Format of the data portal and file naming conventions for the C20C+ Detection and Attribution Project

[1]Dáithí Stone (dstone@lbl.gov)

This document provides details on the categories and labels for use in the data portal of the International CLIVAR C20C+ Detection and Attribution (D&A) project. This is compliant with the CF (NetCDF Climate and Forecast Metadata Convention) standard.

Location

The project output data can be accessed at `http://portal.nersc.gov/c20c/data.html`. There are three methods:

- All data can be accessed through the dedicated project portal interface at `http://portal.nersc.gov/cascade/data/downloader.php`.
- A subset of the data is stored on disk and can be accessed directly at `http://portal.nersc.gov/c20c/data`.
- Most of the data is stored on tape and can be accessed directly at `http://portal.nersc.gov/archive/home/projects/cascade/www/C20C`. Any data not on disk can be found here. (We are gradually moving toward a status where all data is available on tape, irrespective of whether it is also on disk or not.)

Classification

Data published on the data portal are classified according to the categories outlined in Table 1. The category values for some basic climate output from the core experiment should be fairly clear, following from the CMIP5 conventions. However, it is planned that the C20C+ D&A Project will go beyond the pure dynamical climate model output included in CMIP5 and include derivative output, such as extreme value indices or river basin model output. We will need to introduce new labels for such output; please consult `dstone@lbl.gov` for devising these labels.

File names

It would be useful if filenames took the following form, essentially following the format used by CMIP5 but with the <experiment_family> field added:
`<variable>_<realm_frequency>_<model_id>_<experiment_family>_<experiment>_<subexperiment>_run<run_id>_<start>-<end>.nc`
<start> and <end> should be numerical and of the form YYYYMM for monthly data, for instance “196001” for January 1960, YYYYMMDD for daily data, and YYYYMMDDHHMM for 3-hourly data.

Category	**Description**	**Example**
Institute	The institution performing the experiments, preferably a short form	*LBNL* (for Lawrence Berkeley National Laboratory)
Model	The climate model used for the simulations, including an indication of version if different versions of a model are run or if a model is run at different resolutions	*CAM5-1-2degree* (for model CAM5.1-2degree, i.e. CAM5.1 run at $\sim$2-degree horizontal resolution)
Experiment family	The label for the scenario being used	For the core experiment this will either be *All-Hist* for the "real-world" simulations or "Nat-Hist" for the simulations of the world that might have been had humans never interfered with the climate. See Table 2.
Experiment	The label for the estimate of the scenario being used, or equivalently the sub-scenario. For instance, for the "Nat-Hist" scenario one estimate will use one particular plausible attributable ocean warming estimate, while another estimate will use a different plausible warming estimate.	See Table 3.
Subexperiment	The label for the version of this model and scenario estimate. Note that this label is unique to each model/experiment_family/experiment combination, i.e. "v1-0" for one model may have no relation to "v1-0" for another model. This category can be used to distinguish between ensembles of simulations driven with different original forcing datasets, for instance different estimates of the sea surface temperatures.	*v1-0* (but no standard format specified)
Time Frequency	The time period represented by each time step in the data.	*fix* (for time-invariant data), *mon* (for monthly data), *day* (for daily data), or *3hr* (for 3-hourly data)
Realm	The part of the climate system or other system characterised by this variable	*atmos* (for atmosphere) or *land* (for land). Additional values may be generated to classify derivatives calculated from model output in subsequent analyses, for instance extreme value indices or agricultural indices.
Variable	The CF label of the variable	*tas* (for near-surface air temperature). See document `http://portal.nersc.gov/c20c/input_data/C20C_attribution_output_metadata.pdf`.
Ensemble	An identifier for a simulation within the Model/Experiment_family/Experiment/Subexperiment's ensemble of simulations.	*run007*

Table 1: The classification categories for the C20C+ D&A Project, in order of directory hierarchy. Consult dstone@lbl.gov if values for certain categories are not obvious.

Label	**Description**
All-Hist	Historical including changes in "all" known external forcings (anthropogenic and natural)
Nat-Hist	Historical including changes in natural external forcings only

Table 2: Labels of options for the **Experiment_family** classifier for the core experiment of the C20C+ D&A Project.

Experiment_family label	**Experiment label**	**Description**
All-Hist	est1	Driven by the standard set of the boundary conditions, i.e. sea surface temperatures, CO_2 concentrations, etc.
Nat-Hist	CMIP5-est1	Altering SSTs and SICs according to the baseline comparative estimate of the attributable warming derived from the CMIP5 ensemble, without application of optimal fingerprinting
	CanESM2-p25-est1	Using a certain set of original external forcings (e.g. HadISST1 SST and SIC) and altering SSTs and SICs according to a certain estimate of the attributable warming due to anthropogenic activites (in this case for the 25th percentile of the estimate from optimal fingerprinting using CMIP5 simulations of the CanEMS2 AOGCM)

Table 3: Labels of possible entries for the **Experiment** classifier for the core experiment of the C20C+ D&A Project. The values for these properties will be provided with the SST and SIC dataset for running the simulations.

For example, for near-surface air temperature (*tas*), an atmospheric variable (*A*), averaged monthly (*mon*), from model CAM5.1-2degree (*CAM5-1-2degree*), in simulation #15 (*run015*) run under version 1-0 (*v1-0*) of scenario All-Hist/est1 (*All-Hist*, *est1*) with data from January 1960 (*196001*) through December 2011 (*201112*), the filename would be:
`tas_Amon_CAM5-1-2degree_All-Hist_est1_v1-0_run015_196001-201112.nc`
Note that CAM5.1-degree has been converted to "CAM5-1-2degree" for the filename: it is convenient to leave dots out of the filename except in the ".nc" at the end.

Directory structure

Data portal directories for the C20C+ D&A project can be placed:

- On the central server at NERSC.
- On a separate server where all files have a URL. The address of this server will need to be added to the main portal interface.

Within the portal directory, place data under the following subdirectory structure, exactly following the values assigned for the categories listed in Table 1:
`./<Institute>/<Model>/<Experiment family>/<Experiment>/<Subexperiment>/<Time frequency>/<Realm>/<Variable>/<Ensemble>/`

For instance, for the example given above for file names, the subdirectory under the parent directory would be:
`./LBNL/CAM5-1-2degree/All-Hist/est1/v1-0/mon/atmos/tas/`