Obs4MIPs Dataset Requirements
Checklist for output requirements of observational data products prepared for obs4MIPS

The intent of this document is to provide a checklist for teams preparing observational datasets for the obs4MIPS project. It only lists requirements that apply to observational datasets. These have been extracted from the requirements specified in the CMIP5 output requirement documents at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html, and include additional requirements specifically for obs4MIPs. Refer to the documents on the website for more detailed information. In particular, this site has links and explanations of “standard_output.xls” and “CMOR2” mentioned below.

The dataset files can be prepared using the CMOR2 software (see the link above). It is expected that by using that software, the output file will be automatically compliant with PCMDI requirements for model output. However, CMOR requires input for some global attributes that are not relevant to obs4MIPs, some adjustments are required to CMOR tables in order for it to run on an observations dataset, and the output will need to be post-processed to remove these irrelevant attributes. Files can also be prepared by strictly following this guide. This guide is also useful for collecting all of the information that will be required for input to CMOR2.

This document and other documents referenced here can be found through links maintained at http://obs4mips.llnl.gov:8080/wiki/requirements . It is also recommended that a potential dataset provider download a well formed example of an obs4MIPs dataset from the ESGF such as the AIRS ta dataset, and examine its content and structure, before proceeding to construct a new product.

1. General Requirements

- The output file must be written in NetCDF version 3.
- The data output must follow the standard NetCDF Climate and Forecast (CF) Metadata convention (http://cf-pcmdi.llnl.gov/).
- The output file must pass a CF compliance check. You can find a checker at http://puma.nerc.ac.uk/cgi-bin/cf-checker.pl. Choose the latest CF version when submitting the file for checking.
- Each output file must contain a time series of ONLY ONE physical variable (e.g sea surface temperature, specific humidity).
- If the entire time series can be stored in less than 2GB, it must be stored in a SINGLE file. If it requires more than 2GB, it should be split into the minimum number of files required, with the size of each file being less than 2GB. Each file should contain a contiguous time series of complete data grid blocks. Each file must contain all of the required metadata applicable to the data subset contained in the file.
- Each physical variable and coordinate variable must use the specified output/coordinate variable name given in the CMIP5 Requested Output list (standard_output.xls). For example, the latitude output name must be “lat”, and the air temperature output variable name must be “ta”.

2. Coordinate Variables

- The output file must contain temporal and spatial dimension parameters: (e.g time, latitude, longitude, and pressure-levels, if applicable). See worksheet “dims” in the CMIP5 output requirement document “standard_output.xls” for details.
- All coordinate variables (e.g. time, latitude, longitude, pressure levels) must be written as double precision floating point numbers.
- The time variable must be stored in the increasing direction and in units of days since a fixed date (e.g. January 1st, 1990), if it is a daily or monthly averaged product. The fixed date can be any date. However, it makes most sense for the fixed date to be the date of the first spatial date block in the file.
- The boundaries of the time variable values must be included in the file as a two-dimensional array of values that define the time period extent for each time value. I.e., for monthly averaged data, the time value would be a date in that month, and the boundaries would be the dates of the 1st and last days of the month.
- For time-averaged data (e.g., monthly averages of observations), a time variable value must be defined for each data block as the mid-point of the interval over which the average is computed.
- The latitude variable must be stored in the direction of increasing value, and in the range [-90 to +90] in units of north-bound degrees.
- The boundaries of the latitude variable values must be given as a two-dimensional array that defines the spatial extent of each latitude value. I.e., for 1 degree gridded data, the boundaries are the latitudes that bound each grid box.
- The longitude variable must be stored in the direction of increasing value, and in the range [0 to 360] in units of east-bound degrees.
- The boundaries of the longitude variable values must be given as a two-dimensional array that defines the spatial extent of each longitude value. I.e., for 1 degree gridded data, the boundaries are the longitudes that bound each grid box.
- While there is no specific resolution for latitude and longitude, it is recommended to use 1 degree by 1 degree resolution if the observations sampling rate is sufficient to produce meaningful average values. Otherwise, lower resolutions are acceptable.
- The pressure level variable (if applicable) must consist of all the mandatory levels, and may include optional levels. For example, in Amon-type products, the mandatory levels are 100000, 92500, 85000, 70000, 60000, 50000, 40000, 30000, 25000, 20000, 15000, 10000, 7000, 5000, 3000, 2000, and 1000, in units of Pa. See the worksheet “dims” in the CMIP5 output requirement document “standard_output.xls” for details.
- The pressure level coordinates must be stored in the order of decreasing pressure and in units of Pa.
- The boundaries of each pressure level value are required for cfMon and cfDay products but not required for other products. Boundaries are stored as a two dimensional array that defines the pressure extent of each pressure level.
- Metadata required for the dimension parameters (e.g. time, latitude, longitude) are:
    - axis = "X", "Y", "Z", or "T"
    - bounds (when appropriate)
    - calendar (for time coordinates only)
    - units

- long name
- standard name
- formular_timers (for dimensionless vertical coordinates only)
- positive (for vertical coordinates only).

3. Physical Variables

- The physical variable must be stored in a regularly gridded array with specific dimensions and coordinates in the order given in the CMIP5 Requested Output (standard_output.xls : "dims" tab, column called "CMOR dimension"). For example, the air temperature should be stored in a 4 dimensional array with indexes [time, pressure-level/height, latitude, longitude] in row major order (i.e., longitude values are adjacent). This is the standard storage method in C, and opposite of FORTRAN.
- The value for missing data **must be set** to 1.0E20.
- All physical variables (e.g. air temperature, surface upward latent heat flux) must be written consistent with the data type specified in the "type" column (P) of the CMIP5 requested output tables in document "standard_output.xls".
- Metadata required for the physical variable are:
  - Units
  - long name
  - standard name
  - cell methods
  - cell measures
  - missing value
  - fill value
  - associated files
  - coordinates (when appropriate)
  - flag_values (when appropriate)
  - flag_meanings (when appropriate)
  - grid_mapping (when appropriate).
- See the CMIP5 requirement document "CMIP5_output_metadata_requirements.pdf" for the definitions of the metadata. "when appropriate" metadata only applies to certain specific physical variables.

4. Global Attributes

There are required attributes, optional standardized attributes, and the user may define any additional attributes thought to be useful. Additionally, some global attributes that are written by CMOR do not pertain to observations, so they must be removed through a post-processing phase. Because of limitations of the current ESG publishing software, each global attribute should have only one value (i.e. multiple values are not allowed). The global attributes requirements are evolving, and are maintained in a separate document:

obs4MIPs Global Attributes Requirements.pdf

5. Ancillary Data (for products that are averages of individual observations)

- If the uncertainty estimate for the dataset has a significant per datum variation, provide the number of observations per grid cell and the "standard error of the mean" (see http://en.wikipedia.org/wiki/Standard_error_(statistics)) of the data used to calculate the physical variable mean value at each grid cell. For retrieval quantities that are binary or fractional (e.g. total cloud fraction), the standard deviation may be provided instead of the standard error of the mean. If the number of observations and/or the standard error have simple latitude dependence and no time dependence, one may report them in a table in the accompanying technical note instead of computing them for each data value.

- The CF names for these quantities are {var}Nobs, {var}Stderr, and {var}Stddev, where {var} is replaced with the physical variable output name given in "standard_output.xls". Each should be stored in a separate array with the same ordering as the physical variable array. The data type for Nobs should be integer, and the data type for Stderr/Stddev should be single precision floating point. Metadata required for these variables are:
  - Units
  - long name ({variable long name} {number of observations | standard error | standard deviation})
  - standard name (*constructed as above*)
  - missing value (*should be 1.0E20 for Stderr/Stddev and 0 for Nobs*)
  - fill value (*should be 1.0E20 for Stderr/Stddev and 0 for Nobs*)
- This ancillary data can be included within the dataset file, or provided as separate files. If these are provided as separate files, they must include all of the metadata and time, latitude, longitude, and pressure level data as the physical variable file.
- If inclusion of this ancillary data causes the file size to exceed the 2 GB limit, break the file into separate files in the manner described above under **General Requirements** while keeping the ancillary data subset with their corresponding physical variable subset in each file.

6. Naming Convention for File Names

Files containing observations to be included in the CMIP5 archive must be constructed according to a naming convention that is designed to quickly represent the content of the file. The naming convention is slightly different for different kinds of observations, to reflect information that is specific and important to each kind, specifically:

- Satellite datasets: filename = <*variable*>_<*instrument*>_<*processing_level*>_<*processing version*>_<*start_date*>-<*end_date*>.nc
- In-situ datasets: filename = <*variable*>_<*project*>_<*location*>_<*instrument*>_<*processing_level_and product_version*>_<*start_date*>-<*end_date*>.nc

Where

<*variable*> is the physical variable output name listed in standard_output.xls

<*instrument*> is a short name which is the generally recognized name for the instrument used to collect the data (e.g., AIRS, MLS, MODIS, etc.)

<*processing_level*> is the standard satellite product level designator (L1, L2, L3, etc)

<*processing version*> is a short string with no spaces that uniquely identifies the version of this dataset. It may also indicate the processing version of the parent dataset from which this product is produced. In either case, this version indicator MUST be updated if a revised dataset is published

<*start_date*> is a numeric string of the form yyyymm(dd) that corresponds to the time value of the first data block in the file. (dd) is not used for monthly data products.

<*end_date*> is a numeric string of the form yyyymm(dd) that corresponds to the time value of the last data block in the file. (dd) is not used for monthly data products.

<*project*> identifies the In-situ data collection (e.g, ARM)

<*location*> is a short string identifying the location at which the data was collected

<*processing_level_and product_version*> is a short string that uniquely identifies the version of this dataset. It may also indicate the processing version of the parent dataset from which this product is produced. In either case, this version indicator MUST be updated if a revised dataset is published

NB – these strings MUST NOT contain spaces, underscores, slashes (/) or backslashes (\) and should be amenable to FTP wildcard selection.

These names must be assigned to files at creation time, as they are not renamed during the publication process on the ESGF. The information contained in a filename cannot be necessarily inferred from the file content, so the provider must deliver a properly constructed and named file.

7. Revision History of this document

- 8 Nov 2012 - Released as v1