CMIP6 DATA SETS

Pangeo's Google Cloud collection so far consists of some basic datasets, but is growing rapidly.

The full CMIP6 collection is HUGE. There are so many datasets that you will need to learn special techniques to select and use the appropriate data to address your science issue. So here is a quick description.

As for the CMIP5 data, there is a 'controlled vocabulary' of keywords which allow us to organize the millions of data files being produced under CMIP6. Both CMIP5 and CMIP6 use keywords which result in the following file system structure (on NCAR/GLADE). In this Google Cloud Store (GCS) collections, the zarr stores stop at <grid_label>.

CMIP6:

```
<mip_era>/
      <activity_id>/
          <institution_id>/
              <source_id>/
                  <experiment_id>/
                      <member_id>/
                          <table_id>/
                              <variable_id>/
                                  <grid_label>/
                                      <version>/
                                        <CMOR filename>.nc
```

CMIP5:

```
<activity>/
  <product>/
      <institute>/
          <model>/
              <experiment>/
                  <frequency>/
                      <modeling realm>/
                          <MIP table>/
                              <ensemble member>/
                                  <version number>/
                                      <variable name>/
                                          <CMOR filename>.nc
```

Note that the keyword 'model' that we used for CMIP5 corresponds to 'source_id' in the CMIP6 world. The CMIP5 'frequency', 'modeling realm' and 'MIP table' are now combined in CMIP6 and embedded in a (very terse) 'table_id' keyword. A new CMIP6 keyword, 'grid_label', was needed to distinquish between data on a model (natural) grid vs. data regridded to a standard grid.

A data request can be made with 3 keywords: 'experiment_id', 'table_id', and 'variable_id'.

You can visit https://esgf-node.llnl.gov/search/cmip6/ to explore the data and make a note of the experiment_id, table_id, variable_id (listed in left column as 'Experiment ID', 'Table_ID' and 'Variable' respectively). Note the following:

- The popular scenarios in CMIP5 are almost equivalent to similar named CMIP6 ones, but the experimental setup could be slightly different, for example:

CMIP5 'experiment'	CMIP6 'experiment_id'
piControl	piControl
1pctCO2	1pctCO2
amip	amip
historical	historical
rcp45	ssp245
rcp85	ssp585
past1000	past1000

- The CMIP5 'model' keyword matches the CMIP6 'source_id'. Of course there are many more models available in CMIP6.

- The 'variable name' and 'variable_id' are also similar, but again, CMIP6 has a lot more, see below

- The 'table_id' is the hardest to explain, so here I give a brief description of each:
  3hr : atmosphere sampled every 3 hours
  6hrLev : 6-hourly data on atmospheric model levels
  6hrPlev : 6-hourly atmospheric data on pressure levels (time mean)
  6hrPlevPt : 6-hourly atmospheric data on pressure levels (instantaneous)
  AERday : Daily atmospheric chemistry and aerosol data
  AERfx : Fixed atmospheric chemistry and aerosol data
  AERhr : Hourly atmospheric chemistry and aerosol data
  AERmon : Monthly atmospheric chemistry and aerosol data
  AERmonZ : Monthly atmospheric chemistry and aerosol data
  Amon : Monthly atmospheric data
  CF3hr : 3-hourly associated with cloud forcing
  CFday : Daily data associated with cloud forcing
  CFmon : Monthly data associated with cloud forcing
  CFsubhr : Diagnostics for cloud forcing analysis at specific sites
  E1hr : Hourly Atmospheric Data (extension)
  E1hrClimMon : Diurnal Cycle
  E3hr : 3-hourly (time mean, extension)
  E3hrPt : 3-hourly (instantaneous, extension)
  E6hrZ : 6-hourly Zonal Mean (extension)
  Eday : Daily (time mean, extension)
  EdayZ : Daily Zonal Mean (extension)
  Efx : Fixed (extension)
  Emon : Monthly (time mean, extension)
  EmonZ : Monthly zonal means (time mean, extension)
  Esubhr : Sub-hourly (extension)
  Eyr : Daily (time mean, extension)
  IfxAnt : Fixed fields on the Antarctic ice sheet
  IfxGre : Fixed fields on the Greenland ice sheet
  ImonAnt : Monthly fields on the Antarctic ice sheet
  ImonGre : Monthly fields on the Greenland ice sheet
  IyrAnt : Annual fields on the Antarctic ice sheet
  IyrGre : Annual fields on the Greenland ice sheet
  LImon : Monthly fields for the terrestrial cryosphere
  Lmon : Monthly land surface and soil model fields
  Oclim : Monthly climatologies of ocean data
  Oday : Daily ocean data
  Odec : Decadal ocean data
  Ofx : Fixed ocean data
  Omon : Monthly ocean data
  Oyr : Annual ocean variables
  SIday : Daily sea-ice data
  SImon : Monthly sea-ice data
  day : Daily Data (extension - contains both atmospheric and oceanographic data)
  fx : Fixed variables

- **member_id: a key constructed from 4 indices stored as global attributes:**

member_id = r<k>i<l>p<m>f<n>

where

k = realization_index
l = initialization_index
m = physics_index

n = forcing_index

- **grid_label: a key indicating if on native grid, regridded, etc**

Modeling groups may choose to report their output on the model’s native grid and/or regrid it to one or more target grids. To distinguish between output reported on different grids, a “grid_label” attribute is defined.

The rules for assigning grid labels should make it easy for users to select (using the ESGF search tools) CMIP output that is on a grid considered by each modeling group to best represent its model -- the so-called “primary” grid. If output is reported on the native grid, this is always deemed the “primary” grid. If output is not reported on the native grid, then modeling groups should regrid the data to some primary grid of its choosing For the “primary” grid the following labels apply:

grid_label = "gn" (output is reported on the native grid)
grid_label = "gr" (output is not reported on the native grid, but instead is regridded by the modeling group to a “primary grid” of its choosing)
grid_label = “gm” (global mean output is reported, so data are not gridded)

As noted below sometimes a “z” or “a” or “g” is appended to the labels to indicate “zonal means” or grids limited to Antarctica or Greenland.
If besides the “primary” grid, output is regridded to an additional grid, then for this output:
grid_label = "gr[i]" (a “secondary” grid), where <i> should be replaced by a positive integer less than 10, which distinguishes this output from other regridded output.

CMIP6 experiments ("experiment_id'):

- List of Tier 1 Experiments
- List of Tier 2 Experiments
- List of Tier 3 Experiments
- List of Tier 4 Experiments

CMIP6 models ("source_id"):

- List of all Models

CMIP6 variables ("variable_id'):

- List of all Variables