IS-ENES3 Milestone M3.4

Summary of workshops on standards

Reporting period: 01/01/2022 – 31/03/2023

Authors: Martin Juckes (UKRI), Lars Bärring (SMHI), Antonio S. Cofiño (UNICAN), Martina Stockhause (DKRZ)

Release date: 21/12/2022

ABSTRACT

This milestone document records a series of meetings held to discuss a broad range of data standards which are used to enhance the efficiency of IS-ENES services and to ensure and enhance FAIR (Findable, Accessible, Interoperable, Reproducible) data policies.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Table of contents

1. Objectives

One of the objectives of IS-ENES3 is to expand community standards which serve the ENES community. Standards in Earth system science are numerous and it is key to evaluate which can be used or adapted for IS-ENES use and which need to be developed. The existing standards used by ENES are not necessarily the same as the standards in user communities (e.g.
impacts) and a venue for exchanging on these standards is needed to collect user requirements and provide training.

The series of workshops described in this document targeted different areas of standards development and addressed areas of work which were associated with bottlenecks in operational procedures.

- Building on the CMIP6 data request combined requirements (technical and community engagement from 22 international consortia), workshops/meetings were organized that reviewed the challenges and achievements of CMIP6 and tried to provide recommendations for CMIP7.
- Technical experts on the specific metadata standards and on climate indices defined by Expert Team on Climate Change Detection and Indices (ETCCDI[1]) and Expert Team on Sector-Specific Climate Indices (ET-SCI[2]) were brought together to review the achievements made in IS-ENES2 and CLIPC[3], to solve key issues and to make recommendations for additions and extensions to existing community standards, e.g. the Climate and Forecasting Conventions, the CMIP Data Request. The outcome of this workshop informs the development of the tools and software building blocks of the C4I portal in WP10/JRA3, thus enabling production and dissemination of climate indices to be integrated into the IS-ENES infrastructure.
- Climate and Forecast Convention (CF) underpins interoperability of climate model data by providing a rich framework for detailed, machine-operable, technical information about the data products. The CF Convention is maintained by a community network to support the international climate research community. The workshop reviewed the status and outlook of CF and tried to formulate recommendations for next steps.

[1] https://www.wcrp-climate.org/etccdi

[2] https://public.wmo.int/en/events/meetings/expert-team-sector-specific-climate-indices-et-sci

[3] Climate Information Platform for Copernicus (http://www.clipc.eu/)

2. Description of work: Methodology and Results

• Vocabularies and standards for CMIP and other WCRP activities

Climate and Forecast Conventions (CF)

Two well attended virtual meetings of the CF community have been organised, in 2020 and 2021 respectively.

The workshops were organised by Uni. Santander, through the CF Committee, with strong engagement from US partners. The first meeting was planned as an event in Santander, but the disruption caused by the COVID 19 pandemic led to its being held online.
More details are given in the workshop report (http://cfconventions.org/Meetings/2020-Workshop.html, Cofiño, 2020).

For the 2021 workshop see: http://cfconventions.org/Meetings/2021-Workshop.html.

Both meetings attracted more than 100 participants, compared to a typical attendance of 30 - 40 at previous face to face meetings. Virtual meetings are proving to be a useful method of engaging with newer members of the CF community and more will be held in the future. The organising committee have agreed plans to alternate future virtual and face-to-face meetings. A face-to-face meeting was held in Santander in September 2022[4]. An online CF training workshop was held a few days in advance of the main event[5]. Recordings of presentations are available on the CF Conventions YouTube channel[6].

CMIP6 Data Request

A series of meetings on the CMIP6 data request were held from May 2020 onwards. The meetings brought together specialists from IS-ENES, PCMDI and WCRP to discuss priorities and direction of evolution for the request. This series of meetings fed into Milestone 10.2 "CMIP Data Request Schema 2.0".

- 20th May, 2020, online: Review of status and outline of requirements: This meeting was primarily about taking stock. Three objectives for the data request were agreed:
    - Develop and maintain a registry of defined physical parameters appropriate for the analysis of climate simulations and exchange of climate model output.
    - Develop and maintain a registry of file metadata specifications (including templates and parametric templates) to facilitate interoperable exchange of climate model output.
    - Support WCRP-endorsed model intercomparison projects by facilitating the specification of output requirements for climate model intercomparison and evaluation efforts.

[4] https://drive.google.com/drive/folders/18BzZixnRUhV8nDqbeRE_ltVvWTF1BzVG
[5] https://cfconventions.org/Training/2022-Training-Workshop.html
[6] https://www.youtube.com/channel/UCKLq7PCVonFJA0ec98SMFZA

The main outcomes were a request for a draft roadmap (and a follow-up meeting scheduled in July). The draft roadmap[7], which included strategic and technical requirements as well as an outline implementation plan, later evolved into a section of M10.2 (cited above). Meeting notes are available at: https://zenodo.org/record/3901310

- 2nd July, 2020, online: Data Request Review. This meeting identified:
    - Four key reusable assets in the data request:
        - Physical Variables, such as Near-surface Atmospheric Temperature, which can be reused in multiple CMOR Variables.
        - CMOR Grids, such as "plev8", which are also re-used in many CMOR variables
        - CMOR Variables which identify a physical variable and a sampling feature, including specification of spatial and temporal averaging as well as masking options.
        - Sampling Structures which define combinations of CMOR grids and other sampling parameters.
    - Reviewed initial paper on variable prioritisation.
    - Reviewed draft list of external vocabularies used in the data request.
    - Meeting notes are available at: https://zenodo.org/record/6347548
- 1th September, 2020, online: https://zenodo.org/record/6347579
    - The development of the Milestone 10.2 document was discussed and it was resolved to add a section on the strategic view and then send the document out for review (now published as Juckes 2020).
    - The creation of a registry for horizontal grids was discussed.
- 21st Sept. 2020, online: https://zenodo.org/record/7104958
    - There was further discussion of a grid registry, with reference to:
        - A list of structures supported by CMOR: https://docs.google.com/spreadsheets/d/1Njhnp-rAhQNb9lo4d_0YFUOdPttcPpjQec4R_7BMUH8/edit#gid=1854910047
        - A repository for collecting results: https://github.com/cmip6dr/GridsAndProfiles and a mindmap for collecting model grid information which was developed in the METAFOR project: https://github.com/cmip6dr/GridsAndProfiles/issues/1
- 8th October 2020, online. https://zenodo.org/record/6347569
    - Discussion of grids: comparing, referencing and guidance
    - Recommendations on avoiding multiple "fx" fields, completion of grid guidance, and harmonization of file naming.

[7] https://docs.google.com/document/d/15ikAVVOCGXFZqLNQ-xsW2673SAnbUJUDwwwF_ZUhB1U

The outcomes of the first series of meetings led to the creation of M10.2 "CMIP Data Request Schema 2.0"[8]. M10.2 sets out strategic and technical requirements and reviews a range of relevant ISO standards. From these standards and requirements, a set of design principles is derived, specifying a layered approach to the development of a schema which can provide both an element of stability to support parallel development of software tools and flexibility to adapt to new requirements.

- Supporting Documents and Correspondence

Two further documents which were contributed outside the meeting series provide supporting material:

- Analysis of download statistics from HighResMIP from March 2020 to February 2021 shows that the model-level, 3d fields dominate.[9]
- Meteo-France provided feedback by email, reproduced in Appendix 1.

Ongoing Discussions at PCMDI and the WGCM Infrastructure Panel

The meetings above, convened by the IS-ENES3 project team, have been complemented by activities listed here which were led by the WIP with contributions from the IS-ENES3 team.

- Climate model data sharing beyond CMIP6: CMIP6Plus

This discussion paper[10] looks at the potential for continuing a CMIP-like activity after CMIP6, with data publication for MIPs which follow a rationalised and simplified version of the CMIP6 protocol.

- Harmonizing Metadata and Filenames Across CMIP Eras and WCRP Activities

The discussion paper[11] on harmonization across CMIP era and WCRP activities aims to revise the Data Reference Syntax to ensure greater continuity between CMIP eras.

[8] Juckes, Martin. (2020). CMIP Data Request Schema 2.0. Zenodo. https://doi.org/10.5281/zenodo.4287148
[9] https://github.com/PRIMAVERA-H2020/stream2-planning/blob/master/esgf_vars_downloaded_20210204.ipynb
[10] Live document:
https://docs.google.com/document/d/1KSJ5zkKOnN1EgfkIyjc1F4tE4LAroGkJ7y7t6shD9pg
(and https://docs.google.com/document/d/1rnSIuZgjfcAD5GICcQ5R1TSaKDR7ijB89WksoLyNG1A – check comments to see which is current); Copy taken 23 Sep 2022:
https://docs.google.com/document/d/1aulgE19IirK66bA-vNwsN2SgGDQwWAs9PgzoNTfYwGA/edit?usp=sharing
(see version Ref01).
[11] Live document:
https://docs.google.com/document/d/1WLxaxQmGuAf757lqgX1bVp7T1U08VO8IuR2esD11jvw/edit ; Copy taken 23 Sep 2022:
https://docs.google.com/document/d/1l2mrBQh2kUCHsLtjMrhCx9mtFPm2ulM47mOLL7hTKTk/edit?usp=sharing
(see version Ref01).

- ## Data Standards for Climate Indices

A series of online meetings focussed on specific issues related to how well-established precipitation indices are described in various commonly used reference documents and implemented in popular software tools. This resulted in a common suggestion for how to unify these differences that is currently drafted. The new index definitions have recently been included in the CLIX-META[12] repository of climate indices definitions. In parallel, some general aspects of climate indices have been identified as not possible to describe at the required detail using the currently existing CF mechanisms. Suggested extension to CF has been worked out and raised as discussion issues in the CF github forum. Several of these issues have generated extensive interaction with many contributions from across the CF community, not least from members also being partners in IS-ENES3. These discussions have deepened the CF community understanding both of some aspects that are specific to climate indices, and to some more fundamental aspects where CF in general will benefit from having more specific mechanisms for describing the data. The outcome of all this in terms of further development of the climate indices standard will be described in the upcoming revision of Milestone M10.3.

Two online meetings were held with a focus on eastern European users. The outcomes of these meetings and associated surveys are reported in M3.5: Workshop on climate indices - Eastern Europe perspective. The work on data standards for climate indices has progressed along two parallel lines: the development of a database covering the key sets of climate indices, and a range of user interaction activities. The former activity that is reported in Milestone M10.3, focuses on the technical development of a community-based data standard for climate indices. The direction of the technical development has been informed by the user interaction, which aims to span the gap between stakeholders producing and using climate indices as part of climate information and climate service activities, and the data-oriented communities knowledgeable about climate metadata in general and the Climate and Forecasting (CF) Conventions in particular. The former stakeholder category often makes use of some software tool for calculating and/or analysing climate index datasets without being particularly interested in the highly technical and formalistic aspects that by necessity is an intrinsic component of metadata standards. For the latter group, many of the concepts and elements that are central for describing climate indices are new or not a natural part of the existing metadata standards. Examples of user interaction activities related to the first group of producers and users of climate indices are reported in Milestone M3.5, as well as informal discussions with Expert Team on Sector-Specific Climate Indices (ET-SCI[13]) representatives at University of New South Wales, Australia, in charge of one of the key set of climate indices, and persons involved in producing Copernicus C3S sectoral climate information. User interactions with the latter group of stakeholders include presentations and targeted breakout discussions at the two CF Workshops 2020 and 2021 (see Section 3 below), as well as online discussions in CF github issues

[12] https://github.com/clix-meta/clix-meta
[13] https://climpact-sci.org/about/project/

- ## File Metadata in the IPCC AR6 WGI Atlas

The inclusion of a dynamic climate projections atlas within the body of the IPCC Assessment Report for the first time in AR6 created a range of challenges. IS-ENES3 partners supported the work of CSIC-UC which led to a solution implemented through the IPCC Data Distribution Centre. Meetings were held on 21st October and 19th November 2021[14].

3. Difficulties overcome

- ## Engagement

The value of good community standards is widely recognised, but creation of good standards depends on having broad community involvement in the development of requirements and approaches. This involvement is often difficult to achieve, particularly in the context of international standards with multiple communities.

IS-ENES3 has achieved good levels of participation by exploiting the breadth of the project consortium and community links:

- WGCM Infrastructure Panel: established by the WCRP Working Group on Coupled Models, the WIP brings together representatives of key CMIP infrastructure elements with a mandate to oversee the coordination of CMIP infrastructure activities.
- Data Request Mailing Lists and Github issues[15]: The mailing lists include contact points from Model Intercomparison Projects participating in CMIP6 and from all the modeling groups providing data. The discussions in github were used to refine details of the data request and correct errors and issues of clarity. Through these emails and lists, the Data Request activity engaged with 23 independent CMIP-endorsed Model Intercomparison Projects (MIPs). These include 20 climate science MIPs and 3 MIPs supporting climate assessment and climate impact science: the CMIP6 CORDEX MIP (defining the CMIP6 data needed as inputs for the CORDEX downscaling activity), ScenarioMIP (defining future scenarios for assessment studies) and the Vulnerability, Impacts, Adaptation, and Climate Services Advisory Board (VIACS-AB). VIACS-AB is representing the interests of a large number of climate impacts and climate services experts and consortia, including a number of impact model intercomparison projects, such as AgMIP and ISIMIP.
- CF Conventions discussion forum.

[14] https://docs.google.com/document/d/1Saqn9OmkrCckCoIZXLwQc2BqZPHXkSrHsekSpPVFjko

[15] https://github.com/cmip6dr/CMIP6_DataRequest_VariableDefinitions/issues, https://github.com/cmip6dr/Request/issues and https://github.com/cmip6dr/cmip7_forward_look/issues

- ## COVID Impact

The COVID pandemic changed the nature of the workshop series, with more work being done online. This lowered costs, but the core work could still be completed. Informal assessment indicates a balance of advantages associated with reduced travel overheads and disadvantages associated with the reduced quality of interaction when meeting online instead of in person.

- ## Requirements capture

Achieving clarity on the requirements for standard in sufficient details to enable reliable implementation is an enduring problem. Technical complexity in the content and process around data standards presents a barrier to meaningful stakeholder engagement. The requirements capture through expert elicitation, rather than directly from users, worked effectively for the Data Request schema 2.0. The experts provide requirements based on an understanding of the role of standards in supporting the user community. Achieving consensus and/or closure of the expert elicitation process remains a challenge.

4. Next Steps - Recommendations

The work documented here needs to be taken forward into the process for setting up standards and protocols for the next phase of the CMIP project. Ongoing engagement between ENES members and the CMIP governance structures is providing a sound basis for achieving this.

There is a strong likelihood that evolution in model technology (e.g. advance in representing dynamic ice sheets) and societal pressure (e.g. the need for more information about increasing fire-risk and the feedback of fires on the climate) will bring new communities into CMIP and expand the scope of the domain that needs to be covered by standards. The standards bodies need to build in capacity and flexibility to deal with an expanding user base.

The Horizon Europe ClimatEurope2[16] project provides an excellent framework for continuing and broadening the discussion on standards related to climate data. Several ENES partners are involved.

Task Teams convened to support the CMIP7 process by the CMIP International Project Office (CMIP IPO)[17] provided an excellent opportunity to feed knowledge and expertise related to data standards into the global conversation about CMIP7 and deepen community engagement in the standards process.

[16] https://climateurope2.eu/

[17] CMIP IPO Director, personal communication

Appendix 1: Meteo-France Feedback

By gathering various data requirements into a single interface, the Data Request is a useful and even necessary tool for an exercise of the scale of CMIP6. At CNRM-CERFACS and IPSL we used a python tool (dr2xml[18]) based on DreqPy that generates XIOS[19] XML files (model output configuration, including temporal and spatial operation encoding, gluing of attributes) according to the CMIP6 DR. This tool was specifically developed for CMIP6 needs . We really appreciated the existence of a machine-readable Data Request without which we would not have been able to reach such a level of automation. We are aware that DreqPy was not initially designed to be used in that way, the normal usage being the xls file writing. But such documents do not contain all the information we need for XIOS. From what we have heard from other CMIP6 modeling groups, there is a real need for such an interfaceable Data Request. In this perspective and for general purpose, we identified potential sources of improvement:

The Data Request development should be steered in order to rationalise and harmonise its content on one hand and to operationalise the distribution of the versions on the other hand. To aim that, it would be necessary to:

- Define a core set of essential variables to be produced in all simulations. One or more (but not many) additional groups of variables can be defined to allow for further analysis. None of the essential variables must be missing and accessory (not very useful and/or easily re-computable) variables should not be requested (day cycle, climatologies…).
- Harmonise the diagnostics requested between members of the same experiment, between experiments of the same MIP as well as between experiments of different MIPs that are intended to be compared.
- Limit the use of time-slices and harmonise it when they are really useful so that it is possible to compare the different simulations.
- Harmonise and standardise both tables' name (currently, some indicate the associated realm and the output frequency, others only the frequency or frequency and time method), and variables' name. It will ease the identification of interesting elements.

The ergonomics of the Data Request needs to be reviewed:

- The search interface of the Data Request does not allow us to easily know the amount of variables requested for a given simulation at a given resolution. The use of numerous RequestVarGroup and time-slices makes this information all the more difficult to obtain. As things stand, it is difficult to see if any variables are missing.

[18] https://github.com/rigoudyg/dr2xml
[19] http://forge.ipsl.jussieu.fr/ioserver/wiki/training_2020

- The names of the Data Request entries should be renamed to make them more explicit (e.g. name of dimensions provided by altlabels).
- Tools such as dr2xml have made it possible to hide the difficulties of using the Data Request for purposes other than simply generating excel files, but there are few of them and they have been lacking for CMIP6.

For a post-CMIP6 use, it would also be necessary to:

- Set up tools to check the Data Request (to avoid, for instance, regressions, loss of fundamental variables during an upgrade).
- Keep the Data Request interfaceable with code but re-think it as a python module with functions allowing quick access to attributes (e.g. spatial_shape, cell_measures) or to list of variables or other.
- Be able to view in an html interface all the versions and not just the latest. It would be useful to be able to compare two versions.

In addition, in the CMIP6 specifications, two attributes should be added to the files produced:

- an identifier for the forcings used (the forcing_index "f" does not give any information to users).
- the version of the dataset published (which is not always present in the downloaded data and which could be hard to recover).

Last but not least, it would be extremely useful to have a joint procedure with the CF convention for defining new variables.

References

Cofiño, A.S. (2020, June 11). 2020 CF Workshop - Wrap-up and Conclusions. 2020 CF Workshop, Virtual. Zenodo. https://doi.org/10.5281/zenodo.3900253

Juckes, Martin. (2020). CMIP Data Request Schema 2.0, IS-ENES3 Milestone M10.2. https://doi.org/10.5281/zenodo.4287148