**Earth System
Dynamics**
Discussions

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

Anonymous Referee #2

Received and published: 26 July 2016

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.

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

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?

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. . .'

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.

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?

P12, first paragraph: What are the plans to detail the tuning process forCMIP6. Is this going to be part of the standard documentation?

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.

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.

Minor comments

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

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

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

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

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

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

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 . . . '

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

P14, l21: 'process understanding' should read 'process-level understanding'

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

Interactive comment on Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016.