Coordinating an operational data distribution network for CMIP6 data

Ruth Petrie[1], Sébastien Denvil[2], Sasha Ames[3], Guillaume Levavasseur[2], Sandro Fiore[4,5], Chris Allen[6], Fabrizio Antonio[4], Katharina Berger[7], Pierre-Antoine Bretonnière[8], Luca Cinquini[9], Eli Dart[10], Prashanth Dwarakanath[11], Kelsey Druken[6], Ben Evans[6], Laurent Franchistéguy[12], Sébastien Gardoll[2], Eric Gerbier[12], Mark Greenslade[2], David Hassell[13], Alan Iwi[1], Martin Juckes[1], Stephan Kindermann[7], Lukasz Lacinski[14], Maria Mirto[4], Atef Ben Nasser[2], Paola Nassisi[4], Eric Nienhouse[15], Sergey Nikonov[16], Alessandra Nuzzo[4], Clare Richards[6], Syazwan Ridzwan[6], Michel Rixen[17], Kim Serradell[8], Kate Snow[6], Ag Stephens[1], Martina Stockhause[7], Hans Vahlenkamp[16], and Rick Wagner[14]

[1]UK Research and Innovation, Swindon, UK
[2]Institut Pierre Simon Laplace; Centre National de Recherche Scientifique, Paris, France
[3]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, USA
[4]Euro-Mediterranean Center on Climate Change Foundation, Lecce, Italy
[5]University of Trento, Trento, Italy
[6]National Computational Infrastructure, Australian National University, Canberra, Australia
[7]German Climate Computing Center, Hamburg, Germany
[8]Barcelona Supercomputing Center, Barcelona, Spain
[9]Jet Propulsion Laboratory, National Aeronautics and Space Administration, Pasadena, USA
[10]Energy Sciences Network, Lawrence Berkeley National Laboratory, Berkeley, USA
[11]Linköping University, Linköping Sweden
[12]Centre National de Recherches Météorologiques, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[13]National Centre of Atmospheric Science, Leeds, UK
[14]Argonne National Laboratory, Chicago, USA
[15]National Center for Atmospheric Research, Boulder, USA
[16]Geophysical Fluid Dynamics Laboratory, Princeton, USA
[17]World Meteorological Organization, Geneva, Switzerland

Correspondence: Ruth Petrie (ruth.petrie@stfc.ac.uk)

Abstract. The distribution of data contributed to the Coupled Model Intercomparison Project Phase 6 (CMIP6) is via the Earth System Grid Federation (ESGF). The ESGF is a network of internationally distributed sites that together work as a federated data archive. Data records from climate modelling institutes are published on the ESGF and then shared around the world. It is anticipated that CMIP6 will produce approximately 20PB of data to be published and distributed via the ESGF. In addition to this large volume of data a number of value-added CMIP6 services are required to interact with the ESGF, for example the Citation and Errata services both interact with the ESGF but are not a core part of its infrastructure. With a number of interacting services and a large volume of data anticipated for CMIP6 a CMIP Data Node Operations Team (CDNOT) was formed. The CDNOT coordinated and implemented a series of CMIP6 preparation data challenges to test all the interacting components in the ESGF CMIP6 software ecosystem. This ensured that when CMIP6 data were released it could be reliably distributed.

1 Introduction

This paper describes the collaborative effort to publish and distribute the extensive archive of climate model output generated by the Coupled Model Intercomparison Project Phase 6 (CMIP6). CMIP6 data form a central component of the global scientific effort to update our understanding of the extent of anthropogenic climate change and the hazards associated with that change. Peer-reviewed publications based on analyses of these data are used to inform the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports. CMIP data and the underlying data distribution infrastructures are also being exploited by climate data service providers, such as the European Copernicus Climate Change Service (C3S; Thépaut et al., 2018). Eyring et al. (2016) describe the overall scientific objectives of CMIP6 and the organisational structures put in place by the World Climate Research Programme's (WCRP) Working Group on Coupled Modelling (WGCM). Several innovations were introduced in CMIP6, including the establishment of the WGCM Infrastructure Panel (WIP) to oversee the specification, deployment and operation of the technical infrastructure needed to support the CMIP6 archive. Balaji et al. (2018) describe the work of the WIP and the components of the infrastructure that it oversees.

CMIP6 is expected to produce substantially more data than CMIP5, it is estimated that approximately 20PB of CMIP6 data will be produced (Balaji et al., 2018), compared to 2PB of data for CMIP5 and 40TB for CMIP3. The large increases in volumes from CMIP3 to CMIP5 to CMIP6 are due to a number of factors: the increases in model resolution, an increase in the number of participating modelling centres, an increased number of experiments from 11 in CMIP3 to 97 in CMIP5 and now to 312 experiments in CMIP6 and the overall complexity of CMIP6 where an increased number of variables are required by each experiment (for full details see the discussion on the CMIP6 Data Request; Juckes et al., 2020). From an infrastructure view, the petabyte (PB) scale of CMIP5 necessitated a federated data archive system, which was achieved through the development of the global Earth System Grid Federation (ESGF; Williams et al., 2011, 2016; Cinquini et al., 2014). The ESGF is a global federation of sites providing data search and download services. In 2013 the WCRP Joint Scientific Committee recommended the ESGF infrastructure as the primary platform for its data archiving and dissemination across the programme, including for CMIP data distribution (Balaji et al., 2018). Timely distribution of CMIP6 data is key in ensuring that the most recent research is available for consideration for inclusion in the upcoming IPCC 6th Assessment Report (AR6). Ensuring the timely availability of CMIP6 data raised many challenges and involved a broad network of collaborating service providers. Given the large volume of data anticipated for CMIP6, a number of new related value-added services entering into production (see subsection 3.3) and the requirement for timely distribution of the data, in 2016 the WIP requested the establishment of a CMIP Data Node Operations Team (CDNOT) with representatives from all ESGF sites invited to attend. The objective of the CDNOT (WIP, 2016) was to ensure the implementation of a federation of ESGF nodes that would distribute CMIP6 data in a flexible way such that it could be responsive to evolving requirements CMIP process. The remit of the CDNOT described in (WIP, 2016) is the oversight of all the participating ESGF sites ensuring that:

  – all sites adhere to the recommended ESGF software stack security policies of ESGF [1];

---

[1]https://esgf.llnl.gov/esgf-media/pdf/ESGF-Software-Security-Plan-V1.0.pdf (last accessed: 11th Septemeber 2020)

- all required software must be installed at the minimum supported level, including system software which must be kept up-to-date with any recommended security patches applied;

45 - all sites have the necessary policies and workflows in place for managing data, such as acquisition, quality assurance, citation, versioning, publication and provisioning of data access [2];

- the required resources are available, including both hardware and networking [3];

- there is good communication between sites and the WIP.

The organisational structure of the CMIP data delivery management system
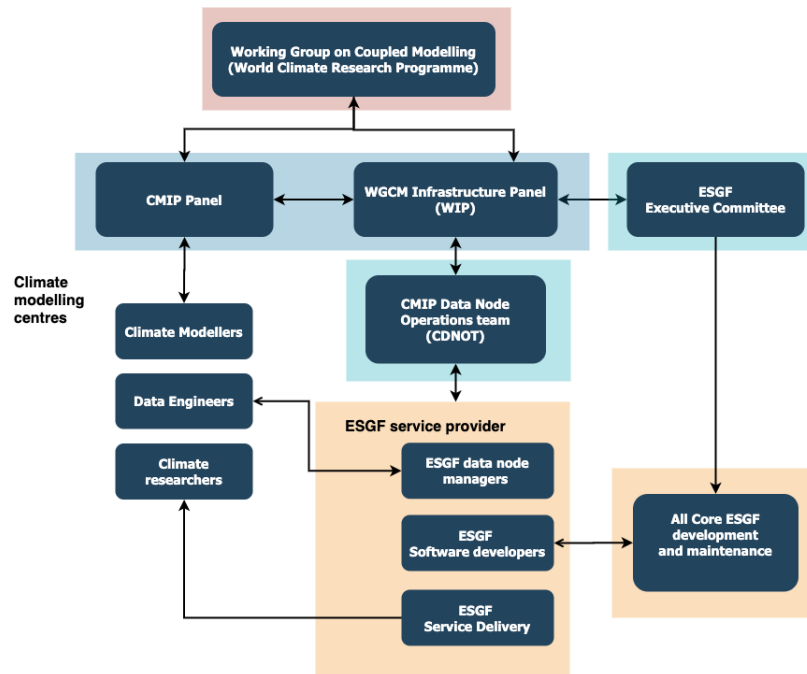
Figure 1. Organogram of the CMIP data delivery management structure.

50   In order to understand the overall climate model data delivery system and where the CDNOT sits in relation to all of the organisations previously mentioned a simplified organogram (only parties relevant to this paper are shown) is shown in Figure 1. At the top is the Working Group on Coupled Modelling (WGCM) mandated by the World Climate Research Programme (WCRP). The CMIP Panel and the WGCM Infrastructure panel (WIP) are commissioned by the WGCM. The CMIP panel oversees the scientific aims of CMIP including the design of the contributing experiments to ensure the relevant scientific

55   questions are answered. Climate modellers that typically sit within climate modelling centres or universities take direction

---

[2]https://esgf.llnl.gov/esgf-media/pdf/ESGF-Policies-and-Guidelines-V1.0.pdf (last accessed: 11th Septemeber 2020)

[3]https://esgf.llnl.gov/esgf-media/pdf/ESGF-Tier1and2-NodeSiteRequirements-V5.pdf (last accessed: 11th Septemeber 2020)

from the CMIP panel. The WIP oversees the infrastructure needed to support CMIP data delivery and includes the hardware and software components. The WIP is responsible for ensuring that all the components are working and available so that the data from the climate modelling centres can be distributed to the data users. The WIP and CMIP panel liaise to ensure that CMIP needs are met and that any technical limitations are clearly communicated. In 2016 the WIP commissioned the CDNOT to have oversight over all participating ESGF sites. Typically an ESGF site will have a data node manager that work closely with the data engineers at the modelling centers to ensure that CMIP data that is produced is of a suitable quality (see section 3.2.2) to be published on the ESGF federation. At each site ESGF software maintainers/developers work to ensure that the software is working at their local site and often work collaboratively with the wider ESGF community of software developers to develop the core ESGF software and how it interacts with the value-added services. The development of the ESGF software is overseen by the ESGF Executive Committee. Finally, the ESGF network provides the data access services that make the data available to climate researchers; the researchers may sit within the climate modelling centers, universities or industry. The aim of this paper is not to discuss in detail all of the infrastructure and different components supporting CMIP6, much of this is discussed elsewhere, such as Balaji et al. (2018); Williams et al. (2016), rather the focus of this paper is to describe the work of the CDNOT and the work done to prepare the ESGF infrastructure for CMIP6 data distribution.

The remainder of the paper is organized as follows. Section 2 briefly describes the ESGF as this is central to the data distribution and the work of the CDNOT. Section 3 discusses each of the main software components that are coordinated by the CDNOT, whereas section 4 describes a recommended hardware configuration for optimizing data transfer. Section 5 describes a series of CMIP6 preparedness challenges that the CDNOT coordinated between January and July 2018. Finally, section 6 provides a summary and conclusions.

2   Overview: The Earth System Grid Federation (ESGF)

The Earth System Grid Federation (ESGF) is a network of sites distributed around the globe that work together to act as a federated data archive (Williams et al., 2011, 2016; Cinquini et al., 2014), as shown in Figure 2. It is an international collaboration that has developed over approximately the last 20 years to manage the publication, search, discovery and download of large volumes of climate model data (Williams et al., 2011, 2016; Cinquini et al., 2014). As the ESGF is the designated infrastructure (Balaji et al., 2018) for the management of and distribution of the Coupled Model Intercomparison Project (CMIP) data and it is core to the work of the CDNOT, it is briefly introduced here.

ESGF sites operate data nodes from which their model data is distributed, some sites also operate index nodes that run the software for data search and discovery. Note, it is not uncommon for the terms sites and nodes to be used somewhat interchangeably, although they are not strictly identical. At each site there are both hardware and software components which need to be maintained. Sites are typically located at large national-level governmental facilities as they require robust computing infrastructure and experienced personnel to operate and maintain the services. The sites interoperate using a peer-to-peer (P2P) paradigm over fast research network links spanning countries and continents. Each site that runs an index (search) node

Figure 2. Geographic distribution of Earth System Grid Federation sites participating in CMIP6 as of April 2020. The figure is generated using a world map taken from Shutterstock and the country flags are taken from Iconfinder.

publishes metadata records that adhere to a standardised terminology[4] allowing data search and download to appear the same irrespective of where the search is initiated. The metadata records contain all the required information to locate and download data. Presently four download protocols are supported, HTTP (Hyper Text Transfer Protocol Transport Layer Security (TLS) is supported to ensure best security practice), GridFTP[5], Globus Connect[6], and OpenDAP[7]. The metadata records held at each site are held on a Solr shard, Solr[8] is an open-source enterprise-search platform, the shards synchronisation is managed by Solr, it is this synchronisation which allows the end user to see the same search results from any index node. There is a short latency due to the shard synchronisation but consistency is reached within a couple of minutes and there are no sophisticated consistency protocols that require a shorter latency. All data across the federation is tightly version controlled, this unified approach to the operation of the individual nodes is essential for the smooth operation of the ESGF federation.

The ESGF is the authoritative source of published CMIP6 data as data are kept up-to-date with data revisions and new version releases carefully managed. It is typical for large national-level data centres to replicate data to their own site so that their local users have access to the data without many users having to download the same data. While it is the responsibility of modelling centres to ensure that their data are backed-up (for example to a tape archive), these subsets of replica data are republished to the ESGF and provide a level of redundancy in the data, typically there will be two or three copies of the data. It is the responsibility of the publishing centres to ensure that the metadata catalogues are backed-up, however the federation itself can also act as back-up (though in practice this has never been necessary). Anyone can take a copy of a subset of data from ESGF, if these are not republished to the EGSF these mirrors (sometimes referred to as "dark archives" as they are not visible to all users) are not guaranteed to have the correct, up-to-date versions of the data.

---

[4]https://github.com/WCRP-CMIP/CMIP6_CVs (last accessed: 11th Septemeber 2020)

[5]https://en.wikipedia.org/wiki/GridFTP (last accessed: 11th Septemeber 2020)

[6]https://www.globus.org/globus-connect (last accessed: 11th Septemeber 2020)

[7]https://www.opendap.org/ (last accessed: 11th Septemeber 2020)

[8]https://lucene.apache.org/solr/ (last accessed: 11th Septemeber 2020)

In CMIP5 the total data volume was around 2PB; while large, it was manageable for government facilities to hold near full copies of CMIP5. For CMIP6 there are approximately 20PB of data expected to be produced, making it unlikely that many (if any) sites will hold all CMIP6 data, most sites will hold only a subset of CMIP6 data. The subsets will be determined by the individual centres and their local user community priorities.

Figure 3. The ESGF data search and delivery system schematic.

The ESGF user interface shown in Figure 3 allows end users to search for data that they wish to download from across the federation. Typically users narrow their search using the dataset facets as displayed on the left hand side of the ESGF web page screen shown in Figure 3. Each dataset that matches a given set of criteria is displayed. Users can further interrogate the metadata of the dataset or continue filtering until they have found the data that they wish to download. Download can either be via a single file, a dataset or a basket of different data to be downloaded. In each case the search (index) node communicates with the data server (an ESGF data node) that in turn communicates with the data archive where the data is actually stored, once located this data is then transferred to the user.

3 CDNOT: The CMIP6 software infrastructure

The CDNOT utilises and interacts with a number of software components that make up the CMIP6 ecosystem as shown in Figure 4. Some of these are core components of the ESGF and the ESGF Executive Committee has oversight of their development and evolution; such as the ESGF software stack and quality control software. Other software components in the ecosystem are integral to the delivery of CMIP6 data but are overseen by the WIP; such as the Data Request, the Earth System Documentation (ES-DOC) service and the Citation Service. The CDNOT had the remit of ensuring that all these components were able to interact where necessary and acted to facilitate communication between the different software development groups. Additionally, the CDNOT was responsible for the working implementation of these interacting components for CMIP6; i.e. ensuring the software was deployed and working and at all participating sites. The different software components are briefly described here, however each of these components is in its own right a substantive piece of work. Therefore, in this paper only

a very high-level overview of the software components are given, and the reader is referred to other relevant publications or documentation in software repositories for full details as appropriate.

Figure 4. Schematic showing the software ecosystem used for CMIP6 data delivery. Note the darker blue bubbles are those that represent the value-added services.

3.1 Generating the CMIP6 data

130 Before any data can be published to the ESGF, the modelling centres decide which CMIP experiments they wish to run. To determine what variables at a given frequency is required from the different model runs, the modelling centres use the CMIP6 data request as described in 3.1.1 and Juckes et al. (2020).

3.1.1 Pre-publication: The CMIP6 Data Protocol

The management of information and communication between different parts of the network is enabled through the CMIP6 135 data protocol which includes a Data Reference Syntax (DRS), Controlled Vocabularies (CVs) and the CMIP6 Data Request (see Taylor et al., 2018; Juckes et al., 2020). The DRS specifies the concepts which are used to identify items and collections of items. The CVs list the valid terms under each concept, ensuring that each experiment, model, MIP, etc has a well defined name which has been reviewed and can be used safely in software components.

Each MIP proposes a combinations of experiments and variables at relevant frequencies required to be written out from the model simulations to produce the data needed to address a specific scientific question. In many cases, MIPs request data not only from experiments that they are proposing themselves, but also from experiments proposed by other MIPs, so that there is substantial scientific and operational overlap between the different MIPs.

The combined proposals of all the MIPs, after a wide ranging review and consultation process, are combined into the CMIP6 Data Request (DREQ: Juckes et al., 2020). The DREQ includes an XML (extensible markup language) database and a tool available for modelling centres (and others) to determine which variables and at what frequencies should be written out by the models during the run. It relies heavily on the CVs and also on the CF Standard Name table[9] which was extended with hundreds of new terms to support CMIP6.

Each experiment is assigned one of three tiers of importance, and each requested variables is given one of three levels of priority (the priority may be different for different MIPs). Modelling centres may choose which MIPs they wish to support, and which tiers of experiments and priorities of variables they wish to contribute, so that they have some flexibility to match their contribution to available resources. If a modelling centre wishes to participate in a given MIP they must at a minimum run the highest priority experiments.

3.2 Publication to ESGF

The first step in enabling publication to the ESGF is that a site must install the ESGF software stack that comes with all the required software for the publication of data to the ESGF as described in 3.2.1. Data that comes directly from the different models is converted to a standardised format. It is this feature that facilitates the model intercomparison and is discussed in 3.2.2. The data are published to the ESGF with a strict version controlled numbering system for traceability and reproducability.

3.2.1 The ESGF software stack

The ESGF Software Stack is installed onto a local site via the "esgf-installer". This module was developed initially by Lawrence Livermore National Laboratories (LLNL) in the US with subsequent contributions made from several other institutions, including Institute Pierre Simon Laplace (IPSL) in France and Linköping University (LIU) in Sweden. Initially, the installer was a mix of Shell/Bash scripts to be run by a node administrator with some manual actions also required. It installed all the software required to deploy an ESGF node. During the CMIP6 preparations as other components of the system evolved, many new software dependencies were required. To capture each of these dependencies into the ESGF Software Stack, multiple versions of the "esgf-installer" code were released to cope with the rapidly evolving ecosystem of software components around CMIP6 and with the several operating systems used by the nodes. The Data Challenges, as described in Section 5, utilise new "esgf-installer" versions at each stage; the evolution of the installer is tracked on the related project in GitHub[10]. The ESGF software stack includes many components e.g. the publishing software, node management software, index search software and security software.

---

[9]http://cfconventions.org/standard-names.html (last accessed: 4th May 2020)
[10]https://github.com/ESGF/esgf-installer(last accessed: 4th May 2020)

Given the difficulties experienced installing the various software packages and managing their dependencies, the ESGF community has moved toward a new deployment approach based on Ansible, an open-source software application-deployment tool. This solution ensures that the installation performs more robustly and reliably than the bash-shell based one by using a set of repeatable (idempotent) instructions that perform the ESGF Node deployment. The new installation software module, known as "esgf-ansible", is a collection of Ansible installation files (playbooks) which are available in the related ESGF GitHub repository[11].

3.2.2 Quality control

Once all the underlying software required to publish data to ESGF is in place, the next step is to begin the publication process using the "esg-publisher" package that is installed with the ESGF software stack. Before running the publication, the data must be quality controlled. This is an essential part of this process and it is meant to ensure that both the data and the file metadata meet a set of required standards before the data is published. This is done through a tool called PrePARE developed at PCMDI/LLNL[12]. The PrePARE tool checks for conformance with the community agreed common baseline metadata standards for publication to the ESGF. Data providers can use the Climate Model Output Rewriter[13] (CMOR) tool to convert their model output data to this common format or in-house bespoke software, in either case the data must pass the PrePARE checks before publication to ESGF.

For each file to be published, PrePARE checks:

- the conformance short, long and standard name of the variable;

- the covered time period;

- several required attributes (e.g., CMIP6 activity, member label, etc.) against the CMIP6 controlled vocabularies;

- the filename syntax.

3.2.3 Dataset versioning

Each ESGF dataset is allocated a version number in the format `"vYYYYMMDD"`. This version number can be set before publication or allocated during the publication process. The version number forms a part of the dataset identifier, which is a "." separated list of CMIP6 controlled vocabulary terms that uniquely describe each dataset. This allows any dataset to be uniquely referenced. Versioning allows modelling centres to retract any data that may have errors and replace it with a new version by simply applying a new version number. This method of versioning allows all end users to know which dataset version was used in their analysis, making data versioning critical for reproducibility.

---

[11]https://github.com/ESGF/ (last accessed: 4th May 2020)
[12]https://github.com/ESGF/esgf-prepare (last accessed: 4th May 2020)
[13]https://cmor.llnl.gov/ (last accessed: 4th May 2020)

3.3 Value-added user services for CMIP6

The core EGSF service described in section 3.2 covers only the basic steps of making data available via the ESGF. The ESGF infrastructure is able to distribute other large programme data using the infrastructure described above (though project specific modifications are required). However, this basic infrastructure does not provide any further information on the data such as documentation, errata, or citation. Therefore, in order to meet the needs of the CMIP community, a suite of value-added services have been specifically included for CMIP6. These individual value-added services were prepared by different groups with the ESGF community and the CDNOT coordinated the implementation of these. The data challenges as described in section 5 were used to test these service interactions and identify issues where the services were not integrating properly or efficiently. The value-added services designed and implemented for CMIP6 are described in this section.

3.3.1 Citation Service

It is commonly accepted within many scientific disciplines that the data underlying a study should be cited in a similar way to literature citations. The data author guidelines from many scientific publishers prescribe that scientific data should be cited in a similar way to the citation described in Stall (2018). For CMIP6, it is required that modelling centres register data citation information, such as a title and list of authors with their ORCIDs (a unique identifier for authors) and related publication references with the CMIP6 Citation Service Stockhause and Lautenschlager (2017). Since data production within CMIP6 is recurrent, a continuous and flexible approach to making CMIP6 data citation available alongside the data production was implemented. The Citation Service[14] issues DataCite[15] Digital Object Identifiers (DOI) and uses DataCite vocabulary and metadata schema as international standards. The service is integrated into the project infrastructure and exchanges information via international data citation infrastructure hubs. Thus, CMIP6 data citation information is not only visible from the ESGF web front end and on the "further_info_url" page of ES-DOC (see 3.3.3) but it is also visible in the Google Dataset Search and external metadata catalogues. The Scholarly Link Exchange[16] (Scholix) allows information on data usage in literature to be accessed and integrated.

The data citations are provided at different levels of granularities of data aggregations to meet different citation requirements within the literature. For the first time within CMIP it will be possible for scientists writing peer-reviewed literature are able to cite CMIP6 data prior to long-term archival in the IPCC Data Distribution Centre (DDC)[17] Reference Data Archive. Details of the CMIP6 data citation concepts are described in the WIP White Paper (Stockhause and Lautenschlager, 2015).

3.3.2 PID Handle Service

Every CMIP6 file is published with a Persistent IDentifier (PID) in the "tracking_id" element of the file metadata. The PID is of the form "hdl:21.14100/<UUID>" where the UUID is the unique id generated by CMOR using the general open source

---

[14]http://cmip6cite.wdc-climate.de (last accessed: 4th May

[15]https://datacite.org/ (last accessed: 4th May 2020)

[16]http://www.scholix.org/ (last accessed: 4th May 2020)

[17]http://ipcc-data.org; (last accessed: 12th May 2020)

UUID library. A PID allows for persistent references to data including versioning information as well replica information. This is a significant improvement over the simple "tracking_id" that was used in CMIP5 which did not have these additional functionalities that are managed through a PID Handle Service[18]. The PID Handle service aims to establish a hierarchically organized collection of PIDs for CMIP6 data. In a similar way to the Citation Service, PIDs can be allocated at different levels

230   of granularity. At high levels of granularity a PID can be generated that will refer to a large collection of files (that may still be evolving), such as all files from a single model, or from a model simulation (a given MIP, model, experiment and ensemble member). This type of PID would be useful for a modelling group to refer to a large collection of files. Smart tools will be able to use the PIDs to do sophisticated querying and allows for the integration with other services to automate processes.

3.3.3 Earth System Documentation

235   The Earth System Documentation (ES-DOC; **?**) is a service that documents in detail all the models and experiments and the CMIP6 endorsed MIPs. Modelling centres are responsible for providing the ES-DOC team with the detailed information of the model configurations such as the resolutions, physics and parameterizations. With the aim of aiding scientific analysis of a model intercomparison study, ES-DOC allows scientists to more easily compare what processes may have been included in a given model explicitly or in a parameterised way. For CMIP6 ES-DOC also provides detailed information on the computing

240   platforms used.

   The ES-DOC service also provides the "further_info_url" metadata, which is an attribute embedded within every CMIP6 NetCDF file. It links to the ESGF search pages that ultimately resolves to the ES-DOC service, providing end-users with a web page of useful additional metadata. For example, it includes information on the model, experiment, simulation and MIP and provides links back to the source information making it far easier for users to find further and more detailed information than

245   was possible during CMIP5. The simulation descriptions are derived automatically from metadata embedded in every CMIP6 NetCDF file, ensuring that all CMIP6 simulations are documented without the need for effort from the modelling centres.

3.3.4 Errata Service

Due to the the experimental protocol and the inherent complexity of projects like CMIP6, it is important to record and track the reasons for dataset version changes. Proper handling of dataset errata information in the ESGF publication workflow has

250   major implications on the quality of the metadata provided. Reasons for version changes should be documented and justified by explaining what was updated, retracted and/or removed. The publication of a new version of a dataset, as well as the retraction of a dataset version, has to be motivated and recorded. The key requirements of the Errata Service [19] are to:

   – Provide timely information about newly discovered dataset issues. As errors cannot always entirely be eliminated, the Errata Service provides a centralized public interface to data users where data providers directly describe problems as

255   and when they are discovered;

---

[18]https://www.earthsystemcog.org/site_media/projects/wip/CMIP6_PID_Implementation_Plan.pdf (last accessed: 4th May 2020)

[19]https://errata.es-doc.org/ (last accessed: 4th May 2020)

- Provide information on known issues to users through a simple web interface.

The Errata Service now offers a user-friendly front-end and a dedicated API. ESGF users can query about modifications and/or corrections applied to the data in different ways:

- Through the centralized and filtered list of ESGF known issues;
- Through the PID lookup interface to get the version history of a (set of) file/dataset(s)

3.3.5 Synda

Synchronise Data[20] (Synda) is a data discovery, download and replication management tool built and developed for and by the ESGF community. It is primarily aimed at managing the data synchronisation between large data centres, though, it could also be used by any CMIP user as a tool for data discovery and download.

Given a set of search criteria, the Synda tool searches the metadata catalogues at nodes across the ESGF network and return results to the user. The search process is highly modular and configurable to reflect the end-users requirements. Synda can also manage the data download process. Using a database to record data transfer information, Synda can optimize subsequent data downloads. For example, if a user has retrieved all data for a given variable, experiment, model combination and then a new search is performed to find all the data for the same variable and experiment but across all models, Synda will not attempt to retrieve data from the model it has already completed data transfers for. Currently Synda supports the HTTP and GridFTP transfer protocols and integration with Globus Connect is expected to be available during 2020. Data integrity checking is performed by Synda. The ESGF metadata includes a checksum for each file, after data has been transferred by Synda the local file checksum is computed and then compared with the checksum of the master data to ensure data integrity. Any user can check at any time the checksums of the data as this key piece of metadata is readily available from the ESGF published catalogue. Synda is a tool under continual development with performance and feature enhancements expected over the next few years.

3.3.6 User perspectives

While many users of CMIP6 data will obtain data from their national data archive and will not have to interact with the ESGF web-based download service, many users will have to interact with the ESGF front-end website (see Figure 3 for web-based data download schematic). Great care has been taken to ensure that all nodes present a CMIP6 front end that is nearly identical in appearance to present users with a consistent view across the federation.

3.3.7 Dashboard

The ESGF also provides support for integrating, visualizing and reporting data downloads and data publication metrics across the whole federation. It aims to provide a better understanding on the amount of data and number of files downloaded; i.e. the

---

[20]http://prodiguer.github.io/synda/ (last accessed: 4th May 2020)

285 most downloaded datasets, variables and models, as well as on the number of published datasets (overall and by project). During the data challenges as described in section 5, it became apparent that the dashboard implementation in use at that time was not feasible for operational deployment. Since the close of the data challenges, work has continued on the dashboard development. A new implementation of the dashboard[21] is now available and has been then successfully deployed in production[22]. Figure 5 shows how the Dashboard statistics are collected.
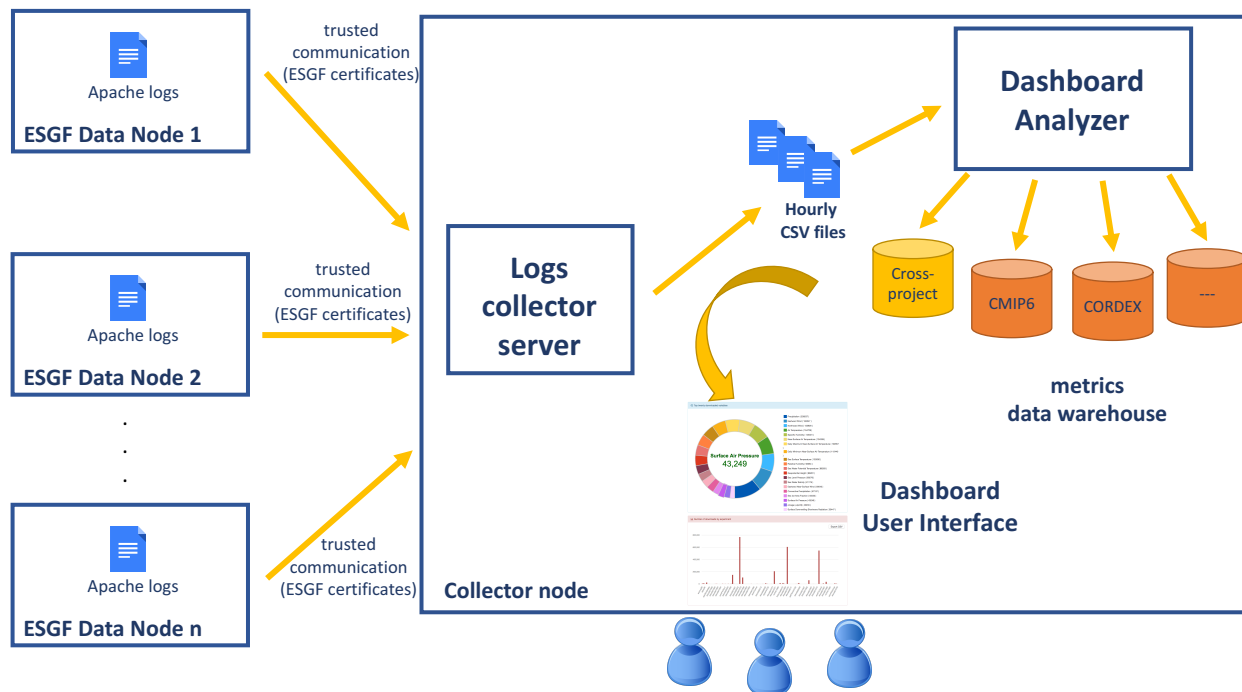
Figure 5. The ESGF Dashboard architecture

290   The deployed metrics collection is based on industry-standard tools made by Elastic[23], namely Filebeat and Logstash. Filebeat is used for transferring log entries of data downloads from each ESGF data node. It sends log entries (after filtering out any sensitive information) to a central collector via a secured connection. Logstash collects the log entries from the different nodes, processes and stores them in a database. Then, the analyzer component extracts all the relevant information from the logs to produce the final aggregated statistics. The provided statistics fulfil the requirements gathered from the community on
295 CMIP6 data download and data publication metrics.
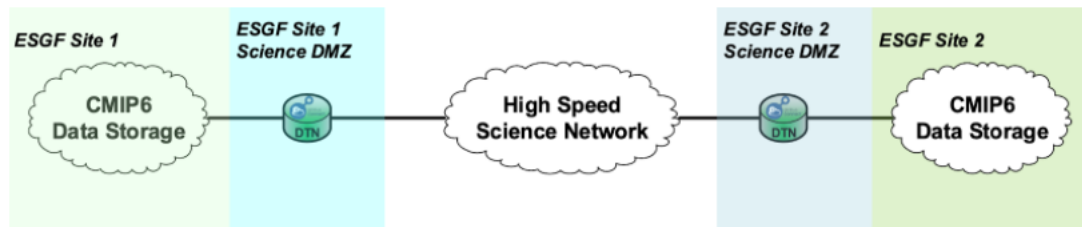
Figure 6. Data transfer architecture

4 CDNOT: Infrastructure

In order to efficiently transfer data between ESGF sites, the CDNOT recommended a basic hardware infrastructure as described in (Dart et al., 2013) and shown in a simplified format in Figure 6. Testing of this was included in the CMIP6 data challenges 5. In the schematic of Figure 6 it is demonstrated how two data centres should share data. Each High Performance Computing (HPC) environment has a local archive of CMIP6 data and ESGF servers that hold the necessary software to publish and replicate the CMIP6 data. It was additionally recommended that sites utilise science data transfer nodes (DTN) that sit in a science Demilitarized Zone (DMZ) also referred to as a Data Transfer Zone (DTZ), designed according to the Energy Sciences Network (ESnet) Science DMZ (Dart et al., 2013). A DTZ has security optimized for high-performance data transfer rather than for the more generic HPC environment, given the necessary security settings it has less functionality than the standard HPC environment. To provide high performance data transfer, it was recommended that each site put a GridFTP server in the DTZ. Such configuration enhances the speed of the data transfer in two ways:

1. the data transfer protocol is GridFTP, which is typically faster than HTTP or a traditional FTP of a file;

2. the server sits in a part of the system that has a faster connection to high-speed science networks and data transfer zones.

Having this hardware configuration was recommended for all the ESGF data nodes, but in particular for the larger and more established sites that host a large volume of replicated data. During the test phase, it was found that >300Mb/s was achievable with well-configured hardware; this implies approximately 25TB/day. It is possible with new protocols and or additional optimisations and tuning that this can still be improved upon. However, during the operational phase of CMIP6 thus far it has proven difficult to sustain this rate due to: (i) only a small number of sites using the recommended infrastructure, (ii) heterogeneous (site-specific) network and security settings and deployment and (iii) factors outside of the system-administration team control such as the background level of internet traffic and the routing from the end users' local computing environment. Sites that are unable to fulfill the recommendations completely, for any reason, such as lack of technical infrastructure, expertise or resource are likely to have a degraded performance in comparison to other sites. The degree to which it is impacted is reflect in how far the deployed infrastructure deviates from the recommendations outlined in this section. In this respect, it is important

---

[21]https://github.com/ESGF/esgf-dashboard (last accessed: 11th Septemeber 2020)

[22]http://esgf-ui.cmcc.it/esgf-dashboard-ui/ (last accessed: 4th May 2020)

[23]https://www.elastic.co/ (last accessed: 4th May 2020)

to remark that the CDNOT has recommended this deployment as best practice, but it can not be enforced; however the CDNOT and its members continue to try and assist all sites in the infrastructure deployments.

5 CMIP6 Data challenges

Tasks 1-16 vs Challenges 1-5	1	2	3	4	5
1. Install (or update) the ESGF software stack	✔	✔	✔	✔	✔
2. Run quality control on primary data	✔	✔	✔	✔	✔
3. Publish primary data	✔	✔	✔	✔	✔
4. Publish replica data	✔	✔	✔	✔	✔
5. Verify search and download is functional	✔	✔	✔	✔	✔
6. Register data with PID assignment service		✔	✔	✔	✔
7. Verify Citation Service registers DOIs for published data		✔	✔	✔	✔
8. Populate "further_info_url" through ES-DOC scanning		✔	✔	✔	✔
9. Replicate published data			✔	✔	✔
10. Apply the "test suite"			✔	✔	✔
11. Verify the metrics collection for the dashboard			✔	✔	✔
12. Register an errata with the Errata Service			✔	✔	✔
13. Retract a version of the data				✔	✔
14. Publish a new version of the data				✔	✔
15. Ensure homogeneity across ESGF CoG sites				✔	✔
16. Move testing to production environment					✔

Table 1. Table of the data challenge tasks and the challenge at which they were implemented. Note the grey tick of Task 11 in Challenge 5 indicates that the task was optional.

In preparation for the operational CMIP6 data publication, a series of five data challenges were undertaken by the members of the CDNOT between January and June 2018, with approximately one month for each data challenge. The aim was to ensure that all the ESGF sites participating in the distribution of CMIP6 data would be ready to publish and replicate CMIP6 data once released by modelling centres. These five data challenge phases became increasingly more complex in terms of the software ecosystem tested and ever increasing data volumes and number of participating source models.

Table 1 shows the tasks that were performed in each data challenge. It is important to note that not every step taken during these data challenges has been reported here; only the most important high-level tasks are listed; the summary of these high-level tasks have been described in such a way as to not be concerned with the many different software packages involved and the frequent release cycles of the software that occurred during the data challenges. The participating sites were in constant contact

throughout the data challenges often developing and refining software during a particular phase to ensure that a particular task was completed.

It is important to note that the node deployment software (the "esgf-installer" software) was iterated at every phase of the data challenge. As issues were identified at each stage and fixes and improvements to services were made available, the installer had to be updated to include the latest version of each of these components. Similarly, many other components, such as PrePARE, and the integration with value-added services, were under continual development during this time. All the individual details have been omitted in this description, as these technical issues are beyond the scope of this paper.

5.1 Data challenge 1: Publication testing

The aim of the first phase was to verify that each participating site could complete tasks 1-5:

1. Install (or update) the ESGF software stack;

2. Run quality control on primary data;

3. Publish primary data;

4. Publish replica data;

5. Verify search and download was functional.

These tasks constitute the basic functionality of ESGF data publication and it is therefore essential that these steps are able to be performed by all sites using the most recent ESGF software stack.

In order for each EGSF site participating in this data challenge to test the publication procedure, a small amount ($\sim$10GB) of pseudo-CMIP6 data provided by modeling centers was circulated "offline", i.e. not using the traditional ESGF data replication methodology. These data were prepared specifically as test data for this phase of the data challenge. The data contained pseudo-data prepared to look as if it had come from a few different modelling centres and was based on preliminary CMIP6 data.

In normal operations ESGF nodes publish data from their national modelling centre(s) as "primary data", these are also referred to as "master copies", some larger ESGF sites act as primary nodes for smaller modelling centres outside their national boundaries. Once data records are published, other nodes around the federation are then able to discover this data and replicate it to their local site. This is typically done using the Synda replication tool (see subsection 3.3.5), the replicating site then republishes the data as a "replica" copy. In this data challenge, using only pseudo-data, sites published data as a primary copy if the data came from their national centre, otherwise it was published as a replica copy.

All participating sites were successful in the main aims of the challenge. However, it was noted that some data did not pass the quality control step due to inconsistencies between the CMOR controlled vocabularies and PrePARE. PrePARE was updated to resolve this issue before the next data challenge.

5.2 Data challenge 2: Publication testing with an integrated system

The aims of the second phase was for each participating site to complete steps 1-5 from phase 1 and additionally complete the following steps:

7. Registration of data with the PID assignment service;

8. Verification of citation service DOIs for published data;

9. Populate "further_info_url" through ES-DOC scanning.

This second phase of the data challenge was expanded to include additional ESGF sites, additionally a larger subset (20GB) of test data were available. Some of the test data were the same as in the first data challenge, though some of the new test data were early versions of real CMIP6 data. The increase in the volume and variety of data at each stage in the data challenges is essential to continue to test the software to its fullest extent. The additional steps in this phase tested the connections to the Citation Service, the PID Handle Service and ES-DOC, which was used to populate the "further_info_url". A valuable piece of metadata linking these value-added services.

In this challenge, the core aims were met and the ESGF sites were able to communicate correctly with the value-added services. Although data download services were functional, it was noted that some search facets had not been populated in the expected way, and this was noted to be fixed during the next phase.

5.3 Data challenge 3: Publication with data replication

The aims of the third phase was for each participating site to complete steps 1-8 from phases 1 and 2 and additionally complete the following steps:

10. Replication of published data;

11. Apply the "test suite" (described below);

12. Verify the metrics collection for the dashboard;

13. Register data errata with the Errata Service.

The third data challenge introduced a step-change in the complexity of the challenge with the introduction of step 9: "Replication of published data". The replication of published data was done using Synda (see subsection 3.3.5) mimicking the way in which operational CMIP6 data is replicated between ESGF sites. In the first two challenges, test data were published at individual sites in a stand-alone manner, where data were shared between sites outside the ESGF network, thus not fully emulating the actual CMIP6 data replication workflow.

This third phase of the data challenges was the first to test the full data replication workflow. The added complexity in this phase also meant that the phase took longer to complete; all sites had to first complete the publication of their primary data before they could search for and replicate the data to their local sites and then republish the replica copies. This was further

complicated as the volume of test data (typically very early releases of CMIP6 data) was much larger than the previous phases with approximately 400GB of test data. Sites also tested whether they could register pseudo errata with the Errata Service. As errata should be operationally registered by modelling centres, this step required contact with the modelling centres' registered users.

An additional test was included in this phase which was to run the *test suite*[24]. The *test suite* is used to check that a site is ready to be made operational. It was required that sites run the ESGF *test suite* on their ESGF node deployment. The *test suite* automated checking that all the ESGF services running on the node were properly operational. The *test suite* was intended to cover all services. However, the GridFTP service was often a challenge to deploy and configure within the test infrastructure and therefore as an exception it was not included in or tested through the *test suite*.

At the time that this phase of the data challenge was being run, a different implementation of the ESGF dashboard was in use to the one described in subsection 3.3.7. Sites were required to collect several data download metrics for reporting activities and most sites were unable to integrate the dashboard correctly. Feedback collected during this phase contributed to the development of the new dashboard architecture now in use. Apart from the dashboard issues and despite the additional steps and complexity introduced in this phase, all sites were able to publish data, search for replicas, download data and publish the replica records.

5.4 Data challenge 4: Full system publication, replication with republication and new version release

The aims of the fourth phase was for each participating site to complete steps 1-12 from phases 1-3, the additional steps to complete were:

14. Retraction a version of a dataset;

15. Publication of a new version of a dataset.

This data challenge was performed with approximately 1.5TB of test data. This increase in volume increased the complexity and the time taken for sites to publish their primary data and the time taken to complete the data replication step. This larger volume and variety of data (coming from a variety of different modelling centres) was essential to continue to test the publication and quality control software with as wide a variety of data as possible. The inclusion of the two new steps, the retraction of a dataset and the publication of a new version, completes the entire CMIP6 publication workflow. When data are found to have errors, the data provider logs the error with the Errata Service and the primary and replica copies of the data are retracted. After a fix is applied and new data are released they are published with a new version number by the primary data node, then new replica copies are taken by other sites who then republish the updated replica versions.

No issues were found in the retraction and publication of new versions of the test data.

Not all sites participated in the deployment of the dashboard. Those that did participate provided feedback on its performance, which identified that a new logging component on the ESGF data node would be required for the dashboard to be able

---

[24]https://github.com/ESGF/esgf-test-suite (last accessed: 11th Septemeber 2020)

to scale with the CMIP6 data downloads. Given the difficulties in sites deploying the dashboard it was recommended that it be optional until the new service was deployed.

5.5 Data challenge 5: Full system publication, replication with republication and new version release on the real infrastructure

425 The aims of the fifth phase was for each participating site to complete steps 1-14 (step 11, deployment of the dashboard was not required) from phases 1-4 and additionally complete the following steps:

16. Ensure homogeneity across the ESGF web user interface;

17. Move testing on to production environment (i.e. operational software, hardware and federation).

The search service for CMIP6 is provided through the ESGF web user interface which cannot be fully configured program-
430 matically. For this reason, on sites that host search services the site administrators had to perform a small number of manual steps to produce an ESGF landing page that was near identical at all sites (with the exception of the site-specific information), this ensured consistency across the federation. While this is a relatively trivial step, it was important to make sure that the end users would be presented with a consistent view of CMIP6 data search across the federation.

Finally, but most importantly, all the steps of the data challenge were tested on the production hardware. In all of the previous
435 four phases the data challenges were performed using a test ESGF federation, which meant that a separate set of servers at each site were used to run all the software installation, quality control, data publication and data transfer. This was done to identify and fix all the issues without interrupting the operational activity of the ESGF network. The ESGF continued to supply CMIP5 and other large programme data throughout the entire testing phase. In this final phase, the CMIP6 project was temporarily hidden from search and all testing steps were performed within the operational environment. Once this phase was completed
440 in July 2018, the CMIP6 project was made available for the actual publication of real CMIP6 data to the operational nodes with confidence that all the components were functioning well. While at the end of this phase a number of issues remained to be resolved such as the development of a new Dashboard and improvements to data transfer performance rates, none of them were critical to the publication of CMIP6 data on the operational nodes.

5.6 Supporting data nodes through the provision of comprehensive documentation

445 Not all ESGF nodes involved with the distribution of CMIP6 data took part in the CMIP6 preparation work, this was mainly done by the larger and more well-resourced sites. In order to assist sites that were not able to participate to effectively publish CMIP6 through following the mandatory procedures, a comprehensive set of documentation of the different components was provided. This fulfilled one of the criteria of the CDNOT; documenting and sharing good practices. The documentation "CMIP6 Participation for Data Managers"[25] was published in February 2019 and is hosted by PCMDI as part of their "Guide to CMIP6
450 participation".

---

[25]https://pcmdi.llnl.gov/CMIP6/Guide/dataManagers.html (last accessed: 5th May 2020)

The Guide to CMIP6 participation is organized into three parts; (1) for CMIP6 participating climate modelers, (2) for CMIP6 ESGF data node managers and operators and (3) for users of CMIP6 model output. The CDNOT coordinated and contributed the documentation for the second component. This documentation had a number of contributing authors and was reviewed internally by the CDNOT team before being made public. This reference material for all CMIP6 data node managers provides guidelines describing how the CMIP6 data management plans should be implemented. It details for example how to install the required software, the metadata requirements, how to publish data, how to register errata, how to retract data and how to publish corrected data with new version numbers. The collation of the documentation was an important part of the CMIP6 preparation. This information is publicly available and provides valuable reference material to prepare CMIP6 data publication.

6 Summary and conclusions

In this paper the preparation of the infrastructure for the dissemination of CMIP6 data through the Earth System Grid Federation (ESGF) has been described. The ESGF is an international collaboration in which a number of sites publish CMIP data. The metadata records are replicated around the world and a common search interface allows users to search for all CMIP6 data irrespective of where their search is initiated.

During CMIP5 only an ESGF node installation and publication software were essential to publishing data. The ESGF software is far more complex for CMIP6 than for CMIP5 with the ESGF software required to interact with a software ecosystem of value-added services required for the delivery of CMIP6 such as the errata service, the citation service, Synda and the dashboard all being new in CMIP6. Existing value-added services including the data request, installation software, quality control tools and the Earth System documentation (ES-DOC) services have all increased in functionality and complexity from CMIP5 to CMIP6 with many of these components being interdependent.

The CMIP6 data request was significantly more complex than in CMIP5 and required new software to assist modelling centres to determine which data they should produce for each experiment. After modelling centres had performed their experiments the data were converted to a common format and the data were required to meet a minimum set of metadata standards, this check is done using PrePARE. Due to the complexities in CMIP6, the PrePARE tool has been under continual review to ensure all new and evolving data standards are incorporated. The installation software provided by the ESGF development team allows sites around the world to set up data nodes and make available their CMIP6 data contributions. The Synda data replication tool is used by a few larger sites to take replica copies of the data, providing redundancy across the federation. The Errata Service allows modelling centres to document issues with CMIP6 data providing traceability and therefore greater confidence in the data. The Citation Service allows scientists writing up analyses using CMIP6 data to fully cite the data used in their analysis. The Earth System Documentation Service allows users to discover detailed information on the models, MIPs and experiments used in CMIP6. The dashboard service allows for the collation of statistics of data publication and data download.

While these many new features enhance the breadth and depth of information available for CMIP6 data, it also increases the complexity of the ESGF software ecosystem and many of these services need to interact with one or more other parts of the system. Many of these services have been developed to sit alongside the main ESGF rather than be fully integrated into the

ESGF. Full integration of services such as the Errata Service and the Citation Service could further enhance the user experience and is under consideration for future developments.

The ultimate goal of the infrastructure enabled by the ESGF, and the preparation work performed by the CDNOT is to ensure a smooth data distribution to end users who are of interest to the CMIP6 data, for example, researchers comparing multiple climate models. However, ESGF and the CDNOT can only guarantee the reliability of the service provided by CMIP6 data nodes. The generic internet traffic and the routing from end users' local computing environment to CMIP6 data nodes are too diverse to be guaranteed. The Quality of Service (QoS) of the network of the nation where CMIP6 data node resides, physical distance between the CMIP6 data node and the testing site are all factors that need to be considered when distributing the CMIP6 to end users.

The distinctive nature of the ESGF can be seen by comparing it with other international data exchange projects. The Sea-DataNet[26] project is a European Research Infrastructure, though it also serves a global community. The use of standards data which are specified in a machine-interpretable form gives them an efficient framework for enforcing standards using generic tools. This is not possible in the current ESGF system, however the data volumes handled by SeaDataNet are substantially smaller, and the degree of human engagement with each published dataset is consequently larger. The problems of dealing efficiently with large volumes of data while maintaining interoperable services across multiple institutions is addressed by a more recently funded project the US NSF research project OpenStorageNetwork[27]. This project did not start until CMIP6 planning was well underway. As ESGF looks to develop and evolve the federation are open to exploring new ideas that come out of this and other projects where they have potential to improve the workflows involved with the management of ever-growing data volumes.

The development of an operational data distribution system for CMIP6 incorporating all the new services described in this paper represents a substantial amount of work and international collaboration. The result of this work is the provision of global open data access to CMIP6 data in an operational framework. This is of great importance to climate scientists around the world to perform analyses for the IPCC Assessment Report 6 Working Group 1 report and would not have been possible without the infrastructure described here. It is important to note that the work described was not centrally funded and relied heavily on project specific funding, which raises some concerns as to the operational sustainability of such an effort for a large ecosystem of software required to perform at an operational service level. This WCRP-led research effort has developed an impressive ecosystem of software which could be further integrated to ensure sustained support of CMIP to policy and services. This can only be achieved through strong international coordination, collaboration and sharing of resources and responsibilities.

Code availability. **ESGF** https://github.com/ESGF/; last accessed: 4th May 2020,

Data Request https://www.earthsystemcog.org/projects/wip/CMIP6DataRequest; last accessed 7th May 2020,

Citation Service http://cmip6cite.wdc-climate.de; last accessed: 4th May 2020,

---

[26]https://www.seadatanet.org (last accessed: 25th Novemeber 2020)
[27]https://www.openstoragenetwork.org (last accessed: 25th Novemeber 2020)

515 **Errata Service** https://github.com/ES-DOC/esdoc-errata-ws, https://github.com/ES-DOC/esdoc-errata-client; last accessed 7th May 2020,

Synda http://prodiguer.github.io/synda/; last accessed: 4th May 2020,

ESGF front-end web page https://github.com/EarthSystemCoG/COG ; last accessed: 4th May 2020,

Dashboard https://github.com/ESGF/esgf-dashboard ; last accessed: 4th May 2020,

Author contributions. RP led the preparation of this manuscript, participated in the organisation of the CDNOT data challenges. SD at the
520 time of the data challenges was the CDNOT chair and led the data challenges and was responsible for the smooth delivery of CMIP6 data.
All of the authors participated in and contributed to the CMIP6 data distribution preparedness data challenges and/or to the preparation of
this manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank the two anonymous reviewers for their helpful comments and the members of the
525 WGCM Infrastructure Panel for their support and assistance during the CMIP6 data challenges work.

Ruth Petrie acknowledges additional support from the European Union Horizon 2020 IS-ENES3 project (grant agreement no. 824084)
and would like to thank Bryan Lawrence for his comments on the manuscript.

References

Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., Blanton, C., Cinquini, L., Denvil, S., Elkington, M., Guglielmo, F., Guilyardi, E., Hassell, D., Kharin, S., Kindermann, S., Nikonov, S., Radhakrishnan, A., Stockhause, M., Weigel, T., and Williams, D.: Requirements for a global data infrastructure in support of CMIP6, Geosci. Model Dev., 11, 3659–3680, https://doi.org/10.5194/gmd-11-3659-2018, 2018.

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., Ananthakrishnan, R., Miller, N., Denvil, S., Morgan, M., Pobre, Z., Bell, G. M., Doutriaux, C., Drach, R., Williams, D., Kershaw, P., Pascoe, S., Gonzalez, E., Fiore, S., and Schweitzer, R.: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data, Future Generation Computer Systems, 36, 400–417, https://doi.org/10.1016/j.future.2013.07.002",, 2014.

Dart, E., Rotman, L., B., T., Hester, M., , and Zurawski, J.: The Science DMZ: A network design pattern for data-intensive science., SC13: Proceedings of the International Conference on High Performance Computing, Networking, pp. 1–10, https://doi.org/10.1145/2503210.2503245, 2013.

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

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.

Stall, S., e. a.: Advancing FAIR data in Earth, space, and environmental science, vol. 99, https://doi.org/10.1029/2018EO109301, 2018.

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

Stockhause, M., T. F. and Lautenschlager, M.: CMIP6 Data Citation and Long-Term Archival, https://zenodo.org/record/35178, 2015.

Taylor, K., Juckes, M., Balaji, V., Cinquini, L., Denvil, S., Durack, P., Elkington, M., G. E. K. S. L. M., Lawrence, B., Nadeau, D., and Stockhause, M.: CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's, https://www.earthsystemcog.org/site_media/projects/wip/CMIP6_global_attributes_filenames_CVs_v6.2.7.pdf, 2018.

Thépaut, J., Dee, D., Engelen, R., and Pinty, B.: The Copernicus Programme and its Climate Change Service, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 1591–1593, https://ieeexplore.ieee.org/document/8518067, 2018.

Williams, D. N., E., T. K., L., C., Evans, B., Kawamiya, M., M., L., N., L. B., D., M., and Contributors, E.: The Earth System Grid Federation: Software Framework Supporting CMIP5 Data Analysis and Dissemination, 2011.

Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D., Evans, B., Ferraro, R., Hansen, R., Lautenschlager, M., and Trenham, C.: A Global Repository for Planet-Sized Experiments and Observations, Bulletin of the American Meteorological Society, 97, 803–816, https://doi.org/10.1175/BAMS-D-15-00132.1, 2016.

WIP, W. I. P.: CDNOT Terms of Reference, http://cedadocs.ceda.ac.uk/id/eprint/1470, 2016.