Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-291-AC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "ESMValTool v2.0 – Extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP" *by* Veronika Eyring et al.

Veronika Eyring et al.

veronika.eyring@dlr.de

Received and published: 19 March 2020

Reply to Anonymous Referee #1

We thank the reviewer for the very detailed and helpful comments. We have now revised our manuscript in light of these and the other review comments we have received. A pointwise reply is given below.

General comments:

This paper documents the latest version of the ESMValTool v2.0. The paper and the tool have lots of very good points. For example, one can reproduce IPCC Working Group 1 figures fairly easily. As an IPCC author and also a user of the IPCC, this alone makes the tool very valuable to our community. The ability to quickly assess the new CMIP6 models is equally valuable and so on. I strongly support the development of the tool.

That said, I found the paper very hard to read and review. The authors want to discuss each option (what they call recipes) found in the tool. In describing the recipe, it is good to show a figure or 2 to help the reader appreciate what the tool can do. Given the large number of recipes, this makes the paper very long with many figures. There is a tension between showing the figures as examples and trying to make certain scientific points/conclusion with those figures. I understand this tension well as I have also tried to produce papers like this one. Scientists want to talk about the science, not just use the figure as an example. The problem is the caveats and sometimes the complete justification for the conclusion is left out because the science is not the point. It is an example.

In this paper, there are conclusion sentences in the figure captions and figure caption text in the main body in many, many places. Cleaning these up would greatly help the readability of the paper. As a reader, I strongly dislike conclusions in the figure caption. This is especially true when the conclusion is not restated in the main text. The technique of giving conclusions in the caption is fine in a talk, but not in a paper. It just adds to the length.

Also, many figure captions are missing units and other details needed to document the figure. Sometimes these details are found in the text, but they belong

in the caption. Again, this would help the readability of the paper. I highlight some of these in my specific comment section below. As a suggestion to greatly shorten the paper, one could have the nice summary section as section 3 and move all of what is now section 3 to an appendix. I feel this would help most readers and give the authors room to add more overview type of text. If the readers need the details, they can find them in the appendix.

Another general comment is that there needs to be some documentation of the models used in the paper. Shorthand names are given in most cases. Somewhere these names have to be connected to references. Most likely another table is needed.

Finally, it needs to be made clear that most users will not run the tool themselves. The users mainly interact with a browser that displays information which has been all ready computed. This is addressed in section 4 but needs to be made clearer in the abstract, introduction and summary sections. There is a lot of misunderstanding in the community about this tool. Clearing up this aspect will help.

Thanks for valuing the development of the tool and for your detailed comments! We have followed your suggestion to move the conclusions from the figure captions to the text. We have also revisited all figure captions to include units and other details as requested in the specific comments below. We have however not moved Section 3 to an Appendix, as the structure of the paper is clear and readers who are not interested in the scientific motivation (first paragraph in each subsection) or in a particular recipe can skip this. For the majority of the readers we expect, however, that this is important information. Also from previous discussions with the community we feel it is important to scientifically justify why a certain diagnostic or recipe is included which is why we

prefer keeping the description of diagnostics and the structure of Section 3 as is. The goal of the paper is to document all large-scale diagnostics and recipes that are newly included from v1 to v2 for the community in a peer-reviewed paper. This will allow those interested in a particular diagnostic or recipe to find some more background in this paper. To consider the comment on length, we have therefore not removed individual sections and have not created an Appendix, but we have streamlined the text further, following the specific comments below and by the other reviewers. Following the reviewer's suggestion, we have included another table on the CMIP5 models used in the paper and included the references to the model papers. On the last point, there are different types of users, those who only look at the results provided through the web browser, and those who use the tool for their scientific work. We have made this clearer throughout the text as suggested.

Specific comments:

1. Line 57 – end-to-end provenance – This is jargon. What does this mean?
End-to-end provenance is the history of an item of data from its creation to its present state. It includes details about the steps that were executed in order to produce the data in its current form. For the ESMValTool this means that it is kept track of which input data have been used and which processing steps have been applied to produce a given product, i.e. typically a netCDF file or a plot. This is explained in more detail in the companion paper Righi et al. (2020), which we now explicitly refer to in the text in the revised paper.

2. Line 58 – ensure reproducibility – Of what? Need to mention it is reproducibility of the analysis/figures.
Extended as suggested.

3. Line 72 – broad user community - Who do the authors have in mind? Non-specialists would struggle with most of the analysis presented in the paper.
User community further clarified.

4. Line 77 – 2.1 and 4.7C for a doubling – By when? Is this transient (then a year is needed along with the time averaging period) or equilibrium?
Clarified.

5. Line 105 – to achieve this – To me this sounds like the tool is the solution. It is only part of the solution. Reword.
Reworded.

6. Line 131 – Reference needed for CMOR tables and definitions. I assume the CMIP6 web address is fine.
Reference added.

7. Line 137 – Section 2 needs to mention that the tool helps address a major CMIP focus – to evaluate model performance by comparing models and observations. It should also mention the OBS4MIPS effort here.
Good point and added.

8. Line 163-164 – The caption and the text use different words to describe what is shown: error versus deviation. Make them match. The text in the caption (lines 1022-1026) need moved to the text. An important thing to note in the caption is that the colors will change if models are added or removed. (Defining what is meant by "relative").
Changed as suggested. As said above, the text in the caption has been moved to the text. The term "relative" has been further described in the caption.

9. Lines 170 – 185 – The high correlation for TAS is likely related to mountains and the land-sea mask. Is there anyway to removed these imposed boundary conditions from the analysis? For precipitation, the observations over the ocean are uncertain. Likely more uncertain than the land obs. Is it possible to weight the land more in this analysis?
This diagnostic resembles the analysis of Gleckler et al. (2008) as also used in IPCC AR5. In their analysis, no weighting of specific regions or geographic features has been applied. Theoretically, this could be done as an extension of the portrait diagram of Gleckler et al. (2008). Since this diagram is rather intended as a summary providing an overview of the models' performance as a starting point for more detailed analyses, such weighting is (in our opinion) better suited for additional (other) diagnostics.

10. Figure 2, lines 1035-1036 – The sentence that starts "The figure shows both" belongs in the text.
Sentence moved as suggested.

11. Lines 185-201 – What variables were used to make the index? In the caption the phrase that starts "and in this case . . ." (line 1045) belongs in the text.
Variables clarified and sentence moved as suggested.

12. Line 218 – 220 – Figure 4 shows . . . - This is figure caption text. In the figure caption, only the first and last sentences should remain. The middle sentences belong in the main body text. What are the units?
Changed as suggested.

13. Lines 237 – 243 – Much of this text belongs in the figure caption. The sentence in the caption "Larger biases can be seen" (lines 1062 – 1064) belongs

in the text. * Note: I stop commenting on figure caption text in the main body and main body text in the caption below this point. It occurs in most captions. ****

Changed as suggested.

14. Line 253 – peculiar – What is peculiar about these regions? Change to "some".

Changed as suggested.

15. Line 260 – From the caption of figure 8, I have no idea what is plotted in figure 8. What are the units? The main text again describes the figure and belongs in the caption.

The text provides a definition and a reference for the quantile bias. The quantile bias, being a ratio, is unitless, we added this in the figure caption. We also changed the figure caption to clarify that the discussion in the caption is an example of how such a figure could be used.

16. Line 284, figure 9 – The units for both plots are missing. It is unclear to me what is being plotted.

Units clarified and caption improved.

17. Line 312, figure 10 – Units? The label says percent. Percent of what? Occurrence?

Blocking events frequency measures the percentage of blocked days in the 1979-2008 period (considering only the winter period DJF). We modified the caption of fig. 10 to specify this.

18. Line 311 – ZIP – Is this shorthand standard? "Compressed" seems better but less precise.

Indeed *zip* is standard compression format. We added "compressed" in the text.

19. Line 323 – severely impacts – This seems way too strong. It could impact prediction or projections. I have not seen many examples where it does. Change to "could severely impact" or reword.

Changed as suggested.

20. Lines 325 – 365 – I think this discussion could be greatly shortened. Just need to reference Lembo et al. 2019. I do not get much from figure 10. There is a lot of text for little gain. Even the analysis sentence in the figure caption (which belongs in the text) only states a known conclusion. Figure 14 and 15 – what are the units and shadings?

Discussion shortened as suggested and units clarified.

21. Line 385 – Somewhere in this discussion, the point that the observational record in many cases is too short to reliably assess the variability.

Added.

22. Line 419 – observed – Change to "reported".

Changed as suggested.

23. Line 421 – unsuitable – This seems too strong. It depends on the questions being asked.

Agreed, we changed the text to say "Forecast systems .... may have difficulties in reproducing climate variability and its long-term changes....".

24. Line 410 – Weather regimes – One has to use these with caution. For large climate changes, they may not be useful/reliable. Caveats are needed in this section.

Yes, indeed the tool should be used with caution when applied (for example) to future scenario simulation where large climate changes are implied. We added a sentence to point this out and to explain how one could proceed in this case.

25. Line 436 – Figure 17 – Units?
Units clarified.

26. Line 445 – ZIP – See comment no 18 above. I have stopped commenting on the use of ZIP.
Changed, see also response above.

27. Line 446 – Figure 18 – Units?
The units are already specified close to the colorbar, it is meters [m] (written rotated along the colorbar).

28. Line 464 – Figure 19 – Units?
Units clarified.

29. Line 492 – strength of northward current – This is incorrect. It is the strength of the overturning circulation – near surface and deep.
Changed as suggested.

30. Line 494 – Figure 22 caption – but it is not clear . . . - This is a funny statement for this paper. It could be computed using the model spread as an estimate. Also, this phrase belongs in the main body text.
Moved to the main body of the text and sentence rephrased as suggested.

31. Line 561 – Add "of temperature" after "Arctic amplification".
Changed to "... processes are Arctic atmospheric temperature warming amplification

...".

32. Line 570 – Figure 29 caption – A. PHC3 is used in the caption. The label is PHC. Make the labels the same. B. Eurasian Basin – Needs defined in some way. Latitude-longitude? C. Add "in Arctic" after "Eurasian Basin". D. Atlantic water too depth – sentence belongs in text. In the caption, need to say how Atlantic water is defined.
A. Changed as suggested.
B., C. Changed to "Eurasian Basin of the Arctic Ocean (as defined in Holloway et al., 2007)"
D. We have removed the sentence from the figure captions, following the reviewer's general suggestion above.

33. Line 577 – Eurasian and Amerasian basins – These basins need defined in some way.
Added "(as defined in Holloway et al., 2007)".

34. Line 589 – 590 – linearly interpolated to climatology levels – Of what? Observations? Need reference.
Clarified and reference added.

35. Lines 593 – 597 – Discussion seems to suggest that velocities are interpolated and then transports are computed. If this is the case, this is calculation is wrong. The transport needs to be computed first on the native grid and then interpolated. Doing the velocity first can and will lead to incorrect transports because of issues with the dot product.
We do not compute transports, only transects of scalar properties. Although it is clear that the confusion comes from the mention of the exchange between basins. We removed this part of the sentence and add explicitly that only temperature and salinity

are considered. The new sentence reads as follows: "For each point, a vertical profile of temperature or salinity on the original model levels is interpolated."

36. Lines 600 – 605 – Why use only T to define Atlantic water? It seems like S should be used too.
Usually, in the literature salinity is not used when Atlantic Water of the Arctic Ocean is defined.

37. Line 689 – Figure 33 – Relative bias (percent) needs better defined. I am not sure what it means. I also do not know what accumulated and averaged bare soil covered area mean.
The relative bias is computed as the difference between simulated and observed land cover area divided by the observed area. It is further converted into percentage (multiplying by 100). The accumulated land cover area is the sum of the surface area covered with a specific land cover type (here bare soil) in a given hemisphere or region, while the average covered fractions gives the ratio of the accumulated area to the total area of the hemisphere or region and, again, is converted to percentage.

38. Line 694 – LUC – Defined somewhere? I could not find it.
This was a typo, thanks for spotting. Corrected to LCC.

39. Line 712 – 713 – 5X5 model grid cells – Model grid cells are not 5X5. This is some interpolated grid.
The reviewer is right that the provided figure had been produced with an algorithm applied on an interpolated grid. However, the diagnostic also provides the option to run the algorithm on the native grid. For the new version of the manuscript, we now provide the figure obtained by running the algorithm on the native grid of the MPI-ESM-LR model. This figure version exhibits very small differences compared to the previous one.

40. Line 728 – Almost only snow-free areas are visible – This makes no sense.
We now provide more background to support this outcome: during the month of July, there is a very limited number of grid cells where the snow area fraction exceeds 0.9 (the criterion for them to be considered snow-covered). Moreover, in order to ensure that the reconstructed signal is of good quality, some more criteria need to be fulfilled for a grid cell to be included in the regression algorithm: at least 15 grid cells – either snow-free or snow-covered – need to be present in a moving window, and the sum of the considered land cover types has to equal at least 90% of the grid cell. This explains why no values are available for snow-covered areas in July for the MPI-ESM-LR model. A full description of the methodology has been included in the following paper submitted to Earth System Dynamics: Lejeune et al., Biases in the albedo sensitivity to deforestation in CMIP5 models and their impacts on the associated historical Radiative Forcing. This recently submitted paper is now referred to in the new version of the manuscript. Consistently with this new background information, we have revised the formulation for this part of the manuscript: "It can calculate albedo estimates for each of these two cases and each of the three land cover types, given that some criteria are fulfilled: the regressions are only conducted in the big boxes with a minimum number of 15 grid cells (either snow-free or snow-covered), taking only into account the grid cells where the sum of the area fractions occupied by the three considered land cover types exceeds 90%. The algorithm eventually plots global maps of the albedo changes associated with..."

41. Lines 725 – 732 – This result seems suspect to me. It needs checked against models where land cover can change during the integration. Why is the sign of the change the same in both hemispheres given that the data are for July.
Our algorithm provides the potential albedo change associated to a transition between

two land cover types. In the case of Fig. 34, it is the change associated to a transition from trees to crops/grasses, which is positive in all seasons in most areas. This is in line with satellite-derived estimates of this albedo change for July, as shown by the right-hand side of Fig. 34. To make this clearer, we added the word "potential" at the beginning of the legend of Fig. 34.

42. Line 770 – not well reproduced – This seems too strong given the large observational uncertainty.
The text has been revised to clarify the statement: "This emergent ecosystem property, calculated, for example, as a ratio of long-term average total carbon stock to gross primary productivity, has been extensively used to evaluate ESM simulations (Todd-Brown et al., 2013, Carvalhais et al., 2014; Koven et al., 2015, Koven et al., 2017). Despite the large range of observational uncertainties and sources, ESM simulations consistently exhibit a robust correlation with the observation ensembles, but with a substantial underestimation bias."

43. Line 777 – Figure 36 – The observational uncertainty shading band seems way too narrow. There are many estimates for observed carbon fluxes and they disagree a lot. The internal uncertainty estimates for any given observed data set is typically quite small relative to the disagreement between obs data sets. Therefore the figure and the text are quite misleading. Revise.
The figure and the observational data used for the figure have been updated so that it is using exactly the same as the original study of Carvalhais et al., 2014. The reviewer is correct that the observational uncertainty in the submitted version of the manuscript was way too narrow. We have updated the figure. In the updated figure as well, there is a clear difference between the models and observation, and most models are outside the observation range. We have revised the text for possible uncertainties with observation-based estimates. The text has been changed to: "Most CMIP5 models (and multi-model ensemble) have a much shorter turnover time than the observation-

based estimate across the whole latitudinal range. Even though different estimates of observation-based carbon fluxes and stocks can vary significantly, a recent study (Fan et al., 2020; Figure 5a) shows that the zonal distributions of observation-based estimates of turnover time is robust against the differences in observations."

44. Line 802 – Figure 39 – Units?
The figure has been updated. Note that this figure now uses a linear scale on the x and y axis.

45. Line 807 – but also the spatial distribution . . . - What does this phrase mean?
The spatial distribution is the distribution of the values of data in space (a map). To clarify, the text has been changed to "When viewed together, Figures 38 and 39 show the biases between the model and the observations in the surface layer relative to each other, both in terms of their spatially-independent distribution in fig. 38 and their spatially-dependent distribution in figure 39."

46. Line 807 – 809 – Figure 40 – Why are there the colored lines on the right side of the figure? The main text refers to a color scale. I do not see a color scale.
For clarity, this has been changed to: "Figure 40 shows the global average depth profile of the dissolved nitrate concentration in the CMIP5 HadGEM2-ES model and against the World Ocean Atlas dataset. The colour scale indicates the annual average, although in this specific case there is little observed inter-annual variability so the annual averages are closely overlaid. Nevertheless, this class of figure can be useful to evaluate biases between model and observations over the entire depth profile of the ocean and can also be used to identify long term changes in the vertical structure of the ocean models."

47. Line 845 – 858 – Can one access the data plotted in the figures if needed? This would be useful both to IPCC authors and anybody who needs to replot the data.
Yes, this is possible, as every diagnostic in the ESMValTool generates one (or more) netCDF file(s) containing the data resulting from the analysis and used for plotting.

48. Line 963 - How does the tool handle "bad data"? By bad I mean having the wrong units or the grid is wrong or the data is missing - as examples. How much of the error checking is human and how much is automated? Can other figures be generated if the input data is ok while bad data exists for some other variables?
Checking for errors in the input data is mostly automated in the CMOR checking module of the ESMValCore preprocessor (see Righi et al., 2020 for details). This module checks for CMOR compliance of the input data and helps to identify common errors such as inconsistent units, wrong coordinates, bad missing values, etc. Errors in the actual data are however hard to detect in an automated way.

49. Line 972 – Are the observations available? Saying they are not distributed with the tool does not address this important question.
Due to license issues, redistribution of observational data is not allowed in most cases, but the tool provides CMORizing scripts for each observational dataset used in the recipes that is not an obs4MIPs dataset. These scripts include detailed information on how to download and process the data for usage with the tool (see Righi et al. (2020) for details).

References

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler,

B., Koldunov, N., Little, B., Loosveldt Tomas, S., and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, Geosci. Model Dev., 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-291, 2019.