INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*CMCC*: SELECTED CASE STUDY FROM THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELING (ENES)

INPUT TO EU DELIVERABLE: D 2.1

Document identifier:	INDIGO-WP2-D2.1-ANNEX-1P0-V7
Date:	**27/05/2015**
Activity:	**WP2**
Lead Partner:	**EGI.eu**
Document Status:	**DRAFT**
Dissemination Level:	**CONFIDENTIAL (INTERNAL)**
Document Link:	

Abstract

This report summarizes the findings of T2.1 and T2.2 **for partner CMCC** along the first three months of the project. It is an integrated document including a general description of the research communities involved and the selected Case Studies proposed, in order to prepare deliverable D2.1, where the requirements captured will be prioritized and grouped by technical areas (Cloud, HPC, Grid, Data management) etc. The report includes an analysis of DMP (Data Management Plans) and data lifecycle documentation aiming to identify synergies and gaps among different communities.

I. COPYRIGHT NOTICE

Copyright © Members of the INDIGO-DataCloud Collaboration, 2015-2018.

II. DELIVERY SLIP

	Name	Partner/Activity	Date
From	Sandro Fiore, Giovanni Aloisio	**CMCC**/WP2	June 3, 2015
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco **Internal Reviewers:** <<To be completed by project office on submission to PMB>>		
Approved by	**PMB** <<To be completed by project office (no submission)>>		

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	5-may-2015	First draft, v01	J.Marco, F.Aguilar CSIC
2	7-may-2015	Initial feedback on structure from all partners	F.Aguilar CSIC, A.Bonvin UUtrecht
3	18-may-2015	Draft discussed in f2f meeting in Lisbon	P.Solagna, EGI.eu F.Aguilar, CSIC
4-7	28-may-2015	Draft ready for initial community input, to be iterated with JRA, v07	P.Solagna, EGI.eu J.Marco, F.Aguilar, CSIC, I.Blanquer UPV
8	4-june-2015	Draft after input from community, v08	JRA?
9	7-june-2015	Draft revised also with JRA, v09	P.Solagna, EGI.eu F.Aguilar, CSIC
10	10-june-2015	Draft to be circulated for internal review, v10	P.Solagna, EGI.eu
11	20-june-2015	Comments included, version for release v11	P.Solagna, EGI.eu

TABLE OF CONTENTS

0 INTRODUCTION AND CONVENTIONS

PLEASE, READ CAREFULLY BEFORE COMPLETING THE ANNEX:

*This Annex is an example of compilation of the information needed to support adequately a **Case Study** of interest in a Research Community. Each partner in INDIGO WP2 is expected to provide such information along the first three months of the project (i.e. by June 2015), and it will be used to compile Deliverable D2.1 on Initial Requirements from Research Communities.*

There will be around 10 Annexes, for example Annex 1.P1 for partner 1 in WP2 (i.e. UPV), will cover Case Studies from EuroBioImaging research community.

The initial version will be discussed with INDIGO Architectural team to agree on a list of requirements.

Some relevant definitions:

***A Case Study** is an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions.*

We should focus on Case Studies that are representative both of the research challenge and complexity but also of the possibilities offered by INDIGO-DataCloud solutions on it!

*The Case Study will be based on a set of User Stories, i.e. how the researcher describes the steps to solve each part of the problem addressed. **User Stories** are the starting point of **Use Cases**, where they are transformed into a description using software engineering terms (like the actors, scenario, preconditions, etc). Use Cases are useful to capture the Requirements that will be handled by the INDIGO software developed in JRA workpackages, and tracked by the Backlog system from the OpenProject tool.*

The User Stories are built by interacting with the users, and a good way is to do it in three steps (CCC): Card, Conversation and Confirmation[1].

Use Cases can benefit from tools like "mock-up" systems where the user can describe virtually the set of actions that implement the User Story (i.e. by clicking or similar on a graphical tool).

Different parts of this document should be completed with the help/input of different people:

RESEARCH MANAGERS

-Section 1, SUMMARY, is to be reviewed/agreed with them as much as possible

RESEARCHERS

*-Section 2, INTRODUCTION is designed to be filled with direct input from (senior) researchers describing the interest of the application, and written in such a way that it can be included in related technical papers. It is likely that such introduction is already available for some communities (for example, for several research communities in WP2 like DARIAH, CTA,EMSO, Structural Biology, one may start from the **Compendium of e-Infrastructure requirements for the digital ERA[2] from EGI***

APPLICATION DEVELOPERS AND INTEGRATORS WITHIN THE RESEARCH COMMUNITIES

-Sections 3, 4, 5, 6: should be discussed from their technical point of view (including data management as much as possible).

MIDDLEWARE DEVELOPERS AND E-INFRASTRUCTURE MANAGERS

-Sections 7, 8: should be discussed with them

[1] For a nice intro, see: https://whyarerequirementssohard.wordpress.com/2013/10/08/when-to-use-user-stories-use-cases-and-ieee-830-part-1/ , and also https://whyarerequirementssohard.wordpress.com/2015/02/12/how-do-we-write-good-user-stories/ etc.

[2] *https://documents.egi.eu/public/ShowDocument?docid=2480*

The logical order to fill the sections is: 2,3,4,5,6,1,7,8. Sections 1 and 8 will go into deliverable D2.1.

Other conventions and instructions for this document:

As this document/template is to be reused, the convention to use it as a questionnaire is that:

1) -text in italics provides its structure and questions,

2) -input/content should be written using normal text, replacing **<input here>**

Also the following conventions are used to identify the purpose of some parts of the questionnaire:

Bold text in blue corresponds to indications/suggestions to complete the questionnaire

Bold text in dark red marks technical issues particularly relevant that should be carefully considered for further analysis of requirements

Text in red indicates pending issues or ad-hoc warnings to the reader

1 EXECUTIVE SUMMARY ON THE CASE STUDY

Summarize the research community applications/plans/priorities (max length 2 pages).
To be completed after section 2 and reviewed later. Supervision by a senior researcher is required.

1.1 Identification

- *Community Name*: **ENES -** European Network for Earth System Modeling
- *Institution/partner representing the community in INDIGO:* CMCC
- *Main contact persons*: **Sandro Fiore, Giovanni Aloisio**
- *Contact email:* sandro.fiore@cmcc.it, giovanni.aloisio@cmcc.it
- *Specific Title for the Case Study:* Climate models intercomparison data analysis

1.2 Brief description of the Case Study and associated research challenge

Please include also a brief description of the community regarding this Case Study: partners collaborating, legal framework, related projects, etc.

Describe the research/scientific challenge that the community is addressing in the Case Study

The scientific community working on climate modelling is organized within the European Network for Earth System modelling (ENES)[3]. The institutions involved in this network include university departments, research centres, meteorological services, computer centres and industrial partners.

A major challenge for this community is the development of comprehensive Earth system models capable of simulating natural climate variability and human-induced climate changes. Such models need to account for detailed processes occurring in the atmosphere, the ocean and on the continents including physical, chemical and biological processes on a variety of spatial and temporal scales. They have also to capture complex nonlinear interactions between the different components of the Earth system and assess, how these interactions can be perturbed as a result of human activities.

The development and use of realistic climate models requires a sophisticated software infrastructure and access to the most powerful supercomputers and data handling systems. In this regard, the increased models resolution is rapidly leading to very large climate simulations output that pose significant scientific data management challenges in terms of data processing, analysis, archiving, sharing, visualization, preservation, curation, and so on[4,5,6].

[3] European Network for Earth System modelling - https://verc.enes.org/community/about-enes

[4] J. Dongarra, P. Beckman, T. Moore, P. Aerts, G. Aloisio, J. C. Andre, D. Barkai, J. Y. Berthou, T. Boku, B. Braunschweig, F. Cappello, B. M. Chapman, X. Chi, A. N. Choudhary, S. S. Dosanjh, T. H. Dunning, S. Fiore, A. Geist, B. Gropp, R. J. Harrison, M. Hereld, M. A. Heroux, A. Hoisie, K. Hotta, Z. Jin, Y. Ishikawa, F. Johnson, S. Kale, R. Kenway, D. E. Keyes, B. Kramer, J. Labarta, A. Lichnewsky, T. Lippert, B. Lucas, B. Maccabe, S. Matsuoka, P. Messina, P. Michielse, B. Mohr, M. S. Mueller, W. E. Nagel, H. Nakashima, M. E. Papka, D. A. Reed, M. Sato, E. Seidel, J. Shalf, D. Skinner, M. Snir, T. L. Sterling, R. Stevens, F. Streitz, B. Sugar, S. Sumimoto, W. Tang, J. Taylor, R. Thakur, A. E. Trefethen, M. Valero, A. van der Steen, J. S. Vetter, P. Williams, R. Wisniewski, K. A. Yelick: "The International Exascale Software Project roadmap". International Journal of High Performance Computing Applications (IJHPCA) 25(1): 3-60 (2011), ISSN 1094-3420, doi: 10.1177/1094342010391989.

In such a context, large scale global experiments for climate model intercomparison (CMIP) have led to the development of the Earth System Grid Federation (ESGF[7]) a federated **data infrastructure** involving a large set of data providers/modeling centres around the globe (the IS-ENES project provides the European contribution to the ESGF infrastructure). ESGF provides support for search & discovery, browsing and access to climate simulation data and observational data products.

An example, ESGF has been serving the Coupled Model Intercomparison Project Phase 5 (CMIP5) experiment, providing access to 2.5PB of data for the IPCC[8] AR5[9], based on consistent metadata catalogues. More specifically, the Coupled Model Intercomparison Project (CMIP) has been established by the Working Group on Coupled Modelling[10] (WGCM) under the World Climate Research Programme[11] (WCRP).

It provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation and data access. This framework enables a diverse community of scientists to analyse GCMs in a systematic fashion, a process that serves to facilitate models improvement.

Running a *climate models intercomparison data analysis* is very challenging, as it usually requires the availability of large amount of data (multi-terabyte order) from multiple climate models. Multiple classes of data analysis can be performed (e.g. trend analysis) as described in Section 2.1.

In the current scenario, these datasets have to be downloaded (e.g. from the ESGF data nodes) on the end-user's local machine before starting to run the analysis steps. This is a strong barrier for climate scientists, as this phase can take (depending on the amount of data needed to run the analysis) from days, to weeks, to months. The current client-side nature of the analysis workflow also needs end-users to have system management/ICT skills to install and update all the needed data analysis tools/libraries on their local machines. Another point relates to the complexity of the data analysis process itself. Analysing large datasets involves running multiple *data operators*, from widely adopted set of command line tools. This is usually done via scripts (e.g. bash) on the client side and also requires climate scientists to take care of, implement and replicate workflow-like control logic aspects (which are error-prone too) in their scripts - along with the expected application-level part.

The large volumes of data and the strong I/O requirements pose additional challenges related to performance. In this regard, production-level tools for climate data analysis are mostly sequential and there is a lack of solutions implementing fine-grain data parallelism or adopting stronger parallel I/O strategies, caching and data locality.

[5] European Exascale Software Initiative roadmap - http://www.eesi-project.eu/pages/menu/project/eesi-1/publications/final-report-recommendations-roadmap.php

[6] PRACE – The Scientific Case for High Performance Computing in Europe 2012-2020 - http://www.prace-ri.eu/IMG/pdf/prace_-_the_scientific_case_-_full_text_-.pdf

[7] Earth System Grid Federation - http://esgf.llnl.gov

[8] Intergovernmental Panel on Climate Change – http://www.ipcc.ch

[9] IPCC Fifth Assessment Report - https://www.ipcc.ch/report/ar5/

[10] Working Group on Coupled Modelling - http://www.wcrp-climate.org/wgcm/

[11] World Climate Research Programme - http://www.wcrp-climate.org/

1.3 Expectations in the framework of the INDIGO-DataCloud project

What do you think could be your main objectives to be achieved within the INDIGO project in relation to this Case Study?

The main objectives to be achieved within the INDIGO project in relation to the Case Study on "*climate models intercomparison data analysis*" are:

- a software framework deployable on heterogeneous infrastructures (e.g. HPC clusters and cloud environments) to run distributed, parallel data analysis;
- provisioning of efficient big data analysis solutions exploiting server-side and declarative approaches;
- an interoperable solution with regard to the already existing community-based software eco-system and infrastructure (IS-ENES/ESGF);
- the adoption of workflow management system solutions for large-scale climate data analysis;
- the exploitation of cloud computing technologies offering easy-to-deploy, flexible, elastic, isolated and dynamic big data analysis solutions;
- the provisioning of interfaces, toolkits and libraries to develop high-level interfaces/applications.

1.4 Expected results and derived impact

Describe the research results and impact associated to this Case Study.

The main research results and impacts associated to this Case Study are:

- the ability to deal in an easy manner with large scale, massive climate model intercomparison data analysis experiments;
- the opportunity to run complex data analysis workflows across multiple data centers, by also integrating well-known existing tools, libraries and command line interfaces;
- the possibility to strongly reduce the time-to-solution and complexity associated to this class of large-scale experiments;
- the possibility to address the re-use of final products, intermediate results and workflows.

1.5 References useful to understand the Case Study

Include previous reports, articles, and also presentations describing the Case Study

[1] Taylor K. E. , R. J. Stouffer G. A. Meehl : 2012 " An overview of CMIP5 and the experiment design" , Bulletin of the American Meteorological Society 93 , (4) , 485 - 498 , doi:10.1175/BAMS-D-11-00094.1 , http://journals.ametsoc.org/doi/abs/10.1175/BAMS-D-11-00094.1

[2] Luca Cinquini, Daniel J. Crichton, Chris Mattmann, John Harney, Galen M. Shipman, Feiyi Wang, Rachana Ananthakrishnan, Neill Miller, Sebastian Denvil, Mark Morgan, Zed Pobre, Gavin M. Bell, Charles M. Doutriaux, Robert S. Drach, Dean N. Williams, Philip Kershaw, Stephen Pascoe, Estanislao Gonzalez, Sandro Fiore, Roland Schweitzer: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems 36: 400-417 (2014).

2 INTRODUCTION TO THE RESEARCH CASE STUDY

Summarize the Case Study from the point of view of the researchers (max length 3 pages + table). Input by the research team in the community addressing the Case Study is required.

2.1 Presentation of the Case Study

Describe the Case Study from the research point of view

The case study on *climate models intercomparison data analysis* is directly connected to the Coupled Model Intercomparison Project (CMIP). CMIP studies output from 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. CMIP output is available for study by approved diagnostic sub-projects. 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. The CMIP5 experiment design has been finalized with the following suites of experiments: (i) Decadal Hindcasts and Predictions simulations, (ii) "long-term" simulations, and (iii) "atmosphere-only" (prescribed SST) simulations for especially computationally-demanding models.

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 differences in model projections, including quantifying some key feedbacks such as those involving clouds and the carbon cycle.

CMIP5 notably provides a multi-model context for 1) assessing the mechanisms responsible for model differences in poorly understood feedbacks associated with the carbon cycle and with clouds, 2) examining climate "predictability" and exploring the ability of models to predict climate on decadal time scales, and, more generally, 3) determining why similarly forced models produce a range of responses[12].

With specific regard to the CMIP* context, the Case Study will focus, in particular, on a specific set of data analysis. More specifically:

- Anomalies analysis
- Trend analysis
- Climate change signal analysis

Moreover, the output related to these three classes of data analysis will be considered as a basis for additional data analysis experiments, such as:

[12] CMIP5 - http://cmip-pcmdi.llnl.gov/cmip5/

- Tracking analysis (e.g. tropical cyclones, oceanic water masses)
- Transport analysis (e.g. Moc, oceanic transport, atmospheric transport, atmospheric rivers identification)

2.2 Description of the research community including the different roles

Please include a description of the scientific and technical profiles, and detail their institutions

Describe the research community specifically involved in this Case Study

The Case Study relates to the final step in the Earth System Modeling workflow, which is associated to the "Analysis by the Community". So stated, the large research community mainly involves climate change scientists (even including scientists working closely to the climate impact community), computational scientists, university researchers, and also students. Multiple research institutions worldwide have downloaded CMIP5 data to perform data analysis experiments. To give an idea about the real users, the ESGF federation today has more than 25K registered users and the CMIP experiments have generated many hundreds of peer-reviewed publications.

The main types of research institutions are: climate modeling/data centers, climate service centers, and universities. The institutions doing research on this topic are considerably a lot. Figure1 shows a geo-referenced map of the clients that have downloaded CMIP5 data from the CMCC node of the Earth System Grid Federation in the 2012-2014 timeframe. About 500TB of data were downloaded during the two years.

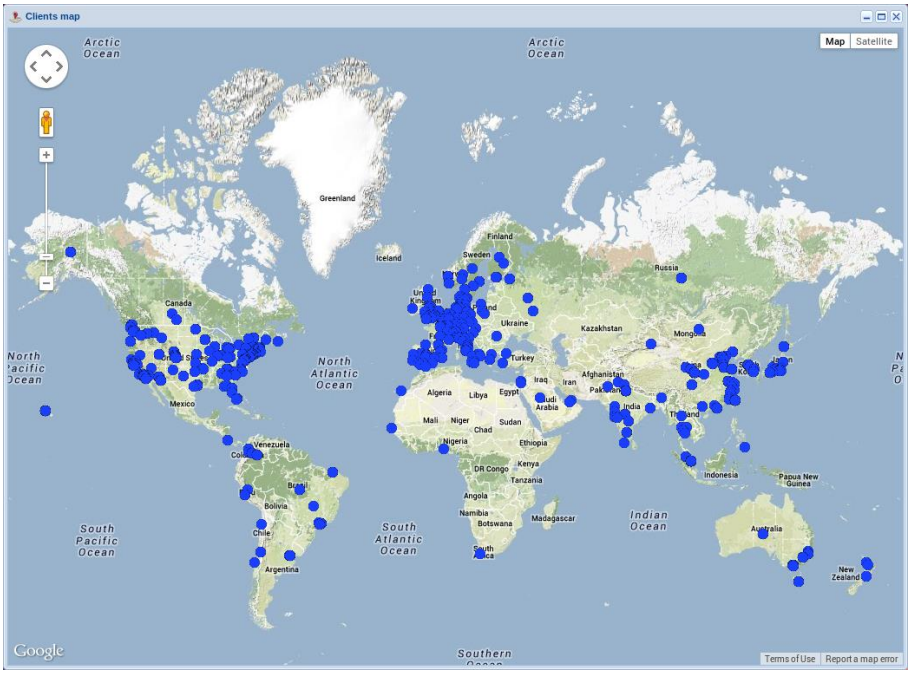

Figure1. Geo-referenced map of the clients that have downloaded CMIP5 data from the CMCC ESGF node in the 2012-2014 timeframe
Source: CMCC data statistics dashboard

2.3 Current Status and Plan for this Case Study

Please indicate if the Case Study is already implemented or if it is at design phase.

Describe the status of the Case Study and its short/mid term evolution expected

Throughout the entire project, the general "environment" of the Case Study will relate to: (i) data analysis inter-comparison challenges, (ii) addressed on CMIP5 data, (iii) which are made available through the IS-ENES/ESGF infrastructure.

The evolution of the Case Study in the short term (end of 2015) will be related to simple inter-comparison workflows facing specific classes of data analysis (see Section 2.1), by exploiting the output of a single climate model. That will provide preliminary insights about the feasibility of the approach. Data from a single data centre (e.g. CMCC) will be exploited to validate this phase.

In the mid term (October 2016), ensemble analysis will be added to the use cases in order to move forward in the research challenges of the Case Study (e.g. uncertainty assessment). Data from multiple models (just a few of them – 2 or 3 - from the same data centre) will be exploited to validate this phase.

In the long term (until the end of the project), the Case study will face data analysis challenges over a larger number of models, which will involve data from multiple data centres. This will represent the more mature phase of the Case Study and will address the general research challenges described in Section 2.1.

The scientific research community will be involved at each stage to provide feedback.

2.4 Identification of the KEY Scientific and Technological (S/T) requirements

Please try to identify what are the requirements that could make a difference on this Case Study (thanks to using INDIGO solutions in the future) and that are not solved by now.

Indicate which are the KEY S/T requirements from your point of view

The Case Study for this community has several requirements:

- Efficiency/Scalability. Running massive inter-comparison data analysis can be very challenging due to the large volume of the involved datasets (e.g. multi-terabyte order). There is a strong need to provide scalable solutions (e.g. HPC-, HTC-based) and different paradigms (e.g. server-side).
- Interoperability/legacy systems. There is a general eco-system for the scientific community that has be taken into account (e.g. existing data repositories, interfaces, security infrastructure, data formats, standards, specifications, tools, etc.). Interoperability with the existing ESGF/IS-ENES infrastructure is key.
- Workflow support. Data analysis inter-comparison experiments are based on multiple (e.g. tens/hundreds) data operators. Workflow tools could help managing the complexity of these experiments at different levels (multi-site and single-site) and increase the re-usability of specific workflow templates in the community.
- Metadata management. It represents a complementary (w.r.t to "data") aspect that must be taken into consideration both from a technical (e.g. metadata tools) and a scientific (e.g. data semantics) point of view.
- Easy to use analytics environments. Providing an easy-to-use and integrated analytics environment for climate model inter-comparison could represent an added value to enable scientific research at such large scale. From a technical point of view it also relates to having easy deployment procedures (e.g. cloud-based) to enable a larger adoption by the community.

- Flexible, elastic and dynamic environments. It must be considered that the data analysis workload can considerably vary over time (in this regard the CMIP* experiments are a very significant example).

2.5 General description of e-Infrastructure use

Please indicate if the current solution is already using an e-Infrastructure (like GEANT, EGI, PRACE, EUDAT, a Cloud provider, etc.) and if so what middleware is used. If relevant, detail which centres support it and what level of resources are used (in terms of million-hours of CPU, Terabytes of storage, network bandwidth, etc.) from the point of view of the research community.

Detail e-Infrastructure resources being used or planned to be used.

The e-Infrastructure resources used for this Case Study will consist of CMIP5 data provided by CMCC (CMCC publishes about 100TB of climate simulations datasets in the CMIP5 federated data archive related to the following three models: CMCC-CM, CMCC-CESM, CMCC-CMS).
CMCC will also provide an ESGF data node for testing activities and an already existing one (production-level) to run specific data challenges (adm07.cmcc.it). Additional computational (initially a cluster with 100 cores) and storage resources (about 100TB) will be provided to support the testing activities of the data analysis use cases. The mid term plan is also to try to get involved additional external sites from ESGF, as soon as a preliminary prototypes, software products will be available for deployment and testing.

2.6 Description of stakeholders and potential exploitation

Please summarize the potential stakeholders (public, private, international, etc.) and relate them with the exploitation possibilities. Provide also a realistic input to table on KPI.

The ENES community at the European level and in general the ESGF partners worldwide represents key stakeholders for INDIGO. Additional stakeholders that could be interested in the INDIGO outcomes are Space agencies (e.g. ESA) involved in climate change related initiatives (e.g. ESA CCI - Climate Change Initiative). It should be also considered that potential stakeholders are climate change consultancy agencies, politicians, educators, and also people from the private sector.

With particular regard to the presented Case Study, CMIP* related projects could be interested in exploiting the INDIGO software. CMIP5 is just a reference example in the climate change community, but the same approach could be applied to Observations for Model Inter-comparisons, which are related to observational products. Exploitation scenarios are mainly connected to providing data analysis functionalities at the data centre level (server-side), which go beyond the current federated or centralized data sharing and access facilities already existing and today available in production (e.g. ESGF).

The provided expected impact in the table is based on a **conservative** approach.

Area	***Impact Description***	***KPI Values***
Access	*Increased access and usage of e-Infrastructures by scientific communities, simplifying the "embracing" of e-Science.*	• *Number of ESFRI or similar initiatives adopting advanced middleware solutions* ESFRIs: **ESGF/IS-ENES** • *Number of production sites supporting the software* **at least 1 in INDIGO (CMCC), but we'll try to reach 3. Some external institutions from ESGF will be also contacted to**

		test the software
Usability	*More direct access to state-of-the art resources, reduction of the learning curve. It should include analysis platforms like R-Studio, PROOF, and Octave/Matlab, Mathematica, or Web/Portal workflows like Galaxy.* *Use of virtualized GPU or interconnection (containers).* *Implementation of elastic scheduling on IaaS platforms.*	• *Number of production sites running INDIGO-based solutions to provide virtual access to GPUs or low latency interconnections* **from at least 1 (CMCC), to 3** • *Number/List of production sites providing support for Cloud elastic scheduling* **we'll target from 2 to 3** • *Number of popular applications used by the user communities directly integrated with the project products:* **5-10** • *Number of research communities using the developed Science Gateway and Mobile Apps:* **Climate change scientists. Potential interest from the Space agencies/community** • *Research Communities external to INDIGO using the software products*: **Climate change scientists. Potential interest from the Space agencies/community**
Impact on Policy	*Policy impact depends on the successful generation and dissemination of relevant knowledge that can be used for policy formulation at the EU, or national level.*	• *Number of contributions to roadmaps, discussion papers:* **2**
Visibility	*Visibility of the project among scientists, technology providers and resource managers at high level.*	• *Number of press releases issued*: **1 per year** • *Number of download of software from repository per year:* **5-10-20** • *List of potential events/conferences/workshops:* ***1) ESGF Annual Meeting (generally planned in December, Livermore, CA, USA).*** ***2) European Geosciences Union (generally planned in April, Vienna, Austria. Venue = Austria Center Vienna (ACV) http://www.egu2015.eu/*** ***3) American Geophysical Union (planned in December, S. Francisco, USA. Venue = Moscone Center) http://fallmeeting.agu.org/2015/*** ***4) ESA workshops (e.g. http://www.eoscience20.org/)*** ***5) GO-ESSP workshops (http://go-essp.gfdl.noaa.gov/)*** ***6) Events organized by EU projects closely related to the climate change domain (e.g. IS-ENES, EUBrazilCC)*** ***7) Research Data Alliance (RDA). Next even planned in Paris 23-25 September, 2015. https://rd-alliance.org/***

		• *Number of domain exhibitions attended* **10** • *Number of communities and stakeholders contacted* **5-10**
Knowledge Impact	*Knowledge impact creation: The impact on knowledge creation and dissemination of knowledge generated in the project depends on a high level of activity in dissemination to the proper groups.*	• *Number of journal publications:* **3-5 referencing INDIGO** • *Number of conference papers and presentations*: **20-30 referencing INDIGO**

Table 1 Key Performance Indicators (KPI) associated to different areas. Add in this table how your community would contribute to the KPIs. Note: this table will NOT be included in the deliverable.

3 TECHNICAL DESCRIPTION OF THE CASE STUDY

Describe the Case Study from the point of view of developers (4 pages max.)
Assemble it using preferably an AGILE scheme based on User Stories.

3.1 Case Study general description assembled from User Stories

Please describe here globally the Case Study. If possible use as input "generic" User Stories built according to the scheme: short-description (that fits in a "card") + longer description (after "conversation" with the research community). Provide links to presentations in different workshops describing the Case Study when available. Include schemes as necessary.

Describe the Case Study showing the different actors and the basic components (data, computing resources, network resources, workflow, etc.). Reference relevant documentation.

To address the Case Study challenges several classes of data analysis will be addressed (see Section 2.3). To pursue this objective, an incremental approach will consider:

- simple inter-comparison workflows exploiting the output of a single climate model (single model, single data center).
- workflows involving both inter-comparison and ensemble analysis to address new research challenges of the Case Study (e.g. uncertainty assessment). Data from 2-3 models (from the same data centre) will be considered at this stage.
- complex workflows involving data from a larger set of models (this will involve datasets from multiple data centres).

Scientists will be able to create, re-use/adapt and submit analysis workflows using a command line and/or a graphical interface. For stage 1 and 2, the data repository and the computing resources to run the analysis will be provided by CMCC (this also includes a private cloud environment for testing the *cloud-based* solutions provided by INDIGO as well as some HPC nodes). For stage 2 and 3 additional sites will be considered; in these two cases the scientists will be also able to (i) perform an outlier analysis on the models ensemble, (ii) identify and remove "outliers", and (iii) re-run the workflow on the reduced set of models. For stage 1, 2 and 3 single and multiple variables selection will be considered. The experiments could include both analysis and visualization tasks.

The actors involved in the Case Study are: (i) end-users, (ii) climate scientists running the simulations and generating the data (they could answer questions about the scientific aspects of the models used in the simulation runs and the output data) and (iii) IT administrators (at the data centre level) that manage (e.g. store, publish) the datasets (they could support end users about the technical aspects, like data availability, network issues, downtime, etc.).

3.2 User categories and roles

Describe in more detail the different user categories in the Case Study and their roles, considering in particular potential issues (on authorization, identification, access, etc.)

Two main end-user categories could be involved in the case study:

- Expert users (e.g. climate scientists). Researchers that access to the platform and data. They should be able to access to the resources of the infrastructure run their analysis. AuthN and AuthZ aspects should be properly addressed.

- Non-expert users. In this case specific training datasets and demo accounts should be made available for allowing external people to have a better understanding about the proposed solutions.

3.3 General description of datasets/information used

List the main datasets and information services used (details will be provided in next section)

The main datasets are made available for the project purposes by CMCC (100TB of climate simulations datasets in the CMIP5 federated data archive, related to the three models CMCC-CM, CMCC-CESM, CMCC-CMS). Additional ones could be related to new ESGF sites willing to test the INDIGO solutions.

3.4 Identification of the different Use Cases and related Services

Identify initial Use Cases based on User Stories, and describe related (central/distributed) Services

With specific regard to the CMIP* context, the Case Study will focus, in particular, on a specific set of data analysis use cases. More specifically:

- **Anomalies analysis**: anomalies are defined as an incidence or occurrence when the actual result under a given set of assumptions is different from the expected result. An anomaly provides evidence that a given assumption or model does not hold in practice.
- **Trend analysis**: analyzing data trends is an age old and powerful tactic, which is used to measure the performance of marketing campaigns over time and to predict future outcomes. Trend Analysis is the practice of collecting information and attempting to spot a pattern, or trend, in the information. Although trend analysis is often used to predict future events, it could be used to estimate uncertain events in the past.
- **Climate change signal analysis**: The treatment of "signal" and "noise" in constructing climate scenarios is of great importance in interpreting the results of impact assessments that make use of these scenarios. If climate scenarios contain an unspecified combination of signal plus noise, then it is important to recognise that the impact response to such scenarios will only partly be a response to anthropogenic climate change; an unspecified part of the impact response will be related to natural internal climate variability.

All the three aforementioned classes of data analysis are strongly related to model inter-comparison and will involve the access to one or more (distributed) data repositories (e.g. managed by IS-ENES/ESGF data nodes), as well as running complex analytics workflows with tens/hundreds of data operators. Intra-data and inter-data centre aspects will need to be properly considered. Specific workflows will be designed jointly with user community experts to address all the relevant scientific challenges and reproduce real experiments. Workflows will also include the execution of already existing tools (e.g. CDO, NCO, NCL, Grads, etc.) widely adopted by the community (e.g. for data processing and visualization). New interfaces will be also needed to provide end-users with a proper environment for running climate data analysis experiments.

It is important to state that, the output related to these three classes of data analysis will be considered as a basis for studying additional data analysis experiments, related to:

- *Tracking analysis* (e.g. tropical cyclones, oceanic water masses).

- *Transport analysis* (e.g. Moc, oceanic transport, atmospheric transport, atmospheric rivers identification).

Architecture and Services

The ESGF architecture is based on a system of autonomous and distributed Nodes, which interoperate through common acceptance of federation protocols and trust agreements. Data is stored at multiple Nodes, and served through local data and metadata services. Nodes exchange information about their data holdings and services, trust each other for registering users and establishing access control decisions. The net result is that a user can use a web browser or rich desktop client, connect to any Node, and seamlessly find and access data throughout the federation[13].

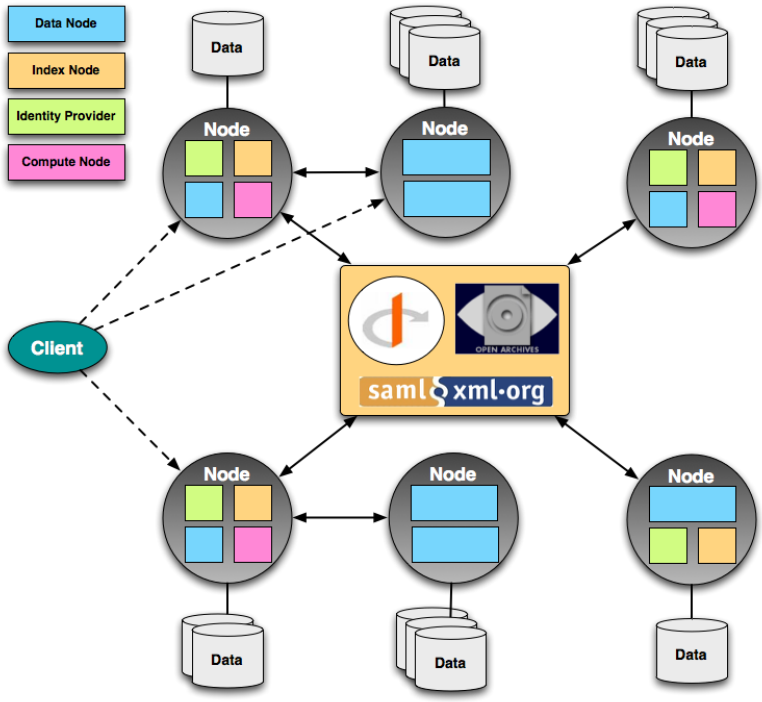

Figure2. ESGF Federation Design.
Source: http://esgf.llnl.gov/media/images/FederationArchitecture.png

The ESGF services are logically grouped in four areas of functionality, which determine the Node ‘‘flavors’’:

- *Data node*: is a collection of open source components packaged and developed as part of the ESGF initiative to provide basic data access functionality (HTTP, OPeNDAP) associated to metadata catalogues (Thredds). Its main components are the data Publisher application that generates the metadata catalogs, the THREDDS and GridFTP servers (with security filters at the front end) to serve the data, and the OpenID Relying Party and Authorization Service to ensure proper authentication and authorization.
- *Identity Provider*: it is a component providing User Authentication service (OpenID based federation wide authentication). It allows user authentication and secure delivery of user attributes. It includes an OpenID Provider for browser-based authentication, a MyProxy server

[13] ESGF Federation Design - http://esgf.llnl.gov/federation-design.html

for requesting limited-lifetime user certificates used for non-browser based access, and User Registration and Attribute Services for distributed access control. In the ESGF context, there is a working group (IdEA - "*Identity Entitlement Access*"), which is discussing and addressing security challenges.

- *Index nodes*: an Apache Solr based search index provides the indexing functionalities needed to enable the search and discovery of scientific datasets.
- *Compute node*: it is the one in the architectural design that will be devoted to providing processing capabilities. It contains higher-level services for data analysis and visualization. Currently its only components are the Live Access Server, the Ferret-THREDDS Data Server and the Ferret engine. Also in this case, there is an ESGF working group (CWT - "*Compute Working Team*"), which is discussing and proposing the interfaces for server-side processing/analysis engines.

3.5 Description of the Case Study in terms of Workflows

Summarize the different Workflows within the Case Study, and in particular Dataflows. Include the interaction between Services.

In the following, the main steps related to a general workflow example for our Case Study:

1. **Experiment definition**: starting from a user interface (graphical or command-line) the climate scientist should be able to choose/define a specific type of data analysis. In this regard, an associated data analysis workflow could be: (i) either selected from a repository (addressing re-usability), customized and re-used, or (ii) composed on the fly by the user (and then – eventually - stored in the workflows repository for further re-use). It should include the entire workflow description (a detailed tasks and dependencies definition related to the expected data analysis process). Input parameters are provided at this stage.
2. **Experiment run**: The data analysis workflow should be submitted to the infrastructure. Computational/storage resources should be allocated for the data analysis taking into special account data locality aspects. Access to the data from multiple data centres as well as reduction tasks could be required for multi-model ensemble analysis. The data analysis tasks produce intermediate data as well as final products. Workflow solutions would be strongly needed to support running these experiments, jointly with tasks monitoring capabilities.
3. **Results access, visualization, and publication**: The results should be made easily available to the end-user through a dedicated interface for download, visualization and possibly further analysis. It should be also possible to publish the results of a specific analysis on dedicated catalogues. The user interface should provide analytics, exploration and visualization capabilities. To this end, already existing and well-known tools in the community should be integrated in the general eco-system.

Security aspects related to AuthN and AuthZ should be also part of the workflow.

3.6 Deployment scenario and relevance of Network/Storage/HTC/HPC

Indicate the current deployment framework (cluster, Grid, Cloud, Supercomputer, public or private) and the relevance for the different Use Cases of the access to those resources.

In the current configuration the ESGF/IS-ENES infrastructure provides a large-scale, federated and production-level data sharing facility. Data analysis is mainly performed by the end-users on their own environment (e.g. desktop machines, supercomputers, etc.).

The future deployment framework should support at the data centre level analysis features. In such a context, the different use cases will mainly rely on two scenarios:

- single-site: the data analysis experiment could run at a single site (e.g. CMCC) providing both HPC and private cloud facilities.
- multi-site: the data analysis experiment could run at multiple sites (as a global, distributed experiment) by analysing datasets from several models.

Sites from WP3 could be also involved to extend the testbed and reproduce a real, geographically distributed environment.

4 DATA LIFE CYCLE

INDIGO-DataCloud is a DATA oriented project. So the details provided in this complex section are KEY to the project. Please try to be as complete as possible with the relevant information.

As already stated in previous sections, the Case Study relates to the final step in the Earth System Modelling workflow, which is associated to the "*Analysis by the Community*" (see Figure 3)

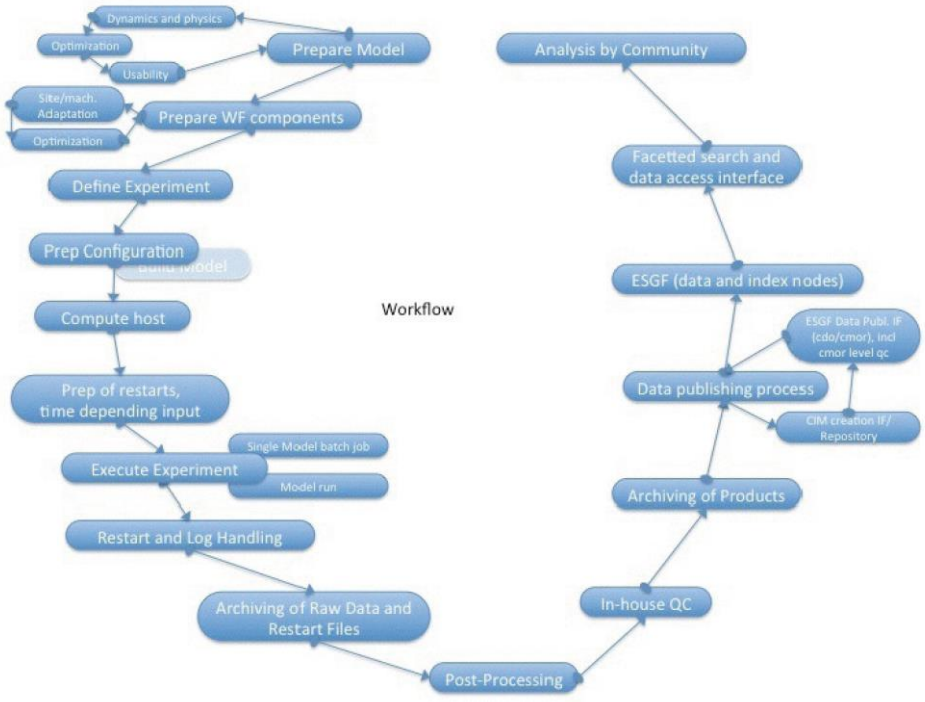

Figure3. Earth System Modelling Workflow
Source: "ISENES2 Workshop on Workflow Solutions in Earth System Modelling", by Reinhard Budich (Strategic IT Partnerships Scientific Computing Lab MPI-M) and Kerstin Fieg (Applications Deutsches Klimarechenzentrum DKRZ).
June 3-5 2014, DKRZ, Hamburg.

The datasets used in the case study will relate to the CMIP5 experiment. In particular 100TB of data will be available from CMCC models output.

4.1 Data Management Plan (DMP) for this Case Study

According to EU H2020 indications[14], following UK DCC tool indications

[14] *In Horizon 2020 a limited pilot action on open access to research data will be implemented. Projects participating in the Open Research Data Pilot will be required to develop a Data Management Plan (DMP), in which they will specify what data will be open. Other projects are invited to submit a Data Management Plan if relevant for their planned research. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project. The first version of the DMP is expected to be delivered within the*

4.1.1 Identification of the DMP

4.1.2 DMP at initial stage (to be prepared before data collection)

4.1.3 DMP at final stage (to be ready when data is available)

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

NetCDF

• Why have you chosen to use particular formats?

This is the format used by the community in the CMIP5 experiment.

• Do the chosen formats and software enable sharing and long-term validity of data?

Yes, they do. There is a continuous effort by the community on conventions, formats, vocabularies, which aims at giving long-term validity to the data.

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc?, e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins

Yes, they are. In particular the NetCDF format, CF conventions[15] allow data exchange between researchers, institutions, organisations, countries.

The conventions for CF (Climate and Forecast) metadata are designed to promote the processing and sharing of files created with the NetCDF API. The CF conventions are increasingly gaining acceptance and have been adopted by a number of projects and groups as a primary standard. The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities[16].

---

first 6 months of the project. More elaborated versions of the DMP can be delivered at later stages of the project. The DMP would need to be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. The templates provided for each phase are based on the annexes provided in the Guidelines on Data Management in Horizon 2020 (v.1.0, 11 December 2013).

[15] http://cfconventions.org/Data/cf-documents/overview/article.pdf

[16] CF Conventions and Metadata - http://cfconventions.org/

The CF conventions generalize and extend the COARDS conventions[17].

4.2 Data Levels, Data Acquisition, Data Curation, Data Ingestion

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study:

Examples of analysis are: anomalies analysis, trend analysis, and climate change signal analysis.

Specify if software packages/tools like MATLAB, R-Studio, iPython,etc. are used:

As the analysis is usually performed on the client side, scientists work with a lot of different tools, with complementary features, and with several of them (for the same analysis) to get more insights from the output of an experiment.
iPython is one of the tools used by the researchers in the community.

4.3.2 Data analytics and Big Data

Describe relevant examples of advanced analysis in the Case Study (like for example application of neural networks, series analysis, etc.)

Some examples are: intercomparison, statistical, and ensemble analysis. They could be part of the same experiment/workflow to analyse extreme events, anomalies, etc. Ensemble analysis is particularly relevant due to the scientific aspects related to the uncertainty assessment. It often includes additional steps like re-gridding as the output of climate simulations of different models are usually related to different grids/resolutions.

Specify the resources and additional software required:

Several tools, libraries, command line interfaces, and frameworks can be exploited to run an analysis experiment. CDO, NCO, NCL, Grads, and Ophidia provide support for analysis, processing, visualization, and big data analytics. There are also a variety of libraries in Python language, which allows further data analysis and visualization capabilities.

Identify analysis challenges that can be classified as "Big Data"

Analysis challenges relate to managing large volumes of data, run I/O intensive data analysis, and complex workflows with potentially tens/hundreds of operators. These aspects have been also covered in other section of the document.

[17] http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html

List Big Data driven workflows:

Anomalies analysis, trend analysis, climate change signal analysis are some types of analysis regarding climate change intercomparison. We can associate specific workflows to each of them. These workflows will include, for instance, operators to perform: sub-setting, statistical analysis, data intercomparison, ensemble analysis, outlier analysis. Percentile calculation (e.g. for heavy precipitation, >90p of precipitation for heavy events, >99p of precipitation for very intense events) could be another example.

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization

Data visualization is an important task in the case study as it allows a visual inspection of the final result of an experiment.

Indicate how visualization is made and if interactivity/steering is needed

Visualization is performed for instance through specific command line tools like NCL. Interactivity is an interesting point, but it is not mandatory in the considered case study.

Specify the User Interfaces (web, desktop, mobile, etc.)

User interfaces are mainly desktop and web-based. Simplified examples could also run on mobile applications.

4.4 Data Publication

Describe the information flow from the analysis to the publication

The output of a data analysis experiment could be published for further re-use, exploitation, download. It should be compliant with the adopted conventions (e.g. CF) and formats (NetCDF). It could be published under simple HTTP server, Opendap/Thredds services. However it would be relevant to link the workflow to the ouput (and possibly the papers related to the scientific result), in order to provide more insights about the experiment behind the final result and address strongly the experiment reproducibility. This is a step not implemented yet in production-level environments like ESGF/IS-ENES, but of great relevance.

Indicate the requirements from publishers/editors to access data, and how it is made available (open data?)

It could be very interesting to make available the results of the data analysis experiments to the broader community. Indeed, it would also face key challenges related to open data and a better dissemination of the data analysis results (linking papers to analysis workflows and final output, targeting “executable papers” and so data analysis reproducibility). That would help improving the way scientific information is communicated and used.

5 DATA INTENSIVE COMPUTING

Describe the Simulation/Modelling requirements in this Case Study. Please identify also any other intensive CPU mainly activity as required.

5.1 General description of data intensive needs

A generic coupled model intercomparison data analysis experiment could involve/incorporate a large set of data operators into a single workflow (from tens to hundreds).

Some operators examples include: selection, comparison, metadata management (CRUD - Create-Read-Update-Delete), arithmetic and statistical operators, regression, etc. A more detailed description is provided in the following sections related to the software/tools used by the community.

5.2 Technical description of data analysis and visualization tools/software

There are several tools and libraries for running data analysis. In the following section we briefly describe: (i) the Climate Data Operators (CDO) and (ii) NCO for data analysis and (iii) NCL for data visualization. They are well known and largely exploited by the community.

1)

Identify the simulation software

Climate Data Operators (CDO)[18]

CDO is a collection of command line Operators to manipulate and analyse Climate and NWP model Data. NetCDF 3/4, GRIB 1/2 including SZIP and JPEG compression, EXTRA, SERVICE and IEG are supported as IO-formats. There are more than 600 operators available.

As reported in the project web site[19], there are operators for the following topics:

- File information and file operations
- Selection and Comparison
- Modification of meta data
- Arithmetic operations
- Statistical analysis
- Regression and Interpolation
- Vector and spectral Transformations
- Formatted I/O
- Climate indices

Provide a link to its documentation, and describe its maturity and support level

https://code.zmaw.de/projects/cdo/, mature, widely adopted by the community.

[18] CDO 2015: Climate Data Operators. Available at: http://www.mpimet.mpg.de/cdo

[19] CDO - https://code.zmaw.de/projects/cdo/wiki/Cdo#Documentation

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) –

Tag the simulation software as HTC or HPC: -

HTC data analysis experiments (parameter sweep) could be implemented.

OpenMP support for CDO is also available. As reported in the documentation page[20]: "Some of the **CDO** operators are shared memory parallelized with OpenMP. An OpenMP-enabled C compiler is needed to use this feature. Users may request a specific number of OpenMP threads nthreads with the '-P' switch. Many **CDO** operators are I/O-bound. This means most of the time is spend in reading and writing the data. Only compute intensive **CDO** operators are parallelized."

List the input files required for execution and how to access them

Command line tool, files are provided in input as arguments. NetCDF 3/4, GRIB 1/2 including SZIP and JPEG compression, EXTRA, SERVICE and IEG are supported as IO-formats.

Describe the output files and how they will be stored

NetCDF 3/4, GRIB 1/2 including SZIP and JPEG compression, EXTRA, SERVICE and IEG are supported as IO-formats.

Reference an existing installation and performance indicators -

Specify if the simulation software is parallelized (or could be adapted):

Sequential software

Specify if the simulation software can exploit GPUs -

Specify how the simulation software exploits multicore systems -

Specify if parametric runs are required -

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) -

2)

Identify the simulation software

NetCDF Operators (NCO)

The NCO toolkit manipulates and analyzes data stored in netCDF-accessible formats, including DAP, HDF4, and HDF5. It exploits the geophysical expressivity of many CF (Climate & Forecast) metadata conventions, the flexible description of physical dimensions translated by UDUnits, the network transparency of OPeNDAP, the storage features (e.g., compression, chunking, groups) of HDF (the

[20] https://code.zmaw.de/projects/cdo/wiki/OpenMP_support

Hierarchical Data Format), and many powerful mathematical and statistical algorithms of GSL (the GNU Scientific Library)[21].

NCO utilities include:

- ncap2 netCDF Arithmetic Processor
- ncatted netCDF ATTribute EDitor
- ncbo netCDF Binary Operator (addition, multiplication...)
- nces netCDF Ensemble Statistics
- ncecat netCDF Ensemble conCATenator
- ncflint netCDF FiLe INTerpolator
- ncks netCDF Kitchen Sink
- ncpdq netCDF Permute Dimensions Quickly, Pack Data Quietly
- ncra netCDF Record Averager
- ncrcat netCDF Record conCATenator
- ncrename netCDF RENAMEer
- ncwa netCDF Weighted Averager

NCO utilities have as a goal being as generic as possible, imposing no limitations on data dimensionality, size, or type.

Provide a link to its documentation, and describe its maturity and support level

NCO - http://nco.sourceforge.net/ mature and widely adopted by the community.

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) –

Tag the simulation software as HTC or HPC: -

NCO is a sequential tool. HTC data analysis experiments (parameter sweep) could be implemented.

List the input files required for execution and how to access them

Command line tool, files are provided in input as arguments. File format NetCDF/HDF, DAP. Output in text, binary, or netCDF formats.

Describe the output files and how they will be stored

Output files are NetCDF/Grib format

Reference an existing installation and performance indicators -

Specify if the simulation software is parallelized (or could be adapted):

Sequential software

[21] NetCDF Operators (NCO) - http://nco.sourceforge.net/

Specify if the simulation software can exploit GPUs -
Specify how the simulation software exploits multicore systems -
Specify if parametric runs are required -
Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) -

3)
Identify the simulation software
NCAR Command Language (NCL)
NCL is an interpreted language designed specifically for scientific data analysis and visualization. Portable, robust and free, NCL is available as binaries or open source. It supports NetCDF 3/4, GRIB 1/2, HDF 4/5, HDF-EOS 2/5, shapefile, ASCII, binary. Numerous analysis functions are built-in. High-quality graphics are easily created and customized with hundreds of graphic resources[22].

Provide a link to its documentation, and describe its maturity and support level
NCL - http://www.ncl.ucar.edu/ - mature, widely adopted by the community.

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) –

Tag the simulation software as HTC or HPC: -
CDO is a sequential tool. HTC data analysis experiments (parameter sweep) could be implemented.

List the input files required for execution and how to access them
Command line tool, files are provided in input as arguments. It supports NetCDF 3/4, GRIB 1/2, HDF 4/5, HDF-EOS 2/5, shapefile, ASCII, binary. A complete documentation about the supported data format is available at[23]

Describe the output files and how they will be stored
A complete documentation about the supported data format is available at[24]

Reference an existing installation and performance indicators -

Specify if the simulation software is parallelized (or could be adapted):
Sequential software

[22] NCAR Command Language (NCL) - http://www.ncl.ucar.edu/
[23] https://www.ncl.ucar.edu/Document/Manuals/Ref_Manual/NclFormatSupport.shtml
[24] https://www.ncl.ucar.edu/Document/Manuals/Ref_Manual/NclFormatSupport.shtml

Specify if the simulation software can exploit GPUs -
Specify how the simulation software exploits multicore systems -
Specify if parametric runs are required -
Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) -

4)
Identify the simulation software
Ophidia – big data analytics framework[25]
The Ophidia framework provides a high performance data analytics solution. It targets high performance (tightly coupled) data analysis through workflows orchestration on its big data analytics framework. With regard to existing solutions for data analysis, it exploits a different paradigm (server-side) and provides workflow and parallel I/O support to properly address performance. It joins inter-task and intra-task parallelism combining HTC and HPC paradigms. Ophidia can run external solutions for data analysis and visualization thanks to a specific operator addressing loosely coupled integration of legacy software. As an example Ophidia could run a workflow integrating CDO and NCO operators, Ophidia native operators and the NCL commands for visualization.

Provide a link to its documentation, and describe its maturity and support level
Ophidia - http://ophidia.cmcc.it/ - novel big data framework for parallel data analysis. Supported by CMCC.

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) –
Ophidia implements a stack that can run both on a single node, and on a cluster (HPC or commodity).

Tag the simulation software as HTC or HPC: -
Ophidia includes both parallel and sequential operators. It combines both HPC and HTC approaches.

List the input files required for execution and how to access them
Command line tool, files are provided in input as arguments. NetCDF is mainly supported as I/O format.

Describe the output files and how they will be stored
NetCDF is mainly supported as I/O format.

Reference an existing installation and performance indicators

Specify if the simulation software is parallelized (or could be adapted):
Parallel software. Supports both MPI and OpenMP.

[25] http://ophidia.cmcc.it/

Specify if the simulation software can exploit GPUs
Yes, to run array-based primitives for data analysis

Specify how the simulation software exploits multicore systems
For the data analytics tasks at the "I/O server" level. The I/O server is a specific component in the Ophidia software stack. It is responsible for the I/O and data analytics tasks.

Specify if parametric runs are required-
Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) -

5.3 Workflows and tools for big data analysis

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing

A simplified "Trend analysis" workflow could involve in a first stage running pipelines of data operators (e.g. spatio-temporal sub-setting, percentile calculation, linear regression) on different input datasets (e.g. multiple models, historical/future scenarios). That would produce reduced inputs for a second stage including, for instance, intercomparison, ensemble and outlier analysis.

6 DETAILED USE CASES FOR RELEVANT USER STORIES

This section tries to put the focus on the preparation of detailed Use Cases starting from User Stories most relevant to the Case Study considered.

6.1 Identification of relevant User Stories

Examples of relevant User Stories linked to roles like for example Final User, Data Curator, etc.

In the following some user stories related to the case study on coupled model intercomparison data analysis and related to:

- running a large scale coupled model intercomparison data analysis experiment
- configuration/installation of the Software components
- dynamic instantiation of big data clusters and persistent store/publication of the results
- programmatic support for specific languages largely used by the community
- interactive and batch support

Running a large scale coupled model intercomparison data analysis experiment

1. the user should be able to define a data analysis experiment through a specific interface (scientific gateway or command line interface).
2. the experiment should be defined in terms of a set of input and a workflow for data analysis. Input related to the target platform (e.g. CPU cores for data analysis) could be provided too. Workflows should be selected from a workflow repository and then customized, re-used, and published again in the repository for further re-use. To this end, market-place tools should be used enabling a more collaborative and community-based approach. Several types of data analysis should be supported (e.g. anomalies analysis, trend analysis, climate change signal analysis) in order to address a large set of user needs.
3. The experiment workflow could involve datasets at a single site or at multiple sites. Search engines from the existing eco-system (e.g. ESGF Index Node) will enable the data discovery. Based on the experiment workflow, the INDIGO platform solution could provide (e.g. instantiate) a workflow management system (WMS) support, according to as a Workflow as a Service (WaaS) model. To run such a big data analysis experiment, two different levels of WMS would be needed: one coarse grain for large scale, distributed, and multi-service tasks orchestration (e.g. Kepler) and another one fine grain, which relates to the core data analytics part of the experiment and would coordinate the workflow execution at the level of big data analytics cluster instance (e.g. Ophidia). Different scenarios - directly related to the components deployment - could be supported, from (i) static configurations, to (ii) dynamic and (iii) elastic ones. The computational and storage resources should be allocated by the INDIGO middleware according to the user needs and the data locality constraints. The deployment of the data analysis environment should to be platform-agnostic.
4. Some of the tasks in the workflow could be related to running existing code like visualization tools well-known in the community (NCL, NCAR Command Language) or other ones for data manipulation and analysis (CDO, Climate Data Operators). Moreover, specific workflow interfaces should support running big data analytics experiments on large datasets by exploiting declarative approaches.

5. the results of the experiments should to be easily accessible by the end user for inspection, download, visualization. Additionally, the results should be also published on existing catalogues and be discovered by other users through search engines provided by the available eco-system.
6. security aspects should be taken into account and be interoperable with the available security eco-system/infrastructure.
7. the user interface (e.g. scientific gateway) should provide specific support for data analytics and visualization.

Configuration/installation of the Software components

1. the needed software components could be available through pre-configured virtual machine images. This would involve both single machines running specific services (like workflow engines) and clusters to support big data analytics experiments/runs.
2. the VMIs should be regularly updated according to the new software releases and users needs in terms of additional software (e.g. for data analysis/visualization) to be provisioned.
3. the VMIs should be published on well-known repositories in order to be easily accessible and discoverable.

Dynamic instantiation of big data clusters and persistent store/publication of the results

1. the coupled model intercomparison data analysis experiments will take advantage of the cloud solutions provided by INDIGO, which will support dynamic instantiation of clusters for big data analytics. That enables flexible scenarios where the resources can be instantiated/released at the begin/end of a user session.
2. the results of an "experiment" should be stored and made available to the end users on external and persistent services/storages, to make them even accessible after the lifetime of the virtual cluster. Data publication support would be needed, taking into account the already existing eco-system.

Programmatic support for specific languages largely used by the community

1. the user should be able to define and run an analytics experiment in the cloud
2. the results could be further analysed and visualized by exploiting programmatic support for specific languages like python. That would help re-using a very interesting set of already available libraries/classes for data analysis and visualization largely used by the community.
3. additionally, the user should be able to implement and add new data analysis operators to the analytic environment provided by INDIGO.

Interactive and batch support

1. interactive and batch support should be provided to address different scenarios.
2. batch support would help running in background large experiments that do not need a specific monitoring during the workflow execution (e.g. production-level experiments);
3. interactive support would provide immediate feedback to the end-user and the opportunity to include a human intervention during the experiment (e.g. research-level experiments or experiments needing additional input from the end users at run-time).
    a. interactivity also relies on real-time support in terms of experiment monitoring that have to be provided in the user interface.

7 INFRASTRUCTURE TECHNICAL REQUIREMENTS

Describe the Case Study from the point of view of the required e-infrastructure support. INDIGO Data-Cloud will support the use of heterogeneous resources.

7.1 Current e-Infrastructures Resources

Start from the current use of e-infrastructures.

7.1.1 Networking

Describe the current connectivity

Exploiting a server-side approach for data analysis, the network connection to transfer the data outside the data center is not key. On the contrary the connection between the storage and the computational resources inside the data center is related. In this regard, the connectivity at the level of the data center is based on InfiniBand w.r.t the cluster used at CMCC for the implementation of the Case Study.

Describe the key requirements (availability, bandwidth, latency, privacy, etc)

The bandwidth is the most relevant requirement, as very large datasets (terabytes order) have to be properly accessed, analysed, processed, etc. and the experiment workflow is strongly I/O demanding.

Specify any current issue (like last mile, or access from commercial, etc) -

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage

Private cloud and supercomputing resources in the provided test environment at CMCC. Initially a cluster with 100 cores for the private cloud part and 240 cores on the Athena supercomputer (Sandy Bridge Cluster).

Indicate if there is any mode of "orchestration" between them

Workflow tools could address a basic type of orchestration. However, so far, they different environments are used separately.

7.1.3 Storage

Describe the current resources used

100TB storage

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator)

I/O performance is the key requirement due to the I/O intensive nature of the analysis operations.

7.2 Short-Midterm Plans regarding e-Infrastructure use

Plans for next year (2016) and in 5 years (2020).

7.2.1 Networking

Describe the proposed connectivity:

Infiniband at the level of the data center

Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc)

Bandwidth is still a key requirement, due to the data volume expected for the next CMIP experiments.

Specify any potential solution/technique (for example SDN) -

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage

Private cloud and supercomputing resources in the provided test environment at CMCC. A cluster with about 100 cores for the private cloud part and 480 cores on the Athena supercomputer (Sandy Bridge Cluster). Additionally, private cloud infrastructures will be also made available, as the supercomputing centre will certainly expand this kind of resources.

Detail potential "orchestration" solutions

Beyond current workflow approaches (for instance with solutions more coupled with the infrastructure layer).

7.2.3 Storage

Describe the resources required

It is expected to double the storage resources devoted to the project activities (from 100 to 200TB).

Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator)

I/O performance will still represent a key requirement. Workflow complexity is expected to increase, based on the new tools (e.g. the ones provided by INDIGO), which will make easier building and extending a data analysis experiment/workflow.

7.3 On Monitoring (and Accounting)

Please outline any requirements for monitoring of the platforms and the applications.

If you have specific tools already in use, please outline them.

Please also specify monitoring, metrics at different levels: system, performance, availability, network QoS, website, security, etc.

Availability of services is monitored by using Ganglia, Cacti or Nagios tools.

7.4 On AAI

(From EGI, revise and check with WP4/5/6)

Describe the current AAI status of your community/research infrastructure

• Does your community/research infrastructure already use AAI solutions?

yes

• Can you describe the solutions you have adopted highlighting as applicable: Technology adopted (e.g. X509, SAML Shibboleth,...), Identity Providers (IdP) federations integrated (e.g. eduGAIN) or approximate number of individual IdPs integrated, Solution for homeless users (users without an insitutional IdP), Solutions to handle user attributes

In the current ESGF/IS-ENES infrastructure, the climate community exploits a federated identity provider solution based on OpenID. There is a specific node type, called Identy Provider which allows user authentication and secure delivery of user attributes. It includes an OpenID Provider for browser-based authentication, a MyProxy server for requesting limited-lifetime user certificates used for non-browser based access, and User Registration and Attribute Services for distributed access control.

A complete description of the security infrastructure is reported in Luca Cinquini et al[26].

*Describe the potential needs and expectations from an AAI integration in the **services and platforms provided by INDIGO***

In general, it is expected to be compliant/interoperable with the IdP supported/managed by the ESGF/IS-ENES community (this answers to all the following sub-points).

[26] Luca Cinquini, Daniel J. Crichton, Chris Mattmann, John Harney, Galen M. Shipman, Feiyi Wang, Rachana Ananthakrishnan, Neill Miller, Sebastian Denvil, Mark Morgan, Zed Pobre, Gavin M. Bell, Charles M. Doutriaux, Robert S. Drach, Dean N. Williams, Philip Kershaw, Stephen Pascoe, Estanislao Gonzalez, Sandro Fiore, Roland Schweitzer: The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. Future Generation Comp. Syst. 36: 400-417 (2014)

- *Type of IdP to be integrated (e.g. institutional IdP part of national federations and eduGAIN or non federated, social media credentials, dedicated research community catch-all IdP, ...)*
- *Preferred authentication technology, and requirements for support of multiple technology and credential translation services (e.g. SAML -> X509 translation)* **<input here>**
- *Community level authorization/attribute based authorization to support different authorization levels for the users* **<input here>**
- *Web access and/or non-web access* **<input here>**
- *Need for delegation (e.g. execute complex workflows on behalf of the user)* **<input here>**
- *Support for different level of assurance credentials, and need to use the information about users with lower level of assurance credentials to limit their capability* **<input here>**
- *Requirements for high level of assurance credentials (e.g. to access confidential/sensitive data)* **<input here>**

7.5 On HPC

Describe any specific issue related to the use of supercomputers.

A more dynamic and flexible use of the supercomputer resources in a cloud-based environment

7.6 Initial short/summary list for "test" applications (task 2.3)

Software used	***Software/applications/services required, configuration, dependencies (Describe the software/applications/services name, version, configuration, and dependencies needed to run the application, indicating origin and requirements.)*** CDO, NCO, Grads, NCL, Ophidia
Operating system requirements	Linux
Run libraries requirements	***Run API/libraries requirements (e.g., Java, C++, Python, etc.)*** Python, Java, C++
CPU requirements (multithread,MPI, "wholenode")	*multithread, MPI*
Memory requirements	>40GB per node

Network requirements	Fast communication between storage and computing nodes
Disk space requirements (permanent, temporal)	***Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)*** Storage requirement >50TB
External data access requirements	No external access, once the data is selected. The analysis is performed closed to the data. Nevertheless, intermediate results could be moved to a central site if a specific/final ensemble analysis tasks is foreseen in the workflow of the experiment.
Typical processing time	It depends on the workflow complexity and input data size
Other requirements	***Requirements for data synchronization*** ***Requirements for data publication*** Output results could be published on existing catalogues ***Requirements for depositing data to archives and referring them*** ***Requirements for mobile application components for data storage and access*** ***Requirements for data encryption and integrity control-related functionality***
Other comments	**<input here>**
Relevant references or URLs	**<input here>**

8 CONNECTION WITH INDIGO SOLUTIONS

<To be filled by INDIGO JRA >

8.1 IaaS / WP4

8.2 PaaS / WP5

8.3 SaaS / WP6

8.4 Other connections

9 FORMAL LIST OF REQUIREMENTS

<this will be further edited within WP2>

10 REFERENCES

R 1	
R 2	
R 3	
R 4	
R 5	