CP4CDS Data quality control checks performed prior to ESGF publication

Ruth Petrie, Martin Juckes, Ag Stephens

13th July 2017

The climate projections data for the Climate Data Store (CP4CDS) will provide a subset of quality controlled CMIP5 data, where files have metadata amended this will be noted. Two primary tools will be used to quality control the metadata of the CMIP5 data, the CEDA compliance checking (CC) tool and the Climate and Forecast (CF) conventions checker. These tools will provide objective analyses of file metadata, providing a quality score to every file provided to the CDS, these tests are fully outlined in this document. The file level metadata quality control scores can then be aggregated in different ways such as to provide quality control scores relating to e.g. variable level datasets, ensembles or simulations. Further subjective analyses of the data may be made available through the use of quality control plots where variable values across multiple models can be compared and gross errors (such as out of normal range or sign errors).

The objective of the quality control tests are to provide information on files in the CEDA CMIP5 data archive that are to be provided to the CDS (that are currently published to ESGF). The information provided will detail whether the files contain metadata errors. Where it is possible and within licensing arrangements data will be corrected before being published to the CP4CDS ESGF index node. Details of corrections applied will be made available within the file and held in a database at CEDA that will be made available to the CDS. This will ensure consistency and confidence across the CP4CDS subset of CMIP5 data.

It is important to note that passing of these quality control tests should not be confused with validity: for example, it will be possible for a file to be fully CF compliant and have fully compliant CMIP5 metadata but contain gross errors in the data that have not been noted. It will be possible for users of these data to report errors through the CDS to CEDA and the data will can then be marked as reported in error. This could be done through the CDS helpdesk, the CEDA CP4CDS QC web tool or CHARMe.

The metadata tests performed on each file are outlined in detail below; some of the quantities tested are required for data management reasons, others to impose consistency across the archive to facilitate scientific analysis.

The tests outlined in this document are derived from data checks prior to ESGF publication of CORDEX data, 28th Oct. 2013, in turn derived from https://www.enes.org/data/projects/documents/quality-control-checks-for-the-cordex-archive

1. File name

Table 1.1: File name		
Test id	**Test**	**Tool**
T1.1	File name must consist of 8 or 9 components, separated by "_", followed by ".nc".	CEDA-CC

Table 1.2: File name components				
Test id	**Position in file name**	**Component id**	**Test**	**Tool**
T1.2a	1	VariableName	Contained in variable list.*	CEDA-CC
T1.2b	2	Realm	Contained in domain list.*	CEDA-CC
T1.2c	3	ModelName	Contained in driving model list.*	CEDA-CC
T1.2d	4	CMIP5 ExperimentName	Contained in experiment list.*	CEDA-CC
T1.2e	5	CMIP5Ensemble Member	Of the form "rxiypz", for integers x,y,z.	CEDA-CC
T1.2f	6	TimeRange	Of the form"yyyy[MM[dd[hh[mm[ss]]]]][-suffix]" Is required if Frequency is not "fx", see also Table 1.3	CEDA-CC

*: CMIP5 controlled vocabularies see references at end of document.

<table>
<tr><th colspan="6">Table 1.3: Time range</th></tr>
<tr><th>Test id</th><th colspan="4">Test</th><th>Tool</th></tr>
<tr><td>T1.3a</td><td colspan="4">TimeRange (see T1.2f) must consist of two integers ("Start" and "End") separated by "-" where start and end have the form: "yyyy[MM[dd[hh[mm]]]]".</td><td>CEDA-CC</td></tr>
<tr><td rowspan="3">T1.3b</td><td colspan="4">The rules governing "Start" and "End" depend on the Frequency (see T1.2f):</td><td>CEDA-CC</td></tr>
<tr><td rowspan="2">Frequency</td><td rowspan="2">Pattern (for integers y, M,d,h,m): start and end**</td><td colspan="2">Valid temporal values</td><td rowspan="2">CEDA-CC</td></tr>
<tr><td>Temporal element</td><td>Valid values</td></tr>
<tr><td>yr</td><td>yyyy</td><td>yyyy</td><td>1800-2500</td><td></td></tr>
</table>

	mon	yyyyMM	MM	01-12	CEDA-CC
	day	yyyyMMdd	dd	01-31	CEDA-CC
	6hr	yyyyMMddhh	hh	00-24	CEDA-CC
	3hr	yyyyMMddhhmm	mm	00-60	CEDA-CC
	sub-hourly	yyyyMMddhhmm			CEDA-CC
T1.3c***	Time axis data values have regular increments (for monthly data the increments may vary between 28 and 31 days).				CEDA-CC
T1.3d***	Consistency between time axis values and "Start" and "End" in file name.				CEDA-CC
T1.3e***	Multi-file dataset time series consistency, continuity of time axis between files.				CEDA-CC

** Restrictions in first and 2nd column do not apply to the first and last file in a series respectively, but the length (i.e. the number of characters) of both time range elements should be equal.

*** Test is under development.

T1.3c: Confirm if sub-hourly data has to have a regular time increment.

T1.3d: need to check guidance/rules on tolerances …possibly only check to nearest day

2. Required global attributes

Table 2: Required global attributes

Test Id	Global attribute name	Test	Tool
T2.1	contact	Free text	CF-Checker
T2.2	Conventions	e.g. 'CF-1.4'	CF-Checker
T2.3	creation_date	Should specify the date of creation of the file. If a file which has previously published in ESGF is modified and re-published, this attribute should be updated.	CF-Checker
T2.4	experiment	String providing a title for the experiment*	CEDA-CC
T2.5	experiment_id	A short string identifying the experiment*	CEDA-CC
T2.6	forcing	a string containing a list of the "forcing" agents that should cause the climate to change in the experiment	CEDA-CC
T2.7	frequency	Temporal frequency of output*	CF-Checker CEDA-CC
T2.8	model_id	a string containing an acronym that identifies the model used to generate the output.	CEDA-CC

T2.9	initialisation_method	an integer (≥1) referring to the initialization method	CEDA-CC
T2.10	institute_id	a short acronym describing “institution”*	CEDA-CC
T2.11	institution	character string identifying the institution that generated the data	CEDA-CC
T2.12	modeling_realm	Equal to filename component CMIP5ExperimentName	CEDA-CC
T2.13	parent_experiment_id	experiment_id indicating which experiment this simulation branched from.	CEDA-CC
T2.14	parent_experiment_rip	identifier indicating which member of an ensemble of parent experiment runs	CEDA-CC
T2.15	physics_version	an integer (≥1) referring to the physics version used by the model	CEDA-CC
T2.16	product	“output”, which indicates that the data you are writing is model output.	CEDA-CC
T2.17	project_id	"CMIP5" for CMIP5.	CEDA-CC
T2.18	realization	an integer (≥1) distinguishing among members of an ensemble of simulations	CEDA-CC
T2.19	source	character string fully identifying the model and version used to generate the output	CEDA-CC
T2.20	table_id	should be assigned a character string identifying the CMIP5 Requested Output table where this variable appears	CEDA-CC
T2.21	tracking_id	a string that is almost certainly unique to this file and must be generated using the OSSP utility	CEDA-CC

3. Optional Global Attributes

Table 3.1: Optional Global Attributes

Test id	**Global attribute**	**Test (if attribute is present)**	**Tool**
T3.1	comment	A character string containing additional information about the data or methods used to produce it.	CF-Checker
T3.2	history	A character string containing an audit trail for modifications to the original data.	CF-Checker
T3.3	references	A character string containing a list of published or web-based references that describe the data or the methods used to produce it.	CF-Checker
T3.4	title	→ A sample title is: 'IPSL-CM5 model output prepared for CMIP5 historical”	CF-Checker CEDA-CC

4. Dimensions

Table 4: Dimensions			
Test id	**Dimension name**	**When required:**	**Tool**
T4.1	time	if frequency not “fx”;	CF-Checker CEDA-CC
T4.2	plev	Units of pressure	CF-Checker CEDA-CC
T4.3	height	Dimensional height or depth axes must always explicitly include the units, includes positive direction up/down.	CF-Checker CEDA-CC
T4.4	lat	Variables representing latitude must always explicitly include the units attribute.	CF-Checker CEDA-CC
T4.5	lon	Variables representing longitude must always explicitly include the units attribute.	CF-Checker CEDA-CC

5. Dimension attributes

Table 5: Dimension attributes			
	Attribute name	**Value or rule**	**Tool**
Test id	**time attributes (if dimension time present)**		
T5.1a	units	“days since 1949-12-01 00:00:00Z” or equivalent (e.g. “days since 1949-12-01”)	CF-Checker CEDA-CC
T5.1b	standard_name	time	CF-Checker CEDA-CC
T5.1c	long_name	time	CF-Checker CEDA-CC
T5.1d	calendar	Must be a valid CF Convention calendar name.	CF-Checker CEDA-CC
T5.1e	time_bnds	Required for non-instantaneous fields (time means, sum and extrema), must equal “time_bnds” if present. See also T7.3.	CF-Checker CEDA-CC
plev attributes (if dimension plev present)			
T5.2a	units	Pa	CF-Checker

T5.2b	standard_name	air_pressure	CF-Checker
T5.2c	long_name	pressure	CF-Checker
T5.2d	positive	down	CF-Checker
T5.2e	axis	Z	CF-Checker
T5.2f	bounds	Required for variables clh, clm, cll; must equal “plev_bnds” if present. See also T7.3.	CF-Checker CEDA-CC
height attributes (if dimension height present)			
T5.2a	units	m	CF-Checker
T5.2b	standard_name	height	CF-Checker
T5.2c	long_name	height	CF-Checker
T5.2d	positive	up	CF-Checker
T5.2e	axis	Z	CF-Checker
lat attributes (if dimension lat present)			
T5.4a	units	degrees_north	CF-Checker
T5.4b	standard_name	latitude	CF-Checker
T5.4c	long_name	latitude	CF-Checker
lon attributes (if dimension lon present)			
T5.5a	units	degrees_east	CF-Checker
T5.5b	standard_name	longitude	CF-Checker
T5.5c	long_name	longitude	CF-Checker

6. Variable name and attributes

Table 6a: Variable name and attributes			
Test id	**Quantity tested**	**Value or rule**	**Tool**
T6.1	Variable name	Same as variableName component of file name (see T2.1a).	CEDA-CC
Test id	**Variable attribute**	**Value or rule**	
T6.2	standard_name	From the CF Standard name table	CF-Checker
T6.3	units	From the CF Standard name table	CF-Checker
T6.4	long_name	From the CF Standard name table	CF-Checker
T6.6	cell_methods	Must contain “time: mean” for time averaged fields, or “time: point” for instantaneous fields. See also Table 8 for special cases of this attribute. From CF conventions.	CF-Checker CEDA-CC

Table 6b: Special cases for cell_methods attributes			
Test id	**Variable**	**cell_methods string**	**Tool**
T6.9a	<var>min	time: minimum within days time: mean over days	CF-Checker
T6.9b	<var>max	time: maximum within days time: mean over days	CF-Checker
T6.9c	<var>d	time: sum within days time: mean over days	CF-Checker

7. General rules

Table 7: General rules		
Test id	**Rule**	**Tool**
T7.1	Variables must be single precision	CF-Checker CEDA-CC
T7.2	Dimensions must be double precision.	CF-Checker CEDA-CC
T7.3	“plev_bnds” and “time_bnds” variables, if present, must follow the CF Convention rules for bounds variables.	CF-Checker CEDA-CC

8. Exceptions

Table 8: Exceptions: the following are not considered for the CP4CDS project	
Test id	**Rule**
T8.1	Identified by the CF-checker: WARNING (6.2): cell_measures referring to variable 'areacella' that doesn't exist in this netCDF file. INFO (6.2): This is strictly an error if the cell_measures variable is not included in the dataset.
T8.2	Identified by the CF-checker: INFO: attribute 'history' is being used in a non-standard way
T8.3	Bounds on non-coordinate variables

References

[1] http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_Appendix1-1.pdf

[2] http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

[3] https://github.com/PCMDI/cmip5-cmor-tables

[4] http://cfconventions.org/standard-names.html

[5] http://cfconventions.org/cf-conventions/v1.6.0/cf-conventions.html

[6] https://verc.enes.org/ISENES2/ Infrastructure for the European Network of Earth System Modelling.