# A Climate Change Community Gateway for Data Usage & Data Archive Metrics across the Earth System Grid Federation

Sandro Fiore[1,*], Paola Nassisi[1,*], Alessandra Nuzzo[1,*], Maria Mirto[1,*], Luca Cinquini[2],
Dean Williams[3], Giovanni Aloisio[1,4]
[1]Euro-Mediterranean Center on Climate Change Foundation, Italy
[2]Jet Propulsion Laboratory/Caltech, USA
[3]Lawrence Livermore National Laboratory, Livermore, California, USA
[4]University of Salento, Italy

***Abstract*— The ESGF Dashboard is a key component of the Earth System Grid Federation (ESGF). It provides a distributed and scalable software infrastructure responsible for capturing a comprehensive set of data usage and data archive metrics both at the single site and federation level. The data usage information is related to the number of downloads and successful downloads and the number of distinct downloaded files, grouped by variable, model, experiment, etc. On the other hand, the data archive information is related to the total number of published datasets, total data volume and CMIP5 models and modelling institutes. All the above metrics relate to both cross and specific projects that are very notable in the climate community, such as, CMIP5, CMIP6, Obs4MIPs and CORDEX. From a Science Gateway perspective, the ESGF Dashboard presents the collected metrics through its Community Gateway (ESGF Dashboard User Interface).**

***Keywords*— *Earth System Grid Federation, Data Usage Metrics, Dashboard Community Gateway, CMIP experiments.***

## I. Introduction

The increased models' resolution in the development of comprehensive Earth System Models is rapidly leading to a very large climate simulations output that poses significant scientific data management challenges in terms of data sharing, processing, analysis, visualization, preservation, curation, and archiving [1][2][3].

In this domain, community efforts like the Coupled Model Intercomparison Projects (CMIP [4]) represent very challenging and relevant large-scale global experiments for climate change research.

The Coupled Model Intercomparison Project (CMIP) has been established by the Working Group on Coupled Modelling [5] (WGCM) under the World Climate Research Programme (WCRP). CMIP studies the output of coupled ocean-atmosphere general circulation models that also include interactive sea ice. These models allow the simulated climate to adjust to changes in climate forcing, such as increasing atmospheric carbon dioxide. CMIP began in 1995 by collecting output from model "control runs" in which climate forcing is held constant. Later versions of CMIP have collected output from an idealized scenario of global warming, with atmospheric CO2 increasing at the rate of 1% per year until it doubles at about Year 70.

The WCRP CMIP3 multi-model dataset archived at PCMDI, included realistic scenarios for both past and present climate forcing. The research based on this dataset has provided much of the new material underlying the IPCC 4th Assessment Report (AR4).

The WCRP CMIP5 experiment has provided the bases for the IPCC AR5. CMIP5 has promoted a standard set of model simulations in order to:

- evaluate how realistic the models are in simulating the recent past,
- provide projections of future climate change on two-time scales, near term (out to about 2035) and long term (out to 2100 and beyond), and
- understand some of the factors responsible for the differences in model projections, including quantifying some key feedback such as the instances involving clouds and the carbon cycle.

CMIP has led to the development of the Earth System Grid Federation (ESGF [6]). ESGF is one of the largest-ever collaborative data efforts in Earth system science that develops, deploys and maintains software to facilitate advancements in geophysical science. With its collection of independently funded national and international projects, ESGF manages the first ever decentralized database for accessing geophysical data at dozens of federated sites. ESGF involves a large set of data providers/modelling centers around the globe and includes the European contribution through the IS-ENES [7] project (by the European Network for Earth System Modelling (ENES [8]) community).

With respect to CMIP, it should be noted that:

[*] These authors have contributed equally to this work

- ESGF has been serving the Coupled Model Intercomparison Project Phase 5 (CMIP5 [9]) experiment, providing access to about 2PB of data produced around the globe by 26 institutes (groups) and 60 models, and
- ESGF is supporting the CMIP6 [10] experiments, which are expected to publish around 20PB of data (a 10X factor with respect to CMIP5).

From an infrastructural perspective, ESGF provides production-level support for search & discovery, browsing and secure access to climate simulation data and observational data products [6]. Still, computing capabilities are being added to the ESGF framework stack to enable server-side data analysis and to complement the data access functionalities mainly available in the current service offering.

Besides that, and in relation to this paper, ESGF also includes a software component named ESGF Dashboard, which provides support for federating, tracking, visualizing, and reporting data usage information. While initially the ESGF Dashboard was primarily meant to address monitoring challenges [6], its focus, over the last few years, has mainly moved towards a distributed and scalable software infrastructure responsible for collecting data usage and archive community metrics both at single site and federation level. This component provides coarse and fine grain information on how much, how frequently and how intensively the whole federation is being exploited by the end-users, by capturing the level of interest of the ESGF community on the available datasets. As such, the ESGF Dashboard provides a complete understanding about the amount of downloaded data, the most downloaded ones, the data published across the federation, etc. Still, geo-referenced metrics add an interesting client-side perspective to this multi-dimensional analysis. At the same time, the ESGF Dashboard gives feedback on the less-accessed datasets and variables, which can both help to design larger-scale future experiments and to get insights on the long tail of research. All the metrics reported above are related to both cross and specific projects that are very relevant in the climate community, such as, among the others: CMIP5 [9], Obs4MIPs [11], CORDEX [12,13], and CMIP6 [10].

From a Science Gateway perspective, the ESGF Dashboard presents the collected metrics through a rich set of attractive widgets (i.e. charts, maps and reports) available via its brand new (w.r.t. [6]) Community Gateway (ESGF Dashboard User Interface) [14]. It allows end-users (i.e. climate scientists) to visualize the data usage and data archive metrics offering a very different perspective (more user oriented) about the scientific experiments data exploitation.

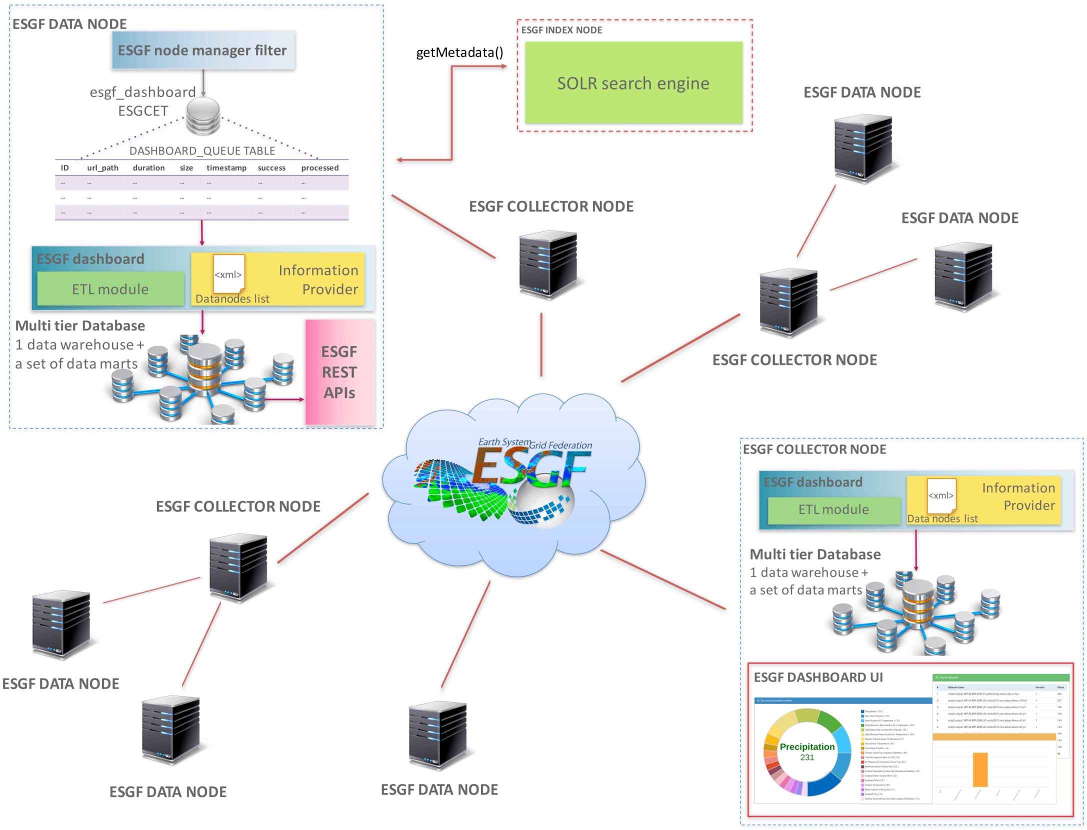


**Figure 1. ESGF Dashboard Architecture. The picture shows also in the small architectural details about the ESGF Data Node and the Collector node.**

The rest of the paper is organized as follows. Section 2 presents the ESGF Dashboard architecture in terms of requirements and architectural design, whereas Section 3 describes in detail its Community Gateway providing a comprehensive description of all the supported views. Section 4 provides some of the most relevant achievements in terms of metrics that can be easily inferred from the Community Gateway. Finally, Section 5 draws the conclusions and hints the future work.

## II. ESGF Dashboard Architecture

The ESGF architecture consists of a set of services, which are logically grouped into four types of nodes: *data node* (providing access to data), *index* (supporting indexing and searching of datasets), *identity provider* (supporting federated user authentication) and *compute* (providing data analysis capabilities). Related to this work is the *data node* type, which is a collection of open source components providing basic data access functionality via HTTP/OPeNDAP [15] services associated with metadata catalogues (THREDDS [16]). Its main components are the data Publisher application that generates the metadata catalogs, the THREDDS and GridFTP [17] servers as well as the ESGF Dashboard.

This section dives into the details of the ESGF Dashboard design, highlighting the requirements (both functional and non-functional) and the architecture. The next two sub-sections specifically address these two aspects.

### A. Requirements analysis

The ESGF Dashboard has been designed by considering a set of functional and non-functional requirements mainly gathered from the ESGF/CMIP community.

More specifically, with regard to the *functional* requirements, the ESFG Dashboard has to provide:

(i) data download statistics (both per node, per project, institution-based and federated-level view) provided according to several analysis dimensions or a combination of two or three of them;

(ii) client statistics (grouped by country and/or continent and also over time) related to all of the clients that carried out at least one download from the ESGF data nodes;

(iii) status of the federation in terms of data volume and number of published datasets at project and global level.

Additionally, the most relevant *non-functional* requirements are related to:

(i) tolerate unpredictable or invalid input (*robustness*);

(ii) hide the back-end complexity (*transparency*);

(iii) properly scale with regard to an increasing number of metrics and data nodes (*scalability*);

(iv) efficiently manage and store the large set of data usage statistics (*efficiency*);

(v) provide a security layer able to address both authentication (in strong synergy with the authentication mechanisms currently available in the IS-ENES/ESGF federation) and authorization (in terms of previously defined classes of users) (*security*);

(vi) be easily extensible with new metrics and interfaces, based on new user/system requirements, and the new elements should be straightforwardly added to the system with few code changes/additions (*extensibility* and *reusability*);

(vii) be highly configurable and flexible to allow site administrators to carry out site-specific configurations, preserving local autonomy & federation-level needs (*configurability/flexibility*);

(viii) provide a user interface aimed at computational scientists as well as domain-based experts (climate change scientists) (*usability*);

(ix) provide a complete set of APIs to give users the opportunity to programmatically access the dashboard metrics (*programmability*)

(x) be "zero-conf" (w.r.t. the list of nodes to be monitored, geo-location information, list of available services, deployment information, etc.) to allow the node administrator to test and use the dashboard right after the installation without any intermediate setup/configuration steps (*zero-conf*).

### B. Architectural design

This sub-section describes the ESGF Dashboard architectural design (see Figure 1), as a result of the requirement elicitation phase mentioned in the previous sub-section. In particular, two key components are relevant in the proposed design for, respectively, (i) metrics collection (data nodes level) and (ii) metrics aggregation across the federation (collectors level).

#### 1) ESGF Data nodes

At the data nodes level, the ESGF Dashboard is responsible for processing every new download entry occurring at the site, as well as inferring and storing the whole set of associated metadata into several multi-dimensional databases (data marts) running in the Dashboard back-end. To do that, the Dashboard queries the proper ESGF index node and retrieves the full metadata description related to the downloaded file. The metrics stored in the data marts are available to any application or service via the Dashboard REST API.

#### 2) ESGF Dashboard Collector

To gather all the metrics across the federation, the ESGF Dashboard relies on a collector node, which exploits a lazy hierarchical pull protocol based on leaves and collector nodes. The former are the data nodes, whereas the latter are intermediate nodes, which hierarchically aggregate metrics. The highest node in the hierarchy (top collector) aggregates the whole set of metrics across the federation. The collector protocol exploits the REST API provided by the Dashboard to expose the metrics. Each collector manages a local set of data marts (with basically the same structure of the leaves nodes) and it provides an aggregated view of the metrics related to the nodes running at the underlying level. All the collectors (at the different levels of the hierarchy) expose the metrics stored in their data marts via the Collector REST API. The top collector exposes its federation-level back-end database of metrics to the ESGF Dashboard User Interface, the community gateway providing user-friendly access to the federated metrics (see next Section).

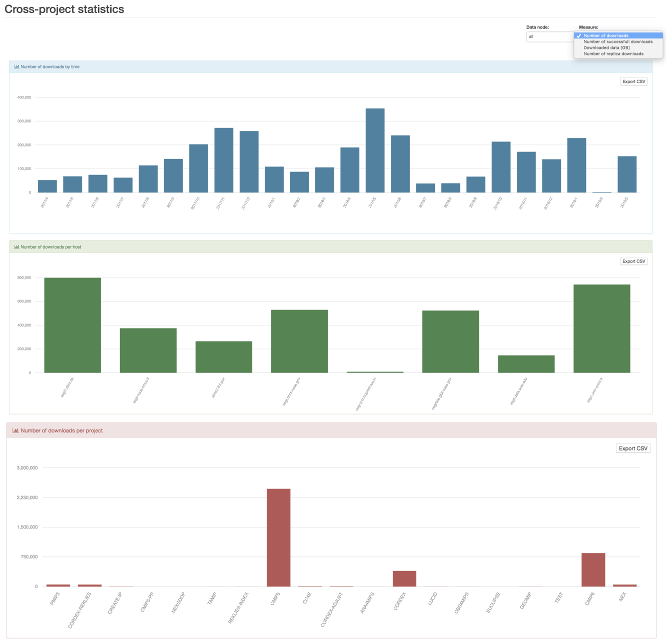
## III. ESGF Dashboard Community Gateway

This section presents in detail the ESGF Dashboard User Interface, which is the central hub of the ESGF Dashboard, providing user-friendly access to a comprehensive set of data usage and data archive metrics across the whole ESGF. From a software stack point of view the ESGF Dashboard UI is built on top of the following technologies: Java 8 and Spring 5 MVC, Bootstrap template, jQuery (UI objects), Morris.js (for graphs), Google APIs (for Maps) and the PostgreSQL RDBMS (its database). The Dashboard User Interface is open to all users and provides an easy-to-use and highly interactive interface.

In the following sub-sections, the main views provided by this gateway are presented and discussed in detail.

### A. *Cross-project metrics*

This view provides information about the data downloads (number of downloads, size, number of successful downloads) over three different dimensions: time, data node and project.



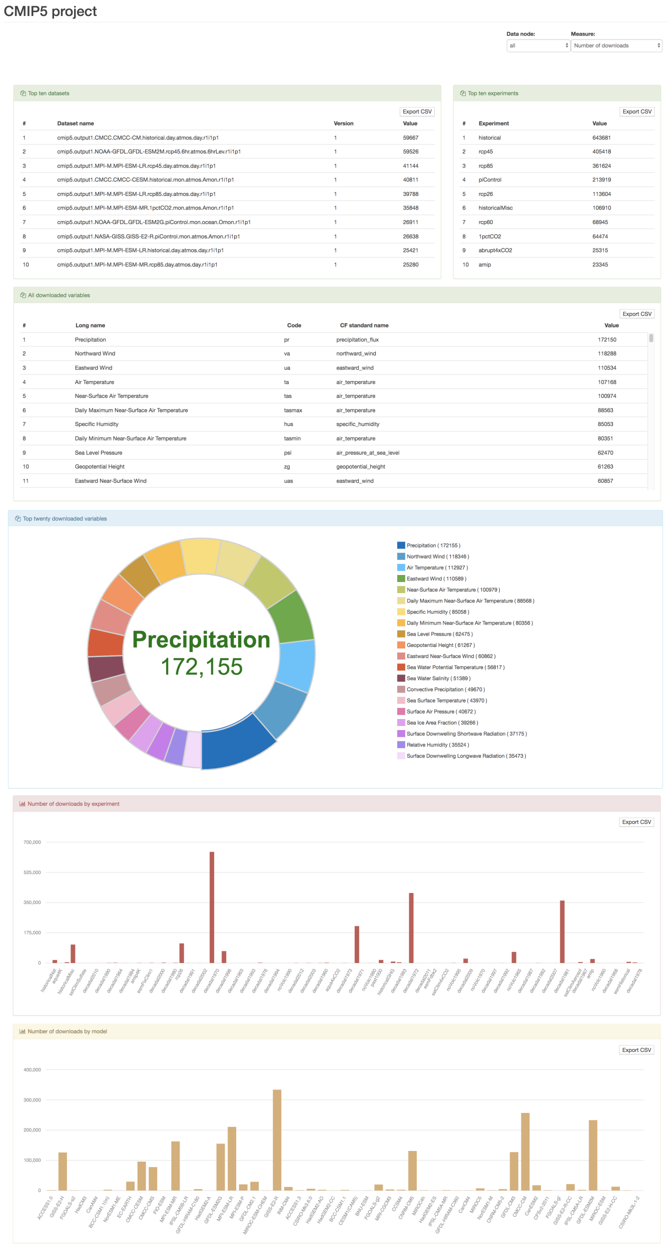
Cross-project statistics

**Figure 2. Cross-project view. This view spans across all the registered projects and provides information of key data download metrics over three different dimensions: time, data node and project.**

The user can choose one of the metrics and visualize it as bar charts from different points of view (See Figure 2). The analysis/reporting can be limited to a single host as well as performed on the entire federation. Charts are dynamic, light and very suitable for analysis and reporting. The metrics can also be downloaded as CSV file for further re-use or analysis outside the gateway.

### B. *Project-specific view*

This view provides in-depth information and details about a specific project. As an example, Figure 3 shows a CMIP5 project-specific view. Similarly, to the previous case, the user can choose one of the following metrics: number of downloads, data size, and number of successful downloads.



CMIP5 project

Data node: all
Measure: Number of downloads

Top ten datasets

| # | Dataset name | Version | Value |
|---|---|---|---|
| 1 | cmip5.output1.CMCC.CMCC-CM.historical.day.atmos.day.r1i1p1 | 1 | 59667 |
| 2 | cmip5.output1.NOAA-GFDL.GFDL-ESM2M.rcp45.6hr.atmos.6hrLev.r1i1p1 | 1 | 59526 |
| 3 | cmip5.output1.MPI-M.MPI-ESM-LR.rcp45.day.atmos.day.r1i1p1 | 1 | 41144 |
| 4 | cmip5.output1.CMCC.CMCC-CESM.historical.mon.atmos.Amon.r1i1p1 | 1 | 40811 |
| 5 | cmip5.output1.MPI-M.MPI-ESM-LR.rcp85.day.atmos.day.r1i1p1 | 1 | 39788 |
| 6 | cmip5.output1.MPI-M.MPI-ESM-MR.1pctCO2.mon.atmos.Amon.r1i1p1 | 1 | 35848 |
| 7 | cmip5.output1.NOAA-GFDL.GFDL-ESM2G.piControl.mon.ocean.Omon.r1i1p1 | 1 | 26911 |
| 8 | cmip5.output1.NASA-GISS.GISS-E2-R.piControl.mon.atmos.Amon.r1i1p1 | 1 | 26638 |
| 9 | cmip5.output1.MPI-M.MPI-ESM-LR.historical.day.atmos.day.r1i1p1 | 1 | 25421 |
| 10 | cmip5.output1.MPI-M.MPI-ESM-MR.rcp85.day.atmos.day.r1i1p1 | 1 | 25280 |

Top ten experiments

| # | Experiment | Value |
|---|---|---|
| 1 | historical | 643681 |
| 2 | rcp45 | 405418 |
| 3 | rcp85 | 361624 |
| 4 | piControl | 213919 |
| 5 | rcp26 | 113604 |
| 6 | historicalMisc | 106910 |
| 7 | rcp60 | 68945 |
| 8 | 1pctCO2 | 64474 |
| 9 | abrupt4xCO2 | 25315 |
| 10 | amip | 23345 |

All downloaded variables

| # | Long name | Code | CF standard name | Value |
|---|---|---|---|---|
| 1 | Precipitation | pr | precipitation_flux | 172150 |
| 2 | Northward Wind | va | northward_wind | 118288 |
| 3 | Eastward Wind | ua | eastward_wind | 110534 |
| 4 | Air Temperature | ta | air_temperature | 107168 |
| 5 | Near-Surface Air Temperature | tas | air_temperature | 100974 |
| 6 | Daily Maximum Near-Surface Air Temperature | tasmax | air_temperature | 88563 |
| 7 | Specific Humidity | hus | specific_humidity | 85053 |
| 8 | Daily Minimum Near-Surface Air Temperature | tasmin | air_temperature | 80351 |
| 9 | Sea Level Pressure | psl | air_pressure_at_sea_level | 62470 |
| 10 | Geopotential Height | zg | geopotential_height | 61263 |
| 11 | Eastward Near-Surface Wind | uas | eastward_wind | 60857 |

**Figure 3. Project-specific view for the CMIP5 project. This view provides a comprehensive set of key metrics like top ten variable, dataset, top ten datasets, downloads by model, by experiment and the full list of variables.**

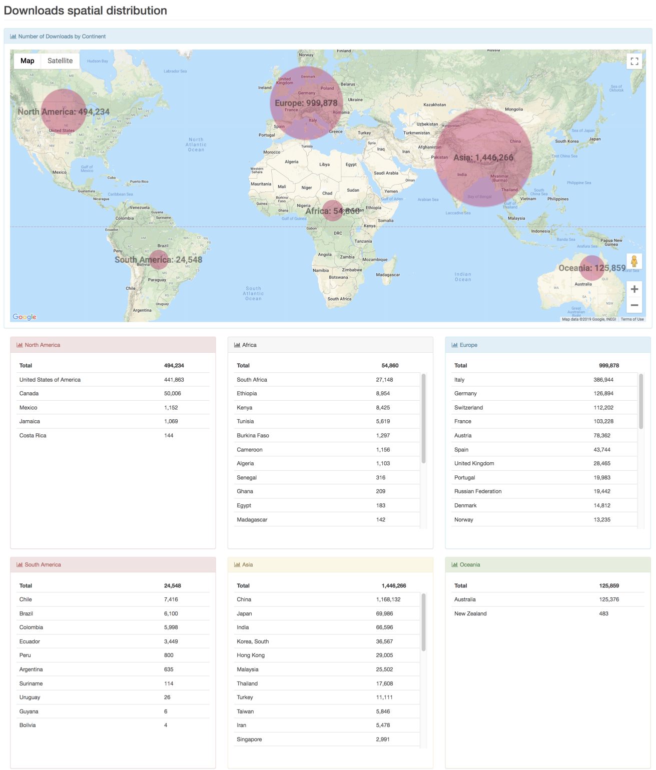
Based on these metrics, this view displays the different charts/tables: (i) top ten datasets; (ii) top ten experiments; (iii) all downloaded variables; (iv) top twenty downloaded variables; (v) number of downloads by experiments; (vi) number of downloads by models.

Besides displaying data aggregated across the whole federation, this view also allows a more selective display of the same metrics for a specific host, thus limiting the reporting/analysis to a single node.

As it can be easily argued, such articulated and complementary views allow a strong and deep understanding of each metrics as they are examined at the same time from different perspectives. Even in this case, the metrics can be downloaded as CSV file for further re-use or analysis outside the gateway.

### C. *Downloads spatial distribution*

The geo-downloads section (Figure 4) aims to display on a map information related to the downloads distribution (size and number of files) per continent as well as additional metrics at a finer level (grouped by countries) on different tables.



North America

| | |
|---|---|
| **Total** | **494,234** |
| United States of America | 441,863 |
| Canada | 50,006 |
| Mexico | 1,152 |
| Jamaica | 1,069 |
| Costa Rica | 144 |

Africa

| | |
|---|---|
| **Total** | **54,860** |
| South Africa | 27,148 |
| Ethiopia | 8,954 |
| Kenya | 8,425 |
| Tunisia | 5,619 |
| Burkina Faso | 1,297 |
| Cameroon | 1,156 |
| Algeria | 1,103 |
| Senegal | 316 |
| Ghana | 209 |
| Egypt | 183 |
| Madagascar | 142 |

Europe

| | |
|---|---|
| **Total** | **999,878** |
| Italy | 386,944 |
| Germany | 126,894 |
| Switzerland | 112,202 |
| France | 103,228 |
| Austria | 78,362 |
| Spain | 43,744 |
| United Kingdom | 28,465 |
| Portugal | 19,983 |
| Russian Federation | 19,442 |
| Denmark | 14,812 |
| Norway | 13,235 |

South America

| | |
|---|---|
| **Total** | **24,548** |
| Chile | 7,416 |
| Brazil | 6,100 |
| Colombia | 5,998 |
| Ecuador | 3,449 |
| Peru | 800 |
| Argentina | 635 |
| Suriname | 114 |
| Uruguay | 26 |
| Guyana | 6 |
| Bolivia | 4 |

Asia

| | |
|---|---|
| **Total** | **1,446,266** |
| China | 1,168,132 |
| Japan | 69,986 |
| India | 66,596 |
| Korea, South | 36,567 |
| Hong Kong | 29,005 |
| Malaysia | 25,502 |
| Thailand | 17,608 |
| Turkey | 11,111 |
| Taiwan | 5,846 |
| Iran | 5,478 |
| Singapore | 2,991 |

Oceania

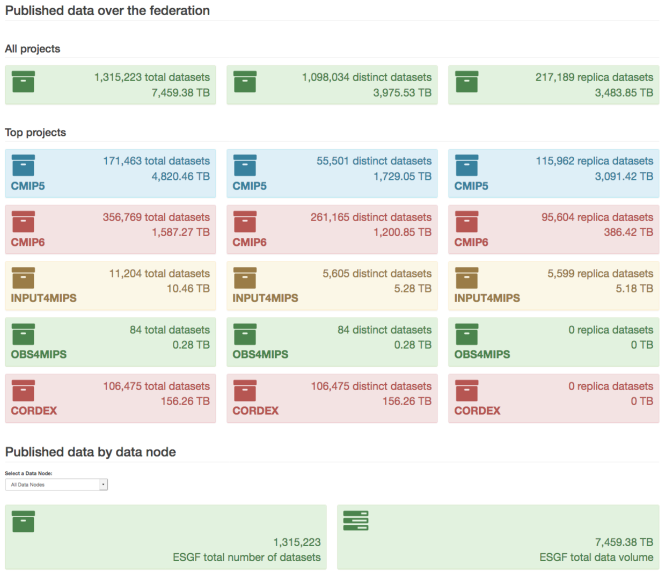
| | |
|---|---|
| **Total** | **125,859** |
| Australia | 125,376 |
| New Zealand | 483 |

**Figure 4. Downloads spatial distribution. This view provides geo-referenced information on data downloads from a client-side perspective. Metrics are gathered by continent and by country.**

This view is particularly interesting because it gives a geo-referenced client-side perspective of the data usage. Of course, sensitive information is not provided being filtered out at the level of each single data node (leaves). Moreover, the country level of details can reveal how much a specific country is actually involved using/exploiting the CMIP data and in general the ESGF infrastructure. Drilling down more than the country level (e.g. regional) is not in the current roadmap.

### D. *Published data over the entire federation*

This section (see Figure 5) provides a summary view of the total amount of data published on the ESGF federated archive in terms of total number of datasets as well as distinct and replica datasets with the related total data volume (in TB); moreover, it also displays the number of datasets, along with distinct and replica datasets and their related data volume for CMIP5, CMIP6, INPUT4MIPs, Obs4MIPs and CORDEX projects. Also, it is possible to select a specific data node and obtain information on the total number of published dataset and the total amount of data for that node. The initial list of this view was set with the 5 projects mentioned before, being them very relevant to a large number of users. This view is meant to be extended over time with additional projects as soon as new ones raise a significant interest from the community.



Published data over the federation

All projects

| | | |
|---|---|---|
| 1,315,223 total datasets<br>7,459.38 TB | 1,098,034 distinct datasets<br>3,975.53 TB | 217,189 replica datasets<br>3,483.85 TB |

Top projects

| | | |
|---|---|---|
| 171,463 total datasets<br>4,820.46 TB<br>CMIP5 | 55,501 distinct datasets<br>1,729.05 TB<br>CMIP5 | 115,962 replica datasets<br>3,091.42 TB<br>CMIP5 |
| 356,769 total datasets<br>1,587.27 TB<br>CMIP6 | 261,165 distinct datasets<br>1,200.85 TB<br>CMIP6 | 95,604 replica datasets<br>386.42 TB<br>CMIP6 |
| 11,204 total datasets<br>10.46 TB<br>INPUT4MIPS | 5,605 distinct datasets<br>5.28 TB<br>INPUT4MIPS | 5,599 replica datasets<br>5.18 TB<br>INPUT4MIPS |
| 84 total datasets<br>0.28 TB<br>OBS4MIPS | 84 distinct datasets<br>0.28 TB<br>OBS4MIPS | 0 replica datasets<br>0 TB<br>OBS4MIPS |
| 106,475 total datasets<br>156.26 TB<br>CORDEX | 106,475 distinct datasets<br>156.26 TB<br>CORDEX | 0 replica datasets<br>0 TB<br>CORDEX |

Published data by data node

Select a Data Node: All Data Nodes

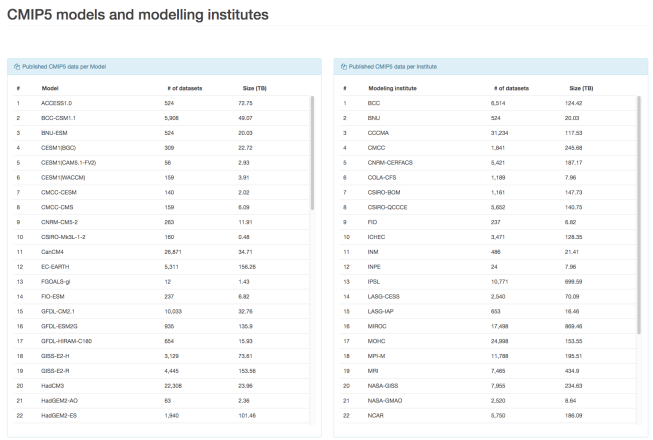
| | |
|---|---|
| 1,315,223<br>ESGF total number of datasets | 7,459.38 TB<br>ESGF total data volume |

**Figure 5. Published data across the entire ESGF federation. This view provides information about published data for some community projects. Info about distinct and replica datasets is also provided.**

### E. *Federated data archive*

The Federated Statistics section (Figure 6) provides specific views on the CMIP5 experiment across the entire federation.

In particular, it addresses a specific requirement, from the CMIP community, related to the CMIP5 data available in the ESGF federated data archive rather to its data usage; more precisely, it provides a very interesting global view of the CMIP5 published data (this information is indeed captured from the ESGF index nodes). Two tables provide respectively a model-based and an institute-based view for the CMIP5 project. As such, end users can easily infer the total number of models (and modelling institutions) and for each model (as well as institution) metrics about the size and number of datasets published across the entire federation.



CMIP5 models and modelling institutes

Published CMIP5 data per Model

| # | Model | # of datasets | Size (TB) |
|---|---|---|---|
| 1 | ACCESS1.0 | 524 | 72.75 |
| 2 | BCC-CSM1.1 | 5,908 | 49.07 |
| 3 | BNU-ESM | 524 | 20.03 |
| 4 | CESM1(BGC) | 309 | 22.72 |
| 5 | CESM1(CAM5.1-FV2) | 56 | 2.93 |
| 6 | CESM1(WACCM) | 159 | 3.91 |
| 7 | CMCC-CESM | 140 | 2.02 |
| 8 | CMCC-CMS | 159 | 6.09 |
| 9 | CNRM-CM5-2 | 263 | 11.91 |
| 10 | CSIRO-Mk3L-1-2 | 160 | 0.48 |
| 11 | CanCM4 | 26,871 | 34.71 |
| 12 | EC-EARTH | 5,311 | 156.26 |
| 13 | FGOALS-gl | 12 | 1.43 |
| 14 | FIO-ESM | 237 | 6.82 |
| 15 | GFDL-CM2.1 | 10,033 | 32.76 |
| 16 | GFDL-ESM2G | 935 | 135.9 |
| 17 | GFDL-HIRAM-C180 | 654 | 15.93 |
| 18 | GISS-E2-H | 3,129 | 73.61 |
| 19 | GISS-E2-R | 4,445 | 153.56 |
| 20 | HadCM3 | 22,308 | 23.96 |
| 21 | HadGEM2-AO | 63 | 2.36 |
| 22 | HadGEM2-ES | 1,940 | 101.46 |

Published CMIP5 data per Institute

| # | Modeling institute | # of datasets | Size (TB) |
|---|---|---|---|
| 1 | BCC | 6,514 | 124.42 |
| 2 | BNU | 524 | 20.03 |
| 3 | CCCMA | 31,234 | 117.53 |
| 4 | CMCC | 1,841 | 245.68 |
| 5 | CNRM-CERFACS | 5,421 | 187.17 |
| 6 | COLA-CFS | 1,189 | 7.96 |
| 7 | CSIRO-BOM | 1,161 | 147.73 |
| 8 | CSIRO-QCCCE | 5,652 | 140.75 |
| 9 | FIO | 237 | 6.82 |
| 10 | ICHEC | 3,471 | 128.35 |
| 11 | INM | 486 | 21.41 |
| 12 | INPE | 24 | 7.96 |
| 13 | IPSL | 10,771 | 699.59 |
| 14 | LASG-CESS | 2,540 | 70.09 |
| 15 | LASG-IAP | 653 | 16.46 |
| 16 | MIROC | 17,498 | 869.46 |
| 17 | MOHC | 24,998 | 153.55 |
| 18 | MPI-M | 11,788 | 195.51 |
| 19 | MRI | 7,465 | 434.9 |
| 20 | NASA-GISS | 7,955 | 234.63 |
| 21 | NASA-GMAO | 2,520 | 8.64 |
| 22 | NCAR | 5,750 | 186.09 |

**Figure 6. Federated data archive. This view relates to CMIP5 published data only and provides metrics from models and institutions perspective.**

## IV. Insights

Thanks to the ESGF Dashboard Community Gateway some insights in terms of metrics targets can be straightforwardly inferred and visualized through the available web interface. Here is some of them (captured at the time the paper is being written), which in some cases can be considered as *milestones* for the community:

- Published data: overall 1,316,528 datasets and 7.3PB in total. Around 300K datasets are replicas.
- 1PB published datasets for CMIP6 reached in March 2019, with about 260K (distinct) datasets.
- Datasets from 61 models and 30 different institutions have been published for CMIP5.
- The CMIP5 most downloaded variable is the *precipitation* (184,212 downloads) followed by *northward wind* and *near-surface air temperature*.
- From a geo-downloads point of view the number of downloads from Asia is currently three times the one from North America and 1,5 times the one from Europe.

## V. Conclusions and Future Work

This paper presents the ESGF Dashboard (a key component of the Earth System Grid Federation), which provides a distributed and scalable software infrastructure responsible for capturing a set of data usage and data archive metrics both at the single site and federation level. The architectural details as well as an in-depth view of the ESGF-Dashboard Community Gateway are comprehensively discussed. This work highlights the relevance of the data usage metrics, the complexity of gathering them across the whole ESGF federation from a distributed systems standpoint as well as the quantitative and visualization aspects related to them from a Science Gateway perspective. Some insights in terms of targets for a few relevant metrics are also presented. Future work will address new requirements from the ESGF community such as for examples, new metrics related to the novel data analysis and compute services, expected to be included in the ESGF infrastructure by the end of 2019. This work is currently ongoing in the context of the ESGF Compute Working Team and will deliver preliminary results during the CMIP6 timeframe.

## Acknowledgment

This work was supported by the EU FP7 Infrastructure for the European Network for Earth System modelling - Phase 2 project (IS-ENES2, Grant Agreement 312979) and it is currently supported by the EU H2020 IS-ENES Phase 3 (IS-ENES3, Grant Agreement 824084).

## References

[1] J. Dongarra, P. Beckman et al., "The international exascale software project roadmap," Int. J. High Perform. Comput. Appl., vol. 25, no. 1, pp. 3–60, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1177/1094342010391989

[2] PRACE - the scientific case for high performance computing in europe 2012-2020. PRACE. [Online]. Available: http://www.prace-ri.eu/IMG/pdf/prace - the scientific case - full text -.pdf

[3] G. Aloisio and S. Fiore, "Towards exascale distributed data management," Int. J. High Perform. Comput. Appl., vol. 23, no. 4, pp. 398–400, Nov. 2009. http://dx.doi.org/10.1177/1094342009347702

[4] WCRP Coupled Model Intercomparison Project (CMIP). [Online]. Available: https://www.wcrp-climate.org/wgcm-cmip

[5] Working Group on Coupled Modelling. [Online]. Available: https://www.wcrp-climate.org/wgcm-overview

[6] L. Cinquini, et al., "The earth system grid federation: An open infrastructure for access to distributed geospatial data," Future Generation Computer Systems, vol. 36, pp. 400 – 417, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X13001477

[7] EU IS-ENES Project (Infrastructure for the European Network for Earth System modelling). IS-ENES Phase2 final report available online at: https://portal.enes.org/ISENES2/documents/contractual-documents/is-enes2-final-report

[8] S. Joussaume, B. Lawrence and F. Guglielmo, Update of the ENES infrastructure strategy 2012-2022, ENES Report Series 2, 2017, 20 pp.

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

[10] Balaji, et al.: 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.

[11] Teixeira, J., D. Waliser, R. Ferraro, P. Gleckler, T. Lee, and G. Potter, 2014: Satellite observations for CMIP5: The genesis of Obs4MIPs. Bull. Amer. Meteor. Soc., 95, 1329–1334, doi:10.1175/BAMS-D-12-00204.1.

[12] Coordinated Regional Climate Downscaling Experiment (CORDEX). Available online at: http://www.cordex.org/

[13] F. Giorgi & W. J. Gutowski. (2015). Regional dynamical downscaling and the CORDEX Initiative. Annual Review of Environment and Resources, 40, pp. 467–490. https://doi.org/10.1146/annurev-environ-102014-021217

[14] ESGF Dashboard Community Gateway. Available online at: http://esgf-ui.cmcc.it:8080/esgf-dashboard-ui/

[15] Cornillon, P., Gallagher, J. and Sgouros, T., 2003. OPeNDAP: Accessing data in a distributed, heterogeneous environment. Data Science Journal, 2, pp.164–174. DOI: http://doi.org/10.2481/dsj.2.164.

[16] Unidata. THREDDS Data Server (TDS) [software]. Boulder, CO: UCAR/Unidata. (http://doi.org/10.5065/D6N014KG).

[17] W. Allcock, et al. The Globus Striped GridFTP Framework and Server. Proceedings of Super Computing 2005 (SC05), November 2005.