Interactive comment on "Towards improved and more routine Earth system model evaluation in CMIP" by Veronika Eyring et al.

Response to Anonymous Referee #2

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

The main changes compared to the previous version are:

- We have clarified that this is a viewpoint paper (see comments by Reviewer 3)
- We made the distinction between what is planned for CMIP6 and what is a long-term vision clearer in the text
- We have expanded the paper with additional information on the tools that will be applied to CMIP6 model simulations as soon as the output is submitted to the ESGF. We have also included two more example figures from these tools. However, we note that this paper is not a detailed documentation of specific evaluation tools that are described elsewhere in the literature. To make it easier for the reader, we have included an additional table with references and links to these tools.

This paper describes the desired modeling community goal to build a routine model evaluation into the Coupled Model Intercomparison Project (CMIP). It argues that the time is right within CMIP6 to make a start on this and describes the different aspects that are needed to achieve it. These include openly available evaluation software for standardized metrics of performance that can be built into community-based diagnostic packages; common formats for model data; integration of the evaluation tools into the ESGF infrastructure; documentation and visualization. The paper describes the current position on these aspects and the vision for the future.

I strongly support the ideals of the paper and think it is a useful contribution to the debate that can provide the community with some clarity on the way forward towards its goal of continuous and standardized evaluation of model performance. However my main criticism of the paper is that it blurs the lines between what is happening now as part of CMIP6 and what the future vision is. I think the authors need to make a clear distinction between the limited (but still useful) progress in developing standard tools (e.g ESMvalTool etc), progress since CMIP5 on developing CMOR and access to data in the ESGF and progress on documentation from the desired long term goals. Notable here are sections on visualisation (end of section 2.4) and all of section 3 which appear to be more aspirational than what might hope to achieved for CMIP6. A figure showing specifically the expected situation for CMIP6 would be helpful, I think.

We have made the distinction between what is possible in time for CMIP6 and the long-term vision clearer in the manuscript. Figure 4 in the manuscript displays the expected situation for CMIP6 whereas Figure 5 displays the long-term situation, so this comment is already addressed and no new figure has been added. We have however included a new table with examples of evaluation tools that will be available for CMIP6.

Specific comments

P3, l3: Here you say you are proposing a plan but I think it needs to be clearer exactly what can be done for CMIP6 and what is on the longer term

This distinction has been made more explicit.

P3, l25: You say that parts of the evaluation have 'demonstrated their value..' but then go on to say that they have 'not provided much guidance in reducing systematic biases nor have they reduced uncertainty in future projections' so what value have they demonstrated?

Model evaluation has still identified many model errors, both in individual models, as well as collectively in CMIP ensembles. Some systematic biases however remain. We refer to the most recent IPCC climate model evaluation chapter where the progress in model evaluation is assessed.

P6, ls15-20. Here are examples of vague statements about what be achieved on the CMIP6 vs longer timescales. e.g. '. . .perhaps even be hosted alongside. . .' and 'The hope is that obs4MIPs can be extended. . .'

Statement has been strengthened.

P7, l20-22. Nowhere in the paper do you mention the possibility of using these easily available evaluation packages and metrics by those seeking to chose a few models e.g. for driving regional models or as 'best estimates' for impact studies etc. This seems an issue that will raise some concerns, notably because as you say the current set of tools are basic evaluation. I think it is worth some discussion.

The risks of choosing a small subset out of the larger ensemble based on a limited or wrong set of metrics or diagnostics has been added to the discussion. We highlight that indeed the metrics need to be sufficiently broad in scope in order to avoid tuning towards a small subset of metrics. As an example of broad metrics applied successfully on a process-based manner to models, we refer to the SPARC CCMVal report. The diagnostics and performance metrics that are available already now for CMIP6 via ESMValTool, ILAMB, NCAR CVDP, PMP, etc. are broad in scope, please see the examples in the paper and by the various tools. Over time the set of diagnostics and metrics will increase and it will be more and more possible to identify compensating errors.

P10 first paragraph: Given the issues with availability of computing within ESGF to run the evaluation software why isn't a first step to make the software available to modeling groups and ask them to run the evaluation software on their own systems and then upload the results to the ESGF?

The evaluation packages are available for the model groups and we suggest that model groups run them locally on their model before submitting the model output to the ESGF. However, we are not making this a requirement.

P12, first paragraph: What are the plans to detail the tuning process for CMIP6. Is this going to be part of the standard documentation?

This is something that will be defined by the ES-DOC initiative that is mentioned.

P14, first paragraph: I think another benefit would be to have a long-standing set of agreed metrics by which we could measure more systematically the progress across the modeling community in time. This would be analogous to the standardized WMO measures for NWP performance.

This is something that is discussed within the WGNE/WGCM diagnostics and metrics panel and not the topic of this paper. We note however that finding such a generic set of standard metrics may not be possible since the metrics will depend on the specific application. The community is actively working on identifying metrics that point to a model getting the response to changes in forcings correct. This is discussed in the paper under the topic of 'Emergent Constraints'.

P14, l28: It might be good to comment on the risks of modeling groups using this diagnostic set of measures to 'tune' their models to. This has the risks that we deliberately use compensating errors to optimize performance for certain metrics.

The evaluation tools here actually offer another opportunity since they include a broad set of diagnostics and metrics so tuning to a small set of metrics is avoided. The goal of this broad characterization is to spot compensating errors. This could be successfully demonstrated for example as part of CCMVal activity (see for example the SPARC CCMVal Report at http://www.sparc-climate.org/publications/sparc-reports/sparc-report-no5/). We have added a comment on this issue in the discussion.

Minor comments

P1, l30: 'more efficiently. . .' and more consistently (perhaps more important)?

Changed as suggested.

P1, l33: 'to develop evaluation tools' Do you really mean to gather evaluation tools?

Both. Tools are being developed by several groups and will be collected to run alongside the ESGF.

P5, l23: 'resulting in a database between 20 and 40 petabytes' should be 'resulting in a database of between 20 and 40 petabytes'

Changed as suggested.

P10, l3: 'A catalogue shall be created'. Is this a goal or will happen in CMIP6?

Changed for clarity.

P11, l11: 'identify strength' should read 'identify strengths'

Changed as suggested.

P11, l25: 'We point to Stouffer et al (2016) who summarize..' should read 'Stouffer et al (2016) summarize..'

Changed as suggested.

P12, l12: here you say ‘the focus is on’ as if this is always the case when comparing models with observations (as you describe at the start of the paragraph) but you are referring to just some specific examples. I think you need to say something more like ‘For many studies, the evaluation is limited to the end result of the combined effect . . . ‘

Changed as suggested.

P14, l14: ‘seem destined’ this sounds as if you think its wrong?

Changed.

P14, l21: ‘process understanding’ should read ‘process-level understanding’

Changed as suggested.

P14, l33: ’need encouragement for contributing..’ should read ’need encouragement to contribute..’

Changed as suggested.