CORDEX-CMIP6 Archiving specifications

First Order Draft

February 2023

This document provides Data Reference Syntax (DRS) elements necessary for post-processing CORDEX-CMIP6 simulations and publishing them on the Earth System Grid Federation (ESGF). The document includes file and directory naming conventions, global attributes and ESGF Search Facets Mappings.

Blue links mean final versions while red links mean documents that are still in development.

Red text and/or "???" mean that a specific issue is still under discussion.

1. DRS elements

The DRS element values must consist of the characters a-z, A-Z, 0-9 and '-' (dash). No other character is allowed. The terms in brackets following the DRS element names in the list below indicate whether the values are prescribed ('single value'), have to be taken from a controlled vocabulary, i.e. a fixed list of values ('CV'), have to be registered within CORDEX ('CV to register'), or can be chosen freely ('free string'). Note that most elements must have the same value as the mandatory NetCDF global attribute.

`variable_id` (CV) is the short name of the variable. The name is taken from the CORDEX-CMIP6 Variable List or CORDEX-CMIP6 CMOR tables.

`domain_id` (CV) is the name assigned to each of the CORDEX regions and includes a flag for resolution as listed in the CORDEX-CMIP6 domain id CV.

`driving_source_id` (CV) is an identifier of the driving data. The name consists of a model identifier. For reanalysis driven runs this is the name of the reanalysis data (ERA5). For runs driven by CMIP6 model data this is the associated CMIP6 `source_id`, which can be found in the CORDEX-CMIP6 driving source id CV.

`driving_experiment_id` (CV) is either "`evaluation`" for the ERA5-driven experiment or the value of the CMIP6 `experiment_id` from the ScenarioMIP activity or "`historical`" for the historical experiment from CMIP. The values for experiment_id can be found in the CORDEX-CMIP6 experiment id CV.

`driving_variant_label` (CV) identifies the ensemble member of the CMIP6 simulation that produced the forcing data. It has to have the same value as the CMIP6 `variant_label`. For the evaluation experiment driven by ERA5 it has to be "`r1i1p1f1`".

`institution_id` (CV to register) is an identifier for the institution that is responsible for generating and providing CORDEX simulations. All CORDEX institutions must be registered to publish their simulations on ESGF. See instructions on how to register an institution (...link...) and the actual state of the CORDEX-CMIP6 institution id CV.

`source_id` (CV to register) is an identifier (acronym) of the CORDEX RCM. All CORDEX RCMs have to be registered to publish their simulations on ESGF. See instructions on how to register a RCM (...link...) and the actual state of the CORDEX-CMIP6 source id CV. Different configurations of the same RCM such as different combinations of parameterization schemes or changes in parameters for existing schemes must be reflected in `source_id` by a free text suffix (e.g. `RCM123`, `RCM123A` for Africa or `RCM123T` for the tropics). RCM simulations with spectral nudging must use the "`SN`" suffix in `source_id` (e.g. `RCM123-SN`, `RCM123A-SN`, `RCM123T-SN`).

`version_realisation` (build rules) is a combination that identifies i) version of CORDEX datasets (simulations) related to technical, configuration or postprocessing errors and ii) realisations with different initial conditions for RCMs. This DRS element has the form "`vN-rM`". "N" in the version part "`vN`" is 1 for the first release of dataset (v1) and subsequent numbers (2, 3, 4, etc.) for any rerun or re-processing of the dataset (v2, v3, v4, etc.). The later version always supersedes the earlier version. "M'' in the realisation part "`rM`" is subsequent numbers (1, 2, 3 etc.) that reflect multiple RCM simulations with perturbed initial conditions (r1, r2, r3, etc.) driven by the same GCM and the same GCM member. The version and realisation parts are separated by a dash "-" (e.g. v1-r1, v1-r2, v1-r3). The version part of this DRS element should not be confused with the ESGF-related DRS element `version` that has the form "vYYYYMMDD '' and is only included in the ESGF directory structure.

`frequency` (CV) is the output frequency indicator: 1hr - 1 hourly, 3hr - 3 hourly, 6hr - 6 hourly, day=daily, mon=monthly, and fx=invariant fields.

`StartTime` and `EndTime` (build rules) indicate the time span of the file content. The format is YYYY[MM[DD[hh[mm]]]], i.e. the year is represented by 4 digits, while the month, day, hour, and minutes are represented by exactly 2 digits, if they are present at all (monthly output - YYYYMM, daily – YYYYMMDD, sub-daily - YYYYMMDDhhmm). The `StartTime` and `EndTime` of sub-daily instantaneous and average data are based on the time values of the first and last record in the file. The two dates are separated by a dash. All time stamps refer to UTC. Constant fields (Frequency=*fx*) do not have the `StartTime-EndTime` element in their file names.

`activity_id (CV)` – an identifier of different CORDEX activities (RCM, ESD, FPS, Adjust ???)

`mip_era (CV)` – determines what cycle of CMIP defines experiment and data specifications (always 'CMIP6')

`project_id` (CV) – project identifier (always 'CORDEX')

2. Global attributes

`Conventions` (CV) - Climate and Forecast (CF) convention version (always '1.10')

`activity_id` (CV) - an identifier of different CORDEX activities (RCM, ESD, FPS, Adjust ???) (part of DRS)

`comment` (free text, not mandatory)

`contact` (free text) - contact information of the institution that is responsible for CORDEX simulations (avoid personal contact information)

`creation_date` (build rules) - date when file was created in format YYYY-MM-DDTHH:MM:SSZ (e.g., "2023-01-15T14:30:23Z")

`domain` (CV) - name of the CORDEX region (CORDEX-CMIP6 domain id CV)

`domain_id (CV)` - an identifier assigned to each CORDEX region including a flag for resolution (CORDEX-CMIP6 domain id CV) (part of DRS)

`driving_experiment (CV)` – short description of the CMIP6 experiments (CORDEX-CMIP6 experiment id CV)

`driving_experiment_id (CV)` – root identifier of the CMIP6 experiments (CORDEX-CMIP6 experiment id CV) (part of DRS)

`driving_institution_id (CV)` – an identifier of the institution that is responsible for the driving CMIP6 simulation (CORDEX-CMIP6 institution id CV) (part of DRS)

`driving_source_id` (CV) - CMIP6 model identifier (CORDEX-CMIP6 driving source id CV) (part of DRS)

`driving_variant_label` - "variant" label of the driving CMIP6 simulation (e.g. "r1i1p1f1" etc.) (part of DRS)

`frequency (CV)` – sampling frequency (day, mon, 6hr, 3hr, 1hr) (part of DRS)

`further_info_url` (..) - location of documentation in CMIP6 (based on the ES-DOC service) ???

`history` (free text, not mandatory) - a timestamped trail for modifications to the original data, as suggested by the CF conventions.

`institution (CV` to register`)` - full name of institution that is responsible for CORDEX simulations (CORDEX-CMIP6 institution id CV) (part of DRS)

`institution_id (CV` to register`)` – an identifier of the institution that is responsible for CORDEX simulations (CORDEX-CMIP6 institution id CV) (part of DRS)

`mip_era` (CV) - determine what cycle of CMIP defines experiment and data specifications (always 'CMIP6') (part of DRS)

`native_resolution` (free text) - provides information about resolution of native model grids in km or deg or more detailed description of unstructured grids

`product` (CV) - product type ('model-output' and 'esd-output' ???) (part of DRS)

`project_id` (CV) - project identifier (always 'CORDEX') (part of DRS)

`realm` (CV) - realm(s) where variable is defined, not part of DRS, CORDEX-CMIP6 CMOR tables are defined per output frequency ???

`references` (not mandatory) - published or web-based references that describe the data, model or methods used.

`source` (some build rules) - full model name/version and components (aerosol, atmos, land etc.) (CORDEX-CMIP6 source id CV)

`source_id` (CV to register) - model identifier (CORDEX-CMIP6 source id CV) (part of DRS)

`source_type` (CV) - model configuration or ESD class (RCM, AGCM, RESM ???)

`version_realisation` (build rules) - identifies version of CORDEX datasets and RCM realisations (part of DRS)

`title` (free text, not mandatory)

`tracking_id` (build rules with some CV) - unique file identifier. It should be of the form "hdl:21.14100/<uuid>" (e.g., "hdl:21.14100/02d9e6d5-9467-382e-8f9b-9300a64ac3cd"). The tracking_id should be unique for each file published in ESGF. The <uuid> should be generated using the OSSP utility which supports a number of different DCE 1.1 variant UUID options. For CORDEX-CMIP6, version 4 (random number based) is required. Download the software from OSSP uuid. (see note 15 in CMIP6 DRS)

`variable_id` (CV) - variable identifier (CORDEX-CMIP6 CMOR tables) (part of DRS)

3. File naming

file_name=<variable_id>_<domain_id>_<driving_source_id>_<driving_experiment_id>_<driving_variant_label>_<institution_id>_<source_id>_<source_configuration_id>_<frequency>[_<StartTime>-<EndTime>].nc

Examples:

tas_AFR-22_ERA5_evaluation_r1i1p1f1_INST_RCM123_v1-r1_mon_201101-202012.nc

tas_AFR-22_GCM_historical_r1i1p1f1_INST_RCM123_v1-r1_mon_201101-201412.nc

tas_AFR-22_GCM_ssp370_r1i1p1f1_INST_RCM123_v1-r1_mon_201501-202012.nc

In contrast to CORDEX-CMIP5:

i) the institution that is responsible for CORDEX simulations (`institution_id`) and model acronym (`source_id`) are 2 different DRS elements, i.e. separated by the underscore "_" in the file name

ii) the institution that is responsible for the driving CMIP6 simulation (`driving_institution_id`) is not a part of DRS and not included in the file name and ESGF directory structure.

4. ESGF Directory structure

directory_structure=<project_id>/<mip_era>/<activity_id>/<domain_id>/<institution_id>/<driving_source_id>/<driving_experiment_id>/<driving_variant_label>/<source_id>/<source_configuration_id>/<frequency>/<variable_id>/<version>/

Examples:

/cordex/cmip6/rcm/AFR-25/ …

/cordex/cmip6/esd/AFR-25/ …

/cordex/cmip6/fps/"fps_name"/ …

/cordex/cmip6/adjust/AFR-25/…

How to synchronise with the existing CORDEX-CMIP5 simulations that have different DRS and have been published in ESGF under the same CORDEX project ? The CORDEX-CMIP5 ESGF directory structure begins with "<activity>/<product>/<Domain>/" i.e. /cordex/output/AFR-22/ … ???

5. File format

Data files must be in NetCDF format, version 4, using the NetCDF 4 classic data model. It is recommended that data should be compressed by using "deflate level" 1 and with "shuffle" turned on. Data files must conform to the CF Conventions v1.10.

Each file may contain only one output field (target variable) from a single simulation. It must include attributes and coordinate variables. The entire time series of a target variable is to be distributed over several files as described in section 8 Time period for each data file.

All output fields must be single precision (type NC_FLOAT), while all coordinate variables (time and space) must be double precision (type NC_DOUBLE). All missing data must be assigned the single precision floating point value of 1.e20.

6. CORDEX domains and horizontal coordinates

The CORDEX domains are defined in (domain doc). A domain must lie fully inside the RCM interior computational domain, i.e. in the area left once the relaxation zone is excluded. It is strongly recommended that RCMs using the rotated-pole coordinate system exactly follow the CORDEX grid definition provided. All variables from one simulation have to be provided on the same grid

The domain acronym is 'domain name'-'resolution', where 'resolution' is the nearest grid spacing in km of the 3 resolutions used in CORDEX-CMIP5 and CORDEX-CMIP6 (50, 25 and 12 km). For example, "AFR-25" means the CORDEX-Africa domain with 25 km resolution in a projected coordinate system and 0.22° resolution in the rotated pole coordinate system. The domain acronyms for the regular grids are the same as those for the corresponding model grid with the letter 'i' appended to the resolution (e.g. "AFR-25i").

Data must be provided for the CORDEX domain only, i.e. the relaxation zones must be removed before the data is delivered. Names of the CORDEX domains are provided in CORDEX-CMIP6 domain id CV .

Data files must contain full description of native coordinate systems used by RCMs:

- the 1-dimensional coordinate variables (e.g. `rlon` and `rlat` for the rotated pole coordinate system or `x` and `y` for the Lambert Conformal Conic (LCC) projection),
- coordinate variable `crs` describing the coordinate reference system and
- the variable attribute - `grid_mapping = "crs"`

in accordance with CF-1.10 (see examples).

The 2-dimensional geographic latitudes and longitudes of the model grid cells (`lon` and `lat`) must be also provided as auxiliary coordinates. Longitudes must have absolute values as small as possible, be monotonic and be confined to the range -180 to 360.

For models with native unstructured grids, it is up to the regional CORDEX communities to decide on whether data must be remapped to one of the regular grids or to the most common native RCM grid used for a specific CORDEX domain.

7. Time coordinate

The unit of the time coordinate is '`days since 1950-01-01T00:00:00Z`' or '`days since 1950-01-01`' for all files. All time dependent variables must have an attribute '`cell_methods: time`' with values provided in the CORDEX-CMIP6 CMOR tables.

The time value of the instantaneous data is [0Z, 6Z, 12Z, 18Z], [0Z, 3Z, 6Z, 9Z, 12Z, 15Z, 18Z, 21Z] and [0Z, 1Z, 2Z, 3Z, … , 20Z, 21Z, 22Z, 23Z] of each day for the 6-, 3- and 1-hourly data respectively.

Variables that are representative for an interval (averages, maxima, minima) must use the midpoint of time intervals as time coordinate values Therefore, these variables have the time values 3Z, 9Z, 15Z, 21Z (6-hourly), 1.5Z, 4.5Z, 7.5Z, 10.5Z, 13.5Z, 16.5Z, 19.5Z, 22.5Z (3-hourly) and 0.5Z, 1.5Z, 2.5Z, … , 21.5Z, 22.5Z, 23.5Z (1-hourly).

Furthermore, interval variables must have a `time_bnds` field of dimensions (`ntimes,2`), where `ntimes` is the dimension of the time coordinate. Intervals for daily and monthly should start and end at 00:00:00 UTC of the appropriate day. Intervals for sub-daily data should start and end at 00:00:00 UTC or an integer multiple of the frequency (1, 3, or 6 hours) from that point.

The time variable must have a `calendar` attribute. Use of the `[proleptic-]gregorian` or `standard` calendar is strongly recommended when possible. Other calendars (`360_day` and `365_day`) inherited from the driving models are also allowed.

8. Time period for each data file

The time spans that have to be included into a single file depend on the aggregation, which is 1-hourly, 6-hourly, daily, monthly, or invariant:

- 1-hourly or 6-hourly: one year,
- daily: 5 years or less,
- monthly: 10 years or less,
- invariant: single file.

Files should always contain full years if the data are available.

Files with monthly data start with years that end with '1' or the first year of the experiment; they end with '0' or the last year of the experiment.

Daily data files start with years that end with '1' or '6' or the first experiment year; the last year they contain ends with '5' or '0' or is the last experiment year. For example, the ERA5-driven evaluation experiment for 1979-2021 with 1979 as a spin-up:

monthly	daily	subdaily
1980-1980	1980-1980	1980-1980
1981-1990	1981-1985	1981-1981
1991-2000	1986-1990	1982-1982
2001-2010	……..	……..
2011-2020	2016-2020	2020-2020
2021-2021	2021-2021	2021-2021

9. License

All CORDEX modeling groups choose a license for their CMIP6-driven simulations depending on institutional and/or funding agency policies. This information is necessary to register a RCM in the CORDEX RCM CV (more details about registration will be provided later). It is strongly recommended to use the Creative Commons Attribution 4.0 International (CC BY 4.0) license, as currently in CMIP6. Note, that any kind of "non-commercial" license will significantly limit the use of the data in downstream climate mitigation and adaptation applications.

Similar to CMIP6 (see note 12 in CMIP6 Global Attributes, DRS, …), information about the license must be provided in the global attribute `license` with the following statement (with segments in square brackets optional, and with required, appropriate text entered in place of <*>):

`license` = "CORDEX model data produced by <institution_id> is licensed under a Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses). Consult (..link..) for terms of use governing CORDEX output, including citation requirements and proper acknowledgment. The data producers and data providers make no warranty, either express or implied, including, but not limited to, warranties of merchantability and fitness for a particular purpose. All liabilities arising from the supply of the information (including any liability arising in negligence) are excluded to the fullest extent permitted by law.";

Or the global attribute `license` can simply lead to a table with information about the license for all RCMs, easier to manage (e.g. in CMIP6_source_id_licenses) ???

10. Registration

All institutions (modelling groups) that contribute or plan to contribute to CORDEX-CMIP6 must i) register their institution and model and ii) provide information about their planned simulations in CORDEX-CMIP6 downscaling plan. The modelling groups will not be able to publish their CORDEX-CMIP6 simulations on ESGF without first registering their institution and model. More details about the registration will be provided later.

11. Examples

11.1. Rotated Pole Coordinate System

```
char crs ;
    crs:grid_mapping_name = "rotated_latitude_longitude" ;
    crs:grid_north_pole_latitude = 39.25 ;
    crs:grid_north_pole_longitude = -162. ;
double rlon(rlon) ;
    rlon:standard_name = "grid_longitude" ;
    rlon:long_name = "longitude in rotated pole grid" ;
    rlon:units = "degrees" ;
double rlat(rlat) ;
    rlat:standard_name = "grid_latitude" ;
    rlat:long_name = "latitude in rotated pole grid" ;
    rlat:units = "degrees" ;
double lon(rlat, rlon) ;
    lon:standard_name = "longitude" ;
    lon:long_name = "longitude" ;
    lon:units = "degrees_east" ;
double lat(rlat, rlon) ;
    lat:standard_name = "latitude" ;
    lat:long_name = "latitude" ;
    lat:units = "degrees_north" ;
float pr(time, rlat, rlon) ;
    pr:standard_name = "precipitation_flux" ;
    pr:long_name = "Precipitation" ;
    pr:units = "kg m-2 s-1" ;
    pr:coordinates = "lon lat" ;
    pr:_FillValue = 1.e+20f ;
    pr:missing_value = 1.e+20f ;
    pr:cell_methods = "time: mean" ;
    pr:grid_mapping = "crs" ;
```

11.2. Lambert Conformal Conic projection

```
char crs ;
    crs:grid_mapping_name = "lambert_conformal_conic" ;
    crs:standard_parallel = 49.5 ;
    crs:longitude_of_central_meridian = 10.5 ;
    crs:latitude_of_projection_origin = 49.5 ;
    crs:false_easting = 2925000. ;
    crs:false_northing = 2925000. ;
    crs:earth_radius = 6371229. ;
double x(x) ;
    x:standard_name = "projection_x_coordinate" ;
    x:long_name = "X Coordinate Of Projection" ;
    x:units = "m" ;
double y(y) ;
    y:standard_name = "projection_y_coordinate" ;
    y:long_name = "Y Coordinate Of Projection" ;
    y:units = "m" ;
double lon(y, x) ;
    lon:standard_name = "longitude" ;
    lon:long_name = "longitude" ;
    lon:units = "degrees_east" ;
double lat(y, x) ;
    lat:standard_name = "latitude" ;
    lat:long_name = "latitude" ;
    lat:units = "degrees_north" ;
float pr(time, y, x) ;
    pr:standard_name = "precipitation_flux" ;
    pr:long_name = "Precipitation" ;
    pr:units = "kg m-2 s-1" ;
    pr:coordinates = "lon lat" ;
    pr:_FillValue = 1.e+20f ;
    pr:missing_value = 1.e+20f ;
    pr:cell_methods = "time: mean" ;
    pr:grid_mapping = "crs" ;
```