INDIGO-DataCloud

INITIAL REQUIREMENTS FROM RESEARCH COMMUNITIES ANNEX 1.*P0*: SELECTED CASE STUDY FROM *LIFEWATCH: ALGAE BLOOM IN A WATER RESERVOIR*

INPUT TO EU DELIVERABLE: D 2.1

Document identifier:	INDIGO-WP2-D2.1-ANNEX-1P0-V10
Date:	**13/06/2015**
Activity:	**WP2**
Lead Partner:	**CSIC**
Document Status:	**First Complete Version V10**
Dissemination Level:	**INTERNAL**
Document Link:	

Abstract

This report summarizes the findings of T2.1 and T2.2 for partner P0, **CSIC,** along the first three months of the project, regarding a Case Study within LifeWatch: Algae Bloom in a Water Reservoir. It is an integrated document including a general description of the research communities involved and the selected Case Study 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	F.Aguilar, J.Marco, (CSIC) A.Monteoliva (ECOHYDROS)	**P0 CSIC**/WP2	13 June 2015
Reviewed by	**Moderators:** P.Solagna, F.Aguilar, J.Marco		
Approved by	**WP2 internal**		

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
4-8	28-may-2015 6-june-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
10	10-june-2015	Draft to be circulated for internal review, v10	J.Marco, F.Aguilar CSIC
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*: **ESFRI LifeWatch**
- *Institution/partner representing the community in INDIGO:* **CSIC**
- *Main contact person*: **Fernando Aguilar (technical) / Agustin Monteoliva (researcher)**
- *Contact email:* **aguilarf@ifca.unican.es**
- *Specific Title for the Case Study:* **Monitoring and Modelling Algae Bloom in a Water Reservoir**

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 problem addressed is the monitoring, modelling and prediction of algae bloom in a water reservoir. The prototype has been developed by an SME, ECOHYDROS, a consulting company working in the environmental area, lead by a biologist. It has been implemented in the Cuerda del Pozo water reservoir, near Soria in Spain. The development is supported by the River Authorities (Confederacion Hidrografica del Duero), by a LIFE+ project (ROEM+). IFCA-CSIC has been collaborating in this initiative since 2008, along the FP7 project DORII.

Eutrophication in water reservoirs, resulting in algae bloom, is an increasing serious problem in many water reservoirs in Europe and in the whole world due to the increase of anthropogenic pressure (human activities, including also farming) and climate change (warmer summers favour algae bloom).

The prediction of eutrophication and of the development of algae bloom requires modelling the water reservoir from the hydrological perspective, predicting in detail the temperature profile of the water and its composition, and also the modelling of all processes related to algae growth from the biological point of view. Validating the model can only be done thanks to the in-situ measurements, requiring a very complex instrumentation. A picture describing the approach followed and the different components is shown below in figure 1.

A key component is the instrumentation. The current setup includes a central platform that is installed in the middle of the water, and instrumented with meteorological sensors (wind, temperature, solar radiation, rain, etc), and water quality sensors (conductivity, temperature, dissolved oxygen, turbidity, pH, etc.). The water quality sensor probe is placed in a cage connected to a wincher system allowing vertical profiling (range 1-30 m. in depth) that is critical to monitor the evolution of the water stratification, clearly reflected in the thermocline curves). More complex instrumentation, including radiometers, spectrometers and absorbance sensors are also included to monitor the abundance of green and cyano-green algae directly, through the correlation with their luminescence.

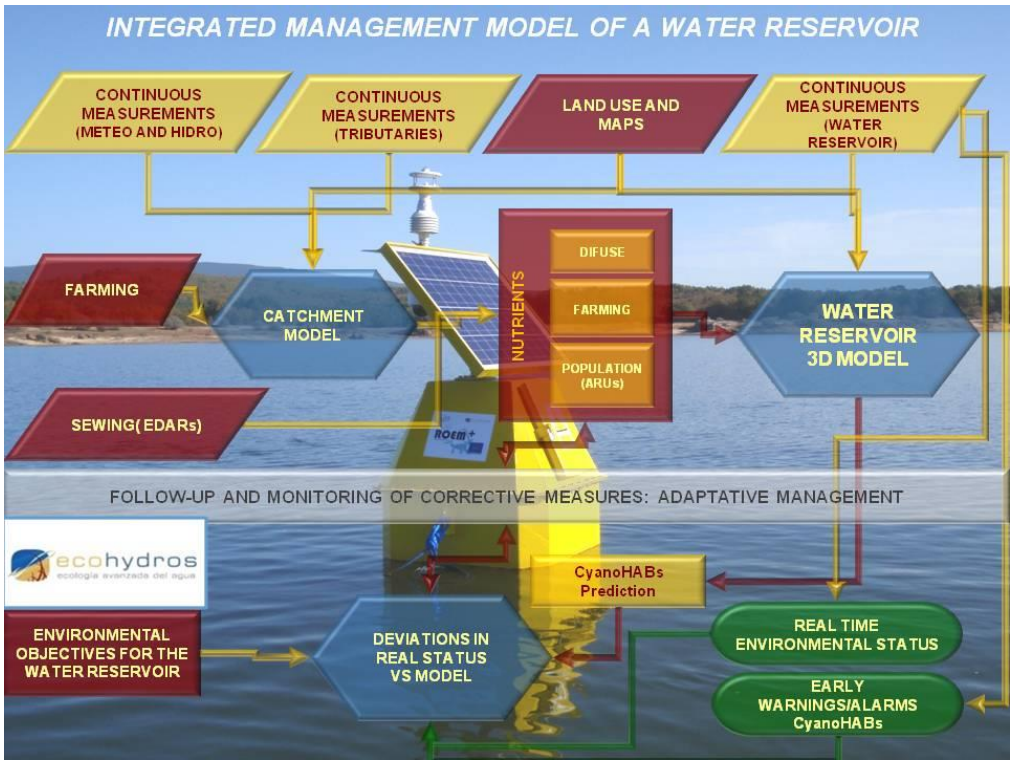


Figure 1: Scheme of the Case Study for the Prediction of Algae Bloom in a Water Reservoir

Another key ingredient is the simulation model and software suite. Currently the software used is the open source suite **Delft3D** that includes a module providing the simulation of the hydrodynamics of the water reservoir (FLOW) and another module for the simulation of the water quality (DELWAQ).

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 idea is to install and execute automatically the whole monitoring and modelling platform on demand in Cloud resources, including HPC resources and analysis capabilities, to be able to use it at different resources by the (SME) company that has developed it.

To adequately "tune" the model and also to explore the predictions under different conditions, the iterative/coupled execution of the different simulation components is required. A SaaS/PaaS solution, including also the possibility of "finding the best solution through a parametric exploration", could be of great interest. The SME plans to use intensively the Cloud as the basic framework, and they would need to migrate their usual tools (visualization using matlab free package, analysis using excel/python/R scripts) and run them on the quite large simulation output (order of hundredths of terabytes). They also need scalability to offer the solution for as many water reservoirs as needed.

1.4 Expected results and derived impact

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

The current ongoing project covers the reservoir at Cuerda del Pozo, with bathing facilities and also providing drinking supply to a city (Soria). So it has clear impact on environmental management. The results of the project will be published and they are at the status of the art at world level, in particular for the advanced instrumentation used. There are many other water reservoirs in Spain and in many other places in the world where eutrophication is problem and algae blooms represent a clear challenge.

1.5 References useful to understand the Case Study

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

Presentation of the latest status of the ongoing LIFE+ project is available at:

http://www.roemplus-life.eu/

A summary was presented within the slides of the EGI LW CC at Lisbon EGI Conf 2015, see pages 4-15; it can be downloaded from:

http://indico.egi.eu/indico/materialDisplay.py?contribId=196&sessionId=55&materialId=slides&confId=2452

Publication on the instrumentation platform:

Coterillo et al., "*Integrating a Multisensor Mobile System in the Grid Infrastructure*", in "*Remote Instrumentation for eScience and Related Aspects*", Springer 2012.

See also:

Modelling of a watershed: a distributed parallel application in a grid framework, Campos et al, in Computing and Informatics, Vol. 27, 2008, 285–296

Poster at the XVII Congreso de la Asociación Ibérica de Limnología, in 2014:

The Advanced Autonomous Water Quality Monitoring System of Cuerda del Pozo Reservoir: Four Years of High Resolution, by Agustín P. Monteoliva, Fernando Aguilar, Alberto Criado, Jesús Marco, Tamara Santiago.

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 researchers (typically biologists) want to use the data collected in an instrumented platform in the water reservoir and also in some tributaries to

a) Monitor the evolution of the potential eutrophication of the water reservoir

b) Use these data as input to a model (hydrological and biological)

c) Execute the model and understand how well it compares with real data, and what are the main parameters that may affect the eutrophication evolution

d) Implement a predictive/alarm framework to enable warnings to the water management authorities about the water quality and the expected evolution in the time framework of weeks or months

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 community around this Case Study includes:

a) The researchers (typically linked to the limnology community) that study the evolution of the water quality, an in particular eutrophication. In this case, the core team is integrated by biologists and environmental researchers working at an SME, Ecohydros.

b) Water management authorities, that include also biologists, chemists and civil engineers. Main institution in this example is the Confederación Hidrográfica del Duero (CHD) in Spain.

c) ICT groups, like the one at IFCA, supporting the implementation of the instrumentation and the simulation. Also other companies, like ITG, involved in the Life+ project.

d) Other groups in limnology, an area that covers a wide spectrum of professionals in other sectors. For example, other biologists interested in the evolution of the cyanophicea, including genetic topics. Typical institutions: public research organizations (like CSIC, and also CEDEX in Spain), Universities (like the University of Cantabria and the Universities Autónoma and Complutense de Madrid in Spain)

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

This Case Study is well advanced. The main current associated project, LIFE+ ROEM+, will end by mid 2016. A basic support project will follow-up, and new installations will be continued in new water

reservoirs (like Cogotas in Avila) and lakes (like Sanabria in Zamora). In the framework of LifeWatch an international extension is foreseen, using also networking channels like those supported by the NetLake initiative.

The results of the current model based in DELFT3D, working on HPC and HTC resources, but yet to be fully migrated to the Cloud platform (see section 6) have been recently presented.

The validation of the hydrological model is almost ready, while yet substantial work on the biological results is required.

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 integration of the full data workflow is needed to enable the biologists in the SME to execute it wholly in the Cloud: from the storage of the instrumentation data to its use as input for the DELFT-3D model, that requires HPC resources for execution, to the postprocessing using standard software like Excel or R or Matlab to process large outputs.

Remote access is a must for postprocesing.

Enabling HPC/HTC workflows in the Cloud is required.

Parametric runs would largely benefit the validation.

Currently the output visualization requires Matlab plug-in

Implementation of the DMP would help a lot to make the solution scalable and sustainable.

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.

Currently the monitoring/data collection e-infrastructure is not in the Cloud, but the plan is to provide such a setup by the end of this year. The resources will be provided by IFCA nodes integrated in the EGI.eu FedCloud.

The simulation can run in FedCloud as well, but an optimized remote access has to be provided to researchers to be able to examine/visualize the large data outputs. Remote interactive postprocessing using either Matlab, R or Excel, is required.

Preferably all the components should run in the Cloud.

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.

Describe the exploitation plans related to this Case Study

The whole platform is part of the research/innovation portfolio of the SME company (ECOHYDROS).

It is currently implemented for Cuerda del Pozo water reservoir, and will follow for Cogotas reservoir, where the instrumentation is already deployed. There are plans to implement it also for Lake Sanabria, also in Spain. There is a large potential of water reservoirs and lakes in Spain and in Europe where this problem is critical, but also in many other water reservoirs around the world. A recent example is new water reservoirs being built in Bolivia, where a preliminary study has to be made.

From the point of view of basic research, the initial results of the project have been presented at the Iberian Limnology meeting, and it is expected that further results will be published, giving origin also to 1-3 PhD thesis. It is also foreseen to present the results to the LTER international community as an example of monitoring relevant ecological variables for inland waters.

New projects are foreseen related also to the expansion of alien invasive species, although this step requires the implementation of further modules in the ecological simulation chain.

Please indicate (as realistic as possible) the expected impact for each topic in the following table:

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: **LIFEWATCH, LTER** • *Number of production sites supporting the software* **at least 5 (in FedCloud)**
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* **1 with IB, 1 with 10G Eth** • *Number/List of production sites providing support for Cloud elastic scheduling* **IFCA, EBD, BIFI, CESGA LIP, within IBERGRID** • *Number of popular applications used by the user communities directly integrated with the project products:* **MATLAB, R** • *Number of research communities using the developed Science Gateway and Mobile Apps:* **1** • *Research Communities external to INDIGO using the software products*: **likely 2 (GLEON)**
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:* **Unknown**
Visibility	*Visibility of the project among scientists, technology providers and resource managers at high level.*	• *Number of press releases issued*: **5 (related to ROEM+ and Sanabria work)** • *Number of download of software from repository per year:* **Unknown** • *List of potential events/conferences/workshops:* **2-3 per**

		year • *Number of domain exhibitions attended* **2-3 per year** • *Number of communities and stakeholders contacted* >**5 per year**
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:* **1-2 per year** • *Number of conference papers and presentations*: **3-5 per year**

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.

SEE SECTION 6 FOR DETAILS
The whole Case Study includes several components:

Real-time data measurements for monitoring

Hydrological simulation of the water reservoir

Biological simulation of the water reservoir

Analysis of results and predictive model

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.)

As previously hinted in section 2.2, the different user categories include

a) The researchers (typically linked to the limnology community) that study the evolution of the water quality, an in particular eutrophication. In this case, the core team is integrated by biologists and environmental researchers working at an SME, Ecohydros.

   They will usually run the monitoring programs and the postprocessing analysis.

b) Water management authorities that includes also biologists, chemists and civil engineers. Main institution in this example is the Confederación Hidrográfica del Duero (CHD) in Spain. They are mainly interested in accessing the monitoring data through a web portal, and potentially they could be interested in executing different simulations under various scenarios, when the eutrophication problem becomes dangerous to the health.

c) ICT groups, like the one at IFCA, under a previous support contract and current collaboration projects with ECOHYDROS, supporting the implementation of the instrumentation and the simulation. They are currently executing the simulation and the water quality model analysis.

3.3 General description of datasets/information used

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

There are three main datasets:

INSTRUMENTED PLATFORM DATA, where the information from the sensors is stored and curated

EXTERNAL DATA from tributaries, meteo and hydrological agencies, maps, etc

SIMULATION AND POSTPROCESSING DATA produced by the different programs being executed.

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

We have considered two main User Stories

User Story A): SME team wants to model the hydrodynamic behaviour of the water reservoir, to reproduce the thermocline and predict the onset and completion times of the water column stratification, with special interest in its final phase (september/october).

User Story B): SME team wants to predict algae bloom based on model, and validate against previous year detailed analytical measurements

See section 6 for further details on Use Cases.

From the point of view of "services" definition, the following ones have been identified:

-Access/Configure all input data for the hydrodynamic model

In particular a service can be defined for the access to the data from the INSTRUMENTED PLATFORM that is placed in CdP water reservoir. These platform data are stored in a database located also next to CdP reservoir and replicated in a server at IFCA. Some other observations and buoys data are stored in independent files (xls in general) or databases. Files formatted in a specific way so they can be used by the model are created and stored in private computers as well as configuration files to run the model. All this files are uploaded and replicated when needed in HPC resource like the Altamira supercomputer or Cloud VMs. So a service could be defined to define the configuration parameters and access/transfer all the data required by the model.

-Execution in HTC/HPC resources

The different model components are currently processed in a wide range of computing resources: from personal computers (water quality model, low resolution hydrodynamics), Altamira (medium-high resolution hydrodynamics and water quality when needed) or cloud VMs (medium-high resolution hydrodynamics and water quality when needed).

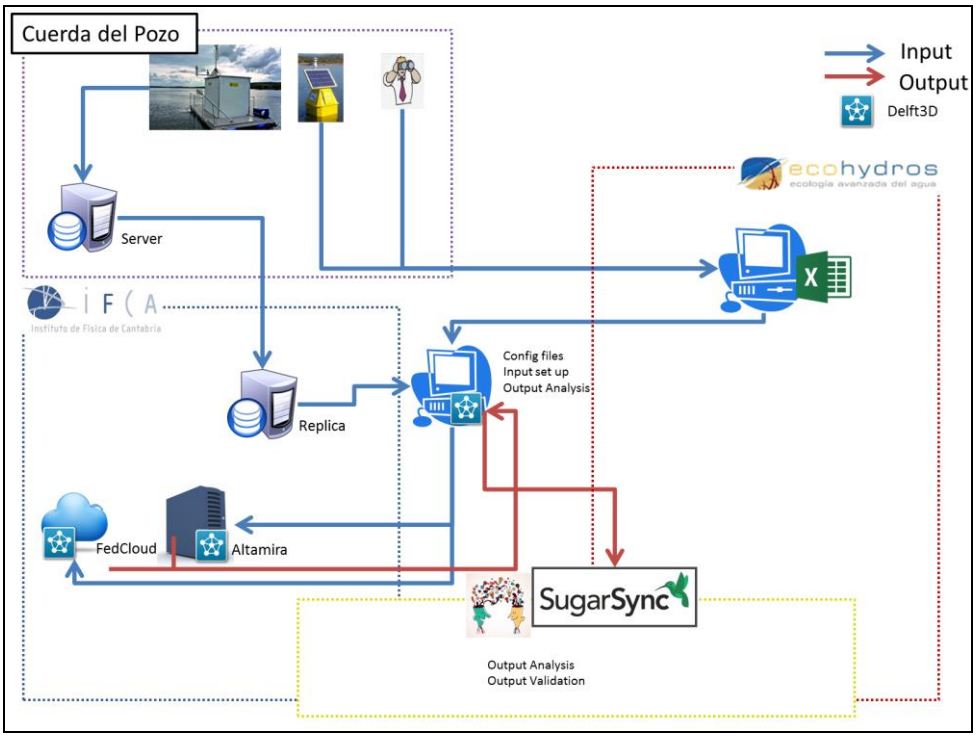
-Remote post processing of output

After processing, the (large/very large) output is stored where the processing was made and any replication requires high bandwidth connectivity. Analysis is made from personal computers using Excel or Matlab. Currently data analysis are stored in PCs and uploaded to a "SugarSync" folder for sharing with ECOHYDROS researchers when it is tagged as relevant model (see next figure). This process should be transformed into a remote post processing service in the Cloud.

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.

The general data workflow is depicted in the following figure:



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.

As previously indicated, both Cluster, Cloud and Supercomputing resources are used.

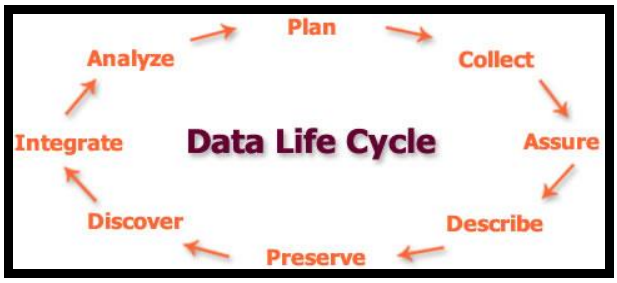
The execution of the DELFT-3D models needs low-medium storage resources: inputs size is around 1GB, while outputs may need up to 1TB per simulation in high resolution (~50 GB in medium resolution)

The main requirement for Cloud or HPC resources is memory, a minimum of 12 GB are needed.

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[3], following UK DCC tool indications

4.1.1 Identification of the DMP

Plan identification: *<Code, ID>* **ECOH-CdP-DMP, ID=0001 (proposed, not implemented yet)**

Associated grants: <Funded Projects, other grants> **LIFE+ ROEM+, CHD-2010, DORII**

Principal Researcher: **Agustin Monteoliva (apmonteoliva@ecohydros.com) (ECOHYDROS)**

DMP Manager: **F. Aguilar (aguilarf@ifca.unican.es) (IFCA-CSIC)**

Description:

[3] *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).*

Data from the instrumentation are collected in the platform in the water reservoir and also in other river stations along the different tributaries. These data have to be curated, integrated and then compared to the results from the simulation. Data are collected in hourly basis are grouped in different sets: atmospheric data, water physical data and water quality data.

External data is integrated from different sources: analytical measurements, real-time measurements from water management authorities, maps and statistics from official sources.

Simulation data is obtained from DELFT3D model output.

Post-processing information is elaborated using both simulation output and real data collected.

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

INSTRUMENTED PLATFORM DATA

Data set description

The platform provides the following data:

METEO DATA: Pressure, Air Temperature, Humidity, Wind speed and direction, Rain mean intensity and peaks, IR solar radiation direct/incident and reflected.

WATER QUALITY DATA: Temperature, Conductivity, pH, Oxygen concentration, Pressure, Redox, Salinity. Carbonates, Nitrates, Total Organic matter, Chlorophyll, Cyano, incident radiation attenuation at different wave lengths,

Standards and metadata

All data are stored in relational databases (MySQL) in different tables.

Metadata Standards are currently not yet used, EML (Ecological Metadata Language) is planned.

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)

The platform is placed in the middle of the reservoir. The platform cabin supports in its structure a weather station, a net radiometer, and a GPS receiver, as well as an external directional antenna for communication with the PC server in the office in the shore. Also there is installed an altimeter and a probe measuring the depth of the water reservoir and the depth. It also supports two renewable energy

systems: an aero generator and a set of solar panels. Furthermore, there is a profiler infrastructure that takes data from the bottom to the top of the water column. All this information is stored and sent to the PC server. The system is currently taking data and has been so since 2010.

For a detailed reference see I. Coterillo et al., "Integrating a Multisensor Mobile System in the Grid Infrastructure", in "Remote Instrumentation for eScience and Related Aspects", Springer 2012.

Calibration is done regularly following the protocols of the different sensors.

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups. (pending definitive form, see next sections)

Vocabularies are pending.

Ontologies are not defined yet.

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

EXTERNAL DATA:

1) Data from Tributaries: Buoys in tributaries deployed within the ROEM+ project provide data of physical, chemical and biological status of rivers that flow into the reservoir. Data includes flow measurement, physical-chemical data, and some indicators related to the biological status of the rivers.
2) Data from Water Management Authorities: the SAI system provides real time information about water level and other measurements at different points of the water reservoir and tributaries
3) Maps from IGN (National Geographical Institute): provide coverage and land use
4) Bathymetry and sediments maps: internal info provided by ECOHYDROS.

Data set description

This information is required to feed the DELFT3D model and validate some outputs (for example, water level).

Standards and metadata

Tributaries: Metadata are currently not used but it will be. Reference to OGS is done(?)

IGN: Inspire should in principle apply (see http://inspire-regadmin.jrc.ec.europa.eu/dataspecification/DataSpecification.action)

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)

Not defined

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups. (pending definitive form, see next sections)

Not defined

Data set reference and name

SIMULATED DATA:

5) DELFT3D Hydro dynamical Model Ouput: Maps in 3D corresponding to the different variables used, with a given mesh resolution and time step (1h, 6h) and range (2010 full year, 2014 full year. 2015 ongoing)
6) DELFT3D DELWAQ Water quality model ouput: Maps in 3D corresponding to the different variables used, like oxygen level or species (algae-cyano/chloro) concentration

Data set description This information is required to compare with the INSTRUMENTED PLATFORM DATA and validate the model, prepare predictions, etc.

Standards and metadata

DELFT3D set of variables are used.

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)

Does not apply

Vocabularies and Ontologies

Are they relevant? Internal vocabularies related to the specific fields. RDA groups. (pending definitive form, see next sections)

Not defined

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?* **See above**

• *What standards or methodologies will you use?* **Pending**
• *How will you structure and name your folders and files?* **Pending**
• *How will you ensure that different versions of a dataset are easily identifiable?* **Pending**

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?*

INSTRUMENTED PLATFORM:

There are two groups of metadata: metadata that defines the different parameters and metadata to identify different datasets. The first group have to be done "by hand" meanwhile the second can be done automatically.

The metadata used will be EML (or similar) because it is a metadata language widely used for ecology and biodiversity initiatives.

• *Can any of this information be created automatically?* **See above**
• *What metadata standards will you use and why?* **See above**

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).

INSTRUMENTED PLATFORM DATA:

The data are public and can be accessed through the web portal (http://doriiie02.ifca.es), that provides access to a Flex tool for data visualization and download.

This portal has two different versions: a public version and an extended version internal to the project. The public version supports direct data downloading and also provides access to charts or plots, in historical or profile mode. The extended version also provides information in real time and some other different parameters and calculated values.

For eventual queries, we will provide the data in formats such as HDF, accessible as well through email requests.

Access to databases and associated software tools generated under the project will be available for educational, research and non profit purposes. Such access will be provided using web based applications, as appropriate.

Materials generated under the project will be disseminated in accordance with University/Participating institutional and CSIC policies. Depending on such policies, materials may be transferred to others under the terms of a material transfer agreement.

Those that use the data (as opposed to any resulting manuscripts) should cite it the data as follows:

ECOHYDROS, IFCA-CSIC, CHD: Monitorización de la Calidad del Agua en Cuerda del Pozo,Soria, SPAIN (2010-2015) [http://doriiie02.ifca.es]; accessed on ddmmyyyy

No PID is provided yet for time-series data.

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?* **See above**

• *With whom will you share the data, and under what conditions?* **See above**

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.* **See above**

• *What restrictions are needed and why?* **See above**

• *What action will you take to overcome or minimise restrictions?* **See above**

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?*

A copy of the INSTRUMENTED PLATFORM DATA will be deposited in the CSIC institutional repository in due time for long time preservation.

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.

Both data, database and web based visualization tool are preserved not only for maintenance purposes, but also for security.

Current system comprises two servers located at different places (one in Soria, close to the instrumentation, another one at IFCA in Santander, separated by 400km) and a backup system on magnetic tape that ensures data conservation in case of problems with disks or access of hackers.

Soria server is the first and main server, where the data is firstly stored. It supports two different databases: one for raw data and another one for processed/curated data. In case of failure, an "rdiff" based system is used to restore the databases to the last stable state.

There is an asynchronous full DB replica hosted at doriiie02.ifca.es. If the replication process is broken, it can be restarted following MySQL documentation. This server, located at IFCA, hosts database replication and web server files as well (included Access front page, Flex files, php script to get data, etc.). Therefore, it is so important to keep this data in backups, guaranteeing the full recovery in case of failure of the online systems.

There are two different types of backups on this server based on rdiff and on Bacula (stored in tapes, for long term preservation).

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. **Current volume for INSTRUMENTED PLATFORM DATA (2010-2015) is well below 1TB. These data will be preserved as mySQL for 20 years using an LTO-5 WORM magnetic tape (double copy).**

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> We are not yet using DOI**

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)

ECOHYDROS SL is the owner of the data. Data is not licensed.

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 **PENDING**

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? **PENDING**

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

The data can be exported as tabular data with minimal metadata in CSV (comma-separated values) format from the monitoring interface, or directly from the MySQL system on demand.

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.

INSTRUMENTED PLATFORM DATA:

LEVEL 0 refers to RAW data, i.e. data as collected online from the instruments and stored in the MySQL database at the water reservoir. This data is NEVER modified after being collected.

LEVEL 1 applies to CURATED data, and although it follows the same format as RAW DATA and the database tables have the same format, they are modified for example subtracting periods with wrong calibrations, or corresponding to artificial measurements.

LEVEL 2 will apply to VALIDATED-DERIVED data, i.e. data certified after comparison with other external measurements. For example some of the EUtrohication parameters.

No DOI are provided until now.

Pending: DATA INGESTION TO DELFT-3D

4.2.2 Collection/Acquisition

Gathering RAW data

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

See above, collected data corresponds to an INSTRUMENTED PLATFORM

How do you pre-process, transfer and store your RAW data?

RAW data is collected from the instrument using a single LabView application that handles multiple sensors and stores the information as RAW data in a MySQL database running in the servers at the water station and at IFCA. Transfer of data from the platform to the server is done via Wifi, and then via Wimax up to IFCA.

From RAW Data to Calibrated Data

Describe the processes applied for Data Calibration, Validation, Filtering, etc.

As indicated above, RAW DATA is usually already calibrated. Corrections to the calibration may apply, and are already included in the CURATED DATA.

CURATED DATA excludes some outliers, and also periods when instrumentation is not properly working.

VALIDATED/DERIVED DATA is obtained from CURATED DATA by contrasting with analytical and in situ measurements, like those from chemical analysis at labs or from other probes. External sources like AEMET Meteo measurements have been also used (see internal work on meteo time series). Reference plots like those corresponding to thermocline profiles along 6 months, or the comparison with water level measurements from water management authorities, are also applied. The oxygen profile is another reference plot directly correlated with many other measurements that is being used for the VALIDATED DATA tag.

4.2.3 Access to external data

Describe the identification and access to External Data

Data from tributaries can be downloaded via another web portal (from ROEM+, see http://roem.itg.es/data_tables)

AEMET data can be downloaded diary from the web page www.aemet.es

IGN data can be downloaded from IGN repositories.

Indicate if there is a procedure for validation of External Data

Meteo data were contrasted with AEMET data

Analytical and in situ measurements, like those on chemical analysis, have also been contrasted

4.2.4 Data curation

Specify any automatic check applied, like completing series, detecting outlier

A basic outlier detection routine is applied (details to be included)

Peaks detection scripts and corresponding alarms are set (details to be given)

Manual filtering is applied to periods where the instrumentation shows problems.

Describe manual quality checks **Profiles are regularly checked by ECOHYDROS operators online**

Are there quality flags applied to the data?

Yes, very simple tags apply in the CURATED database: 0=yet not curated, 1=curated

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. **Do not apply yet (see connection to DELFT3D)**

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 **See above**

4.3 Analysis

4.3.1 Basic analysis and standard analysis suites

Describe usual examples of basic analysis in the Case Study

The analysis of the evolution of thermoclines and algae bloom (see section 6)

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

Production and analysis of thermoclines (multi fit with several parameters in Excel, R, MATLAB). See section 6

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.) **Time series analysis: NN to provide prediction based on historic registers**

Specify the resources and additional software required **Pending**

Identify analysis challenges that can be classified as "Big Data" **See section 6**

List Big Data driven workflows **See section 6**

4.3.3 Data visualization and interactive analysis

Indicate the need for data and analysis results visualization

Flex plots in monitoring interface are also used for data visualization

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

See section 6

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

See section 6

4.4 Data Publication

Describe the information flow from the analysis to the publication

Pending

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

Not directly linked up to now to the use of e-infrastructure

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)

Delft3D Open Source software, see http://oss.deltares.nl/web/delft3d

Installed in ALTAMIRA (HPC) and in local cluster (HTC)

Installed also in local workstations and laptops

We use the Delft3D to simulate the water body, in our case Cuerda del Pozo reservoir in Soria (Spain). The program runs with 2D and 3D meshes with a number of layers that can be edited by the user. Hence mesh resolution is an important factor for program performance. With a low or medium resolution mesh (cells larger than 250x250 meters with few vertical layers) execution can be successfully accomplished with standard PCs. However, when a detailed simulation is required, as it is the case to model the complex conditions' leading to eutrophication, the resolution has to be increased (e.g. 100x100 meter cells with more than 30 vertical layers) and more powerful computers are needed. Given our project requirements in CPU (>=2.5Ghz, few cores), memory (>12Gb) and disk (up to a few Terabytes), we could exploit Cloud services providers like EGI FedCloud to manage the entire workflow of data processing and analysis. Also, given the needs for the output we need a service able to support the storage of few terabytes and let us to transfer it using an easy and fast way.

Although some modules in Delft3D have been parallelized, we use a special kind of model named "Z model" where every single layer has the same size and so it cannot be easily parallelized. However, the model needs to be calibrated by running the simulation with different parameters or submodels (e.g. murakami or ocean sub-model for heat flux modelling), so EGI with a large and distributed offer of powerful resources provides the ideal platform for improving productivity. Moreover, water quality model can be parallelized so using multiple cores if available also improves performance.

Indicate the type and quantity of simulations needed in the Case Study, and how they are incorporated in the general workflow of the solution

Due to memory needs, Delft3D has been installed in our Supercomputer Altamira. In medium or high resolution models a single standard PC cannot run the hydrodynamic model so we need more powerful resources. Simulation takes up to 24 hours in processing high resolution models.

5.2 Technical description of simulation/modelling software

For each simulation package:

Identify the simulation software **DELFT3D, http://oss.deltares.nl/web/delft3d**

Provide a link to its documentation, and describe its maturity and support level See http://oss.deltares.nl/web/delft3d , maturity level is HIGH, support level through forum is GOOD.

Indicate the requirements of the simulation software (hardware: RAM, processor/cores, extended instruction set, additional software and libraries, etc.) **at least 12GB RAM**

Tag the simulation software as HTC or HPC **HTC and HPC for some simulation types**

List the input files required for execution and how to access them **around 50 different files, provided in the local directory**
Describe the output files and how they will be stored **several large files, again locally stored**
Reference an existing installation and performance indicators **TBD**
Specify if the simulation software is parallelized (or could be adapted) **YES, except for Z-model option (that is the one we currently use)**
Specify if the simulation software can exploit GPUs **NO**
Specify how the simulation software exploits multicore systems **UNKNOWN**
Specify if parametric runs are required **YES for the optimization of parameters**
Estimate the use required of the resources (million-hours, # cores in parallel, job duration, etc) **Estimation is 100K-1M core-hours per year, longest jobs are typically <24h**

5.3 Simulation Workflows

Describe if there are workflows combining several (HTC/HPC) simulations or simulations and data processing **Yes, the workflow will**

a) **combine the output of the hydrodynamic simulation to the input to the water quality**
b) **the water quality model requires a multi-parametric scan , ideally it could be an optimization scan (and starting with around 1000 points) (see below)**

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

User Story A): SME team wants to model the hydrodynamic behaviour of the water reservoir, to reproduce the thermocline and predict the onset and completion times of the water column stratification, with special interest in its final phase (september/october).

A.1) An ICT expert sets all the input parameters and maps to model in 3D the water reservoir using Delft3D. The input comes from ENV experts in the SME.

A.2) An ICT expert launches the simulation in HPC resources in the cloud. Each simulation is run from April to October for a given year, in 1h step, 3D.

A.3) The ICT expert checks with the ENV experts in the SME that the reference distributions (Temperature profiles, water level, etc) make sense and compare them to previous monitoring data (when existing). The SME experts use the programs output using EXCEL (currently, MATLAB or iPYTHON in the short future) on cloud resources via remote interactive portal (currently TeamViewer, next VNC/Thinlinc).

A.4) If the comparison makes sense, the output is stored. If a prediction is required, different runs are executed using previous year meteo scenarios to provide an estimation of the probability of an algae bloom. An statistical model is then applied to estimate the expectations.

A.4) If the comparison doesn't qualify or the output is not ok, the model is re-run varying the input parameters with larger incertitude, and a set of new measurements are transmitted to the BIO team.

User Story B): SME team wants to predict algae bloom based on model, and validate against previous year detailed analytical measurements

B.1) The ICT and BIO experts confirm the required input, including those from A). The experts also define the model metrics (difference in the evolution of relevant functions: oxygen profile, algae profile, etc.).

B.2) The ICT expert prepares the simulation of the DELWAQ module to estimate how the algae will grow. All input parameters to be tuned (around 60) are listed, including their uncertainty range. Also different input maps have to be used when there is an uncertainty (like for sediments). An initial random selection using uniform sampling (after variable normalization) for 1000 different simulation points is prepared, each one requiring around 5h. These 1000 MC points are used as a first estimation of the dependence with the different variables.

B.3) An optimization of the parameters, using 2010 data, is made using a multivariable gradient technique: different simulation results guide the convergence. An optimized set of parameters (including for example mortality rates) is obtained.

B.4) The DELWAQ model is run on different clima scenarios to estimate a prediction.

B.5) The model is run weekly and contrasted with real evolution. Prediction of scenarios are updated. Estimation of the impact of different pressures and corrective measures (for example, water level management, by pass of waste water treatment plants, artificial wetlands, cattle management) is done to propose the best path to the management authorities.

Provide details from conversation with the researchers' teams

The input is based in 5 years of cooperation, but some details refer to last analysis strategy defined in a meeting on 2nd june 2015.

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>**

Current user interfaces to check monitoring data run on a portal using Flex. The simulation output is handled using the Delft3D scripts and MATLAB visualization.

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 **Internal network connectivity is 10GB / FDR IB**
However the current problem is the connectivity for the final user to be able to copy the data (not feasible, o(100 GB)) or use remote visualization (feasible using VNC, or TeamViewer).
Describe the key requirements (availability, bandwidth, latency, privacy, etc) **Interactivity**
Specify any current issue (like last mile, or access from commercial, etc) **commercial access for SME researchers**

7.1.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the current use of each of these type of resources: size and usage **The simulation steps are run on local clusters and eventually in ALTAMIRA supercomputer (but parallel mode is usually not used, as z-mode is not parallelized)**
Indicate if there is any mode of "orchestration" between them **by hand, all input/output runs on the same group GPFS accounts**

7.1.3 Storage

Describe the current resources used **mainly GPFS storage at IFCA, O(few TB) in total**
Discuss the key requirements (I/O performance, capacity, availability, reliability, any other QoS indicator) **none seems key**

7.2 Short-Midterm Plans regarding e-Infrastructure use

Plans for next year (2016) and in 5 years (2020).
Migration to the Cloud for scalability/service offer for the SME (new water reservoirs)

7.2.1 Networking

Describe the proposed connectivity **Check VNC is adequate for remote access to large maps**
Describe new/old key requirements (availability, bandwidth, latency, QoS, private networking, etc) **VNC requirements**
Specify any potential solution/technique (for example SDN) **VNC**

7.2.2 Computing: Clusters, Grid, Cloud, Supercomputing resources

Describe the evolution expected: which infrastructures, total "size" and usage
if data from SENTINEL-2 starts to be used for complementary approach, we will need larger storage and computing power. If more water reservoirs are modelled, we may need more space

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?* **NO , we need a solution for external users (SME researchers)**

• *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* **Username/Password for external users, X509 certificates and IPA for local team**

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, ...)* **institutional authentication for CSIC / Universidad de Cantabria**
- *Preferred authentication technology, and requirements for support of multiple technology and credential translation services (e.g. SAMLX509 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)* **YES**
- *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.

Delft3D or any similar model requires HPC (as many other engineering applications)

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> Delft3D, Spreadsheets (MS Excel, LibreOffice), Matlab.**
Operating system requirements	**<input here> Linux/Windows**
Run libraries requirements	*Run API/libraries requirements (e.g., Java, C++, Python, etc.)* **For compiling Delft3D code:** **1. Subversion client required to communicate with the Subversion server.** **2. GNU Autotools (use Package Manager). Currently, fully tested binaries are build using Autoconf version 2.68 and Automake 1.11.1.** **3. GNU Libtool. Currently, fully tested binaries are build using version 2.4.2. Including libtool, libtool-ltdl and libtool-ltdl-devel** **4. GNU C++ compiler (use Package Manager), version 3.4.6** **5. expat-devel (Expat is a library to develop XML applications)(use Package Manager)** **6. GNU Fortran compiler 4.6.2. Alternative: Intel Fortran compiler, version 11.0 or higher. Binaries produced with the Intel compiler are faster. Currently, fully tested binaries are build using version 11.** **7. Mpich, version 3.1.2 or higher.** **8. Lex & Yacc** **9. OpenSSL.** **10. "readline-devel" (RedHat/Fedora) or "libreadline6-dev" (Debian/Ubuntu)** **11. Ruby** **12. NetCDF version 4.1.3 for Intel Fortran 11.1, version 4.3.2 or above for Intel Fortran 14.0.3**
CPU requirements (multithread,MPI, "wholenode")	**Single core for Z-models** **Multicore, MPI for sigma-models**

Memory requirements	**8GB in medium resolution (40x40 m)** **>12GB in high resolution (10x10 m)**
Network requirements	**IB for parallel execution**
Disk space requirements (permanent, temporal)	***Include the requirements for data transferring (upload and download of data objects: files, directories, metadata, VM/container images, etc.)*** **~50GB per model (medium resolution)** **Up to 1 TB per model (high resolution)**
External data access requirements	**Not critical, some external files have to be copied as input**
Typical processing time	**Medium resolution (Hydrodinamics+Water Quality) up to 12 hours** **High resolution (Hydrodinamics+Water Quality) up to 3 days**
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*** **Specific hypervisors? (cf. VMware image translation?)**
Other comments	**<input here>**
Relevant references or URLs	**http://oss.deltares.nl/web/delft3d**

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	