IS-ENES3 Deliverable D7.3

Second KPI and TA report for ENES CDI data services

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

Authors: Stephan Kindermann (DKRZ), Martina Stockhause (DKRZ)
Alessandra Nuzzo (CMCC)
Martin Juckes, David Hassel (UKRI)
Guillaume Levavasseur (CNRS-IPSL)
Alessandro Spinuso (KNMI)

Reviewer(s): Sylvie Joussaume (CNRS-IPSL), Paola Nassisi (CMCC)
Release date: 10/02/2022

ABSTRACT

The ENES Climate data infrastructure provides services for data search, data access and FAIR data management (including support for persistent identifiers and data citation). Associated processing services are established at larger sites and are provided via the Virtual Access and Transnational Access mechanism. Data standards services are provided for the CMIP data request (specifying the variables and controlled vocabularies characterizing the data collections) and the Climate and Forecast Convention (CF). Also dedicated support services are provided with respect to CMIP6 documentation.

In this deliverable we summarize characteristic usage information for each service category and provide associated statistics. The service provisioning is based on eleven installations distributed across Europe. The installation specific details are provided as part of the associated IS-ENES3 Reporting Period 2 (RP2) report. The evolution of the services is coordinated in close cooperation with WP5/NA4 and WP10/JRA3. Sustainability aspects of the services are agreed and discussed further as part of the sustainability work in WP2/NA1.

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

Revision table			
Version	**Date**	**Name**	**Comments**
V.0.1	01/11/2021	Stephan Kindermann	Initial version: structure and contribution collection
V 0.2	05/12/2021	Stephan Kindermann, Alessandra Nuzzo, Alessandro Spinuso, Nikulin Grigory, Guillaume Levasseur, Martin Juckes, Martina Stockhause, David Hassel	Updates for individual KPI/PI sections
Release for review	6.12.2021	Stephan Kindermann	Pre-final version, not including the final december statistics
Final	10.02.2022	Stephan Kindermann, Alessandra Nuzzo, Andrey Dara	Final numbers updated, resolved reviewer comments

Dissemination level		
PU	Public	X

Table of contents

Executive Summary

For each of the services provided by the ENES Climate data infrastructure a set of performance indicators is provided. The performance indicators as well as key performance indicators were defined as part of the first report (Deliverable D7.1[1]) and are listed in section 1.1. The service provisioning is based on eleven installations distributed across Europe. More installation specific service details will be provided as part of the IS-ENES3 Reporting Period 2 (RP2) report.

The data delivery related services (see section 2) show a continued high demand in CMIP6 data access (especially from non-European users), whereas European users can access the large CMIP6 replica data pools hosted at DKRZ, CNRS-IPSL and UKRI directly without the need to rely on the ENES CDI ESGF data delivery services. The service related to the establishment of these data pools is described in section 2.2. Statistics also show a growing need for CORDEX data delivery and access via the ENES CDI ESGF nodes.

The services supporting the FAIR data principles with respect to data identification, citation and long term archival and access are provided in section 2.3, 2.4 and 2.5.

The Climate4Impact portal services were completely upgraded and are characterized in section 2.6.

Statistics for the provisioning of data near compute services are provided in section 3; section 4 summarizes the data standards and documentation related support services.

[1] IS-ENES3 deliverable D7.1 "First KPI and TA report for ENES CDI data services", https://is.enes.org/documents/deliverables/d7-1-first-kpi-and-ta-report-for-enes-cdi-data-services

1. ENES CDI data and metadata services: Objectives and Overview

1.1 Overview

In IS-ENES3, installations across Europe join to provide a consistent set of services to the European climate research community, including downstream communities like climate impact research. The ENES Climate Data Infrastructure (ENES CDI) provides: (1) access services on CMIP and CORDEX data from the Earth System Grid Federation (ESGF), the archival system (the IPCC Data Distribution Centre, DDC), and the Climate4Impact portal, (2) processing services, and (3) services on data documentation and standards. These services are mainly offered through virtual access (VA). Users have also the possibility to apply for virtual workspaces through a trans-national access (TA), which allows them to remotely access not only the data pools but also the IS-ENES3 computing infrastructure (high performance computers (HPC) and clusters) hosting the data. In comparison to the first reporting period, the compute service offered via TA is now accompanied by a stronger lightweight virtual access-based service to support users in shorter term (computationally inexpensive) analysis and testing activities without the need to apply for TA resources. The overall goal of the IS-ENES3 data service activities is to provide operational support to the climate and climate impact research communities and other communities using the model data and tooling provided by IS-ENES3. The VA/TA activities will also provide support for communities that are new to using climate model data.

The following service activity report is structured according to the individual data service areas, reflected in different service tasks: data dissemination, archival and user support (Task1), compute services (virtual access in Task2 and TA based in Task3) and data standards related services (CF convention and data request in Task4 as well as ES-DOC in Task5).

1.2 Service statistics and performance indicators

The performance indicators (PIs) and key performance indicators (KPIs) are summarized in the following table and did not change in comparison to the previous reporting period:

ESGF data download KPIs and PI	KPI: Number of downloads (EU/no-EU/no geo-located)
	KPI: Downloaded data volume (EU/no-EU/no geo-located)
	KPI: Number of distinct users (EU/no-EU/no geo-located)
	KPI: Number and percentage of emails answered in the user support mailing list by an ENES member
	PIs: CORDEX specific number of downloads, volume, and distinct users and number of answers to the new CORDEX user support mailing list
Replication and archival PIs	Number of TB of original data
	Number of TB of replicated data
	Number of TB of overall volume
Data citation PIs	Number of registered DOI registered to DataCite
	Number of revisions of citation information published to DataCite
	Number of citation entries added to the service database
Persistent data identification PIs	Number of original and number of replica CMIP6 datasets
	Number of original and number of replica CMIP6 files
DCC PIs	Number of downloads
	Downloaded data volume
	Number of distinct users (EU/no-EU/no geo-located)
	KPI: Unique Users
	KPI: Number of access to the users' personal space (Basket Requests)

Climate4Impact KPIs and PIs	PI: Number of map visualisations requested by users (WMS Get Map Requests)
	PI: Number of processing functions executed by users (WPS Execute Requests)
	PI: Number of data subsetting requests by users (WCS GetCoverage Requests)
	PI: Number of hits
CF data model PI	Release of package updates
CF Standard Name PIs	Publication of new versions of the table
	New terms published
CMIP Data Request PIs	Issues resolved
	Releases
ES-DOC PIs	Issues registered on the web service
	Number of documentation search and web site visits
	Number of questions to the helpdesk
	Metadata generated by the cdf2cim process of the ESGF publisher

Table 1: KPIs and PIs for the ENES CDI data services

2. ESGF data dissemination, data archival and Climate4impact services (Task 1)

The ESGF data dissemination is based on a federation of European data node and data portal installations. Three tier 1 sites (DKRZ, UKRI, CNRS-IPSL) host data portals and additionally act as replica sites, such that also non-European CMIP6 data can be accessed from European sites. The data access statistics are based on the European data node installations at DKRZ (Germany), UKRI (England), CNRS-IPSL and CNRM (France), BSC and UNICAN (Spain), CMCC (Italy) as well as LIU/SMHI (Sweden, concentrating on CORDEX data distribution).

2.1 ENES CDI ESGF data download KPIs and PIs

The ESGF data download KPIs quantify the monthly *number of files* downloaded from the European ESGF data nodes and the associated *data volume* (with a distinction between complete and partial downloads), as well as the monthly *number of distinct users* successfully performing the downloads.

The KPIs are collected as part of the ESGF Data Statistics service[2] developed and hosted at CMCC. They are summarized in Figure 1.

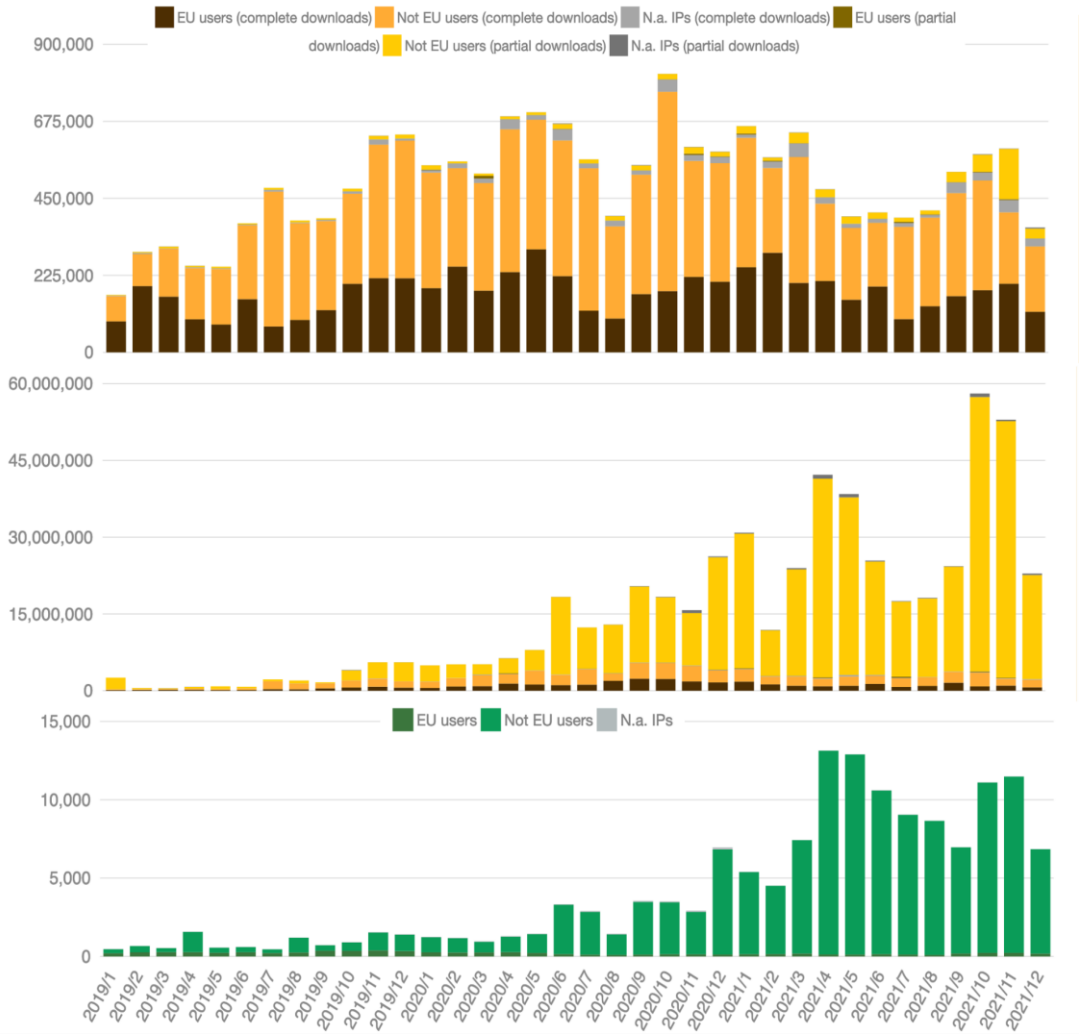


Figure 1: ESGF data download KPIs (stacked charts): number of downloaded files (top), associated data volumes (in GB) (center), and distinct clients per month[3] (bottom).

[2] http://esgf-ui.cmcc.it/esgf-dashboard-ui/isenes3-kpi.html

[3] Due to the EU General Data Protection Regulation (GDPR) and the new CMIP6 open data policy, by monthly distinct users we mean the "average number of monthly distinct clients per data node". With respect to the other two

The decreased overall data download volume since March 2021 can be attributed to the finalization of the CMIP6 input data requirements in support of the IPCC working groups and the associated availability of these CMIP6 data collections as part of the established CMIP6 data replica pools at DKRZ, CNRS-IPSL and UKRI. Therefore the need to download and access these data via European ESGF nodes was reduced significantly. The growing number of distinct users after March 2021 is probably also related to the before mentioned IPCC Working group timeline: after the finalization of the IPCC related CMIP6 input data collections these data are very interesting for downstream community usage, and thus new users outside the core climate modeling community accessed the data.

During the reporting period, 9.5 PB of data were downloaded from European ESGF data nodes through over 473 million downloads by a mean of 7191784583 distinct users per month. Of these, 3.2 PB were downloaded by a mean of 15220572 distinct users in the EU per month through 26 million downloads, while 0.4 PB were downloaded by a monthly mean of 316534 non-geolocated users through over 6 million downloads. On average, 530 TB of data (179 TB in the EU) was downloaded per month. Monthly data download volume peaked in October 2020 at 814 TB, while in the EU it peaked in February 2021 at 293 TB. In comparison to the first reporting period the averaged monthly downloaded data volume stayed nearly unchanged, whereas the mean number of monthly users increased by more than a factor of 10 (especially non-European users). This can be explained by the increased exploitation of the large CMIP replica pools in Europe, eliminating the need to download data from the ESGF data nodes.

KPIs (number of files and data volume), the distinct users metric is non-additive, which explains why we calculated the average instead of the total.

- **KPI: Number and percentage of emails answered in the user support**

The ESGF user support is mainly handled using the esgf user support mailing list. We have approx. 10 new requests per week and the majority is answered by IS-ENES3 partners (in particular DKRZ).

Since July 2020 there were around 900 user requests of which approx. 80% could be resolved.

In Figure 2 the number of CORDEX file downloads are summarized, whereas in the Figure 3 the associated data volume is illustrated. This statistic includes CORDEX datasets for all CORDEX domains. User from Europe are dominated in the statistics since the CORDEX datasets for different domains are very actively used in many European projects. In contrast, users from other continents are mostly interested in CORDEX datasets for their region only. Additionally, the Euro-CORDEX ensemble (EUR-11) is the largest one among other CORDEX domains and provides many sub-daily datasets that also can explain a large number of downloads by users from Europe. Because of organizational issues it was not possible to include the latest December numbers in the figures.

- **PI: CORDEX: Number of downloads (EU/no-EU/no geo-located)**
-

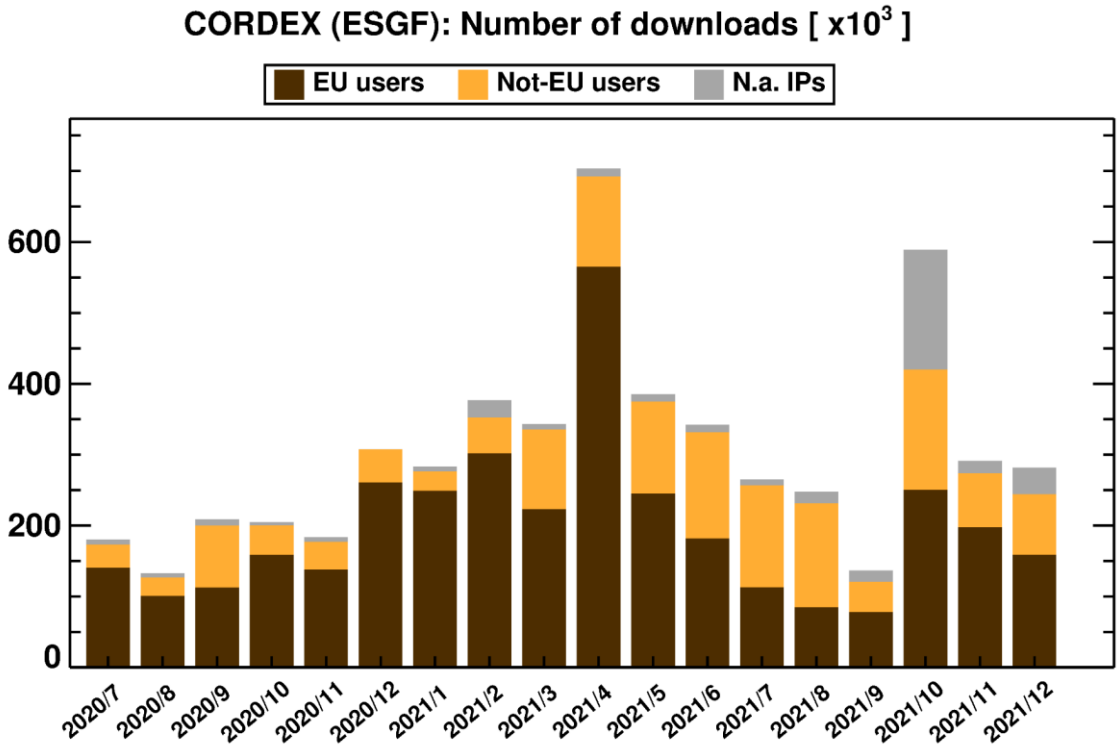


Figure 2: Number of ESGF CORDEX data downloads

- **PI : CORDEX: Downloaded data volume (EU/no-EU/no geo-located)**

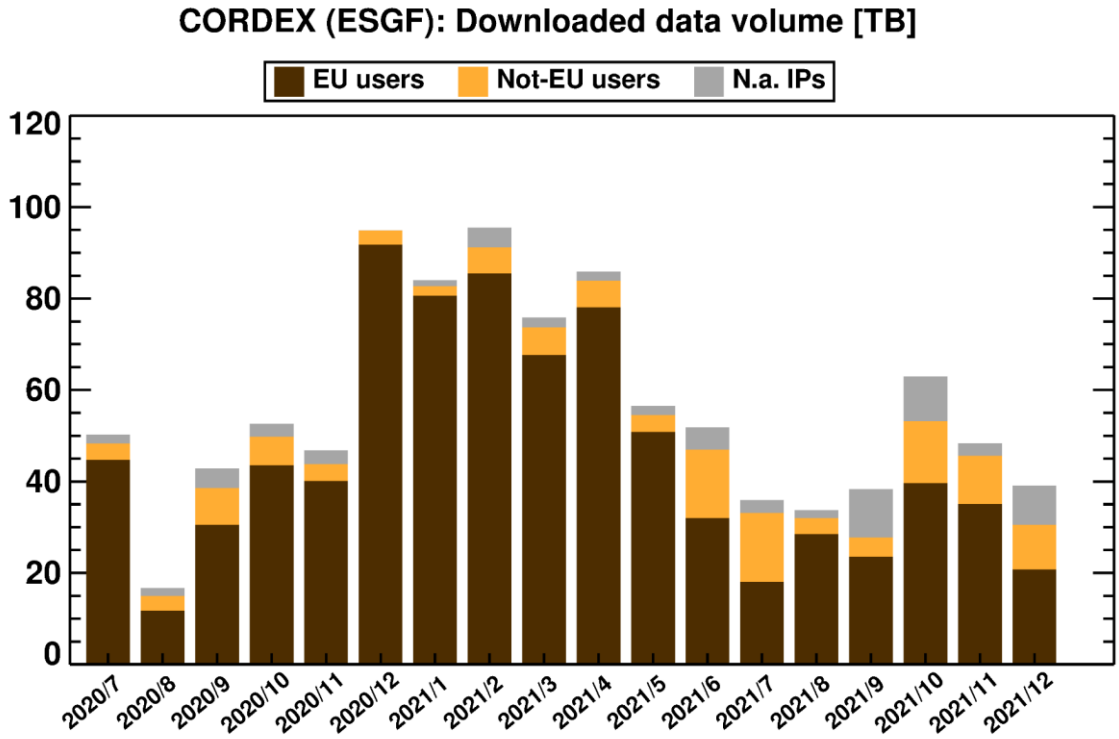


Figure 3: Volume of ESGF CORDEX data downloads

- **PI : CORDEX: Number of answers to the CORDEX data support mailing list**

The CORDEX data support mailing list (datasupport@cordex.org) is handled by the International Project Office for CORDEX (IPOC) hosted by SMHI. All questions received are sorted first and then forwarded to relevant experts from the CORDEX community. About 100 questions have been received since July 2020. About 10 of them, related to complex ESGF issues have been forwarded to the ESGF support while the rest have been answered.

- **Other interesting metrics coming from the ESGF Data Statistics service**

Other interesting data statistics coming from the ESGF Data Statistics service are shown below in Figure 4 and Figure 5 illustrating the distribution by continent of the clients which downloaded CMIP6 and CORDEX data over the European data nodes; Figure 6 shows the top twenty CMIP6 variables downloaded (in GB) from the European data nodes. This statistics is only an indication on the scientific interest in specific variables as the ranking is also strongly influenced by the data organization. Thus e.g. we see that the first 4 most popular variables (ua, va, ta, hus) are 3D

variables. Size of one 3D dataset is much larger than for a 2D dataset as precipitation or tas. 3D datasets are also chunked in many files compared to 2D ones. The statistics in the donut chart are based on the total size of the downloaded files, not on a number of datasets. If one 3D dataset has been downloaded (e.g. ua) it gives a higher contribution to the statistic than for example a 2D dataset like precipitation or temperature. If we count only datasets, in both cases (ua and pr), only one dataset has been downloaded.

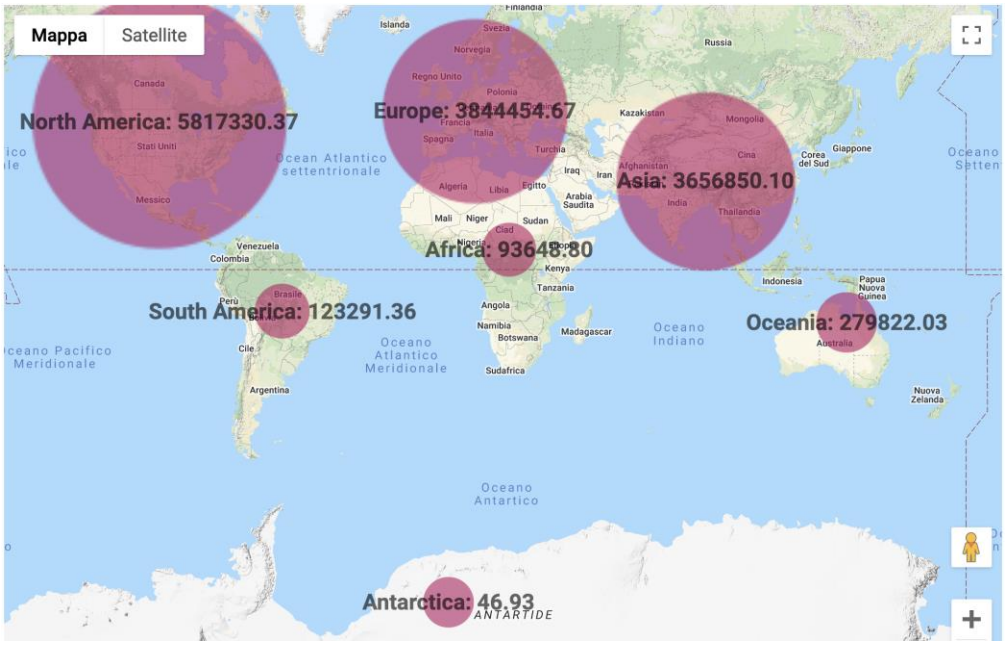


Figure 4: CMIP6 downloaded data volume (in GB) by Continent (from European data nodes)

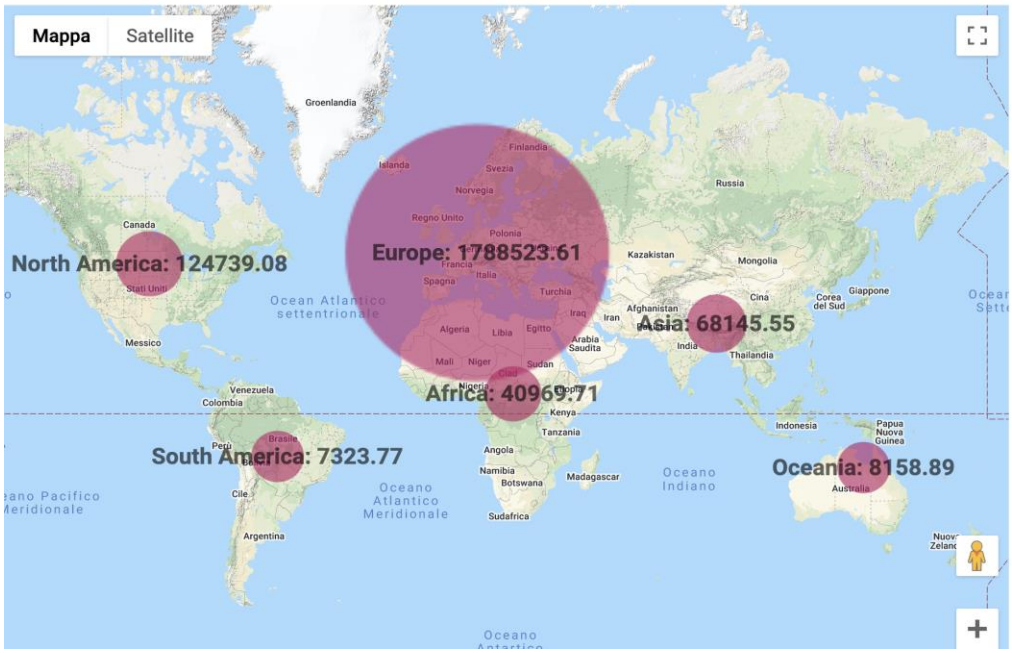


Figure 5: CORDEX downloaded data volume (in GB) by Continent (from European data nodes)

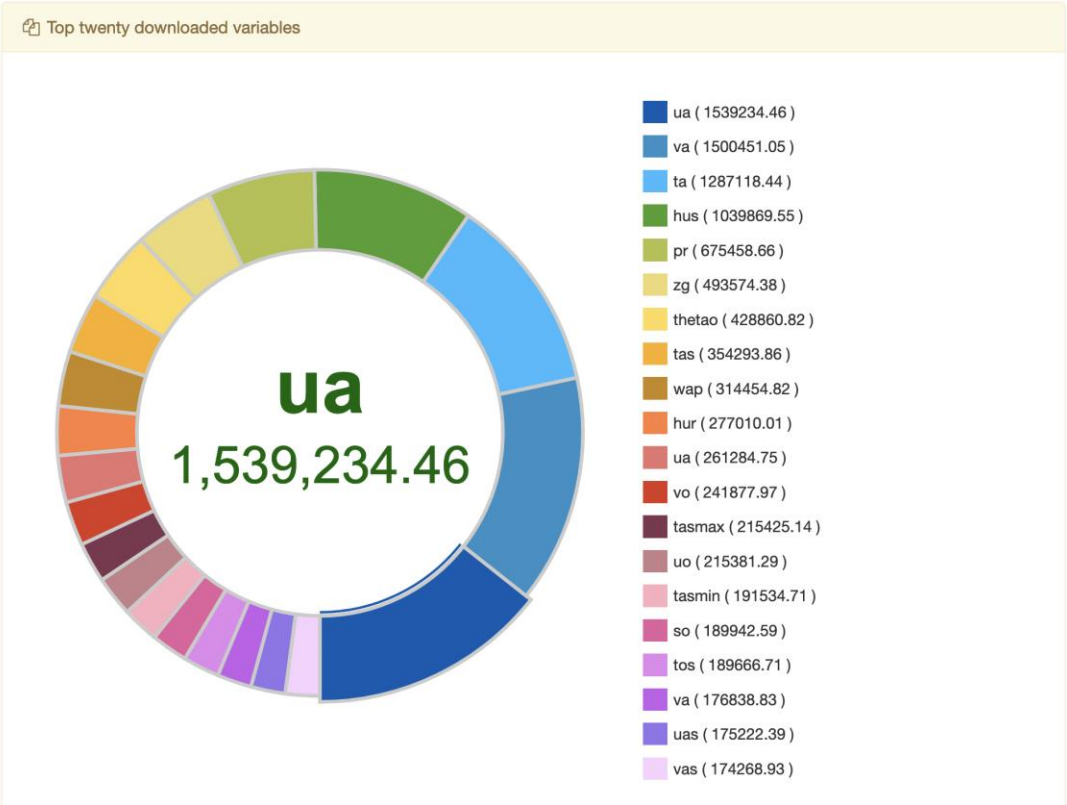


Figure 6: Top twenty CMIP6 downloaded variables (from European data nodes)

The ESGF Data Statistics service also provides general information about the data published over the federation and the projects and data nodes included until the end of 2021 into the environment. Figure 7 depicts an overview of the available metrics over 24 data nodes, consisting of about 13M of published datasets and 830M of downloaded files corresponding to more than 27 PB coming from 172 countries and 32 projects.

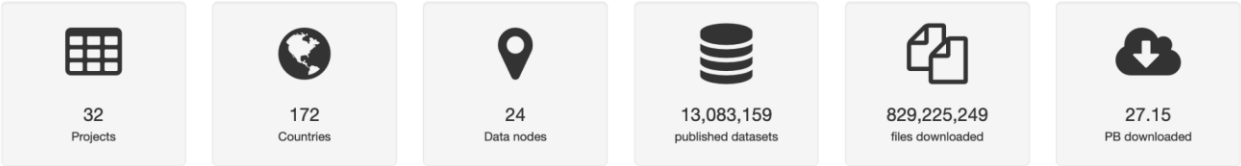


Figure 7: Overview of ESGF

In particular, a total of about 12M datasets and 22 PB of CMIP6 data are available over the whole federation and about 180 thousand datasets and 1,5 PB of CORDEX data (see Figure 8).

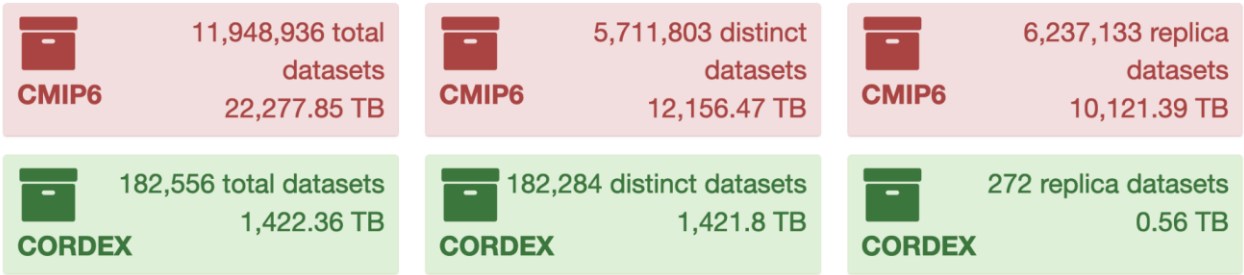


Figure 8: CMIP6 and CORDEX data available over ESGF

CMIP6 constitutes the major part of published data in the ESGF. Indeed, the total amount of published data is about 31 PB with CMIP6 covering about 22 PB of this data.
On the European nodes, about 3,3M datasets and more than 10 PB of CMIP6 data are available. Regarding CORDEX, European nodes provide about 124 thousand datasets and 760 TB.

2.2 Replication and Archival PIs

The three tier 1 ESGF installations at DKRZ, UKRI and CNRS-IPSL maintain large CMIP data pools which provide not only original CMIP6 model outputs from their associated modeling centers but also collect and provide large CMIP6 collections replicated from remote ESGF nodes. This enables the European research community to access the vast majority of CMIP6 data at European data providers, which also provide data near processing facilities thus reducing the need to download data by individual scientists and research groups. This reduced need to download large volumes of data by European users is clearly reflected in the ESGF access statistics provided above.

Institution	Original data (TB)	Replicated data (TB)	Overall volume (TB)
DKRZ	1400	2400	3800
UKRI	1700	1700	3400
CNRS-IPSL	1500	1300	2800

Table 2: CMIP data pool volume (original and replicated data collections) at European ESGF sites

Associated to these data pools there are user support and data curation services: users often request additional data to be included in the data pools to cover the needs of their data analysis activities. Additionally the content of the data pool needs to be continuously updated to include new versions of data and to remove replicated data which was retracted from ESGF.

The archival process of key CMIP6 data collections just started at DKRZ and will be completed in the final year of IS-ENES3. The archival process initially concentrates on often used core data identified by the IPCC Technical Support Unit (TSU) and will be extended to include data used within the IPCC AR6.

2.3 Data citation PIs

The citation service provided a complete coverage of data citations for all CMIP6 data published in the ESGF for IPCC WGI AR6[4] on the data and literature cut-off date 2021-01-31. The CMIP6 references became included in the Appendix II of the AR6 as well as in the metadata of IPCC WGI Interactive Atlas[5].

In the 1.5 years reporting period, entries due to new participant registrations were added to the database, DOIs were registered for completed citation information and available data in the ESGF, and metadata was updated and curated. The metadata changes include user changes and automatic curation measures such as adding paper references using the Scholix interface. All metadata updates are also sent to DataCite (Table 3). Manual metadata curation efforts are not included. The temporal development of DOI registrations for the projects CMIP6 and input4MIPs is depicted in Figure 9. The frequency of CMIP6 data published in the ESGF and therefore the number of published DOIs decreased after 2019 from an annual registration of 1 375 DOIs in 2019 via 991 in 2020 to 334 in 2021. The short term increase of DOI registration rate beginning of 2021 is related to the closing date of IPCC working group contributions and their associated need to complete related CMIP6 data citation information. Table 3 provides a summary of the activities related to the DOI provisioning in the reporting period: number of DOIs registered, DOI metadata updates (related to participant info, model info, author info) and DOI reference updates.

[4] https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/
[5] https://interactive-atlas.ipcc.ch/

Nr.	Activity	Count
	Registered DOIs	669
1.3	Inserted entries due to new participant registrations	1 707
1.4	Model description inserts and updates	16
1.6	Default author and contact information added for experiment entries based on model/MIP information	1 776
1.7	Paper references added, provided by Scholix	325
2.15	Metadata changes synchronized with DataCite	5 813
3.3	Missing (required) contactPerson added using first author	18

Table 3: Activities of the citation service (DOI registration and metadata curation) in the reporting period; numbers according to quality checks documented for Copernicus CDS at https://bit.ly/CMIP6_Citation_Quality.

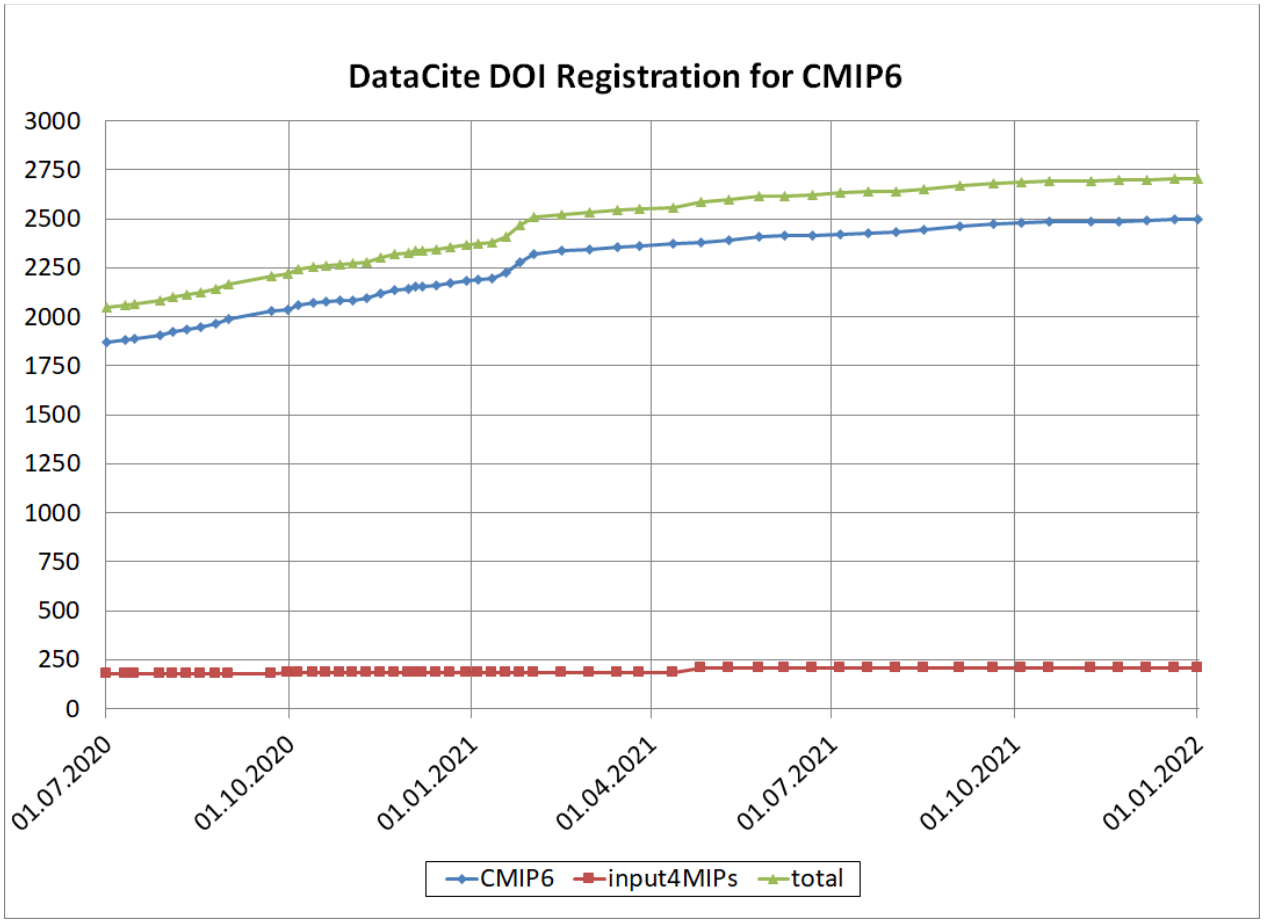


Figure 9: Temporal development of DOI registration for the projects CMIP6 and input4MIPs and metadata updates sent to DataCite in the reporting period.

2.4 Persistent Identification PIs

For each CMIP6 file and each CMIP6 dataset (set of files, corresponding to full time-span) a persistent identifier (PID) is registered as part of the ESGF publication process. Thus in the following the overall number of file and dataset publications is used as a basic measure for the persistent identifier service usage. The actual PID service usage is larger as there are continuous curation processes to resolve PID registration problems (e.g. update erroneous registrations) in addition also all unpublication activities involve PID service interactions.

- There are currently approx. 12 million CMIP6 datasets published to the ESGF. Around 5.5 million of those are original data and 6.5 million are replica.
    - Around one million of the original datasets have been published between July 2020 and November 2021
- There are currently approx. 55 million CMIP6 files published to the ESGF. Around 26.5 million of those are original data and 28.5 million are replica.
    - Around 10 million of the original files have been published between July 2020 and November 2021
- Thus the number of registered PIDs for CMIP6 is approximately 32 million (original data sets and original data files - replicas are not assigned dedicated new PIDs, yet they are linked to the originals in the PID metadata). The distribution number and EU/nonEU distribution of assigned PIDs is illustrated in Figure 10.

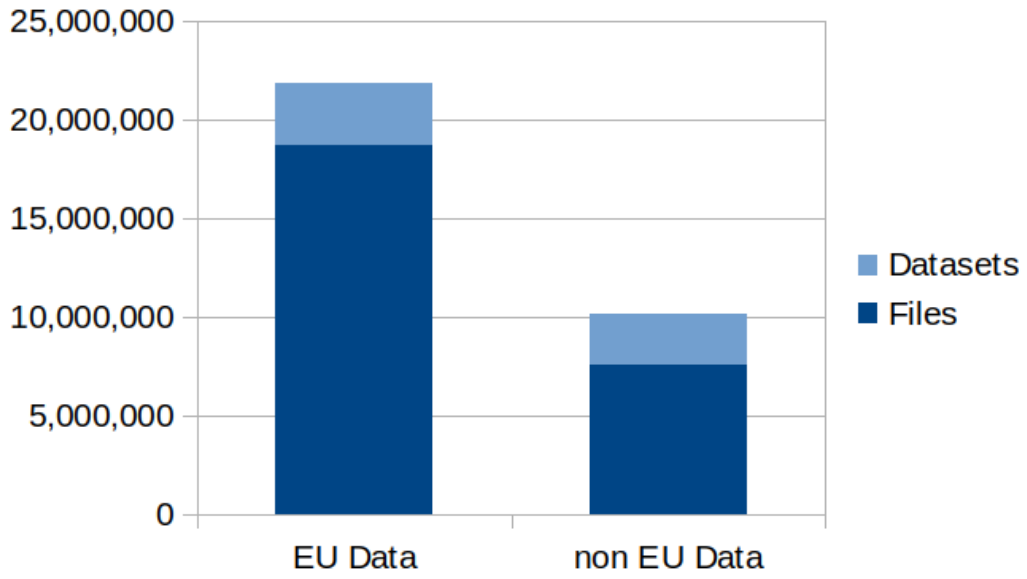


Figure 10: Number of PIDs assigned to Files and Datasets of European / non European data nodes

2.5 DDC PIs

The Data Distribution Center (DDC) of the Intergovernmental Panel on Climate Change supported IPCC authors through the Virtual Workspaces (see section 2.2), which provided direct access to DKRZ's data pool (subset of CMIP6, CMIP5, CORDEX datasets requested by IPCC authors, DKRZ users or IS-ENES users), common software packages and some amount of storage for collaboration on figure creation.

IPCC WGI Technical Support Unit (TSU) was supported in gathering information on CMIP6 data usage from the IPCC authors and in compiling a list of CMIP6 datasets for long-term preservation in the DDC at DKRZ. The data usage information provided by the authors remained incomplete. Therefore the CMIP6 data subset specified by the IPCC authors at the start of AR6, which is also used by Copernicus CDS, is used for archival. Data archival of these datasets was started in November 2021 as a first data archival step. This data will be complemented with the datasets in the list of WGI TSU, which are not part of the Copernicus CDS, and complemented with references to AR6 WGI chapters/figures, where the data was used.

DDC users of data for past IPCC assessment reports (FAR to AR5; AR5 data can be downloaded from the DDC and ESGF portals) downloaded a total volume of 570 TBytes in 1.1 million individual dataset downloads (Figure 11) or average monthly downloads of ca 32 TByte/month and ca. 61 000 dataset downloads per month. No trend in the monthly downloads is observed over the reporting period. Raw dataset downloads during the second half of 2020 until mid of January 2021 included the massive (every few seconds) automated single dataset downloads from a single European site via the ESGF. After contacting the institution at the beginning of 2021, this stopped. These technical downloads were excluded from the statistics, which might lead to differences to the results from the ESGF dashboard.

The continental user distribution of dataset downloads shows a share of European users on the total dataset download during the reporting period of about one third (33 %; Figure 12). The percentages of users located in other continents were: 52 % Asian, 10 % North American, 4 % African and less than 1 % each from Australian and South American users.

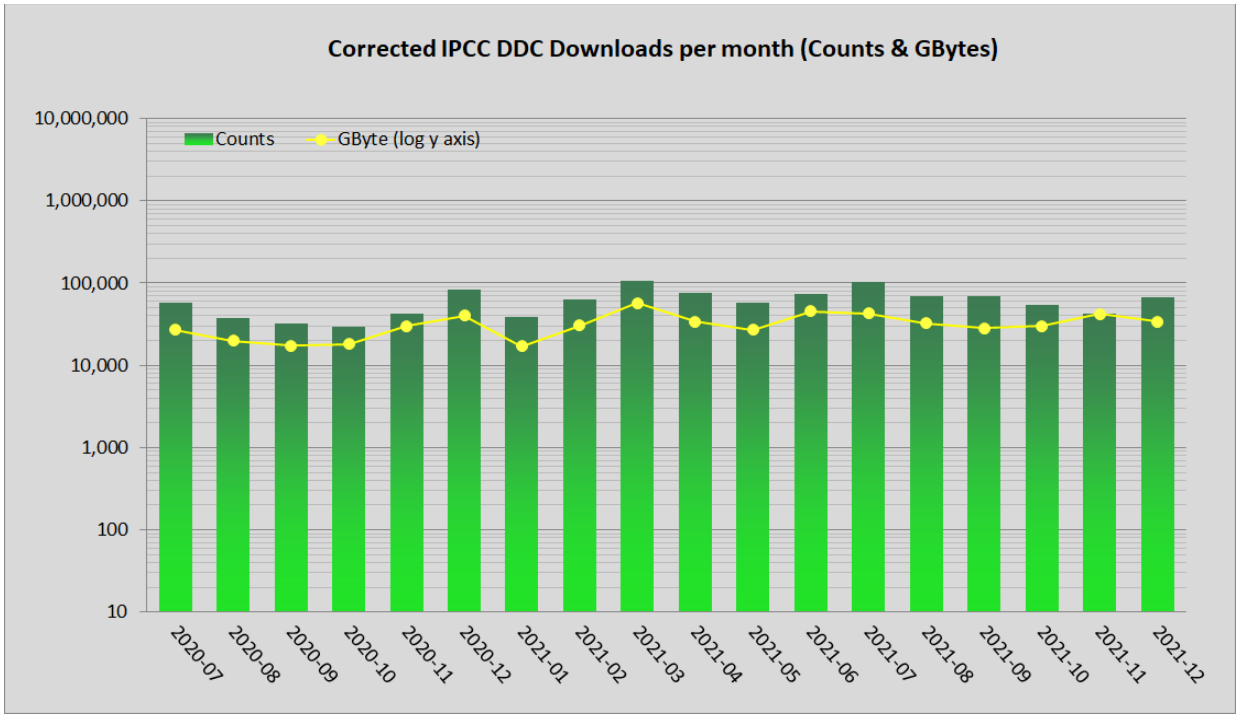


Figure 11: Monthly data download volume and dataset counts from the IPCC DDC at DKRZ in the reporting period. Numbers are corrected by subtraction of massive automated downloads of a single dataset by an ESGF user.

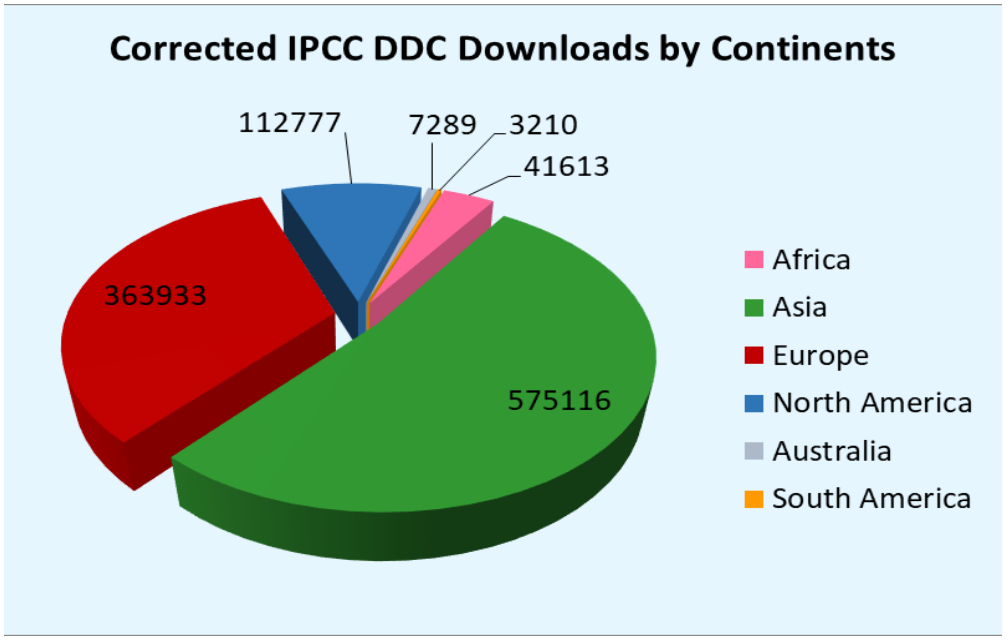


Figure 12: DDC dataset download counts via ESGF and the DDC portal per continent of user's residence in the reporting period. Numbers are corrected by subtraction of massive automated downloads of a single dataset by an ESGF user.

2.6 Climate4Impact KPI and PIs

The Climate4Impact[6] portal is undergoing a large reimplementation towards the release of a new version, Climate4Impact v2[7], which is already available to *alpha* testers and underwent expert review reported in D7.2. Despite the ongoing upgrades, we have kept the first version of the portal operational and collected statistics according to the updated KPIs presented in D7.1. These are summarised again below:

- KPI Unique Users, that is, the actual number of registered users who accessed the service.
- KPI Number of access to the users' personal space (Basket Requests).
- PI Number of processing functions by users (WPS Execute Requests).
- PI Number of map visualisations requested by users (WMS Get Map Requests).
- PI Number of data subsetting requests were done by users (WCS GetCoverage Requests).
- PI Number of hits.

In Table 4 we report the numbers associated with these metrics.

Month	Hits	wps Execute Requests	wms Get Map Requests	wcs Get Coverage Requests	Basket Requests	Unique/Logged Users
10-2020	26597	41	676	0	1669	41
11-2020	28233	5	1886	0	3109	54
12-2020	28078	17	942	0	3515	41
1-2021	17413	13	1373	0	2865	39
2-2021	13469	6	1553	0	1901	48
3-2021	34930	52	5801	79	4206	63
4-2021	14675	65	799	26	1955	46
5-2021	3063	3	215	0	578	25
6-2021	33781	107	1265	0	2493	48
7-2021	19568	73	2883	0	3397	30
8-2021	16235	99	706	0	2822	32

[6] https://climate4impact.eu
[7] https://dev.climate4impact.eu

9-2021	3915	13	142	0	802	25
10-2021	15407	2	398	0	2424	63
11-2021	2711	0	82	0	373	21

Table 4: User activity at the Climate4Impact portal in 2020/2021 (RP2)

For the new version of the portal, we are updating the collection of KPIs. Besides the most traditional web based KPIs, refining the current ones. These will take into account metrics associated with new underlying components aimed at the provision of JupyterLab workspaces, as well as the execution of data-staging and processing workflows, see D7.2 and D10.3.

Climate4Impact User support

We provided continuative support for operations of both Climate4Impact's versions, reacting to downtimes and supporting experts in the evaluation and use of the new portal. In v2 we have implemented an updated user feedback page[8], besides producing help material explaining how to search and download data[9] and how to proceed to perform custom analysis[10]. For the latter, we also published a collection of sample analysis notebooks that can be executed in C4I v2[11]. Finally, we have activated a dedicated email address where users can ask questions and get support on any particular issue related to the portal directly by the C4I development team at KNMI.

[8] https://forms.gle/m9FRTABa7Xq79Kop8
[9] https://dev.climate4impact.eu/c4i-frontend/helpC4I
[10] https://dev.climate4impact.eu/c4i-frontend/helpSwirrl
[11] https://gitlab.com/is-enes-cdi-c4i/notebooks

3 Compute services

Compute services are provided by 4 installations at DKRZ (Germany), UKRI (UK), CNRS-IPSL (France) and CMCC (Italy). They are separated into light weight low resource usage access services provided under the virtual access mechanism as well as access to compute platforms which are provided under the transnational access mechanism (and which involve an application review procedure, which is explained in detail in the previous PI and KPI report (Deliverable D7.1). Since the beginning of 2021 only two installations (DKRZ and UKRI) provide this transnational access possibility (see IS-ENES3 second amendment) because of the lack of appropriate service application proposals.

3.1 Compute service: derived data products and web services (VA, Task2)

CMCC

In the following, two tables provide some information about the exploitation of the Virtual Access services hosted at CMCC. In particular, the first table refers to the ESGF Data Statistics service web interface, reporting data usage and publication statistics about the ESGF federation. The second one reports some access information to the CMCC Analytics Hub web portal, the web access point to the CMCC Analytics Hub, which provides a user-friendly data analytics environment to support scientists in their daily research activities on top of 11TB of CMIP6 data available so far.

ESGF Data Statistics service web interface

From 01/07/2020 to 31/12/2021 (2nd Reporting Period).

Total users	Sessions	Page view	Countries
546	1655	5520	31, mostly from Italy, United States, Australia, France, Germany, United Kingdom

CMCC AnalyticsHub web portal

From 01/01/2021 to 31/12/2021 (from the first release of the portal).

Total users	Sessions	Page view	Countries
1128	1856	3821	112, mostly from Italy, United States, Germany, Japan, India

DKRZ

DKRZ virtual access platform (login nodes, Jupyter hub portal, CMIP data pool)
From 01/07/2020 to 29/11/2021 (2nd Reporting Period)

Registered Users	Active Users (> 0 node hours)	Node hours used	Countries
77	21	> 1300	**Europe:** Germany, Sweden, Spain, France, Italy, UK, Switzerland, Ireland, Austria, Netherland, Kroatia, Norway, Sweden **Eastern Europe:** Estonia, Hungary, Czech republic **Other:** USA, Canada, Hong Kong, Afrika, China, Egypt

Based on the support questions asked by the service users the main motivation for them to use the service was the direct access to the CMIP6 data pool and CORDEX data collections. Additionally other collections hosted at DKRZ were requested and used e.g. CMIP5 data and ERA5 reanalysis data.
A public climate service center in Africa and a private climate service provider in Spain especially expressed their gratitude to being provided with free computing resources closely attached to a large CMIP and Cordex data pool.
Also authors of the IPCC Working groups (especially WG 1 and 3) have been supported by this VA service to be able to (re-)generate data products.

CNRS-IPSL

The generic VA compute access at IPSL has been reinforced during this 2nd reporting period and fully described in the "Second Release of the ENES CDI software stack document" - D10.3. During this period, more than 200 new users have registered to the IPSL Compute and Data center (Figure 13) with 5% of them from European partners (especially from the Complutense University of Madrid and in the Copernicus context).

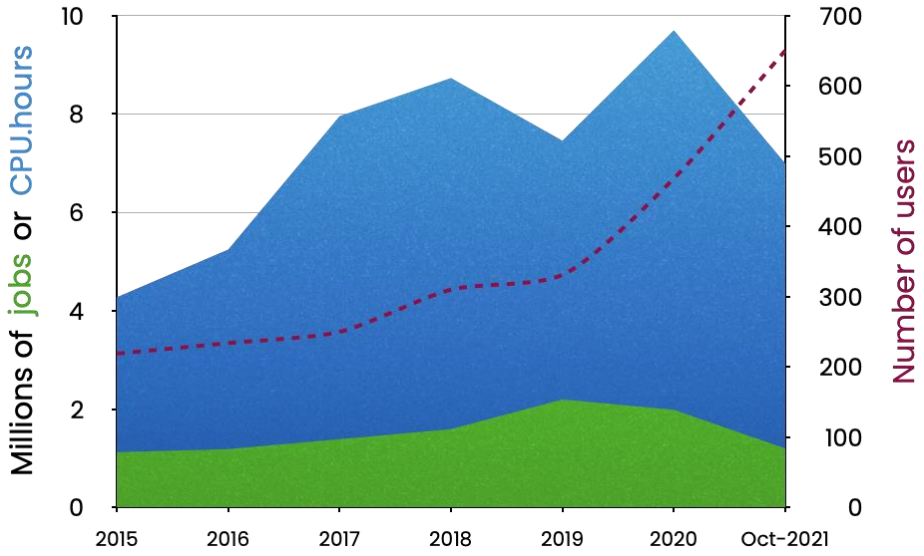


Figure 13. Compute service usage at CNRS-IPSL in terms of jobs and CPU-hours

The WPS deployment will be enforced in the coming year with additional dedicated computing resources to be opened to CNRS-IPSL computing center users and extended VA access to a broader EU community (not only in the Copernicus context).

UKRI

The UKRI virtual access platform (login nodes, Jupyter hub portal) provides users with access to the extensive CEDA data holdings and is also used as a collaborative workspace. The data holdings include the CMIP6 archive, large volumes of Earth observation data from Sentinel and other missions, re-analyses and many other observational datasets. From 01/07/2020 to 29/11/2021 (2nd Reporting Period) access was granted to 1677 users.

3.2 Compute service: Virtual workspaces (Transnational Access - TA, Task3)

The reporting period covers three transnational access allocation periods (July 2020 - January 2021, Feb - July 2021 and July 2021 - January 2022). The following table summarizes the number of successful applications during these periods and their allocation to TA service providers (DKRZ and UKRI).

successful applications	allocation period	assigned to DKRZ	assigned to UKRI/Jasmin
5	July 2020- January 2021	3	2
6	February 2021 - July 2021	3 (1 used VA)	2
3	July 2021 - January 2022	2	1

Table 5: Successful TA applications

DKRZ

Overall the successful applications used a small amount of compute resources (very few job submissions to the HPC system, mostly interactive sessions e.g. via Jupyter hub). Only two applications showed higher activity, and also applied for extensions of their allocations. Most groups named the effects of Covid as the main reason for delay/cancellation of their projects. The unused compute resources were allocated to the VA activity as some groups (especially one company (SME) and an IPCC activity related research group) requested additional capacity to generate derived data products. Two additional pre-access allocations were provided, one switched to the VA access mechanism, the other did not apply for a full allocation. The experience so far showed no clear picture why this type of service was not exploited fully. User support showed high interest in the access to the large data pools, yet researchers did not use their TA compute resource allocations. One possibility is that they are associated with research groups already having allocations at DKRZ. Others (a concrete example is a climate service center located in Africa) do not meet the requirements for TA allocations (majority of researchers coming from Europe) and thus were supported via VA. A short summary for the activities of the individual applications (characterized by their identification code used in the evaluation process) is provided in the following:

2nd allocation period (July 2020-January 2021) :

- **(100620_hu)** "Vulnerability assessment of the energy sector (electricity, gas and district heating) and its infrastructure based on geology and climate data - Energy infrastructure in Hungary": No compute resource usage
- **(310520_se)** "Changes in extreme precipitation and droughts over East Asia: historical attribution and future projection" (CEPAD): no compute resource usage, extension granted
- **(220520_no)** "Atmosphere Sea Ice interactions in the new Arctic" (ARIA): no resource usage because of Covid

- **(290520_no)** "Multi-model assessment of Arctic Ocean Heat" (ArcHEAT): used pre-access test allocation, re-applied for 3rd allocation period as "NORTH"

3rd allocation period (February 2021 - July 2021):

- **(281020)** "CRED/MedCORDEX": no compute resource usage
- **(301020_no)** "NORTH, A NORTHern perspective on CMIP6 climate model variability": supported and small compute resource usage
- **(301020_1_no)** "DecNorth, Multi-model assessment of decadal climate predictability in the North Atlantic": supported and small compute resource usage
- **(311020_cz)** "MCS LOVE CCS,Multimodel Climate Simulations - LOcalization, Validation and Evaluation of Climate Change Signal": supported, switched to ECAS (VA) service

4th allocation period (July 2021 - January 2022):

- **(23_05_21_no)** "Evaluation Weakening Overturning Circulation": activity delayed
- **(28_05_21_nl)** "Accessing Climate Change impact on the Hydrological Cycle using the eWaterCycle Platform for Open Hydrology": related project activity moved to 2022, allocation extended

UKRI JASMIN

2nd allocation period:

- **CLI-SEE-6 [100620_ro]:** "Exploring Climate extremes over SE Europe in CMIP6 projections", A small storage and compute resource has been assigned to the project. The PI has not been in touch.

3rd allocation period:

- **PRIMAVERA (271020_uk):** "PRocess-based climate sIMulation: AdVances in high resolution modelling and European climate Risk Assessment (PIMAVERA)". The HighResMIP team has used a JASMIN shared workspace (up to 350TB) for analysis of CMIP6 and PRIMAVERA data. Additional tape storage has been used to store additional simulations. A JASMIN virtual machine is used to host a web-application to allow users to request/manage migration to/from tape.
- **MODELANT (201020_es)**: "Multi-model comparison of interannual-to-decadal Atlantic variability modes", A small storage and compute resource has been assigned to the project. The project was delayed because of staff availability issues on the side of the project team, rescheduled to run from Feb. to June 2022.

4th allocation period

- **SNAPSI (30_03_21_il)**: "Stratospheric Nudging And Predictable Surface Impacts", CEDA have been providing science support to the project to enable the implementation of metadata standards based on the CMIP6 data standards. The transfer of example SNAPSI datasets began in December 2021 to test the quality control and ingestion pipeline..

CNRS-IPSL and CMCC

CMCC, and CNRS-IPSL did not provide TA access in the second reporting period, focusing their efforts into VA access only as agreed in the last amendment.

4. Data standards and documentation

4.1 Support for CF convention and data request (Task 4)

CF Data Model and CF standard names

The demand for new terms in the CF Standard Name list has been relatively light during this reporting period, with just 73 new terms added. This reflects the usual pattern of having high demand in the early stage of the IPCC Assessment cycle, when new scientific questions are being explored, and a quiet period when results are being analyzed.

In the reporting period 5 new versions of the CF table were generated, which are listed below associated with the version date. 73 new terms were added in this period, taking the total to 4495.

- 1.8.6 July 24, 2020
- 1.8.7 Oct 9, 2020
- 1.8.8 Dec 18, 2020
- 1.8.9 May 25 2021
- 1.9.0 Sep 21, 2021 and 1.9.0.1 Oct 12, 2021

To support data analysis a python library is maintained implementing the complete CF data model called cf-python[12]. In the reporting period 5 releases and one minor upgrade were produced . The release dates are as follows:

- Version 74, 04 August 2020
- Version 75, 15 September 2020
- Version 76, 13 October 2020
- Version 77, 19 January 2021
- Version 78, 21 September 2021

4.2 The CMIP Data Request

The CMIP6 Data Request service has operated smoothly, with low levels of demand as expected at this stage of the CMIP and IPCC cycles. There has been one new release resolving four issues. The resolved issues are summarized in the following with references to the full technical descriptions:

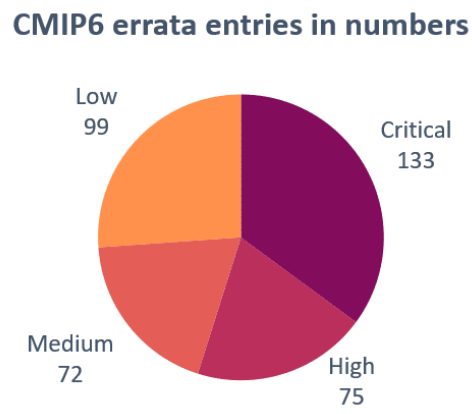
- Corrected specification of duration of the "past1000" experiment #126;
- Added experiment details for the PMIP project #123;
- Added experiment details fpr the CovidMIP project #128;
- Definitions of two variables, "tntrl27" and "tntrs27", corrected #408;

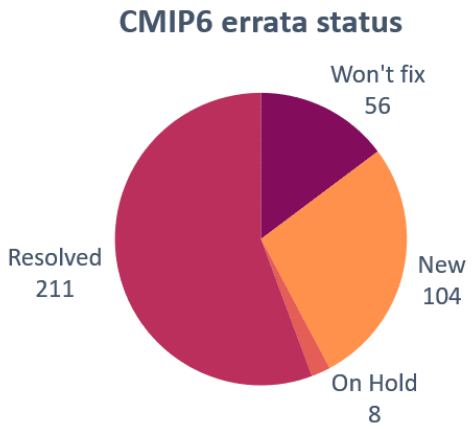
4.3 ES-DOC operational support for CMIP6 (Task 5)

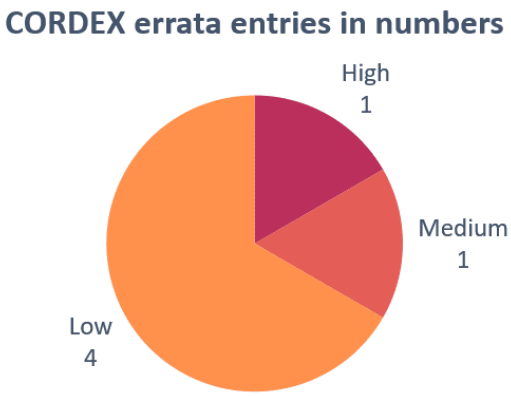
The Errata Service documentation[13] has been entirely revised and improved to accompany data providers, one side, to register and manage issues and the model data users, on the other side, to query the service about the datasets they downloaded. The next ESGF user survey includes several questions about the Errata Service usage to gather all users' needs and requirements. Many users already suggested improvements of the service or new features that are under discussion together with ES-DOC, WIP and WGCM and that will be developed and deployed during RP3.

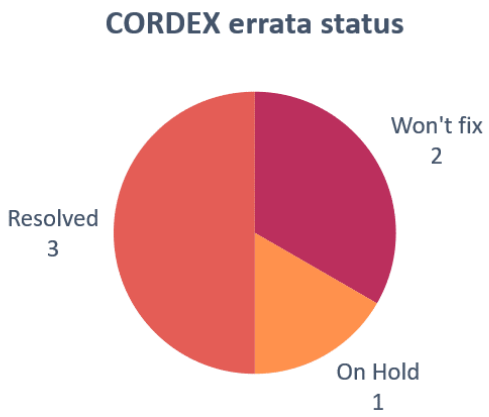
Breakdown of the errata entries:
The ES-DOC Errata service indexes 378 CMIP6 issues, 6 CORDEX issues and 1 input4MIPS issue. These issues are of different severities (low, medium, high, critical) and in different stages of the errata lifecycle (new, on hold, resolved or won't fix). The following pie charts show how these issues are split across these factors. The "won't fix" issues are a rather rare occurrence within the service and generally speaking describe low severity problems within the data itself.

[12] cf python package: https://ncas-cms.github.io/cf-python/index.html
[13] https://es-doc.github.io/esdoc-errata-client/









- **Number of documentation search and web site visits**

The ES-DOC model documentation service is designed for the modeling centers that need to provide their model documentation which is useful for a large range of users of model data. The service has two main objectives:

i) supporting the creators of model documentation and

ii) supporting the users of the documentation.

The creators are supported via the ES-DOC liaison at each CMIP6 modeling institute. This person (or persons) has been trained by ES-DOC and organizes the creation of documentation locally. Their primary contact with ES-DOC is via the support email (support@es-doc.org) and a dedicated ES-DOC-liaison mailing list.

Since July 2020, there have been 45 user support queries; 30 new institutional GitHub repositories have been created for the CORDEX project. In total, documentation for 36 models from 14 institutes, and documentation for the machines from 11 institutes have been published. Users of

these documents access them via the ES-DOC website, which is maintained by the ES-DOC team. Due to the enforced change of web service provider, the web service are unfortunately unavailable. However, what records were retained suggest a similar usage to the previous 18 months (21405).

- Number of questions to the helpdesk

The primary contact with ES-DOC is via the support e-mail address (support@es-doc.org), and a dedicated ES-DOC-liaison mailing list. Between July 2020 and December 2021, there have been 45 user support queries, compared with 80 queries in the previous 18 months.

- Metadata generated by the cdf2cim process of the ESGF publisher

The cdf2cim service that collects automatically collects CMIP6 simulation descriptions from CMIP6 datasets published to ESGF now has 2813991 simulation records, which are ready to be transformed into on-line documents, accessible via the further_info_URL service, during the first quarter of 2022.

5 Conclusions

This report focuses on the collection of performance indicators for the individual data services of the ENES CDI. Overall data distribution service usage shows a continued high user uptake with peaks related to individual and larger coordinated (e.g. IPCC related) data analysis activities. General trends are thus difficult to derive, yet some general observations can be concluded: The establishment of coordinated European CMIP data(-replica) pools contributed to a better coordination of analysis activities (which were also supported by the new VA and TA processing services offered) thus e.g. preventing the duplicate and uncoordinated download of data collections. This observation is backed up by the low number of distinct European users of European ESGF data nodes shown in Figure 1 (in comparison to non-EU users) which also decreased over time even though data analysis activities intensified (e.g. because of the IPCC deadlines in 2021). Many users were able to directly exploit the pre-collected data replicas in the pools and only very few needed to access data via the ESGF data services at European data nodes. An additional observation relates to the increased number of distinct ESGF users since March 2021 (see Figure 1, bottom chart), which probably relates to an increased downstream community usage of the CMIP6 data after IPCC working groups finalized their contributions for the IPCC report.