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

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. The figure below shows just 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.

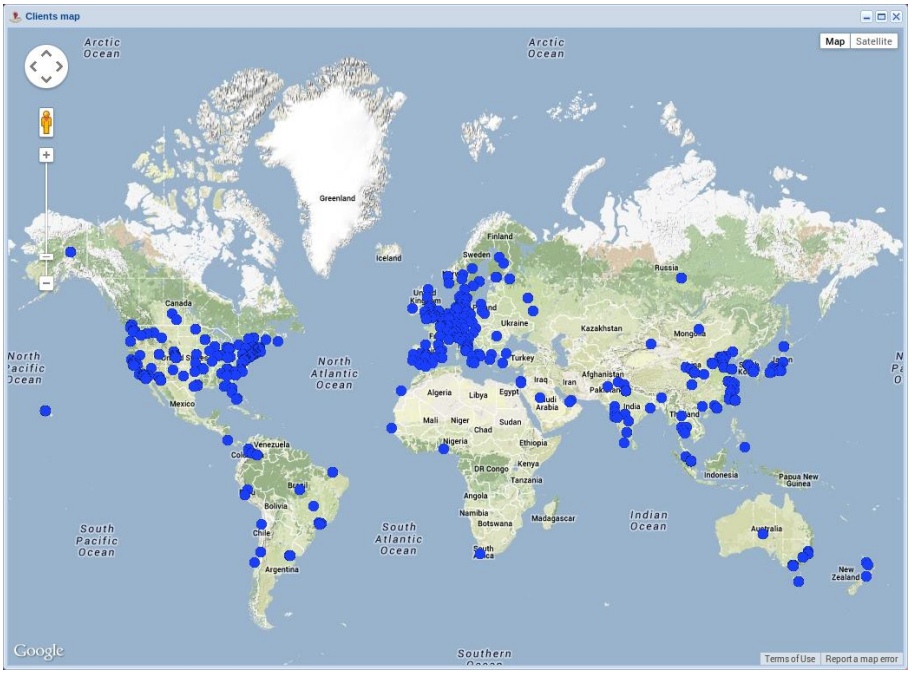


Figure1. Geo-referenced map of the clients that have downloaded CMIP5 data from the CMCC ESGF node in the 2012-2014 timeframe

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-*	• *Number of production sites running INDIGO-based*

	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.*	*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:* ***- ESGF Annual Meeting*** ***- European Geosciences Union*** ***- American Geophysical Union*** ***- ESA CCI workshops*** ***- GO-ESSP workshops*** ***- Events organized by EU projects closely related to the climate change domain: IS-ENES, EUBrazilCC*** • *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, 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.

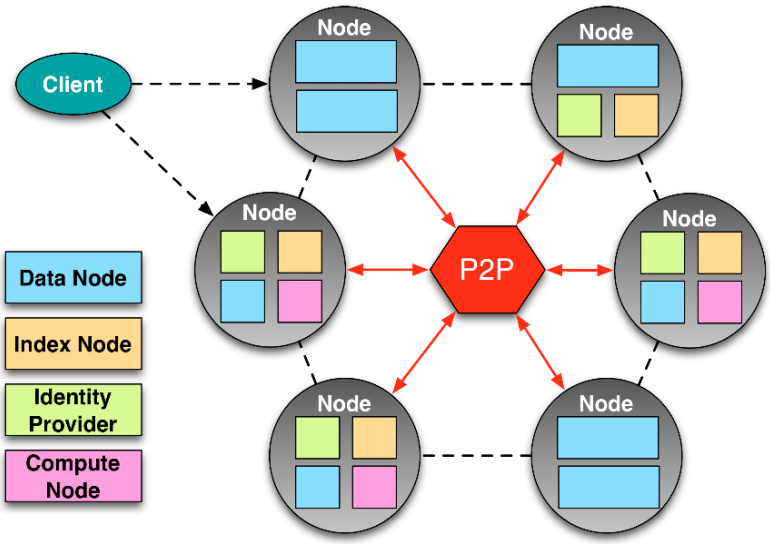
It is important to state that, the output related to these three classes of data analysis will be considered as a basis for studying additional data analysis experiments, related to:

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

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

With regard to the provided set of services, the current ESGF/IS-ENES infrastructure includes the following key components:

- Data nodes: 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).
- Idp nodes: components providing User Authentication service (OpenID based federation wide authentication)
- Index nodes: a solr based search index provides the indexing functionalities needed to enable the search and discovery of scientific datasets.
- The compute node is the one in the architectural design that will be devoted to providing processing capabilities.



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 (specific aspects related to the infrastructure have been reported in italic):

1. Starting from a user interface the climate scientist should be able to define a workflow. A workflow could be taken from a repository (addressing re-usability), or composed on the fly by the user (and then – eventually - stored in the workflows repository). It should include tasks and dependencies definition related to the expected data analysis process. It should include references to the datasets, variables, models, resolutions, etc., and a complete definition for each task in terms (e.g. the data operator, inputs, outputs). Computational/storage requirements should be also provided.
2. The workflow will be submitted to the infrastructure (*target data repositories should be identified and accordingly, the computational/storage resources should be allocated for data analysis taking into special account data locality aspects. Workflow as a Service solutions will be also considered. Deployment should be platform-agnostic and supporting elastic scenarios. Resources could be allocated dynamically*).
3. The workflow tasks will run remotely and the experiment results should be made available through the user interface. It should be also possible to publish the results of the analysis on dedicated catalogues. The user interface should provide analytics, exploration & visualization capabilities.

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:

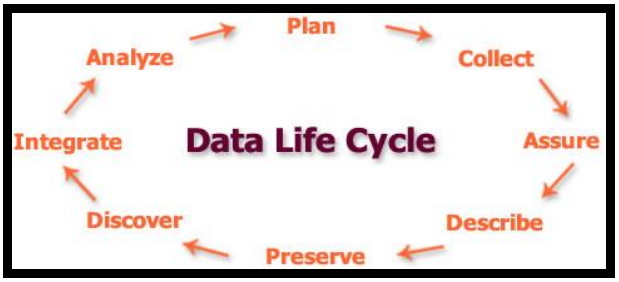
- single-site: the data analysis experiment could run at a single site (e.g. CMCC) providing both HPC and private cloud facilities.
- multi-site: the data analysis experiment could run at multiple sites (as a global, distributed experiment) by analysing datasets from several models.

In general additional 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.

Using the DataONE scheme, shown below, the different stages in the data life cycle are considered under the perspective of preparation of a DMP (Data Management Plan) following the recommendations of the UK DCC and H2020 guidelines.



BEFORE FILLING NEXT SECTIONS, CONSIDER CONSULTING:
https://www.dataone.org/all-best-practices-download-pdf and https://dmponline.dcc.ac.uk/

4.1 Data Management Plan (DMP) for this Case Study

According to EU H2020 indications[13], following UK DCC tool indications

4.1.1 Identification of the DMP

Plan identification: *<Code, ID>* **<input here>**

Associated grants: <Funded Projects, other grants> **<input here>**

Principal Researcher: **<input here>**

DMP Manager: **<input here>**

Description: **<input here>**

[13] *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 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).*

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

The DMP should address the points below on a dataset by dataset basis and should reflect the current status of reflection within the consortium about the data that will be produced.

For each data set provide:

Description of the data that will be generated or collected; indicate its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

Data set reference and name **<input here>**

Data set description **<input here>**

Standards and metadata **<input here>**

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created (see also below).

Connection to Instrumentation,

Sensors, Metadata, Calibration, etc (pending definitive form, see next sections)

<input here>

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups.

(pending definitive form, see next sections)

<input here>

Data Capture Methods

Outline how the data will be collected / generated and which community data standards (if any) will be used at this stage. Indicate how the data will be organised during the project, mentioning for example naming conventions, version control and folder structures. Consistent, well-ordered research data will be easier for the research team to find, understand and reuse.

- *How will the data be created?* **<input here>**
- *What standards or methodologies will you use?* **<input here>**
- *How will you structure and name your folders and files?* **<input here>**
- *How will you ensure that different versions of a dataset are easily identifiable?* **<input here>**

Metadata

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc. Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The UK Data Curation Center offers a catalogue of disciplinary metadata standards.

- *How will you capture / create the metadata?* **<input here>**

• *Can any of this information be created automatically?* **<input here>**
• *What metadata standards will you use and why?* **<input here>**

Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

<input here>

Method for Data Sharing

Consider where, how, and to whom the data should be made available. Will you share data via a data repository, handle data requests directly or use another mechanism? The methods used to share data will be dependent on a number of factors such as the type, size, complexity and sensitivity of data. Mention earlier examples to show a track record of effective data sharing.

• *How will you make the data available to others?* **<input here>**
• *With whom will you share the data, and under what conditions?* **<input here>**

Restrictions on Sharing

Outline any expected difficulties in data sharing, along with causes and possible measures to overcome these. Restrictions to data sharing may be due to participant confidentiality, consent agreements or IPR. Strategies to limit restrictions may include: anonymising or aggregating data; gaining participant consent for data sharing; gaining copyright permissions; and agreeing a limited embargo period.

• *Are any restrictions on data sharing required? e.g. limits on who can use the data, when and for what purpose.* **<input here>**
• *What restrictions are needed and why?* **<input here>**
• *What action will you take to overcome or minimise restrictions?* **<input here>**

Data Repository

Most research funders recommend the use of established data repositories, community databases and related initiatives to aid data preservation, sharing and reuse. An international list of data repositories is available via Databib or Re3data.

• *Where (i.e. in which repository) will the data be deposited?* **<input here>**

Archiving and preservation (including storage and backup)

Questions to consider before answering:

•What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

•Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

• What additional resources are needed to deliver your plan?

• Is additional specialist expertise (or training for existing staff) required?

• Do you have sufficient storage and equipment or do you need to cost in more?

• Will charges be applied by data repositories?

• Have you costed in time and effort to prepare the data for sharing / preservation?

Carefully consider any resources needed to deliver the plan. Where dedicated resources are needed, these should be outlined and justified. Outline any relevant technical expertise, support and training that is likely to be required and how it will be acquired. Provide details and justification for any hardware or software which will be purchased or additional storage and backup costs that may be charged by IT services. Funding should be included to cover any charges applied by data repositories, for example to handle data of exceptional size or complexity. Also remember to cost in time and effort to prepare data for deposit and ensure it is adequately documented to enable reuse. If you are not depositing in a data repository, ensure you have appropriate resources and systems in place to share and preserve the data.

Describe the procedures that will be put in place for long-term preservation of the data.

<input here>

Indicate how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered. **<input here>**

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

SCIENTIFIC RESEARCH DATA SHOULD BE EASILY ***DISCOVERABLE***

Questions to consider:

• How will potential users find out about your data?

• Will you provide metadata online to aid discovery and reuse?

Guidance: Indicate how potential new users can find out about your data and identify whether they could be suitable for their research purposes. For example, you may provide basic discovery metadata online (i.e. the title, author, subjects, keywords and publisher).

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. ***Digital Object Identifier****)?* **<input here>**

SCIENTIFIC RESEARCH DATA SHOULD BE **ACCESIBLE**

Questions to consider:

• Who owns the data?

• How will the data be licensed for reuse?

• If you are using third-party data, how do the permissions you have been granted affect licensing?

• Will data sharing be postponed / restricted e.g. to seek patents?

State who will own the copyright and IPR of any new data that you will generate. For multi-partner projects, IPR ownership may be worth covering in a consortium agreement. If purchasing or

reusing existing data sources, consider how the permissions granted to you affect licensing decisions. Outline any restrictions needed on data sharing e.g. to protect proprietary or patentable data. See the DCC guide: How to license research data.

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses? (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc) **<input here>**

*SCIENTIFIC RESEARCH DATA SHOULD BE **ASSESSABLE** AND INTELLIGIBLE*

• What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?

• What information needs to be retained to enable the data to be read and interpreted in the future?

Describe the types of documentation that will accompany the data to provide secondary users with any necessary details to prevent misuse, misinterpretation or confusion. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, any assumptions made, the format and file type of the data.

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review?, e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them **<input here>**

***USABLE** BEYOND THE ORIGINAL PURPOSE FOR WHICH IT WAS COLLECTED*

• What is the long-term preservation plan for the dataset? e.g. deposit in a data repository

• Will additional resources be needed to prepare data for deposit or meet charges from data repositories?

Researchers should consider how datasets that have long-term value will be preserved and curated beyond the lifetime of the grant. Also outline the plans for preparing and documenting data for sharing and archiving. If you do not propose to use an established repository, the data management plan should demonstrate that resources and systems will be in place to enable the data to be curated effectively beyond the lifetime of the grant.

Guidance on Metadata:

• How will you capture / create the metadata?

• Can any of this information be created automatically?

• What metadata standards will you use and why?

Metadata should be created to describe the data and aid discovery. Consider how you will capture this information and where it will be recorded e.g. in a database with links to each item, in a 'readme' text file, in file headers etc.

Researchers are strongly encouraged to use community standards to describe and structure data, where these are in place. The DCC offers a catalogue of disciplinary metadata standards.

Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data? e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum

software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists? **<input here>**

INTEROPERABLE** TO SPECIFIC QUALITY **STANDARDS

• What format will your data be in?

• Why have you chosen to use particular formats?

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

Outline and justify your choice of format e.g. SPSS, Open Document Format, tab-delimited format, MS Excel. Decisions may be based on staff expertise, a preference for open formats, the standards accepted by data centres or widespread usage within a given community. Using standardised and interchangeable or open lossless data formats ensures the long-term usability of data?

See the UKDS Guidance on recommended formats

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

<input here>

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

4.2.1 General description of data levels

Indicate if the DATASETS are organized into different levels (LEVEL-0, 1, 2, 3,4) and if so what are the relevant definitions and how DOI are provided. **<input here>**

4.2.2 Collection/Acquisition

Gathering RAW data

Specify how do you gather/collect your data (e.g. sensors, observations, satellites, etc.)? **<input here>**

How do you pre-process, transfer and store your RAW data? **<input here>**

From RAW Data to Calibrated Data

Describe the processes applied for Data Calibration, Validation, Filtering, etc. **<input here>**

4.2.3 Access to external data

Describe the identification and access to External Data **<input here>**

Indicate if there is a procedure for validation of External Data **<input here>**

4.2.4 Data curation

Specify any automatic check applied, like completing series, detecting outlier **<input here>**

Describe manual quality checks **<input here>**

Are there quality flags applied to the data? **<input here>**

4.2.5 Data ingestion / integration

Describe transformations applied to data taking into account ontologies/metadata. Indicate also if there is any "harmonization procedure" (to share/integrate data) and how linking internal and external data is made if relevant. **<input here>**

4.2.6 Further data processing

Describe, if relevant, the different additional processing steps (and the associated software and resources) applied to the (collected/curated) datasets to provide a "final" dataset collection that can be used in the analysis **<input here>**

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study **<input here>**

Specify if software packages/tools like MATLAB, R-Studio, iPython,etc. are used **<input here>**

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.) **<input here>**

Specify the resources and additional software required **<input here>**

Identify analysis challenges that can be classified as "Big Data" **<input here>**

List Big Data driven workflows **<input here>**

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization **<input here>**

Indicate how visualization is made and if interactivity/steering is needed **<input here>**

Specify the User Interfaces (web, desktop, mobile, etc.) **<input here>**

4.4 Data Publication

Describe the information flow from the analysis to the publication **<input here>**

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

5 SIMULATION/MODELLING

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 simulation/modelling needs

Describe the different models used (including references) **<input here>**

*Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution***<input here>**

5.2 Technical description of simulation/modelling software

For each simulation package:

Identify the simulation software **<input here>**

Provide a link to its documentation, and describe its maturity and support level **<input here>**

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) **<input here>**

Tag the simulation software as HTC or HPC **<input here>**

*List the input files required for execution and how to access them***<input here>**

Describe the output files and how they will be stored **<input here>**

Reference an existing installation and performance indicators **<input here>**

Specify if the simulation software is parallelized (or could be adapted) **<input here>**

Specify if the simulation software can exploit GPUs **<input here>**

Specify how the simulation software exploits multicore systems **<input here>**

Specify if parametric runs are required **<input here>**

Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) **<input here>**

5.3 Simulation Workflows

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing **<input here>**

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.

List User Stories based on data collection, curation, processing, analysis, simulation, etc, that are considered most relevant for the Case Study being analyzed **<input here>**

For each relevant User Story:

Draft a basic card **<input here>**

Provide details from conversation with the researchers' teams **<input here>**

Draft as a Use Case **<input here>**

Analyze tools to support the definition of the Use Case (like mockups). Integrate in the analysis the requirements on user interfaces (like the use of mobile resources, under different flavours, access through web interfaces, etc.) **<input here>**

Describe the way to extract requirements and define acceptance criteria **<input here>**

Include if possible an example of support for Big Data driven workflows for e-Science, with requirements for scientific workflows management, under a "Workflow as a Service" model, where the proper workflow engines will be selected according to user needs and requirements.

In such case please describe the scenario for Big Data analysis, and assure that the Use Case considers which levels of workflow engines are needed (e.g., "coarse gran", which targeting distributed (loosely coupled) experiments, through workflow orchestration across heterogeneous set of services; "fine grain", which targeting high performance (tightly coupled) data analysis through workflows orchestration on big data analytics frameworks)

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 **<input here>**
Describe the key requirements (availability, bandwidth, latency, privacy, etc) **<input here>**
Specify any current issue (like last mile, or access from commercial, etc) **<input here>**

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage **<input here>**
Indicate if there is any mode of "orchestration" between them **<input here>**

7.1.3 Storage

Describe the current resources used **<input here>**
Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **<input here>**

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 **<input here>**
Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc) **<input here>**
Specify any potential solution/technique (for example SDN) **<input here>**

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage **<input here>**
Detail potential "orchestration" solutions **<input here>**

7.2.3 Storage

Describe the resources required **<input here>**
Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **<input here>**

7.2.4 SPECIFIC QUESTIONS REGARDING USE OF EGI.eu (FROM EGI DOC 2478)

Sample questions to capture details of a support case

These questions can help case supporters interview the case submitter and the NGIs to refine the technical details of the case and ultimately to move towards a suitable technical setup. These questions aim at understanding the user's need, the technical and other requirements/constrains of the case, and the impact that a solution would bring to the scientific community. These questions provide only guidance – Ticket owners can use other questions or even other methods to identify details of their support case(s).

- *What does the user/community want to achieve? (What's the user story?)*
- *For who does the case request resources for? (CPU/storage capacity, SW tools, consultant time, etc.) For a group? For a project? For a collaboration? Etc.*
- *What is the size of the group that would benefit from these resources, and where these people are? (which country, institute)*
- *Approximately how much compute and storage capacity and for how long time is needed? (may be irrelevant if the activity is for example assessment of an EGI technology)*
- *Does the user need access to an existing allocation (→ join existing VO), or does he/she needs a new allocation? (→ create a new VO)*
- *What is the scientific discipline?*
- *Which institute does the contact work for (or those he/she represents)?*
- *Does the case include preferences on specific tools and technologies to use?*
    - *For example: grid access to HTC clusters with gLite; Cloud access to OpenStack sites; Access to clusters via standard interdafaces; Access to image analysis tools via Web portal*
- *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)*
- *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)*
- *Does the user (or those he/she represent) have the resources, time and skills to manage an EGI VO?*
- *Which NGIs are interested in supporting this case? (Question to the NGIs)*

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.

<input here>

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? **<input here>**

• 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 **<input here>**

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

- *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, ...)* **<input here>**
- *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.

<input here>

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.)*** **<input here>**
Operating system requirements	**<input here>**
Run libraries requirements	***Run API/libraries requirements (e.g., Java, C++, Python, etc.)*** **<input here>**
CPU requirements (multithread,MPI, "wholenode")	**<input here>**
Memory requirements	**<input here>**
Network requirements	**<input here>**
Disk space requirements (permanent, temporal)	***Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)*** **<input here>**
External data access requirements	**<input here>**
Typical processing time	**<input here>**
Other requirements	***Requirements for data synchronization*** ***Requirements for data publication*** ***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*** **<input here>**
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	