Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

Earth System Dynamics Discussions
Open Access
EGU

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 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 systematic, open and rapid performance assessment of the large and diverse number of models that will participate in current and future phases of CMIP. 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 evaluation tools that 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 accessible via ESGF, additionally publishing observations (obs4MIPs) and reanalyses

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

(ana4MIPs) for Model Intercomparison Projects using the same data structure and organization. This largely facilitates routine evaluation of the models, 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 for model evaluation to the obs4MIPs archive. The intention is to produce through ESGF a widely accepted quasi-operational evaluation framework for climate models that would routinely execute a series of standardized evaluation tasks. Over time, as the 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)). 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 scientists (over 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.

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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 a strategy 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 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 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 community-tools for ESM analysis and evaluation and the reliance of those tools on 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 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 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).

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

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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 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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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.

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 than under CMIP3. The data volume is expected to grow by another factor of 10-20 for CMIP6, resulting in a database between 20 and 40 Petabytes, depending on model resolution and the number of modelling centres ultimately participating in CMIP6 (Table 1). The CMIP6 routine model evaluation activity 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

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

technical 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. 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 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 is that obs4MIPs can be extended to better characterize observational uncertainty.

2.2 Community-tools for Earth system model evaluation

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.

Some tools are being developed specifically to address targeted applications or phenomena. 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

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

simulations 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) 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). Still other tools 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 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 overlap in function between the ESMValTool and PMP and the other tools mentioned above, but efforts are underway to provide 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, 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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

community prepares for CMIP7 and beyond.

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).

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 operate directly on the data served by the major ESGF data nodes. This functionality did not exist in CMIP5 and 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 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.

The second option that we envisage to be feasible for CMIP6 is schematically displayed in Fig. 3. 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]

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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 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)). These supernodes have the necessary storage and computing resources and are 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 also requires 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.

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-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 is not yet clear, it is likely that early ESGF computing with the evaluation tools will be used more to provide diagnostic products centrally performed by the tool developers rather than to provide open computing resources on demand for multiple users. Multiple

[11] http://www.opendap.org

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

users can however still make substantial use of the tools by downloading the open source versions 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[12].

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. 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 evolution and performance through different CMIP phases would be preserved and tracked over time (see Fig. 4).

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. 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 are currently underway. With the focus here on model evaluation, we anticipate in the longer term expanding model documentation to include metrics of model performance, which would 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

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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. 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 strength 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.

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.

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). 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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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), which necessitate comprehensive documentation of the tuning approaches and observations used. In evaluating a model simulation, it is important to consider 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 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)), 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 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; SPARC-CCMVal, 2010; Williams and Webb, 2009). Other targeted diagnostics can determine the extent to which specific phenomena (such as natural, unforced modes of climate variability) 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 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. Moreover, we must rule out the possibility that some apparent relationship might simply

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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).

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 discussed here.

4 Summary and discussion

We have advocated 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). Over time, an expanding collection of performance metrics and diagnostics would be produced for successive model generations. These baseline measures of model performance, calculated 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.

A more routine and systematic approach to model evaluation 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, 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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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. 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. 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 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 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 Horizon 2020 European Union's Framework Programme for Research and Innovation "Coordinated Research in Earth Systems and Climate: Experiments, kNowledge, Dissemination and Outreach (CRESCENDO)" project under Grant Agreement No 641816. 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. 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.

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.

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.

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.

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.

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.

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.

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.

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., 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. 673-687, 2016.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

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.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

Earth System Dynamics Discussions Open Access EGU

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

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

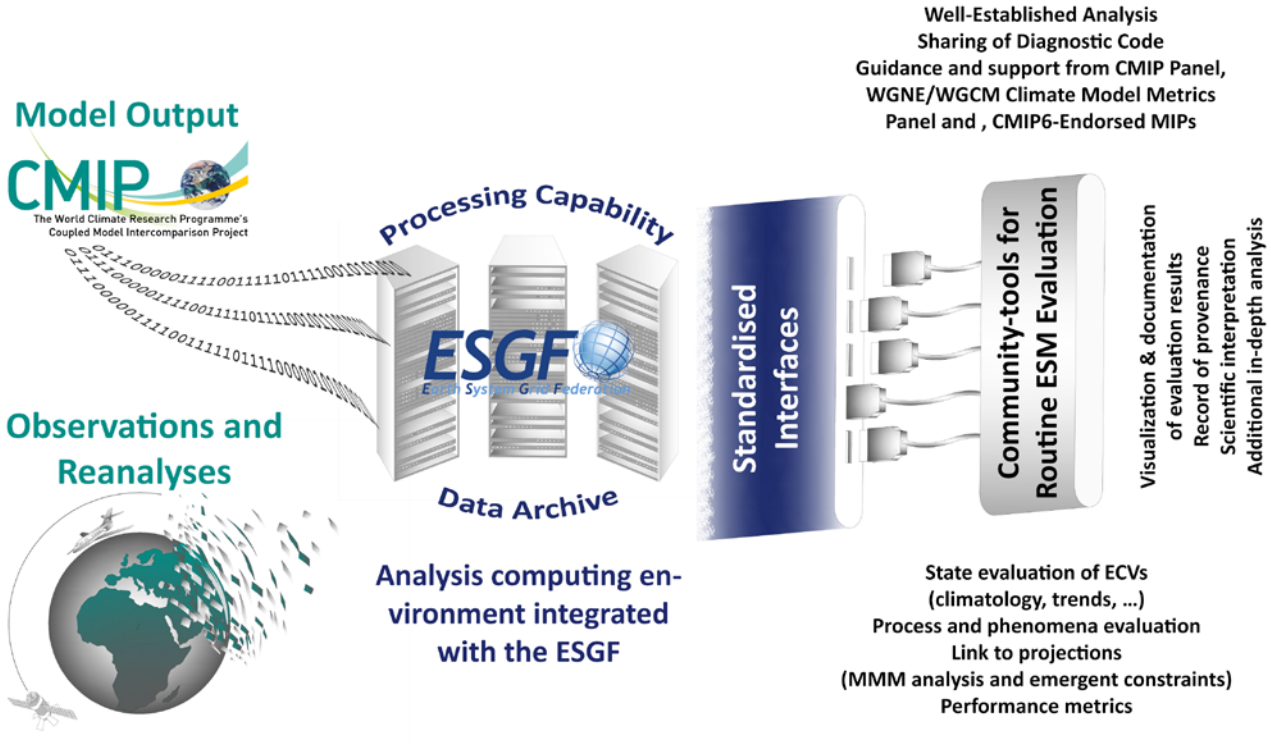

Figure 1: Schematic diagram of the workflow for routinely producing a broad characterization of model performance for CMIP model output using community evaluation tools that utilize relevant observations and reanalyses and rely on the ESGF infrastructure.

5

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

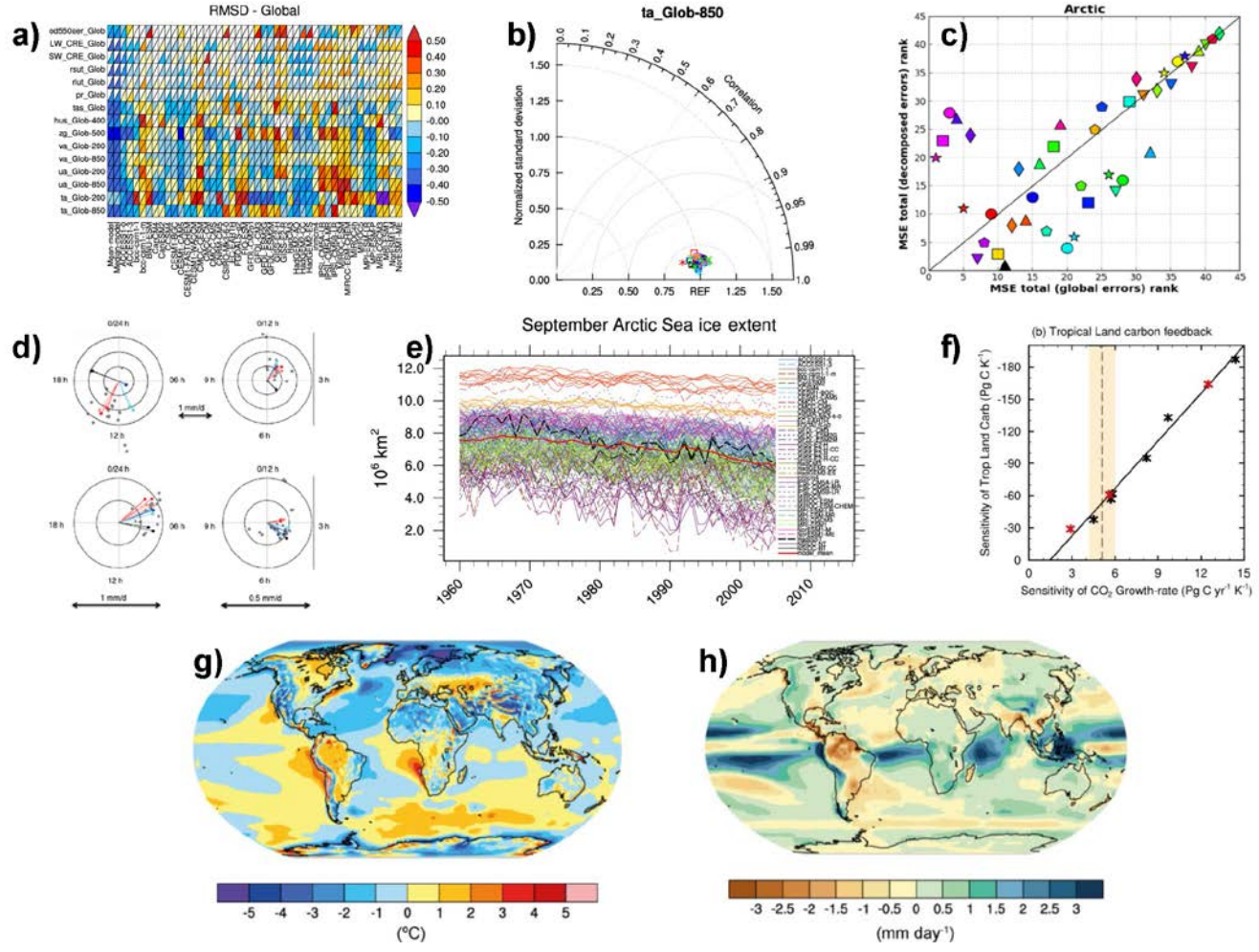

Figure 2: Examples of performance metrics and diagnostics that will be routinely calculated on CMIP models. Figures produced with ESMValTool version 1.0 (Eyring et al., 2016b) and PMP (Gleckler et al., 2016). (a) Multi-model, multi-variable summary of relative root-mean square error (RMSE) for CMIP5 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) 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 precipitation rate (mm day^{-1}) bias from the CMIP5 multi-model mean compared to ERA-Interim and the Global Precipitation Climatology Project, respectively.

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

Earth System Dynamics
Discussions
Open Access
EGU

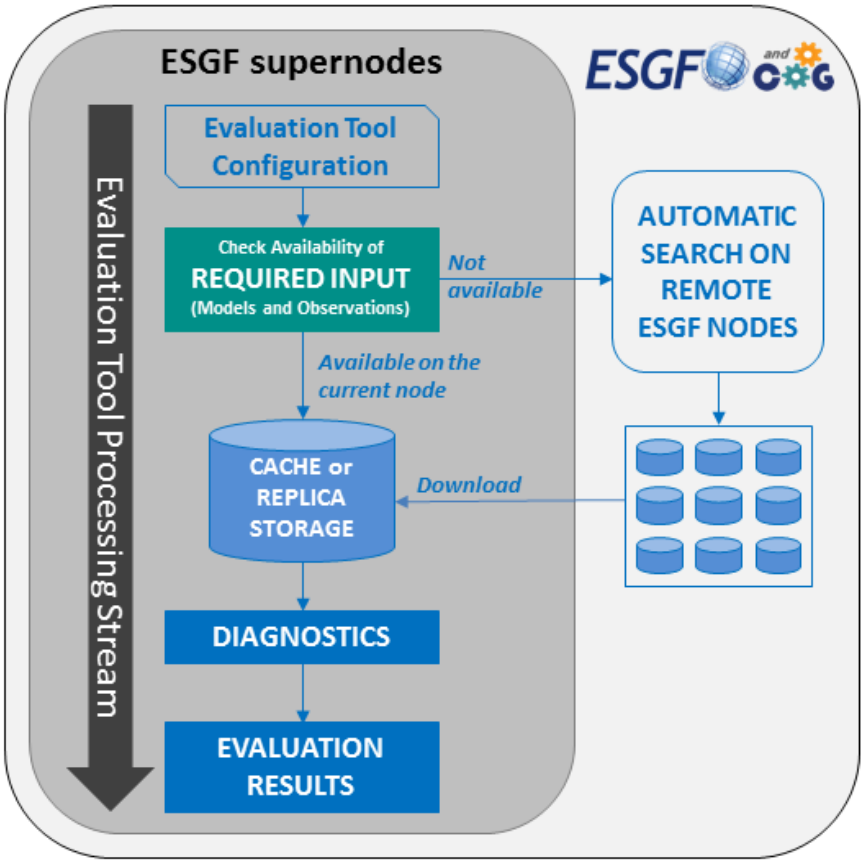

Figure 3: Schematic diagram of the envisaged evaluation tool processing stream for CMIP6. The schematic displays how the tools will be executed directly on ESGF supernodes exploiting optimized ESGF data organization and software solutions (see details in Sect. 2.3).

Earth Syst. Dynam. Discuss., doi:10.5194/esd-2016-26, 2016
Manuscript under review for journal Earth Syst. Dynam.
Published: 16 June 2016

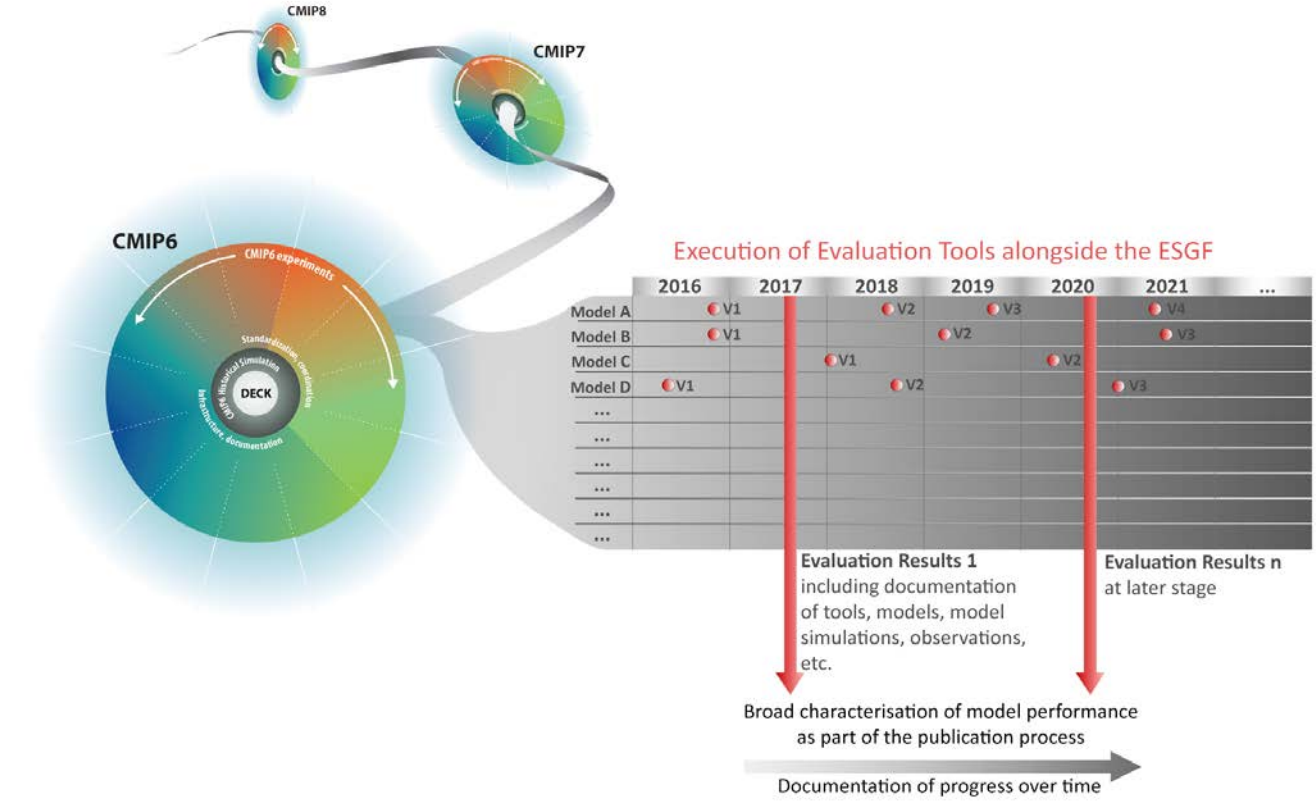

Figure 4: Schematic diagram of routine evaluation of CMIP DECK experiments and the CMIP historical simulations that is envisaged on the long-term. The evaluation tools would be executed quasi-operationally to produce a broad characterization of model performance as part of the ESGF publishing workflow, 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.