obs4MIPs data specifications (ODS) v2.1

Peter J. Gleckler, Karl E. Taylor, Paul J. Durack, Denis Nadeau, Jim Biard, Robert Ferraro, Stephan Finkensieper, Scott Stevens, Matthias Tuma
and
contributing members from the WDAC, WDAC task team, and others…

Latest version finalized : September 22, 2017 (v2.1.0)

Short URL: https://goo.gl/jVZsQl

Link to obs4MIPs tables and controlled vocabulary 2.0.0

ODS v2.1 strives to improve upon the original obs4MIPs data specifications (ODS v1.0) by capturing additional useful information while at the same time retaining consistency with both ODS v1.0 (for backwards compatibility) and the CMIP6 data specifications. We believe most, if not all of the required global attributes that will be relied on through CMIP6 are described below. However, minor updates to this document can be expected throughout 2017 as some of the underlying metadata tables are revised and ESGF search facets are upgraded. Any changes will be noted with an increment to the ODS version number (Appendix 3), which is currently "2.1.0".

We strongly encourage obs4MIPs data providers to communicate with us (obs4MIPs-admin@llnl.gov) as they prepare their data contributions with CMOR3 (https://goo.gl/LePm8H) following the metadata specifications below.

CONTENTS

Introduction

The purpose of obs4MIPs is to facilitate comparison of observational data to model output from WCRP intercomparison projects (notably the Coupled Model Intercomparison Project, CMIP). To accomplish this, the organization of CMIP and obs4MIPs data are closely aligned, including the data structure and metadata requirements, and the infrastructure relied on for archival, searching and accessibility via the Earth System Grid Federation (ESGF) using the CoG.

The original set of obs4MIPs contributions adhered to guidelines (ODS V1.0, circa 2012) that were based on the CMIP5 data specifications. Now, as described below, the obs4MIPs data specifications have been refined to be closely aligned with the CMIP6 data specifications, which are overseen by the Working Group on Coupled Models (WGCM) Infrastructure Panel (WIP). Obs4MIPs governance is provided by WCRP's Data Advisory Council's "Observations for Model Evaluation Task Team" (or simply "obs4MIPs TT").

To avoid naming conflicts in the identification of data sources and institutions, it is necessary for data providers to register the name of their institution and information about their data sets *prior to generating* obs4MIPs data sets. This is done by submitting an "issue" on the obs4MIPs-cmor-tables Github repository. (See Appendix 2 for the information requested.)

obs4MIPs metadata specifications for the following categories are described in this document:

1. Global attributes: Explicitly defined metadata included in netCDF files that describe the contents of the dataset and provenance (e.g., "variable_id").
2. Data Reference Syntax (DRS): Some attributes (e.g.,"institution_id', "source_id", "variable_id") comprise the data reference syntax (DRS) for obs4MIPs, which closely parallels the DRS of CMIP6. The DRS is used, for example, in file names, directory structures, the further_info_url, and in facets of some search tools such as ESGF.
3. Directory structure: This can be thought of in the traditional sense, and includes selected DRS information used as subdirectories. It is explicitly defined in obs4MIPs to facilitate searching for data with the ESGF.
4. Filenames: Like the directory structure, filenames are constructed based on a template that relies mainly on a subset of the CV entries.

The global attributes are constructed to facilitate organization of the obs4MIPs datasets, and in particular for providing a useful set of options (or facets) for data exploration via CoG/ESG. The purpose of the DRS is more behind the scenes than the search functions - but it is critical in defining how data from the complex CMIP6 and obs4MIPs databases is organized (notably the directory structure). These obs4MIPs data structures are curated by the authors of this document in consultation with the obs4MIPs TT and the WIP.

Satisfying these (or the CMIP6) data requirements can be facilitated by using the [Climate Model Output Rewriter](#) (CMOR3), as described in a [separate document](#). Use of CMOR3 is not required for producing obs4MIPs data, but it is strongly recommended because CMOR3 ensures that the necessary metadata for distributed data searching is included. A user populates an input table with the entries of the required global attributes, and from this, CMOR3 automatically produces the DRS and filenames.

Beyond metadata, this document includes a discussion of the inclusion (or not) of grid cell bounds. This information makes it possible for users of the data to weight the influence of each cell based on its surface area (or volume), e.g., for calculating a global mean. A user can estimate the grid cell bounds, but in general those closest to the original data are better positioned to estimate them. We recommend that data providers include grid bounds ([Appendix 1](#)).

Global attributes

Table 1 contains the list of obs4MIPs global attributes, indicating which are required and which are optional. The values for many of the global attributes must be drawn from special obs4MIPs "controlled vocabularies" (CVs). A CV, in simplest form, is a list of the permitted values that can be assigned to a given global attribute. The lists of permitted values can be found in the [reference CVs for obs4MIPs](#) maintained on a github repository.

Table 1: obs4MIPs global attribute description

and comparison with the original obs4MIPs conventions (ODS V1.0, circa 2012)

Table color key:

Name or form has been changed relative to CMIP5 and/or ODS V1.0

New attribute for obs4MIPs

Controlled Vocabulary (CV)

Registered Content (RC)

obs4MIPs required global attributes see note 1	description	examples	related attribute in original obs4MIPs circa 2012	form see note 2	when required?	further information and rationale
activity_id **(see note 3)**	activity identifier	only value permitted is "obs4MIPs"	project_id	CV	always	Renamed more generically, since not all activities are projects; also multiple activities may now be listed separated by single spaces
comment	see note 9	see note 9	comment	free form	never	No change from original obs4MIPs or CMIP5; CF-convention standard
contact	see note 9	see note 9	contact	free form	always	Still required with ODS-2.0, but anticipated to be superseded by "further_info_url" (below) maintained by ESDOCs
Conventions	convention version	"CF-1.7 ODS-2.1" or higher	Conventions	CV	always	Updated version from ODS-1.0 with a list of conventions separated by single spaces now allowed
creation_date	date file was created	see note 5	creation_date	structured form	always	No change from original obs4MIPs specs (automatic with CMOR3)

data_specs_version	version identifier of obs4MIPs CV's and CMOR tables	2.1.0	*NEW*	CV	always	This version number is associated with the version or github "tag" for the obs4MIPs CMOR tables and the CV's. See CV discussion and PCMDI/obs4MIPs-cmor-tables which originate from the CMIP6 tables
external_variables	external cell measures	"areacella", "areacello", "volcella", "volcello", as defined in CMIP6	*NEW*	CV	rarely needed, but include when appropriate	List of cell measure variables (separated by single spaces) that are referenced but not included in the file. These variables will be stored independently in the obs4MIPs data archive. Use of this attribute is expected to be infrequent in contrast to the recommended inclusion of grid bounds (Appendix 1)
frequency	sampling frequency	"day"	frequency	CV	always	No change from original obs4MIPs. The current options are given in obs4MIPs_frequency.json
further_info_url	location of documentation	see note 6	*NEW*	CV based	always	Efforts are underway to coordinate with ES-DOCs to create urls for each dataset. It is expected that this will point to documentation that can be updated by the data provider
grid	grid	see note 7	*NEW*	free form	always	Briefly describes output grid characteristics

grid_label	grid identifier	"gn", "gr1" see note 8	*NEW*	CV (note 11)	always	Used in file name to distinguish among files when the variable is reported on more than one grid. See obs4MIPs_grid_label.json
history	see note 9	see note 9	history	free form	never	No change; CF-convention standard
institution	institution name	"NOAA's National Centers for Environmental Information, Asheville, NC 28801, USA"	institution	CV with registered content	always	Can be used to identify institute responsible for dataset including a curator role if data is no longer managed by original individual or entity. These entries must be registered in: obs4MIPs_institution_id.json
institution_id Will have ESGF alias with institution_id to be backwards compatible with ODS v1.0	institution identifier	"NCEI"	institute_id	CV with registered content	always	Was "institute_id"; name changed to parallel other global attributes; this string is constructed only using the character set: a-z, A-Z, 0-9, and "-". These entries must be registered in: PCMDI/obs4MIPs-cmor-tables/blob/master/obs4MIPs_institution_id.json
license	license restrictions	see note 13	*NEW*	some required text	always	Ensures that anyone using the files has access to the terms of use
nominal_resolution	approximate horizontal resolution	"50 km", "100 km", "250 km", "1x1 degree". (See Appendix 2 of CMIP6 specifications)	*NEW*	CV	always	Added in CMIP6 to provide an indication of approximate output grid resolution. See obs4MIPs_nominal_resolution.json
product	product type	"observations", "reanalysis"	product	CV	always	As in original obs4MIPs and CMIP5

realm	realm(s) where variable is defined	"atmos", "ocean", "land", "sealce" "atmosChem"	realm	CV	always	As in original obs4MIPs. See obs4MIPs_realm.json
references	see note 9	see note 9	references	free form	never	No change; CF-convention standard
region	Pre-defined (CF conventions) approximate regions	"north_america", "global", "global_land"	*NEW*	CV	always	See obs4MIPs_region.json
source (modified form)	full observational dataset name/version	see Appendix 2	source	CV (Generated from registered content)	always	More comprehensive than original obs4MIPs and CMIP5. See Tables/obs4MIPs_CV.json and search for source_id
source_id (modified form)	Observations identifier	"GPCP-2-4-1" See Appendix 2	source_id	CV (Generated from registered content)	always	Edited version of first part of "source" (with forbidden characters like spaces and periods replaced with hyphens); used in constructing the file name. See obs4MIPs_source_id.json for examples
source_label	label used to identify source (independent of source version)	"GPCP" See Appendix 2	*NEW*	CV with registered content	always	This should be the same as source_id, but without a version number. It will likely be used in faceted searches to get a truncated list of sources (without all the different versions listed)

source_type	Observational class	'satellite_retrieval', 'satellite_blended', "gridded_insitu", "reanalysis"	Mixture of source_type and product	CV see Appendix 2	always	See obs4MIPs_source_type.json. Additional entries may be added over time
source_version_number	Numeric version identifier	v1.0, 1.2 ver2.3.1 See Appendix 2	*NEW*	CV with registered content	always	See obs4MIPs_source_id.json and search for source_version_number
title	see note 4	see note 4	title	free form	never	no change; CF-convention standard
tracking_id	unique file identifier (automatically generated by CMOR)	see note 14	tracking_id	structured form with some CV	always	Form modified to facilitate use by ESGF
variable_id	variable identifier	"tas", "pr", "ua"	variable_id	CV	always	Added to direct users and software to the primary variable of interest in the file. The complete list of CMIP6 variable_ids is here
variant_info	description of run variant	"Best Estimate", "Sphere of influence = 20km" See note 15	*NEW*	free form	as appropriate	Provides brief description of variant differences. Not needed in most cases, when there is only a "Best Estimate" (BE)
variant_label	"variant" label	Usually "BE" (for "Best Estimate"); rarely "r1", "r2", r123" See note 16	*NEW*	structured form	always	In most cases the value is "BE". Only when there are multiple estimates of the same source_id (e.g., constructed with alternate processing choices) is the value different in which case it should follow CMIP6, e.g., with "BE", "r1", "r2", "r3", ...

Table Notes:

1. If CMOR is used to write output files (strongly recommended), an additional global attribute will automatically be included: cmor_version, but this is not an obs4MIPs required global attribute

2. "CV" means content must be taken from a "controlled vocabulary" defined by the WDAC's task team in coordination with the WIP. "registered content" (RC) is special controlled vocabulary defined by data contributor and monitored by the obs4MIPs TT. Data contributors can submit RC at (https://github.com/PCMDI/obs4MIPs-cmor-tables) and contact obs4MIPs-admin@llnl.gov

3. For backwards compatibility with ODS V1, project_id will be an alias of activity_id on ESGF.

4. Since some software uses the 'title' for default plotting or describing the contents of a file, it can be used to provide a description that is similar or equivalent to the source. A common entry may be the same as the source but without the release_year.

5. creation_date form: YYYY-MM-DDTHH:MM:SSZ (e.g., "2010-03-23T05:56:23Z")

6. further_info_url has the form
https://furtherinfo.es-doc.org/<activity_id>.<institution_id>.<source_label>.<source_id>.<variable_id>
(e.g., "https://furtherinfo.es-doc.org/obs4MIPs.RSS.REMSS-PRW.REMSS-PRW-6-6-0.prw"). The further_info_url page will be maintained by the ES-DOCs project and will simply be a rendering by the Viewer tool of information provided by modeling groups and recorded in so-called CIM documents.

7. The "grid" global attribute can be used to describe the horizontal grid and regridding procedure. There is no standard form used to record this information, but it is suggested that when appropriate the following be indicated: brief description of native grid and resolution, and if data have been regridded, regridding procedure and description of target grid. Here are some examples:

      grid = "data regridded to a CMIP6 standard 1x1 degree latxlon grid from the native T63 grid using an area-average preserving method"
      grid = "data regridded via bilinear interpolation to a 3x3 deg latxlon grid from the native atmosphere T63 gaussian grid (64x128 latxlon)"
      grid = "data regridded to a CMIP6 standard 1x1 degree latxlon grid from the native T63 grid using an area-average preserving method"
      grid = "data regridded via bilinear interpolation to a 3x3 deg latxlon grid from the native atmosphere T63 gaussian grid (64x128 latxlon)"

8. Data providers may choose to report their output on one or more grids, or with special care provide an alternate projection. To distinguish between output reported on different grids, a "grid_label" attribute is defined. The original grid should be labeled with "gn". Additional grids should be labeled using the form "gr[i]" where i is a positive integer. If the data is subsequently regridded to a second and third grid, "gr2" and "gr3" could be used to distinguish these grids from "gr1".

9. A description and examples of the contact, comment, history and references global attributes may be found in the document:
CMIP5_output_metadata_requirements (http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf).

13. The wording of the "license" attribute is up to the dataset creator, but it is recommended that you reference one of the "Creative Commons" licenses. Here is some possible text: "Data in this file produced by <Your Centre Name> is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/). Use of the data must be acknowledged following guidelines found at <a URL maintained by you>. Further information about this data, including some limitations, can be found via <some URL maintained by you>."

14. tracking_id should be of the form hdl:21.14102/<uuid> (e.g., "hdl:21.14102/02d9e6d5-9467-382e-8f9b-9300a64ac3cd"). The tracking_id should be unique for each file published in ESGF. CMOR automatically generates a tracking_id. You can review the python built-in UUID library documentation here.

15. Except when variant_label="BE", it is recommended that variant_info include information identifying major distinguishing features of a variant, but care should be taken to record correct information.

16. When multiple versions are provided with the same source_id, the first one should be labeled "r1" and all subsequent versions (derived from different processing choices) will follow the "realization nomenclature of CMIP with "r2", "r3", … "rN".

Data Reference Syntax (DRS) components:

The DRS is used, for example, in file names, directory structures, the further_info_url, and in facets of some search tools. The following components are needed for obs4MIPs:

activity_id   (original obs4MIPs "activity")
realm   (original obs4MIPs "activity")
frequency   (original obs4MIPs: "frequency")
product   (original obs4MIPs: "product")
institution_id   (original obs4MIPs: "institute")
source_id   (original obs4MIPs: "model")
source_label   (new in obs4MIPs)
variable_id   (original obs4MIPs: "variable name")
region   (new in obs4MIPs)
grid_label   (new in obs4MIPs)
variant_label   (original obs4MIPs: "ensemble member")
version   (original obs4MIPs: "version number")
grid_label   (new in obs4MIPs)

File name template:

The obs4MIPs file name must be constructed consistent with the following template.

New obs4MIPs file name template =
<variable_id>_<frequency>_<source_id>_<variant_label>_<grid_label>[_<time_range>].nc

For time-invariant fields, the last segment (time_range) above is omitted.

All strings appearing in the file name are constructed using only the following characters: a-z, A-Z, 0-9, and the hyphen ("-"), except the hyphen must not appear in variable_id. Underscores are prohibited throughout except as shown in the template.

Note that the last segment of the file name indicates the time-range spanned by the data in the file, and is omitted when inappropriate. The format for this segment is the same as in CMIP6 (see Table 2 of the CMIP6 specs document: http://goo.gl/v1drZl).

For comparison, here is the CMIP6 file name template:
<variable_id>_<table_id>_<source_id>_<experiment_id >_<member_id>_<grid_label>[_<time_range>].nc

and the legacy obs4MIPs file name template:
<variable>_<instrument>_<processing_level>_<processing version>_<start_date>-<end_date>.nc

Directory structure template:

The obs4MIPs directory structure must be constructed consistent with the following template.

New obs4MIPs directory structure =
<activity_id>/
    <institution_id>/
        <source_id>/
            <frequency>/
                <variable_id>/
                    <grid_label>/
                        <version>

For comparison, here is the CMIP6 directory structure:

<mip_era>/
    <activity_id>/
        <institution_id>/
            <source_id>/
                <experiment_id>/
                    <member_id>/
                        <table_id>/
                            <variable_id>/
                                <grid_label>/
                                    <version>

and the legacy obs4MIPs directory structure:

obs4MIPs/
    observations/
        <realm>/
            <variable_id>/
                <frequency>/
                    <grid>/
                        <Institution_id>
                            <instrument>/
                                <version>/

Notes:
<version> here refers to the CMOR-assigned version number which has the form "vYYYYMMDD" (e.g., "v20170921"), indicating a representative date for the version was produced for obs4MIPs. For those not using CMOR, the convention must be followed.

Sample file header

The example below was produced from the following demo: https://github.com/PCMDI/obs4MIPs-cmor-tables/demo

KEY: <mark>yellow</mark> means "absolutely essential"

```
netcdf prw_mon_REMSS-PRW-6-6-0_BE_gn_198701-198812 {
dimensions:
    time = UNLIMITED ; // (24 currently)
    lat = 72 ;
    lon = 144 ;
    bnds = 2 ;
variables:
    double time(time) ;
            time:bounds = "time_bnds" ;
            time:units = "days since 1987-1-1 0:0:0" ;
            time:calendar = "gregorian" ;
            time:axis = "T" ;
            time:long_name = "time" ;
            time:standard_name = "time" ;
    double time_bnds(time, bnds) ;
    double lat(lat) ;
            lat:bounds = "lat_bnds" ;
            lat:units = "degrees_north" ;
            lat:axis = "Y" ;
            lat:long_name = "latitude" ;
            lat:standard_name = "latitude" ;
    double lat_bnds(lat, bnds) ;
    double lon(lon) ;
            lon:bounds = "lon_bnds" ;
            lon:units = "degrees_east" ;
            lon:axis = "X" ;
            lon:long_name = "longitude" ;
            lon:standard_name = "longitude" ;
    double lon_bnds(lon, bnds) ;
    float prw(time, lat, lon) ;
            prw:standard_name = "atmosphere_water_vapor_content" ;
            prw:long_name = "Water Vapor Path" ;
            prw:comment = "vertically integrated through the atmospheric column" ;
            prw:units = "kg m-2" ;
            prw:cell_methods = "area: time: mean" ;
            prw:missing_value = 1.e+20f ;
            prw:_FillValue = 1.e+20f ;
            prw:valid_min = 2.f ;
            prw:valid_max = 3.f ;

// global attributes:
            :Conventions = "CF-1.7 ODS-2.1" ;
            :activity_id = "obs4MIPs" ;
```

```
    :contact = "RSS (support@remss.com)" ;
    :creation_date = "2017-11-08T20:36:13Z" ;
    :data_specs_version = "2.1.0" ;
    :frequency = "mon" ;
    :further_info_url = "https://furtherinfo.es-doc.org/obs4MIPs.RSS.REMSS-PRW.REMSS-PRW-6-6-0.prw" ;
    :grid = "2.5x2.5 degree latitude x longitude" ;
    :grid_label = "gn" ;
    :history = "2017-11-08T20:36:13Z; CMOR rewrote data to be consistent with obs4MIPs, and CF-1.7 ODS-2.1 standards" ;
    :institution = "Remote Sensing Systems, Santa Rosa, CA 95401, USA" ;
    :institution_id = "RSS" ;
    :nominal_resolution = "250 km" ;
    :product = "observations" ;
    :realm = "atmos" ;
    :region = "global" ;
    :source = "REMSS-PRW 6.6.0 (2017): Water Vapor Path" ;
    :source_id = "REMSS-PRW-6-6-0" ;
    :source_label = "REMSS-PRW" ;
    :source_type = "satellite_blended" ;
    :source_version_number = "6.6.0" ;
    :table_id = "obs4MIPs_Amon" ;
    :table_info = "Creation Date:(04 October 2017) MD5:9516b952d4e1c5084ec8c8e481e6cf35" ;
    :title = "PRW v6.6.0 prepared for obs4MIPs (ODS-v2.1.0)" ;
    :tracking_id = "hdl:21.14102/0dc9218d-6923-4415-8fb0-93cd9fa734c1" ;
    :variable_id = "prw" ;
    :variant_info = "Best Estimate" ;
    :variant_label = "BE" ;
    :license = "Data in this file produced by <Your Centre Name> is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (https://creativecommons.org/licenses/).
Use of the data must be acknowledged following guidelines found at <a URL maintained by you>. Further information about this data, including some limitations, can be found via <some URL maintained
by you>" ;
    :cmor_version = "3.2.8" ;
}
```

Appendix 1: Coordinate bounds

While climate models have grid cells with boundaries that have exact definitions allowing analysts to precisely compute area integrals (such as a domain average), they are often not well defined for gridded observational products. If the analyst wants to compute a similar integral from observations, the area associated with each gridded value must be estimated. We therefore recommend that the data provider include their own estimates of the latitude-longitude grid cell bounds from which the grid cell areas can be calculated (since they are most familiar with the data). This will help insure uniformity among different researchers analyzing the data.

Appendix 2: Guidance for defining and registering source information (product description)

We need 7 pieces of information from obs4MIPs data providers to enable search functioning and other services on ESGF. (Note that we are referring here to only the "source" information which collectively is used to identify a dataset; data providers will also need to register their institution, and if they are contributing a new variable, not already defined by CMIP6, they will need to register the variable.). The 7 attributes are listed below, each with 3 examples:

1. **source_name** ("brand name") of the dataset. (See below for further guidance.)
   "REMSS PRW"
   "GPCP"
   "NOAA NCEI AVHRR NDVI"

2. **release_year** (year this version of the data was produced and made available; the year that the data was reformatted for obs4MIPs is irrelevant):
   "2017"
   "2003"
   "2013"

3. **source_description** of the dataset which usually is simply an expansion of the acronyms or abbreviations appearing in source_name
   "Remote Sensing Systems precipitable water"
   "Global Precipitation Climatology Project"
   "NOAA Nat Cent ... AVHRR Normalized difference vegetation index"

4. **source_version_number** of the dataset (following whatever convention the data provider prefers)
   "V6.6.0"
   "2.3"
   "V4.0"

5. **institution_id** (acronym(s) used to identify the institution, group, or consortium responsible for producing the data set)
   "RSS"
   "UofMD"
   "NOAA NCEI"

6. the **region** covered by the dataset (see obs4MIPs_region.json for options; the smallest appropriate region among the options should be used)
   "global"
   "global_ocean"
   "global_land"

7. the **source_type** (see obs4MIPs_source_type.json for options)
    "gridded_insitu"
    "satellite_blended"
    "satellite_retrieval"

From the information above, the obs4MIPs team constructs the following 3 items:

8. **source_label** = <source_name>, but substituting "-" for certain forbidden characters (including ".", "_", "(", ")", "/", and " ").
    "REMSS-PRW"
    "GPCP"
    "NOAA-NCEI-AVHRR-NDVI"

9. **source_id** = <source_label>-<source_version_number> but substituting "-" for certain forbidden characters (including ".", "_", "(", ")", "/", and " ").
    "REMSS-PRW-6-6-0"
    "GPCP-2-3"
    "NOAA-NCEI-AVHRR-NDVI-4-0"

10. **source** = <source_name> <source_version_number> (<release_year>): <source_description>
    "REMSS PRW v6.6.0 (2017): Remote Sensing Systems precipitable water"
    "GPCP 2.3 (2003): Global Precipitation Climatology Project"
    "NOAA NCEI AVHRR NDVI v4.0 (2013): NOAA Nat Cent ... AVHRR Normalized difference vegetation index"

In summary, data providers need to *register content* for the following: 1) source_name, 2) release_year, 3) source_description, 4) source_version_number, 5) institution_id, 6) region and 7) source_type. They can do this by submitting an issue on the obs4MIPs-cmor-tables github repository. From this information the source_label, source_id and source are constructed. CMOR will record as global attributes all items except the first 3: source_name, release_year, and source_description (because they are included in "source").

Notes:

1) The *source_name* must be unique across all obs4MIPs datasets. It should be as short as possible because it appears in file names, directory structures, search lists, and the like.

2) For an existing dataset that is now being prepared for obs4MIPs, the *source_name* might be an established name associated with the dataset and recognizable by the community (e.g., GPCP, GHCN). For new data sets that do not yet have a recognized brand name, *source_name* might (but is not required to) include an acronym with some indication of: who the data provider is, what instrument the observations are based on, and/or the name of the observed variable. The *source_name* should not be too generic because each separate contribution to obs4MIPs (current or future) must have a unique source_name. [If, for example, you are currently contributing a sea surface temperature dataset, but expect later to prepare a surface salinity dataset, you should anticipate that when constructing your source_name. In this case you wouldn't want the source_name to just be the name of your group. Similarly, you wouldn't want your source name to be just the name of the variable because it is likely other groups will also want to contribute the same variable to obs4MIPs.] Consider what name you would like the dataset to be known by in the years to come.

3) If an obs4MIPs contribution includes multiple, related datasets (i.e., more than one variable), coming, for example, from a self-consistent analysis procedure, then all the variables would normally share a common source_name and source_version_ number.

4) If a single group plans to prepare multiple *unrelated* datasets, then each dataset should be assigned a different source_label. By "unrelated" we mean they would normally be independently assigned "release" (or version) numbers. If these multiple, unrelated datasets each deals with a different variable, then one option would be to include the variable name as part of the source_label, but this is not a requirement. The provider might have other ways to distinguish between these datasets.

Note: If a minor problem is found in obs4MIPs datasets (e.g., a mistake in one of the global attributes), the mistake can be corrected without issuing a new "release" with a new version. The data can be republished on ESGF, and ESGF's versioning system will hide the deprecated file(s) (and make it easy for users to determine whether they have the most up-to-date files).

Appendix 3: Document version information

0.1 (23 October, 2016) - Document initiated with CMIP6 analog. *Not public*
0.2 (23 October - 22 February 2017) - working on draft document in relation to CMIP6, CMIP5 and original obs4MIPs specs
0.3 (March 01 - April 05 2017) - Added "region" as global attribute, finessing Table 1 footnotes
0.4 (April 25) - Draft made public on obs4MIPs CoG site enabling broader feedback
2.0 beta (June 25) - Public on obs4MIPs CoG site

[2.0 released (June 30, 2017)](#)

2.1 released (September 22, 2017)
- source_id now is a compound structure with source_label followed by the version number. Example: source_label="GPCP", source_id="GPCP-2-4-1". This is described in Appendix 2
- table_id was eliminated since table names have very little meaning now in CMIP. "frequency" will be used in place of table name to distinguish between a variable sampled at different frequencies. This is expected to be more compatible with CMIP7
- source_version_number was added
- "region" is now a required attribute (so that we can populate a search facet)
- typos and links were corrected and clarifying text and links were added
- CMOR software has been incremented from 3.2.x to 3.2.7 with changes made to implement the revised specifications