

IS-ENES – WP10/JRA4

D10.4: User service package

Abstract

The stabilisation of powerful and flexible core services described in D10.3 (Core Service Package) underpins a range of user services which have been developed and deployed in the last period of the project. The services deliver the flexibility and reliability that users will need to deal with large and complex archives of climate model data such as the maturing CMIP5 archive and the developing CORDEX archive. Facilitating the use of 3rd party software tools or software developed by users is central to the approach followed here.

Task 4, “Development of the User Services Package”, of WP10/JRA4 aimed to produce user services on top of the core archive services developed in task 3 (and described in D10.3). The work is split into four main subtasks: (1) access support, (2) visualisation, (3) data manipulation and efficiency and (4) support for regional modelling. In section 2, the work done on these 4 sub-tasks is described in the corresponding subsections.

Grant Agreement	228203	Proposal Number:	FP7-INFRA-2008-
Project Acronym:	IS-ENES		
Project Co-ordinator:	Dr Sylvie JOUSSAUME		

Document Title:	User Service Package			Deliverable:	D10.4
Document Id N°:		Version:	Final	Date:	27/03/2013
Status:					

Filename:	IS-ENES_D10.4_Final.doc
------------------	-------------------------

Project Classification:	Public
--------------------------------	--------

Approval Status		
Document Manager	Verification Authority	Project Approval

REVISION TABLE

Version	Comments
Draft04	Circulated for review
Draft05	Final revision

Table of Contents

1	Sub-tasks.....	4
1.1	Access Support	4
1.2	Visualisation.....	6
1.3	Data manipulation and efficiency	6
1.4	Support for regional modelling.....	11
2	Outlook	11
3	Summary.....	11
4	ANNEX 1: Software libraries.....	12
5	ANNEX 2: Glossary.....	13

1 Sub-tasks

1.1 Access Support

A range of tools and services have been developed to support access to the archive. The complexity and volume of the data collection is such that users could lose much time in a maze of options. Flexible and robust tools allow users to quickly select a well constrained sub-set of the data, avoiding lengthy searches in a browser interface or the overhead of downloading more data than required.

Synchro-data

Synchro-data is a python program managing discovery/authentication/certificate/download and versioning processes from the [CMIP5](#) archive in an easy way. Furthermore it is compatible with all projects currently available from ESGF (e.g. Obs4MIPs, CORDEX, ...). It's a command line tool designed to facilitate the download of files hosted by the distributed digital repositories of the ESGF Federation. It has been developed by the Institut Pierre Simon Laplace (IPSL).

The download of files is achieved through multi-threaded exploration of the ESGF data repositories governed by the configuration files in which user defines the search criteria for the file selection. These criteria pertain to metadata attributes used for discovery of climate data. These attributes are defined by Data Reference Syntax (DRS). Thus, a user can enter lists of values for variables, frequencies, experiments and ensemble numbers (including wild-cards) into a configuration as directed by already provided templates. Before running the application, the user also needs to enter the ESGF user name and password that can be obtained by registering at any of the ESGF gateways, which are required for the data download.

Further details are available here: <https://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synchro-data>

Esgf-pyclient

Esgf-pyclient is software library for use in the Python programming language providing an interface to ESGF web services. The current release focuses on supporting ESGF's application programming interfaces (APIs) for login, search and download. Esgf-pyclient is designed to support the same "faceted search" workflow available through the ESGF web sites, only in a scriptable environment suitable for automation.

The library is targeted at advanced end users, developers and archive administrators. Through the library end users can automate complex searches to explore the archive catalogue. The integrated login API allows scripts using esgf-pyclient to obtain ESGF security credentials on demand. Once the user has constrained their search to their requirements they can extract download URLs or OPeNDAP service URLs to download or interact with the data further through python or external tools. Thus esgf-pyclient enables scripting of the complete workflow from data search, through download, to visualisation and analysis.

Developers can use esgf-pyclient in their applications to query the ESGF system so that they can provide value added services on top of ESGF data or domain specific views of the ESGF system. Archive administrators can use the library as a tool for archive maintenance and bulk information retrieval. Through esgf-pyclient statistics about the state of the archive can be collated including

number and location of replicas, data volumes, file counts and archive integrity information such as file checksums.

The library is maintained using an open development approach with online documentation, source code, and development process.

- Download from <http://pypi.python.org/pypi/esgf-pyclient>
- See the documentation at <http://esgf-pyclient.readthedocs.org>
- Follow development at <http://github.com/stephenpascoe/esgf-pyclient>

ESGF Download Script

The primary mechanism for users to access CMIP5 data starts with a search in the browser. After finding and selecting data, users may obtain a script which, run on their own computer, will transfer data to the local system. This script, developed at DKRZ for IS-ENES and implemented in all ESGF archive centres, deals with authorisation (passing the users credentials to the archive servers), repeated requests for files when transfers fail (a frequent occurrence when transferring large files over large distances), and verification of file check sums.

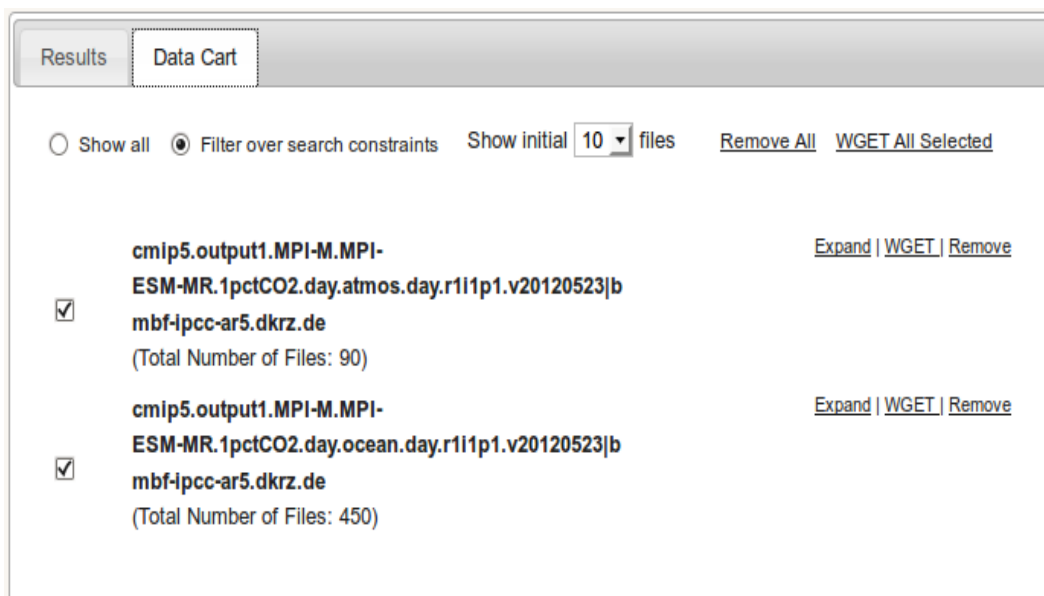


Illustration 1 View of the “data cart” generated by the ESGF portal – the “WGET” link will deliver a version of the ESGF download script to the user, with an embedded list of files taken from the search results.

OpenDAP – secured

The OpenDAP library provides a framework for direct access to remote data from a wide range of applications. This library, developed and maintained by UNIDATA, is now embedded in the UNIDATA NetCDF library¹ which is widely used in the climate science community. UNIDATA

¹ First available in netcdf-4.1.2-beta2 and first stable version in netcdf-4.1.3.

have adopted the PKI security configuration which IS-ENES co-developed with US ESGF partners. Consequently, 3rd party applications designed to read and manipulate NetCDF files using the UNIDATA libraries will automatically have the capability to read data directly from the CMIP5 archive through the embedded OpenDAP library.

A sample script showing how the header of a remote file can be viewed with “ncdump” and the data visualised with the “ncview” utility, without needing to transfer a complete file has been made available at <http://home.badc.rl.ac.uk/mjuckes/esgfNcview/>. These scripts have very limited functionality, but they demonstrate the ability to access the archive through 3rd party software tools.

1.2 Visualisation

Visualisation of data is provided through the interface developed in JRA5. JRA4 has enable the OpenDAP access to data to support this service. Further details are given in JRA5 reports (D11.4 Software Code and e-impact-portal full documentation). Visualization can be accessed through the climate4impacts portal on <http://www.climate4impact.eu/> . In order to support this JRA5 development work and the deployment of the prototype impacts portal an upgrade to the security infrastructure was designed and implemented, to support delegated authority (in order that the impacts portal could access data on behalf of users, rather than passing data directly to users).

Visualisation is also supported by the Live Access Server (LAS) described below.

1.3 Data manipulation and efficiency

OGC Web Processing Service

The Open Geospatial Consortium provides a global standard for web processing services. Adhering to these standards make the services accessible to a wide range of clients. The prototype service deployed within IS-ENES allows data from the CMIP5 archive to be transformed onto one of two standard grid resolutions (one and two degree grid squares in the current configuration). Within ExArch this service will be extended to provide ensemble means of the re-gridded data. The ENES service is an implementation of the CEDA OGC Web Services (COWS) system which is developed and maintained at STFC. IS-ENES has supported the configuration of the system for the CMIP5 archive.

The prototype regriding service is described at:

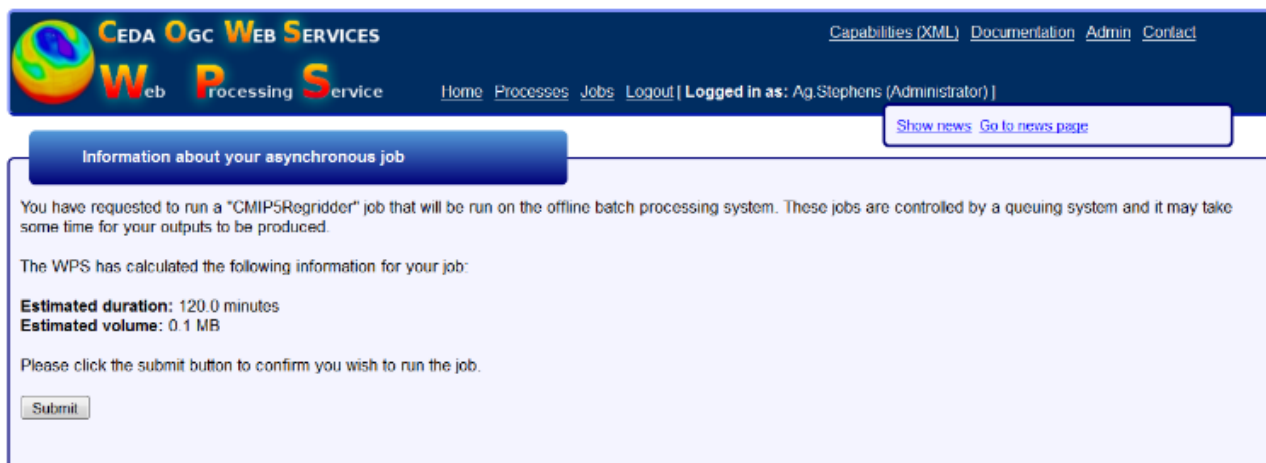
<http://proj.badc.rl.ac.uk/exarch/wiki/ExArchProcessing/CDOProcsForWPS/CMIP5Regridder>

The service split into two independent WPS components. The first allows the user to identify a set of files, and the second takes such a set of files, transforms them to a common grid and calculates the ensemble mean. This modularisation will allow both integration into other services which might provide a list of files by another means and chaining of processes by establishing a protocol through which files produced by one process can be fed into the re-gridder. The current prototype will work on files held at STFC but can be extended using OpenDAP to work on the distributed archive. However, a key feature of the processor is resource estimation and implementation of this feature with mixed local and remote data requires more work, including gathering of reliable information on network performance.

The regriding process is designed to run asynchronously, providing the user with a web page recording process. Before submission the user is given a resource estimation, indicating how long

the job is likely to run (Illustration 2).

During execution a job results page displays the job status and, on completion, provides links to output files (Illustration 3).



Information about your asynchronous job

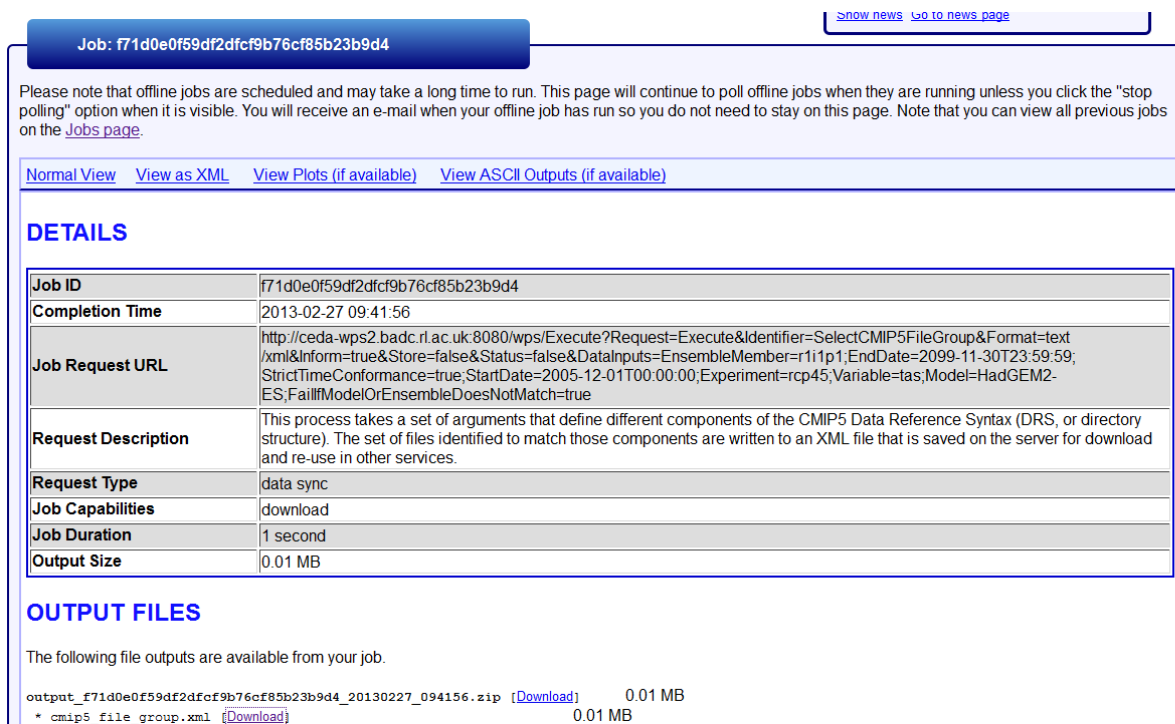
You have requested to run a "CMIP5Regridder" job that will be run on the offline batch processing system. These jobs are controlled by a queuing system and it may take some time for your outputs to be produced.

The WPS has calculated the following information for your job:

Estimated duration: 120.0 minutes
Estimated volume: 0.1 MB

Please click the submit button to confirm you wish to run the job.

Illustration 2: Resource estimation from the regridder



Job: f71d0e0f59df2dfcf9b76cf85b23b9d4

Please note that offline jobs are scheduled and may take a long time to run. This page will continue to poll offline jobs when they are running unless you click the "stop polling" option when it is visible. You will receive an e-mail when your offline job has run so you do not need to stay on this page. Note that you can view all previous jobs on the [Jobs page](#).

[Normal View](#) [View as XML](#) [View Plots \(if available\)](#) [View ASCII Outputs \(if available\)](#)

DETAILS

Job ID	f71d0e0f59df2dfcf9b76cf85b23b9d4
Completion Time	2013-02-27 09:41:56
Job Request URL	http://ceda-wps2.badc.rl.ac.uk:8080/wps/Execute?Request=Execute&Identifier=SelectCMIP5FileGroup&Format=text/xml&Inform=true&Store=false&Status=false&DataInputs=EnsembleMember=r1i1p1;EndDate=2099-11-30T23:59:59;StrictTimeConformance=true;StartDate=2005-12-01T00:00:00;Experiment=rcp45;Variable=tas;Model=HadGEM2-ES;FailIfModelOrEnsembleDoesNotMatch=true
Request Description	This process takes a set of arguments that define different components of the CMIP5 Data Reference Syntax (DRS, or directory structure). The set of files identified to match those components are written to an XML file that is saved on the server for download and re-use in other services.
Request Type	data sync
Job Capabilities	download
Job Duration	1 second
Output Size	0.01 MB

OUTPUT FILES

The following file outputs are available from your job.

output_f71d0e0f59df2dfcf9b76cf85b23b9d4_20130227_094156.zip [\[Download\]](#) 0.01 MB
 * cmip5_file_group.xml [\[Download\]](#) 0.01 MB

Illustration 3: Job report page, where user can monitor progress and access results when the process is completed.

Implementation of the Live Access Server

The Live Access Server (LAS) developed by NOAA and PMEL has been integrated into the ESGF software stack. This has been deployed at the IPSL data node for evaluation. LAS provides quick-looks and some sampling and analysis capabilities. LAS is also popular because it provides data export to several widely used format (CSV, ASCII, arcGrid). Every IPSL datasets has been republished to enable LAS access. IS-ENES has supported the configuration of the system and the publication steps. See for example:

http://vesg.ipsl.fr/las/localGetUI.do?dsid=2775C793BD2CBFA8D451089FC11EC4C6_ns_cmip5.output1.IPSL.IPSL-CM5A-MR.historical.mon.atmos.Amon.r3i1p1.v20120804&varid=cmip5.output1.IPSL.IPSL-CM5A-MR.historical.mon.atmos.Amon.r3i1p1.tas.20120804.aggregation.1-tas&auto=true

Illustrations 4-6 show example views of the data provided by the LAS server.

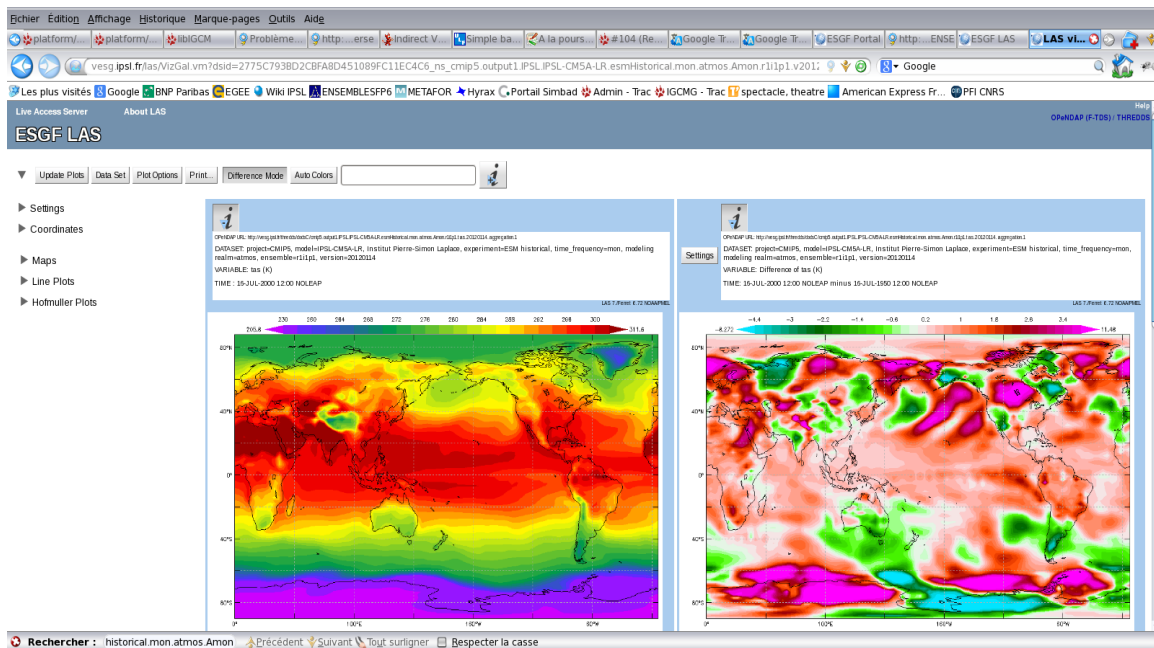


Illustration 4. LAS showing two a reference plot and an anomaly (“difference mode”).

In illustration 4, the server has been used to show a user selected field (on the left) and a difference plot between that and a 2nd user selected field on the right. Differencing is one of the most common user operations, allowing quick comparison of different fields. In illustration 5 a 3 way scatter plot is shown. The difference plot of illustration 4 can be used to identify spatial patterns in differences, but the scatter plot is more effective for users interested in statistical correlations and associated quantities. Finally, illustration 6 shows a CMIP5 data field viewed in Google Earth, after being exported by the LAS server. Export to Google Earth makes the data accessible to a range of user groups who do not wish to deal with formats and protocols favoured by the climate science community.

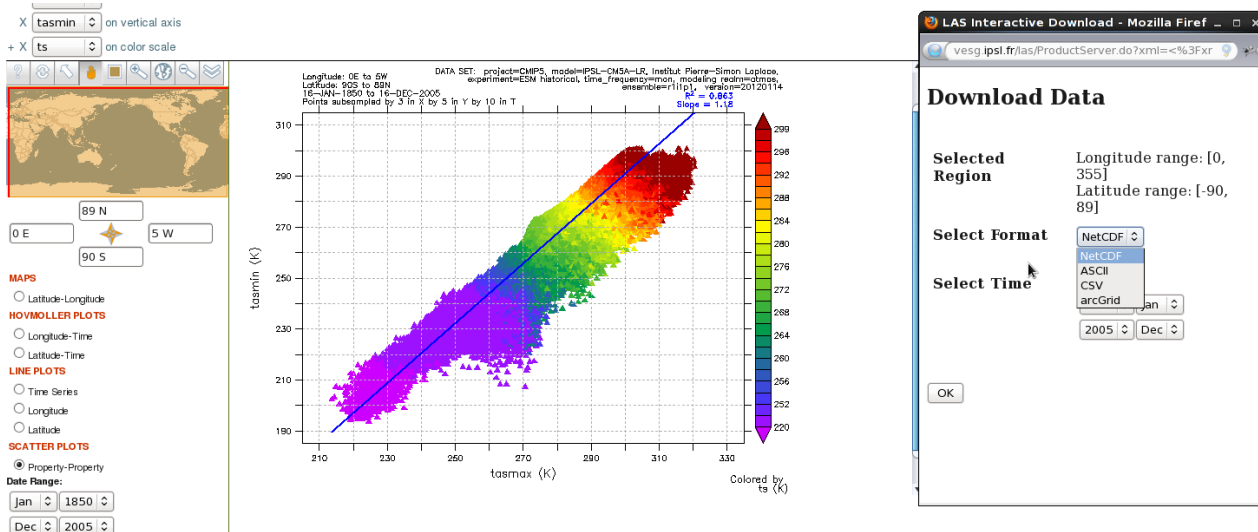


Illustration 5: Three way scatter plot (2 meter temperature minimum [y-axis], maximum [x-axis] and mean [colouring]) and data export possibilities

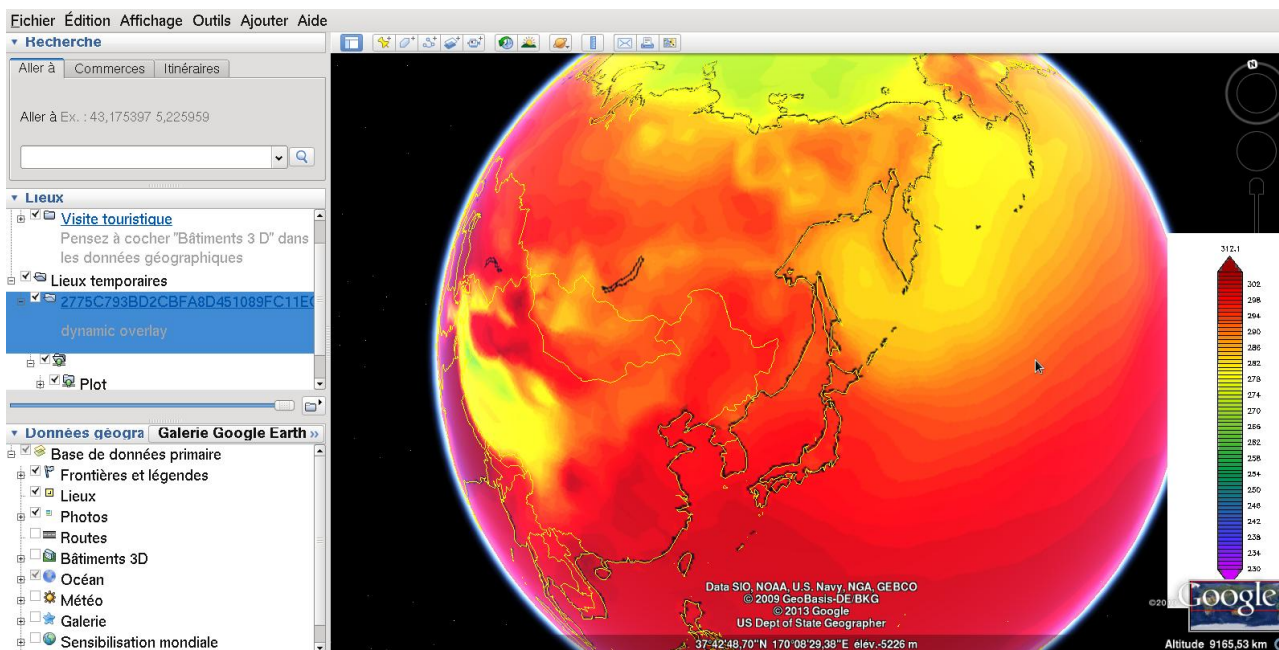


Illustration 6: Display of 2 meter temperature in July over Japan exported by LAS to Google Earth.

1.4 Support for regional modelling

At the start of the project it became clear that the WCRP CORDEX project (<http://wcrp-cordex.ipsl.jussieu.fr/>) would have a major influence on down-scaling work and that WP10/JRA4 resources for supporting regional modelling work would be used most productively if targeted to CORDEX activities. Consequently, the sub-task described in the description of work was re-focused on provision of key tools and protocols to support integration of CORDEX data into the ENES archive, as described below.

CORDEXwriter: Formatting data for the CORDEX archive

A set of scripts is designed to produce the CORDEX CORE and Tier 1 output in netcdf format.

All scripts are bash-shell scripts and use the CDO and NCO packages.

Input can be in NetCDF or GRIB format. There are 3 main scripts, dealing with daily monthly and seasonal data respectively. The scripts will form the data into the required time slices and ensure that global attributes and dimensions are correctly placed in the files. For some variables time averaging is also performed.

Configuring ESGF for CORDEX

The CORDEX project will produce regional climate projections down-scaled from the CMIP5 global projections. Managing the data from CORDEX requires some additional features in the archive configuration. Several meetings were held with CORDEX scientists, resulting in a clear statement of data requirements for the CORDEX regional climate model data. An ESGF data node was then configured to publish data complying with the requirements and test data published to the ESGF federation.

CORDEX “MIP” tables

“Model intercomparison project” tables (MIP tables) are a de-facto standard for describing the data to be archived in a model intercomparison project. The system has been developed by PCMDI, and the tables are used by the CMOR software developed and maintained at PCMDI. All CMIP5 data in the ENES archive has been produced in a standard file format using the CMOR tool. For CORDEX a more flexible approach has been adopted (by the CORDEX community) with a less rigorous file format specification. Nevertheless, key information is encoded in the CORDEX MIP tables:

http://www2-pcmdi.llnl.gov/cmor/tables/copy4_of_cmip5-tables/ – produced by IPSL.

2 Outlook

The CMIP5 archive has been evolving throughout the IS-ENES project, and corrections and additions are still being published. This has created a challenging environment for user software development. The tools and services here will need to be developed further to ensure that their potential is fulfilled.

3 Summary

The user services provide services and tools which give users flexible access to the data and enable

access by 3rd party software. The flexibility of the access control system opens up many possibilities for further developments. The link to the climate4impacts portal developed in JRA5 is one important illustration of this flexibility. The tools developed in this task support efficient selection and access to data, visualisation for easy data inspection, and processing of data before download to increase efficiency. Finally, specific support has been provided for the CORDEX regional modelling project since many important user groups cannot make direct use of data from Earth System Models and require downscaled products. The support provided for CORDEX will expedite the provision of a comprehensive range of downscaled products based on the CMIP5 climate projections.

4 ANNEX 1: Software libraries

Synchro-data (current version 2.5)	
Category	User tool; java library with command line interface
Location	https://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synchro-data
Documentation	https://forge.ipsl.jussieu.fr/prodiguer/wiki/docs/synchro-data
Licence	CeCILL license. Free software http://dods.ipsl.jussieu.fr/jripsl/synchro_data/LICENSE

ESGF-python	
Category	User library; python library
Location	http://pypi.python.org/pypi/esgf-pyclient
Documentation	http://esgf-pyclient.readthedocs.org
Licence	BSD

COWS CMIP5REGRIDDER	
Category	Service package; python
Location	Accessible through “easy_install” and “pip” tools
Documentation	http://proj.badc.rl.ac.uk/exarch/wiki/ExArchProcessing/CDOProcsForWPS/CMIP5Regridder , see also http://cows.badc.rl.ac.uk/cows_wps.html
Licence	BSD

ESGF Download Script	
Category	Bash script
Location	Generated by ESGF portals (see documentation).

Documentation	http://esgf.org/wiki/ESGF_scripting
Licence	Open source (unrestricted)

5 ANNEX 2: Glossary

Term	Description
ArcGrid	A widely used proprietary format for gridded data from ESRI.
ASCII	American Standard Code for Information Interchange character-encoding
BSD	Berkley Software Distribution
CECILL	CEA CNRS INRIA Logiciel Libre (http://www.cecill.info/)
CEDA	Centre for Environmental Data Archival (http://www.ceda.ac.uk)
CORDEX	Coordinated Regional climate downscaling experiment. (http://wcrp-cordex.ipsl.jussieu.fr/)
CSV	Comma-separated variables
DRS	Data Reference Syntax
ESGF	Earth System Grid Federation
ExArch	Climate analytics on distributed exascale data archives (http://proj.badc.rl.ac.uk/exarch)
LAS	Live Access Server
NetCDF	Network Common Data Form
NOAA	National Oceanographic and Atmospheric Administration
Obs4MIP	A pilot activity to make observational products more accessible for climate model inter-comparisons (http://obs4mips.llnl.gov:8080/wiki/ _
OGC	Open Geospatial Consortium
OPeNDAP	http://www.opendap.org/
PMEL	Pacific Marine Environmental Laboratory
UNIDATA	Unidata is a diverse community of over 250 institutions vested in the common goal of sharing data: http://www.unidata.ucar.edu
WPS	Web Processing Service