Review of "Revisiting the critical role of stabilized Criegee intermediates (sCIs) in sulfuric acid formation: coupling mechanistic updates with interpretable machine learning"

Author(s): Yuhuan Zhu et al.

This study is relevant to the field of atmospheric chemistry and addresses an important topic: the contribution of the stabilized Criegee intermediates (sCI) reactions with SO_2 to the formation of H_2SO_4 in the atmosphere, which subsequent contributes to the production of sulfate aerosols. The paper provides new findings coupling updates of the Master Chemical Mechanism (MCM), a benchmark mechanism within the atmospheric chemistry community, and machine learning technique. I recommend publication after addressing the comments below.

The text lacks clarity at times and there is need to provide more detail all throughout the manuscript, particularly about the application of machine learning method. The terms used in the description of the machine learning method should be clearly defined, given that the intended readership may not be familiar with this specialised terminology and the associated methodology. In general, the figure captions should be expanded to more clearly describe the plotted variables and symbols.

At the end of the Introduction the authors should state that they used field observations obtained at the Wuhai City Super Monitoring Station and explain why this site was chosen as representative for their study.

Specific comments

- line 15: '…are recognised to be atmospheric intermediates responsible for the oxidation of sulfur dioxide' instead of '…are recognised to be one of the free radicals oxidated sulfur dioxide'. It is well known that Criegee intermediates have a zwitterionic character and, thus referring to them as 'intermediates' rather than 'free radicals' is more appropriate.

- line 15: It is not clear what the word 'dominant' means in here. OH radicals are not the most abundant radicals in the atmosphere. The authors should clarify that the intended meaning is that OH is typically the dominant oxidant of SO_2.

- line 40: add the word 'Earth' before 'surface', i.e. 'near the Earth surface'.

- line 50: The reference 'Anon, n.d.' (line 423 in the reference list) should include the source of the pdf document, such as a webpage and the date the webpage was accessed by the authors.

- line 54: the cited reference, Boy et al., 2013, is not in the list of references given at the end of the manuscript

- line 58-59: lack of references for: 'Existing studies reported …environments'.

The authors should comment on a number of other relevant previous studies addressing the contribution of the sCI + SO_2 reactions to the production of H_2SO_4 and sulfate aerosols in the

atmosphere, such as: Mauldin Iii et al. *Nature* 2012; Kim, S et al. *Environ. Sci. Technol.* 2015; Kukui, A. *Atmos. Chem. Phys.* 2021; Sarwar, G.et al *Atmos. Environ.* 2014.

- lines 63-64: The authors should change '…unimolecular decomposition as well' to '…unimolecular decomposition/isomerisation as well'

- line 67 states that there is an 'intense competition' between $H_2O/(H_2O)_2$ and SO_2 for reaction with sCI'. The authors should clarify which type of Criegee intermediates they are referring to here as the intensity of this competition depends on the sCI structure; for example CH_2OO reacts much more rapidly with water vapour than other sCIs.

- lines 70-71: It is not a direct reaction of VOCs with NOx producing O_3. Therefore, please change the statement about the O_3 formation in the troposphere at day time to one such as '…O_3 is primarily formed through chemistry involving VOCs and NOx, where NO_2 produced in these reactions photolyses to generate ozone…'.

- line 72: The O_3 photolysis in the presence of water vapour is an important source of OH at day time and should be mentioned: '…through the photolysis of precursors such as HONO and O_3 in the presence of water vapour…'

- lines 74-75: The statement 'OH and sCIs are intermediate products generated via different reaction pathways' is not completely true. The authors state just before this that alkene ozonolysis is an important source of OH at night time; the decomposition of sCI formed following alkene ozonolysis can be a significant source of OH.

- line 85: Reference for MCM v3.3.1 is missing

- lines 106-107: The statement 'the concentration of bimolecular water is 10^4 times the concentration of H_2O in the atmosphere' is wrong. I think the authors meant the other way around. There is no explanation why [water monomer] was considered 10^4 times larger than [water dimer]. From the equilibrium constant for the dimerisation, $K = [dimer]/[monomer]^2$, it follows that at $T = 25^oC$ $[dimer] = 10^{-4}$ [monomer] if RH is around 13%. How much was the relative humidity and temperature at the observation site (Wuhai city)? A clear explanation about the choice of the 10^4 factor is needed, in both the manuscript and the supplement, where this factor is included in the rate coefficients for the sCI + water dimer reactions.

- Tables 1 and 2: Remove the word 'bimolecular' from the title as the tables show pseudo-first rate coefficients for the sCI decomposition/isomerisation too. Suggest adding notes under the tables showing the units of the rate coefficients **(**Please consult how tables are presented in other papers published in Atmos. Chem. Phys.) The errors associated with the rate coefficient values should be included, as well as the temperature, pressure, and a reference for MCM v3.3.1.

- line 119: Please add 'see Section 2.2.2 after '(…and PAN)'

- lines 122-133 (Observation data): The key observations time series as well as a chart showing the percentage contributions of the alkenes shown in Table 2 to the total alkene concentration during the campaign should be included in the supplement.

Were all the observations used in the present study? I suggest removal of the ones which were not used.

- line 135, regarding 'We treated the AtChem inputs as features'. Is the meaning that part of the AtChem outputs represented input variables ('features') in the machine learning method? Please re-write the sentence to clarify. The term 'feature' should be explained here.

- line 142: Please explain what is meant by 'L1 and L2 penalties'

- line 143: Please add a reference after Python xgboost library.

- line 144: The term 'hyperparameter' should be explained.

- line 152: Suggest to replace 'xi and xc constitute…' with 'the sum of features xi and xc constitute…'

- line 153: Add '(equation 1)' at the end of the sentence, i.e. '…model (equation 1)'

- line 154: What does *E* represent?

- line 161: The authors should provide examples of the specific fields they are referring to in: '…adopted across abroad range of fields'.

- lines 175-176 states that 'Figures 1a and 1b displays the global SHAP values for each feature, ranked from top to bottom by their mean |SHAP| values.' However, the high to low axes in those figures are labelled 'feature value' instead of SHAP. The meaning of SHAP in the present study should be explained.

- lines 200-201: The authors should provide references—such as Onel et al, *Phys.Chem.Chem.Phys.* 2021 and Lade et al. *J. Phys. Chem. A* 2024—that discuss the dominant removal of E-CH_3CHOO and CH_2OO by reaction with water vapour, in comparison with their losses via reaction with SO_2 under tropospherically relevant conditions.

- Figures 1(a) and 1(b): The meaning of the x axis labels are confusing and should be explained.

- Figure 2: The unit of SHAP value is mole cm^{-3} (unit of concentration) while in Figure 1 is mole cm^{-3} s^{-1} (unit of rate). Why is this difference? The authors should clarify why moles were used rather than number of molecules, the latter being more commonly used in atmospheric chemistry. What are D1(n=120) and D2(n=128) in the legend of top left figure? There are no units for 'chemical feature concentrations'. The authors should clarify what is meant by 'chemical feature concentrations' in both the main text and Figure 2 capture. Do these represent the initial alkene concentration inputs in the machine learning model?

-line 235: The authors should specify which 'specific region' they are referring to

- lines 244 - 245 states: that 'the strongest pairs' are 'O_3 × alkene%, O_3 × VOCs, and VOCs × alkene%'. However, the Top 10 Interaction Effects in Figure 3b shows that the contribution of NO_x × NO_2% is larger than the contribution of VOCs × alkene%. Why NO_x × NO_2% is not listed in 'the strongest pairs'?

- Figure 3b: I recommend including error bars in the contribution values showed in the Main Effects Contribution and Top 10 Interaction Effects figures. The authors should describe more clearly the methodology used to generate the pie chart in in both the main text and Figure 3 capture. What is the meaning of the numbers on the right vertical axis of the figure in the bottom right corner? The legend includes only the text '(b) ANOVA effect analysis for all features', which is not sufficiently explanatory.

- line 258, regarding: 'The three factors with the largest main effects …in the order O3 > alkene% > VOCs.' This sentence refers to Figure 4 where the Main Effects Contribution plot shows that the order is O3 > VOCs > alkene%.

- line 258: The authors should explain why the NO_x × NO_2% interaction is not listed in 'the strongest pairs' because the Top 10 Interaction Effects plot shows that its contribution is the largest (see similar comment about Figure 3 above).

- line 263: 'the promoting potential of O_3, VOCs, and alkene% on μsCIs% was unlocked' is confusing and should be clarified.

- Figure 4 capture: The word 'assessing' should be replaced by 'assessment of'. Regarding both Figures 3 and 4 captures: Please include how LSO_2(g) was generated and what machine learning method was used to generate the plots in (a).

- lines 273-284: The authors should clarify what is meant by low – high NO_2%. The entire paragraph is somewhat confusing and should be reorganised for better clarity.

- Figure 5: There are no explanations for the numbers shown in any of the schematics (a-f), making their meaning unclear. Please clarify in both the main text and the figure capture.

- Figure 6: What are the differences between SA-sCI and SA-sCIg and between SA-OH and SA-OHg? The main text should clearly state what SA-sCI and SA-OH represent and the figure capture should explain the meaning of all 4 notations (SA-sCI, SA-sCIg, SA-OH and SA-OHg) and which version of MCM corresponds to each of the plot line (black and purple).

- line 342: Please state the meaning of WS.

- line 348: The authors should clarify the rationale for including Fe in the features contributing to the sulfate formation.

- Figure 7(a): See comments on Figure 1a and b above.

- Figure 7(b): Explain 'high-sCIs and low-sCIs datasets'.

- line 354: A reference for the methodology used to generate comparative 'beeswarm' plots should be given.

- line 359: Replace the word 'indispensable' with 'significant'.

- line 378: Explain the word 'paradoxically'.

- line 383: Replace the word 'indispensable' with 'important'.

- lines 386-387 state: ‘While our box model simulations …they are limited by the exclusion of meteorological factors..’. Please clarify what meteorological factors were excluded as line 118 states: ‘the model was constrained by the observed meteorological parameters (T, RH and p)’.

-