Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-225-RC2, 2021

Review of the manuscript "ML-TOMCAT: Machine-Learning-Based Satellite-Corrected Global Stratospheric Ozone Profile Dataset from a Chemical Transport Model" by S.S. Dhomse et al.

Anonymous Referee #2

Referee comment on "ML-TOMCAT: machine-learning-based satellite-corrected global stratospheric ozone profile data set from a chemical transport model" by Sandip S. Dhomse et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-225-RC2, 2021

The manuscript "ML-TOMCAT: Machine-Learning-Based Satellite-Corrected Global Stratospheric Ozone Profile Dataset from a Chemical Transport Model" by S.S. Dhomse describes a new stratospheric ozone profile dataset and the machine-learning method used for its creation. The basis for the new dataset are profiles from a simulation provided by the chemical transport model TOMCAT. Biases between these profiles and profiles from the well-established ozone profile dataset SWOOSH are then calculated and used as input for a Random Forest regression method with five learners (or basis functions). The obtained coefficients for these five learners are then used to bias correct the entire TOMCAT simulation. The resulting ML-TOMCAT dataset is then compared to several different available observational datasets and differences are discussed. The manuscript is very well written, well structured, and the topic lays within the scope of the ESSD journal. However, there are a few things that I think would help to improve the manuscript, and that I would suggest the authors to consider while revising the manuscript. Comments highlighting these issues are outlined below.

General comments:

- As the authors state in the manuscript, merging techniques for ozone datasets can result in significant uncertainties within the created dataset, and can be a considerable source of differences between merged datasets. The applied machine-learning method to correct the biases between the TOMCAT simulation and SWOOSH is therefore one of the main parts of the manuscript. However, at the moment it feels too short and not detailed enough considering its importance for the creation of the dataset. Are the five learners applied in the same form to all latitudes and altitudes/pressures? How big are the residuals after the Random Forest regression has been applied? Why was the

tropospheric part of the profile also put through the Random Forest regression although there is no data from SWOOSH and the ML-TOMCAT dataset is recommended only for stratospheric values?

- It is not clear enough where the focus of the ML-TOMCAT dataset is placed. The title of the manuscript suggests that it would the stratosphere, and in the summary section the recommendation is given to only use the dataset between the tropopause and 0.1hPa, but the creation of the dataset is described for all pressure levels between 1000hPa and 0.1hPa. This is confusing. Please add more explanations why either the tropospheric levels are necessary for the creation of the dataset or why they were created but should not be used.

More specific comments:

- Line 18: Should be "…within uncertainties of the…" rather than "…within uncertainties in the…"?
- Line 29: Should be "…similar or larger magnitude…" rather than "…similar magnitude or larger…"?
- Line 37: Word missing? "…which could variability in ozone trends."
- Line 66: Remove "s" from "profiles"
- Line 102: Add "the" between "…throughout troposphere."
- Line 117: "Random Forest" has been introduced already and does not need to be written out here anymore.
- Line 121: The selection of the years to train the machine-learning model (1991-1998 and 2005-2016) seems very random here and is not explained. It is explained later in the manuscript, but I think it would be good to add an explanation here as well (not just that it is a 20-yr period, but WHY exactly these 20-yrs were chosen).
- Line 129: Please explain here what you mean with "passive ozone"
- Line 169: Here it states that the climatology was calculated for the period 2006-2020, but the figure caption (Figure 1) states that the climatology was calculated for the period 2001-2020. Which is correct?
- Line 201: the sentence part about positive and negative ozone biases is confusing here since it refers to Figure 1 and the paragraphs explains details of Figure 2. Maybe add a reference to Figure 1 at the end of the sentence?
- Line 209: What could be the reason for the significant HCl coefficients in the tropical lower stratosphere? You mention that they are present there but don't offer an explanation.
- Line 225: Remove the comma after "Since, …"
- Line 246: Reference to Figure 1 is not correct, I guess, since Figure 1 does not show ML-TOMCAT data.
- Line 321: ";" between the two references should probably be an "and"
- Line 322: I guess you mean "North Polar region" instead of "North Pole"? It might be good to mention here as well which latitudes are covered by the North Polar regions.
- Line 323: I think "South Polar region" is more appropriate than "South Pole".
- Figure 2: The differences in the blue color range a very hard to distinguish. It might be helpful to change the color scale somewhat to help the reader understand the discussion about these differences from Section 4.2.
- Figure 5: The title says "w.r.t. SAGE-CCI-OMPS", but it should be "w.r.t. GOZCARDS"
- Figure 9: Last line of the figure caption: the ")" is missing after "70°S".