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

Reply to Anonymous Referee #1

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.

The authors advocate the very laudable goal of developing a set of diagnostic tools that could be applied during future phases of the Coupled Model Intercomparison Project (CMIP). Unfortunately, however, the authors remain rather vague regarding some elementary design features of the proposed framework in which these tools should be implemented.

For example, the question whether the set of tools should be easily portable to users' platforms or whether it will be more or less tied to the ESGF framework is only answered somewhat implicitly on page 9, especially since "open source" clearly does not imply portability.

The tools shall run both in the ESGF structure and locally. The use of "open source" was inadequate in this context; we have revised the text to clearly state that users can download the source codes of the tools and run them on their local systems. The tools are designed to be portable and have been tested on a number of systems already, which is explicitly stated now in the paper.

Also, it would be interesting to know whether the proposed code can also be used to interpolate and/or extract data or whether it will only provide ready made plots.

The evaluation tools are not targeted towards flexibility, rather they produce as we say standard evaluation plots. Some options will still be left to the user choice, concerning for example regridding methods, confidence level in statistical tests, etc. The focus of the evaluation tools is however currently not on simply extracting data, although they could in principle be used to do this and we have added this to the text.

I think it would be good to answer these questions already at an early stage in the design of the tool.

We have expanded the text in Sect. 2.2 with details on the evaluation tools and provided some additional examples and figures. However, we note again that we are not describing the design of specific tools in this paper, rather the framework for a more routine model

evaluation in CMIP. Specifics of the design of the tools are detailed in the corresponding documentations of the tools. A few examples of available evaluation tools are given throughout the manuscript with the corresponding references that we refer to and available evaluation tools for CMIP6 are now also listed in a new Table 2.

I also think an overview of the key requirements/specifications (or maybe it is better to say "goals" instead of "requirements" since after all the authors are planning to provide another valuable and voluntary service to the community) in a list that might help the reader to understand what might be coming and also to serve as a guide for the further development of the framework would be helpful.

We have framed the discussion in terms of goals to deploy initially two available software package (also including other tools) that we plan to apply to CMIP6, to then build an infrastructure capable of supporting additional codes in the future. More details, in terms of achievable goals and actual specifications will depend on what the ESGF consortium can actually deploy in 2017 and beyond – which to the most part is still unknown.

Specific comments and questions:

1) could the proposed code also be used to interpolate and extract data or will it only provide ready made plots? Interpolating and extracting seems to be one capability that is needed anyway in designing the evaluation tool, and that by itself would be of great benefit to the user community. Would it make sense to construct the tool in order to eventually do data extraction and/or interpolation on the server side and plotting on the user side?

The priority at the moment is to provide the specific diagnostics already identified by the scientific community. Additional functionality, such as generic sub setting and interpolation might be desirable, but this is up to the teams that develop the tools. The ESMValTool, for example, allows for a separation of the data processing and graphic part. The results of each diagnostic are stored in a NetCDF file before drawing the plots. The users can therefore ignore the graphical part of the tool and make use of the NetCDF output and plot the results offline with a different software. Since this is quite a specific comment that is treated differently in different tools, we have not included these details in the text, but rather only say that in principle some of the tools can also be used to extract data, noting that this is not their main purpose at this stage.

2) in case a standard set of plots will be provided during CMIP, will the plots be archived permanently, will they be citable, and should they be copied and included in publications?

The plots produced by the standardize evaluation process outlined here will eventually be archived and become part of model documentation. In the meantime they can also be included in publications on model evaluation: since the tools that produce them are open source, the resulting plots are also effectively freely available. However, we would expect users to cite both software versions and technical papers produced by the tool developers to provide the formal provenance for the plots. This has been added to the text.

3) p. 3, line 26f: I find that there has been absolutely no lack of studies that have pointed out model deficits, and modelers do know where the problems are. One thing that has been missing is sufficient investments in climate model developers. My impression is that having a standard suite of evaluation tools will not do much to actually ameliorate the problem described in line 26f.

We highlight that model development is important. The evaluation activity proposed here can however provide guidance for model development which is mentioned. It can help focus a group of developers which span institutions for instance.

4) p. 11, line 7: could you either give a reference and/or describe what these "well defined standard interfaces" might look like? (Please see also my "additional comment" below).

This paragraph has been removed, but the fundamental point is that tools will need to declare inputs and outputs, via an interface – and an appropriate one doesn't yet exist, but is a named task in one of the many projects aiming at delivering this functionality.

5) would it be better to use just a single language (python) that can in principle could replace the potpourri of all the other languages in the ESMValTool? Please discuss.

What languages are used is a decision of the corresponding development teams of the tools. While there are certainly advantages of having tools all written in python or another language, experience in the project EMBRACE has shown that new diagnostics can be more readily added if different languages are used. Several existing diagnostic packages (such as the CVDP developed at NCAR and then included as part of the ESMValTool) are written in languages other than python. There are no mechanisms to fund the rewriting and supporting of the existing tools.

6) who should users contact if they are interested in contributing to the tool? Who will decide which diagnostics are to be included and which diagnostics are not to be included in the framework? Will there be "standard" and "user supplied" diagnostics? I understand that some strategy still needs to be developed, but it would also be nice to know what the possible outcomes of these developments might be.

The users should contact the respective PIs of the evaluation tools. What diagnostics are included in the tools is decided by the tool development teams. As mentioned in the paper, they correspond to " well-established parts of ESM evaluation that have demonstrated their value in the peer-reviewed literature". The distinction between standard and user-supplied is misleading in this context, since user-supplied diagnostic can also become part of a standard set after they have been included and tested in the respective tools.

7) p. 10, line 2: "users can however make substantial use of the tools by downloading the open source versions and by running them locally on their machines" -> this seems to me a major design requirement and it should be mentioned already at an early on in the manuscript. Is the code meant to be portable? Or will it tied to the ESGF servers?

As mentioned above, the tools are designed to be portable, i.e. not tied to ESGF. We changed the text and mention the portability issue early on in Sect. 2.2.

8) is it thought that individual users will eventually be able to adapt the code that they run on the ESGF machines? In other words, will users eventually operate their own version of the code on the servers in which they can adapt not only namelist settings, but also add diagnostics? Will it be possible to use additional data that might not be stored on the servers in these diagnostics? If yes, how could this be achieved? Would it make sense to do data extraction and/or interpolation on the server side and plotting on the user side as suggested in point 1 above?

Servers side extraction is already possible via OPENDAP, but as yet we have no specification of (or plans for) how to define a required interpolation operation on the server from a client. At the moment the easiest route is to bundle interpolation into the diagnostic itself – and yes, we expect it would be possible for users running their own versions of the code to add inputs and make configuration changes. Ideally where such changes extend beyond configuration, they would contribute their changes and extensions back to the community. The details of how that should be done will be tool dependent (and is therefore in the tool documentation). We think details of this sort go beyond what should appear in the paper itself.

9) p. 7, line 25: how are the groups supposed to use the tool during model development if it is run on the ESGF nodes? Will there be a stand-alone and an online version or will it just be one tool that can do both jobs? And also, how are you planning to deal with the dual requirements that the code should facilitate automatic processing while at the same time be user friendly, highly portable, and easily adaptable and expandable?

As now said in the manuscript, these tools should operate in both modes: integrated into the ESGF structure, and locally during model development. Concerning the dual requirement of facilitating automatic processing while being user-friendly, portable, adaptable and expandable, we do not see a contradiction here. The automatic processing of model data is based on the standardization of model output (CF/CMOR) which has been already successfully achieved within CMIP5 and other MIPs, such as ACCMIP and AEROCOM. Standardization of observations is a more complicated issue, but progress has been made within obs4mip and ana4mips, and some of the existing tools such as the ESMValTool or the CIS Tools (http://www.cistools.net/) proposed methods for dealing with the large variety of formats in the observational community. The portability aspect has been clarified in the text.

10) p. 2, line 26: I can not find any useful documentation of CMIP5 models under the web site given for ES-DOC. This reminds me of another project that has received funding for collecting meta-information on CMIP5 models, but that provided a poorly designed questionnaire and website to the model developers and as far as I can see has ultimately failed to be useful to users as well.

The http://es-doc.org provides access to model documentation through search and compare functions in the documentation folder. The information gathered through the CMIP5 questionnaire has been displayed with easy menu tables. Information can be displayed per model or compared between models. The website is already given in our manuscript that points the reader to further information on ES-DOC.

11) on p.8 in line 26 you are suggesting that the software will be able to acquire cache data from other servers. Will this cache data be kept for when one of the other servers is down? My experience has been that due to the distributed nature of ESGF it is sometimes very difficult to have access to all the data sets one wants to analyze at a given time. Would it be useful to cache processed (interpolated/extracted) data on the user side once the tool is opened up to users?

That's why we're suggesting that the data will be replicated to the supernodes, so the supernode tooling will have local copies of the data. Users running their own local copies will of course be able to manually cache their data as necessary.

12) p. 9 line 7f: "these supernodes have the necessary storage and computing resources". In line 17 it says: "requires the extension of current hardware" and in line 30, it says that the computing resources might not suffice for users to base their own analysis on this tool.

Thanks for spotting this. We rephrased these sentences for consistency. Our point here is that the existing supernodes are already providing large storage and computing resources, but may need to be extended to handle the larger CMIP6 data amount compared to CMIP5.

13) p. 10 line 7: "whereby new diagnostics developed by individual scientists can quickly and routinely" -> how will porting diagnostic tools be handled? Especially, what do the scientists have to do in order to port their diagnostic tools to the framework or to have them considered?

There will be two routes: (1) adding diagnostics to existing tools, and (2) adding new tools. With respect to (1): The ESMValTool for example offers a development environment that is open for the community to join. Both PMP and ESMValTool are also on github and are thus open to anyone contributing and sending a pull request. This has been clarified by extending the description of both tools. With respect to (2), the mechanism for new codes to be added into the system is another area where work is underway, but details are not yet available. It's likely this will be rudimentary at best for CMIP6.

14) p. 11 line 15: to me it seems important that the data version should be somehow documented. Yet, this is not mentioned here. As far as I can see, with CMIP5, finding out the version of a data set can only be achieved via sending a query with a checksum to an ESGF server. Maybe the users' interest in version numbers for the data sets has been underestimated? Also, will old versions of the data be stored so that one can reproduce results later without having to keep a local copy of the data? With CMIP5 this is not clear to me.

There are several questions here which go beyond the scope of our paper. However, it will never be practical to keep all the CMIP data, the volume would be prohibitive. That said, some of the supernodes are committed to keeping as much as is possible, and all are committed to keeping metadata about what data did exist, and how it differed. There will be a new errata system in CMIP6 to help this. The existing DOI system has been enhanced, and new work is in place to address workflow provenance to identify which data was used (even if it is not longer available) – however much of this is still under development. Please follow the CMIP6 Special Issue for further details that the WIP will provide, and more generally, the CMIP Panel website for up-to-date information on CMIP6.

15) p. 11, line 3: I think for all practical purposes, this would require either a new electronic data base format for citing the data or else summary doi's, which I don't think would work. I don't think that having 500 references to data sets each with its own doi would make much sense in something that might be printed on a printer, even if it would certainly be possible to automatically generate the corresponding list.

There has been considerable discussion of this in the community (e.g. http://home.badc.rl.ac.uk/lawrence/blog/2013/08/23/gavin%27s_proposal). The bottom line is that you're right, it won't be 500 references per paper.

16) p. 12, line 2f: "Model evaluations must take into account the details of any model tuning" -> how? Are you planning to archive output from all the untuned model versions? I don't understand what this sentence and also the following sentences might mean in practical terms. I also don't quite understand why this might be useful at all.

We simply mean that clear and concise information about what tuning went into setting up the model needs to be made available, so evaluations can be cognizant of any consequences. ES-DOC will be collecting some information to aid this process.

I think that it might be nice to have output for the same model tuned in different ways (maybe as "physics options" p1, p2,..."). But the sentences in the manuscript sounds like you are advocating the archiving of data for untuned models? If yes, please explain what you expect to learn from this. I do not think that archiving the data of untuned models within the framework of a model intercomparison projects makes much sense. Untuned models do not generally simulate a realistic radiation balance at the top of the atmosphere, and I think it is save to discard them in for the sake of model intercomparisons, especially since you are talking about comparisons with observations.

Sentence revised for clarity. See above.

17) p. 14, line 15: "requires ongoing maintenance" -> very good point. How can this be achieved?

A close collaboration between the ESGF system manager and tool developers is essential. To that end, the WIP has established the CDNOT: the climate data node operations team to directly serve such requirements for WCRP.

18) Fig 1: given this centralized approach, how can sufficient reliability and redundancy be achieved? Just recently, the ESGF nodes have been completely unavailable for several months.

The unavailability of ESGF was due to a hacker attack and not to infrastructural or technical issues. As pointed out in a recent letter to the community, “while ESGF cannot guarantee immunity from future dedicated hackers, the project has taken several steps to minimize the likelihood of a future security incident, and to recover much faster in case such an event should happen. At the same time, we have upgraded our infrastructure in many respects, to make it faster, more resilient, and more reliable.”
https://verc.enes.org/community/announcements/ESGFOperationalLettertotheCommunity.pdf

19) Notwithstanding my criticisms above, I do think that the ESGF people have on the whole done a great job and that their efforts have been extremely useful to the community. I also very much appreciate the initiative for the standard model evaluation tool, and I am confident that it will ultimately be very useful as well. I was also glad to have find other sources of the CMIP5 data while the ESGF servers were down.

Thanks!

Minor Points:

1) p. 12, line 27f: for an "emerging constraint", one needs a relationship between climate sensitivity and a model diagnostic that varies between models but can be constrained by observations. I think the formulation in the manuscript is not entirely clear.

Sentence revised for clarity.

2) p. 12., line 31: "might" -> could be considered more likely to

Changed as suggested.

3) p. 12, line 32: "A question raised ... " -> I don't understand what is meant here. Please re-formulate.

Sentence revised for clarity.

4) p. 12, line 33: "Moreover, ..." -> I think that this is a very good point.

Noted.

5) p. 13, line 3: "studies need not lead to contradictory results" -> I don't understand this sentence. Please re-formulate.

Sentence revised for clarity.

6) p. 13, line 21: in my opinion, one key question might be how easily adaptable this platform is by individual users

We are not sure what the reviewer means with "adaptation". The evaluation tools shall be designed in such a way that the user can easily extend them with additional diagnostic and metrics. In terms of customization of the available software, that will depend on the individual tools. As mentioned above, however the goal is not to provide fully-flexible analysis tools (such as CDO, NCO, CISTools etc.,) which already exist, but rather to share diagnostic codes reproducing well-established analyses for climate model evaluation.

7) p. 24, line 24: could you please specify what you mean by "revolutionary"?

The evaluation task will be transferred from the local modelling groups to the whole community, which will share agreed-upon methodologies and approaches and apply well-established and well-tested analysis, diagnostics and metrics to evaluate models. The observational data used for the evaluation will also be shared and, thanks for the obs4mips and ana4mips efforts, well documented. In our opinion, this will represent a revolutionary step forward to the approaches which has been followed so far.

Additional comment:

I am using an analysis framework in which placeholders such as "###(obs_data_path)###" are used for variables in analysis scripts (which are e.g. in ncl, R, python, etc) which are then inserted e.g. based on values specified in .xml files. In other words, the xml file and the analysis scripts are parsed by a preprocessor that then inserts whatever values are provided by the .xml file into the scripts (e.g. paths to data, etc.) before the scripts are automatically executed. I liked this more than the interface approach in which various interfaces are used for the various languages. In my diagnostic package, one can combine diagnostics into packages by specifying the package name in the .xml file and then run a package of scripts. I do, however, sometimes ask myself whether I should convert to a language such as python that would make the whole construct more uniform.

See our response above.

Technical comments:

p.1 line 4: Scientifically more research -> nice pleonasm

Changed.

p.9 line 7: was the list intended to be in alphabetical order?

Changed to alphabetical order.

p. 10, line 2: can -> could

Changed as suggested.

p. 13, line 18 this is -> this would be

Changed as suggested.

p. 13, line 19f in shared -> a shared

We think this sentence is correct and did not change it.

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.

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.

Towards improved and more routine Earth system model evaluation in CMIP

Veronika Eyring[1], Peter J. Gleckler[2], Christoph Heinze[3], Ronald J. Stouffer[4], Karl E. Taylor[2], V. Balaji[4,5], Eric Guilyardi[6,7], Sylvie Joussaume[8], Stephan Kindermann[9], Bryan N. Lawrence[7,10], Gerald A. Meehl[11], Mattia Righi[1], and Dean N.Williams[2]

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]Program for Climate Model Diagnosis and Intercomparison (PCMDI), Lawrence Livermore National Laboratory, Livermore, CA, USA
[3]Geophysical Institute, University of Bergen and Bjerknes Centre for Climate Research, Norway; Uni Climate, Uni Research AS, Bergen, Norway
[4]Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ, USA
[5]Cooperative Institute for Climate Science, Princeton University
[6]Institut Pierre Simon Laplace, Laboratoire d'Océanographie et du Climat, UPMC/CNRS, Paris, France
[7]National Centre for Atmospheric Science, University of Reading, United Kingdom
[8]Institut Pierre Simon Laplace, Laboratoire des Sciences du Climat et de l'Environnement, CNRS/CEA/UVSQ, Saclay, France
[9]Deutsches Klimarechenzentrum, Hamburg, Germany
[10]Centre for Environmental Data Analysis, STFC Rutherford Appleton Laboratory, United Kingdom
[11]National Center for Atmospheric Research (NCAR), Boulder, USA

Correspondence to: Veronika Eyring (veronika.eyring@dlr.de)

Abstract. The Coupled Model Intercomparison Project (CMIP) has successfully provided the climate community with a rich collection of simulation output from Earth system models (ESMs) that can be used to understand past climate changes and make projections and uncertainty estimates of the future. Confidence in ESMs can be gained because the models are based on physical principles and reproduce many important aspects of observed climate. ~~Scientifically more~~More research is required to identify the processes that are most responsible for systematic biases and the magnitude and uncertainty of future projections so that more relevant performance tests can be developed. At the same time, there are many aspects of ESM evaluation that are well-established and considered an essential part of systematic evaluation but are currently implemented ad hoc with little community coordination. Given the diversity and complexity of ESM model analysis, we argue that the CMIP community has reached a critical juncture at which many baseline aspects of model evaluation need to be performed much more efficiently ~~to enable a~~and consistently. Here, we provide a perspective and viewpoint on how a more systematic, open and rapid performance assessment of the large and diverse number of models that will participate in current and future phases of CMIP~~.~~ can be achieved, and announce our intention to implement such a system for CMIP6. Accomplishing this could also free up valuable resources as many scientists are frequently "re-inventing the wheel" by re-writing analysis routines for well-established analysis methods. A more systematic approach for the community would be to develop and apply evaluation tools that are based on the latest scientific knowledge and observational reference, are well suited for routine use and provide a wide range of diagnostics and performance metrics that comprehensively characterize model

behaviour as soon as the output is published to the Earth System Grid Federation (ESGF). The CMIP infrastructure enforces data standards and conventions for model output and documentation accessible via ESGF, additionally publishing observations (obs4MIPs) and reanalyses (ana4MIPs) for Model Intercomparison Projects using the same data structure and organization~~.~~ as the ESM output. This largely facilitates routine evaluation of the ~~models~~ESMs, but to be able to process the data automatically alongside the ESGF, the infrastructure needs to be extended with processing capabilities at the ESGF data nodes where the evaluation tools can be executed on a routine basis. Efforts are already underway to develop community-based evaluation tools, and we encourage experts to provide additional diagnostic codes that would enhance this capability for CMIP. At the same time, we encourage the community to contribute observations and reanalyses for model evaluation to the obs4MIPs ~~archive~~and ana4MIPs archives. The intention is to produce through ESGF a widely accepted quasi-operational evaluation framework for ~~climate models~~CMIP6 that would routinely execute a series of standardized evaluation tasks. Over time, as ~~the~~this capability matures, we expect to produce an increasingly systematic characterization of models, which, compared with early phases of CMIP, will more quickly and openly identify the strengths and weaknesses of the simulations. This will also expose whether long-standing model errors remain evident in newer models and will assist modelling groups in improving their models. This framework will be designed to readily incorporate updates, including new observations and additional diagnostics and metrics as they become available from the research community.

1 Introduction

High-profile reports such as the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5, IPCC (2013)) attest to the exceptional societal interest in understanding and projecting future climate. The climate simulations considered in IPCC AR5 are mostly based on Earth System Model (ESM) experiments defined and internationally coordinated as part of the World Climate Research ~~Program (WCRP) Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al. (2012)~~Programme (WCRP) Coupled Model Intercomparison Project Phase 5 (CMIP5, Taylor et al. (2012)). The objective of CMIP is to better understand past, present and future climate changes in a multi-model context. However, intelligent use of the simulations requires an awareness of their limitations. Therefore it is essential to systematically evaluate models with available observations (Flato et al., 2013). More generally, model evaluation and intercomparison provides a necessary albeit not sufficient perspective on the reliability of models, and also facilitates the prioritization of research that aims at improving the models.

Output from CMIP5 models is archived in a common format and structure and is accessible via a distributed data archive, namely the Earth System Grid Federation (ESGF[1]). The scientific contents of the models and the details of the simulations are further described via the Earth System Documentation (ES-DOC) effort[2]. This has enabled a diverse community of

[1] http://esgf.llnl.gov/
[2] http://es-doc.org

scientists ~~(over~~with more than 27,000 registered users (Williams et al., 2015) to readily search, retrieve and analyse these simulations. Since CMIP5, there has also been a large effort to provide observations and reanalysis products to end-users of CMIP results as part of the observations (obs4MIPs, Teixeira et al. (2014)) and reanalysis (ana4MIPs) for Model Intercomparison Projects. Together, these efforts have the potential to facilitate comparisons of model simulations with observations and reanalyses. However, the full rewards of the coordinated experiments and data standards have yet to be realized to further capitalize on the CMIP multi-model and observational infrastructure already in place (Williams et al., 2015).

Here, we ~~propose~~provide a ~~strategy~~perspective for developing standardized analysis procedures that could routinely be applied to CMIP model~~-~~ output at the time of publication on the ESGF, and we announce our intention to implement such a system in time for the sixth phase of CMIP (CMIP6, Eyring et al. (2016a)). The goal is to produce - along with the model output and documentation - a set of informative diagnostics and performance metrics that provide a broad, albeit incomplete, overview of model performance and simulation behaviour. ~~An important element of our strategy is~~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). The CMIP standard evaluation procedure should ~~comprise~~utilise open-source and community-based evaluation tools, flexibly designed in order to allow their improvement and extension over time. Our discussion here specifically addresses the crucial infrastructure requirements ~~of~~generated by such community-tools for ESM analysis and evaluation~~ and the~~, including how such requirements lead to reliance ~~of those tools~~ on the infrastructure supporting ESM output and relevant Earth system observations. An overarching theme is that if we are to capitalize on the ~~enormous~~ community effort devoted to model development, analysis, documentation and evaluation and if we are to fully exploit the value of coordinated multi-model simulation activities like CMIP, then further infrastructure development and maintenance will be needed. Given ~~CMIP6's~~the CMIP6 timeline and the complex and integrated nature of the infrastructure, it is expected that requirements will have to be satisfied by modifications and additions to the current infrastructure, rather than development and deployment of a completely new approach. ~~This~~The proposed infrastructure relies on conventions for data and~~ conventions~~ for recording model and experiment documentation that have been developed over the last two decades. Its backbone is the distributed data archive and the delivery system developed by the ESGF, which with CMIP5's success and WCRP's encouragement is increasingly being adopted by the climate research community. We hope the overview presented here inspires additional, focused efforts toward improved and more routine evaluation in CMIP.

We emphasize that routine evaluation of the ESMs cannot and is not meant to replace the cutting-edge and in-depth explorative analysis and research that makes use of CMIP output which will remain essential to close gaps in our scientific understanding. ~~Rather we suggest to make the well-established parts of ESM evaluation that have demonstrated their value in the peer reviewed literature more routine in order to leave more time for innovative research. For example, the current suite of evaluation procedures have generally not provided much guidance in reducing systematic biases, nor have they reduced the uncertainty in future projections (Stouffer et al., 2016).~~Rather we suggest to make the well-established parts of

ESM evaluation that have demonstrated their value in the peer-reviewed literature for example as part of the IPCC climate model evaluation chapters (Flato et al., 2013) more routine. This will leave more time for innovative research, for example on additional guidance in reducing systematic biases and on new diagnostics that can reduce the uncertainty in future projections.

Our assessment draws substantially on responses to a CMIP5 survey[3] of representatives from the climate science community and some additional related documents (Eyring et al., 2010; Mitchell et al., 2012). The summer 2013 survey was developed by the CMIP Panel, a sub-committee of the WCRP Working Group on Coupled Modelling (WGCM), which is responsible for direct coordination of CMIP. The scientific gaps and recommendations for CMIP6 that were identified through this community survey are summarized by Stouffer et al. (2016).

This paper is organized as follows. In Section 2 we argue for the development of community evaluation tools that would be routinely applied to CMIP model output as soon as it becomes available on ESGF, and we identify the associated software infrastructural needs. In Section 3, we discuss some of the scientific gaps and challenges that might be addressed through innovative diagnostic analysis that could be incorporated into future, more comprehensive evaluation tools. Section 4 closes with a summary and outlook.

2 Evaluation tools and corresponding infrastructure needs for routine model evaluation in CMIP

With the increasing complexity and resolution of ESMs, it is a daunting challenge to systematically analyse, evaluate, understand and document their behaviour. Thus, it is an especially attractive idea to engage a wide range of scientific and technical experts in the development of community-based diagnostic packages. The value of a broad suite of performance metrics that summarize overall model performance across the atmospheric, oceanic, and terrestrial domains is recognized by model developers, among others, as one way to obtain a broad picture of model behaviour. An obvious way to avoid duplication of effort across the model development and research community would be to adopt open source, community-developed diagnostic packages that would be routinely applied to standardized model output produced under common experiment conditions. The CMIP Diagnostic, Evaluation and Characterization of Klima (DECK) experiments and the CMIP historical simulations (Eyring et al., 2016a) lend themselves to this purpose.

The workflow for routinely analysing and evaluating the CMIP DECK and historical simulations is shown in Fig. 1. It utilizes community tools and relies on the ESGF infrastructure and relevant Earth system observations. The workflow assumes CMIP model output and observations are accessible in a common format on ESGF data nodes (Sect. 2.1), open-source software evaluation tools exist (Sect. 2.2), and that the existing ESGF infrastructure, which is now mainly a data archive, is enhanced with additional processing capabilities enabling evaluation tools to be directly executed on at least some

[3] http://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6

of the ESGF nodes (Sect. 2.3). Plans for making evaluation results traceable, well documented and visually rendered are also discussed (Sect. 2.4).

2.1 Access to CMIP model output and observations in common formats

The CMIP5 archive of multi-model output constitutes an enormous and valuable resource that efficiently enables progress in climate research. This diverse repository, in excess of 2 PB (see Table 1), of commonly-formatted climate model data also has proved valuable in the preparation of climate assessment reports such as the IPCC and in serving the needs of downstream users of climate model output such as impact researchers. The CMIP data format requirements are based on the Climate and Forecast (CF) self-describing Network Common Data Format (NetCDF) standards and naming convention[4] and tools such as Climate Model Output Rewriter (CMOR[5]). As a result, the CMIP model output conforms to a common standard with metadata that enables automated interpretation of file contents. The layout of data in storage and the definition of discovery metadata have also been standardized in the Data Reference Syntax (DRS[6]), which provides for logical and automated ways to access data across all models. This has enabled development of analysis tools capable of treating data from all models in the same way. and effectively independent of the platform on which they are executed.

The infrastructure supporting the publication of CMIP5 data was developed by the ESGF, which archives data accessible via a common interface but distributed among data nodes hosted by modelling and data centres. The CMIP5 survey noted that this first generation of a distributed infrastructure to serve the model data did not initially perform well, which retrospectively is not surprising given that it was a first major application of a distributed approach to archiving CMIP data and given the limited time and resources available for development and testing. Storing, testing, and delivering this data has relied on a distributed infrastructure developed largely through community-based coordination and short-term funding. This relatively fragile approach to providing climate modelling infrastructure will face even stiffer challenges in the future. Climate modelling and evaluation, which already involves management of enormous amounts of data, is a big data challenge confronted with demands for prompt access and availability (Laney, 2012). Unless we meet the challenge of dealing with increasing volumes of data, it will be difficult to routinely and promptly evaluate CMIP models.

Improvements in the functionality of the ESGF require a coordinated international undertaking. Priorities for CMIP are set by the WGCM Infrastructure Panel (WIP), and through ESGF's own governance structure these are integrated with demands from other projects. The individual, funded projects comprising ESGF ultimately determine what can be realized by volunteering to respond to the prioritized needs and requirements, and their efforts that are coordinated by ESGF working teams. The model evaluation activity advocated here depends on ESGF providing automated and robust access to all published model output and relevant observational data. The data made available under CMIP5 was about 50 times larger

[4] http://cfconventions.org
[5] https://pcmdi.github.io/cmor-site/
[6] http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

than under CMIP3. The data volume is expected to grow by another factor of 10-20 for CMIP6, resulting in a database of between 20 and 40 Petabytes, depending on model resolution and the number of modelling centres ultimately participating in ~~CMIP6~~the project (Table 1). The CMIP6 routine model evaluation activity discussed here will initially rely mostly on well-observed and commonly analysed fields, so this activity is not expected to increase the CMIP6 data request beyond the CMIP6-Endorsed MIP demands.

The convenience of dealing with CMIP output that adheres to well-defined standards and conventions is a major reason why the data have been used extensively in research. Another requirement of any model evaluation activity is well-~~-~~characterized observational data. Traditionally, observations from different sources have been archived and documented in a variety of ways and formats. To encourage a more unified approach, the obs4MIPs initiative ~~(Teixeira et al., 2014) has defined a set of technical~~(Teixeira et al., 2014) has defined a set of specifications and criteria for technically aligning observational data sets with CMIP model output (with common file format, data and metadata structure). Over 50 gridded datasets that conform to these standards are now archived on the ESGF alongside CMIP model output, and the archive continues to rapidly expand (Ferraro et al., 2015). Data users have enthusiastically received obs4MIPs, and the WCRP Data Advisory Council's (WDAC) has established a task team to encourage the project and provide guidance and governance at the international level. The expansion of the obs4MIPs project, with additional observational products directly relevant to Earth's climate system components and process evaluation, is a clear opportunity to facilitate routine evaluation of ESMs in ~~CMIP6~~CMIP. A sister project, ana4MIPs, provides selected fields well suited for model evaluation from major atmospheric reanalyses. The obs4MIPs protocol requires every dataset submitted to be accompanied by a technical note, which includes, for example, discussion of uncertainties and guidance as to aspects of the data product that are particularly relevant to model evaluation. Similar documentation efforts for observations specifically meant for use in model evaluation can be found at the National Center for Atmospheric Research (NCAR) climate data guide[7]. Ideally, standard technical documentation as defined by obs4MIPs will be adopted broadly by the international observational community and ~~perhaps even~~will be hosted alongside (or integrated with) the CMIP model and simulation standard documentation (ES-DOC). Additionally, there are proposals being considered to include non-gridded data in obs4MIPs (e.g., data collected by ground stations or during aircraft campaigns), and the possibility that auxiliary data such as land-sea masks, averaging kernels, and additional uncertainty data might also be provided. Whatever datasets are used for model evaluation, it will be important to determine the size of observational error relative to the errors in the models. One approach being developed is to provide ensembles of observational estimates, all based on a single sensor or product and generated by making many different choices of retrieval algorithms or parameters, all considered to be reasonable. The ~~hope~~goal is that obs4MIPs can be extended to better characterize observational uncertainty.

[7] https://climatedataguide.ucar.edu

2.2 Community-tools for Earth system model evaluation ready for CMIP6

There is growing awareness that community-shared software could facilitate more comprehensive and efficient evaluation of ESMs and that this could help increase the pace of understanding model behaviour and consequentially also the rate of model improvement. Here we highlight several examples of capabilities that are currently under development and relevant to the goal of developing routine ~~testing of CMIP simulations~~evaluation of CMIP simulations. Table 2 provides examples for existing diagnostic tools that can be used within CMIP6. Specifics of the design and the diagnostics included in these tools are detailed in the corresponding documentations of the tools that we refer to in the text and Table 2. Here, we only provide a brief overview and show a few examples of the type of plots that could be produced as soon as the model output is submitted to the ESGF. An up-to-date version of available tools will be catalogued by the WCRP's Working Group on Numerical Experimentation (WGNE)/WGCM Climate Model Diagnostics and Metrics Panel and maintained via an Earth System CoG site.

It is envisaged that well-established plots produced by the standardize evaluation process outlined here will eventually be archived and become part of model documentation. In the meantime they can also be included in publications on model evaluation: since the tools that produce them are open source, the resulting plots are also effectively freely available. However, we would expect users to cite both software versions and technical papers produced by the tool developers to provide the formal provenance for the plots.

2.2.1 Evaluation tools targeting the broad characterization of ESMs in CMIP6

Our initial goal is that two capabilities will be coupled to the ESGF to produce a broad characterization of CMIP6 DECK and historical simulations as soon as new model experiments are published on the CMIP6 archive: the Earth System Model Evaluation Tool (ESMValTool) that includes other model evaluation packages such as the NCAR Climate Variability Diagnostics Package (CVDP, see Section 2.2.2 and Table 2), and the PCMDI Metrics Package (PMP). The foundation that will enable this quasi-operation evaluation of the models to be efficient and systematic is the community-based experimental protocols and conventions of CMIP, including their aforementioned extensions to obs4MIPs and ana4MIPs (see Sect. 2.1). These evaluation tools are designed to exploit the data standards used in CMIP.

Both software packages are open source, have a wide range of functionalities, and are being developed as community tools with the involvement of multiple institutions. CMIP6 modelling groups and users of the CMIP6 data can make use of the evaluation results that are produced with these tools which will be made available to the wider community. They can also download the source code and can run the tools locally before submission of the results to the ESGF for an additional quality check of the simulations.

Here we summarize some aspects of these tools but refer to their respective documentation in the literature for further details.

- The ESMValTool (Eyring et al., 2016b) consists of a workflow manager and a number of diagnostic and graphical

output scripts. The workflow manager is written in Python, whereas a multi-language support is provided for the diagnostic and graphic routines. The ESMValTool workflow is controlled by a main namelist file defining the model and observational data to be read, the variables to be analyzed, and the diagnostics to be applied. The priority of the effort so far has been to target specific scientific themes focusing on selected Essential Climate Variables (ECVs), a range of known systematic biases common to ESMs, such as coupled tropical climate variability, monsoons, Southern Ocean processes, continental dry biases and soil hydrology-climate interactions, as well as atmospheric CO_2 budgets, tropospheric and stratospheric ozone, and tropospheric aerosols. ESMValTool v1.0 includes a large collection of standard namelists for reproducing the analysis of many variables across atmosphere, ocean, and land domains, with diagnostics and performance metrics focusing on the mean-state, trends, variability and important processes, phenomena, as well as emergent constraints. The collection of standard namelists allows to reproduce, for example, the figures from the climate model evaluation chapter of IPCC AR5 (Chapter 9, Flato et al. (2013)) and parts of the projection chapter (Chapter 12, Collins et al. (2013b)), a portrait diagram comparing the time-mean root mean square difference (RMSD) over different sub-domains as in Gleckler et al. (2008) and for land and ocean components of the global carbon cycle as in Anav et al. (2013). ESMValTool v1.0 also includes stand-alone packages such as the NCAR CVDP and the cloud regime metric developed by the Cloud Feedback MIP (CFMIP) community (Williams and Webb, 2009), as well as detailed diagnostics for monsoon, El Nino Southern Oscillation (ENSO), the Madden-Julian Oscillation (MJO). Example plots that illustrate the type of plots that will be produced with ESMValTool for CMIP6 are illustrated in Fig. 2, and we refer to the corresponding literature and the ESMValTool website (see Table 2) for full details.

- The PMP (Gleckler et al., 2016) includes a diverse suite of summary statistics to objectively gauge the level of agreement between model simulations and observations across a broad range of space and time scales. It is built on the Python and Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT, Williams (2014)), a powerful software tool kit that provides cutting-edge data management, diagnostic and visualization capabilities. Example plots produced with PMP are shown in Fig. 3. The first examines how well simulated sea-ice agrees with measurements on sector scales and demonstrates that the classical measure of total sea-ice area is often misleading because of compensating errors (Ivanova et al., 2016). The second highlights the amplitude and phase of the diurnal cycle of precipitation (Covey et al., 2016), and the third example is given by a simple “portrait plot” comparing different versions of the same model (Gleckler et al., 2016) in Atmospheric Model Intercomparison Project (AMIP) mode.

Both tools are under rapid development with a priority of providing a diverse suite of diagnostics and performance metrics for all DECK and historical simulations in CMIP6 to researchers and model developers suitable for use soon after each simulation is published on the ESGF. Since these tools are freely available, modelling groups participating in CMIP can additionally make use of these packages. They could choose, for example, to utilize the tools during the model development process in order to identify relative strengths and weaknesses of new model versions also in the context of the performance of other models or they could run the tools locally before publishing the model output to the ESGF. The tools are therefore

highly portable and have been tested across different platforms. The packages are designed to enable community contributions, with all results made highly traceable and reproducible. Collectively, the ESMValTool, PMP, and other efforts such as those mentioned in Section 2.2.2 below offer valuable capabilities that will be crucial for the systematic evaluation of the wide variety of models and model versions contributing to CMIP6.

2.2.2 Evaluation tools targeting specific applications or phenomena

Some other tools are being developed specifically to address targeted applications or phenomena. ~~The~~For example, the European Network for Earth System Modelling (ENES) portal [8] provides open source evaluation tools for specific applications that include chemistry-climate models (Gettelman et al., 2012), the aerosol component of ESMs, a satellite simulator package for satellite observations of ocean surface fluxes, and an objective recognition algorithm for properties of mid-latitude storms.

Other examples are the NCAR Climate Variability Diagnostics Package (CVDP), that has been designed to work on CMIP ~~simulations~~output and provides analysis of the major modes of climate variability in models and observations ~~(Phillips et al., 2014), and the International Land Modeling Benchmarking Project (ILAMB)~~(Phillips et al., 2014). The NCAR CVDP is also implemented as a stand-alone namelist in the ESMValTool. Fig. 4 shows a comparison of the CMIP5 models with observations for the Pacific Decadal Oscillation (PDO) to illustrate the kind of plots that can be produced with CVDP. Other available model evaluation packages that could be applied to CMIP6 output are the International Land Modeling Benchmarking Project (ILAMB), focusing on the representation of the carbon cycle and land surface processes in climate models via extensive comparison of model results with observations (Luo et al., 2012~~).~~), with a newer version under development. Still other ~~tools~~packages target model evaluation methods that are computationally demanding such as the parallel toolkit for extreme climate analysis (TECA, Prabhat et al. (2012)).

~~A few packages specifically target the broad and comprehensive characterization of CMIP DECK experiments and the CMIP historical simulations with the goal to run these tools at the ESGF as soon as the model output is published. The foundation that will enable this to be efficient and systematic is the community based experimental protocols and conventions of CMIP, including their extension to obs4MIPs and ana4MIPs (see Sect. 2.1). The evaluation tools can be designed to exploit the data standards used in CMIP. Examples of available tools that target routine evaluation in CMIP are the Earth System Model Evaluation Tool (ESMValTool, Eyring et al. (2016b)) and the PCMDI Metrics Package (PMP, Gleckler et al. (2016)). The ESMValTool includes diagnostics and performance metrics on the mean state, trends, variability and important processes, phenomena, and emergent constraints, including reproduction of the analysis in the IPCC AR5 model evaluation chapter (Chapter 9, Flato et al. (2013)) and parts of the projection chapter (Chapter 12, Collins et al. (2013b)). Version 1.0 of the ESMValTool also includes other packages such as the aforementioned NCAR CVDP, diagnostics for monsoon, El Nino~~

[8] https://verc.enes.org/models/support-service-for-model-users-1

~~Southern Oscillation (ENSO), the Madden Julian Oscillation (MJO) and the cloud regime metric developed by the Cloud Feedback MIP (CFMIP) community (Williams and Webb, 2009). The PMP is implementing a diverse suite of summary statistics to objectively gauge the level of agreement between model simulations and observations across a broad range of space and time scales (Gleckler et al., 2008). Both software packages are open source, have a wide range of functionality, and are being developed as community tools with the involvement of multiple institutions. Collectively, the ESMValTool, PMP, and other efforts such as those mentioned above offer valuable capabilities that will be crucial for the systematic evaluation of the wide variety of models and model versions contributing to CMIP6. Examples of such model – observation comparisons that will be produced for CMIP6 with the ESMValTool and PMP are shown in Fig. 2.~~

~~Since these tools are freely available, modelling groups participating in CMIP can additionally make use of these packages. They could choose, for example, to utilize the tools during the model development process in order to identify relative strengths and weaknesses of new model versions also in the context of the performance of other models or they could run the tools locally before publishing the model output to the ESGF. The wider community is being encouraged to contribute to the development of these tools by adding code for additional diagnostics. The free availability of the codes should facilitate this task and also help to increase code quality.~~

There is ~~slight~~some overlap in function between the ESMValTool and PMP and the other tools mentioned above, but efforts are underway to provide some coordination between these developing capabilities to reduce duplication of effort and to help ensure they advance in a way that best serves the CMIP modelling and research communities including the modelling groups themselves. ~~Nevertheless~~ In any case, encouraging a diversity of technical approaches and tools rather than a single one may at this stage be beneficial as it will provide experience that will help guide a more integrated approach in the longer term, perhaps as the community prepares for CMIP7 and beyond. Current testing with the same RMSD and ENSO metrics implemented in both the ESMValTool and PMP should inform such comparisons and reliability tests of the same scientific metrics incorporated into different technical frameworks.

~~To provide an overview of existing tools that target ESM evaluation for the community and the modelling groups, a central catalogue for model evaluation software is being populated by the WCRP's Working Group on Numerical Experimentation (WGNE)/WGCM Climate Model Diagnostics and Metrics Panel. An internationally coordinated strategy is required to document, organize and present results from these tools, and also to identify the metrics most relevant for climate change and impact studies (see also discussion in Sect. 3).~~

The wider community is being encouraged to contribute to the development of these tools by adding code for additional diagnostics. We refer to the literature of the individual tools for details how the development teams invite these contributions. The free availability of the codes should facilitate this task and also help to increase code quality. We stress again that the focus of these evaluation tools is on reproducing standard evaluation tasks and not on performing generic data processing task, such as extracting for example monthly or zonal means, reducing or regridding model data. Although they could be in principle used just for data processing, this is not their main goal and they may not include all the functionalities typically covered by pure data-processing tools.

2.3 Integration of evaluation tools in ESGF infrastructure

In order to connect multivariate results from multiple models and multiple observational data sets (Sect. 2.1) with tools for a quasi-operational evaluation of the CMIP models (Sect. 2.2), an efficient ESGF infrastructure is needed that can handle the vast amount of data and execute the evaluation tools. At the same time the workflow should be captured so that the evaluation procedure can be reproduced as new model output becomes available. This will allow changes in model performance to be monitored over a time frame of many years. Our expectation is that for CMIP6 the ESMValTool and PMP, with contributions from other efforts such as the NCAR CVDP and ILAMB packages, will be able to operate directly on the data served by the major ESGF data nodes. ~~This~~While it was and is possible to run analysis tools over the CMIP5 archive, it was difficult, error prone, and not widely done. The proposed new functionality ~~did not exist in CMIP5 and~~for CMIP6 is a step toward what should become a tighter integration of model analysis tools with data servers. This advancement will be particularly advantageous given the very large and complex CMIP data archive. Here we describe the necessary associated infrastructural changes that need to be made to enable this for CMIP6. As we provide an overview of the challenges emerging from the desire to move towards more routine evaluation of the models in future CMIP phases, it should be understood that actual implementation will require specification of many important technical details not addressed here.

It is envisaged that the evaluation tools will be executed at one or more of the ESGF sites that host copies (i.e. 'replicas') of most of the required CMIP datasets and the observations used by the evaluation tools. Although these replicas typically represent a significant subset of the data volume available on the ESGF, especially at the larger ESGF nodes, the complete replication of the entire CMIP model output at a single ESGF site cannot be achieved. As a consequence, some of the required CMIP model output used in the evaluation tools might still not be available even on the largest ESGF nodes. There are two practical solutions: (1) to distribute the processing of the evaluation tools at different ESGF nodes, and (2) to acquire and potentially cache data as needed for the evaluation tools. We regard the first option as not being practical in the CMIP6 timeframe~~.~~, but a possibly promising option on the long-term.

The second option that we envisage to be feasible for CMIP6 is schematically displayed in Fig. ~~3~~5. The evaluation tools are executed with specific user configurations (e.g., the ESMValTool namelists (Eyring et al., 2016b)). These user configurations also include the list of model and observational data to be analysed. Tools such as esgf-pyclient[9] and synda[10] exist which allow interrogation of local and distributed node data, and which could transfer the necessary data into either a cache or the ESGF replica storage. OPeNDAP[11] could also be used without the necessity for a cache. However, the workflow for managing this process does not yet exist and needs to be developed. Given the huge volumes of the ESGF data collections, it is realistic to assume that the requisite data will be maintained only at specific ESGF nodes where the

[9] https://pypi.python.org/pypi/esgf-pyclient
[10] https://github.com/Prodiguer/synda
[11] http://www.opendap.org

evaluation tools will be executed. It is therefore realistic that within CMIP6 the evaluation tools will be installed and operated on selected ESGF supernodes only, ~~which are hosted by seven climate data centers on four continents (Beijing Normal University, CEDA, DKRZ, LLNL, NCI, IPSL, and the University of Tokyo, see Williams et al. (2015)~~currently expected to be those hosted by seven climate data centers on four continents (Beijing Normal University (China), Centre for Environmental Data Analysis (CEDA, UK), Deutsches Klimarechenzentrum (DKRZ, Germany), Institut Pierre Simon Laplace (IPSL, France), Lawrence Livermore National Laboratory (LLNL, USA), National Computational Infrastructure (NCI, Australia), and the University of Tokyo (Japan), see Williams et al. (2015)). These supernodes ~~have~~will need to provide the necessary storage and computing resources and ~~are~~be integrated into the ESGF replication infrastructure, which optimizes data transport between core ESGF sites. Since it will take substantial time to replicate all output from the CMIP DECK and historical simulations to the supernodes (similar replications took months in CMIP5), we have recommended to the ESGF teams that the data used by the CMIP evaluation tools be replicated with higher priority. This should substantially speed up the evaluation of model results after submission of the simulation output to the ESGF. A prerequisite for this is that the evaluation tools provide an overview of the experiments, the subset of data from the CMIP6 data request, and the observations and reanalyses that are used. On the long-term (e.g.~~.~~., in time for CMIP7), more automatic and rapid procedures could be developed so that the evaluation tools could be run as part of the publication process of the model output.

Executing the evaluation tools directly alongside the ESGF may also ~~requires~~require the extension of the current hardware and software infrastructure to implement processing capabilities where the tools can be run. This infrastructure will need to include new interfaces to computers, and should allow for flexible deployment and usage scenarios since we can foresee application in a spectrum of possible environments discussed above. Given the large amount of data involved, parallelization of the data handling in the evaluation tools themselves needs to ~~efficiently process the large amount of data.~~be considered in order to efficiently process the large amount of data. A number of projects are either underway, or soon to start, which are targeting this part of the required infrastructure. A number of possible technical solutions are possible, but in Europe at least, it is likely that supernodes will deploy Web Processing Services[12] exposing the diagnostic codes as “capabilities” to new ESGF portals which exploit backend computing and access to the ESGF data nodes.

A coordinated set of community-based diagnostic packages will require standards and conventions to be adopted governing the analysis interface and the output produced by the diagnostic procedures. Clear documentation of the procedures and codes is required, as are standards for all key interfaces. Because working towards a community-based approach represents a shift in CMIP procedure, like the data standards themselves it will likely take considerable time and effort to establish agreed upon software standards. In the interim, substantial progress can be made by expert teams developing diagnostic tools if they follow a set of best practices and reasonable efforts are made to coordinate them where possible. During this period the different approaches available can be assessed, and further experience with them can help lead to advancing community-

[12] http://www.opengeospatial.org/standards/wps

based interfaces. During this time it will also be possible to experiment with different approaches to delivering the required computing within or alongside ESGF. Given that the amount of ~~computing~~ necessary and~~/or~~ affordable computing resources is not yet clear, it is likely that early ESGF ~~computing with the evaluation tools~~resources will be ~~used more~~allocated to ~~provide diagnostic products centrally performed by~~ the tool developers to provide diagnostics products centrally rather than ~~to provide~~for open computing ~~resources~~ on demand ~~for~~by multiple users. Multiple users ~~can~~could however still make ~~substantial~~profitable use of the tools by downloading the ~~open~~ source ~~versions~~codes and ~~by~~ running them ~~locally~~ on their own local systems. For more information regarding ESGF's infrastructure and progress towards computing and tool integration, please see the 2016 5th Annual ESGF Face-to-Face Conference Report[13].

In support of the ESGF infrastructure, a library will provide a system for indexing the output of the community-based diagnostics packages and automatically generate a user-friendly web interface for looking through the results (i.e., "viewer"). This library will integrate with an ESGF web service to provide a simple workflow for uploading diagnostics results to a server and share them with collaborators. Each diagnostics run will generate provenance data that will track data used for input, version of the community-based package, who ran the diagnostics and at which location, etc. This information would then be bundled with the output automatically and made available within the ESGF web service as well as in the local viewer.

To summarize: we will begin in the CMIP6 timeframe with the deployment of a subset of packages such as ESMValTool (which itself includes other well-known packages such as ~~CVDP) and PMP and run them on or alongside ESGF supernodes.~~NCAR CVDP) and PMP and run them on or alongside ESGF supernodes. Starting with available data in existing CMIP5 replica caches the evaluation package developments are tested at dedicated sites (some of the supernodes) and prepared for CMIP6. In parallel developments with respect to the supporting infrastructure (replication, cache maintenance, provenance recording, parallel processing) are starting. We expect this initial effort to spur developments toward a uniform approach to analytic package deployment. Eventually we aspire to put in place a robust and agile framework whereby new diagnostics developed by individual scientists can quickly and routinely be deployed on the large scale.

2.4 Data documentation, provenance, and visualization

For CMIP6, a specific goal will be to use the analysis tools currently being developed and to execute them on the ESGF once CMIP6 model output is published to provide a comprehensive evaluation of model behaviour. ~~On the long term such an evaluation could be part of the publication workflow and quality control (Sect. 2.3).~~ To document the process and to ensure traceability and reproducibility of the evaluation tool results, a catalogue shall be created, including all the relevant information about models, observations and versions of the tools used for evaluation along with information on the creation date of running the script, applied diagnostics and variables, and corresponding references. In this way a record of model

[13] http://esgf.llnl.gov/media/pdf/2015-ESGF_F2FConference_report_web.pdf

evolution and performance through different CMIP phases would be preserved and tracked over time (see Fig. ~~4~~6). In the long term, such an evaluation could be part of the publication workflow (Sect. 2.3).

The interpretation of the model evaluation results requires a precise understanding of a model's configuration and the experimental conditions. Although these requirements are not new for CMIP, the plan to carry out routine model evaluation increases the priority for enhancing documentation in these respects. In CMIP5 with over one thousand different model/experiment combinations, the first attempt was made to capture structured metadata describing the models and the simulations themselves (Guilyardi et al., 2013). Based upon the Common Information Model (CIM, Lawrence et al. (2012)), the European Metafor and US Earth System Curator projects worked together to provide tools to capture documentation of models and simulations. This effort is now continuing as part of the international ES-DOC activity, which defines common Controlled Vocabularies (CVs) that describe models ~~and~~, simulations, forcings and conformance to MIP protocols. Information from this structured representation of models and experiments can be extracted to provide comparative views of differences across models. Feedback from the CMIP5 survey indicates that improvements in methodology used to record model documentation consistent with the CIM are needed~~, and these~~. These developments are currently underway and will be implemented in time for CMIP6. With the focus here on model evaluation, we anticipate in the longer term expanding model documentation to include metrics of the model scientific performance, ~~which would~~to help characterize the simulations.

In addition, ~~a~~ proper data citation and provenance is required. Both model output and the observations serve as the basis for large numbers of scientific papers. It is recognized that sound science and due credit require: 1) that data be cited in research papers to give appropriate credit for the data creator, and 2) the provenance of data be recorded to enable results to be verified. Although these requirements were recognized in CMIP5, an automated system to generate appropriate data citation information and provenance information remained immature. For CMIP6 the WIP encourages concerted efforts in this area to meet the growing demand for formal scientific literature to cite all data sets used. Visualization of the evaluation diagnostics and metrics generated by the tools is also envisaged for CMIP6.

~~Visualization of the evaluation diagnostics and metrics generated by the tools is also envisaged. Similar to the processing capability supporting the execution of evaluation tools, standardized interfaces are required (Fig. 1). A visualization structure should be defined that can display evaluation results on a website or in form of a report, although a well defined standard interface will allow several visualization tools to coexist.~~

3 Current Earth system model evaluation approaches and scientific challenges

Establishing a more routine evaluation approach based on performance metrics and diagnostics that have been commonly used in ESM evaluation in the peer-reviewed literature will complement model evaluation analyses existing at each individual modelling group and will more rapidly allow modelling groups and users of CMIP output to identify ~~strengths~~strengths and weaknesses of the simulations in a shared and multi-model framework. This will constitute an

important step forward that will help uncover some of the main characteristics of CMIP models. However, in order to fill some of the main long-standing scientific gaps around systematic biases in the models and the spread of the models' responses to external forcings as evident for example in the large spread in equilibrium climate sensitivity in CMIP5 models (Collins et al., 2013b), additional research is required so that more relevant performance tests can be developed~~.~~ that could at a later stage be added to the community tools.

Unlike numerical weather prediction models, which can routinely be tested against observations on a daily basis, ESMs produce their own interannual variability and "weather", meaning that they cannot be compared with observations of a specific day, month or year, but rather only evaluated in a statistical sense over a longer, climate-relevant time period~~. In practice, confidence in ESMs relies on them being based on physical principles and able to reproduce many important aspects of observed climate (Flato et al., 2013). Assessing ESMs' performance is essential as they are used to understand historical and present day climate and to make scenario based projections of the Earth's climate over many decades and centuries. While significant progress has been made in ESM evaluation over the last decades, there are still many important scientific research opportunities and challenges for CMIP6 that will be addressed by the various CMIP6 Endorsed MIPs with the seven WCRP Grand Science Questions as their scientific backdrop (Eyring et al., 2016a). We point to Stouffer et al. (2016) who summarize the main CMIP5 scientific gaps and here we review and discuss briefly only those scientific challenges related specifically to model evaluation.~~, except when they are run in offline mode and nudged towards for example observed meteorology (e.g., Righi et al. (2015)). Confidence in ESMs relies on them being based on physical principles and able to reproduce many important aspects of observed climate (Flato et al., 2013). Assessing ESMs' performance is essential as they are used to understand historical and present-day climate and to make scenario-based projections of the Earth's climate over many decades and centuries. While significant progress has been made in ESM evaluation over the last decades, there are still many important scientific research opportunities and challenges for CMIP6 that will be addressed by the various CMIP6-Endorsed MIPs with the seven WCRP Grand Science Questions as their scientific backdrop (Eyring et al., 2016a). Stouffer et al. (2016) summarise the main CMIP5 scientific gaps and here we review and discuss briefly only those scientific challenges specifically related to model evaluation.

A critical aspect in ESM evaluation is that despite significant progress in observing the Earth's climate, the ability to evaluate model performance is often still limited by deficiencies or gaps in observations ~~(Collins et al., 2013a; Flato et al., 2013)~~(Collins et al., 2013a; Flato et al., 2013). Additional investment in sustained observations is required, while at the same time some improvements can be made by fully exploiting existing observational data and by more thoroughly taking into account observational uncertainty so that model performance can be advanced. In addition, the comparability of models and observations will need to be further improved for example through the development of simulators that take into account the features of the specific instrument (Aghedo et al., 2011; Bodas-Salcedo et al., 2011; Jöckel et al., 2010; Santer et al., 2008; Schutgens et al., 2016). Model evaluations must also take into account the details of any model tuning ~~(Mauritsen et al., 2012)~~(Hourdin et al., 2016; Mauritsen et al., 2012), which necessitate comprehensive information and documentation ~~of the~~about what tuning ~~approaches and observations used. In evaluating a~~ went into setting up the model ~~simulation, it is~~

~~important to consider~~, so evaluations can be cognizant of any consequences. ES-DOC will be collecting the ~~metrics used by the model developers, spanning the range from the parametrization level to holistic simulation to the methods used to initialize and force the model. The details of~~relevant information to aid this ~~tuning~~ process ~~will be documented for CMIP6.~~.

~~A wide variety of observational data sets, including, for example, the already identified Essential Climate Variables (ECVs, GCOS (2010)),~~A wide variety of observational data sets, including, for example, the already identified ECVs (GCOS, 2010)), can be used to assess the evolving climate state (e.g., means, trends, extreme events and variability) on a range of temporal and spatial scales. Examples include the evaluation of the simulated annual and seasonal mean surface air temperature, precipitation rate, and cloud radiative effects (e.g., Figs. 9.2-9.5 of Flato et al. (2013)). In evaluating the climate state, ~~the focus is on~~many studies are limited to the end result of the combined effects of all processes represented in CMIP simulations, and as determined by the prescribed boundary conditions, forcings and other experiment specifications.

While a necessary part of model evaluation, one limitation of this approach is that it rarely reveals the extent to which compensating model errors might be responsible for any realistic-looking behaviour, and it often fails to reveal the origins of model biases. To learn more about the sources of errors and uncertainties in models and thereby highlight specific areas that require improvements, evaluation of the underlying processes and phenomena is necessary. This approach hones in on the sources of model errors by performing process- or regime-oriented evaluations (Bony et al., 2006; Bony et al., 2015; Eyring et al., 2005; Waugh and Eyring, 2008; Williams and Webb, 2009). 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 (SPARC-CCMVal, 2010). Other targeted diagnostics can determine the extent to which specific phenomena (such as natural, unforced modes of climate variability like ENSO) are accurately represented by models (Bellenger et al., 2014; Guilyardi et al., 2009; Sperber et al., 2013).

Another longstanding open scientific question is the missing relation between model performance and future projections~~..~~ While the evaluation of the evolving climate state and processes can be used to build confidence in model fidelity, this does not guarantee the correct response to ~~changed forcing~~changing forcings in the future. One strategy is to compare model results against paleo-observations. The response of ESMs to forcings that have been experienced during, for example, the last Glacial Maximum or the Mid-Holocene can be assessed and compared with the observational paleo-record record (Braconnot et al., 2012; Otto-Bliesner et al., 2009)~~. Another increasingly explored option is to identify apparent relationships between climate sensitivity to anthropogenic forcing and some observable feature of the Earth's climate system. Such relationships are termed "emergent constraints". If physically plausible relationships can be found between, for example, changes occurring on seasonal or interannual time scales and changes found in anthropogenically forced climate change, then models that correctly simulate the seasonal or interannual responses might make more reliable projections (Cox et al., 2013; Fasullo et al., 2015; Hall and Qu, 2006; Sherwood et al., 2014; Wenzel et al., 2014; Wenzel et al., 2016). A question raised concerning the "emergent constraint" approach is whether we should trust the constraints given that they emerge from relationships uncovered in models themselves.~~ . Another increasingly explored option is to identify apparent relationships across an ensemble of models, between some aspect of long-term Earth system sensitivity and an observable trend or

variation in the current climate. Such relationships are termed "emergent constraints" referring to the use of observations to constrain a simulated future Earth system feedback. If physically plausible relationships can be found between, for example, changes occurring on seasonal or interannual time scales and changes found in anthropogenically-forced climate change, then models that correctly simulate the seasonal or interannual responses could be considered more likely to make more reliable projections. For example, Hall and Qu (2006) used the observable variation in the seasonal cycle of the snow albedo as proxy for constraining the unobservable feedback strength to climate warming, and Cox et al. (2013) and Wenzel et al. (2014) found a good correlation between the carbon cycle-climate feedback and the observable sensitivity of interannual variations in the CO_2 growth rate to temperature variations in an ensemble of models, enabling the projections to be constrained with observations. Other examples include constraints on the CO_2 fertilisation effect (Wenzel et al., 2016a), equilibrium climate sensitivity and clouds ~~Moreover, we must rule out the possibility that some apparent relationship might simply occur by chance or because the representation of the underlying physics is too simplistic. The key is whether the processes underlying the constraints are understood and simple enough to likely govern changes on multiple time scales (Caldwell et al., 2014; Karpechko et al., 2013; Klocke et al., 2011). In addition, different studies need not lead to contradictory results and rather should confirm each other. As the approach is fairly new, more work is needed to consolidate its applicability. Related to the topics on emergent constraints, more research is required to explore the value of weighting multi-model projections based on both model performance (e.g., Knutti et al. (2010)) and model interdependence (Sanderson et al., 2015), as well as the statistical interpretation of the model ensemble (Tebaldi and Knutti, 2007)~~(Fasullo et al., 2015; Fasullo and Trenberth, 2012; Klein and Hall, 2015; Sherwood et al., 2014), the Austral jet stream (Wenzel et al., 2016b), total column ozone (Karpechko et al., 2013), and sea ice (Mahlstein and Knutti, 2012; Massonnet et al., 2012). One should keep in mind, however, that the "emergent constraint" approach is based on relationships which are uncovered in models themselves. Moreover, we must rule out the possibility that some apparent relationship might simply occur by chance or because the representation of the underlying physics is too simplistic. The key is whether the processes underlying the constraints are understood and simple enough to likely govern changes on multiple time-scales (Caldwell et al., 2014; Karpechko et al., 2013; Klocke et al., 2011). In addition, different studies shall not lead to contradictory results but rather confirm each other. As the approach is fairly new, more work is needed to consolidate its applicability. ~~With the ever-expanding range of scientific questions and communities using CMIP output, model evaluation also needs to be expanded to develop more downstream, user-oriented diagnostics and metrics that are relevant for impact studies, such as statistics (e.g., frequency and severity) of extreme events that can potentially have a significant impact on ecosystems and human activities (e.g., Ciais et al. (2005)), water management (e.g., Sun et al. (2007)) or energy sector (e.g., Schaeffer et al. (2012)~~Related to the topic on emergent constraints, more research is required to explore the value of weighting multi-model projections based on both model performance (e.g., Knutti et al. (2010)) and model interdependence (Sanderson et al., 2015), as well as the statistical interpretation of the model ensemble (Tebaldi and Knutti, 2007).

With the ever-expanding range of scientific questions and communities using CMIP output, model evaluation also needs to be expanded to develop more downstream, user-oriented diagnostics and metrics that are relevant for impact studies, such as

statistics (e.g., frequency and severity) of extreme events that can potentially have a significant impact on ecosystems and human activities (e.g., Ciais et al. (2005)), water-management (e.g., Sun et al. (2007)) or energy sector (e.g., Schaeffer et al. (2012)) related variables.

In summary, there is a large demand for substantially more research in the area of ESM evaluation. The evaluation tools proposed here will support this by making established approaches more routine thus leaving more time to develop innovative diagnostics targeting ~~the~~ open scientific questions such as the ones discussed ~~here~~above.

4 Summary and discussion

~~We have advocated~~We provide a viewpoint here that advocates the development of community evaluation tools and the associated infrastructure that as part of CMIP6 will enable increasingly systematic and efficient ESM evaluation. This is an improvement over the existing CMIP infrastructure which mainly only supports access to the data in the CMIP database. The initial goal is to make available in shared, common analysis packages a fairly comprehensive suite of performance metrics and diagnostics, including those that appeared in the IPCC's AR5 chapter on climate model evaluation ~~(Flato et al., 2013)~~(Flato et al., 2013). Over time, an expanding collection of performance metrics and diagnostics would be produced for successive model generations. These baseline measures of model performance, ~~calculated~~applied at the time new model results are archived, would also likely uncover obvious mistakes in data processing and metadata information, thereby providing an additional level of quality control on output submitted to the CMIP archive. Routine evaluation of the ESMs cannot and is not meant to replace cutting-edge and in-depth explorative multi-model analysis and research, in particular within the various CMIP6-Endorsed MIPs. Rather, the routine evaluation would complement CMIP research by providing comprehensive baseline documentation of broad aspects of model behaviour. Furthermore, the use of the broad set of diagnostics offered by the tools highlighted here also reduces the risk that model performance is tuned to a single or limited set of metrics.

~~A more routine and systematic approach to model evaluation~~Our experience with past MIPs has been that initially the threshold effort required for standardizing data output (CMORization) is perceived as an obstacle by many groups, but time and experience has shown that this effort is well worth it. We have found that only standardized data gets widely used by the community, and the analysis of that data, especially by researchers outside the major modelling centres, has been central to CMIP's success. Once the output is collected in a common format, a more routine and systematic approach to model evaluation in CMIP has clear benefits for the scientific community. The recording of a set of informative diagnostics and metrics, along with publication of the model output itself and the model and simulation documentation, would enable anyone interested in CMIP model output to obtain a broad overview of model behaviour soon after the simulation has been published to the ESGF, and with a level of efficiency that was not possible before. The information would, for example, help the climate community to analyse the multi-model ensemble and would facilitate the comparison of models more generally. 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. Diagnostic tools like the ESMValTool (Eyring et al., 2016b)~~and the PMP (Gleckler et al., 2016) are now available that will form the starting point for routine evaluation of CMIP6 models.~~, the NCAR CVDP (Phillips et al., 2014), and the PMP (Gleckler et al., 2016) are now available to directly run on CMIP6 model output and observations, and will form the starting point for routine evaluation of CMIP6 models. An international strategy is required to organize and present results from these tools and to develop a set of performance metrics and diagnostics that are most relevant for climate change studies. The WGNE/WGCM Climate Model Diagnostics and Metrics Panel is in the process of defining such a strategy in collaboration with the CMIP Panel and the CMIP community. Such a strategy should also propose a way to mitigate the risk of restricting the evaluation of models to a predefined set of – possibly rapidly aging – metrics, however comprehensive~~.~~, or to a limited subset of models or model ensembles. It should for instance ensure that performance and process-based metrics definitions evolve as scientific knowledge progresses. This requires that the relevant science expert groups be involved in the development so that they can directly feed new metrics into the evaluation infrastructure.

Modelling centres now periodically produce and distribute data compliant with the CMIP data standards and conventions. These standards critically underpin the multi-model analyses that ~~seem destined to~~ play an ever-increasing role in supporting and enabling climate science. Development of an analysis and evaluation framework requires ongoing maintenance and evolution of that existing infrastructure. Observational and reanalysis data are also produced now in accordance with well-defined specifications and are stored on ESGF data nodes as part of obs4MIPs and ana4MIPs. The modelling, observational, and reanalysis communities should continue to nurture these efforts, and ensure that these datasets include documentation in form of technical notes, uncertainty information, and any special guidance on how to use the observations to evaluate models. This encapsulates ongoing efforts of the WCRP's data advisory council. The effort devoted to conforming data to well-defined standards should pay off in the long-term and lead to a better process-level understanding of the models and the Earth's climate system while fully exploiting existing observations. ~~Sustained funding for further developing, running, and maintaining the ESGF system and the development of community evaluation tools needs to be ensured.~~

With an eventual multi-model evaluation infrastructure established, we can look forward to revolutionary advancement in how climate models are evaluated. Specifically, results from a comprehensive suite of important climate characteristics should become available soon after simulations are made publicly available, with extensive documentation and workflow traceability. Moreover, modelling centres will be able to incorporate these codes into their own development-phase workflows to gain a more comprehensive understanding of the performance of new model versions. The infrastructure will enable groups of experts to develop and contribute both standard and novel analysis codes to community-developed diagnostic packages. The ongoing efforts to establish uniform standards across models and observations will lead to standard ways to develop and integrate codes across analysis packages and languages.

Successful realization of these plans will require our community to make a long-term commitment to support the envisioned infrastructure. Moreover, the wider climate research community will need encouragement ~~for contributing~~to contribute innovative analysis codes to augment the community-developed tools already being developed. The resulting suite of

diagnostic codes will constitute a CMIP evaluation capability that is expected to evolve over time and be run routinely on CMIP model simulations. At the same time, continuous innovative scientific research on model evaluation is required if metrics and diagnostics are to be discovered that might help in narrowing the spread in future climate projections.

Acknowledgements

This work was supported in part by the European Commission's 7th Framework Programme "InfraStructure for the European Network for Earth System Modelling Phase 2 (IS-ENES2)" project under Grant Agreement No 312979. VE acknowledges additional funding received from the European Union's Horizon 2020 ~~European Union's Framework Programme for Research~~research and ~~Innovation "Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach (CRESCENDO)" project~~innovation programme under ~~Grant Agreement~~grant agreement No 641816~~.~~ (CRESCENDO). VB acknowledges funding from an ExArch grant (NSF Award 1119308) and support by the Cooperative Institute of Climate Science from the National Oceanic and Atmospheric Administration, U.S. Department of Commerce (Award NA08OAR4320752). GM acknowledges support from the National Science Foundation and from the Regional and Global Climate Modeling Program of the U.S. Department of Energy's Office (DOE) of Biological & Environmental Research (Cooperative Agreement # DE-FC02-97ER62402). KET and PJG acknowledge support from the same DOE program under Lawrence Livermore National Laboratory as a contribution to the U.S. Department of Energy, Office of Science, Climate and Environmental Sciences Division, Regional and Global Climate Modeling Program under contract DE-AC52-07NA27344. The authors thank all representatives of the climate science community who responded to the CMIP5 survey that formed much of the basis for this and an accompanying paper on scientific needs for CMIP6. We thank Ingo Bethke, Björn Brötz, Tony Del Genio, Larry Horowitz, Martin Juckes, John Krasting, and Bjorn Stevens for helpful comments on an earlier version of this manuscript~~.~~, and Sébastien Denvil for many related discussions. Thanks to Luisa Sartorelli for her help with the figures and to Simon Read for helpful discussions and recommendations on the coupling of the evaluation tools to the ESGF. The statements, findings, conclusions, and recommendations are those of the authors and do not necessarily reflect the views of any government agency or department.

References

.

Aghedo, A. M., Bowman, K. W., Shindell, D. T., and Faluvegi, G.: The impact of orbital sampling, monthly averaging and vertical resolution on climate chemistry model evaluation with satellite observations, Atmos Chem Phys, 11, 6493-6514, 2011.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R., and Zhu, Z.: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J Climate, 26, 6801-6843, 2013.

Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., and Vialard, J.: ENSO representation in climate models: from CMIP3 to CMIP5, Clim Dynam, 42, 1999-2018, 2014.

Bodas-Salcedo, A., Webb, M. J., Bony, S., Chepfer, H., Dufresne, J. L., Klein, S. A., Zhang, Y., Marchand, R., Haynes, J. M., Pincus, R., and John, V. O.: COSP Satellite simulation software for model assessment, B Am Meteorol Soc, 92, 1023-1043, 2011.

Bony, S., Colman, R., Kattsov, V. M., Allan, R. P., Bretherton, C. S., Dufresne, J.-L., Hall, A., Hallegatte, S., Holland, M. M., Ingram, W., Randall, D. A., Soden, B. J., Tselioudis, G., and Webb, M. J.: How Well Do We Understand and Evaluate Climate Change Feedback Processes?, J Climate, 19, 3445-3482, 2006.

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., Watanabe, M., and Webb, M. J.: Clouds, circulation and climate sensitivity, Nature Geosci, 8, 261-268, 2015.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nat Clim Change, 2, 417-424, 2012.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophys Res Lett, 41, 1803-1808, 2014.

Ciais, P., Reichstein, M., Viovy, N., Granier, A., Ogee, J., Allard, V., Aubinet, M., Buchmann, N., Bernhofer, C., Carrara, A., Chevallier, F., De Noblet, N., Friend, A. D., Friedlingstein, P., Grunwald, T., Heinesch, B., Keronen, P., Knohl, A., Krinner, G., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J. M., Papale, D., Pilegaard, K., Rambal, S., Seufert, G., Soussana, J. F., Sanz, M. J., Schulze, E. D., Vesala, T., and Valentini, R.: Europe-wide reduction in primary productivity caused by the heat and drought in 2003, Nature, 437, 529-533, 2005.

Collins, M., AchutaRao, K., Ashok, K., Bhandari, S., Mitra, A. K., Prakash, S., Srivastava, R., and Turner, A.: CORRESPONDENCE: Observational challenges in evaluating climate models, Nat Clim Change, 3, 940-941, 2013a.

Collins, M., R. Knutti, J. Arblaster, J.-L. Dufresne, T. Fichefet, P. Friedlingstein, X. Gao, W.J. Gutowski, T. Johns, G. Krinner, M. Shongwe, C. Tebaldi, A.J. Weaver, and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013b.

Covey, C., Gleckler, P. J., Doutriaux, C., Williams, D. N., Dai, A., Fasullo, J., Trenberth, K., and Berg, A.: Metrics for the Diurnal Cycle of Precipitation: Toward Routine Benchmarks for Climate Models, J Climate, 29, 4461-4471, 2016.

Cox, P. M., Pearson, D., Booth, B. B., Friedlingstein, P., Huntingford, C., Jones, C. D., and Luke, C. M.: Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability, Nature, 494, 341-344, 2013.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, 2016a.

Eyring, V., Harris, N. R. P., Rex, M., Shepherd, T. G., Fahey, D. W., Amanatidis, G. T., Austin, J., Chipperfield, M. P., Dameris, M., Forster, P. M. D., Gettelman, A., Graf, H. F., Nagashima, T., Newman, P. A., Pawson, S., Prather, M. J., Pyle, J. A., Salawitch, R. J., Santer, B. D., and Waugh, D. W.: A Strategy for Process-Oriented Validation of Coupled Chemistry–Climate Models, B Am Meteorol Soc, 86, 1117-1133, 2005.

Eyring, V., Manton, M., Stammer, D., and Steffen, K.: Promoting the synergism models with observations and results of process studies, Discussion Paper for WCRP Modelling Coordination Meeting, 2010. 2010.

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S.,

Krasting, J., Kunert, D., Levine, R., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A. S., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., van Ulft, L. H., Walton, J., Wang, S., and Williams, K. D.: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, Geosci. Model Dev., 9, 1747-1802, 2016b.

Fasullo, J. T., Sanderson, B. M., and Trenberth, K. E.: Recent Progress in Constraining Climate Sensitivity With Model Ensembles, Current Climate Change Reports, 1, 268-275, 2015.

Fasullo, J. T. and Trenberth, K. E.: A Less Cloudy Future: The Role of Subtropical Subsidence in Climate Sensitivity, Science, 338, 792-794, 2012.

Ferraro, R., Waliser, D. E., Gleckler, P., Taylor, K. E., and Eyring, V.: Evolving obs4MIPs to Support the Sixth Coupled Model Intercomparison Project (CMIP6), B Am Meteorol Soc, doi: 10.1175/BAMS-D-14-00216.1, 2015. 2015.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T. F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (Ed.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

GCOS: Implementation Plan for the Global Observing System for Climate in Support of the UNFCCC, August 2010, 2010. 2010.

Gettelman, A., Eyring, V., Fischer, C., Shiona, H., Cionni, I., Neish, M., Morgenstern, O., Wood, S. W., and Li, Z.: A community diagnostic tool for chemistry climate model validation, Geosci. Model Dev., 5, 1061-1073, 2012.

Gleckler, P. J., Doutriaux, C., Durack P. J., Taylor K. E. , Zhang, Y., Williams, D. N., Mason, E., and Servonnat, J.: A more powerful reality test for climate models, Eos Trans. AGU, 97, 2016.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, 2008.

Guilyardi, E., Balaji, V., Lawrence, B., Callaghan, S., Deluca, C., Denvil, S., Lautenschlager, M., Morgan, M., Murphy, S., and Taylor, K. E.: Documenting Climate Models and Their Simulations, B Am Meteorol Soc, 94, 623-+, 2013.

Guilyardi, E., Wittenberg, A., Fedorov, A., Collins, M., Wang, C. Z., Capotondi, A., van Oldenborgh, G. J., and Stockdale, T.: Understanding El Nino in Ocean-Atmosphere General Circulation Models Progress and Challenges, B Am Meteorol Soc, 90, 325-+, 2009.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophys Res Lett, 33, 2006.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D.: The art and science of climate model tuning, B Am Meteorol Soc, 0, null, 2016.

IPCC: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Jöckel, P., Kerkweg, A., Pozzer, A., Sander, R., Tost, H., Riede, H., Baumgaertner, A., Gromov, S., and Kern, B.: Development cycle 2 of the Modular Earth Submodel System (MESSy2), Geosci Model Dev, 3, 717-752, 2010.

Karpechko, A. Y., Maraun, D., and Eyring, V.: Improving Antarctic Total Ozone Projections by a Process-Oriented Multiple Diagnostic Ensemble Regression, J Atmos Sci, 70, 3959-3976, 2013.

Klein, S. A. and Hall, A.: Emergent Constraints for Cloud Feedbacks, Current Climate Change Reports, 1, 276-287, 2015.

Klocke, D., Pincus, R., and Quaas, J.: On Constraining Estimates of Climate Sensitivity with Present-Day Observations through Model Weighting, J Climate, 24, 6092-6099, 2011.

Knutti, R., Abramowitz, G., Collins, Eyring, V., Gleckler, P. J., Hewitson, B., and Mearns., L.: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland0165-0009, 2010.

Laney, D.: The Importance of Big Data: A Definition, 2012.

Lawrence, B. N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R. W., Guilyardi, E., Lautenschlager, M., Morgan, M., Moine, M. P., Murphy, S., Pascoe, C., Ramthun, H., Slavin, P., Steenman-Clark, L., Toussaint, F., Treshansky, A., and Valcke, S.: Describing Earth system simulations with the Metafor CIM, Geosci. Model Dev., 5, 1493-1500, 2012.

Luo, Y. Q., Randerson, J. T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., Ciais, P., Dalmonech, D., Fisher, J. B., Fisher, R., Friedlingstein, P., Hibbard, K., Hoffman, F., Huntzinger, D., Jones, C. D., Koven, C., Lawrence, D., Li, D. J., Mahecha, M., Niu, S. L., Norby, R., Piao, S. L., Qi, X., Peylin, P., Prentice, I. C., Riley, W., Reichstein, M., Schwalm, C., Wang, Y. P., Xia, J. Y., Zaehle, S., and Zhou, X. H.: A framework for benchmarking land models, Biogeosciences, 9, 3857-3874, 2012.

Mahlstein, I. and Knutti, R.: September Arctic sea ice predicted to disappear near 2 degrees C global warming above present, J Geophys Res-Atmos, 117, 2012.

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M., and Barriat, P. Y.: Constraining projections of summer Arctic sea ice, Cryosphere, 6, 1383-1394, 2012.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, Journal of Advances in Modeling Earth Systems, 4, 2012.

Mitchell, J. F., Budich, R., Joussaume, S., Lawrence, B., and Marotzke, J.: Infrastructure Strategy for the European Earth System Modelling Community 2012-2022, ENES Foresight Document, 2012. 2012.

Otto-Bliesner, B., Schneider, R., Brady, E., Kucera, M., Abe-Ouchi, A., Bard, E., Braconnot, P., Crucifix, M., Hewitt, C., Kageyama, M., Marti, O., Paul, A., Rosell-Melé, A., Waelbroeck, C., Weber, S., Weinelt, M., and Yu, Y.: A comparison of PMIP2 model simulations and the MARGO proxy reconstruction for tropical sea surface temperatures at last glacial maximum, Clim Dynam, 32, 799-815, 2009.

Phillips, A. S., Deser, C., and Fasullo, J.: Evaluating Modes of Variability in Climate Models, Eos Trans. AGU, 95(49), 453-455, 2014.

Prabhat, Rubel, O., Byna, S., Wu, K. S., Li, F. Y., Wehner, M., and Bethel, W.: TECA: A Parallel Toolkit for Extreme Climate Analysis, Procedia Comput Sci, 9, 866-876, 2012.

Righi, M., Eyring, V., Gottschaldt, K. D., Klinger, C., Frank, F., Jöckel, P., and Cionni, I.: Quantitative evaluation of ozone and selected climate parameters in a set of EMAC simulations, Geosci. Model Dev., 8, 733-768, 2015.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties, J Climate, 28, 5150-5170, 2015.

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., Solomon, S., Free, M., Gleckler, P. J., Jones, P. D., Karl, T. R., Klein, S. A., Mears, C., Nychka, D., Schmidt, G. A., Sherwood, S. C., and Wentz, F. J.: Consistency of modelled and observed temperature trends in the tropical troposphere, International Journal of Climatology, 28, 1703-1722, 2008.

Schaeffer, R., Szklo, A. S., de Lucena, A. F. P., Borba, B. S. M. C., Nogueira, L. P. P., Fleming, F. P., Troccoli, A., Harrison, M., and Boulahya, M. S.: Energy sector vulnerability to climate change: A review, Energy, 38, 1-12, 2012.

Schutgens, N. A. J., Gryspeerdt, E., Weigum, N., Tsyro, S., Goto, D., Schulz, M., and Stier, P.: Will a perfect model agree with perfect observations? The impact of spatial sampling, Atmos. Chem. Phys. Discuss., 2016, 1-32, 2016.

Sherwood, S. C., Bony, S., and Dufresne, J. L.: Spread in model climate sensitivity traced to atmospheric convective mixing, Nature, 505, 37-42, 2014.

SPARC-CCMVal: SPARC Report on the Evaluation of Chemistry-Climate Models, V. Eyring, T. G. Shepherd, D. W. Waugh (Eds.). SPARC Report No. 5, WCRP-132, WMO/TD-No. 1526., 2010.

Sperber, K., Annamalai, H., Kang, I. S., Kitoh, A., Moise, A., Turner, A., Wang, B., and Zhou, T.: The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century, Clim Dynam, 41, 2711-2744, 2013.

Stouffer, R. J., Eyring, V., Meehl, G. A., Bony, S., Senior, C., Stevens, B., and Taylor, K. E.: CMIP5 Scientific Gaps and Recommendations for CMIP6, BAMS, accepted, 2016.

Sun, Y., Solomon, S., Dai, A., and Portmann, R. W.: How often will it rain?, J Climate, 20, 4801-4818, 2007.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of Cmip5 and the Experiment Design, B Am Meteorol Soc, 93, 485-498, 2012.

Tebaldi, C. and Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections, Philos T R Soc A, 365, 2053-2075, 2007.

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., and Potter, G.: Satellite Observations for CMIP5: The Genesis of Obs4MIPs, B Am Meteorol Soc, 95, 1329-1334, 2014.

Waugh, D. W. and Eyring, V.: Quantitative performance metrics for stratospheric-resolving chemistry-climate models, Atmos. Chem. Phys., 8, 5699-5713, 2008.

Wenzel, S., Cox, P. M., Eyring, V., and Friedlingstein, P.: Emergent constraints on climate-carbon cycle feedbacks in the CMIP5 Earth system models, Journal of Geophysical Research: Biogeosciences, 119, 2013JG002591, 2014.

Wenzel, S., Cox, P. M., Eyring, V., and Friedlingstein, P.: Projected land photosynthesis constrained by changes in the seasonal cycle of atmospheric CO2, Nature, doi: 10.1038/nature19772, 2016a. 2016a.

Wenzel, S., Eyring, V., Gerber, E. P., and Karpechko, A. Y.: Constraining Future Summer Austral Jet Stream Positions in the CMIP5 Ensemble by Process-Oriented Multiple Diagnostic Regression, J Climate, doi: 10.1175/JCLI-D-15-0412.1, ~~2016~~2016b. 673-687, ~~2016~~2016b.

Williams, D. N.: Visualization and Analysis Tools for Ultrascale Climate Data, Eos, Transactions American Geophysical Union, 95, 377-378, 2014.

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M., and Trenham, C.: A Global Repository for Planet-Sized Experiments and Observations, B Am Meteorol Soc, doi: 10.1175/bams-d-15-00132.1, 2015. 150904101253006, 2015.

Williams, K. and Webb, M.: A quantitative performance assessment of cloud regimes in climate models, Clim Dynam, 33, 141-157, 2009.

Table 1. Participation statistics for CMIP3, CMIP5 and estimated for CMIP6.

	CMIP3	CMIP5	CMIP6 (estimated)
Modelling groups	17	29	>30
Models	25	60	>60
Mean number of simulated years per model	~2800	~5500	~7500
Data volume (terabytes)	~36	>2,000	~20,000-40,000

Table 2. Examples of existing diagnostic tools that are available for CMIP6. Some of the tools focus on the broad characterization of the models and others on specific scientific applications. The examples marked with * are also included as separate namelists in the ESMValTool and will be applied to CMIP6 models together with the other diagnostics and performance metrics as the output is submitted to the ESGF. For details on diagnostics and performance metrics included in these tools and a description on how to contribute to the developments, we refer to the individual literature and websites given below.

Short Name of Tool	Long Name of Tool	Focus	Reference	Language	Further information
Coupled to the ESGF running locally and on specific supernodes in CMIP6 (see text and Figure 5)					
ESMValTool	Earth System Model Evaluation Tool	Diagnostics and performance metrics for the broad characterization of CMIP6 models	Eyring et al. (2016b)	Backend in python, diagnostics in NCL, R, python	http://www.esmvaltool.org/
PMP	PCMDI Metrics Package	Summary statistics and associated diagnostics for the broad characterization of CMIP6 models	Gleckler et al. (2016)	python	https://github.com/PCMDI/pcmdi_metrics
Examples of evaluation tools targeting specific applications or phenomena					
CREM*	Cloud Regime Error Metric	Clouds	Williams and Webb (2009)	python	http://cfmip.metoffice.com/
CLIMDEX*	Climate Extremes Indices	Core indices of extreme climate	https://pacificclimate.org/resources/software-library	R	https://cran.r-project.org/web/packages/climdex.pcic/
ILAMB	International Land Modeling Benchmarking Project	Carbon cycle and land surface processes i	Luo et al. (2012)	NCL, Python	http://www.ilamb.org/
NCAR CVDP*	NCAR Climate Variability Diagnostics Package	Major modes of climate variability	Phillips et al. (2014)	NCL	https://www2.cesm.ucar.edu/working-groups/cvcwg/cvdp
SOCCOM*	Southern Ocean Climate Model Atlas	Southern ocean	In preparation	Ferret	http://southernocean.arizona.edu/

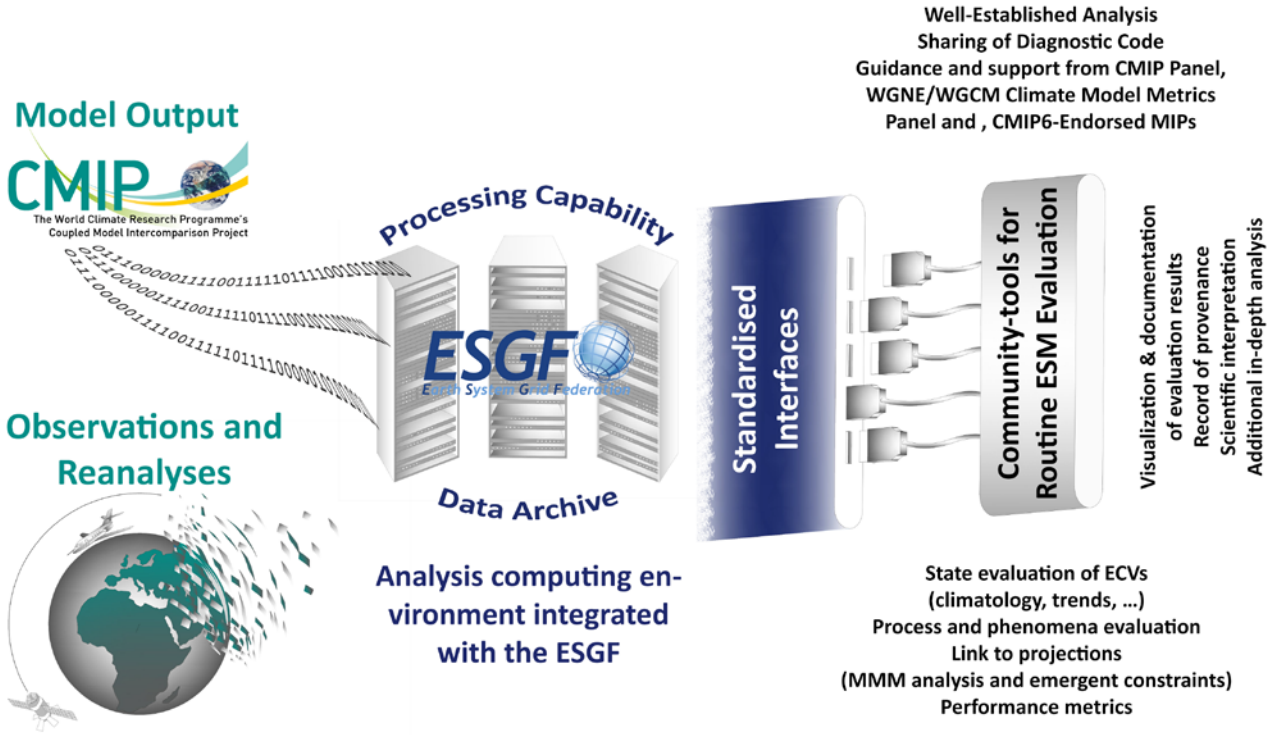
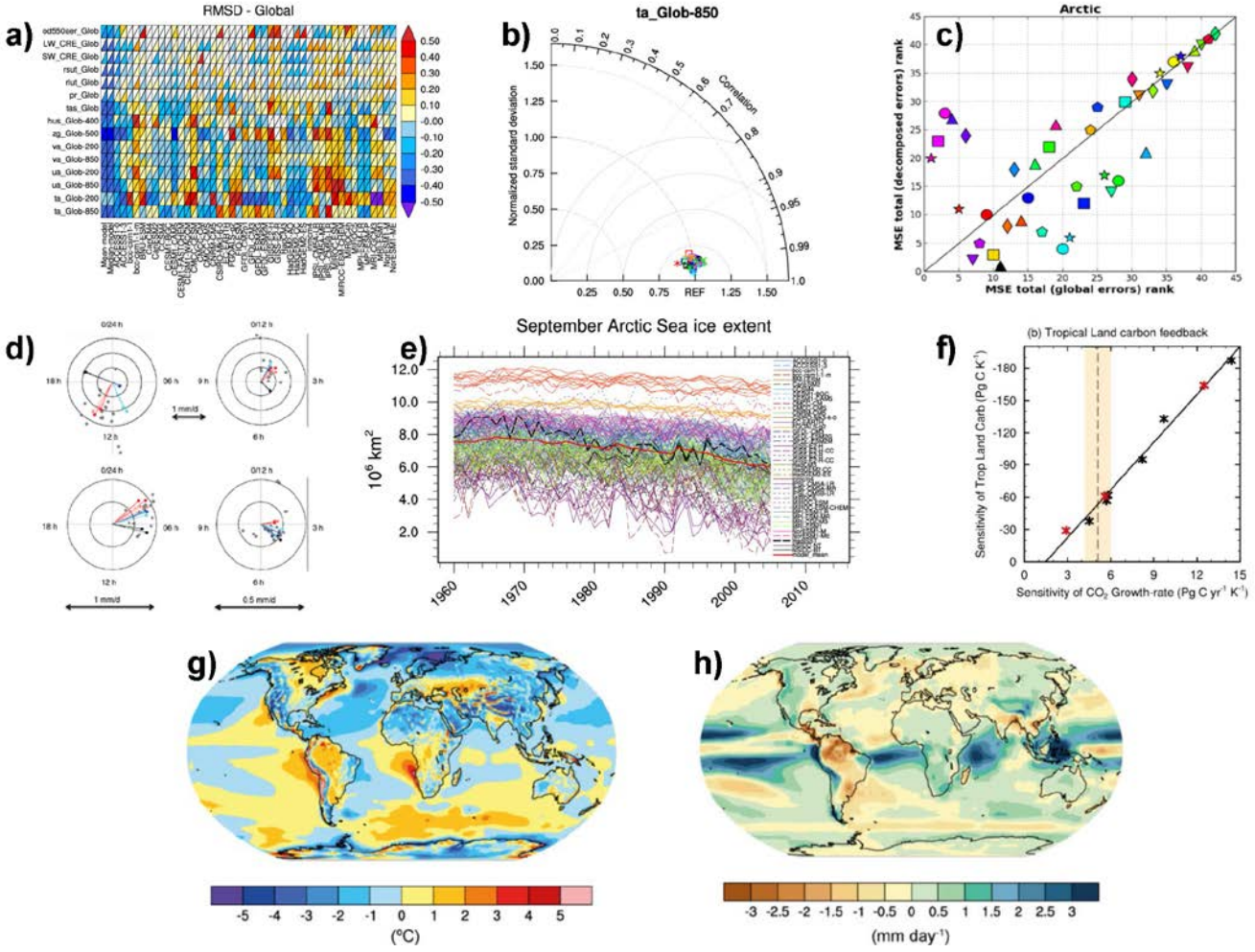


Figure 1: Schematic diagram of the workflow for routinely producing a broad characterization of model performance for CMIP model output using community evaluation tools that utilize relevant observations and reanalyses and rely on the ESGF infrastructure.

5


a)
RMSD - Global
b)
ta_Glob-850
Correlation
Normalized standard deviation
REF
c)
Arctic
MSE total (decomposed errors) rank
MSE total (global errors) rank
d)
e)
September Arctic Sea ice extent
10^6 km^2
f)
(b) Tropical Land carbon feedback
Sensitivity of Trop Land Carb (Pg C K^-1)
Sensitivity of CO2 Growth-rate (Pg C yr^-1 K^-1)
g)
(°C)
h)
(mm day^-1)

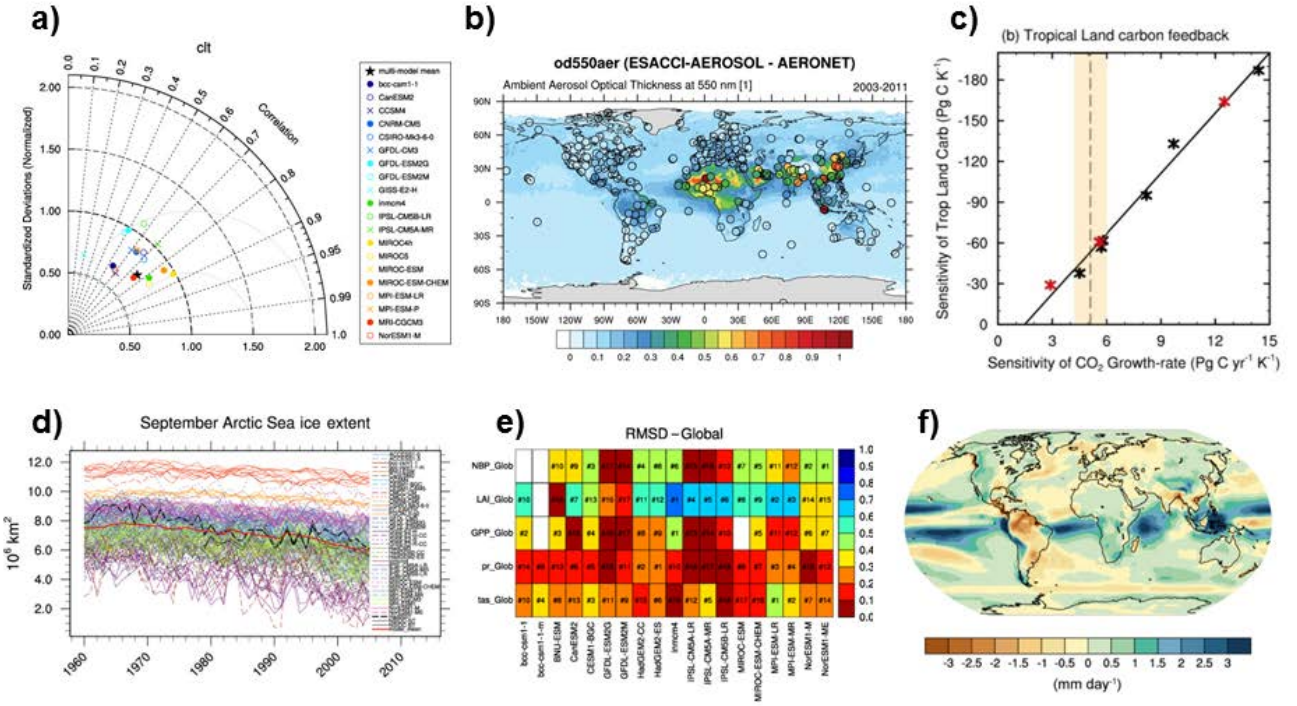
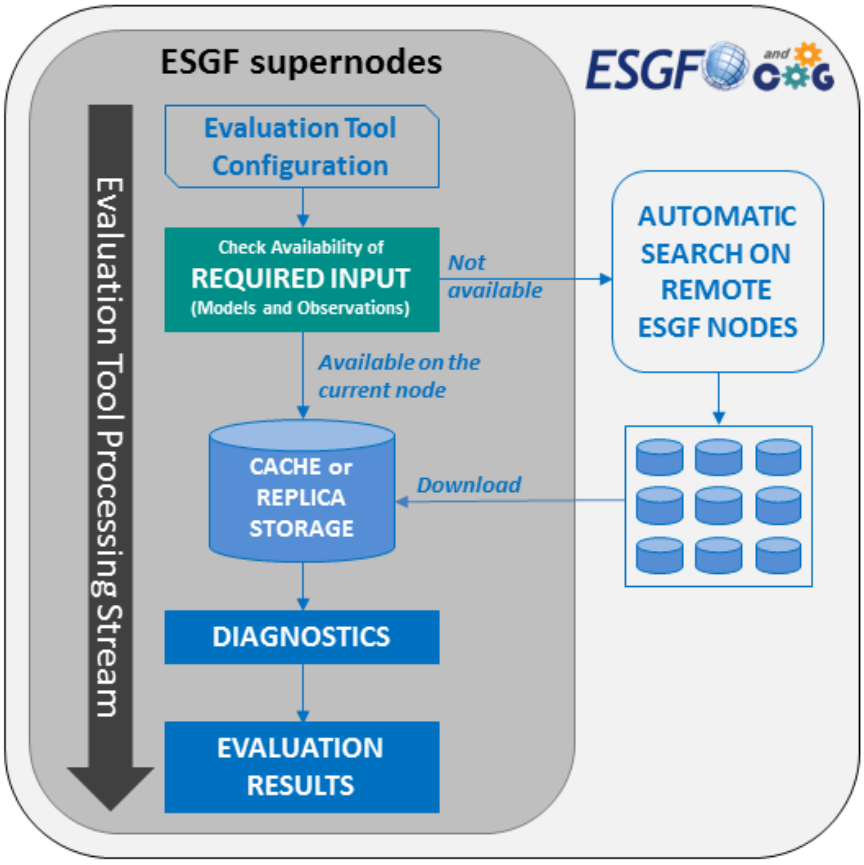


Figure 2: Examples of performance metrics and diagnostics that will be ~~routinely~~ calculated ~~on CMIP~~from CMIP6 models~~. Figures produced~~ with the ESMValTool ~~version 1.0 (Eyring et al., 2016b)~~(Eyring et al., 2016b) ~~and PMP (Gleckler et al., 2016).~~as soon as the output is submitted to the ESGF. (a) ~~Multi-model, multi-variable summary of relative root-mean-square error (RMSE) for CMIP5 models; (b) multi-model~~ Taylor diagram ~~for surface air temperature; (c) multi-model sector-scale sea-ice metrics; (d) diurnal precipitation metrics; (e) modelled and observed time series of September mean Arctic sea-ice extent; (f~~showing the 20-year annual average performance of CMIP5 models for total cloud fraction as compared to MODIS satellite observations; (b) aerosol optical depth from ESA-CCI satellite data (contours) compared with station measurements by AERONET (circles); (c) an emergent constraint on the carbon cycle-climate feedback (γ_{LT}) based on the short-term sensitivity of atmospheric CO_2 to interannual temperature variability (γ_{IAV}) in the tropics; (~~g, h) annual-mean surface air temperature (°C) and~~d) modelled and observed time series of September mean Arctic sea ice extent; (e) RMSD metric of several components of the global carbon cycle and (f) annual- mean precipitation rate (mm day^{-1}) bias from the CMIP5 multi-model mean compared to ~~ERA-Interim and~~ the Global Precipitation Climatology Project~~, respectively.~~.



~~Figure 3~~

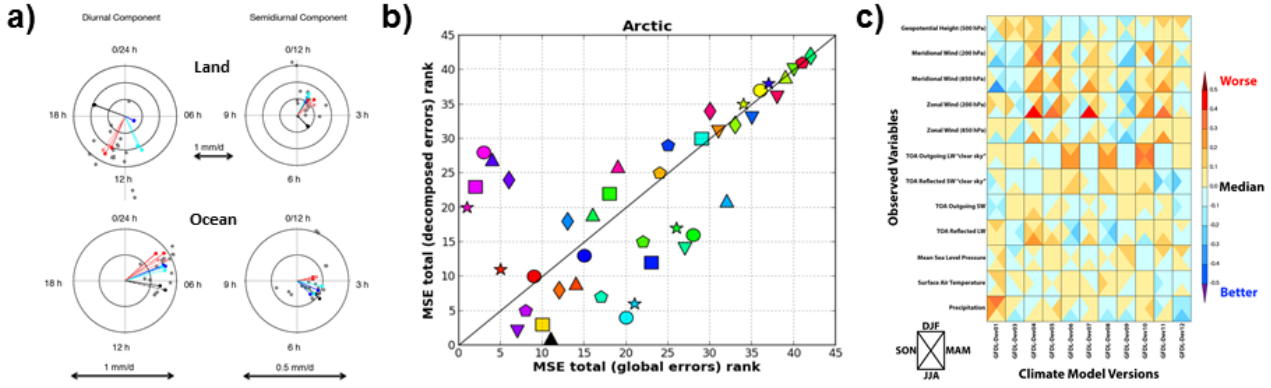


Figure 3: Examples of the kind of summary statistics that will be calculated from CMIP6 models with the PMP (Gleckler et al., 2016) as soon as the output is submitted to the ESGF: (a) harmonic dial plots of the amplitude and phase of Fourier components, after vector averaging over land and ocean areas separately; (a) model ranking using mean-square error (MSE) of the total sea-ice

area annual cycle versus an MSE constructed to include spatial information on sector scales; (c) relative error measures of different developmental tests of the Geophysical Fluid Dynamics Laboratory (GFDL) model in AMIP mode.

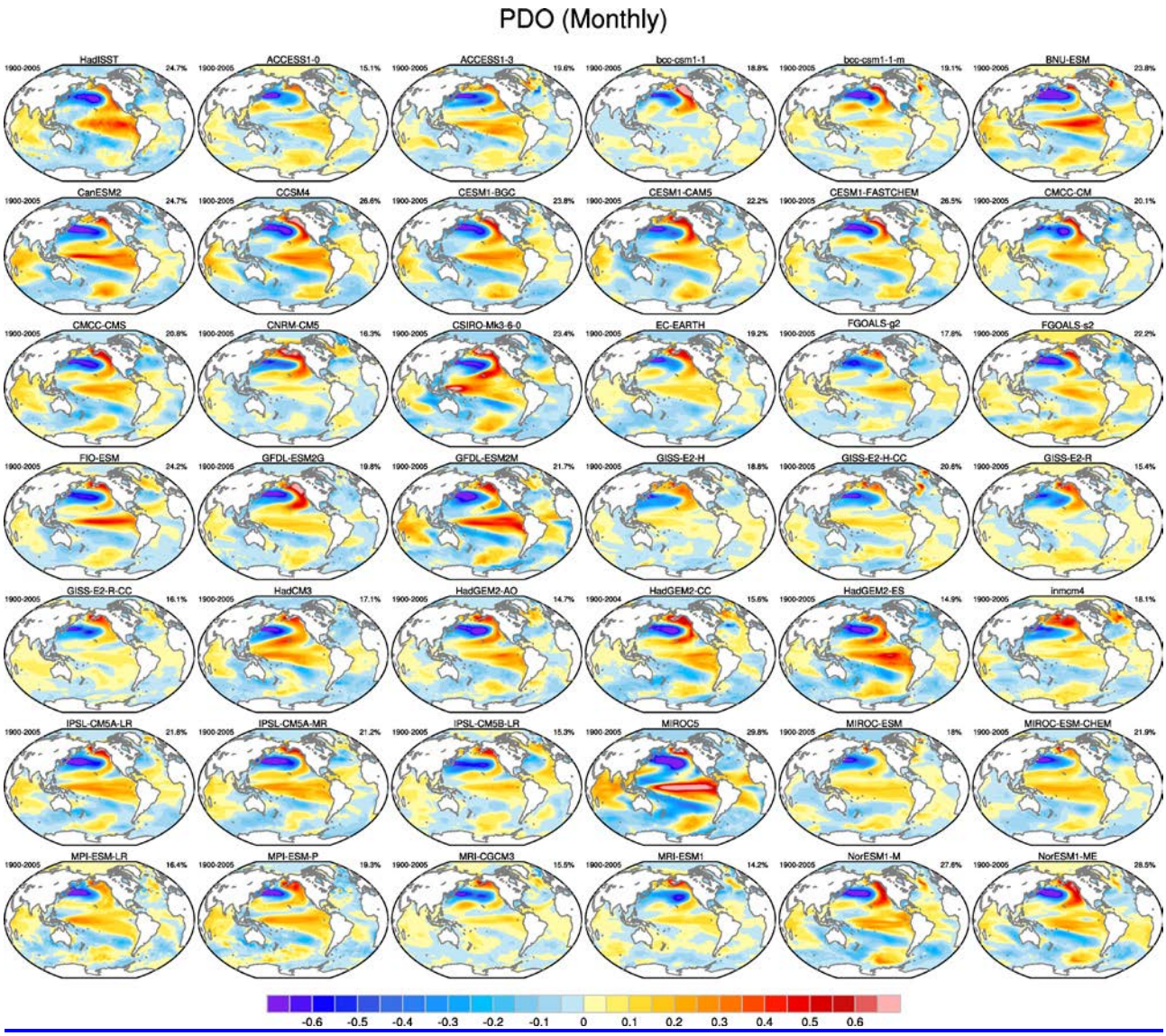


Figure 4. Examples of the kind of plots that will be calculated from CMIP6 models with the NCAR CVDP (Phillips et al., 2014) ESMValTool namelist as soon as the output is submitted to the ESGF The figure shows the PDO as simulated by 41 CMIP5 models (individual panels labelled by model name) and observations (upper left panel) for the historical period 1900-2005.

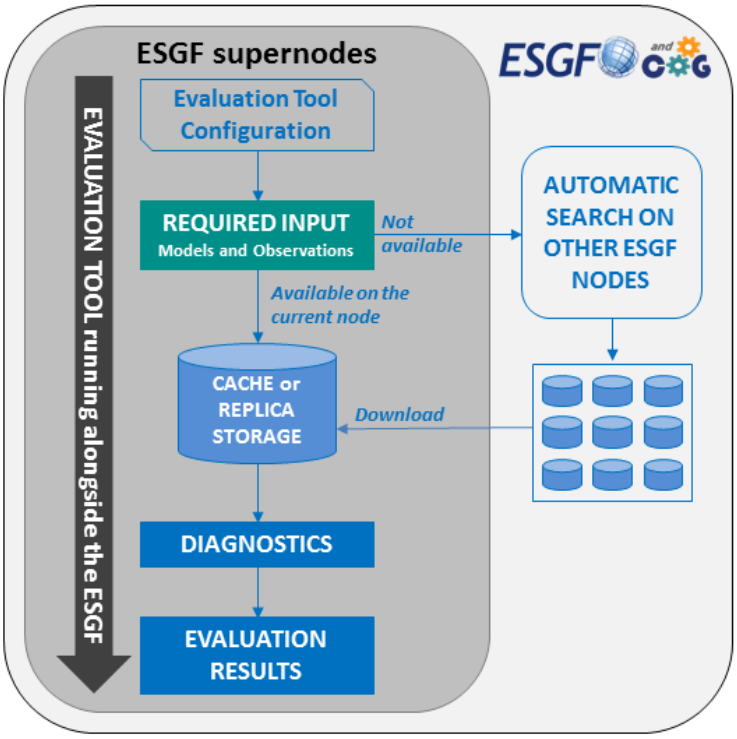
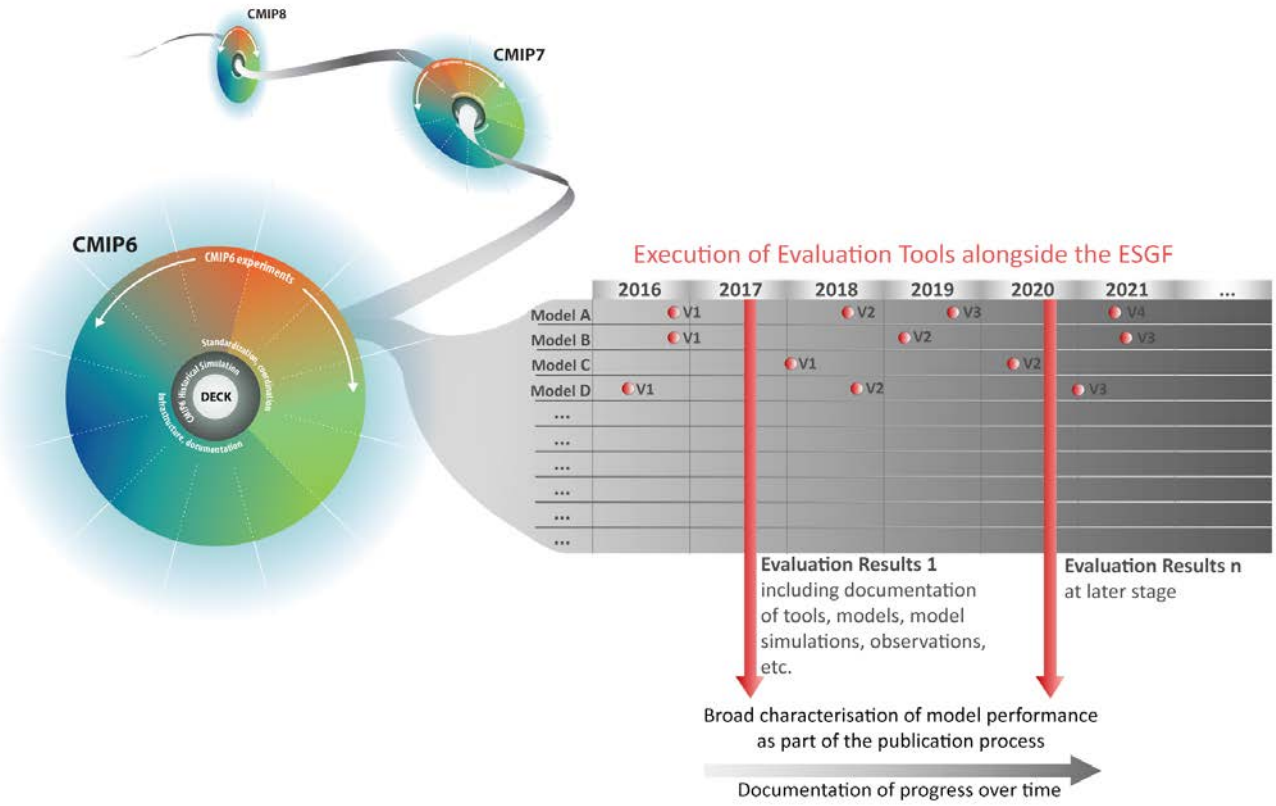


Figure 5: Schematic diagram of the envisaged evaluation tool processing stream for CMIP6. The schematic displays how the tools will be executed directly on ESGF supernodes exploiting optimized ESGF data organization and software solutions (see details in Sect. 2.3).


CMIP8
CMIP7
CMIP6
CMIP6 experiments
DECK
Execution of Evaluation Tools alongside the ESGF
2016
2017
2018
2019
2020
2021
...
Model A
Model B
Model C
Model D
V1
V2
V3
V4
Evaluation Results 1
including documentation
of tools, models, model
simulations, observations,
etc.
Evaluation Results n
at later stage
Broad characterisation of model performance
as part of the publication process
Documentation of progress over time

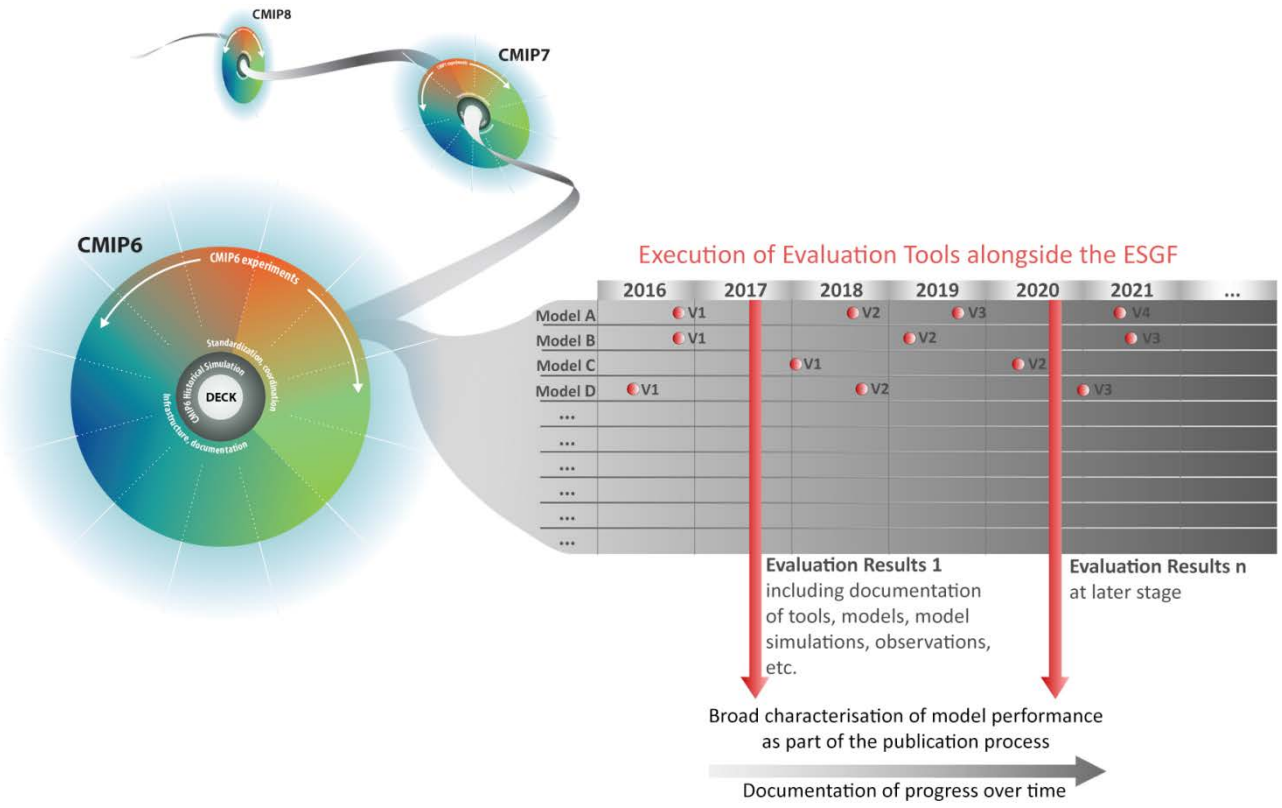


Figure 46: Schematic diagram of routine evaluation of CMIP DECK experiments and the CMIP historical simulations that is envisaged on the long-term. The evaluation tools would be executed quasi-operationally to produce a broad characterization of model performance as part of the ESGF publishing workflow, as could documentation and visual displays of the evaluation results with records of provenance. This example shows four different models that contribute with different model versions (V1-4) over time throughout CMIP6 and following phases.