Status of Climate Impact Indicator global metadata and data reference syntax at the end of CLIPC (November 2016)

Ruth Petrie (STFC), Milka Radojevic (CERFACS)

CLIPC Guidelines

The final version (v1.5) of the "Metadata standards for climate impact indicators for publication in the CLIPC portal" that were produced as part of the Climate information platform for Copernicus (CLIPC; http://www.clipc.eu/) project can be found at https://github.com/cerfacs-globc/impact-indicators. The document contains guidance on the provision of NetCDF global attributes, file-naming conventions and data reference syntax (DRS) for climate impact indicators (CII). The tables from the document are also included in the repository as a spreadsheet for ease of use. The guidelines detail the importance and relevance of the global metadata and DRS, along with some examples from the CLIPC project. In this status document some unresolved problems with the CII metadata are outlined and some suggestions for improved are detailed.

Items that require attention before re-use

At the end of the CLIPC project the *2nd CLIPC/IS-ENES2 Workshop on Metadata/DRS for climate indices* was held in Brussels on the 17th Oct 2016 and two items were identified as in need of immediate attention.

1. **Global attribute "*title*"**: The global attribute "*title*" was not used in a consistent way by data providers. The inconsistent nature of the title meant that in the CLIPC portal the "*title*" was not able to be shown to users when selecting datasets, instead the filename DRS was shown. While this did not prevent users from accessing the data it made understanding which dataset was being selected overly complex. It is suggested that a consistent form of title be agreed upon so that data providers provide all the necessary information in the title in a consistent way. This could be achieved in a number of ways:
   (a) providing examples of good and bad titles,
   (b) clearly specifying the necessary contents of the title,
   (c) providing a formulaic way of constructing a title with only the necessary information.

2. **Global attribute mapping to DRS of "*GCMName*":** The dataset identifier required for publication in the Earth System Grid Federation (ESGF) is constructed of DRS elements. There is a corresponding global attribute for each DRS element. In the case of the DRS element "*GCMName*", it was mapped to the global attributes "*invar_rcm_model_driver*", "*invar_reanalysis_id*" and "*invar_gcm_model_id*" as it was anticipated that only one of these attributes would be specified. However "*invar_rcm_model_driver*" and "*invar_gcm_model_id*" are not mutually exclusive and in some cases both were supplied. This caused problems in the DRS checker developed by KNMI in validation and generation of the DRS. It is suggested that in order to resolve this that a strictly 1-1 mapping of global attributes and

DRS elements be adhered to. This could be achieved, for example, by the addition of a new global attribute that mapped solely to the DRS element "*GCMName*".

Multiple input sources

The DRS element "*source_data_id*" was stretched beyond it's initial purpose. Initially this DRS element was intended to provide an identifier of the source of the data used to create the CII, e.g. the name of a satellite or the name of pre-existing dataset. However this became ever more complicated during the CLIPC project and the "*source_data_id*" element is now of a complex construction covering multiple input sources and ensembles of data (see the Metadata standards for climate impact indicators v1.5 document at https://github.com/cerfacs-globc/impact-indicators for full details). It is recommended that this be looked at in detail before any new data is produced using the metadata specifications guidance v1.5 used within CLIPC.

The DRS checker

The global attribute "*source_data_id*" uniquely maps to the DRS element of the same name. This was used in the DRS checker to decide which one of the two different CII-DRS to check consistency against. However this is not a reliable methodology, if a data provider incorrectly provides this element then the DRS checker could fail. It is suggested that while there are two DRS for CIIs that an additional global attribute be included which states which DRS the data providers intend their data to published under.

It is also important to note that the DRS checker only checks for consistency and does not check for Climate and Forecast (CF) metadata conventions (http://cfconventions.org/) compliance. A separate tool known as the CF checker (https://github.com/cedadev/cf-checker, http://puma.nerc.ac.uk/cgi-bin/cf-checker.pl) exists that will check for CF-compliance of the file and variable metadata. The metadata checking is therefore a two stage process, firstly checking the internal consistency with the DRS checker developed in CLIPC and checking the file for CF compliance.

Longer term opportunities/aims

1. **Create a centralised repository (GitHub) of all work relating to metadata for CIIs**.
   By placing the guidance notes for global attributes and DRS within a centralised repository it will be easier for data providers and publishers to follow the updates to the versioning of the guidance notes. The changes to the guidance note versions should contain statements as to what was updated either within the document or as separate document alongside the guidance notes.
   Within the repository the DRS checker developed by KNMI and controlled vocabularies relating to CIIs should also be stored. This will allow data providers to browse the vocabularies to see what terms they can use and publishers the ability to add to the vocabularies and use a dynamic DRS checking process.

2. **A single CII DRS**.
   The approach taken within CLIPC provided two data reference syntax, one for model data and one for general usage. This is far from optimal and effort should be put into the

development of a single DRS which covers all CIIs. This is not a trivial task and is likely to involve a significant community effort to gain agreement between data providers and publishers. If this is achieved it will make wider publication of CII data through infrastructures, such as ESGF far simpler and more useful for users.

3. **Climate statistics software to produce files with agreed attributes.**
   At present it is the responsibility of the data producer/provider to ensure that the datafiles have the correct global attributes. However given the inconsistent way in which this has been provided within CLIPC it would be of great benefit to the community if the statistical software that is managed by the community, e.g. icclim produced files which had all the agreed global attributes in the file, correctly filled where possible else left blank for data providers to fill in.
   An alternative to this would be a simple tool that manipulated a standard CF-compliant NetCDF file adding the agreed attributes (potentially with user specified entries). This would enhance consistency and therefore reliability.