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

Response to Anonymous Referee #3

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.

One important note at the beginning: the reviewer argues that the paper does not contain sufficient new material or findings to warrant publication as a scientific article and suggests the following possibility: "This makes me feel the paper should be submitted more as an ...opinion piece, or something analogous, to a journal where such articles more regularly appear (e.g. the AGU EoS Transactions is one example).." We were obviously aware that we are not presenting new scientific results but rather a viewpoint paper. We had therefore contacted the ESD chief editors before submission to determine whether they would find such a perspective and viewpoint paper on model evaluation in CMIP suitable for ESD. The chief editors all responded extremely positively and encouraged us to submit to ESD, so ESD welcomes such viewpoint papers. CMIP has a long and successful history of being useful to a wide range of climate scientists and its data has been important in all of the past IPCC and several National Climate Assessments as well as other important studies. The publication of our article in ESD will help in choosing related CMIP research and should help with the communication of the CMIP6 goals to a wide community. To consider the reviewer's comment, we have revised the abstract to make clearer to the reader from the start that this paper is not presenting new scientific results but rather provides a perspective and viewpoint on how a more systematic and efficient model evaluation can be achieved in CMIP. We also announce our intention to implement such a system for CMIP6.

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 stress that this paper is not documentation of specific evaluation tools which 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.

General Comments.

This article describes the ambition to develop a number of standardized evaluation tools for calculating performance metrics to inter-compare model simulations made within CMIP6 and future CMIP cycles. The paper further proposes that these evaluation tools will be developed to run directly on CMIP6 data, stored on a limited number of ESGF super data-nodes, allowing multi-model analysis to be performed "where the data resides" rather than requiring numerous data downloads to local machines prior to analysis.

The authors contend such evaluation tools will:

1. Speed up and make easier the analysis of CMIP multi-model ensembles.
2. Reduce duplication of effort across the international Earth system model analysis community.
3. As a result of (2), free up time within the research community to allow more effort to be expended on model development, thereby accelerating the improvement of Earth system models.
4. Reduce the amount of data download presently occurring with respect to CMIP data.
5. Allow modelling groups to do on-the-fly evaluation of developing models within their local model development cycles.
6. Lead to a significant improvement in the overall level of evaluation of Earth system models.

All of these potential benefits are highly laudable and should be supported.

Thanks for your support!

But, I am not totally convinced the path to these outcomes will be as smooth as envisaged in the paper. I expand on a few of these reservations later in this review.

While the 2 main aims outlined in the paper; (i) developing a community evaluation tool to calculate standard performance metrics on CMIP6 data and (ii) developing such tools as open-source code to run directly on data stored on the ESGF at the storage location, are both excellent aims, the paper does not really present any solid information on how this will be done, nor what type of metrics will be included or how users should go about either using the tools or contributing diagnostic code to them.

We would like to note that this paper is not a description of the evaluation tools themselves nor is the goal to define the metrics that are included in the individual tools. This is decided by the development teams of the tools. The current status of diagnostics and metrics included in the tools is documented in the corresponding publications we refer to. For example, there is a detailed description and user guide of the ESMValTool at http://www.geosci-model-dev.net/9/1747/2016/, for the NCAR CVDP at https://www2.cesm.ucar.edu/working-groups/cvcwg/cvdp, for the ILAMB tool at http://redwood.ess.uci.edu/mingquan/www/ILAMB/index.html, and for PMP at https://github.com/PCMDI/pcmdi_metrics. We do not see value in repeating all these diagnostics and metrics. This manuscript instead focuses on the entire workflow how these tools could be used to improve evaluation within CMIP and announces their application for CMIP6. To address this comment, we have expanded the description on the tools that we expect to apply to CMIP6 output and included some more examples on performance metrics and diagnostics from these tools that will be calculated from the CMIP6 models as soon as the output is submitted.

The paper is very aspirational in content, making a lot of quite reasonable observations of how the present mode of developing and analysing multiple ESMs is not optimal, but there are very few concrete details on what will be done to alleviate these problems. Rather there are a lot of generalized recommendations and, in some places, the paper actually reads more like a lobbying document, e.g. for more funding to be put into either ESGF or ESM development (p 3 lines 7-15). While I support both of these points, I don't think a scientific article is where such lobbying should appear.

We have removed the statement for more funding.

This makes we feel the paper should (i) be significantly reduced in content and (ii) submitted more as an opinion piece, or something analogous, to a journal where such articles more regularly appear (e.g. the AGU EoS Transactions is one example). Equally, the article could be repackaged as a forward-look/recommendation paper from the WGNE Climate Model Metrics Panel (which is referred to a number of times in the article). In its present form I do not feel the paper contains sufficient new material or findings to warrant publication as a scientific article.

This paper describes the existing state of infrastructure and Earth System model (ESM) evaluation strategies in CMIP5 and looks ahead toward CMIP6 and future phases of CMIP. It argues for the development and application of community evaluation tools to allow for routine evaluation of the CMIP models as soon as the output is submitted to the CMIP archive, and outlines the associated infrastructure needs. It then reviews some of the main associated scientific gaps and challenges the community needs to work on to develop relevant metrics for climate change that can guide future model developments and observations. Since the paper is not presenting new research results as the reviewer correctly says, we had contacted the ESD chief editorial board before submission and received confirmation that this article is fully suitable for submission to ESD, see response above. So rather than submitting to another journal, we have addressed the comments from the three reviewers and plan to submit a revised version with the goal that it will be accepted for publication in ESD.

We also do not want to follow the second suggestion by the reviewer to repackage the paper as a forward-look/recommendation paper from the WGNE/WGCM diagnostics and metrics panel. While a paper by the WGNE/WGCM diagnostics metrics panel would certainly be an important contribution to the CMIP process, it would have a different focus and author team.

To achieve this the paper should (i) be shortened in terms of general recommendations and (ii) contain a more actual examples of the type of analysis that can/will be performed with the tools and (iii) details of how the application of these tools directly on the ESGF will be realized.

We have shortened general recommendations and name the tools that we expect to apply for CMIP6. However, we note again that this paper is not about the definition of diagnostics and metrics to be applied to CMIP6 models. The diagnostics and metrics included in the various tools are described elsewhere in the literature and the web, see above.

On the last point (iii): we have outlined the general strategy that is currently envisaged to couple the tools to the ESGF. More details will need to be specified during the actual coupling process but we have expanded the text to consider the reviewer's comment.

More specific comments:

1. With respect to modelling groups using such generalized evaluation tools in their model development cycles.

This is possible, but it assumes groups do not already have such systems locally. Many do, the problem with them is they have been developed over a long time period, assuming a single (local) approach to model output, file naming and file formatting. This makes these analysis systems highly specialised to one model (or modelling institute) but also potentially quite efficient within that institute. The downside is that because institutes have (historically)

developed different approaches to model output/format/filenaming inter-comparison across models is not possible with these localised tools. This is the great benefit that CMIP has brought to the multi-model aspect of Earth system modelling, enforcing a single and common set of diagnostics, format and filenaming, allowing the potential for one evaluation tool to be able to analyse and inter-compare multi-model output. In itself a common output from all models is an enormous step forwards as is a single (all-encompassing) model evaluation tool that could be used by all modelling centres. To realize this aim requires either that modelling centres (i) modify their mode of standard (internal) output to follow CMIP conventions or (ii) the evaluation tools include some form of data converter to convert model output type X into CMIP compliant format, or (iii) with the developers of the evaluation tools, each modelling centre develops an interface between their preferred (local) model output and the required (CMIP-compliant) input to these evaluation tools. Option (i) may gradually happen, although the size of this task should not be underestimated. The authors might like to sound out a number of the larger modelling centres to gauge interest in these 3 options.

We disagree this assumes that the groups do no already have such systems locally. What we describe here is that modelling groups can make use of these additional evaluation tools that will also allow comparing to other CMIP models. The issues the reviewer raises in (i) to (iii) are all correct. Some models indeed move towards modifying their standard output to follow the CMIP conventions (the reviewer's point (i)) and for those that don't some tools are indeed providing data converters that the modelling groups can easily set up by following the given examples (see ESMValTool description for example). We have expanded the description to make this clear.

2. Standardized evaluation tools will free up time for more effort on model development.

If this was achieved it would be an excellent outcome, unfortunately I don't really see this being a natural result of a standardized evaluation tool. Such a tool might free up time for more in-depth evaluation of models across different processes and this in itself would be a good thing. The problem with a lack of model developers is that such work does not easily lead to publications and in the present mode of research funding it therefore becomes very difficult to successfully seek funds for purely a model development/ improvement activity. Furthermore, the required skills are not directly transferable; e.g. someone engaged in model analysis cannot just directly switch to model development, such a switch implies a significant change in tasks, required expertise and takes time to achieve. I feel there is a general misconception in the article as to the amount of effort that goes into converting and quality checking ESM output before it is published on the ESGF. Lines 30-33 on p.13 gives the impression modelling groups do this as a routine exercise. This is not the case and the effort to quality check and publish data onto the ESGF is very significant. Clearly, this level of effort may decrease in the future if models begin to produce CMIP-compliant output directly, but I imagine it will still remain a fair effort and may limit the ease by which groups "routinely" publish data onto the ESGF.

Nowhere in our manuscript have we said that a researcher should change his/her research focus. We solely consider the research topic on 'model evaluation'. We fully agree that model development is important, but this is not the subject of this paper. To address the comment, we now include a sentence on p.13 that comments on the efforts by the modelling groups to make the output CMOR compliant. We further state on p. 14, l. 2: In addition, the diagnostic tools could also be run locally by individual modelling groups to provide an initial check of the quality of their simulations before submission to the ESGF, thereby accelerating the model development/improvement process. Therefore, the standardised evaluation tools will

contribute to reducing also the effort of quality checking before the data is submitted to ESGF. The expansion from using individually in–house built evaluation tools only to using selected new tools does require an initial investment by modelling groups, which is only done so if they conclude it is worth the effort. If successful, this could lead to concrete knowledge transfer from analysts in the form of specialized codes that potentially expand beyond the expertise of any one modelling group. Ultimately a simplification of the workflow is envisaged that facilitates communication across modelling groups and sharing of the diverse expertize of the CMIP analysis community.

3. Standardized evaluation tools will lead to radically improved model evaluation efforts.

Where I do see such standardized evaluation tools contributing is in making somewhat quasi-regular (standard) analysis of multi-model performance more rapid and easier to produce. This could help areas such as IPCC assessments to progress more smoothly and reduce some of the burden on CLAs/LAs. It may be that standardized evaluation tools also leads to an overall improvement in ESM evaluation and/or more novel evaluation methodologies being developed. This will depend on the level of take-up of these tools in place of existing analysis tools already in use at various institutes. Such an uptake will be sensitive to; (i) the ease of use of these new tools, (ii) the ease by which new evaluation methods/metrics can be implemented into these tools, (iii) the flexibility of the tools in terms of what platforms they can run on and what software/libraries are assumed available and (iv) the quality of the output generated (e.g. in terms of publication quality graphics). Some examples of the chain of tasks from model output to publishable figures/results, as well as how one goes about running and implementing new diagnostics into these tools could be useful although I am not sure the latter point lends itself easily to a science article.

See responses above: these details are described in the individual tools and might also vary among the tools themselves. We refer to the available literature on details how to contribute diagnostics etc.

4. A more general concern I have with the use of performance metrics and these being (i) rapidly produced and compared across models on the ESGF and (ii) modelling groups potentially checking these metrics before submission of results, is the risk of models being tuned to “look good” on such metric figures. As the authors acknowledge, while performance metrics do carry useful information on model performance they can be misleading in that good metrics can occur through error compensation. Furthermore, if the metrics are not sufficiently broad in scope (e.g. variables, model domains and processes sampled, time and spatial scales sampled, importance in future feedback response etc) then models that are less ambitious/complete in terms of including important Earth system processes (in particular processes underpinning potential future feedbacks that might be less well constrained by observations, such as carbon cycle feedbacks) may look better on such metric plots than more ambitious/process complete models. This may lead to the opposite effect to the one aspired to, in that the degree of modelling ambition in terms of Earth system process-completeness, may be reduced if the resulting performance metrics show such models in a poor light relative to competitor models (that are more conservative). Such risks need to be acknowledged and carefully considered. The authors partially acknowledge this, for example they briefly make the point that model performance quality, as measured against present day observations/processes, does not necessarily equate to a model being reliable in terms of future projections. There is also a comment on this risk on p14, lines 9-10. Some more discussion as to how this risk will be mitigated seems important.

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.

5. Along similar lines to point 4, performance metrics normally lead to the ensemble mean (of a multi-model ensemble) being judged "the best model". This is because the metrics used are typically based on variables averaged over large spatial and temporal scales (e.g. continental and decadal). This is for done for good reasons (e.g. to average out natural variability and emphasize the evaluation of statistics rather than weather events). A problem arises when models are chosen (based on these performance measures) for driving impact models. Often it is the representation and change (or not) of extreme events (weather variability) that is crucial for impacts. Neither time and space, nor ensemble averaged, variables are suitable for use in impact models. Hence for these models there is an even greater risk that performance metrics (as presently developed) give an erroneous guide to model suitability. Impact models definitely should not use the ensemble mean of ESMs as input, even if this appears the "most accurate". These potential problems also need discussion.

We cannot prevent misuse of the models or model results, but we can do our best to objectively and comprehensively evaluate the models and to provide this information openly to the community. With the workflow described in this paper, this can be achieved much faster and in a more comprehensive manner than this was possible in CMIP5. Specifically for impact related research, some tools are developed further to include more impact relevant metrics and diagnostics, but what will be available in time for CMIP6 depends on the resources of the development teams. This is however not the focus of this paper - what we describe instead here is the workflow how this new framework can work and we announce its application to CMIP6 in this paper.

6. With respect to evaluating ESMs from the perspective of future projection reliability.

The authors introduce the concept of emergent constraints, which offers the potential to link the ability of models to represent key aspects of present day (observed) variability to the reliability of simulated future feedbacks. Unfortunately, the authors do not describe how such constraints will actually be used in model evaluation. This point could be expanded on to bring more scientific content.

We have expanded slightly on this topic, noting however that this is not a review on emergent constraints.

7. P11, lines 16-20: It is true that coupled ESMs cannot be compared in a temporal evolution sense against observations (i.e. simulated natural variability is not necessarily aligned temporally with reality) but the individual components of an ESM (e.g. atmosphere, ocean, land models) can all be run in a constrained setting and successfully compared to observations following a real-calendar time.

Discussion on this point has been added.

8. With reference to the statement (p3, lines 7-9). This is a worthwhile aim but it is never explained how this will be achieved.

We have changed this sentence to "**With this paper we aim to** attract input and development of established, yet innovative analysis codes from the broad community of scientists analysing CMIP results, including the CMIP6-Endorsed Model Intercomparison Projects (MIPs)."

9. With reference to the statement (p3, lines 11-12). Same as point 7.

We do not see what is wrong with this sentence, so kept it: "Our discussion here specifically addresses the crucial infrastructure requirements of community-tools for ESM analysis and evaluation and the reliance of those tools on infrastructure supporting ESM output and relevant Earth system observations."

10. With references to the statement p9, lines 20-21. Again this is true but it is just a wishful statement.

The analysis of the model output is already quite demanding in CMIP5 and will be even more challenging in CMIP6, given the expected growth in the amount of data. Parallelization is a necessary step for efficiently dealing with such data. Developments in this direction are already underway, we have therefore kept the statement.

11. Line 7. It is assumed all these institutional acronyms are known by all readers. Maybe they need defining? Likewise on p11, line 2, it is assumed everyone knows what WIP stands for.

Sorry about this omission. We spelled out the acronyms of the individual institutions. WIP is already defined on page 5 and then used as an acronym.

12. P 13, lines 4-7 and 8-11. Both true statements, but so what, this is known and accepted.

Here we broaden to impact studies and provide examples. Given the previous comment by this reviewer, we decided to keep this in order to make the reader aware (or remind) that more needs to be done in this area.