The ETH Zurich CMIP6 next generation archive: technical documentation

Lukas Brunner, Mathias Hauser, Ruth Lorenz, and Urs Beyerle

ETH Zurich, Institute for Atmospheric and Climate Science, Universitätstrasse 16, 8092 Zurich, Switzerland

March 31, 2020

Abstract. The CMIP6 next generation (CMIP6ng) archive is an update to the raw CMIP6 archive as provided by the Earth System Grid Federation (ESGF). It introduces a range of additional checks for the processed variables and their main dimensions (time, longitude, latitude) as well as incremental optimizations in the file structure and consistency of the files from different institutions. It provides models in their native horizontal resolution and on a common 2.5°×2.5° longitude-latitude grid. Files are provided in monthly time resolution and as annual means calculated from the monthly means. In addition, selected variables are available in daily resolution. Here, the differences between the CMIP6ng and the raw CMIP6 archives are presented, the processing structure is detailed and a list of checks is given.

Citation. Brunner L., M. Hauser, R. Lorenz, and U. Beyerle (2020). The ETH Zurich CMIP6 next generation archive: technical documentation. DOI: 10.5281/zenodo.3734128.

Contact. cmip6-archive@env.ethz.ch

1 Introduction

The sixth phase of the Coupled Model Intercomparison Project (CMIP6) (https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6) is a coordinated effort for collecting, organizing and distributing climate model output. Participating models perform a common set of experiments, including the mandatory DECK (Diagnostic, Evaluation, and Characterization of Klima) and historical simulations as well as a range of optional Model Intercomparison Projects (MIPs). In addition, CMIP6 provides common standards for documentation and output characteristics (Eyring et al. 2016). Despite these efforts, however, further post-processing can be beneficial to, for example,...

... address differences in the file structure (such as one file versus multiple files per experiment)

... address inconsistencies in the dimensions (such as the time periods covered)

... flag files with possibly unrealistic variable values (such as relative humidity exceeding several 100 %)

... exclude files with very unrealistic variable values (such as relative humidity exceeding several 1000 %)

... provide models on a common grid

... provide annual mean files and annual mean files on a common grid

The CMIP6 next generation (CMIP6ng) processor aims to address these points in a consistent and traceable way. In the following an overview of the changes in the file structure is given (section 2.1), the traceability between the CMIP6 and CMIP6ng archives (section 2.2), the special quality flag variable which is added to all files (section 2.3), and the fixes and checks applied (section 2.4).

2 CMIP6ng processor

2.1 CMIP6ng file structure

A given file in the CMIP6ng archive is uniquely defined by a tuple of parameters: experiment, table, variable, model, variant, and grid (see table 1). In the original Earth System Grid Federation (ESGF) structure files are organized in a folder structure, where each of these parameters is one layer. Depending on the model center each path contains one or more files (sliced by time) for a given setting. An example file name pattern is show in (1):

variable_table_model_experiment_variant_grid_sdate-edate.nc (1)

e.g., tas_Amon_CESM2_historical_r1i1p1f1_gn_185001-201401.nc

Table 1: Acronyms, long names, and examples of parameters.

Acronym	**Long name**	**Example**
experiment	Experiment name	historical
table	Model table	Amon
variable	Variable name	tas
model	Model name	CESM2
variant	Ensemble member	r1i1p1f1
grid	Grid resolution	gn
sdate	Start date (yyyymm)	185001
edate	End date (yyyymm)	201401
tres	Time resolution	mon

In the ESGF setup the file path as well as each file name contains the full tuple of parameters that uniquely define a certain setting. In the next generation archive a flatter structure is used, with the file path only separating by variable, time resolution, and grid:

$$\begin{gathered}\ldots/\text{variable/tres/grid} \\ \text{e.g., } \ldots/\text{tas/mon/native}\end{gathered} \tag{2}$$

The time resolution can be one of daily (day), monthly (mon), or annual (ann). The grid in the next generation archive is either native or g025. Different gird identifiers are available for the native case, the first available one is selected following the hierarchy given in table 2. The g025 files have been regridded to a 2.5°×2.5° longitude-latitude grid using second order conservative remapping (cdo remapcon2), except for ocean-grid variables which are regridded using distance-weighted average remapping (cdo remapdis).
The CMIP6ng file name pattern is similar to the original ESGF pattern:

$$\begin{gathered}\text{variable\_tres\_model\_experiment\_variant\_grid.nc} \\ \text{e.g., tas\_mon\_CESM2\_historical\_r1i1p1f1\_native.nc}\end{gathered} \tag{3}$$

With the last part (_sdate-edate) deleted, since all files for a given setting are merged into one single file along the time axis. The time information is no longer necessary as it can be inferred from the experiment for all cases in which it is important. The grid identifier, again, indicates either native or g025.

Table 2: Grid identifiers for the CMIP6 (top) and CMIP6ng (bottom) archives. The full list of grid names can be found here: https://github.com/WCRP-CMIP/CMIP6_CVs/blob/master/CMIP6_grid_label.json

Acronym	**Description**	**Note**
gn	data reported on a model's native grid	1st priority
gr	regridded data reported on the data provider's preferred target grid	2nd priority
pr1	regridded data reported on a grid other than the native grid and other than the preferred target grid	3rd priority
gm	global mean data	4th priority
native	Native grid (one of the above)	
g025	2.5°×2.5° grid	cdo remapcon2

2.2 Processor information, consistency, and traceability

The core part of the CMIP6ng processor is written in Python (using version 3.6.7). Since the processor is necessarily constantly developing (e.g., additional fixes might be added or new variable checks might be introduced) it is important to be able to track each produced file back to the version of the code it was produced with (e.g., to discover potential bugs in the processing). Each time the processor is run repository information and revision hash of the current revision are retrieved and added to the file metadata (see (4)).
General information about the processor, contact information, and licenses as well as the applied fixes and open issues are also added to the file metadata as a new-line separated global attribute:

:cmip6-ng = (4)
contact = cmip6-archive@env.ethz.ch
description = ETH Zurich CMIP6 "next generation" (ng) archive.
disclaimer = This dataset is provided "as is", without warranty of any kind.
fixes = ...
git = yyyy-mm-dd HH:MM:SS repository_url branch revision_hash
ownership = The ownership of this dataset remains with the original provider
unfixed_issues = ...

As detailed in section 2.1 in several instances multiple input files are combined to one output file. To ensure tractability back to the original files, the full path of each file as well as the file hash (sha256) are saved to the metadata as comma separated lists:

$$:\text{original\_file\_names} = \text{path1, path2, } \ldots \tag{5}$$

and

$$:\text{original\_file\_hash\_codes} = \text{hash1, hash2, } \ldots \tag{6}$$

Regridding and time-averaging is done by a CDO (Climate Data Operators version 1.9.6: https://code.mpimet.mpg.de/cdo) sub-routine. CDO operations are added to the "history" attribute of the file, the CDO information can be found in the "CDO" attribute.

2.3 Quality flag

The CMIP6ng processor adds an additional "file_qf" variable to the file to track the quality of the file. The logic of this quality flag follows the NetCDF Climate and Forecast (CF) Metadata Conventions (http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.html#flags). It currently has four possible values:

- 0 File was not changed (beside regridding and time averaging) and has no issues
- 1 Fixes were applied to the file
- 10 The variable exceeds the warning range
- 100 The file has unfixed issues but was still included in the CMIP6ng archive

If multiple points are applicable for a given file the flag values are summed up (e.g., flag value 11 means that fixes were applied to the file and the variable exceeds its warning range). For all non-zero flag values an entry is added in either the fixes or the unfixed_issues attributes (see 4).

2.4 Data checking and bug fixing

Figure 1 shows the CMIP6ng processor workflow. After loading know issues are fixed in the file and tracked in the "fixes" attribute. Then general checks are performed (see table A1 for a list). General checks are mostly independent of the variable (only distinguishing, e.g., between dimensionality) and focus on the file metadata and dimensions. Finally, variable-dependent checks are carried out, testing if a given variable has the correct units and is within its respective warning and error ranges (see table A2). If a variable exceeds its warning range quality flag 10 is set and a corresponding entry in the "unfixed_issues" attribute is added. If a variable exceeds the error range it is considered physically implausible and the file is not saved in the next generation archive.

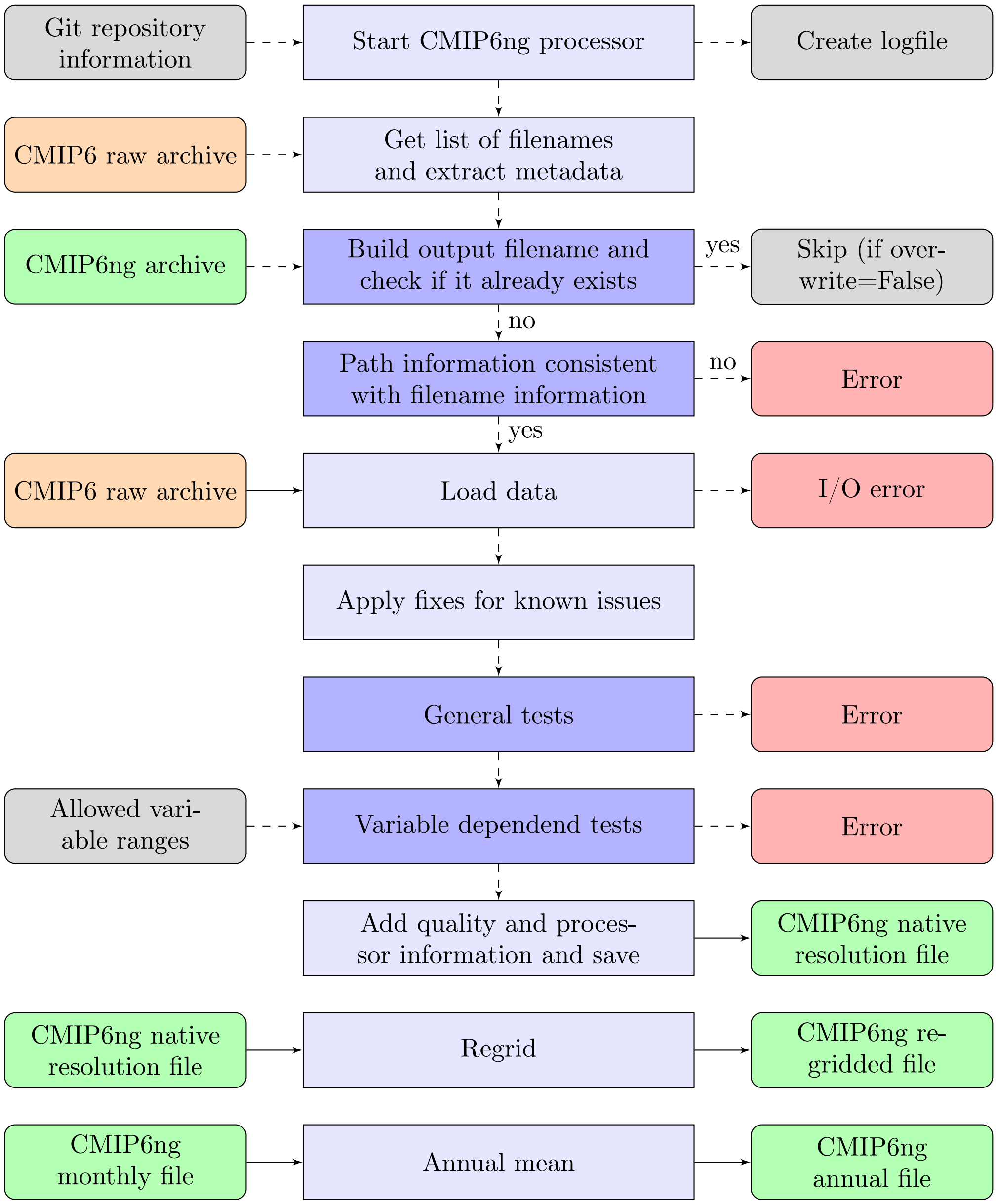

Figure 1: Schematic of the CMIP6ng processor.

2.4.1 Update January 2020: additional checks for relative variables

In using the CMIP6ng archive we have encountered several cases of potentially wrong values in variables which are given in %. Is seems that in these cases the unit stated in the file is correct according to the CMIP6 standards (namely %) but the corresponding values are in fractions (i.e., mainly between 0 and 1). This is not covered by the variable tests since it always falls within the checked range. For variables with unit % we have therefore introduced an additional test which demands that the maximum value (over all grid cells, time steps) is larger than 5. For values that are truly in % (hence covering values mainly between 0 and 100) this test should never fail but values that are mistakenly in fractions are very improbable to exceed this threshold. We currently do not fix nor include such cases into the CMIP6ng archive since we do not want to introduce new errors in case this behaviour has other reasons. However, all cases are reported to the respective model centres.

3 Summary

The CMIP6ng archive is a voluntary effort to create an as-consistend-as possible archive of CMIP6 data for easy use. Included files are checked to follow the same filename and variable name standards and cover the same time periods for experiments with fixed length. In addition, runs, which include potentially unrealistic values are flagged and runs which include extremely unrealistic values are skipped.

Acknowledgements. The authors thank everybody who has contributed to this effort, in particular Neil Allen, Lea Beusch, Doris Folini, Reto Knutti, Ryan Padrón Flasher, Angeline Pendergrass, Martin Stolpe, and Vera Weichlinger. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF.

Data availability. The CMIP6ng data are available upon request to cmip6-archive@env.ethz.ch. We reserve the right to process request only as we are able to or have time for. The original data were downloaded from `https://esgf-node.llnl.gov/projects/cmip6/`.

Code availability. The CMIP6ng processor is tailored to the ETH Zurich system and we therefore do not provide a public code repository. If you are interested in the code send an e-mail to cmip6-archive@env.ethz.ch.

References

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (May 2016). "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5, pp. 1937–1958. ISSN: 1991-9603. DOI: 10.5194/gmd-9-1937-2016. URL: https://www.geosci-model-dev.net/9/1937/2016/.

A Checks

Table A1: List of general checks applied to the files.

Test	**Applied to**
Path information consistent with file metadata	all
Time dimension exists and has correct name	all but fx
Number of years equals endyear - startyear + 1	all but fx
First time step is in January and last time step is in December	all but fx
(startyear, endyear) matches...	(1850, 2014) for historical (2015, 2100) for SSPs (except SSP534-over) (2040, 2100) for SSP534-over
Number of time steps is number of years times 12	monthly files
Each month exists exactly number of years times	monthly files
Number of days per year matches calendar	daily files
Days of year run from 1 to expected number of days based on calendar in each year	daily files
First times step is January 1st and last times step is December 31st (or 30st depending on calendar)	daily files
Latitude dimension exists and has correct name	all but gm
Latitude has unit 'degrees_north'	all but gm
Latitude is within (-90., 90.)	all but gm
Latitude is strictly increasing	all but gm
Longitude dimension exists and has correct name	all but gm
Longitude has unit 'degrees_east'	all but gm
Longitude is within (0., 360.)	all but gm
Longitude is strictly increasing	all but gm
Longitudinal gird is equidistant	all but gm

Table A2: Variables included in the CMIP6 next generation archive for at least one case. Shown are the variable acronym, full name, unit as well as the warning and error ranges which it is tested against in the processing.

Acronym	Long name	Unit	Warning range	Error range
areacella	Grid-Cell Area for Atmospheric Grid Variables	m2	0 to 100 000 000 000.0	0 to 100 000 000 000.0
clt	Total Cloud Cover Percentage	%	0.0 to 100.0	0.0 to 100.1
evspsbl	Evaporation	kg m-2 s-1	0.0 to 0.0005	−0.0005 to 0.005
evspsblveg	Evaporation from Canopy	kg m-2 s-1	−0.0005 to 0.005	−0.0005 to 0.005
evspsblsoi	Water Evaporation from Soil	kg m-2 s-1	−0.0005 to 0.005	−0.0005 to 0.005
gpp	Carbon Mass Flux out of Atmosphere due to Gross Primary Production on Land	kg m-2 s-1	−9999 to 9999	−9999 to 9999
npp	Carbon Mass Flux out of Atmosphere due to Gross Primary Production on Land	kg m-2 s-1	−9999 to 9999	−9999 to 9999
hfss	Surface Upward Sensible Heat Flux	W m-2	−999 to 999	−999 to 999
hfls	Surface Upward Latent Heat Flux	W m-2	−999 to 999	−999 to 999
hurs	Near-Surface Relative Humidity	%	0.0 to 250.0	0.0 to 2300.0
huss	Near-Surface Specific Humidity	kg/kg	0.0 to 1.0	−0.01 to 1.0
pr	Precipitation	kg m-2 s-1	0.0 to 0.01	−0.001 to 0.03
prw	Water Vapor Path (vertically integrated through the atmospheric column)	kg m-2	−999 to 999	−999 to 999
psl	Sea Level Pressure	Pa	80 000 to 120 000	80 000 to 120 000
ra	Carbon Mass Flux into Atmosphere Due to Autotrophic (Plant) Respiration on Land	kg m-2 s-1	−9999 to 9999	−9999 to 9999
rh	Carbon Mass Flux into Atmosphere Due to Heterotrophic Respiration on Land	kg m-2 s-1	−9999 to 9999	−9999 to 9999
rlds	Surface Downwelling Longwave Radiation	W m-2	−1 to 1100	−100 to 2000
rldscs	Surface Downwelling Clear-Sky Longwave Radiation	W m-2	−1 to 11 000	−100 to 2000
rlus	Surface Upwelling Longwave Radiation	W m-2	−1 to 1100	−100 to 2000
rlut	TOA Outgoing Longwave Radiation	W m-2	−1 to 1100	−100 to 2000
rlutcs	TOA Outgoing Clear-Sky Longwave Radiation	W m-2	−1 to 1100	−100 to 2000
rsds	Surface Downwelling Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rsdscs	Surface Downwelling Clear-Sky Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rsdt	TOA Incident Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rsus	Surface Upwelling Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rsuscs	Surface Upwelling Clear-Sky Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000

rsut	TOA Outgoing Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rsutcs	TOA Outgoing Clear-Sky Shortwave Radiation	W m-2	−10 to 1500	−100 to 3000
rtmt	Net Downward Flux at Top of Model	W m-2	−300 to 300	−500.0 to 500
sftlf	Percentage of the grid cell occupied by land (including lakes)	%	−0.01 to 100.01	−0.01 to 100.01
siconc	Sea-ice Area Percentage (Ocean Grid)	%	0 to 100.01	0 to 101
tran	Transpiration	kg m-2 s-1	−0.0005 to 0.005	−0.0005 to 0.005
tauu	Surface Downward Eastward Wind Stress	Pa	−1.5 to 1.5	−20.0 to 20.0
tauv	Surface Downward Northward Wind Stress	Pa	−1.5 to 1.5	−20.0 to 20.0
tas	Near-Surface Air Temperature	K	178.0 to 333.0	159.0 to 353.0
tasmax	Daily Maximum Near-Surface Air Temperature	K	178.0 to 333.0	164.0 to 386
tasmin	Daily Minimum Near-Surface Air Temperature	K	178.0 to 333.0	158.0 to 353.0
tos	Sea Surface Temperature	decC	−50 to 50	−50 to 50
zg500	Geopotential Height at 500hPa	m	4000.0 to 6500	4000.0 to 6500