IS-ENES3 Deliverable D3.4

CMIP documentation requirements

Reporting period: 01/07/2020 – 31/12/2021

Authors: David Hassell (NCAS-University of Reading), Eric Guilyardi (CNRS-IPSL), Charlotte Pascoe (NCAS-STFC), Bryan Lawrence (NCAS-University of Reading), Sadie Bartholomew (NCAS-University of Reading), Atef Bennasser (IPSL), Mark Greenslade (IPSL), Martina Stockhause (DKRZ), Ruth Petrie (NCAS-STFC), Guillaume Levavasseur (IPSL), Allyn Treshansky (University of Colorado), Chris Blanton (GFDL), Sylvia Murphy (NOAA-CIRES)

Reviewers: Olivier Boucher (CNRS-IPSL), Matthew Mizielinski (MOHC)

Release date: 17/12/2021

ABSTRACT

Modelling group participation CMIP6 has been predicated on the understanding that the CMIP6 results will be fully documented and made accessible via the Earth System Documentation (ES-DOC) viewer and comparator interface (https://es-doc.org). Given the wide variety of users of CMIP6 outputs, who are from a wide range of specialist and/or non-technical backgrounds, the need for traceability is seen to be necessary in order to achieve the best understanding of CMIP6 datasets produced by the modelling institutes.

This report describes how the ES-DOC infrastructure has been developed with the collaboration of the scientists who will ultimately be using and benefiting from a comprehensive description of every element of the CMIP6 workflow.

Revision table			
Version	**Date**	**Name**	**Comments**
Release for review	02/12/2021	David Hassell	First version of the document
First revision	06/12/2021	Olivier Boucher	Comments from first reviewer
Second revision	13/12/2021	Matt Mizielinski	Comments from second reviewer
Final version	17/12/2021		
Dissemination Level			
PU	Public		X
CO	Confidential, only for the partners of the IS-ENES3 project		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Table of contents

Executive Summary

This report describes the work of ES-DOC towards delivering a sustainable documentation framework that can be applied for the current of the Coupled Model Intercomparison Project (CMIP6) that will also have utility for future numerical modelling projects (climate-related or otherwise) that are not supported by IS-ENES3. The need for traceability delivered through comprehensive documentation is necessary in order to achieve the best understanding and use of CMIP6 datasets produced by the modelling institutes, and this traceability has to include every aspect of the modelling workflow.

The role of the ES-DOC team within IS-ENES3 is to create the necessary infrastructure for collecting, archiving, and accessing the CMIP6 documentation, as well as ensuring that it is possible to record the information that is appropriate to users' needs. The infrastructure needed to deliver these documentation requirements was created via close consultation with the stakeholder communities to ensure that all of their practicable use cases may be delivered.

Delivery of the ES-DOC documentation infrastructure has been delayed, but will be completed by the end of 2022. Provision of documentation content by the modelling groups is currently patchy, partly as a result of delays in the ES-DOC delivery but also as this time consuming task is often given a low priority in the already busy schedules of the independent institutes.

It is recommended that the ES-DOC web services and software maintenance activities be addressed by the IS-ENES3 sustainability activity; and that ES-DOC should continue to engage with both the WGCM Infrastructure Panel (WIP), the Copernicus Climate Change Service (C3S), and national activities on the best way to document future global climate modelling projects.

1 Objectives

1.1 Introduction

Modelling group participation in CMIP6[1], the latest phase of the Coupled Model Intercomparison Project, has been predicated on the understanding that the CMIP6 results will be fully documented and made accessible via the Earth System Documentation (ES-DOC) viewer and comparator interface (https://es-doc.org). Given the wide variety of users of CMIP6 outputs, who are from a wide range of specialist and/or non-technical backgrounds, the need for traceability is seen to be necessary in order to achieve the best understanding of CMIP6 datasets produced by the modelling institutes. This traceability has to include every aspect of the modelling workflow, so the documentation has been broken down into the distinct types shown in Figure 1.1a, that may be summarised as:

- **Experiments**: The ES-DOC project has already recorded documentation of the CMIP6 experiments including lists of forcings, model configuration, numerical requirements, information about building the ensembles, links to citations and contact information of the principal investigators as well as text descriptions and information about the rationale behind each experiment.
- **Models**: Models will be described on a realm-by-realm basis (i.e. atmosphere, ocean, sea ice, etc.) as well as the top level (coupled model configuration).
- **Experimental conformance**: Each simulation should conform to a number of specific requirements established by the MIP leaders[2]. In some cases there is more than one way to meet the requirements, so modelling groups should record information about how each simulation conforms to the specifications.
- **Individual members of an ensemble**: The description of the ensemble of individual simulations that was created for each experiment.
- **Computer hardware performance**: Information on the hardware used in running simulations and also metrics describing the performance of each simulation on its machine.

[1] Eyring, V, Bony, S, Meehl, G A, Senior, C, Stevens, B, Stouffer, R J and Taylor, K E (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev* 9: 1937–1958, DOI: https://doi.org/10.5194/gmd-9-1937-2016

[2] Special issue on Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization, 2015-2020, https://gmd.copernicus.org/articles/special_issue590.html

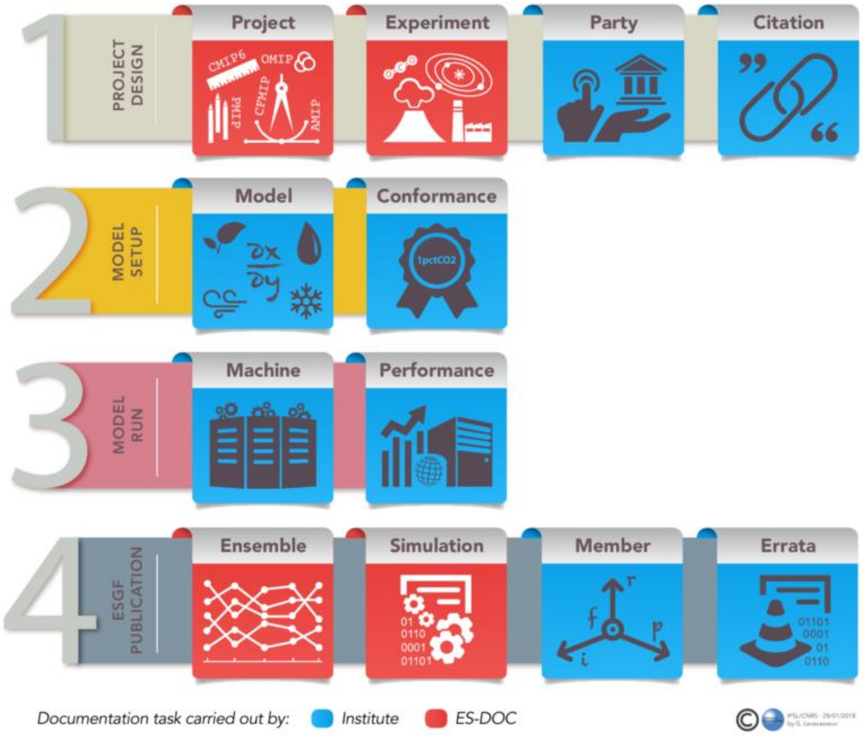


Figure 1.1a: *The CMIP6 documentation types.*

Note that documentation on the conformance to the data request[3], that defines all of the model variables that should be output from each simulation, has not been provided by ES-DOC. This is only because the capability to record this was only created after the ES-DOC work plan was finalised, but should be included for future projects (see section 3.2).

The role of the ES-DOC team within IS-ENES3 is to create the necessary infrastructure for collecting, archiving, and accessing the CMIP6 documentation, as well as ensuring that it is possible to record the information that is appropriate to users' needs. The infrastructure needed to deliver these documentation requirements can only be created via close consultation with the stakeholder communities, to ensure that all of their practicable use cases may be delivered. This report describes the methodologies and results of this engagement.

[3] Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M., and Sénési, S.: The CMIP6 Data Request (DREQ, version 01.00.31), Geosci. Model Dev., 13, 201–224, https://doi.org/10.5194/gmd-13-201-2020, 2020.

If CMIP6 was to achieve successful workflow documentation, necessarily through committed and sustained input from every participating modelling group, addressing these issues was vital, as CMIP6 has increased size, complexity and scope compared to CMIP's previous phase, CMIP5[4]. For instance, CMIP6 has 53 participating institutes running 314 experiments with a total of 140 models producing (to date) ~11.5 Petabytes of distinct archived datasets. In comparison, CMIP5 had 28 participating institutes running 56 experiments with a total of 44 models and produced 1.5 Petabytes of distinct archived datasets.

The CMIP6 experiments in particular embody this added complexity, with a new organisational framework involving the distributed management of a collection of quasi-independently constructed model inter-comparison projects (MIPs), each containing many experiments, some of which are defined in other MIPs (Figure 1.1b)[5]. The documentation of the CMIP6 experiments is described in section 2.8.

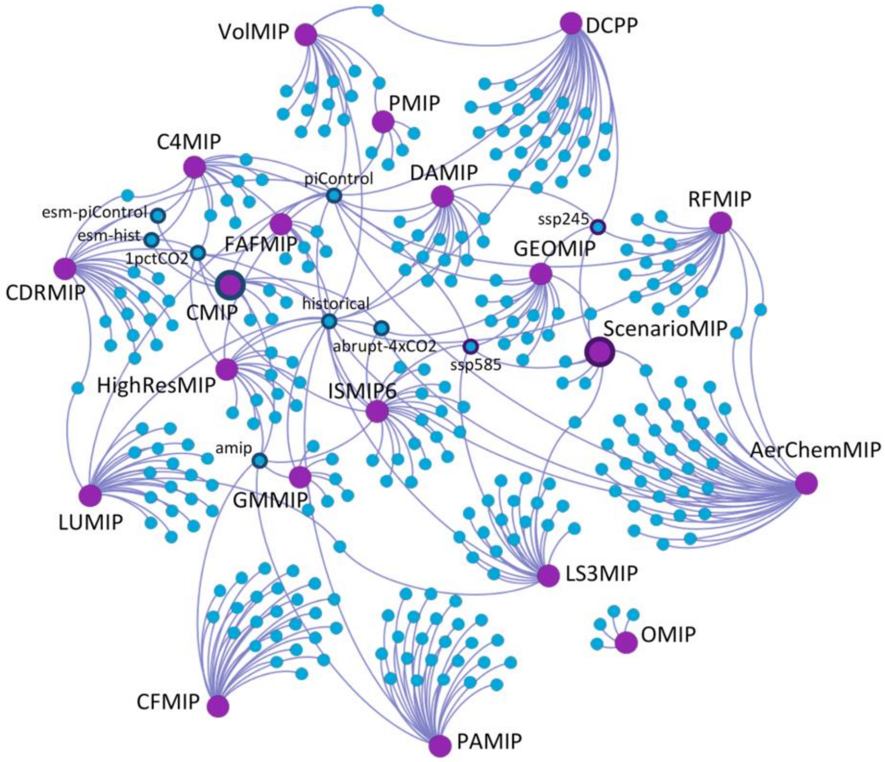


Figure 1.1b: *The CMIP6 MIPs and experiments. Individual MIPs are represented by large purple dots. Lines connect each MIP to the experiments that are related to it, which are shown as smaller blue dots.*

[4] Taylor, K.E., R.J. Stouffer, G.A. Meehl: An Overview of CMIP5 and the experiment design. Bull. Amer. Meteor. Soc., **93**, 485-498, doi:10.1175/BAMS-D-11-00094.1, 2012.

[5] Pascoe, C., Lawrence, B. N., Guilyardi, E., Juckes, M., and Taylor, K. E.: Documenting numerical experiments in support of the Coupled Model Intercomparison Project Phase 6 (CMIP6), Geosci. Model Dev., 13, 2149–2167, https://doi.org/10.5194/gmd-13-2149-2020, 2020.

The details of the software and hardware infrastructure that delivers the collection and access to the CMIP6 documentation is fully described elsewhere (IS-ENES3 deliverables D10.1[6] and D10.2[7]), and only particular aspects of it will be mentioned here, as required.

1.2 Overall strategy

The CMIP5 project was also documented via ES-DOC as part of the IS-ENES2 project[8], but significant issues were identified in that process:

- **Context**:
  - there was inadequate information for information providers and consumers about what information was being collected and why.
- **Interconnection**: there was insufficient linkage between:
  - different documentation elements of a given workflow;
  - the different documentation types.
- **Workload**:
  - creating documentation was an onerous and time consuming task for the modelling groups.
- **Granularity**:
  - it was not possible to publish partially complete documentation.
- **Detail**:
  - for model documentation, the on-line questionnaire was not sufficiently flexible to capture every aspect of all models.
- **Quality**:
  - errors in documentation content need to be reduced;
  - a wider range of individuals should take responsibility for the information and quality control.
- **Delivery:**
  - Much of the software required to capture the CMIP5 simulation workflow arrived late in the process, and the software to exploit that information was even later.

[6] Fiore, S., Nassisi, P., Antonio, F., Barring, L., Ben Nasser, A., Berger, K., Hassell, D., Juckes, M., Kershaw, P., Kindermann, S., Levavasseur, G., Nuzzo, A., Pagé, C., Stephens, A., Som de Cerff, W., Spinuso, A., Stockhause, M., & Weigel, T. (2020). Architectural document of the ENES CDI software stack (D10.1). Zenodo. https://doi.org/10.5281/zenodo.4309892

[7] Spinuso, Alessandro, Som de Cerff, Wim, Nassisi, Paola, & Pagé, Christian. (2021). First release of the ENES CDI software stack (D10.2). Zenodo. https://doi.org/10.5281/zenodo.4450012

[8] https://cordis.europa.eu/docs/results/312/312979/final1-is-enes2-finalreport.pdf

A first step was to invite each modelling group to appoint an “ES-DOC liaison” to act as an intermediary between the ES-DOC team and their modelling group. The ES-DOC liaison would be trained by ES-DOC, coordinate the documentation creation from within their modelling group, and manage the upload and publication of that documentation. The ES-DOC liaisons are further described in section 2 (Methodology and Results) including in the dedicated section 2.4.

Noting the CMIP5 experiences, the requirements for delivering CMIP6 documentation fall into distinct categories:

- Allow for the capture of the increased complexity of the CMIP6 project and of current generation of climate models.
- Appoint an “ES-DOC liaison” from each modelling group to coordinate the documentation process, and allow a wider range of different individuals to take responsibility for the documentation, in turn increasing quality (control).
- Improve the software functionality and interfaces for documentation collection.
- Move the responsibility of creating documentation from the modelling groups to ES-DOC, for as many document types as possible.
- Create a linkage of documents in the same workflow.

A further requirement emerged when it was known that some groups had started maintaining their own databases of documentation relevant to CMIP6, namely

- to allow a group’s own documentation system to be programmatically converted to documents that can be published by ES-DOC, thereby removing the responsibility of creating some documentation from the modelling groups to ES-DOC.

Redesigning the ES-DOC infrastructure to satisfy these requirements had three phases. Firstly, the information model used for representing documentation - the Common Information Model (CIM)[9] - needed to be updated (the new data model became known as CIM2 and is described further in section 2.1); secondly, the back-end tooling for creating, manipulating and publishing content created with the CIM2 was to be developed; and finally, the CMIP6 community had to be consulted in detail on how their requirements could best be delivered.

[9] 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, https://doi.org/10.5194/gmd-5-1493-2012, 2012.

A further requirement emerged when it was known that some groups had started maintaining their own databases of documentation relevant to CMIP6, namely to allow a group's own documentation system to be programmatically converted to CIM-based documents that can be published by ES-DOC, thereby removing the responsibility of creating some documentation from the modelling groups to ES-DOC. This process is described in section 2.7.

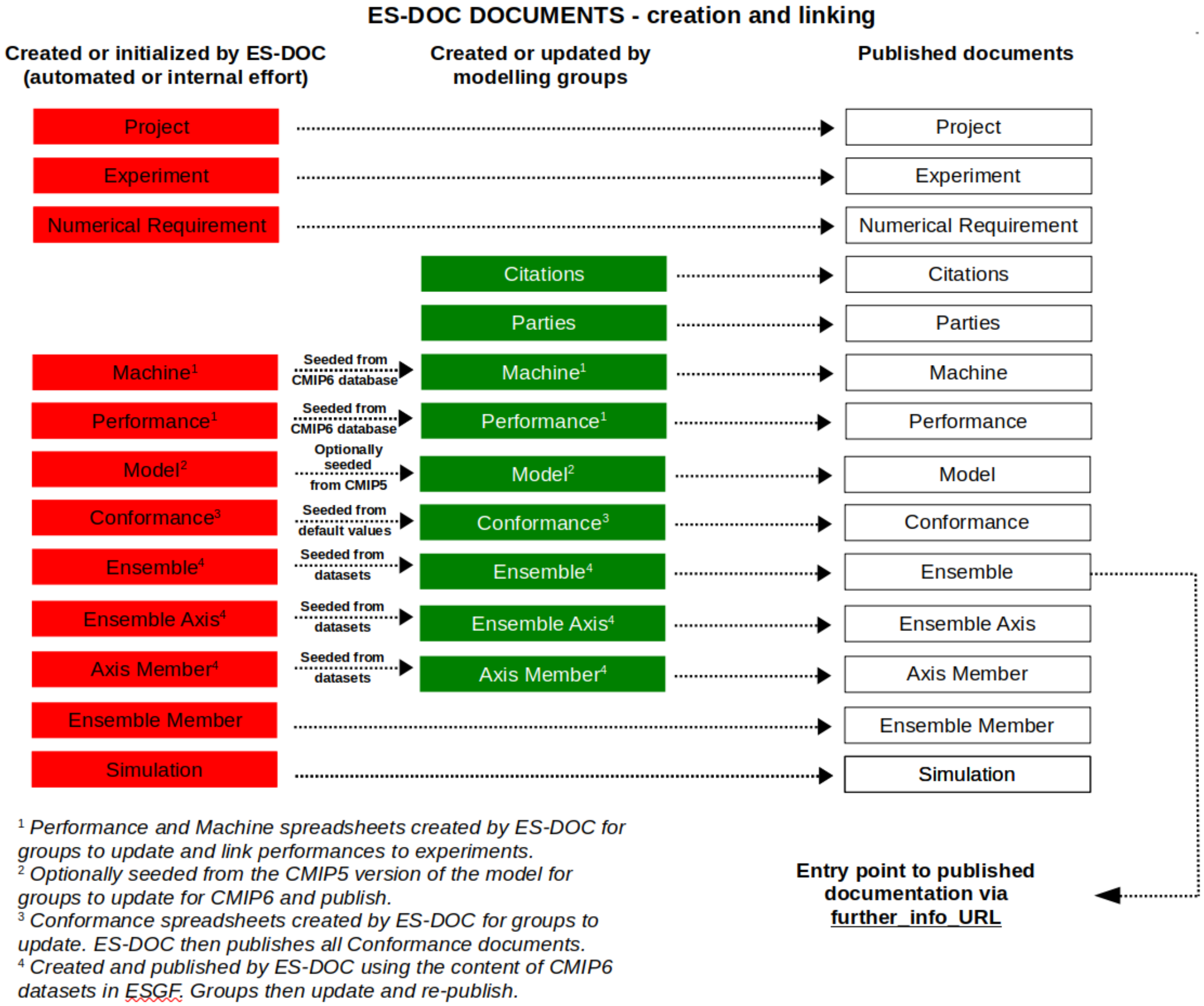
These objectives were contributed to and endorsed by the WGCM Infrastructure Panel (WIP, a WGCM subcommittee charged with coordinating infrastructure support for CMIP)[10] [11], which includes some ES-DOC staff as members.

[10] https://www.wcrp-climate.org/wgcm-cmip/wip

[11] https://wcrp-cmip.github.io/WGCM_Infrastructure_Panel//Papers/CMIP6_ESDOC_documentation.pdf

2 Methodology and Results

The different document types and how they are created, either by ES-DOC or with input from the modelling groups are summarised in Figure 2. The methods and procedures used to create the various documents are covered within this section.



[1] *Performance and Machine spreadsheets created by ES-DOC for groups to update and link performances to experiments.*
[2] *Optionally seeded from the CMIP5 version of the model for groups to update for CMIP6 and publish.*
[3] *Conformance spreadsheets created by ES-DOC for groups to update. ES-DOC then publishes all Conformance documents.*
[4] *Created and published by ES-DOC using the content of CMIP6 datasets in ESGF. Groups then update and re-publish.*

Figure 2: *Overview of ES-DOC documents and creation workflow for CMIP6. The further_info_URL (section 2.9) points to the ensemble document and is the entry point to the full documentation. The documents listed on the left are generated by ES-DOC (either automatically or via internal effort). The central column lists the documents either created by modelling groups (citations, parties, machine and model) or updated toward their final version.*

2.1 Extending the CIM

The CIM identifies the key concepts of a simulation workflow and shows how they interact and depend on each other. Its aim is to describe climate data and the models that produce it in a standard way, minimising fragmentation and gaps in availability of metadata. The first version of the CIM (CIM1) was developed to describe the CMIP5 workflow, but in light of the requirements identified here, a new version (CIM2) has been developed which has been redesigned to make a clearer distinction between the different parts of the system, including ownership and governance, allows for a more modular approach, and is also less "CMIP-specific", i.e. it has been designed to be also applicable to other modelling activities (see also section 2.12).

A key aspect of the modularity is the partitioning of some document types into elements that are themselves stand-alone documents. The most relevant such case is the Model document, i.e. the description of a model's scientific formulation. Under CIM2, the model document is composed of multiple Realm documents, each of which may exist independently, where each realm describes a physically coherent model component. Within CMIP6 there are eight possible realms: atmosphere, atmospheric aerosols, atmospheric chemistry, land surface, land ice, ocean, ocean biogeochemistry, and sea ice. Each realm may be documented independently of the others, thus allowing each realm description to be published as soon as it is complete, without needing to wait for the others to be finished. This framework shifts the responsibility of content creation away from a single person towards allowing each realm's description to be the responsibility of those who best understand it; and also allows each realm to be published separately, without having to wait for others to be completed.

An important feature of the CIM is the ability to "specialise" it for a particular purpose. These specialisations create an instance of the CIM that is familiar and understandable within the confines of a particular project. For example, the base CIM describes a generic realm as having arbitrary processes, and these processes can have arbitrary properties. However, within the CMIP6 project we know, for example, that all atmosphere realms will have a dynamical core (a process) and these dynamical cores will have a timestep of integration (a property), and a "specialisation" can be created that pre-defines this process, with this property, for the realm. This then makes the documentation clear to those who are both creating and consuming content, as well as allowing documentation to be compared through well-defined features.

An issue with the CMIP5 model documentation was that, for some elements, it was not possible to provide accurate content for all models. This arose because the questions being asked of the content providers were sometimes too specific and not enough flexibility was offered for the range of potential answers. To avoid this happening again, it was decided to include in the specialisations more questions which asked for open-ended answers, rather than limiting answers to given options. It is important that these open-ended questions do not wholly replace questions for which the answers are more limited. A limited answer might, for instance, ask for a numeric response or a response from a controlled vocabulary of allowed values. By using the two types of questions alongside each other, as appropriate, we can minimise the chance of a modelling group not being able to describe parts of their models whilst retaining the power of the automated comparison of documentation that comes from the presence of well-defined limited questions and answers.

An example of this approach may be seen in the CMIP6 specialisation for describing the ocean bottom boundary layer. This process of an ocean realm is described by the properties listed in Table 2.1, which demonstrate the mix of open-ended and specifically restricted documentation types.

Property	**Description**
`overview`	A free-text description of the bottom boundary layer in ocean
`lateral_mixing_coef`	The type of bottom boundary layer in ocean, one of “diffusive” or “advective”, or a user-defined setting
`lateral_mixing_coef`	If bottom BL is diffusive, specify the float value of lateral mixing coefficient (in m2/s)
`sill_overflow`	A free-text description of any specific treatment of sill overflows

Table 2.1: *The properties that define the ocean bottom boundary layer of the ocean realm, adapted from the specialisation file* https://github.com/ES-DOC/cmip6-specializations-ocean/blob/master/ocean_uplow_boundaries.py.

2.2 Community designed specialisations for model documentation

It was essential for the user community to be engaged in determining the scientific content required for describing the CMIP6 climate models, as it is they who collectively have the knowledge of what needs to be documented for the CMIP6 generation of models. In this case the primary user community comprises expert model developers from the modelling groups who will also have to create the documentation based on the new specialisations. Individual consultations with the experts on each model realm from each modelling group would not be practical, as that would entail conducting many hundreds of interviews, along with the subsequent work in ensuring all of the results could be incorporated consistently. Instead it was decided to build an initial version of the CMIP6 model descriptions from individual consultations with a small number of scientists, and then invite the wider community to submit comments and suggestions on this initial version, which would then be addressed by ES-DOC to create the final specialisations for the CMIP6 model descriptions.

Climate models are generally developed incrementally, which means that most CMIP6 models have many similarities with a CMIP5 model counterpart. Therefore a good starting point for creating specialisations for each model realm was to copy the CMIP5 documentation structure into the new CIM2 framework[12], at the same time as the introduction of the new open-ended questions. This task was carried out by members of the ES-DOC team, who also took on responsibility for coordinating the further development of the scientific content.

The canonical description of each specialisation is defined as a Python script residing in a GitHub repository (all of which are listed in Table B.1 in appendix B), and it was initially intended that, once briefed, the expert scientists would simply edit these files directly to add new content. This approach was dropped very soon into the process, as taking the time to learn the details on how to do this discouraged engagement, even from people who were already Python users, and in any event the syntax of the specialisation files was also undergoing a rapid development cycle, rendering them unsuitable for modification by those outside of the ES-DOC team. The alternative solution was for ES-DOC to provide more accessible means of viewing the specialisations, and then collecting the required changes during interviews, either in person or via video conferencing. These changes would be later added to the Python specialisation files by the ES-DOC interviewer, and presented back to the interviewees for review.

2.2.1 Accessible views of the specialisation content

A "mind map" view of the documentation content (such as in Figure 2.2.1) was well received by the community during the development of the CMIP5 documentation, and so the same approach was used here. These CMIP5 mind maps contain the structure of the CMIP5 model documents and have been archived to each CMIP6 realm's GitHub repository[13] (see Table B.1 in appendix B). Mind maps are good for viewing the overall structure (e.g. at the process level), but can be harder to use for viewing individual properties within a sub-process, as the expanded graph can quickly become too large to fit on a screen. Therefore, a tabular view was also provided to make it easier to inspect the detail instead of the overall structure[14]. Both of these views, as well as the canonical Python files, were easily accessible on-line via the ES-DOC website[15]. Whilst the tabular view is still available, the mind map view is no longer there since it required a paid service to host it and

[12] E.g. https://github.com/ES-DOC/cmip6-specializations-seaice/tree/master/mappings

[13] E.g. https://github.com/ES-DOC/cmip6-specializations-aerosol/blob/master/CMIP5_Aerosols_bdl.mm

[14] E.g. https://specializations.es-doc.org/cmip6/aerosol

[15] https://es-doc.org/cmip6-specialisations/

the need for them has passed. However, the source files for the mind maps[16] have been retained in the specialisation GitHub repositories (see Table B.1 in appendix B), and so the mind map views may be recreated at any time.

These resources were successfully used as an interactive resource during the interviews between ES-DOC and the realm experts. Changes to the content resulting from an interview were transferred to the specialisation Python files by ES-DOC members, from which the mind maps and tabular views could be easily regenerated and presented back to the scientist for review.

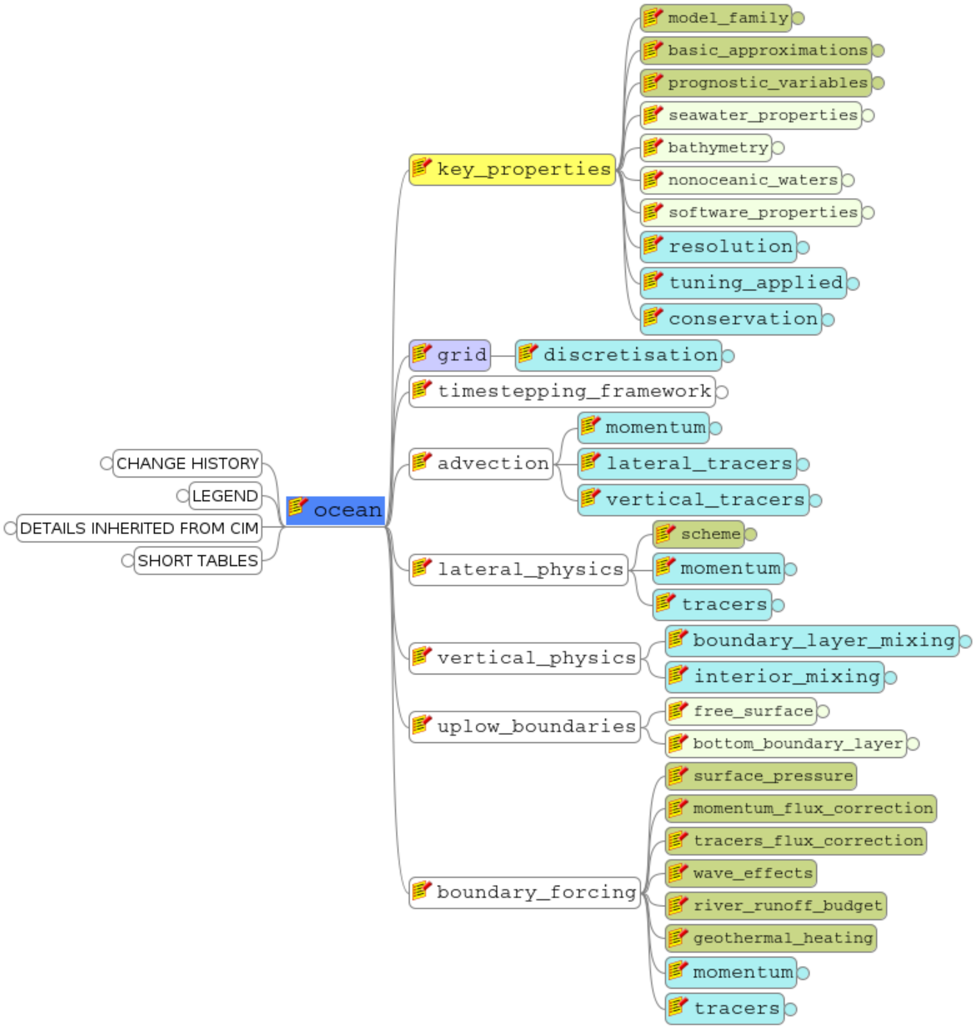


Figure 2.2.1: *The final mind map of the ocean realm, showing only processes, and with all properties collapsed (https://github.com/ES-DOC/cmip6-specializations-ocean/blob/master/_ocean.mm)*

[16] E.g. https://github.com/ES-DOC/cmip6-specializations-seaice/blob/master/_seaice.mm

2.2.2 The scientific experts taking part in the initial interviews

The scientific experts for each realm who kindly gave up their time to take part in these initial interviews were (their host institutions are given in brackets, see Table A.1 in appendix A for details):

Model realm	Scientific experts	ES-DOC interviewers
Top-level	● Roland Seferian (CNRM) ● Tim Johns (MOHC)	David Hassell
Atmosphere	● Robert Pincus (NOAA)	Charlotte Pascoe
Atmospheric aerosols	● Yves Balkanski (LSCE) ● Michael Schulz (MET Norway)	David Hassell
Atmospheric chemistry	● Bill Collins (University of Reading)	David Hassell
Land ice	● Sophie Nowicki (NASA) ● Steve George (NCAS-University of Reading)	David Hassell
Land surface	● Rich Ellis (CEH) ● Phillipe Peylin (IPSL)	Eric Guilyardi, David Hassell
Ocean	● Julie Dehayes (LOCEAN/IPSL) ● Steve Griffies (GFDL) ● Alistair Adcroft (GFDL) ● Gokhan Danabasoglu (NCAR) ● Gurvan Madec (LOCEAN/IPSL)	Eric Guilyardi
Ocean biogeochemistry	● Olivier Aumont (LOCEAN/IPSL) ● Laurent Bopp (LSCE/IPSL)	Eric Guilyardi
Sea ice	● Jamie Rae (MOHC) ● Martin Vancoppenolle (IPSL) ● Alexandra Jahn (University of Colorado)	Ruth Petrie

Table 2.2.2: *Initial contributors to the CMIP6 model specialisations.*

2.2.3 A community-wide review of the specialisations

Once the initial CMIP6 Model documentation content had been created in the specialisations by the selected realm experts, the content was put out to the whole CMIP6 community for review, via the CMIP6 mailing lists. The mind map and tabular views were advertised, and on-line feedback spreadsheets were set up in which anyone could ask questions and make further suggestions. A total of 157 comments were received from 35 reviewers at 14 institutes, all of which were openly addressed by the ES-DOC team. An example comment and response is given in Table 2.2.3. The full contents of each of these spreadsheets are available from the ES-DOC website[17], and have also been archived within each realm's GitHub repository (see Table B.1 in appendix B).

Date	**Reviewer**	**Reviewer's Institution**	**Component**	**Comment**	**ES-DOC Response** ***(include the date, the responder's name, and the new version number if implementing any changes)***
2017-11-21	Christian Rodehacke	Danish Meteorological Institute	Dynamics -> Timestep	If the ice sheet/ice shelf model uses an adaptive time scheme, what time step of the ice scheme shall we report for "Dynamics -> Timestep"; longest, shortest, mean?	Added a question on the presence of an adaptive time scheme, and if there is one then any reasonable, representative timestep may be reported (0.5.0) David Hassell

Table 2.2.3: *An excerpt from the land ice realm feedback spreadsheet (archived at https://github.com/ES-DOC/cmip6-specializations-landice/blob/master/ES-DOC_realm_review_landice.pdf)*

This community-wide review concluded the creation of the model documentation realm specialisation, thus defining the content of the CMIP6 model documentation.

2.3 Reviewing the content of machine and performance documentation

CIM2 contains a new representation for documenting the machines that simulations are run on, and the performance of those model runs. Whilst any element of CIM2 may be modified with specialisations, the structure and properties already built into the CIM may already be sufficient, thus removing the need for extra complexity in document creation infrastructure; this was

[17] https://es-doc.org/cmip6-specialisations/

considered to be the case for machine and performance documentation for which there is a much restricted variety in the items being described.

For machines, CIM2 was designed with the current (i.e. CMIP6-era) collection of computers in mind. It was therefore not surprising that an external review by V. Balaji (GFDL) and Mario Acosta (Barcelona Supercomputing Centre), as well as an internal re-review by Bryan Lawrence, the lead author of CIM2, brought up no major issues. However, the inclusion of machine properties relating to benchmark performances was identified as being universally useful and were added directly as properties to CIM2. These properties were the maximal LINPACK performance achieved and the theoretical peak performance[18].

The CIM2 content for performance of simulations running on these machines was wholly based on the computational performance metrics that were specifically designed to allow comparison between climate models, known as the CPMIP metrics[19], and external review from the authors of the metrics agreed that these were sufficient for capturing the performance of the CMIP6 models.

Understanding the way in which the modelling groups would be collecting their performance documentation would open up possibilities for reducing their documentation overhead, as is described in section 2.6.

2.4 ES-DOC liaisons

The ES-DOC team invited modelling groups to appoint an ES-DOC liaison to act as an interface between their institution and the ES-DOC team during the CMIP6 documentation process. This person’s role was to:

- be in charge of collecting the institute’s CMIP6 documentation, which will involve interaction with scientists and IT experts;
- become familiar with various documentation methodology that ES-DOC is providing for CMIP6 documentation;
- manage any technical issues, liaising with the ES-DOC team as required;
- act as a point of contact for communications from ES-DOC to the CMIP6 modelling groups.

[18] Dongarra J., Luszczek P. (2011) LINPACK Benchmark. In: Padua D. (eds) Encyclopedia of Parallel Computing. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09766-4_155

[19] Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geosci. Model Dev., 10, 19–34, https://doi.org/10.5194/gmd-10-19-2017, 2017.

The creation of such a role would also make it practical for many individuals within a modelling group to contribute to documentation creation, who would otherwise be difficult to identify and reach from the ES-DOC team.

The ES-DOC liaison would be trained in the documentation procedures, thereby freeing the content providers at their institute from learning this technical aspect, and it is expected that the ES-DOC liaison could coordinate the creation of documentation content using methods most familiar to their group, before taking responsibility for uploading and publishing the information.

Training and documentation for ES-DOC liaisons takes the form of overview information on the CMIP6 documentation project (https://es-doc.org/cmip6); a comprehensive technical "how to" guide (https://es-doc.org/cmip6-how-to); and demonstration screencasts, with full transcripts (https://es-doc.org/cmip6-screencasts). In addition a mailing list was set up for all of the liaisons, but most perhaps importantly direct contact with the ES-DOC team was encouraged via the support@es-doc.org e-mail address.

To date, 44 out of 53 institutions participating in CMIP6 have volunteered an ES-DOC liaison contact.

2.5 Creating an interface for documentation creation

The tools that were provided to the CMIP5 modelling groups for collecting their documentation were on-line questionnaires, one for each document type (model, simulation, and ensemble descriptions). This was well received, and a new version was under development that would work within the CIM2 framework[20]. Unfortunately the lead questionnaire developer could not stay with the project and there were no extra resources to continue its development. An alternative was required at short notice, which took the form of a Python Jupyter Notebook[21]. This comprised cells that were pre-populated with the questions defined by the CIM and the specialisations, and formulated using the Python language syntax of the `pyesdoc` library[22]. This fully featured library developed within ES-DOC is the tool for all interactions (creating, reading, publishing) with CIM documentation. The answer to each question would be entered in the same cell following the usual Python language rules. This approach had many perceived advantages:

[20] https://github.com/ES-DOC/esdoc-questionnaire
[21] https://jupyter.org/
[22] https://github.com/ES-DOC/esdoc-py-client

- It was relatively easy to implement at short notice, as the `pyesdoc` library already provided the back-end functionality to the questionnaire that was under development.
- The Jupyter Notebooks could be presented on-line from an ES-DOC server, so no software installation was required from the users.
- The notebooks could be "run" by the user to validate their answers before submitting them.

In addition to the Jupyter Notebooks, extra resources were provided to make it easier to see what content needed to be provided. These took the form of a spreadsheet view of the documentation questions, and a text view in a PDF file. Both of these were made available via GitHub in a bespoke repository that was created for each CMIP modelling group - an "institutional repository" (see Table A.1 in appendix A).

However, it was not known how well this approach would be received by the ES-DOC liaisons who would be using it, therefore a trial of the infrastructure was required before a general release. To take part in this a small number of "beta testers" were recruited to use the Jupyter Notebooks to create a small amount of sample documentation, without any guidance other than that which was available via the ES-DOC website[23]. In addition to the detailed step-by-step instructions on creating documentation content with Jupyter Notebooks[24], the ES-DOC website contains overview information, including flow charts, on how the various processes work. Two aspects were reviewed by the beta testers: the overall CMIP6 documentation presentation and the more specific model documentation process, but this did not include the science relevance of the questions, which had already been reviewed by the CMIP community.

Feedback was gathered verbally and by e-mail from the following ES-DOC liaisons who kindly offered to take part in the beta testing (their institutions are given in brackets - see Table A.1 in appendix A for details):

- Chris Blanton (GFDL)
- Guillaume Levavasseur (IPSL)
- Jin Ba (BCC)
- John Scinocca (CCCma)
- Mahesh Ramadoss (CCCR-IITM)
- Mark Elkington (MOHC)
- Takahiro Inoue (MIROC)
- Thibaut Lurton (IPSL)

[23] https://es-doc.org/cmip6-how-to/
[24] https://es-doc.org/cmip6-models-documenting-with-ipython/

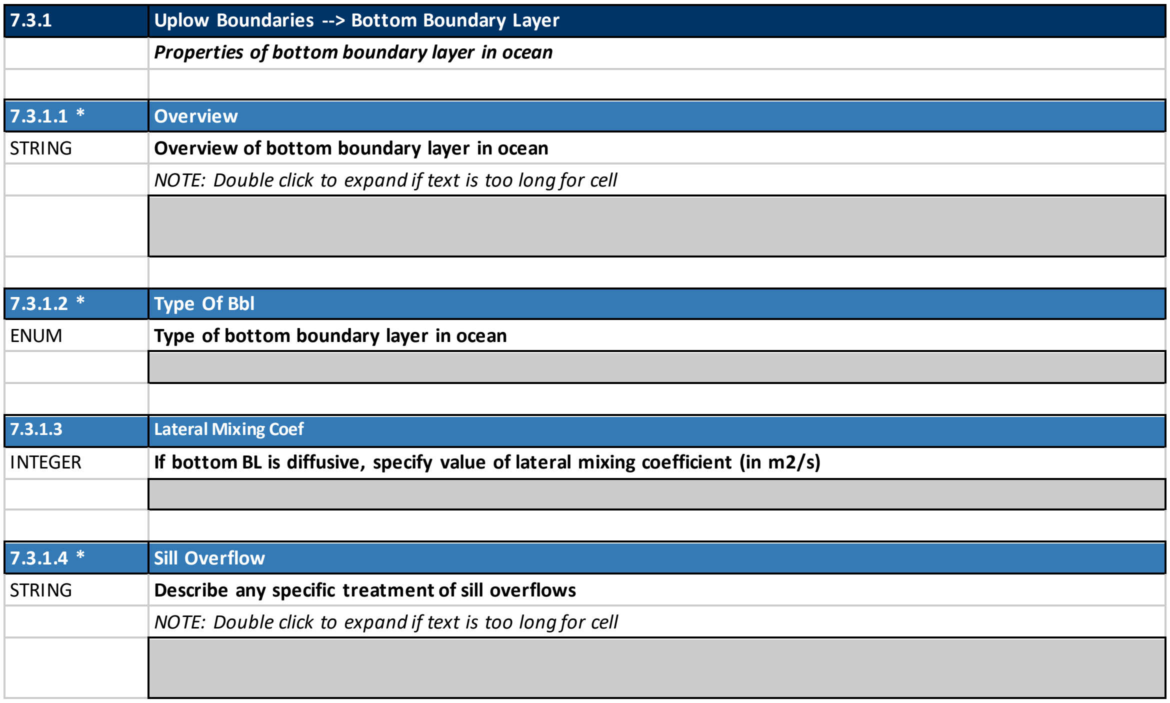
The detailed feedback from these beta testers is given in appendix C. In summary, the ES-DOC documentation process was in general regarded as good, with the workflow diagrams very helpful; and the ES-DOC approach was considered suitable for documenting the CMIP6 workflow.

However, the use of Jupyter Notebooks as the collection tool was seen as unnecessarily complex, and we were encouraged to find an alternative solution.

The obvious solution was to promote the spreadsheet views of the documentation stored in each institutional GitHub repository from a supplementary resource to the primary means for content creation and collection. This decision was universally popular since spreadsheets are universally understood, and could therefore be given directly (via the ES-DOC liaison) to the modelling experts to fill in, without any special training. An example spreadsheet is shown in Figure 2.5.

The ES-DOC liaisons still have a key role to play, even though the need to manually enter all of the information into Jupyter Notebooks has been removed. Internally organising the process and providing context within the modelling groups, and dealing with the technicalities of uploading the completed spreadsheets and requesting when they are ready for publication remain as essential activities.

The institutional GitHub repositories (see Table A.1 in appendix A) now played an integral role in the creation of CMIP6 documentation, as they are means by which artefacts can be delivered to the modelling groups, and by which ES-DOC can easily collect and process the content. The use of GitHub was not a barrier to the ES-DOC liaisons who would be interacting with the repositories, as the CMIP6 project was already requiring GitHub for some interactions, and in general it is rapidly becoming an essential research resource.



7.3.1	Uplow Boundaries --> Bottom Boundary Layer
	Properties of bottom boundary layer in ocean
7.3.1.1 *	Overview
STRING	Overview of bottom boundary layer in ocean
	NOTE: Double click to expand if text is too long for cell
7.3.1.2 *	Type Of Bbl
ENUM	Type of bottom boundary layer in ocean
7.3.1.3	Lateral Mixing Coef
INTEGER	If bottom BL is diffusive, specify value of lateral mixing coefficient (in m2/s)
7.3.1.4 *	Sill Overflow
STRING	Describe any specific treatment of sill overflows
	NOTE: Double click to expand if text is too long for cell

Figure 2.5: *An example from one of the documentation collection spreadsheets. This extract corresponds to the specialisation content for ocean bottom boundary layer described in Table 2.1.*

2.6 Reducing the documentation overhead

2.6.1 Model documentation

As mentioned in section 2.2 (Community designed specialisations for model documentation), climate models are generally developed incrementally, meaning that most CMIP6 models have many similarities with a CMIP5 model counterpart, or else another CMIP6 model. As a consequence, it is likely that a given CMIP6 model will share much of its documentation requirement with another model (or other models), and so it would be highly beneficial to the modelling groups to be able to create this shared information once and then reuse it as required.

To facilitate this, it is possible, by request from the ES-DOC liaison, to “seed” a model documentation spreadsheet for any realm from information that has been created for another CMIP6 model, or that was previously created for a CMIP5 model. The seeding is a copy of the information from the existing documentation which can then be edited if required. For instance, if the ocean realms of two models are identical apart from a single parameter value, then the first of these may be documented and copied to the second. The resulting spreadsheet for the second ocean realm may then be changed for the new parameter value.

This seeding is not restricted to be between models from a single modelling group, as models from the same family are often used at multiple institutes.

2.6.2 Simulation documentation

The creation of accurate documentation content was recognised to be an onerous and time consuming task for the modelling groups, so any means by which the ES-DOC team could reduce this workload would be welcomed. The description of the experiments is a general resource used by all modelling groups, and this has been produced by ES-DOC, as it was for the CMIP5 project. On the face of it, all other documentation types are specific to each modelling group and could therefore only be created with institutional knowledge. However, it was noted that the information required to document the individual simulations carried out with each model is in fact all contained in the metadata of the models' output datasets. This presented the possibility of ES-DOC being able to harvest this information and automatically create and publish the simulation documents.

However, for CMIP5 this was done manually by the modelling centres, and only about 50% of the simulations were described, with known inaccuracies in their content (found by checking their content against the metadata found in the published dataset outputs). A contributing factor to these errors was that the documents were created long after the actual simulations had been carried out, after some internal records were no longer available. An automatic creation approach would not only reduce the effort required from the modelling groups, but also guarantee a complete and accurate record of *all* CMIP6 simulations. Note that in this context, "accurate" means describing what was *actually* done, not which might not always have been what was intended, see below for more details.

It was first investigated whether this information could be retrieved from the publicly available Earth System Grid Federation (ESGF)[25] index of archived data. Such a mechanism was attractive as the information could be obtained without having to access the full datasets, and could be obtained wholly independently by ES-DOC. However, not all metadata was available (such as the identities of child and parent simulations) and, at the time, programmatic access to the ESGF index was unreliable.

Consequently, a different solution was decided upon. All CMIP6 output datasets must be ingested as CF-compliant netCDF files[26] into the ESFG archive, from where they are made available to users. Working with the maintainers of the ESGF software stack[27], code was developed to run as part of the ESGF publication process for all CMIP6 datasets. This code, called `cdf2cim`[28], reads the netCDF datasets as they are being published on an ESGF node so as to automatically extract

[25] https://esgf.llnl.gov

[26] https://cfconventions.org

[27] https://esgf.llnl.gov/software.html

[28] https://github.com/ES-DOC/esdoc-cdf2cim

all of the required metadata and send this information (which is only about 1 Kilobyte per simulation per publication request) to an archive managed by ES-DOC[29]. This procedure generates at least as many simulation descriptions as there are ESGF publications (i.e. millions) and will contain many duplicates, but these will be reduced, at a later date or in real time, to the much smaller unique set of simulation documents that can be published by ES-DOC.

This process is illustrated in Figure 2.6.1, which also describes how a simulation is related to all related documentation (e.g. model and experiment descriptions) via the “further_info_URL” (see section 2.9).

As mentioned above, the assumed accuracy of the resulting simulation documentation is reliant on the netCDF file metadata being correct. However, ESGF manages a very rigorous checking procedure on all CMIP6 datasets, only ingesting those which meet the required standards, so it is a reasonable assumption that the file metadata are correct enough for this purpose (and certainly more accurate than any manual process would likely be).

[29] https://github.com/ES-DOC/esdoc-cdf2cim-archive

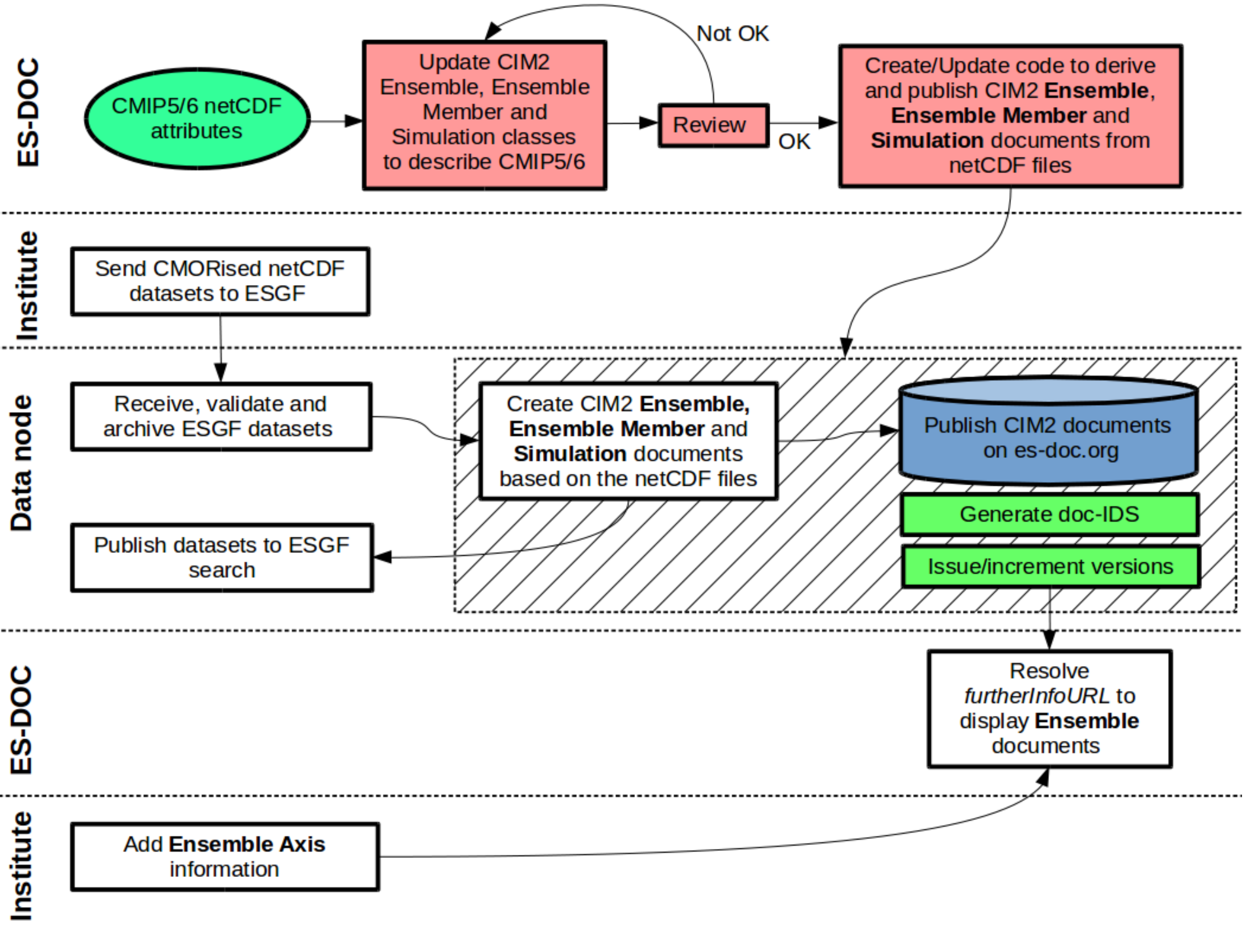

ES-DOC
CMIP5/6 netCDF attributes
Update CIM2 Ensemble, Ensemble Member and Simulation classes to describe CMIP5/6
Review
Not OK
OK
Create/Update code to derive and publish CIM2 **Ensemble**, **Ensemble Member** and **Simulation** documents from netCDF files
Institute
Send CMORised netCDF datasets to ESGF
Data node
Receive, validate and archive ESGF datasets
Create CIM2 **Ensemble, Ensemble Member** and **Simulation** documents based on the netCDF files
Publish CIM2 documents on es-doc.org
Generate doc-IDS
Issue/increment versions
Publish datasets to ESGF search
ES-DOC
Resolve *furtherInfoURL* to display **Ensemble** documents
Institute
Add **Ensemble Axis** information

Figure 2.6.1. *The gathering of simulation documentation from the ESGF publication process (https://es-doc.org/cmip6-ensembles-simulations).*

The `cdf2cim` software has also been designed to create simulation documentation from CMIP5 netCDF datasets, and will be used to retrospectively create a complete and correct set of CMIP5 simulation documentation in a one-off off-line processing of the ESGF CMIP5 archive.

2.6.3 Machine and performance documentation

The CIM2 design initially required a model performance for every simulation of every experiment's ensemble, but it was realised that this information on this granularity would not be collected by many CMIP6 modelling groups. To account for this practicality, it was also allowed for a *representative* performance to be documented that can apply to all member simulations of an ensemble. Furthermore, it was the case that a configured model's performance on a machine would typically only be measured once, and that measured performance assumed to apply to all applicable experiments. Many modelling groups ran hundreds of simulations, but with a small number of models (fewer than ten) on a smaller number of machines (typically one or two), so fewer than twenty performances were typically recorded per group. These performances need to be duplicated across their applicable ensembles of simulations, so the challenge was to create a system whereby each unique performance could be created only once, such that ES-DOC could manage the duplication across the relevant ensembles.

The solution was to present to the modelling groups the collection of machine and performance documentation as a connected two-stage process (illustrated in Figure 2.6.2). First, spreadsheets for collecting machine descriptions were delivered to the modelling groups, but as well as describing the computer, these spreadsheets contained a section in which modelling groups could state which model and experiment combination were run by each of their machines. The list of possible model and experiment combinations was also pre-restricted to only those model and experiment combinations that the institute had agreed to carry out, as defined in the CMIP6 controlled vocabularies[30], to make the task manageable.

[30] https://github.com/WCRP-CMIP/CMIP6_CVs

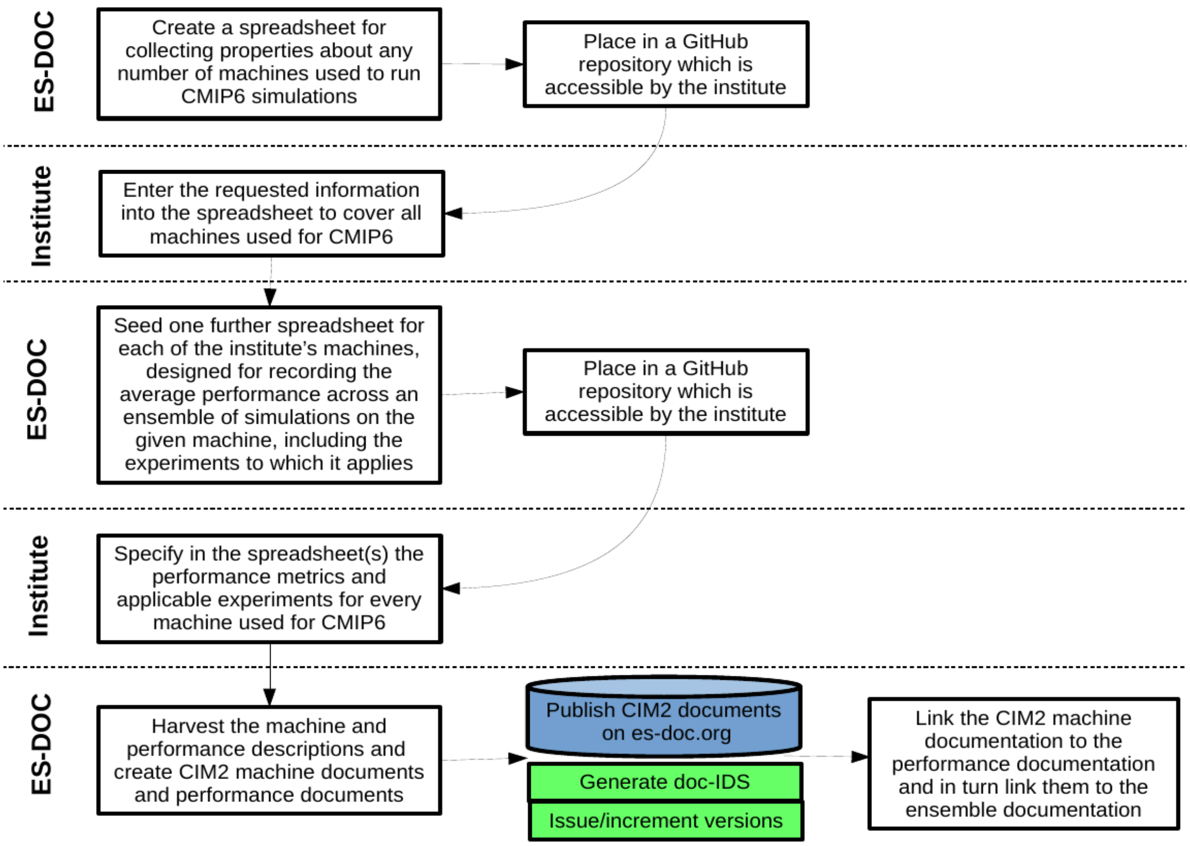

ES-DOC
Create a spreadsheet for collecting properties about any number of machines used to run CMIP6 simulations
Place in a GitHub repository which is accessible by the institute
Institute
Enter the requested information into the spreadsheet to cover all machines used for CMIP6
ES-DOC
Seed one further spreadsheet for each of the institute's machines, designed for recording the average performance across an ensemble of simulations on the given machine, including the experiments to which it applies
Place in a GitHub repository which is accessible by the institute
Institute
Specify in the spreadsheet(s) the performance metrics and applicable experiments for every machine used for CMIP6
ES-DOC
Harvest the machine and performance descriptions and create CIM2 machine documents and performance documents
Publish CIM2 documents on es-doc.org
Generate doc-IDS
Issue/increment versions
Link the CIM2 machine documentation to the performance documentation and in turn link them to the ensemble documentation

Figure 2.6.2: *The two-stage process for creating machine and performance documentation (https://es-doc.org/cmip6-machine-and-performance).*

When these completed machine spreadsheets are received by ES-DOC, not only can the machine descriptions be published, but the model and experiment selection will be used to create customised spreadsheets for collecting the minimum number performance descriptions required to span the modelling groups entire CMIP6 activity. ES-DOC then manages the required duplication of performances (as originally defined by the modelling groups in the machine collection spreadsheet) and publishes the performance descriptions for every experiment.

In order to ensure that the machine spreadsheets designed by ES-DOC, along with the associated guidance on how to fill them out, were as clear as possible and enabled the groups to document information to the level of detail they desired, the alpha draft version of the spreadsheet was shared with the liaisons in advance of the release, to provide the opportunity for feedback. Though only a small number of institutions did provide feedback, as reported in Appendix D, their comments were useful nonetheless to make small but important improvements to the alpha version, and the lack of feedback from most institutions was taken as a sign that there were no major issues with the spreadsheets.

2.6.4 Documentation for conformance to experimental requirements

Each CMIP6 experiment contains detailed requirements (also known as numerical requirements) that state how to configure a climate model in order to correctly define the experiment that is to be run. Each of these requirements may be met, not met, or partially met by a particular model, and the extent to which a model met all of the experiment's requirements can be important information when interpreting the output. An example of not meeting a requirement could be the need to impose pre-industrial concentrations of methane in the atmosphere in a model that can not represent methane in its atmosphere. The model may, however, be able to emulate the effect that methane has by applying a suitable concentration of a different greenhouse gas. It is useful to know that the requirement was not strictly met, but that an approximation to it was used instead.

An analysis of the numerical requirements showed that many are applicable to multiple experiments, and it is likely that all of an institute's models will have conformed in the same manner. Therefore it would be beneficial to only ask the modelling groups to describe these once, and then ES-DOC could manage their distribution to every applicable model and experiment combination. Since the experiments and model in use for each institute are defined in the CMIP6 controlled vocabularies, bespoke spreadsheets can be created by ES-DOC to minimise the effort required by the modelling groups to describe their conformances to experimental requirements.

2.7 Creating new documentation from existing sources

An important consideration in delivering a documentation collection system is to leverage any existing documentation efforts that may have already occurred within an institution or towards a different project activity. If these can be easily converted to CIM-format documents then there is no need for these documents to be manually created via spreadsheets. Or, in cases where a specific subset of relevant information has already been documented elsewhere, the appropriate spreadsheets may still be required to capture further information, but could be pre-populated with known data that needs only to be verified by the institute and not re-recorded.

For CMIP6, only one modelling group—the Met Office Hadley Centre (MOHC)—had an existing documentation database that could be used in the former way. Working with the ES-DOC team, a `pyesdoc`-based library was created at MOHC which could read their bespoke database of model descriptions, responsible parties and citations; convert them to CIM format documents; and push these documents up to a GitHub repository. This repository is monitored by ES-DOC, and any new content is automatically published into the ES-DOC archive of CIM documents. At the same time, the content is injected into the usual spreadsheets within the MOHC institutional repository[31], thus allowing future changes to be made easily without reference to the MOHC database. It was decided that other documentation types (machine, performance and experimental conformance) would be most easily produced via the spreadsheets, which posed no problems, as the document types could be collected independently of each other.

With regards to the latter case, opportunity to extract relevant information from existing documentation arose for the machine and performance documentation, since this information had already been collected for IS-ENES3 from early 2019 onwards, from a subset of eleven institutions on the CPMIP metrics (see section 2.3) measured during the running of CMIP6 experiments on their machines (as part of IS-ENES3 WP4[32]). Most of these institutes are also participating in ES-DOC, and those metrics form the basis of the performance CIM, so it was relatively straightforward to extract the values from the submissions from the existing effort and to allocate them in the spreadsheets towards those groups' machine and performance documentation. For an extra layer of quality control, such groups were asked during the process to confirm that the transferred data was correct.

2.8 Documenting the CMIP6 experiments

ES-DOC experiment documentation was generated from a range of sources that included the initial Overview of CMIP6-Endorsed MIPs (a collection of the proposal documents from the MIPs making the case for their inclusion in CMIP6), and later the published MIP papers[33], the CMIP6 data request[34] and the information collated by the CMIP6 controlled vocabulary (CV) team at PCMDI[35]. Initially the ES-DOC documentation was iterated via e-mail conversations with the principal investigators of the CMIP6 MIPs, later these conversations were mediated by the CV team at PCMDI with issues tracked via GitHub.

ES-DOC kept track as CMIP6 experiment names were changed, experiments were discarded, and new experiments were added, with documentation that allowed experiments to be recognised by their previous names and, later, the aliases that some CMIP6 experiments go by within their

[31] https://github.com/ES-DOC-INSTITUTIONAL/mohc

[32] IS-ENES3 WP4: Networking on Models, Tools and efficient use of HPC, https://is.enes.org/project/wp4-na3

[33] https://gmd.copernicus.org/articles/special_issue590.html

[34] Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M., and Sénési, S.: The CMIP6 Data Request (DREQ, version 01.00.31), Geosci. Model Dev., 13, 201–224, https://doi.org/10.5194/gmd-13-201-2020, 2020.

[35] https://github.com/WCRP-CMIP/CMIP6_CVs

originating research communities. A valuable resource for coordinating the federated effort to support an endeavour such as CMIP6.

The ES-DOC documentation process revealed occasional discrepancies that arose from the parallel nature of the CMIP6 workflow. For example, specifications could vary between what was published in a CMIP6-endorsed MIP's GMD paper and what had been agreed by the MIP authors with the data request and/or the PCMDI team with the controlled vocabulary. ES-DOC documentation exposed such issues, resulting in revisions all round.

The ES-DOC documentation process also revealed where savings could be made by identifying where experiment specifications were duplicated. For instance ES-DOC intervention saved 25 years of unnecessary simulation for the CDRMIP (Carbon Dioxide Removal MIP)[36] experiment esm-ssp534-over[37] by recommending an initialisation from esm-ssp585[38] in the year 2040 rather than esm-historical in the year 2015.

ES-DOC experiment documentation for CMIP6 can be used to see the relationship between the MIPs via their shared experiments (Figure 1.1b), and the relationship between experiments via their shared requirements. ES-DOC uses these relationships to streamline the collection of conformance information from modelling groups as they provide information about how their models were configured to conform to the requirements of the CMIP6 experiments.

2.9 Connecting different types of documentation with the further_info_URL

A key element arising from the feedback on the CMIP5 documentation process (section 1.2) was that of “interconnection”. For CMIP5 the different documentation types were disparate and it was difficult to collate all of the documentation relating to the entire workflow that produced a dataset. A dataset was produced by a simulation made by a model that met the requirements of an experiment, and was run on a machine with performance characteristics; and all (or a subset) of the descriptions of each of these elements will be relevant to the user of the output dataset. This, and other useful information, such as any dataset errata and the dataset citation, should be made available from a single and easy to access location.

[36] https://documentation.es-doc.org/cmip6/mips/cdrmip

[37] https://documentation.es-doc.org/cmip6/experiments/esm-ssp534-over

[38] https://documentation.es-doc.org/cmip6/experiments/esm-ssp585

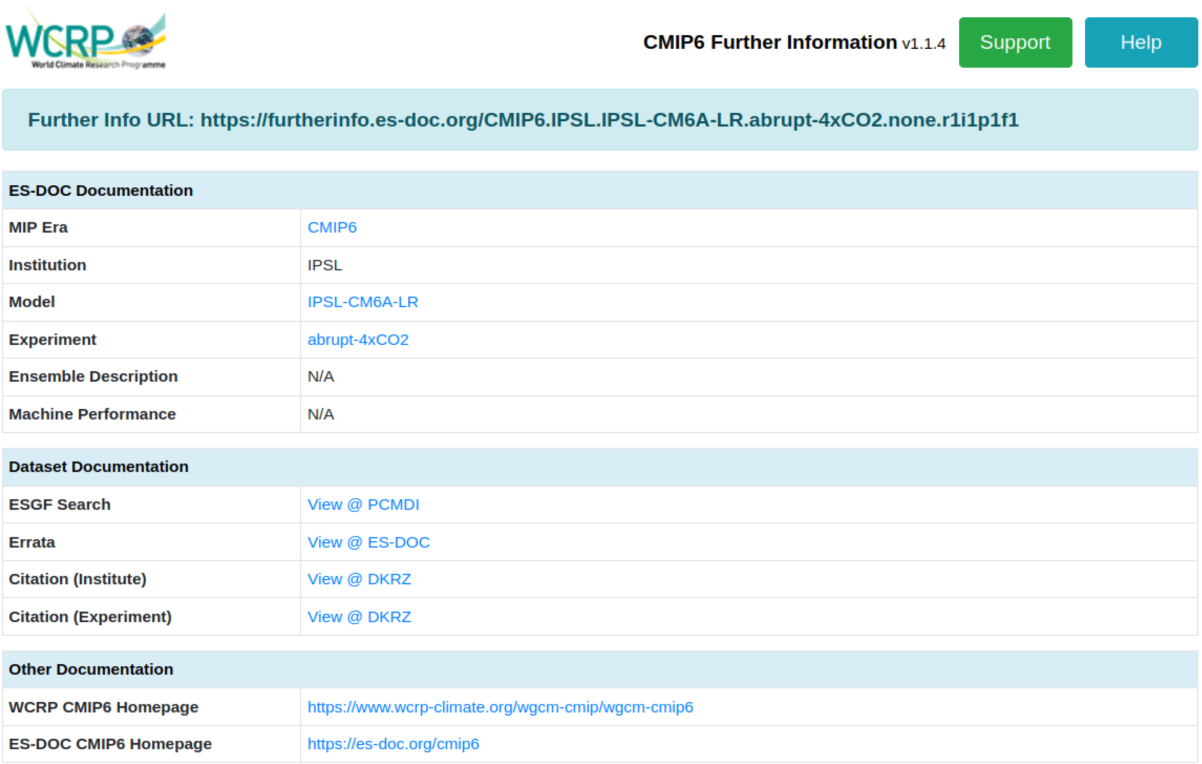
For CMIP6, access to all of these documents is made via the “further_info_URL” web page, which is recorded as part of the metadata in every CMIP6 netCDF file. The further_info_URL, that is unique to every CMIP6 simulation rather than each dataset or netCDF file, names a specific webpage that collates all of the required documents and has the form:

`http://furtherinfo.es-doc.org/`**`<simulation identifiers>`**

where **`<simulation identifiers>`** is a dot-separated collection of elements that define a simulation, comprising the:

- **MIP era** - the activity’s associated CMIP cycle (e.g. `CMIP6`)
- **Institution ID** - the institution identifier (e.g. `GFDL`)
- **Source ID** - the model identifier (e.g. `GFDL-CM2-1`)
- **Experiment ID** - the root experiment identifier (e.g. `abrupt4xCO2`)
- **Sub experiment ID** - the sub-experiment identifier (e.g. `s1965`). In the common case of no sub-experiment this will be `none`.
- **Variant label** - a label constructed from 4 indices describing the position of an individual simulation within the ensemble for the experiment (e.g. `r3i1p2f17` denotes the simulation created by the combination of the 3rd realisation, the 1st initialization of the 2nd physics and the 17th forcing methods.)

An example of an actual further_info_URL web page is https://furtherinfo.es-doc.org/CMIP6.IPSL.IPSL-CM6A-LR.abrupt-4xCO2.none.r1i1p1f1, which can be seen in Figure 2.9.



WCRP
World Climate Research Programme
CMIP6 Further Information v1.1.4
Support
Help

Further Info URL: https://furtherinfo.es-doc.org/CMIP6.IPSL.IPSL-CM6A-LR.abrupt-4xCO2.none.r1i1p1f1

ES-DOC Documentation	
MIP Era	CMIP6
Institution	IPSL
Model	IPSL-CM6A-LR
Experiment	abrupt-4xCO2
Ensemble Description	N/A
Machine Performance	N/A

Dataset Documentation	
ESGF Search	View @ PCMDI
Errata	View @ ES-DOC
Citation (Institute)	View @ DKRZ
Citation (Experiment)	View @ DKRZ

Other Documentation	
WCRP CMIP6 Homepage	https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6
ES-DOC CMIP6 Homepage	https://es-doc.org/cmip6

Figure 2.9: *The web page arrived at from the further_info_URL* *https://furtherinfo.es-doc.org/CMIP6.IPSL.IPSL-CM6A-LR.abrupt-4xCO2.none.r1i1p1f1*

Note that a single simulation will produce many datasets (typically one per output variable), each of which may comprise multiple netCDF files, and the further_info_URL contained in each of these files will be identical.

The connection of the simulation to the ES-DOC errata service is important, as it increases the chance that a user will be aware of any issues with data. See section 2.12 for details, in particular Figure 2.12a to see the errata for the simulation linked to by the further_info_URL of Figure 2.9.

The further_info_URL page also includes links to related services outside of ES-DOC, namely the data citation services for the institute and simulations[39] provided by DKRZ[40], and a connection to download the experiment's data from the ESGF archive.

[39] Stockhause, M. and Lautenschlager, M., 2017. CMIP6 Data Citation of Evolving Data. Data Science Journal, 16, p.30. DOI: http://doi.org/10.5334/dsj-2017-030

[40] Deutsche Klimarechenzentrum

2.10 Documenting CORDEX

The Coordinated Regional Climate Downscaling Experiment (CORDEX)[41] is a WCRP framework to evaluate regional climate models (RCMs) through a set of experiments aiming at producing regional climate projections. CORDEX is connected with CMIP6 through its designation as a Diagnostic MIP in CMIP6[42]. As such it has exactly the same documentation needs as CMIP6, but in practice the documentation process has started much later than for CMIP6 and has focussed solely on documenting the RCM formulations.

The European CORDEX RCMs being documented are all atmosphere-only, comprising the atmosphere, aerosol and land surface realms which generally have equivalent formulations to those of a CMIP GCM. However there are a small number of additional properties that are needed to document the atmospheric horizontal discretization[43] and how boundary conditions were set. Whilst building on the CMIP6 atmosphere specialisation, and using the CMIP aerosol and land surface specialisations to define the CORDEX model documentation content may have been possible, it was not clear that there was a reusability case for such infrastructure. As a result, the relevant CMIP6 specialisations were copied to new CORDEX specialisation repositories (see Table B.2 in appendix B), and documentation collection spreadsheets generated from those.

Completed documentation is published in the same ES-DOC archive as the CMIP6 documentation, and is also available from the ES-DOC website (https://es-doc.org/).

2.11 Documenting Obs4MIPs and input4MIPs

The Observations for Model Intercomparison Project (Obs4MIPs)[44] facilitates the use of observations in climate model evaluation and research, with a particular target being CMIP. As such it is an important part of the analysis workflow and would benefit from being well described for the CMIP6 data users. Obs4MIPs chose to implement the further_info_URL (section 2.9) in the observation dataset files, and the intention is to ensure that these web pages exist. In the first instance, they will contain a link to their existing (non-CIM) documentation. However, the CIM

[41] https://cordex.org

[42] Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F.: WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6, Geosci. Model Dev., 9, 4087–4095, https://doi.org/10.5194/gmd-9-4087-2016, 2016.

[43] https://github.com/ES-DOC/cordex-specializations-atmos/blob/master/atmos_grid.py

[44] Waliser, D., Gleckler, P. J., Ferraro, R., Taylor, K. E., Ames, S., Biard, J., Bosilovich, M. G., Brown, O., Chepfer, H., Cinquini, L., Durack, P. J., Eyring, V., Mathieu, P.-P., Lee, T., Pinnock, S., Potter, G. L., Rixen, M., Saunders, R., Schulz, J., Thépaut, J.-N., and Tuma, M.: Observations for Model Intercomparison Project (Obs4MIPs): status for CMIP6, Geosci. Model Dev., 13, 2945–2958, https://doi.org/10.5194/gmd-13-2945-2020, 2020.

has been designed to be able to capture dataset documentation, and in time it is hoped that we will be able to map the existing Obs4MIPs documentation to CIM format documents so that it can be more widely available via the ES-DOC archive, and can be easily linked to related artefacts.

The Input Datasets for Model Intercomparison Projects (input4mips)[45] initiative provides the forcing datasets that define the CMIP6 experiments. These are based on observations (e.g. greenhouse gas concentrations needed to force simulations of historical climate change), and others are based on hypothetical future conditions, such as in the forcing datasets employed in the future projections called for by ScenarioMIP[46]. These datasets have existing, non-CIM documentation, but investigations have not yet started as to how this documentation can be delivered via ES-DOC.

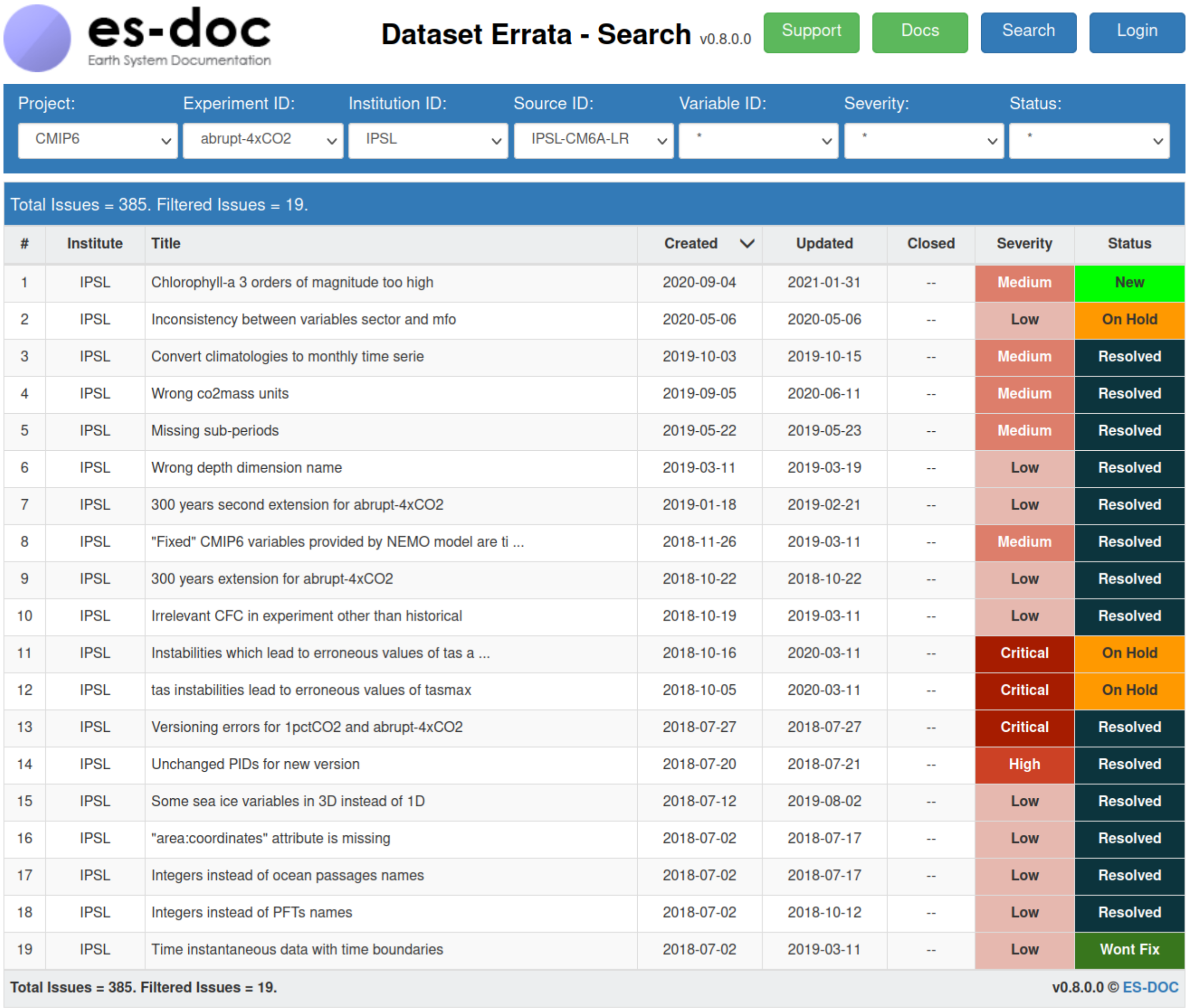
2.12 ES-DOC Errata Service

The ES-DOC Errata service[47] has been the community's answer to an issue arising from the complexity of projects like CMIP5 and CMIP6. The aim is to provide a platform that enables reasons that motivate dataset version change to be recorded and tracked. This, of course, should result in a considerable improvement of data quality, given the proper implementation of errata information handling, i.e. timely and accurate issue reporting with comprehensive updates. Errata are registered by both the modelling centres who created the original datasets, and general users of the data who have downloaded the data from the ESGF archive and discovered potential issues. In both cases, however, it is the responsibility of the data creators to acknowledge or refute the problem. If a problem is refuted, details of what are recorded in the errata database, but when a problem with the data is acknowledged there are typically three courses of action, all of which are fully record in the errata database, that may be taken by the modelling group:

- provide guidance on the issue that state the conditions under which the data should be used,
- retract the incorrect data and issue corrected versions,
- retract the incorrect data with no replacement.

An example of the errata for a particular simulation is shown in Figure 2.12a. The process is well documented[48] (e.g. Figure 2.12b) and the development team at IPSL offers guidance for first-time users to make the most out of the platform.

[45] https://www.osti.gov/servlets/purl/1463030
[46] https://documentation.es-doc.org/cmip6/mips/scenariomip
[47] https://errata.es-doc.org
[48] https://es-doc.github.io/esdoc-errata-client/index.html



Dataset Errata - Search v0.8.0.0

Project:	Experiment ID:	Institution ID:	Source ID:	Variable ID:	Severity:	Status:
CMIP6	abrupt-4xCO2	IPSL	IPSL-CM6A-LR	*	*	*

Total Issues = 385. Filtered Issues = 19.

#	Institute	Title	Created	Updated	Closed	Severity	Status
1	IPSL	Chlorophyll-a 3 orders of magnitude too high	2020-09-04	2021-01-31	--	Medium	New
2	IPSL	Inconsistency between variables sector and mfo	2020-05-06	2020-05-06	--	Low	On Hold
3	IPSL	Convert climatologies to monthly time serie	2019-10-03	2019-10-15	--	Medium	Resolved
4	IPSL	Wrong co2mass units	2019-09-05	2020-06-11	--	Medium	Resolved
5	IPSL	Missing sub-periods	2019-05-22	2019-05-23	--	Medium	Resolved
6	IPSL	Wrong depth dimension name	2019-03-11	2019-03-19	--	Low	Resolved
7	IPSL	300 years second extension for abrupt-4xCO2	2019-01-18	2019-02-21	--	Low	Resolved
8	IPSL	"Fixed" CMIP6 variables provided by NEMO model are ti ...	2018-11-26	2019-03-11	--	Medium	Resolved
9	IPSL	300 years extension for abrupt-4xCO2	2018-10-22	2018-10-22	--	Low	Resolved
10	IPSL	Irrelevant CFC in experiment other than historical	2018-10-19	2019-03-11	--	Low	Resolved
11	IPSL	Instabilities which lead to erroneous values of tas a ...	2018-10-16	2020-03-11	--	Critical	On Hold
12	IPSL	tas instabilities lead to erroneous values of tasmax	2018-10-05	2020-03-11	--	Critical	On Hold
13	IPSL	Versioning errors for 1pctCO2 and abrupt-4xCO2	2018-07-27	2018-07-27	--	Critical	Resolved
14	IPSL	Unchanged PIDs for new version	2018-07-20	2018-07-21	--	High	Resolved
15	IPSL	Some sea ice variables in 3D instead of 1D	2018-07-12	2019-08-02	--	Low	Resolved
16	IPSL	"area:coordinates" attribute is missing	2018-07-02	2018-07-17	--	Low	Resolved
17	IPSL	Integers instead of ocean passages names	2018-07-02	2018-07-17	--	Low	Resolved
18	IPSL	Integers instead of PFTs names	2018-07-02	2018-10-12	--	Low	Resolved
19	IPSL	Time instantaneous data with time boundaries	2018-07-02	2019-03-11	--	Low	Wont Fix

Total Issues = 385. Filtered Issues = 19.

v0.8.0.0 © ES-DOC

Figure 2.12a: *The errata entries for the CMIP6 simulation described by the further_info_URL shown in Figure 2.9.*

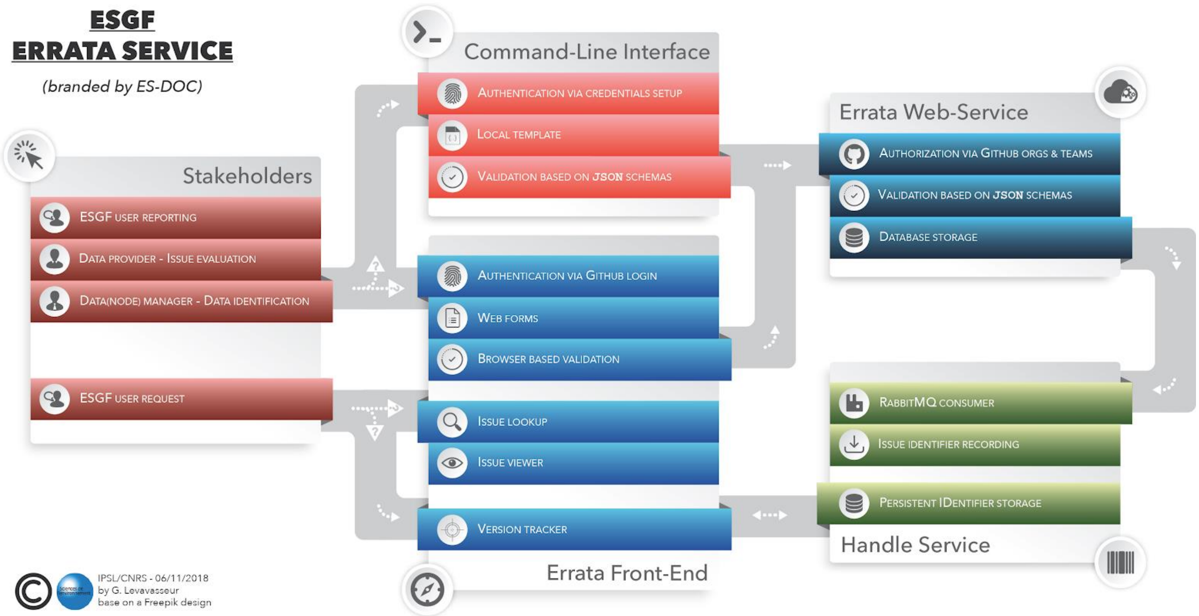


Figure 2.12b: *The components of the errata service*

2.13 ES-DOC beyond CMIP6

The entire ES-DOC infrastructure, whilst primarily created to serve as a documentation framework for CMIP, is intended to be portable to other projects that follow a similar workflow pattern. This pattern of

1. build numerical models
2. design experiments
3. configure the models for each experiment
4. run model simulations for each experiment
5. make data output available to users

is applicable to many projects that use numerical simulation to create its results, and the ES-DOC framework can be used to document each stage, regardless of the scientific nature of the project. Figure 2.13 describes how the documentation process was connected to this pattern for CMIP6.

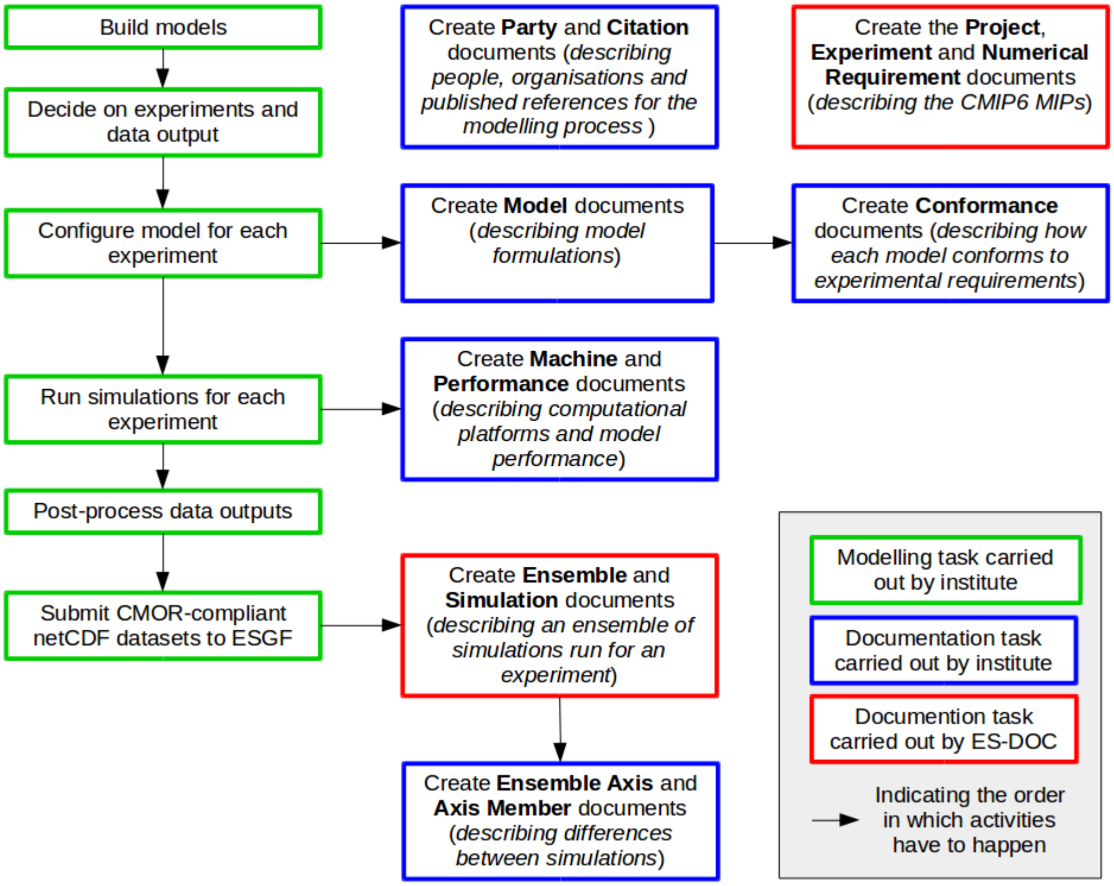

Build models
Decide on experiments and data output
Configure model for each experiment
Run simulations for each experiment
Post-process data outputs
Submit CMOR-compliant netCDF datasets to ESGF
Create Party and Citation documents (describing people, organisations and published references for the modelling process)
Create Model documents (describing model formulations)
Create Machine and Performance documents (describing computational platforms and model performance)
Create Ensemble and Simulation documents (describing an ensemble of simulations run for an experiment)
Create Ensemble Axis and Axis Member documents (describing differences between simulations)
Create the Project, Experiment and Numerical Requirement documents (describing the CMIP6 MIPs)
Create Conformance documents (describing how each model conforms to experimental requirements)
Modelling task carried out by institute
Documentation task carried out by institute
Documention task carried out by ES-DOC
Indicating the order in which activities have to happen

Figure 2.13: *The CMIP6 numerical modelling workflow and the creation of documentation for each step.*

The application of parts of the ES-DOC system to the CORDEX project (described in section 2.10) demonstrated that the ES-DOC infrastructure was certainly very easy to re-apply, but this was perhaps not a very stringent test since the application was carried by the ES-DOC team, rather than a third party with no prior experience in the field. It is recognised that for ES-DOC to have future use for as-yet-unknown projects and by as-yet-unknown people, ES-DOC itself must be well documented so that the it can be understood and implemented without the assumption that the current ES-DOC team will be available to assist. A considerable on-going effort is being put towards creating a collection of ES-DOC technical documentation, which will eventually be collated at https://technical.es-doc.org, the ultimate aim of which is be comprehensive enough to allow any one to create and deliver documentation for any project.

However well it may be documented, it is also crucial that the CIM that underpins all of the documentation process is capable of representing an arbitrary modelling workflow of the type described above. It was noted in section 2.1 that CIM2 is less CMIP-specific than its predecessor, however it is not yet entirely independent of this particular project. Work is underway on CIM3, a new version that not only aims to remove CMIP6 dependencies, but also will feature some redesigned elements that follow from the experience of actually using CIM2 in a real project.

3 Conclusions and recommendations

3.1 Conclusions

This report describes the work of ES-DOC towards delivering a sustainable documentation framework that can be applied for CMIP6 that will also have utility for future numerical modelling projects (climate-related or otherwise) that are not supported by IS-ENES3. What has not been mentioned up to this point is the amount of documentation content that has actually been created by ES-DOC and by the CMIP6 modelling groups.

Comprehensive descriptions of all 314 of the CMIP6 experiments were published very early on in the process. Whilst it was the case that the necessary ES-DOC model documentation infrastructure required by was delivered on time (December 2018), the content that needed to be provided by the modelling centres has not been fully delivered. Only 15 of the 53 participating institutes have so far engaged with the model documentation process, publishing documentation for 37[49] out of the 113 models that have published data to ESGF, with varying degrees of completeness.

The simulation documents are yet to be published, even though information has been successfully harvested from the ESGF publication process (see section 2.6.1), publication is expected in early 2022. The necessary infrastructure for documenting conformance to experimental requirements, has not yet been delivered. The framework for publishing machine and performance documentation was released by ES-DOC in the second half of 2021, much later than hoped. However, some groups had already collected their machine and performance records and ES-DOC was able to ingest this metadata into published documentation on their behalf.

The late delivery of the documentation infrastructure that was a problem for CMIP5 (see section 1.2) that was unfortunately repeated for CMIP6, albeit for different reasons. These delays in delivery of ES-DOC infrastructure were regrettable, but perhaps not surprising given an ambitious program of novel work. Recruitment and contractual issues for key staff, and the impact of COVID-19 on the available work time all contributed delays in delivery. For example, much time was lost when the planned on-line questionnaire for CMIP6 document collection had to be abandoned due

[49] https://documentation.es-doc.org/cmip6/models

to staff leaving the project. A new method (Jupyter Notebooks) was then developed but discarded after beta testing; and a final method (spreadsheets) developed, tested, and put into production (see section 2.5). Moving away from the on-line questionnaire also created unforeseen work in other areas. For instance, a means of automatically updating published documentation with further changes made by the modelling groups was a solved problem with the questionnaire, since that had been developed for the CMIP5 documentation project, but this issue needed to be reconsidered when using spreadsheets in GitHub repositories.

Scientific requirements are the drivers of a numerical modelling project, not the ease of creating its documentation, but looking to the future much effort has been spent on creating a design and software infrastructure that is maintainable and reusable (see section 2.13). Therefore there is good cause to think that applying ES-DOC to another similar project (the proposed next phase of CMIP, CMIP7[50], perhaps) could be done in a timely and efficient manner—an adaptation is a much lower risk activity than the design-and-implement exercise performed here.

It was expected that not all modelling groups would attach a high priority to creating comprehensive documentation for CMIP6, perhaps citing the understandable reasons of the time needed to create it, and perceived lack of benefit to the creators. To counter this, efforts were made to remove technical barriers that could hinder engagement, and to advertise the benefits that could be seen in particular by the documentation creators.

The efforts to make the documentation easier for the modelling groups to create their documentation are covered in detail in section 2. The late delivery of the infrastructure collecting some documentation type (machine, performance, and conformance to experimental requirements) is clearly outside of the modelling groups control, however the collection of model descriptions was delivered in good time, so it is useful to consider possible reasons for why only 33% of the published CMIP6 models (37 out of 113) have so far been documented within ES-DOC. The first thing to note is that a percentage of models with documentation is a very crude measure that hides much detail. For instance, if we look at the 50 models that have participated in the key DECK[51] and ScenarioMIP[52] experiments, then the percentage of documented models rises to 50%.

It is a possibility that the amount of documentation needed for CMIP6 could have been large enough that it appeared to be an intractable problem given the finite resources of the modelling groups, despite the best efforts of ES-DOC to create an easy-to-use system. Another possibility is

[50] E.g. https://www.wcrp-climate.org/images/modelling/WGCM/WGCM23/Presentations/3d_WGCM23_WIP_options_for_the_future.pdf
[51] https://documentation.es-doc.org/cmip6/mips/deck
[52] https://documentation.es-doc.org/cmip6/mips/scenariomip

that the benefits to the modelling centres of having a comprehensive documentation resource may not have been sufficiently advertised. In the second case, quantifying the “benefit” is difficult. It may be argued that most of the benefit goes to dataset users who are outside of the modelling group. However, one could counter this by noting that the modelling group itself will be an external user of other institutions’ data; and having readily available documentation may encourage wider use of the data, so that the data creators also benefit from wider recognition and citation.

3.2 Recommendations

ES-DOC will deliver a complete documentation service for CMIP6 by the end of the IS-ENES3 project (i.e. by the end of 2022), but it is crucial that the archived documentation is available for many years to come, to at least match the useful lifetime of the CMIP6 data archive. In addition, the ES-DOC infrastructure should be available to future projects.

The IS-ENES3 sustainability activity[53] covers ES-DOC and will be looking at three associated issues:

- Service management, and
- Software maintenance, and
- Ability to reuse the infrastructure.

Service management will require ensuring that the web services are managed and upgraded as necessary. While the software should not require any more development after 2023 it will require resources to upgrade (e.g. if the server environment requires upgrades for security reasons, libraries and interfaces may change and need to be updated).

The current services are maintained on a commercial host, but this has been problematic. In 2020 the then commercial provider for all of the ES-DOC web service (WebFaction) ceased to exist, and services were migrated to a new environment hosted by OpalStack[54]. This was a non-trivial task, although one that would be easier now that it has been done before and with a consolidation of the related infrastructure. There are also financial implications, while the sums involved for acquiring a commercial server are small, the mechanisms for community ownership will be difficult to sustain on their own.

Accordingly, it is recommended that:

[53] https://portal.enes.org/ISENES3/project/wp2-na1

[54] https://www.opalstack.com

1. Web services should be transferred to an institution with CMIP and IS-ENES connections (such as CEDA[55] or IPSL[56], who already have close connections with ES-DOC) who could take responsibility to manage and guarantee the provision, and
2. The overall IS-ENES3 sustainability activity should also address ES-DOC software maintenance. How this is to be done is not covered by this report, but this activity should consider creating a hierarchy of ES-DOC services so the most critical ones (by consensus) can be conserved, in the hopefully unlikely event that resources are not forthcoming to provide support for everything that ES-DOC provides

Ideally the investment in ES-DOC will support other activities, both nationally and internationally. Accordingly:

3. ES-DOC should continue to engage with the WGCM Infrastructure Panel (WIP), the new CMIP International Project Office[57], and the Copernicus Climate Change Service (C3S)[58] on the best way to document future global climate modelling projects, such as the proposed CMIP7, and
4. Continue to work with national activities, especially those which might involve large communities sharing simulations carried out on centralised national platforms

All of these actions will of course require funding to provide the required staff time, and the sustainability program is tasked with investigating the best mechanisms to enable this outside of bespoke funding provided by IS-ENES3.

Other recommendations:

5. Future projects should maintain the link from the output datasets to the applicable ES-DOC information, as is done by the further_info_URL embedded within each file (see section 2.9). This should incentivize users to consult ES-DOC as their first source of information, and
6. Include the possibility to link a model's description to the configuration to another model. This is already possible within the CIM but was not been utilised for CMIP6, and
7. Provide the capability to show the differences between published documents, e.g. to see how a model differs from a previous model of the same family.

[55] Centre for Environmental Data Analysis, https://www.ceda.ac.uk

[56] Institut Pierre-Simon Laplace, https://www.ipsl.fr

[57] https://climate.esa.int/en/news-events/esa-to-host-global-climate-modelling-project-office/

[58] https://climate.copernicus.eu

8. Include a link within the netCDF file metadata, similar to the further_info_URL, that connects directly to the errata service for that dataset.
9. Consider sharing the tools to go from spreadsheets to publishable CIM2 documents rather than handling this within the ES-DOC project. This would help with version control and being able to view and understand the differences between comparable documents.
10. Include documentation of conformance to the data request, that defines all of the model variables that should be output from each simulation.

Appendix A. The institutes participating in CMIP6

Table A.1: *The participants in the CMIP6 project, and the location of their ES-DOC institutional GitHub repository, the primary means of creating documentation. The canonical record of the institutes and their official CMIP6 abbreviations is found at https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_institution_id.json.*

Institute/consortium	Abbreviation	ES-DOC institutional GitHub repository
Atmospheric and Environmental Research, USA	AER	github.com/ES-DOC-INSTITUTIONAL/aer
Research Center for Environmental Changes, Academia Sinica, Taiwan	AS-RCEC	github.com/ES-DOC-INSTITUTIONAL/as-rcec
Alfred Wegener Institute, Germany	AWI	github.com/ES-DOC-INSTITUTIONAL/awi
Beijing Climate Center, China	BCC	github.com/ES-DOC-INSTITUTIONAL/bcc
Beijing Normal University, China	BNU	github.com/ES-DOC-INSTITUTIONAL/cams
Chinese Academy of Meteorological Sciences, China	CAMS	github.com/ES-DOC-INSTITUTIONAL/cams
Chinese Academy of Sciences, China	CAS	github.com/ES-DOC-INSTITUTIONAL/cas
CCCR (Centre for Climate Change Research), India IITM (Indian Institute of Tropical Meteorology), India	CCCR-IITM	github.com/ES-DOC-INSTITUTIONAL/cccr-iitm
Canadian Centre for Climate Modelling and Analysis, Canada	CCCma	github.com/ES-DOC-INSTITUTIONAL/cccma
Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Italy	CMCC	github.com/ES-DOC-INSTITUTIONAL/cmcc
CNRM (Centre National de Recherches Météorologiques, France CERFACS (Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, France	CNRM-CERFACS	github.com/ES-DOC-INSTITUTIONAL/cnrm-cerfacs
CSIR (Council for Scientific and Industrial Research), Wits (University of the Witwatersrand), Australia CSIRO (Commonwealth Scientific and Industrial Research Organisation), Australia	CSIR-Wits-CSIRO	github.com/ES-DOC-INSTITUTIONAL/csir-wits-csiro
Commonwealth Scientific and Industrial Research Organisation, Australia	CSIRO	github.com/ES-DOC-INSTITUTIONAL/csiro

CSIRO (Commonwealth Scientific and Industrial Research Organisation), Australia ARCCSS (Australian Research Council Centre of Excellence for Climate System Science), Australia	CSIRO-ARCCSS	github.com/ES-DOC-INSTITUTIONAL/csiro-arccss
CSIRO (Commonwealth Scientific and Industrial Research Organisation), Australia COSIMA (Consortium for Ocean-Sea Ice Modelling in Australia), Australia	CSIRO-COSIMA	github.com/ES-DOC-INSTITUTIONAL/csiro-cosima
Deutsches Klimarechenzentrum, Germany	DKRZ	github.com/ES-DOC-INSTITUTIONAL/dkrz
Deutscher Wetterdienst, Germany	DWD	github.com/ES-DOC-INSTITUTIONAL/dwd
LLNL (Lawrence Livermore National Laboratory), USA ANL (Argonne National Laboratory), USA BNL (Brookhaven National Laboratory), USA LANL (Los Alamos National Laboratory), USA LBNL (Lawrence Berkeley National Laboratory), USA ORNL (Oak Ridge National Laboratory), USA PNNL (Pacific Northwest National Laboratory), USA SNL (Sandia National Laboratories), USA	E3SM-Project	github.com/ES-DOC-INSTITUTIONAL/e3sm-project

AEMET, Spain BSC, Spain CNR-ISAC, Italy DMI, Denmark ENEA, Italy FMI, Finland Geomar, Germany ICHEC, Ireland ICTP, Italy IDL, Portugal IMAU, Netherlands IPMA, Portugal KIT, Germany KNMI, Netherlands Lund University Met Eireann, Ireland NLeSC, Netherlands NTNU, Norway Oxford University, UK surfSARA, Netherlands SMHI, Sweden Stockholm University, Sweden Unite ASTR, Belgium University College Dublin University of Bergen, Norway University of Copenhagen, Denmark University of Helsinki, Finland University of Santiago de Compostela, Spain Uppsala University, Sweden Utrecht University, Netherlands Vrije Universiteit Amsterdam, Netherlands Wageningen University, Netherlands	EC-Earth-Consortium	github.com/ES-DOC-INSTITUTIONAL/ec-earth-consortium
European Centre for Medium-Range Weather Forecasts, UK	ECMWF	github.com/ES-DOC-INSTITUTIONAL/ecmwf
FIO (First Institute of Oceanography, Ministry of Natural Resources), China QNLM (Qingdao National Laboratory for Marine Science and Technology), China	FIO-QLNM	github.com/ES-DOC-INSTITUTIONAL/fio-qlnm

ETH Zurich, Switzerland Max Planck Institut fur Meteorologie, Germany Forschungszentrum Jülich, Germany University of Oxford, UK Finnish Meteorological Institute, Finland Leibniz Institute for Tropospheric Research Center for Climate Systems Modeling (C2SM)	HAMMOZ-Consortium	github.com/ES-DOC-INSTITUTIONAL/hammoz-consortium
Institute for Numerical Mathematics, Russian Academy of Science, Russia	INM	github.com/ES-DOC-INSTITUTIONAL/inm
National Institute for Space Research, Brazil	INPE	github.com/ES-DOC-INSTITUTIONAL/inpe
Institut Pierre Simon Laplace, France	IPSL	github.com/ES-DOC-INSTITUTIONAL/ipsl
Korea Institute of Ocean Science and Technology, Republic of Korea	KIOST	github.com/ES-DOC-INSTITUTIONAL/kiost
Lawrence Livermore National Laboratory, USA	LLNL	github.com/ES-DOC-INSTITUTIONAL/llnl
The Modular Earth Submodel System (MESSy) Consortium, Germany	MESSy-Consortium	github.com/ES-DOC-INSTITUTIONAL/messy-consortium
"JAMSTEC (Japan Agency for Marine-Earth Science and Technology), Japan AORI (Atmosphere and Ocean Research Institute), Japan NIES (National Institute for Environmental Studies), Japan R-CCS (RIKEN Center for Computational Science), Japan	MIROC	github.com/ES-DOC-INSTITUTIONAL/miroc
Met Office Hadley Centre, UK	MOHC	github.com/ES-DOC-INSTITUTIONAL/mohc
Max Planck Institute for Meteorology, Germany	MPI-M	github.com/ES-DOC-INSTITUTIONAL/mpi-m
Meteorological Research Institute, Japan	MRI	github.com/ES-DOC-INSTITUTIONAL/mri
Goddard Institute for Space Studies, USA	NASA-GISS	github.com/ES-DOC-INSTITUTIONAL/nasa-giss
NASA Goddard Space Flight Center, USA	NASA-GSFC	github.com/ES-DOC-INSTITUTIONAL/nasa-gsfc
National Center for Atmospheric Research, USA	NCAR	github.com/ES-DOC-INSTITUTIONAL/ncar

NorESM Climate modeling Consortium: CICERO, Oslo MET-Norway, Norway NERSC, Norway, NILU, Norway University of Bergen, Norway University of Oslo, Norway Uni Research, Norway	NCC	github.com/ES-DOC-INSTITUTIONAL/ncc
Natural Environment Research Council, UK	NERC	github.com/ES-DOC-INSTITUTIONAL/nerc
NIMS (National Institute of Meteorological Sciences), Republic of Korea KMA (Korea Meteorological Administration), Republic of Korea	NIMS-KMA	github.com/ES-DOC-INSTITUTIONAL/nims-kma
National Institute of Water and Atmospheric, New Zealand	NIWA	github.com/ES-DOC-INSTITUTIONAL/niwa
National Oceanic and Atmospheric Administration, USA Geophysical Fluid Dynamics Laboratory, USA	NOAA-GFDL	github.com/ES-DOC-INSTITUTIONAL/noaa-gfdl
National Taiwan University, Taiwan	NTU	github.com/ES-DOC-INSTITUTIONAL/ntu
Nanjing University of Information Science and Technology, China	NUIST	github.com/ES-DOC-INSTITUTIONAL/nuist
Program for Climate Model Diagnosis and Intercomparison, USA	PCMDI	github.com/ES-DOC-INSTITUTIONAL/pcmdi
Pacific Northwest National Laboratory, USA	PNNL-WACCEM	github.com/ES-DOC-INSTITUTIONAL/pnnl-waccem
AER, USA University of Colorado, USA	RTE-RRTMGP-Consortium	github.com/ES-DOC-INSTITUTIONAL/rte-rrtmgp-consortium
ORNL, USA ANL, USA BNL, USA LANL, USA LBNL, USA Northern Arizona University, USA NCAR, USA University of California Irvine, USA University of Michigan, USA	RUBISCO	github.com/ES-DOC-INSTITUTIONAL/rubisco
Seoul National University, Republic of Korea	SNU	github.com/ES-DOC-INSTITUTIONAL/snu

Tsinghua University, China	THU	github.com/ES-DOC-INSTITUTIONAL/thu
University of Arizona, USA	UA	github.com/ES-DOC-INSTITUTIONAL/ua
University of California Irvine, USA	UCI	github.com/ES-DOC-INSTITUTIONAL/uci
Universitat Hamburg, Germany	UHH	github.com/ES-DOC-INSTITUTIONAL/uhh
University of Tasmania, Australia	UTAS	github.com/ES-DOC-INSTITUTIONAL/utas
University of Toronto, Canada	UofT	github.com/ES-DOC-INSTITUTIONAL/uoft

Appendix B. The model document specialisation GitHub repositories

Table B.1: *GitHub repositories that contain the canonical definition of the CMIP6 model documentation specialisations (section 2.2).*

Realm	ES-DOC CMIP6 specialisation GitHub repository
Top-level	github.com/ES-DOC/cmip6-specializations-toplevel
Atmosphere	github.com/ES-DOC/cmip6-specializations-atmos
Atmospheric aerosols	github.com/ES-DOC/cmip6-specializations-aerosol
Atmospheric chemistry	github.com/ES-DOC/cmip6-specializations-atmoschem
Land ice	github.com/ES-DOC/cmip6-specializations-landice
Land surface	github.com/ES-DOC/cmip6-specializations-land
Ocean	github.com/ES-DOC/cmip6-specializations-ocean
Ocean biogeochemistry	github.com/ES-DOC/cmip6-specializations-ocnbgchem
Sea ice	github.com/ES-DOC/cmip6-specializations-seaice

Table B.2: *GitHub repositories that contain the canonical definition of the CORDEX model documentation specialisations (section 2.10).*

Realm	ES-DOC CORDEX specialisation GitHub repository
Top-level	github.com/ES-DOC/cordex-specializations-toplevel
Atmosphere	github.com/ES-DOC/cordex-specializations-atmos
Atmospheric aerosols	github.com/ES-DOC/cordex-specializations-aerosol
Land surface	github.com/ES-DOC/cordex-specializations-land

Appendix C. Feedback from the document creation beta-testing process

The detailed reviews from the groups who beta tested the document creation process (section 2.4) have been collated and are given *in italics* after each of the review questions that was asked.

- General CMIP6 documentation process (for all steps/documents):
    - *The ES-DOC documentation is in general good and the workflow diagrams in particular are excellent. Overall, the ES-DOC approach was considered suitable for documenting CMIP6 models and simulations, and an improvement over CMIP5's metadata documentation.*

    - Clarity of overall process proposed (https://es-doc.org/cmip6)

    - Clarity of detailed steps and related instructions:
        - Liaison checklist (https://es-doc.org/cmip6-liaison-checklist)
            - *Well organised and well documented. A concern was raised over the multiple occasions of having to contact a single person at ES-DOC in order to move to the next stage, as this proved to be a bottleneck on occasion. The liaison checklist was an on-line Google Doc, which was problematic as Google services are not universally available.*
        - IPython model documentation instructions (link above)
            - *Well documented and understandable.*
        - ES-DOC/CMIP6 web site
            - *Well designed, well organised and well documented*

    - Agility/modularity of process:
        - *The modular document structure gives flexibility to input information as it comes in the workflow without having to wait for all information (e.g. publish a simulation document without certainty of the citation information), and then the freedom to correct mistakes in individual documents (update citation information and have the simulation document automatically point to the updated one).*
        - Combined use of github, spreadsheets and Jupyter notebooks
            - *Using github to organise, store, and collect documents was considered convenient, lightweight, and requires no training or introduction. The use of binary files (spreadsheets and PDFs) in the institutional repository was noted to be acceptable, but not ideal.*

    - Capacity for modelling group to organise their CMIP6 documentation work following their own timeline
      - *There was no general feeling for how much effort the process will take, as it may depend on how many experts are in the modelling group and how much time the modelling group is given to prepare the documentation.*

  - Quality of user support
    - *The support during the beta-testing phase was considered satisfactory.*

- Model documentation specific questions:
  - Ease of use of the Jupyter tool (login, look and feel, save/publish,…)
    - *Editing the IPython notebook files was not considered a problem, but there were concerns about only being able to do so on the ES-DOC JupyterHub server. It is a disadvantage that the IPython notebooks do not have an "undo" feature to correct mistakes. Being able to edit the documentation off-line is not possible with the IPython notebooks, but was a requested feature. The IPython notebooks were seen as requiring too much instruction to access and use, but the most tech-savvy scientists were comfortable with the method. It was noted that the "Publish" nor "Validate" buttons were not present in the iPython notebooks. Whilst eminently learnable, the python jargon was a cause for concern, as it limited the number of people who could access them. Also, it would be preferable that the notebooks be more rigid in terms of structure - cells can be added/moved/deleted/etc. which could affect the processing software.*

  - Seeding functionality (from CMIP5 and CMIP6, specification in model configuration file)
    - *Seeding is a feature welcomed by all. Some found seeding from CMIP5 more important than seeding from CMIP6, and vice versa. It was noted that for some realms, seeding from CMIP5 did not answer many of the CMIP6 documentation questions.*

  - Tools to interact with home scientists to collect information (PDF, spreadsheets, html and mindmaps)
    - *The PDF, html and mind maps were considered not useful for collecting model information. The spreadsheets were useful, but it was not ideal that information entered in the spreadsheets had to be manually copied to the IPython notebooks. Mindmaps might be useful if they could be generated from completed model documents and accessed on-line. The PDF printouts were considered largely*

redundant when used alongside the spreadsheets. The spreadsheets were popular and in general required only a bit of personal instruction.

- Clarity of instructions
    - *The IPython instructions are clear and good*

Appendix D. Feedback about the alpha version machine spreadsheet

Institution names and details have been removed from the comments to anonymise responses, and ES-DOC responses provide a summary of our action to address the feedback in question.

- *Comment*: I think pretty much everything is clear for me, except maybe section 1.9.2 ("applicable experiments by MIP"): I'm a bit confused concerning the ENUM items which are given as choices; following the example tab, I expected to find a list of possibly relevant experiments by MIP with their actual nomenclatures, but I can only see some generic "Experiment 1" types of answers. Could you please clarify this point to me?
    - *ES-DOC response*: The generic text was just a facet of the draft sent out, which was a generic template that would be specialised to each institute upon release. This was clarified with the liaison.
- *Comment*: The frontmatter says: "each tab should record the properties of a single machine (or partition thereof). If you have multiple machines to record for your institute, you should make copies of the ‘Machine 1’ tab so there is one tab per machine.". Our institute has two identical halls, each which contain two machines. Since we have gone through an upgrade during CMIP6, this means 8 machines. That is a lot of information to track. The instructions seem to say we should report each machine separately - for us, the machines in each hall are identical, so is this really necessary?
    - *ES-DOC response*: It was explained to the liaison that only minor duplication was necessary in this case, as the spreadsheet was designed in a way to reduce the need to repeat information.
- *Comment*: Under section 1.5 it talks about "storage pools", and says "All institutes have no more than two storage pools per machine as far as we are aware. If you only have one for this machine, please leave all answers to ‘1.5.2.X’ blank to indicate this. In the unlikely case you do have more than two storage pools for any machine, please contact the ES-DOC team via support@es-doc.org." . Looking at the example, we definitely have far more than 2 storage pools by this definition. These include a variety of frontend machine partitions, an independent mass-storage machine, an extensive tape archive, and of course separate storage for the ESGF (one pool for staging and one pool for serving up the data). It is not exactly clear what should be included nor whether this granularity is useful.
    - *ES-DOC response*: Guidance was provided on how this particular institute could classify the amount of storage pools they had according to the CIM and how they should document each of those with the alpha version spreadsheet.
- *Comment*: In our systems, we have a distinction between backend machines, where simulations are run, and frontend machines, where diagnostics are run and longer term storage is accessed. It is not clear to us how to distinguish these on the existing spreadsheets.

  - *ES-DOC response*: Advice was given regarding the distinction between the front-end and back-end machines and how that would be reflected in the alpha version spreadsheet.
- *Comment*: Filling out the list of experiments that were run on each machine is labour intensive and represents a non-trivial task for an already overburdened staff. There should be some sensitivity to the effort required vs. the value of the information gleaned. For example, we can see the value in providing the machine information and perhaps an overview of what was run, but this fine grained detail is expensive to generate. So, for those being asked to undertake this exercise, there needs to be a clear rationale provided for the value of the specific information requested and a connection made to how we might benefit from investing effort in this.
  - *ES-DOC response*: As covered in Section 3.1, a major difficulty for the ES-DOC project is that groups are often not convinced of the benefits of providing documentation, and the best that can be achieved is to provide a strong written case to try to encourage engagement, as was done in this case, where the benefits to the institute in question were specifically outlined.