
875   air handling units, Energy and Buildings, 133, 37-45, https://doi.org/10.1016/j.enbuild.2016.09.039,
876   2016.

877   Zhang, Z., Zimmermann, N. E., Calle, L., Hurtt, G., Chatterjee, A., and Poulter, B.: Enhanced response of
878   global wetland methane emissions to the 2015–2016 El Niño-Southern Oscillation event, Environmental
879   Research Letters, 13, 10.1088/1748-9326/aac939, 2018.
880