# The ESGF Virtual Aggregation (CMIP6 v20240125)

Ezequiel Cimadevilla[1], Bryan N. Lawrence[2], and Antonio S. Cofiño[1]

[1]Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, Spain
[2]National Centre for Atmospheric Science, Department of Meteorology, University of Reading, Reading, UK

**Correspondence:** Ezequiel Cimadevilla (ezequiel.cimadevilla@unican.es)

**Abstract.**

The Earth System Grid Federation (ESGF) holds several petabytes of climate data distributed across millions of files held in data centers worldwide. Obtaining and manipulating the scientific information (climate variables) held in these files is non-trivial. The ESGF Virtual Aggregation is one of several solutions to providing an out-of-the-box aggregated and analysis ready view of those variables. Here we discuss the ESGF Virtual Aggregation in the context of the existing infrastructure, and some of those other solutions providing analysis ready data. We describe how it is constructed, how it can be used, its benefits for model evaluation data analysis tasks, and provide some performance evaluation. It will be seen that the ESGF Virtual Aggregation provides a sustainable solution to some of the problems encountered in producing analysis ready data, without the cost of data replication to different formats, albeit at the cost of more data movement within the analysis than some alternatives. If heavily used, it may also require more ESGF data servers than are currently deployed in data node deployments. The need for such data servers should be a component of ongoing discussions about the future of the ESGF and its constituent core services.

## 1 Introduction

The importance of effective and efficient climate data analysis continues to grow as the demand for understanding the climate system intensifies. Traditionally, climate data repositories have been structured as file distribution systems, primarily facilitating file downloads. However, this conventional approach poses challenges for climate data analysts, requiring them to invest substantial time in managing data access, often unrelated to their ongoing research. The Earth System Grid Federation (ESGF) is a global infrastructure and network that consists of internationally distributed research centers that follows this approach (Williams et al., 2016; Cinquini et al., 2012). While the ESGF provides a critical platform for data sharing, its current architecture lacks integrated tools for advanced data analysis. Thus, researchers must handle data access and analysis independently.

To address this limitation, current research focuses on enhancing climate data infrastructures with built-in data analysis capabilities that streamline data access and processing. Several methodologies are emerging based on Analysis Ready Data (ARD, Dwyer et al. 2018), remote data access and new formats for climate data storage (Abernathey et al., 2021). This paper introduces the ESGF Virtual Aggregation (ESGF-VA), an innovative method for climate data analysis leveraging rarely exploited aspects of the ESGF. It is based on the capabilities of virtual aggregations built on top of the ESGF architecture and designed to be included in the federation as an external service. The ESGF-VA enables scientists to perform efficient, scalable, and remote climate data analysis within the ESGF. Section 2 provides an overview of the current landscape of climate

data analysis and infrastructure. Section 3 introduces the notion of ARD and virtual aggregations in the context of climate data. Section 4 describes a model evaluation use case and the benefits provided by ESGF-VA for the task. Following this, Section 5 delineates the methodology employed in the ESGF-VA. Section 6 presents a performance evaluation of the ESGF-VA comparing it to other data access methods. Section 7 ends with a discussion and concluding remarks.

## 2 Background

In the ESGF, research centers collectively serve as a federated data archive, supporting the distribution of global climate model simulations representing past, present, and future climate conditions (Balaji et al., 2018). The ESGF enables modeling groups to upload model output to federation nodes for archiving and community access at any time. To facilitate multi-model analyses, the ESGF ensures standardization of model output in a specified format. It also facilitates the collection, archival, and access of model output through the ESGF data replication centers. As a result, the ESGF has emerged as the primary distributed data archive for climate data, hosting data for international projects such as CMIP6 (Eyring et al., 2016) and CORDEX (Gutowski Jr. et al., 2016). It catalogues and stores tenths of millions of files, with more than 30 petabytes of data, distributed across research institutes worldwide (Fiore et al., 2021), and it serves as the reference archive for Assessment Reports (AR, Asadnabizadeh 2023) on Climate Change produced by the Intergovernmental Panel on Climate Change (IPCC, Venturini et al. 2023).

The significant growth of data poses a scientific scalability challenge for the climate research community (Balaji et al., 2018). Contributions to the increase in data volume include the systematic increase in model resolution and the complexity of experimental protocols and data requests (Juckes et al., 2020). While these advancements enrich climate modeling and analysis, they also exacerbate difficulties in accessing and processing the resulting large datasets. Currently, the primary method of data acquisition involves downloading files directly from repositories. However, as the number and size of files continue to grow, this approach becomes increasingly impractical, creating bottlenecks that make data analysis inefficient. The ESGF infrastructure is designed as a file distribution system, but scientific research often requires multidimensional data analysis on datasets encompassing multiple variables, spanning the entire time period, multiple model ensembles and different climate model runs. Several ongoing developments in scientific data research try to address the issues of growing data volume and variety and provide new approaches to data analysis.

Climate Analytics-as-a-Service (CAaaS, Schnase et al. 2016), GeoDataCubes (Nativi et al., 2017; Mahecha et al., 2020), cloud native data repositories (Abernathey et al., 2021) and Web Processing Services (WPS, 2015) are some of the systems that are being used to improve climate data analysis workflows. The data consolidation process in building these new systems may involve data duplication of an enormous volume of data, incurring in large costs of operational and storage requirements. However, the cost of data duplication is assumed to be compensated by a gain in efficiency in information synthesis. In order to overcome these costs, several technologies do allow the creation of virtual datasets, which provide ARD capabilities without the need to duplicate the original data sources. These provide the opportunity for more sustainable approaches to enhancing climate data analysis capabilities. Figure 1 illustrates the outcome of a data analysis task that integrates multiple files from the

ESGF, encompassing several model runs of a specific model spanning 85 years of data. By leveraging the advantages of ARD and remote data access, this task can be executed without the need for file downloads, requiring only a few lines of code.

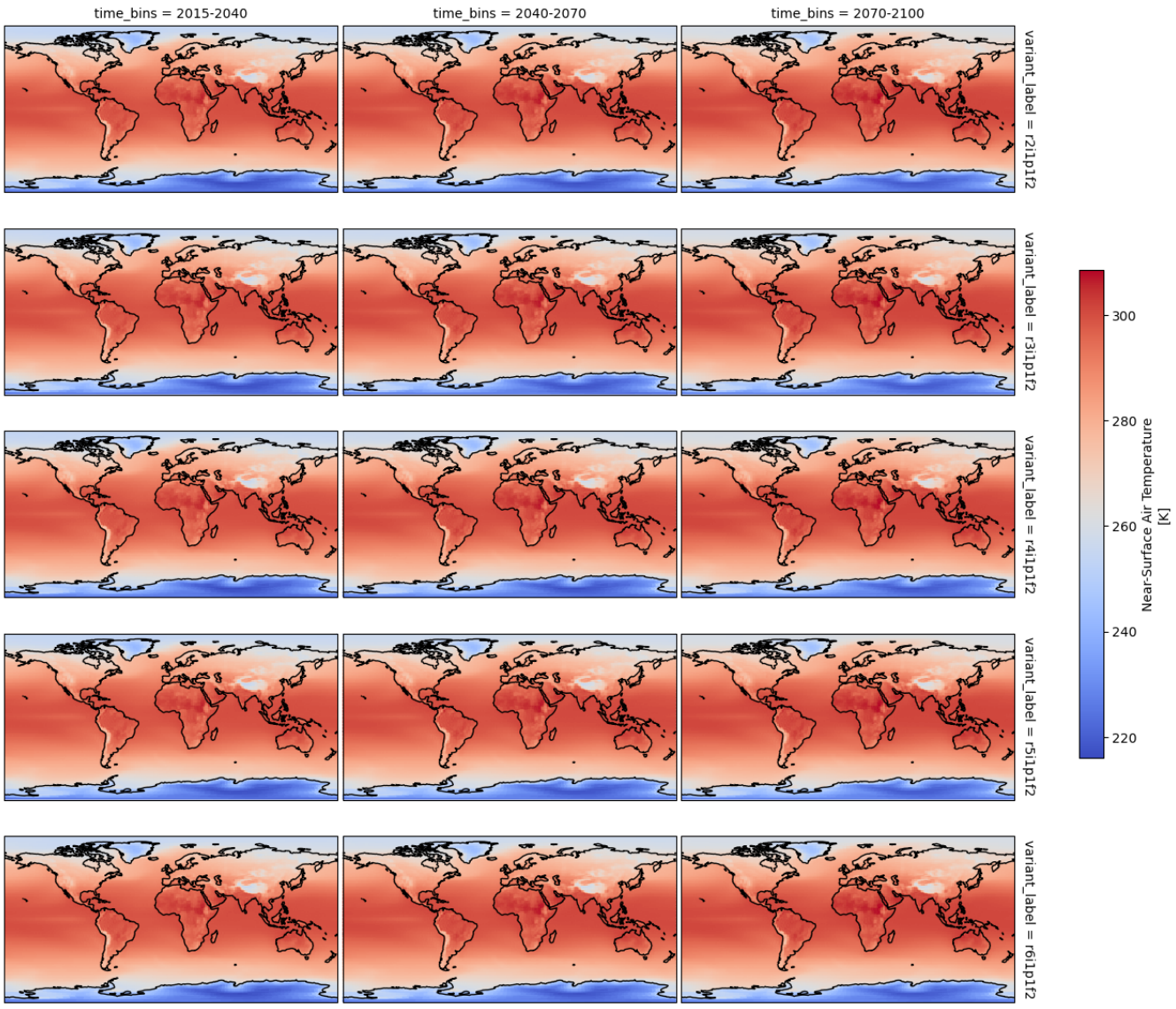


**Figure 1.** Mean Near-Surface Air Temperature for different time periods and different model runs. The code needed to obtain this result is minimal, enabled by the capabilities of the data cube. Because all the information is stored in one single ESGF Virtual Aggregation dataset, only one data source is needed to perform the data analysis. The data is fetched directly from ESGF data nodes on a remote data access basis.

The ESGF-VA serves as a bridge between the current implementation of the ESGF and the development of cloud native data repositories for climate research. Figure 2 shows how it fits into the current ecosystem. It is implemented as an additional

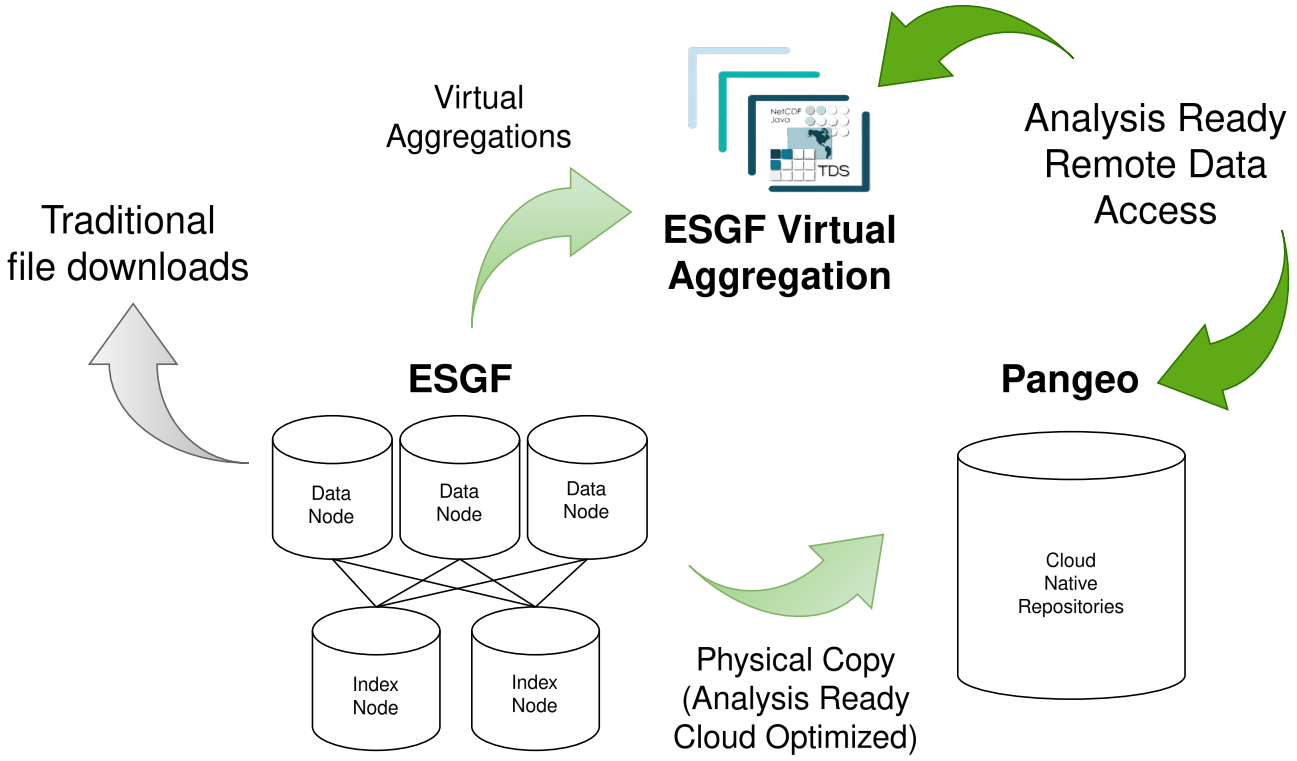


**Figure 2.** The ESGF Virtual Aggregation aims to be a sustainable bridge that eases the technological transition between the current state of the ESGF and more ground-breaking and expensive solutions, based on data replication, such as cloud native repositories.

value-added user service on top of ESGF, running in conjunction with other value-added user services such as the citation and PID handle services (Petrie et al., 2021). To satisfy sustainability requirements, a balanced strategy is adopted to manage operational costs and complexity. The ESGF-VA is aimed to advance the sharing and reuse of scientific climate data by building a catalog of logically aggregated datasets, facilitating remote access to the distributed data hosted in the ESGF. It offers data access (remotely) to convenient and adequate views of the data (ARD) that allow ad hoc complex queries without the need to duplicate data sources.

## 3 Analysis Ready Climate Datasets

ARD refers to datasets that have undergone processing to enable analysis with minimal additional user effort (Dwyer et al., 2018). The climate data offered by the ESGF is stored in netCDF files (Rew et al., 1989), with an atomic dataset defined as a set of netCDF files that are aggregated, containing the data from a single climate variable sampled at a single frequency from a single model running a single experiment (Balaji et al., 2018). These data conform to a file request and a structure controlled

by *Data Reference Syntax* published in partnership with the ESGF. For example, the CMIP6 data conform the CMIP6 data request, (Juckes et al., 2020), and to the CMIP6 Data Reference Syntax (Taylor et al., 2018).

Figure 3 shows an ESGF atomic dataset and a collection of three netCDF files that conform the dataset. Traditional ESGF-based climate data analysis workflows involve downloading the files on the collection for at least one atomic dataset. The files are downloaded to a local workstation or HPC infrastructure. In subsequent steps of the data analysis workflow, developed software tools and scripts, are executed to perform data analysis tasks. However, these programs must often deal with the hierarchical file organization structure of an ESGF repository, introducing complexities unrelated to the primary research analysis task.

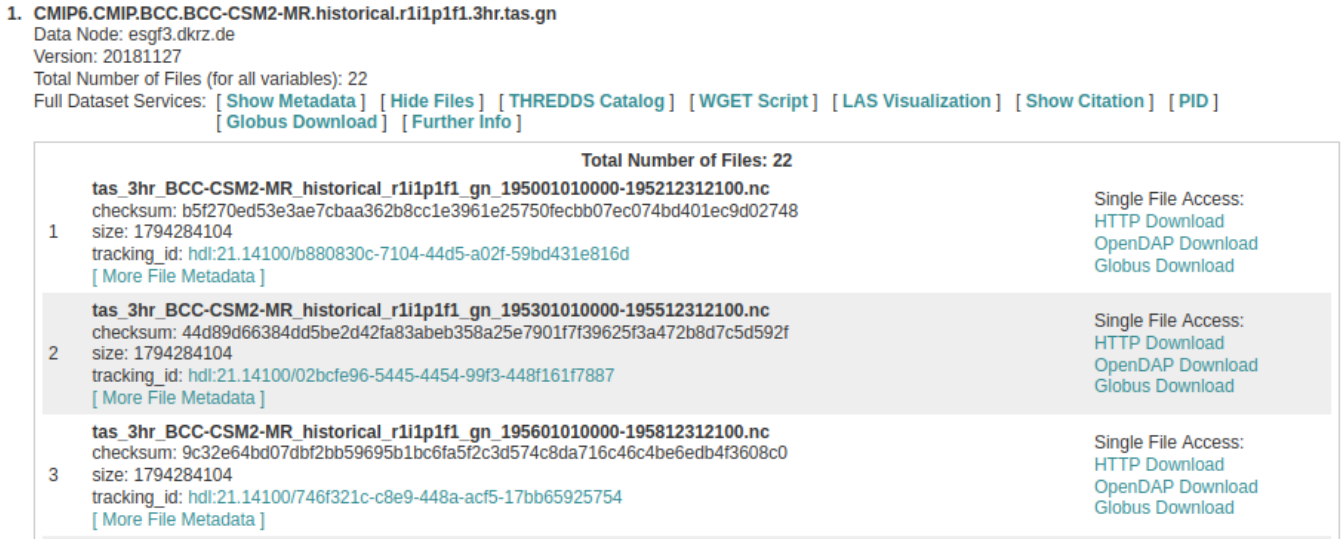


**Figure 3.** ESGF listing three files of a CMIP6 dataset. A common practice in the ESGF consists in splitting the dataset into many files along the time dimension. Smaller files are easier to manage in the federation but performing data analysis becomes harder. This image is a screenshot obtained from the ESGF web portals. Credit is attributed to the ESGF partners supporting these portals. For further details, please refer to https://esgf.llnl.gov/acknowledgments.html.

The goals of ARD are aimed at addressing the inherent complexities associated with file handling. To achieve this, various methodologies are under consideration, based in either aggregations of the original datasets and/or transition to new infrastructures such as cloud providers. Aggregation-based approaches focus on creating either physical or virtual views of data, optimized for efficient analysis, thereby relieving users from the intricacies of directly manipulating netCDF files. On the other hand, performance optimization-based approaches involve leveraging hardware infrastructures, such as cloud computing providers, to enhance the speed and efficiency of data analysis operations. By utilizing these resources, significant improvements in processing capabilities can be achieved, thereby facilitating smoother data analysis workflows.

ARD based on aggregations can be performed at different layers of abstraction and may involve varying levels of complexity depending on the desired outcome. Many approaches are based on data analysis applications offering functionality for

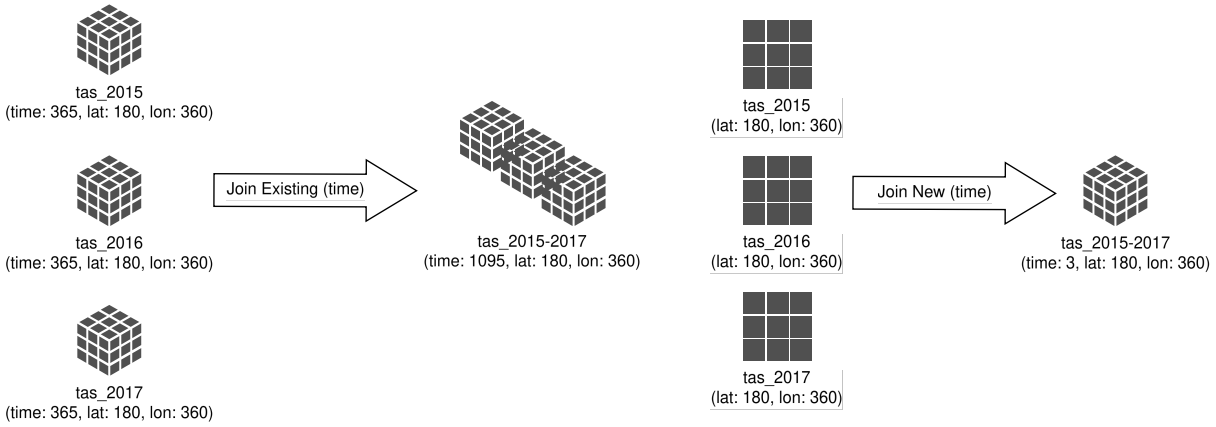


**Figure 4.** Illustration of both *join existing* and *join new* aggregations along the time dimension. In the case of the join exiting, the result of the aggregation is a multidimensional array with the same dimensions (*time*, *lat* and *lon*), in which the size of the time dimension has increased. In the case of the join new aggregation, the time dimension is a new dimension created to aggregate the existing two dimensional arrays into a new three dimensional array.

abstracting the underlying files and hierarchical file system organization from the data user. Examples of this approach include xarray's (Hoyer and Hamman, 2017) `open_mfdataset` function and software applications for climate data analysis. Listing 1 provides an example of the usage of the `open_mfdataset` function. Example of software applications include CF-Python (Hassell et al., 2017), xMIP, (Busecke et al., 2023), intake-esm (Banihirwe et al., 2023) and intake-esgf (Collier et al., 2024). In general, these approaches hold an in-memory representation of the virtual dataset or aggregation, which is manipulated by the data analysis package behind the scenes.

Software packages may offer to persist their aggregated logical view of the underlying files, but these persistence formats are not interoperable between packages, and/or do not provide interchangeable logical view of aggregation. In the process of generating aggregated views, data may be duplicated, or virtual aggregations can be used to avoid the data duplication. The advantage of relying on virtual dataset capabilities is that data duplication is avoided, and the existing infrastructure may be reused to obtain ARD capabilities without huge costs associated. Examples of virtual aggregations that follow this approach include (but are not limited to) NcML (Caron et al., 2009), Kerchunk (Durant), CFA (Hassell et al.), and HDF5 Virtual Datasets (The HDF Group, 2024). The lack of a standard persistence format is also accompanied by different approaches to aggregation methodologies, which arise from a lack of a common data model and a suitable algebra in the context of climate data management.

One attempt to address this issue is the development of the Climate Forecast Aggregation (CFA) conventions, which can describe an aggregated view of netCDF files using the Climate Forecast conventions (Hassell et al.). The CFA conventions provide a formal syntax for storing an aggregation view of file *fragments* using netCDF itself as the storage mechanism. Currently, this syntax is only supported by CF-Python, but libraries and tools are in development to extend CFA support to other packages once the syntax has been through the CF conventions process. CF-Python (Hassell et al., 2017) utilises an underlying

data model from the CF conventions, which extends the original netCDF data model with custom structure types. With this data model, a set of unambiguous rules can be established which allow formal manipulation of netCDF variable fragments.

```
ds=xr.open_mfdataset(
    sorted(glob.glob(
        "/storage/ESGF/CMIP6/.../tas_3hr_BCC-CSM2-MR_historical_r1i1p1f1_gn_*.nc")),
    combine="nested",
    concat_dim=["time"])
```

**Listing 1.** Usage of xarray's *open_mfdataset* to generate an ARD dataset at the application layer from several netCDF files.

Similarly, the software library netCDF-java (Caron et al., 2009) extends the original netCDF data model with additional operations for manipulation of climate datasets. Using the netCDF-java nomenclature, the operations *join existing* and *join new* are defined among others. A *join existing* operation concatenates variables of netCDF datasets on a given input dimension. A *join new* operations merges variables of netCDF datasets by creating a new coordinate dimension, thus extending the dimensionality of the variable. An example of both types of aggregation is shown in Figure 4. Such operations depend on clean notions of variable identity in order to ensure the semantic correctness of the aggregations. In addition, these virtual aggregations may be performed by referencing remote sources of data using the OPeNDAP protocol (Garcia et al., 2009). This particular capability is exploited by the ESGF-VA to provide ARD to the whole ESGF community by exploiting the existence of OPeNDAP access in the federation (Caron et al., 1997).

```
<?xml version="1.0" encoding="UTF-8"?>
<netCDF xmlns="http://www.unidata.ucar.edu/namespaces/netCDF/NcML-2.2">
 <aggregation dimName="time" type="joinExisting">
   <netCDF
      location="tas_3hr_BCC-CSM2-MR_historical_r1i1p1f1_gn_195001010000-195212312100.nc"/>
   <netCDF
      location="tas_3hr_BCC-CSM2-MR_historical_r1i1p1f1_gn_195301010000-195512312100.nc"/>
   <netCDF
      location="tas_3hr_BCC-CSM2-MR_historical_r1i1p1f1_gn_195601010000-195812312100.nc"/>
 </aggregation>
</netCDF>
```

**Listing 2.** NcML file that showcases a logical aggregation by performing a *join existing* aggregation over several local netCDF files.

As already discussed, another approach to ARD is to leverage the capabilities of novel hardware infrastructures such as cloud providers. One notable example of this is the Pangeo initiative, a collaborative effort that brings together diverse communities to address challenges in climate data analysis. Pangeo has facilitated the development of cloud-native repositories tailored specifically for climate data analysis needs. These repositories leverage the capabilities of commercial public cloud providers, such as Amazon Web Services (AWS) or Google Cloud Platform (GCP), to provide scalable, efficient and operational storage

solutions for climate ARD. Pangeo has established a collaboration with the ESGF for further enhancing the accessibility and usability of climate data for researchers and practitioners worldwide (Abernathey et al., 2021; Stern et al., 2022).

Cloud native repositories have enormously facilitated climate data analysis by leveraging the capabilities of remote data access provided by cloud infrastructures and ARD on top of cloud native data formats. As a result, climate data is accessible from anywhere and climate data analysts are able to opt for the computation platform of their choice, either HPC infrastructures from their home institutions, user-paid on-demand cloud resources running close to the cloud repository or even a personal laptop. However, the establishment of these repositories has demanded substantial investments in human resources and financial resources to accommodate storage within the premises of cloud service providers[1]. In addition, in order to keep consistent copies of the source repositories, the cost required to sustain cloud native repositories is increased as long as the source repositories keep updating their datasets. The following section presents a model evaluation data analysis task that demonstrates the benefits of ARD and remote data access provided by ESGF-VA for climate data analysis.

## 4 Model evaluation

ARD enables model evaluation data analysis tasks to be carried out with much greater ease compared to working with raw data files. This section illustrates a model evaluation task focused on studying model member agreement on precipitation outputs from the CanESM5 global climate model (Swart et al., 2019) on the region of Europe. The data analysis tasks computes relative anomalies of precipitation for two future scenarios relative to the historical period. Due to the convenience of dealing with ARD datasets and remote data access, this workflow saves the user from locating and downloading from the ESGF the 54 netCDF files required to perform the task. Instead, only three URLs will be used. These URLs can be easily obtained from the ESGF-VA. The three URLs correspond to the ESGF-VA endpoints of the CanESM5 multi-member data sources of the historical, SSP1-2.6 and SSP5-8.5 of the CMIP and ScenarioMIP ESGF activities respectively. Moreover, spatial and temporal subsetting is automatically performed by OPeNDAP on behalf of the user, regardless of how the netCDF files are split along the time coordinate in the ESGF.

The data analysis task involves calculating model agreement on precipitation anomalies by computing the difference between the climatologies of both future scenarios relative to the historical period. The climatologies for each scenario have been computed as the temporal and model ensemble member mean of 18 model runs, given that this information is available out-of-the-box in the ARD dataset from the ESGF-VA. The years 1995 to 2014 are chosen as the reference for the historical period, and the years 2080 to 2100 represent the future period. Model member agreement will be computed following the *low model agreement simple approach* methodology proposed in the Intergovernmental Panel on Climate Change (IPCC) Sixth Assesment Report (IPCC, 2023). This methodology aims to display robustness and uncertainty in maps of multi-model mean changes. Model agreement is computed using *model member democracy* without discarding/weighting model members. Low model

[1]Refer to Pangeo Showcase talk "How to transform thousands of CMIP6 datasets to zarr with Pangeo Forge - And why we should never do this again!" for further details on this topic.

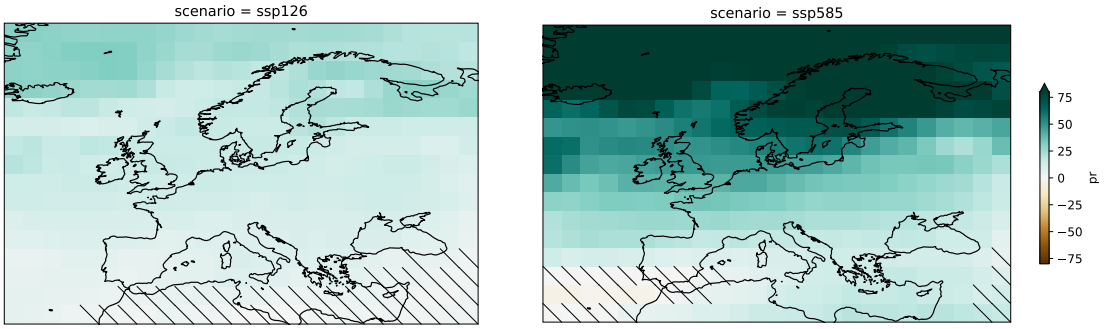


**Figure 5.** Model member agreement on relative precipitation changes for 18 model members across two future scenarios from CanESM5. These changes are calculated as the difference between the projected future scenario period and the historical reference. The results highlight significant projected increases in precipitation over northern Europe by the end of the century, under the fossil-fueled development scenario. Diagonal lines indicate areas in southern Europe where model member agreement is low.

member agreement locations, those with $< 80\%$ agree on sign of change, are marked using diagonal hatched lines (Gutiérrez et al., 2021). The results for both future scenarios are shown in figure 5.

## 5 Implementation

ARD in the form of virtual aggregations or virtual datasets allows users to view the data of their interest as single logical units rather than collections of files. This eliminates the need to navigate through files that necessitate intricate data analysis programming for interpretation. In ESGF-VA, the logical aggregations are based on aggregation capabilities expressed in NcML, and provided by netCDF-Java. With NcML it's not required to inspect the storage internals of the netCDF files in order to perform the aggregation. This is in contrast to other alternatives such as Kerchunk, which currently requires that all the variables or multidimensional arrays are parameterized with the same configuration (chunking, filters, etc). This is often not the case in ESGF. Kerchunk also needs to extract the byte positions of the chunks from the source netCDF files. Given that the ESGF contains millions of netCDF files, avoiding inspection of each of them provides an enormous advantage. ESGF index nodes contain metadata about netCDF files, and they can be used to quickly retrieve metadata from the netCDF files. Thus, the complexity and required time of the process of generating the virtual aggregations is reduced in several orders of magnitude.

The implementation of the ESGF-VA involves the following steps:

1. The search process involves querying the ESGF catalog and indexing service to obtain dataset information and metadata, which is then stored in a local database.

2. The aggregation process queries the local database to create virtual datasets (NcMLs) for the entire federation. These are the ARDs that the user utilizes for remote climate data analysis.

Figure 6 shows how netCDF files from the ESGF that belong to the CMIP6 project are distributed between the virtual datasets. Most virtual datasets contain few references to netCDF files inside (<=100) although some virtual aggregations provide access to hundreds or even thousands of netCDF files. Table 1 shows the ratio of netCDF per NcML for each CMIP6 activity (Eyring et al., 2016). The following sections detail the implementation of both the search and aggregation processes.

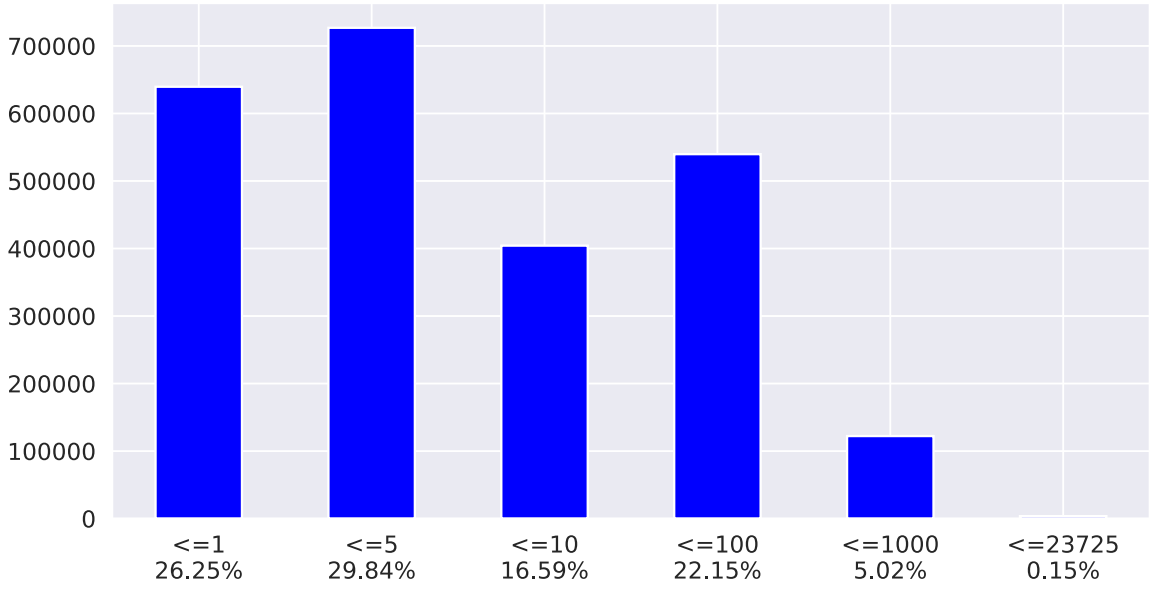


**Figure 6.** Distribution of netCDF files in the virtual datasets (NcMLs). Most of the virtual aggregations are made of a relatively small number of files although some virtual datasets spawn hundreds or thousands of files.

### 5.1 The ESGF search process

For the search process, the ESGF Search REST API (Cinquini et al., 2012) is used by the client to query the contents of the underlying search index, returning results matching the given constraints in the whole federation. The search service provides useful metadata that allow clients to obtain valuable information about the datasets being queried. However, in the context of the ESGF-VA, it is not as efficient as one would like - sufficient for the first implementation and experiments described here, but in an operational context one would want to see time coordinate information held in the index. This is because applications otherwise need to read such information from each and every file in an aggregation, which may be a significant overhead, before any actual data transfer.

The search process is performed by an iterative querying the ESGF search service, requesting small chunks of data that are manageable by the service. The search service limits the number of records that can be obtained from a single request to ten thousand elements. Since the federation contains information on the order of tens of millions of records, several requests

| CMIP6 activity | NcMLs | netCDFs | Ratio (netCDFs / NcML) |
|---|---|---|---|
| ISMIP6 | 2570 | 10864 | 4.23 |
| GMMIP | 9489 | 587501 | 61.91 |
| LS3MIP | 16041 | 188533 | 11.75 |
| OMIP | 17009 | 441578 | 25.96 |
| PAMIP | 19824 | 4931240 | 248.75 |
| CDRMIP | 21189 | 395444 | 18.66 |
| PMIP | 26277 | 645989 | 24.58 |
| GeoMIP | 28470 | 184666 | 6.49 |
| FAFMIP | 41324 | 208881 | 5.05 |
| LUMIP | 57140 | 581573 | 10.18 |
| HighResMIP | 63359 | 5806778 | 91.65 |
| RFMIP | 81548 | 745604 | 9.14 |
| CFMIP | 81599 | 309421 | 3.79 |
| C4MIP | 81964 | 847376 | 10.34 |
| DAMIP | 134708 | 3482721 | 25.85 |
| AerChemMIP | 199307 | 1850392 | 9.28 |
| ScenarioMIP | 250591 | 17317882 | 69.11 |
| CMIP | 505733 | 19090708 | 37.75 |
| DCPP | 506085 | 8152594 | 16.11 |
| **Total** | **2144227** | **65779745** | **-** |

**Table 1.** Number of virtual aggregations (NcMLs), netCDF files for which metadata was retrieved from the federation and ratio of netCDF per NcML generated for CMIP6 in the ESGF Virtual Aggregation. Note that the distribution of number of references to netCDFs files on NcMLs does not follow a uniform distribution (see figure 6).

need to be made. The results are stored in a local SQL database and multiple ESGF Virtual Dataset labels are assigned to the record, in order to identify the virtual dataset in which the records participate in different virtual aggregations. Figure 7 gives an overview of the cost in size of generating the NcMLs.

### 5.2 The aggregation process

The aggregation process is responsible for generating the virtual aggregations and mapping multiple ESGF individual files and their metadata to the appropriate virtual datasets. Although the number of records could be overwhelming, the use of SQL indexes allows the aggregation process to quickly retrieve the granules that belong to the different virtual datasets. The result from the aggregation process in the ESGF Virtual Aggregation is a collection of NcML files that represent the virtual datasets. The virtual datasets are stored in different directories in order to provide appropriate organization. Each virtual dataset is

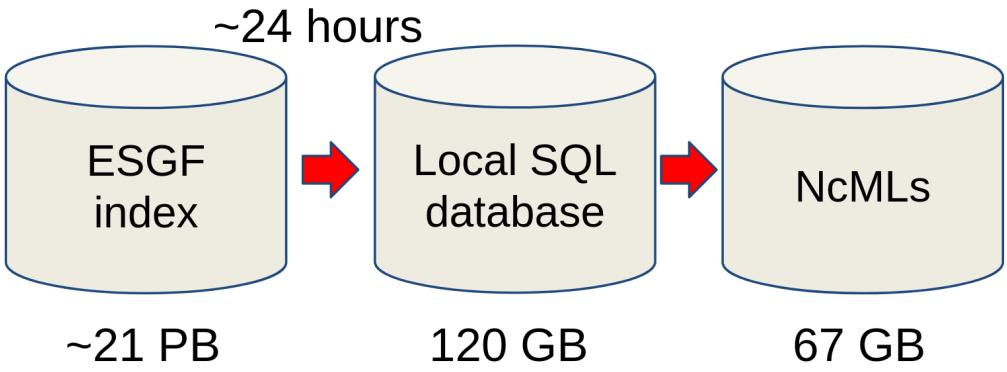


**Figure 7.** From left to right, sizes of the ESGF federation data archive for CMIP6 including replicas, the local database containing the metadata from the federation and the ESGF Virtual Dataset. Note that the 21 PB of data refers to the size of the netCDF files stored within the federation. The metadata in the ESGF indexes requires storage on the order of gigabytes and allow querying metadata about netCDF files in a reasonable amount of time. Storage requirement for the virtual aggregations is reduced in several orders of magnitude compared to original data.

labeled with the data node to where each of the granules that form the virtual dataset belong. Additionally, the virtual datasets are generated in such a way that replicas from the same virtual dataset are easily identifiable.

The virtual datasets of ESGF-VA are made of two kind of aggregations. First, the ESGF *atomic dataset* aggregation is generated by concatenating the time series of each variable along the time dimension. Figure 4 illustrates this operation and listing 2 provides an NcML example. This concatenation does not increment the rank of dimensions of the multidimensional array that represents the variable, it only increases the size of the time dimension. This kind of aggregation is ignored in time independent variables such as orography. Then the variables are aggregated by creating a new dimension that represents the variant label (i.e. ensemble members), the different model runs of a climate model. The rank of dimensions is incremented by one, to accommodate a dimension for the ensemble or variant label. It is important to note that for this kind of aggregation to be performed properly, climate variables involved must share a spatial and temporal coordinate reference system, with the exact same spatial coordinate values. If that were not the case, the resulting multidimensional array would expose incorrect data. Listing 3 shows the NcML for the virtual aggregation in figure 8.

### 5.3 Remote data access

Remote data access capabilities of the ESGF provided by OPeNDAP and THREDDS (Caron et al., 1997) allow the virtual aggregations to load the data directly and transparently from ESGF data nodes with no file downloads. Figure 8 shows a virtual dataset of the ESGF Virtual Aggregation (the NcML file) opened with xarray through an OPeNDAP THREDDS data server, since xarray does not currently support opening NcML files directly. Figure 1 shows the result of a data analysis task from this dataset. Because a single dataset contains all the ensemble members of a particular member run, only one dataset is needed to perform this data analysis.

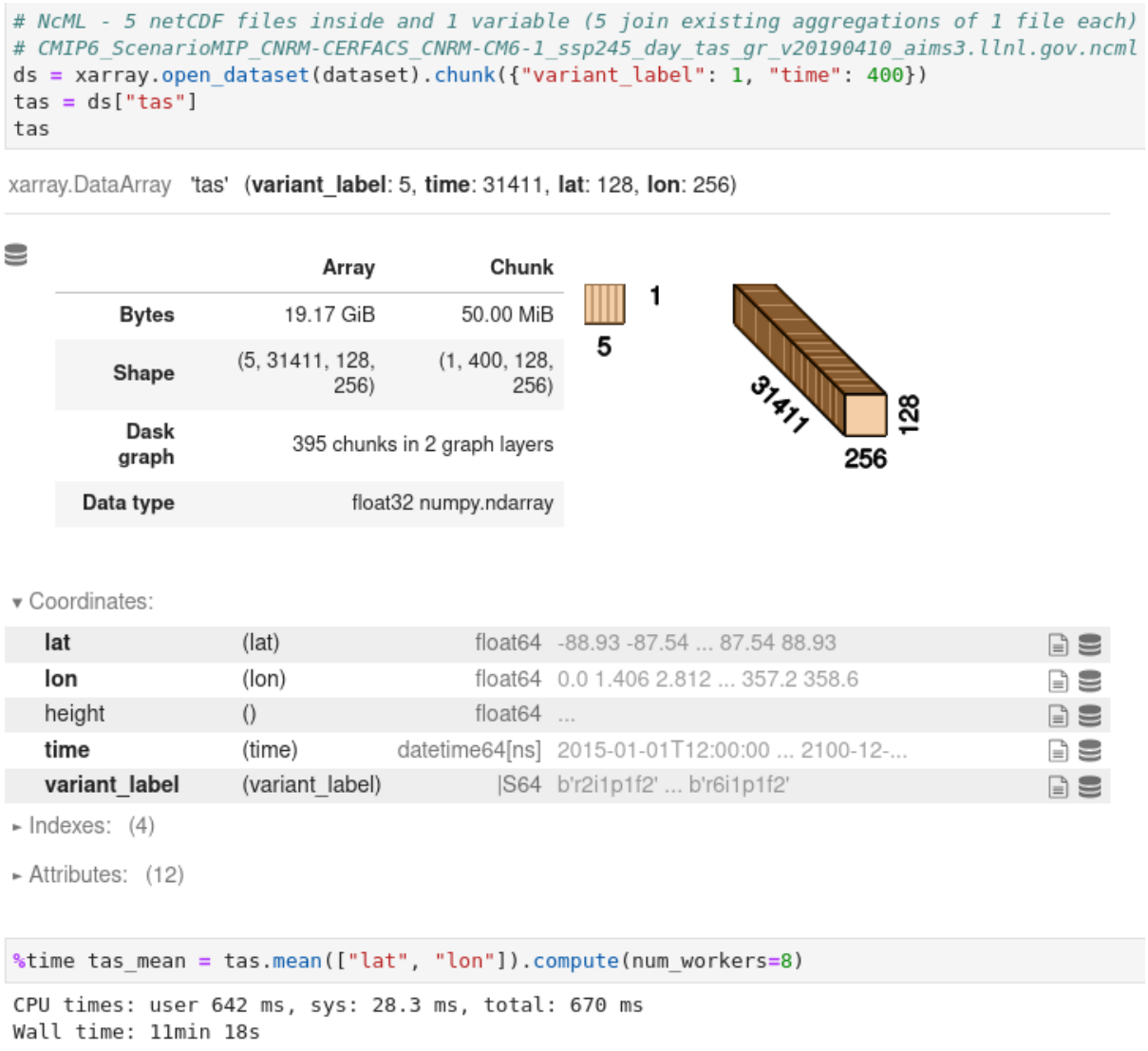


**Figure 8.** Example of an ESGF Virtual Aggregation NcML file of surface temperatures opened with the xarray package through OPeNDAP/THREDDS in a Jupyter Notebook. Readers may notice that surface temperature is a three dimensional variable in ESGF, but it is now a four dimensional variable including the model ensemble member dimension. The user does not need to know the number of files involved in the dataset and it can be analysed as a datacube instead of a series of netCDF files. The data requested by the user through xarray will be fetched on demand directly from ESGF data nodes. The NcML is available in appendix A.

## 6 Performance

To investigate the performance of accessing data using the ESGF-VA, an experiment was carried out to examine data access performance from a xarray client. This limited experiment is enough to show some of the benefits of, and issues with, the ESGF-

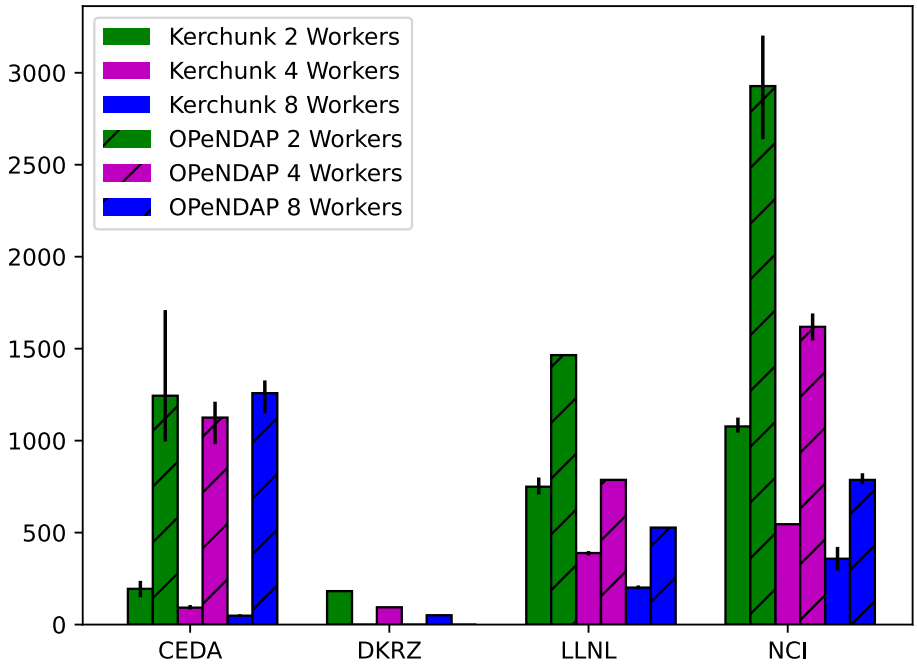


**Figure 9.** The results of the experimental retrieval of data for meaning using Kerchunk and OPeNDAP from a client in Spain (IFCA) to servers in the UK (CEDA), Germany (DKRZ), the US (LLNL) and Australia (NCI). The bars show the mean time, in seconds, taken across experiment replicants for each configuration of number of workers. Where error bars are shown, these reflect the minimum and maximum times taken. Kerchunk data is shown without hatching, and OPeNDAP data with. (Note that there is no OPeNDAP data for DKRZ, and no replicants - and hence no error bars - for the OPeNDAP experiments using the LLNL server.)

VA. The experiment was carried out with the ESGF-VA utilising OPeNDAP, and for comparison, with Kerchunk aggregation. In both cases, virtual aggregations were generated first, and each was performed with varying numbers of Dask worker processes to test the potential scalability (albeit in a situation where we know that there is limited scalability on the servers themselves, and we believe there would have been little or not contention from other users). Here *Kerchunk* refers to the use of Kerchunk files to access individual blocks of compressed data via Zarr and other Pangeo middleware on the client talking directly to an ESGF HTTPS server, whereas OPeNDAP is the vanilla usage of the ESGF-VA on the client talking to an ESGF OPeNDAP server.

The experiment was simple: we read a dataset consisting of the entire atomic datasets (>80 years) of daily values for one spatially two-dimensional variable (the surface temperature, *tas*) from each member of a simulation and carried out a global mean of that data. The actual calculation was done on a cloud hosted virtual machine in Spain at Instituto de Física de Cantabria

(IFCA), while the data was read from each of four ESGF servers. In each case, the dataset was chunked for Dask into segments of 400 daily values (so each chunk was about 50 MB in memory, the default maximum limit for OPeNDAP), in order to examine the benefit of using multiple Dask workers. We attempted to repeat the experiment five times on each of the ESGF servers for each of 2,4, and 8 Dask workers. However, it was not possible to get OPeNDAP results from all four servers, or to get a fullset from each of the servers - the reasons for this are discussed below. We did not attempt to mitigate against file system caching in this design, as while it could have impacted on the comparison, in practice the I/O time for reading the data ($\sim$10 GB on disk, $\sim$20 GB on memory) would be small compared to the overall times reported.

The results are shown in figure 9. There are several obvious results: when using Kerchunk, considerable benefit was gained by using more workers, and that data nodes close to Spain (where the calculation was done) yielded much faster outcomes than remote data nodes. In each case, OPeNDAP is much slower than Kerchunk, and the benefit of geographical proximity on the OPeNDAP results is much less obvious (e.g. using 8 workers to process data loaded from Australia is faster than using 8 works on data from the UK, but for 2 workers, it is much faster to use the UK data). Unfortunately, DKRZ do not offer the OPeNDAP service, and LLNL took the service down after we did our first experiments and before we added the replicas. It is also clear that the OPeNDAP results from the CEDA server are anomalous in terms of having no dependency on the number of workers.

As already noted, proximity matters. The benefit of the client-side decompression used by Kerchunk is clear. A priori, we might have expected the OPeNDAP results to be roughly a factor of two slower (given that OPeNDAP decompresses server-side and sends the uncompressed data over the wire), and this is roughly what is seen at LLNL and NCI. As already noted the CEDA OPeNDAP results are anomalous so we make no attempt to explain the disparity between Kerchunk and OPeNDAP speeds seen there. For this experiment at least, with the fastest times seen (44s and 49s from CEDA and DKRZ respectively), it is clear that the bottleneck is the data flow across the wide area.

Similar experiments with other data highlighted some suboptimal data practices within the ESGF archive. A significant number of CMIP6 datasets stored in the ESGF exhibit poor chunking configurations, specifically related to the time coordinate. Chunking in HDF5 is a crucial technique for optimizing data access performance. It involves organizing how data is stored on disk, enabling different arrangements based on desired data access patterns. Proper chunking can greatly enhance data access efficiency, similar to how SQL indexes improve database query performance. Conversely, incorrect or inappropriate chunking choices can have a detrimental impact on data access performance. Notably, the CMIP6 files within the ESGF often displayed a chunking configuration of *(1,)* for the time coordinate, resulting in severe degradation of dataset access times (figure 10). Sub-optimal chunking configuration negatively affected the efficiency of data retrieval and subsequent analysis tasks. A fix for the standard climate model output writer (CMOR) has been proposed (https://github.com/PCMDI/cmor/pull/733), although not all modelling centres use CMOR.

## 7 Conclusions

We have introduced the ESGF Virtual Aggregation (ESGF-VA), and shown how it can be used to obtain data from the existing ESGF OPeNDAP servers. In doing so, we have showcased how the ESGF federated index and the ESGF OPeNDAP endpoints

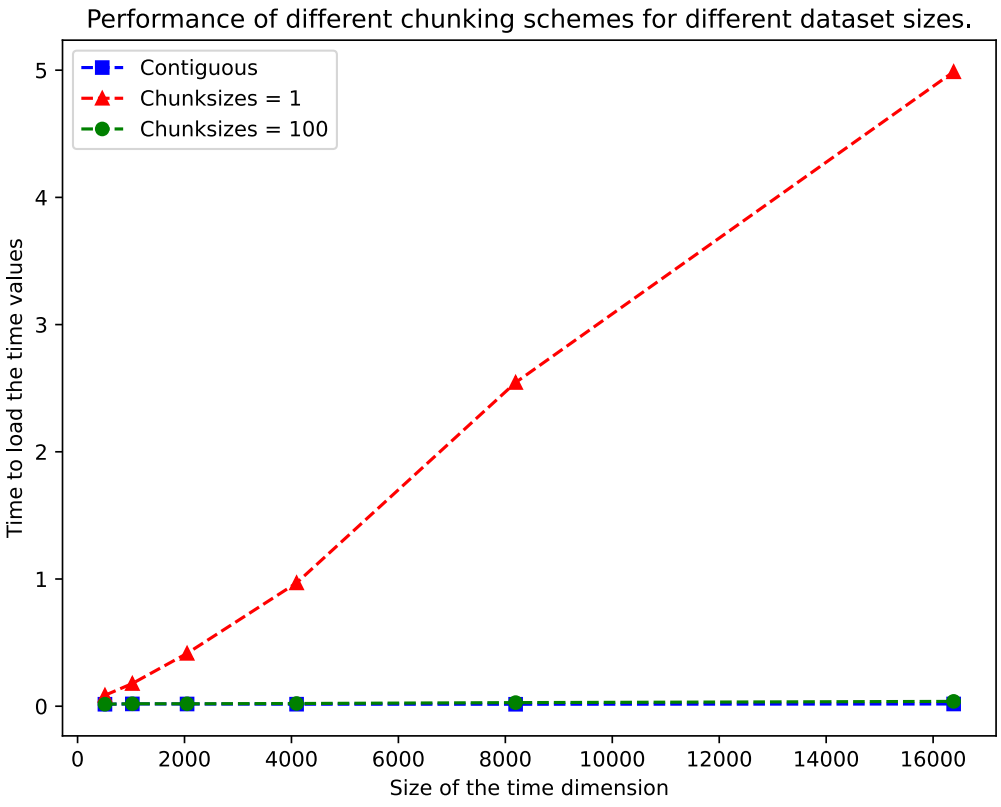


**Figure 10.** Required time to read a temporal coordinate in function of storage type. Contiguous storage does not incur into performance issues. If chunking storage with a bad chunking scheme is used, performance quickly deteriorates.

can be used to deliver capabilities beyond conventional file search and download. By enabling remote data analysis over virtual analysis ready data, the use of the ESGF-VA could enhance the efficiency and productivity of climate data analysis tasks. It could empower researchers to access and analyze data directly within the ESGF environment, eliminating the necessity for time-consuming data transfers and facilitating more streamlined and effective climate data analysis workflows.

### 7.1 Summary of findings

The virtual datasets provided by the ESGF-VA facilitate an aggregated view of the time series as well as the ensemble model members of a particular model. Thus, data analysis comparing different runs of the same model can be performed by loading only the view of one dataset. In doing so, the details of the aggregation are hidden completely from the user, who sees the dataset as a single netCDF. Using the OPeNDAP endpoints of the federation, data analysis can be performed from anywhere. While this implementation of the ESGF-VA exploits NcML and netCDF-java, the concept is readily extensible across any

netCDF client which supports OPeNDAP - i.e. any which utilise the netCDF library itself, rather than directly using HDF5. However, because OPeNDAP performs chunk decompression in the server, it is not as efficient as other data access methods as more data is sent over the network.

The creation of the virtual aggregations presented in this work follows a much more maintainable approach than alternatives focusing on duplication of the data, such as cloud native repositories. The storage requirements of the virtual aggregations are minimal compared to the relative size of the raw data. In addition, the generation of the virtual aggregations can be performed in few hours, where most of the time is spent querying the ESGF distributed index. As the ESGF-VA aggregation information is obtained directly from the existing ESGF index, it can be generated much faster than the process needed to generate Kerchunk indices, which requires access to each file. The speed of creation of new virtual aggregations, coupled with the lack of actual data duplication, means that the system can cope well with an environment where datasets are being updated as data processing issues are found and fixed, since the ESGF-VA can be quickly updated. However, whatever system is used to create analysis ready data, it is necessary to know that such updates are necessary - it would be helpful for a future ESGF to have some sort of automated alert system for data updates.

Certain issues regarding data distribution of the ESGF were identified during the creation of the ESGF Virtual Aggregation. There is often inconsistent use of the version facet, and a significant portion of the data stored in the federation does not adhere to best practices regarding the chunking of HDF5. In the first place, the version facet is supposed to distinguish between allegedly equal datasets that have changed due to different kinds of errors, such as incorrect data due to bad model execution or incorrect publication process. In practice, the version facet may, in some cases, end up dividing granules that should belong to the same aggregation, due to inappropriate usage of the facet. From an ESGF-VA point of view, this could be avoided by using the latest value of the version facet, but that would lead to issues with maintenance. There may be value in both providing better guidance to modelling centres about how to use version facets and in adding some chunk checking to future ingestion processes.

### 7.2 Discussion

The performance analysis presented in this work suggests a declining interest from the ESGF community in supporting OPeNDAP, given the instability of this service compared to data access based on HTTP. While we do not know the details of the individual server configurations, the fact that the CEDA OPeNDAP results are so odd, and that both DKRZ and LLNL no longer offer OPeNDAP servers, it is plausible to conclude that it is a) difficult to deploy OPeNDAP and b) currently not enough usage to justify it. However, our results suggest that there may yet be mileage in deploying properly configured OPeNDAP services in the future ESGF (maybe with a different server, such as that of Gallagher et al. 2022) - at least until such time that remote direct access to chunks via HTTP is available to a much greater proportion of netCDF clients. In doing so, the use of HTTP compression could mitigate the issue of server side decompression of the chunks. This functionality is currently supported by netCDF clients but is currently provided by few, if any, ESGF nodes. Finally, it would also be helpful if the time coordinate information could be stored in the ESGF index to be used by virtual aggregation clients in a way to avoid the need to read time coordinate values from each file when opening the virtual dataset.

While the ESGF-VA provides many benefits for users, albeit with the cost of moving the uncompressed data selections, such benefits would only transpire if there was sufficient server capacity to support demand. Although the ESGF-VA itself requires no change to the ESGF architecture itself, support for access to ESGF data via the OPeNDAP protocol is currently delivered by the use of THREDDS Data Server (a Java web application). While scaling out server infrastructure with THREDDS is possible, it requires both sufficient hardware and significant configuration knowledge. The pros and cons of wider usage of the ESGF-VA or similar OPeNDAP based tools and the consequential need for server capacity and issues of configuration should form part of future ESGF discussions.

It is clear that the ESGF has and will evolve. Our work suggests the the ongoing evolution of the ESGF needs to address not only indexing and data download, but where possible, the provision of direct data access suitable for a wide range of use-cases. Such support may include giving modelling centres good guidance as to how to chunk and organise their data, beyond just relying on CMOR, as not all centres use CMOR. Despite the focus of this work on OPeNDAP, NcML and Kerchunk, future work involves evaluation and assessment of other lightweight data servers and metadata file formats. These will allow better decision making in the process of providing both remote data access and generation of ARD. Examples include but are not limited to xpublish (https://github.com/xpublish-community/xpublish) and DMR++ (https://opendap.github.io/DMRpp-wiki/DMRpp.html).

*Code availability.* The code of the ESGF-VA is open source and freely downloadable at https://doi.org/10.5281/zenodo.14203625. The repository contains the Python scripts that query the ESGF and generate the local database (*search.py*) and generate the NcMLs (*ncmls.py*). Required dependencies are listed in the *environment.yml* file. A tutorial on locating and accessing ESGF-VA datasets is provided in the form of a Jupyter Notebook (*demo.ipynb*). Other notebooks available in the repository provide reproducibility of the figures and performance results of the study (*model_evaluation.ipynb*, *performance.ipynb* and *validation.ipynb*). Kerchunk files used in this study may be found in the *kerchunks* folder. Finally, a sample THREDDS catalog for those interested in setting up a THREDDS Data Server is included in the *content* directory. For further details, refer to the README.md of the repository.

## Appendix A: NcML example

**Listing 3.** NcML generated by the ESGF Virtual Aggregation.

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2">
    <explicit/>
    <attribute name="size" type="int" value="11536844671"/>
    <attribute name="size_human" value="10.7 GiB"/>

    <attribute name="__info__"
               value="Virtual dataset generated by the ESGF Virtual Aggregation"/>
    <attribute name="__license__"
```

```
10:              value="This is a derived dataset product from ESGF, same licenses from original
     datasets apply for this dataset."/>
11:
12:     <!-- only mandatory (when required? = always) attributes from
13:          http://cerfacs.fr/~coquart/data/uploads/cmip6_global_attributes_filenames_cvs_v6.2.6.pdf
14:          are included -->
15:     <!-- mandatory attributes are extracted from netcdf files
16:          BUT creation_date and further_info_url have got custom values relative to EVA -->
17:     <attribute name="activity_id" value="ScenarioMIP"/>
18:     <attribute name="Conventions" value=""/>
19:     <attribute name="data_specs_version" value=""/>
20:     <attribute name="experiment" value=""/>
21:     <attribute name="experiment_id" value="ssp245"/>
22:     <attribute name="forcing_index" value=""/>
23:     <attribute name="frequency" value="day"/>
24:     <attribute name="grid" value=""/>
25:     <attribute name="grid_label" value="gr"/>
26:     <attribute name="initialization_index" value=""/>
27:     <attribute name="institution" value=""/>
28:     <attribute name="institution_id" value="CNRM-CERFACS"/>
29:     <attribute name="license" value=""/>
30:     <attribute name="mip_era" value="CMIP6"/>
31:     <attribute name="nominal_resolution" value=""/>
32:     <attribute name="physics_index" value=""/>
33:     <attribute name="product" value="model-output"/>
34:     <attribute name="realization_index" value=""/>
35:     <attribute name="realm" value="atmos"/>
36:     <attribute name="source" value=""/>
37:     <attribute name="source_id" value="CNRM-CM6-1"/>
38:     <attribute name="source_type" value=""/>
39:     <attribute name="sub_experiment" value=""/>
40:     <attribute name="sub_experiment_id" value="none"/>
41:     <attribute name="table_id" value="day"/>
42:     <attribute name="variable_id" value="tas"/>
43:     <attribute name="cmor_version" value=""/>
44:     <!-- attributes that default to "no parent" if they don't exist -->
45:     <attribute name="branch_method" value="no parent"/>
46:     <attribute name="parent_activity_id" value="no parent"/>
47:     <attribute name="parent_experiment_id" value="no parent"/>
48:     <attribute name="parent_mip_era" value="no parent"/>
49:     <attribute name="parent_source_id" value="no parent"/>
50:     <attribute name="parent_time_units" value="no parent"/>
51:     <!-- attributes that are omitted if they don't exist -->
52:
```

```
    <attribute name="further_info_url" value="See netCDF variable 'further_info_url'"/>
    <attribute name="creation_date" value=""/>
    <attribute name="version" value="v20190410"/>
    <attribute name="replica" value="1"/>

    <dimension name="nfiles" length="5"/>
    <dimension name="file" length="2"/>
    <variable name="further_info_url" type="string" shape="nfiles file">
        <values>http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r2i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r2i1p1f2_gr_20150101-21001231.nc https://furtherinfo.es-doc.org/CMIP6.CNRM-CERFACS.CNRM-CM6-1.ssp245.none.r2i1p1f2 http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r3i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r3i1p1f2_gr_20150101-21001231.nc https://furtherinfo.es-doc.org/CMIP6.CNRM-CERFACS.CNRM-CM6-1.ssp245.none.r3i1p1f2 http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r4i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r4i1p1f2_gr_20150101-21001231.nc https://furtherinfo.es-doc.org/CMIP6.CNRM-CERFACS.CNRM-CM6-1.ssp245.none.r4i1p1f2 http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r5i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r5i1p1f2_gr_20150101-21001231.nc https://furtherinfo.es-doc.org/CMIP6.CNRM-CERFACS.CNRM-CM6-1.ssp245.none.r5i1p1f2 http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r6i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r6i1p1f2_gr_20150101-21001231.nc https://furtherinfo.es-doc.org/CMIP6.CNRM-CERFACS.CNRM-CM6-1.ssp245.none.r6i1p1f2</values>
    </variable>
    <variable name="tracking_id" type="string" shape="nfiles file">
        <values>http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r2i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r2i1p1f2_gr_20150101-21001231.nc hdl:21.14100/732d884b-c3ea-4e0f-ac0b-9a2c36b0144e http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r3i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r3i1p1f2_gr_20150101-21001231.nc hdl:21.14100/f426136c-5f56-40fa-8320-5d85b392d063 http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r4i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r4i1p1f2_gr_20150101-21001231.nc hdl:21.14100/3d1b3f27-306a-4e1d-a132-7560873809fd http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r5i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r5i1p1f2_gr_20150101-21001231.nc hdl:21.14100/85b18e89-37d1-46e3-b18b-b48a6fb5f33f http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r6i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r6i1p1f2_gr_20150101-21001231.nc hdl:21.14100/dfc72483-de8e-47a8-b148-e0d53fbe059a</values>
    </variable>

    <dimension name="variant_label" length="5"/>
    <variable name="variant_label" shape="variant_label" type="string">
        <attribute name="standard_name" value="realization"/>
        <attribute name="_CoordinateAxisType" value="Ensemble"/>
```

```
    </variable>
    <aggregation type="joinNew" dimName="variant_label">
        <variableAgg name="tas"/>
            <netcdf coordValue="r2i1p1f2">
                <aggregation type="joinExisting" dimName="time">
                    <netcdf location="http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r2i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r2i1p1f2_gr_20150101-21001231.nc"/>
                </aggregation>
            </netcdf>
            <netcdf coordValue="r3i1p1f2">
                <aggregation type="joinExisting" dimName="time">
                    <netcdf location="http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r3i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r3i1p1f2_gr_20150101-21001231.nc"/>
                </aggregation>
            </netcdf>
            <netcdf coordValue="r4i1p1f2">
                <aggregation type="joinExisting" dimName="time">
                    <netcdf location="http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r4i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r4i1p1f2_gr_20150101-21001231.nc"/>
                </aggregation>
            </netcdf>
            <netcdf coordValue="r5i1p1f2">
                <aggregation type="joinExisting" dimName="time">
                    <netcdf location="http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r5i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r5i1p1f2_gr_20150101-21001231.nc"/>
                </aggregation>
            </netcdf>
            <netcdf coordValue="r6i1p1f2">
                <aggregation type="joinExisting" dimName="time">
                    <netcdf location="http://aims3.llnl.gov/thredds/dodsC/css03_data/CMIP6/ScenarioMIP/CNRM-CERFACS/CNRM-CM6-1/ssp245/r6i1p1f2/day/tas/gr/v20190410/tas_day_CNRM-CM6-1_ssp245_r6i1p1f2_gr_20150101-21001231.nc"/>
                </aggregation>
            </netcdf>
    </aggregation>
</netcdf>
```

*Author contributions.* **Ezequiel Cimadevilla:** Investigation; methodology; software; visualization; writing – original draft; writing – review and editing.

*Author contributions.* **Bryan N. Lawrence:** Methodology; visualization; writing – original draft; writing – review and editing

*Author contributions.* **Antonio S. Cofiño:** Writing – review and editing

*Competing interests.* The authors declare that they have no competing interests.

*Acknowledgements.* CORDyS project (PID2020-116595RB-I00) and ATLAS (PID2019-111481RB-I00) funded by Ministerio de Ciencia e Innovación / Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033). Ph.D. grant PRE2021-097646 funded by MICIU/AEI/10.13039/501100011033 and by ESF+. PTI-Clima, MITECO and NextGenerationEU (Regulation EU 2020/2094) IMPETUS4CHANGE, grant agreement no. 101081555, from the European Union's Horizon Europe research and innovation programme. European Union's Horizon 2020 research and innovation programme under grant agreement No. 824084 (IS-ENES3).

## References


Abernathey, R. P., Augspurger, T., Banihirwe, A., Blackmon-Luca, C. C., Crone, T. J., Gentemann, C. L., Hamman, J. J., Henderson, N., Lepore, C., McCaie, T. A., Robinson, N. H., and Signell, R. P.: Cloud-Native Repositories for Big Scientific Data, Computing in Science & Engineering, 23, 26–35, https://doi.org/10.1109/MCSE.2021.3059437, 2021.

Asadnabizadeh, M.: Critical findings of the sixth assessment report (AR6) of working Group I of the intergovernmental panel on climate change (IPCC) for global climate change policymaking a summary for policymakers (SPM) analysis, International Journal of Climate Change Strategies and Management, 15, 652–670, https://doi.org/10.1108/IJCCSM-04-2022-0049, 2023.

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, Geoscientific Model Development, 11, 3659–3680, https://doi.org/10.5194/gmd-11-3659-2018, 2018.

Banihirwe, A., Long, M., Grover, M., bonnland, Kent, J., Bourgault, P., Squire, D., Busecke, J., Spring, A., Schulz, H., Paul, K., RondeauG, and Kölling, T.: intake/intake-esm: intake-esm v2023.11.10, https://doi.org/10.5281/ZENODO.3491062, 2023.

Busecke, J., Ritschel, M., Maroon, E., Nicholas, T., and Readthedocs-Assistant: jbusecke/xMIP: v0.7.1, https://doi.org/10.5281/ZENODO.3678662, 2023.

Caron, J., Davis, E., Hermida, M., Heimbigner, D., Arms, S., Ward-Garrison, C., May, R., Madry, L., Kambic, R., and Johnson, H.: Unidata THREDDS Data Server, https://doi.org/10.5065/D6N014KG, language: en Medium: application/java-archive, 1997.

Caron, J., Davis, E., Hermida, M., Heimbigner, D., Arms, S., Ward-Garrison, C., May, R., Madry, L., Kambic, R., Van Dam II, H., and Johnson, H.: Unidata NetCDF-Java Library, https://doi.org/10.5065/DA15-J131, 2009.

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., Drach, B., 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, in: 2012 IEEE 8th International Conference on E-Science, pp. 1–10, IEEE, Chicago, IL, USA, ISBN 978-1-4673-4466-1 978-1-4673-4467-8 978-1-4673-4465-4, https://doi.org/10.1109/eScience.2012.6404471, 2012.

Collier, N., Grover, M., and Stachelek, J.: esgf2-us/intake-esgf, https://github.com/esgf2-us/intake-esgf, 2024.

Durant, M.: fsspec/kerchunk, https://github.com/fsspec/kerchunk.

Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., and Lymburner, L.: Analysis Ready Data: Enabling Analysis of the Landsat Archive, Remote Sensing, 10, 1363, https://doi.org/10.3390/rs10091363, 2018.

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

Fiore, S., Nassisi, P., Nuzzo, A., Mirto, M., Cinquini, L., Williams, D., and Aloisio, G.: A climate change community gateway for data usage & data archive metrics across the earth system grid federation, in: CEUR Workshop Proceedings, vol. 2975, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117857366&partnerID=40&md5=4882870b6cda97c5595337cb15c624b2, 2021.

Gallagher, J., Potter, N., Rmorris2342, Captain James Tiberius Kirk, Kodi Neumiller, Travis, T. R., Tsgouros, Blackone-Sudo, Kyang2014, Slav Korolev, Horák, D., Davis, E., H. Joe Lee, Yuanho, Poplawski, O., Schmidt, R., and Lloyd, S.: OPENDAP/libdap4: libdap 3.20.11 for Hyrax 1.16.8, https://doi.org/10.5281/ZENODO.6878992, 2022.

Garcia, J., Fox, P., West, P., and Zednik, S.: Developing service-oriented applications in a grid environment: Experiences using the OPeNDAP back-end-server, Earth Science Informatics, 2, 133–139, https://doi.org/10.1007/s12145-008-0017-0, 2009.

Gutiérrez, J., Jones, R., Narisma, G., Alves, L., Amjad, M., Gorodetskaya, I., Grose, M., Klutse, N., Krakovska, S., Li, J., Martínez-Castro, D., Mearns, L., Mernild, S., Ngo-Duc, T., van den Hurk, B., and Yoon, J.-H.: Atlas, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 1927–2058, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.021, 2021.

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

Hassell, D., Gregory, J., Massey, N. R., Lawrence, B. N., and Bartholomew, S. L.: NetCDF Climate and Forecast Aggregation (CFA) Conventions, https://github.com/NCAS-CMS/cfa-conventions/blob/main/source/cfa.md.

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., and Taylor, K. E.: A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1), Geoscientific Model Development, 10, 4619–4646, https://doi.org/10.5194/gmd-10-4619-2017, 2017.

Hoyer, S. and Hamman, J.: xarray: N-D labeled Arrays and Datasets in Python, Journal of Open Research Software, 5, 10, https://doi.org/10.5334/jors.148, 2017.

IPCC: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1 edn., ISBN 978-1-00-915789-6, https://doi.org/10.1017/9781009157896, 2023.

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), Geoscientific Model Development, 13, 201–224, https://doi.org/10.5194/gmd-13-201-2020, 2020.

Mahecha, M. D., Gans, F., Brandt, G., Christiansen, R., Cornell, S. E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., Donges, J. F., Dorigo, W., Estupinan-Suarez, L. M., Gutierrez-Velez, V. H., Gutwin, M., Jung, M., Londoño, M. C., Miralles, D. G., Papastefanou, P., and Reichstein, M.: Earth system data cubes unravel global multivariate dynamics, Earth System Dynamics, 11, 201–234, https://doi.org/10.5194/esd-11-201-2020, 2020.

Nativi, S., Mazzetti, P., and Craglia, M.: A view-based model of data-cube to support big earth data systems interoperability, Big Earth Data, 1, 75–99, https://doi.org/10.1080/20964471.2017.1404232, 2017.

OGC: WPS 2.0.2 Interface Standard, http://docs.opengeospatial.org/is/14-065/14-065.html, 2015.

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., Antonio, F., Berger, K., Bretonnière, P.-A., Cinquini, L., Dart, E., Dwarakanath, P., Druken, K., Evans, B., Franchistéguy, L., Gardoll, S., Gerbier, E., Greenslade, M., Hassell, D., Iwi, A., Juckes, M., Kindermann, S., Lacinski, L., Mirto, M., Nasser, A. B., Nassisi, P., Nienhouse, E., Nikonov, S., Nuzzo, A., Richards, C., Ridzwan, S., Rixen, M., Serradell, K., Snow, K., Stephens, A., Stockhause, M., Vahlenkamp, H., and Wagner, R.: Coordinating an operational data distribution network for CMIP6 data, Geoscientific Model Development, 14, 629–644, https://doi.org/10.5194/gmd-14-629-2021, 2021.

Rew, R., Davis, G., Emmerson, S., Cormack, C., Caron, J., Pincus, R., Hartnett, E., Heimbigner, D., Appel, L., and Fisher, W.: Unidata NetCDF, https://doi.org/10.5065/D6H70CW6, language: en Medium: application/java-archive,application/gzip,application/tar, 1989.

Schnase, J. L., Lee, T. J., Mattmann, C. A., Lynnes, C. S., Cinquini, L., Ramirez, P. M., Hart, A. F., Williams, D. N., Waliser, D., Rinsland, P., Webster, W. P., Duffy, D. Q., McInerney, M. A., Tamkin, G. S., Potter, G. L., and Carriere, L.: Big Data Challenges in Climate Science: Improving the next-generation cyberinfrastructure, IEEE Geoscience and Remote Sensing Magazine, 4, 10–22, https://doi.org/10.1109/MGRS.2015.2514192, 2016.

Stern, C., Abernathey, R., Hamman, J., Wegener, R., Lepore, C., Harkins, S., and Merose, A.: Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production, Frontiers in Climate, 3, 782 909, https://doi.org/10.3389/fclim.2021.782909, 2022.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., Von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geoscientific Model Development, 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019, 2019.

Taylor, K. E., Juckes, M., Balaji, V., Cinquini, L., Denvil, S., Durack, P. J., Elkington, M., Guilyardi, E., Kharin, S., Lautenschlager, M., and others: CMIP6 Global Attributes, DRS, Filenames, Directory Structure, and CV's, 2018.

The HDF Group: Hierarchical Data Format, version 5, https://github.com/HDFGroup/hdf5, original-date: 2020-04-24T18:25:20Z, 2024.

Venturini, T., De Pryck, K., and Ackland, R.: Bridging in network organisations. The case of the Intergovernmental Panel on Climate Change (IPCC), Social Networks, 75, 137–147, https://doi.org/10.1016/j.socnet.2022.01.015, 2023.

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.