IS-ENES3 Deliverable D7.2
First external review report for ENES CDI services

Reporting period: 01/07/2020 – 31/12/2021

Authors: A. Spinuso (KNMI), S. Kindermann (DKRZ)
Internal Reviewers: F. Adloff (UREAD-NCAS), S. Joussaume (CNRS-IPSL)
Release date: 15/11/2021

ABSTRACT

A summary report on the review of the services of the ENES CDI conducted by a small team of external reviewers selected by the consortium. The reviews are organised for two different classes of services; the first addressing data and metadata dissemination, in particular the ESGF, ES-DOC, Climate4Impact Data Portals, while the second will cover aspects of the new IS-ENES Trans-national access service. The reviews are complemented by the description of the actions foreseen to address the issues raised by the advisory board.

Dissemination Level		
PU	Public	X
CO	Confidential, only for the partners of the IS-ENES3 project	

Revision Table			
Version	**Date**	**Name**	**Comments**
Document Structure and Contributors	28/05/2021	Alessandro Spinuso	Preliminary Structure
Reviewers on Data and Metadata collected and reported	11/10/2021	Alessandro Spinuso	Feedback requested to contributors to address reviewer's comment
Edited reviewers on Climate4Impact	15/10/2021	Alessandro Spinuso	Document structure reshaped to better represent reviewers' focus.
Executive Summary	21/10/2021	Alessandro Spiunso	Summary of the document structure.
Pre review refinement	4/11/2021	Alessandro Spiunso	General editing
Internal Review	8/11/2021	Alessandro Spinuso, Sylvie Joussaume, Fanny Adloff	General editing addressing revision remarks.

Table of contents

Executive Summary

The document collects and discusses comments and suggestions received by a selected group of reviewers, on the data and metadata services advertised from the pages of IS-ENES3 project. The format of the review was initially developed in M7.2 [3]. The introduction summarises the main review's objectives. It presents the selected reviewers and the guidelines they have been provided with. Section 2 and its subsections report on the reviewers' experience and remarks on the data and metadata services from the IS-ENES3 portal, while Section 3 focuses on the functionalities and documentation of the renewed version of the Climate4Impact portal. In the conclusions, we summarise the main insights gained by this exercise and follow-up actions.

1. Introduction

In this document we report on the assessment of the data and metadata pages and services published on the IS-ENES3 website. This is according to the first review's objective which has been previously elaborated in M7.2 [3], together with the indication of the reviewers. M7.2 also included a second review objective, which aims at setting a Scientific Evaluation Committee to support the selection of users of the new Trans-national access service. This objective has been excluded from this report, being the committee established and continuously involved in the selection and evaluation process.

For Objective 1, the selected reviewers are: Dr. Iuliia Polkova, teaching "Climate Modelling" to undergraduates and researching at the Institute of Oceanography at the University of Hamburg, and Dr. Laurens Bouwer impact researcher at the Climate Service Center Germany (GERICS). They have been provided with the following indications for the main review targets, which have been supported by additional documentation, links and support material.

1.   Please, visit the graphical user interfaces for inspection, searching the data catalogue and accessing the climate data and metadata and report on their functionality.
2.   In respect to the Climate4Impact portal, which is targeted towards the climate impact communities, please assess its fit for purpose.
3.   Infer with respect to understandability and completeness additional guidance material like "How to Use" and "Frequently Asked Questions" and direct support line.

As a preliminary consideration, although these review points were useful for the reviewers, they focussed on the aspects that were more relevant to their own interests and expertise, or where they found more space for improvements. We will report and comment on their remarks in the following sections.

2. Data and Metadata

To address point 1, reviewers were asked to check the services and pages published at the following links.

  I.   Data: https://portal.enes.org/data/enes%20infrastructure%20is-enes/enes_esgf/portals/data-portals
  II.   Metadata: https://portal.enes.org/data/data-metadata-service/metadata_services/services-on-model-doumentation-es-doc

Especially, for the metadata we proposed to address the ES-DOC services, the Climate and Forecast Convention pages and Impact Indicators metadata.

2.1 Data Pages and Services

Before going into the specifics, one reviewer addressed that the amount of information presented on the ENES website could be sometimes overwhelming. It was not obvious to the reviewer how the pages relates to each other. For instance, how does Climate4Impact relates to CLIPC, to ES-DOC etc.? One suggestion from the reviewer would be to publish on the portal a diagram where these relationships would be explicitly represented. This would also contribute to facilitating the review task.

About the data services in point I [1]. One reviewer defined these as the 'standard' way to search for data across nodes, suggesting that these are known and accepted by the community. However, the attempt from one of the reviewers to access data related to "Boundary Condition and Forcing Datasets for CMIP6", on the DKRZ node[1], resulted in receiving confusing feedback from the search pages, which immediately reported 0 results. The confusion was probably created by the selection parameters being hidden in several sub menus. The interface could be more intuitive if the access to detailed parametrisation is made immediately visible and not optional. We will forward this feedback to make sure we prevent such usability flaws in the implementation of the new ESGF Data Access Client, MetaGrid[2].

2.2 Metadata Pages and Services

For the metadata information linked from II, as already anticipated, the reviewers find it hard to relate the information published across the website and the external tools/services. In particular one reviewer considers it not handy to have metadata separate from the data. This assumes a lot of browsing and clicking making the task of finding the right metadata uncertain. The reviewer acknowledges that the description of the content seems to be all there, but the way to get to the needed information, and metadata is still considered a bit painful. Below we report the main comments and suggestions on the particular services addressed by the reviewers.

ES-DOC: Particular attention was given by one reviewer to ES-DOC [1]. This is a software ecosystem and a service, which facilitates both the provision and the consumption of documentation of the CMIP6 workflow. Its value is acknowledged by the reviewers, although one argues that the information could be for most users less relevant than for developers of the GCMs themselves. The reviewer agrees that the technical details are considered important to document, but for interpretation of the simulation data by users it is not immediately needed. What would be

---

[1] https://esgf-data.dkrz.de/search/input4mips-dkrz/
[2] https://metagrid.readthedocs.io/en/latest/

important though, for the reviewer, is to have better descriptions of simulations, such as explanation or guidance on the use of individual members of the ensembles, for instance.

A brief answer to this remake is that such a level of explanation is considered out of scope as far as ES-DOC is concerned in the context of IS-ENES3. Maybe it could be addressed in tasks that target outreach and training material.

The Dataset Errata Pages[3] is a particular service of ES-DOC that records and disseminates the changes applied to a particular dataset, and the rationale that led to a new version. Although this is considered to be a very important source of information, the reviewers find that the warnings associated with the datasets' changes should be tied to ESGF data services. We report here the questions raised in the assessment, justifying the need for an explicit and usable integration between the two:

"Which users will actually look at the errata page, before they download and handle their data? Or am I mistaken about the use of the errata page? What about the implications of errors and masking issues etc. on the use of the simulation data?".

These concerns can be partially addressed by a technical answer, since the errata pages and ESGF are actually linked via persistent identifiers (PID) assigned to the data. This solution allows the implementation of the two services to be decoupled, which brings advantages in terms of their performance and maintainability. It is documented in the CMIP6 user guide[4] (linked from the IS-ENES portal) that those who are interested should use these PIDs (stored in each netCDF file) to query the errata page. For instance, services such as C4I are already using the PIDs to resolve the ES-DOC pages associated with the data, offering users direct access to the documentation. Such a mechanism could be extended also to the *errata* pages. However, the technical "under usage" of PIDs and the associated *errata* information could be related to the lack of training and technical support material (ie. like notebooks). Producing such material and making it accessible to possible different consumers could improve the exploitation of the two services, in combination. We foresee technical "how to" pages, linked from the project portal *ENES Model Data and Metadata* page[5], as a first step towards better communication.

Impact indicators Metadata and CLIPC: One of the reviewers noticed that the documentation on the metadata, which is stored in a GitHub repository[6], is 5 years old. This raised the concern whether a review after so long could be of any help. It also suggests that such important content would gain more attention and communicate higher impact when this is represented by dedicated

---

[3] errata.es-doc.org

[4] https://pcmdi.llnl.gov/CMIP6/Guide/dataUsers.html

[5] https://portal.enes.org/data/enes-model-data

[6] https://github.com/cerfacs-globc/impact-indicators

and descriptive web pages, rather than by a direct link to the Git repository that shows little activity in the project life-span. Especially since these efforts are advertised and fostered by the current IS-ENES3 project. The reviewer dedicated some attention also to the CLIPC pages and their content. The main comment was on the functionality which allows users to visualise individual model runs on a map. The reviewer considers this function is not particularly useful, as it is unclear for the user how to select or compare the different individual model runs. Also, for instance rainfall data, are displayed with their raw title, which is rather technical and not easy to interpret. This suggests that metadata should have been already used in the tool to better describe the 'raw' content presented. Finally, the reviewer highlights that it should be considered what is the essential information to show herein the display menu structure of the CLIPC web pages. This last comment suggests that, being CLIPC a past and closed project, it could be beneficial to focus the information disseminated via the IS-ENES3 website, to the most important contributions of the project and associated background material, in order to bring more clarity to its expected use and value.

To conclude, both reviewers have been asked to comment on the **Climate and Forecast Convention material**, linked from the IS-ENES metadata pages. They did not report any particular remark, suggesting that no imminent action is needed in that respect.

3. Climate4Impact v2

The IS-ENES Climate4Impact[7] is an integrated portal that provides an easier and single access point to climate simulations and analysis methods. To address points 2 and 3 of the review guidelines, we have requested to review its new implementation Climate4Impact v2[8][1], whose innovative aspects have been introduced in [1]. This portal will replace the current website, in coordination with the release of the new components of the ENES-CDI architecture and software stack [2,1]. Reviewers have been provided with help pages that explain how to search[9] and analyse[10] the data via the portal. We address these two functionalities in the sections below.

3.1 Search and select for data

Both reviewers consider the selection process by the variable, time resolution, model, ensemble member straightforward. One highlighted the useful feature offered to users by the *scoped view*, which presents the explanation for the different and most commonly used variables. This is not available in the *extended view*, which is acceptable being many of the variables listed less popular,

---

[7] http://climate4impact.eu
[8] https://dev.climate4impact.eu/
[9] https://dev.climate4impact.eu/c4i-frontend/helpC4I
[10] https://dev.climate4impact.eu/c4i-frontend/helpSwirrl

or available for a limited set of simulations. In the experience of the reviewer, users are often interested in variables with high frequency data, eg. 3-hourly rainfall data. The capability of the search interface, combined with the *scoped view,* which puts on the forefront the most relevant variables and their availability is a very useful feature.

The overview of the experiments is a bit chaotic when it comes to the categories "Hindcast" and "SSP". The reviewer suggests calling the latter SSP and RCP scenarios (some are RCP simulations). The list under SSP is a bit difficult to read, although this is reflecting the new terminology as per CMIP6 and IPCC AR6 WG1, some guidance and explanation would be needed. For the members, it would be helpful if it would be explained what the different abbreviations ("r8i1p1f1" etc.) mean precisely to the user, and how to select them.

One reviewer raised the question about the availability for regional climate model data, being this of interest for many users. Will be useful to have guidance on the selection of either global or regional climate data. The CORDEX button is not yet active. This will be enabled as soon as the new ESGF Future Architecture IdP [1] will be available at the different CORDEX sites, as expected at the beginning of RP3.

The reviewers also reported that, being climate researchers, the data selection step was easy because of their familiarity with the name conventions, especially those used for the CMIP experiments. However, this might not be that easy for other users, who might need help from experts explaining what the naming of data and experiments actually represent. One reviewer noticed that for each dataset there are different data providers. These sometimes show differences in the data size for the same dataset, while it should be expected to be the same. This part is not intuitive and provokes a lot of guessing here. The user may wonder if the data from this particular experiment is really stored in five different places, including locations providing an incomplete dataset. We acknowledge this situation and will take into account whether any improvement can be done in that respect. Currently this is the information provided by the ESGF Network. In C4I we have no insight on the differences between the data provided by the different sites. We can't take judgment neither on their quality nor on the actual performance of the remote site. For instance, many sites could report a complete data availability, while being less performant or reliable when transferring the data. We will use these suggestions to give feedback to the implementation of the ESGF Future Architecture, especially concerning the ongoing improvements of its search capabilities [1], aiming at obtaining a better usability of the data selection.

To mitigate confusion, C4I allows users to narrow down the search on a subset of providers. At the time of writing this report, providers are described in terms of the analysis capabilities that are available on premises (ie. subsetting via remote WPS/OpenDAP services) and used by C4I's data reduction workflows. In that respect, one of the reviewers suggested a new functionality where

users would select a region of interest, for instance via a Shapefile or JSON file addressing a particular location and case study. This would be really helpful to be able to reduce the data load in an early stage. The realisation of this use case would largely be facilitated by a wider support of the data reduction workflows across more nodes of the federation, giving confidence to pursue the ongoing developments.

3.2 Data Analysis, C4I Workspaces via SWIRRL

C4I offers to its users analysis workspaces that are controlled via the SWIRRL API. This interface allows C4I to orchestrate remote cloud resources, obtaining interactive tools, such as jupyter notebooks, and workflows. Via SWIRRL users can stage data of interest and conduct reproducible analysis. According to one reviewer, SWIRRL generating notebook was rather fast. However, after reading the SWIRRL welcome notes, the data selected by the reviewer could not be found in the expected location, probably because of the data-staging workflow being still running. This can be monitored in the GUI, as explained in the help pages.

The help pages also explain that SWIRRL enforces a separation between raw data and user generated content. However, at the first attempt to create a notebook page, the reviewer's did not take that into account and tried to save the notebook in the same location where the selected data was being staged. This resulted in an error. Eventually, by exploring and gaining experience on the system and its user interface the reviewer succeeded to access the data and to upload a sample notebook. This was chosen among those stored in the dedicated repository of notebooks linked from the portal[11]. Finally the reviewer managed to execute the notebook on the selected data, obtaining the visualization of the results. This positive outcome demonstrates, as the reviewer commented, that the workspace capability is functioning. However, the experience reported suggests that the help pages could be refined, in order to explain the implications of workflow execution time and raw data management, aiming at reducing confusion, especially at the first use of the tool.

The other reviewer did not try to use the workspace functionality in every detail, but he found the help pages informative, with clear instructions, graphics etc. He recommends to keep some important information from the "old" site, such as the guidance & use cases page, where can be found explanations on the generic climate-data processing workflow and the roles of different experts[12].

---

[11] https://gitlab.com/is-enes-cdi-c4i/notebooks
[12] https://climate4impact.eu/impactportal/documentation/guidanceandusecases.jsp?q=generic_work_flow

4. Conclusions

The reviewers analysed at different levels of detail the services indicated in our guidelines, highlighting their value, shortcoming and suggestions for improvements. In general, the services are considered useful and functioning. However there is still a need for some guessing to accomplish their productive use. Reading about the way the reviewers approached a service, as well as the dissemination web pages and support material, gave us the indication on how functionalities could be made more usable by providing clearer instructions or by making them findable and more accessible. The most relevant information should be put on the fore-front, with the additional aid of more technical documentation and "how to". This seemed to be relevant especially for the combined use of data and metadata information, offered by the ES-DOC and Errata services.

More generally, one reviewer commented that for someone new to the climate research community, the different data catalogues might still appear somewhat cryptic, as well as the relationships between the metadata pages. One reviewer strongly suggests that publishing a diagram in the ENES website, depicting how the most important services are integrated, could help gaining the overview on how these and the information they deliver are linked to each-other. Another reviewer proposes to create short videos, briefly summarizing the use of the different tools, to help to anticipate what is possible and what one should expect.

5. References

[1] Spinuso, A. et al. *D10.2 - First release of the ENES CDI software stack* https://zenodo.org/record/4450012#.YWQsRqCxWNY

[2] Fiore, S et al. *D10.1 - Architectural document of the ENES CDI software stack,* https://zenodo.org/record/4309892#.X9CSbS2ZMn1

[3] de Castro, M. M. Bessembinder, J. *M7.2 - Set up of review committee and user selection panel* https://is.enes.org/documents/milestones/m7-2-set-up-of-review-committee-and-user-selection-panel/view