Earth Syst. Sci. Data Discuss., referee comment RC1
https://doi.org/10.5194/essd-2021-225-RC1, 2021

Comment on essd-2021-225

Anonymous Referee #1

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-RC1, 2021

Review of "ML-TOMCAT: Machine-Learning-Based Satellite-Corrected Global Stratospheric Ozone Profile Dataset from a Chemical Transport Model" by Dhomse et al., Earth System Science Data

GENERAL COMMENTS

First, I think that what this team has done here is excellent, very much needed, and that the data set arising from this work will be very useful. It is something I have always wanted to do myself but, lacking infinite time, have been unable to. What is presented in this paper is what I see as an ideal combination of our knowledge of stratospheric chemistry with all available stratospheric ozone profile measurements to create a definitive 42-year history of the vertical distribution of ozone in the stratosphere. The only superior approach that I could think of would be a version of TOMCAT that has a 4Dvar assimilation of all available ozone profile measurements. You may want to say something about how your approach differs from what would be achieved from a data assimilating version of TOMCAT and why nobody has gone down the route of creating a data assimilating version of TOMCAT for this purpose.

I found that the description of exactly how the ML-TOMCAT database was created was too terse and I would like to see much more detail on that, i.e. sufficient for a reader to reliably replicate what was done.

SPECIFIC COMMENTS

Line 58: Rumour has it, but who knows whether it is true, that the true expansion for

BDBP is "Birgit's Database of Biblical Proportions".

Lines 95-96: Does this mean that the resultant ML-TOMCAT ozone profile database is not longitudinally resolved? That's a little disappointing. I thought that with TOMCAT being 3D you would be able to create a 3D ozone profile database.

Section 3: I suspect that most readers of your paper will be like me, i.e. experts in stratospheric ozone with some (or maybe even little) expertise or knowledge of machine-learning methods. When I think of a 'random forest' I think of someone who has planted trees somewhere arbitrary to claim carbon credits. All I am saying here is that you may need to be more pedagogical in your approach to Section 3. You may need to describe in greater detail, and at a higher level, the machine-learning methods that you are using. Don't miss the opportunity to educate your reader to the level that they require to understand your paper, being cognizant of the level of knowledge your typical reader is likely to have. If you do it well, they will be eternally grateful. Test the material on non-ML-experts to see whether it has been pitched at the appropriate level. Just as one small example, on line 113 you refer to 'the model'. I suspect here that you are no longer referring to TOMCAT but now to some model underlying the random forest? Line 113 also refers to 'learners' for the first time without describing what these are. You will need to rewrite this section being extra careful to maintain clarity.

Line 117: Isn't the full correct name for the package scikit-learn?

Line 132: In the context of regression modelling, would a better term for 'learner' be 'predictor' or 'basis function' or 'explanatory variable'? I would suggest that, to the extent possible, you use terms that will be more familiar to the majority of your audience (atmospheric scientists) rather than terms used more by the data science community.

Equation (1): shouldn't there be coefficients at the front of each of these terms as in a standard regression model, these being the regression model coefficients? You refer to these coefficients on line 191. Otherwise this equation doesn't make sense to me. It doesn't even make sense in terms of units, e.g., dO3 would be measured in ppm and dTCO in DU. OK, maybe you are let off the hook by the 'and so all the predictor time series are detrended and normalised between 0 to 1' clarification at the end.

Line 140: I assume that TCO fields are not available from ERA5?

Line 163: And this is no small caveat. I think that it would be worth pointing readers to the SPARC S-RIP activity here.

Figure 1: Why start in December and finish in November? Here, our year runs from January to December.

Line 169: You say '(2006–2020)' but the figure caption says 2001-2020.

I think that throughout the paper things would be clearer if you replaced 'the model' with 'TOMCAT' wherever that makes sense since there is also the regression model and you don't want the readers to constantly be having to figure out which one you are referring to when you say 'the model'.

Line 175: I find this confusing. On line 170 you state that positive biases occur in the upper stratosphere and negative biases occur in the lower stratosphere. But then on line 175 you explain the source of negative ozone biases in the upper stratosphere (even though this is not what is observed) and, similarly, on line 180 you explain the source of positive biases in the lower stratosphere whereas what is observed is negative biases. Am I missing something here?

Line 195: And presumably these RF-derived biases extend to regions of the globe for which observed ozone profiles are not available?

Line 202: Here you refer to TOMCAT-SWOOSH differences whereas on line 170 you said 'SWOOSH minus TOMCAT'. I find this confusing.

Figure 2: Presumably the sum of all 5 fields other than the R^2 field should sum to the R^2 field? Has that been confirmed?

Line 209: I think that it would be worth saying more about why the HCl basis function has such high explanatory power in the tropical lower stratosphere.

Line 238: This suggests that in Figure 3 there should be stippling to show where the differences are statistically significantly different from zero at the 2 sigma level.

Line 247: I assume you mean significant improvements compared to TOMCAT? You should then say so.

Line 252: Why the need to use SAGE II and HALOE when UARS MLS always covers the

lower latitudes (34S to 34N)?

Line 262: You say with respect to GOZCARDS but in the title of the figure it says "w.r.t. SAGE-CCI-OMPS"?

Line 266: It wasn't clear to me what you meant by 'uneven variations'.

Line 275: You need to cite a paper to support this assertion that there is a bias in the ERA5 reanalyses in 2020.

Lines 276-278: As it is currently written, this sounds like conjecture. You either need to do the analysis to investigate whether or not this is the case or cite something that gives some credence to the assertion.

Line 279: I would prefer to see this referred to as 'Comparison with ozone concentrations reported on altitude levels'. In fact I would prefer to see the word 'altitude' rather than 'height' used throughout the paper, unless you are specifically referring to geopotential height rather than geometric altitude in which case I think that you should always be explicit and say 'geopotential height'. After all, Figure 6 refers to altitude rather than height.

Line 288: Yes, but if I remember correctly, the TOMCAT profiles are not used directly in the BSVert ozone data set but only as a transfer standard to calculate relative biases. So there is no reason why BSVert ozone cannot be biased against TOMCAT. I would suggest that you read Hassler et al., 2018 very carefully.

Line 289: This is not true. The negative biases extend to 25N.

Line 310: Should this be 'ozone number density profile'?

GRAMMAR AND TYPOGRAPHICAL ERRORS

I have suggested some, but not all, grammar and typographical corrections. I am hoping/expecting that further corrections will be made by the author(s) and the ESSD editorial staff.

Line 3: 'Satellite instruments obtain stratospheric ozone profile measurements' sounds rather convoluted. Why not just say 'Satellite-based instruments measure stratospheric ozone profiles'.

Line 37: Replace 'which could variability in ozone trends' with 'which could induce variability in ozone trends'.

Line 40: Replace 'As there is no' with 'As there are no'.

Line 54: Replace 'data was' with 'data were'.

Line 65: replace 'merged data' with 'merged data set'. Likewise on line 72 replace '(SWOOSH) data' with '(SWOOSH) data set'.

Line 75: Replace 'One of major' with 'One of the major'.

Line 81: Replace 'firstly' with 'first'.

Line 111: Would it not be clearer to replace 'algorithm by splitting observations' with 'algorithm to split observations'?

Line 119: Replace 'on to' with 'onto'.

Line 120: Replace 'data is' with 'data are'.

Line 152: Isn't the IUPAC convention that it is 'sulfate' rather than 'sulphate' irrespective of which side of the Atlantic you're one?

Line 198: Replace 'that RF regression model' with 'that the RF regression model'.

Line 216: Replace 'This means although' with 'This means that although'.

Line 226: Replace 'data was processed' with 'data were processed'.

Line 227: Replace 'geopotential' with 'geopotential height'.

Line 320: Replace 'is able to polar' with 'is able to model polar'.

Lines 323-325: This sentence lacks a verb.

Line 350: I believe this should be 'Hassler et al., 2018'? And please fix the reference accordingly - you appear to have used first names rather than surnames.