CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies

Karl E. Taylor, V. Balaji, Steve Hankin, Martin Juckes, Bryan Lawrence, and Stephen Pascoe

Version 1.0

4 November 2010.

Style Definition: Normal

Deleted: .27

Deleted: 7 April

1 Introduction

1.1 Scope

This document provides a common naming system to be used in files, directories, metadata, and URLs to identify datasets wherever they might be located within the distributed CMIP5 archive It defines controlled vocabularies for many of the components comprising the data reference syntax (DRS).

Deleted: federation.

1.2 Context:

The CMIP5 archive will be distributed among several centers using different storage architectures. As far as possible these differences should be hidden from the user.

The data reference syntax (DRS) should be sufficiently flexible to cover all the services that the archive might wish to offer, even though resource limitations may restrict the services that are actually delivered within the CMIP5 time frame. The DRS needs to take account of the user resources (usually a file system based data store) and the software to be used by the archive (such as OPeNDAP). The context in which the system will be used will require a compromise between brevity and clarity but there should be no ambiguity and easily accessible expansions of all terms.

1.3 Purpose

The Data Reference Syntax (DRS) should provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive and of files delivered to users. The DRS should make use of controlled vocabularies to facilitate documentation and discovery. Providing users with data in files with well structured names will facilitate management of the data on the users' file systems and simplify communication among users and between users and user support. The controlled vocabularies will be useful in developing category-based data discovery services. The elements of the controlled vocabularies will occur frequently in software and web pages, so they should be chosen to be reasonably brief, reasonably intelligible, and avoid characters which may cause problems in some circumstances (e.g. "/", "(", ")").

1.4 Use Case and Requirements

There are 6 specific use cases which the DRS must support:

1. Those responsible for replicating data within the CMIP5 archive should be able to exploit the DRS to guide what needs to be replicated, and to where.
2. Those responsible for the federation catalogues should be able to use the DRS to identify to catalogue users unambiguously which replicants are available for download or for on-line access (such as OPeNDAP).
3. Those responsible for the archives should be able to use the DRS to define a logically structured file layout (if they use file systems as their storage management system).
4. Users should be able to modify download scripts in a completely transparent manner, so that for example, a slow wget from one site, can be repeated (or finished) using a script in which only the hostname part of the DRS has been changed.
5. The names of the core datasets should be predictable enough that, for example, a user having found and downloaded or accessed data on-line from one model simulation using a script can modify that script to download or access another model and/or simulation with only knowledge of the relevant controlled vocabulary terms (in this case, the model and/or simulation names).
6. The DRS should be sufficiently extensible to describe variables and time periods beyond those defined in the CMIP5 core.

Deleted: federation

2 Definitions

2.1 Atomic dataset

Deleted: :

Model archives consist of collections of "atomic datasets", defined as follows:

Atomic dataset definition: a subset of the output saved from a single model run which is uniquely characterized by a single *activity*, *product*, *institute*, *model*, *experiment*, data sampling *frequency*, *modeling realm*, *variable name*, *MIP table*, *ensemble member*, and *version number*.

Deleted: The collection of data constituting

Deleted: single *product*

Deleted: *identifier*

The definition is intended to provide a well-founded naming system to record archive contents in a structured way. An atomic dataset consists of one variable (field). For each variable the atomic dataset contains the entire spatio-temporal domain, with values reported at each included time and location. An "atomic dataset" may be a very large entity, with 1000 years of daily model output or more; it does not necessarily represent a chunk of data that can practically be put into a single file. The first nine components (*activity*, *product*, *institute*, *model*, *experiment*, *frequency*,

Deleted:

Deleted: eight

Deleted: *product*,

modeling_realm, *variable name* and *MIP table*) should all come from controlled component vocabularies, and the structure for the last two components is also controlled.

2.2 Publication-level dataset

When applied to the CMIP5 experiment definition the atomic dataset definition above leads to millions of atomic datasets. This level of granularity is too fine for the data management technologies employed for CMIP5, therefore atomic datasets are aggregated into "publication-level" datasets[1] containing all variables for a single combination of other DRS components.

Publication-level dataset definition: The collection of atomic datasets which share a single combination of all DRS component values except *variable name* but which might include only selected time intervals (i.e., not necessarily the entire temporal domain) of the contributing atomic datasets. The publication-level dataset therefore represents, in general, an intersection of several atomic datasets.

Note that the *version number* component is effectively a property of publication-level datasets.

2.2 Component Definitions and Controlled Vocabularies

After seeking community input, PCMDI has final authority for defining the controlled vocabularies that together with the component categories comprise the DRS. These components and vocabularies are defined below. (See also Appendix 1.1 and Appendix 1.2.).

Activity identifies the model intercomparison activity or other data collection activity. For CMIP5 all the archived data will be discoverable under the "CMIP5" activity. For "Transpose AMIP", the data will be archived under the "TAMIP" activity. In some cases there may be other activities (e.g., CFMIP and PMIP), which have been coordinated with CMIP5, so these activities may be cross-referenced or aliased with CMIP5 for certain portions of the CMIP5 archive.

Product currently has three options: "output", "output1" and "output2". For CMIP5, files will initially be designated as "output". Subsequently, the data will be assigned a version (see below) and placed in either "output1" or "output2". In some cases a continuous sequence of model data will be split between "output1" and "output2" in order to facilitate archive management.

It is likely that various data products derived from this output will be produced subsequently which could be identified by a different term (e.g., "derived" or "processed"), but this is not part of the current DRS.

[1] Publication-level datasets have previously been referred to as "Realm-level datasets" in internet communications related to CMIP5 such as email lists and wiki pages.

Deleted: and

Deleted: *identifier*

Deleted: *This*

Deleted: explicitly does not distinguish between the different temporal portions of an experiment – these are notionally subsets of

Deleted: – even though in CMIP5 different portions of the same experiment may be assigned different priorities (e.g., extensions of some of

Deleted: future scenario runs to the 22nd and 23rd centuries have been designated to be lower priority than

Deleted: 21st century portion of these simulations).

Deleted: : This component will allow the DRS to be extended to other

Deleted: intercomparisons and

Deleted: archives

Deleted: : This allows distinctions between various types of model data products. For CMIP5 the only initially permissible

Deleted: for product are

Deleted: which refers to all the model

Deleted: published, and "requested", which refers to

Deleted: subset of model output specified by the CMIP5

Deleted: request

Deleted: [2]. The "requested" product type allows users (

Deleted: data managers) to focus on the subset of the complete output that is likely to be available from most of the models. Note that an atomic dataset defined under the "requested" classification will also be part of (or in some cases identical to) an atomic dataset defined under the "output" classification. There may also be "output" atomic datasets that do not include any "requested" data. [WILL POSSIBLY MODIFY THE ABOVE IF WE DON'T NEED TO KNOW ABOUT "REQUESTED"].

Institute identifies the institute responsible for the model results (e.g. UKMO), and it should be as short as possible. For CMIP5 the institute name will be suggested by the research group at the institute, subject to final authorization by PCMDI.

Model identifies the model used (e.g. HADCM3, HADCM3-233). Subject to certain constraints imposed by PCMDI, the modeling group will assign this name, which might include a version number (usually truncated to the nearest integer). The model identifier will normally change if any aspect of the model is modified (e.g., if the resolution is changed). An exception may be made if the modifications to the model are clearly implied by the experiment design. If, for example, a coupled atmosphere-ocean model performs an AMIP simulation (which clearly implies prescribed SSTs and sea ice, rather than a fully interactive ocean), then the name may not necessarily be modified. Another exception is when closely-related "perturbed physics" versions of a model are run, in which the different model versions can be uniquely identified by assigning each a different "p" value in defining the "ensemble member" (described below).

Experiment identifies either the experiment or both the experiment *family* and a specific *type* within that experiment family. In CMIP5, for example, "rcp45" refers to a particular experiment in which a "representative concentration pathway" (RCP) has been specified which leads to an approximate radiative forcing of 4.5 W m^{-2}. As another example, "historicalGHG" is a simulation of the historical" period, but with forcing other than anthropogenic "greenhouse gas" forcing suppressed. In this latter case, "historical" is the experiment *family* and "GHG" is used to designate the specific *type* of historical run. These experiment names are not freely chosen, but come from controlled vocabularies defined in the Appendix 1.1 of this document under the column labeled "Short Name of Experiment".

Frequency indicates the interval between individual time-samples in the atomic dataset. For CMIP5, the following are the only options: "yr", "mon", "day", "6hr", "3hr", "subhr" (sampling frequency less than an hour), "monClim" (climatological monthly mean) or "fx" (fixed, i.e., time-independent). These are specified for each variable in the "standard_output" spreadsheet found at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html.

Modeling realm indicates which high level modeling component is of particular relevance for the dataset. For CMIP5, permitted values are: "atmos", "ocean", "land", "landIce", "seaIce", "aerosol" "atmosChem", ocnBgchem (ocean biogeochemical). These are specified for each variable in the "standard_output" spreadsheet which can be accessed at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that sometimes a variable will be equally (or almost equally relevant) to two or more "realms", in which case the atomic dataset might be assigned to a primary "realm", but cross-referenced or aliased to the other relevant "realms".

Variable name and the MIP table component of the DRS (defined next) identify the physical quantity and often imply something about the sampling frequency, and modeling realm. For CMIP5 the variable name and MIP table for requested output appear in the "standard_output" spreadsheet available at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Monthly mean surface air temperature, for example, has a "variable name" of "tas" and is found in the "Amon" MIP table., Note that hyphens (-) are forbidden in CMIP5 variable names.

MIP table: See description under the "variable name" component directly above. For CMIP5 each MIP table contains fields sampled only at a single frequency (although in the case of

Deleted: : This

Deleted: : This

Deleted: : This

Deleted: "

Deleted: run

Deleted: : This

Deleted: : This

Deleted: **identifier**: For CMIP5, each variable is uniquely identified by a combination of two strings: 1) a "variable name" associated generically with the variable (typically, as recorded in the netCDF file – e.g., tas, pr, ua), and 2) a "variable table"

Formatted: component Char

Deleted: (e.g., Amon, day, aero) in which the variable appears. These two components of the variable name are defined in

Deleted: "standard_output" spreadsheet found at http://cmip-pcmdi.llnl.gov/cmip5/output_req.html. Note that for CMIP5 a variable is also uniquely defined by

Deleted: DRS component "

Deleted: "

Deleted: alone (without reference to a specific

Deleted:). Note that within CMIP5 variable names,

Deleted: ('-')

monthly mean data the DRS will place some of the monthly means in the "mon" DRS frequency category and others in the monClim DRS frequency category, as appropriate.

Ensemble member (r<N>i<M>p<L>): This triad of integers (N, M, L), formatted as shown above (e.g., "r3i1p21") distinguishes among closely related simulations by a single model. All three are required even if only a single simulation is performed.

The so-called "realization" number (a positive integer value of "N") is used to distinguish among members of an ensemble typically generated by initializing a set of runs with different, but equally realistic, initial conditions. CMIP5 historical runs initialized from different times of a control run, for example, would be identified by "r1", "r2", "r3", etc.). The data supplier must assign a realization number to each atomic dataset. It is generally recommended that the numbers be assigned sequentially starting with 1 (but other recommendations, specified below, may override this recommendation). In CMIP5, time-independent variables (i.e., those with frequency="fx") are not expected to differ across ensemble members, so for these N should be invariably assigned the value zero ("r0"). For TAMIP ("the Transpose AMIP activity), the "realization" number is used to distinguish among the 16 members of each of 4 ensembles (one for each of 4 "seasons") generated from different observed conditions, spaced 30 hours apart. So, for example, the 16-member ensemble of runs initialized at 00Z on 15 Oct 2008, 06Z 16 Oct 2008, 12Z 17 Oct 2008, and so-on, would be assigned "r1", "r2", "r3", etc.

Models used for forecasts that depend on the initial conditions might be initialized from observations using different methods or different observational datasets. These should be distinguished by assigning different positive integer values of "M" in the "initialization method indicator" (i<M>). For CMIP5 this indicator might in some cases be needed to distinguish among runs in the decadal-prediction suite of experiments (1.1-1.6). The data supplier must assign an initialization method number to each atomic dataset. It is recommended that the numbers be assigned sequentially starting with 1. In CMIP5, time-independent variables (i.e., those with frequency="fx") are not expected to differ across ensemble members, so for these M should invariably be assigned the value zero ("i0"). A key (i.e., a table) should be made available that associates each value of M with a particular initialization method and/or observational dataset.

If there are many closely related model versions, which, as a group, are generally referred to as a perturbed physics ensemble (e.g., QUMP or climateprediction.net ensembles), then these should be distinguishable by a "perturbed physics" number, p<L>, where the positive integer value of L is uniquely associated with a particular set of model parameters (e.g., r3i1p78 is a third realization of the seventy-eighth version of the perturbed physics model). If there are different "forcing" combinations prescribed in experiment 7.3 in CMIP5 (the "historicalMisc" runs), then each of these different runs are also assigned different values of L (in "p<L>"). Note that the data supplier must assign a physics version number to each atomic dataset. It is recommended that the numbers be assigned sequentially starting with 1. In CMIP5, time-independent variables (i.e., those with frequency="fx") are not expected to differ across ensemble members, so for these L should always be assigned the value zero ("p0"). A key (i.e., a table) should be made available that associates each value of L with a particular set of model parameter values and/or, in the case of the "historicalMisc" experiment, a particular suite of "forcing" agents.

Deleted: ¶

Deleted: Different simulations that are equally likely outcomes for a particular simulation (i.e., they typically differ only by being started from equally realistic initial conditions) are distinguished by different

Deleted: values

Deleted: "

Deleted: requirements

Deleted: rule out following

Deleted: . Simulations resulting from initializing a model with different *methods*

Deleted: performed as part of

Deleted: suite of

Deleted:

Deleted: that defines the various initialization methods should be made available so that a user can learn which

Deleted: is associated with each value of M.

Deleted: The

Deleted: that defines the various model versions

Deleted: so

Deleted: user can learn which

Deleted: is associated with each value of L.

Note that for a single model and experiment N, M, and L should be interpretable independently; for all members of the ensemble, the correspondence between the values of N, M, and L and the simulation characteristics they represent should be consistent. For example the two different ensemble members, r3i1p7 and r3i1p8, should both be initialized from *exactly the same initial conditions using the same method* (because the "r" and "i" values are identical) although the subsequent evolution of the simulations will presumably differ since they were produced by two different "perturbed physics" versions of the same model. Note that there may be cases where "gaps" could occur in the list of ensemble members. If, for example, two different initialization procedures were used, but the second procedure was tested with only a subset of the initial condition cases of the first procedure (say, every other case). Then the list of ensemble members would look like: r1i1p1, r2i1p1 r3i1p1, r4i1p1, r5i1p1, r6i1p1, …, r1i2p1, r3i2p1, r5i2p1, …

A recommendation for CMIP5 is that each so-called RCP (future scenario) simulation should when possible be assigned the same realization integer as the historical run from which it was initiated. This will allow users to easily splice together the appropriate historical and future runs. Thus, for example, suppose a 3-member ensemble of historical runs of a model exists, and a single rcp45 simulation was produced, initialized from the third member of the historical ensemble. The recp45 simulation would be designated "r3" (rather than "r1"), even though it is the only existing ensemble member, in order to indicate that it was spawned from member 3 of the historical ensemble. A similar convention should be followed, when appropriate, with other simulations (e.g., the decadal simulations).

Version number (vN): The version number will be 'v' followed by an integer, which uniquely identifies a particular version of the output (e.g., perhaps distinguishing between an original version of the output that might have been found to be flawed in some respect--perhaps due to some improper post-processing procedure-- and a subsequent version in which the data were corrected). The version number will be generated from the date, e.g "v20100105" for a version provided on 5th January 2010.

Deleted: (i.e., across all members of the ensemble), the values assigned to

Deleted: (which together define an individual ensemble member) should each be uniquely associated with, respectively, a specific initial condition, initialization method, and perturbed physics version. Thus, these numbers will

Deleted: across all members of an ensemble.

Deleted: There

Deleted: r1i1p2, r3i1p2, r5i1p2, …

Deleted: Another requirement

Deleted: ¶

2.4 Extended Path

Note that thus far we have not considered datasets which might be spatio-temporal subsets. We expect these to exist both as files in the archive as well as virtual files (that is, URLs representing aggregated time series of files that are accessible by services such as OPeNDAP). The DRS supports the specification of such subsets, however, these represent only "parts" of an atomic dataset, and hence they were not included in the definition of atomic dataset above.

Temporal Subsets: Time instants and periods (N1(-N2))

Time instants and periods will be represented by 'yyyy[mm[dd[hh][mm]]][-clim]', where 'yyyy', 'mm', 'dd', 'hh' 'mm' are integer year, month, day, hour, and minute, respectively, and enough (and just enough) of the suffixes should be added to unambiguously resolve the interval between time-samples contained in the file or virtual file URL. (For example, monthly mean data would include "mm", but not "dd", "hh", or "mm"; "subhr" data would include all suffixes.) The optional "-clim" is appended when the file contains a climatology. For example, a file with sampling frequency of "mo" and the time designation 196001-198912-clim represents the monthly mean climatology (12 time values) computed for the period extending from 1/1960-12/1989. Note that the DRS does not explicitly specify the calendar type (e.g., Julian, Gregorian), but the calendar will be indicated by one of the attributes in each netCDF file.

Geographic Subsets

It is (currently) unlikely that geographical subsets described by bounding boxes will be stored in the archive, but subsets by named location might be. Where these appear in the extended Path, they should appear last as gXXXXX where XXXXX is a name from a specific gazetteer (which is yet to be selected).

2.5 Permitted Characters.

The character set permitted in the components needs to be restricted in order that strings formed by concatenating components can be parsed. For the purposes of this scoping exercise, it will be assumed that the components will be used in URLs, punctuated by "/", "=", ":", and "?", and in the names of files delivered to users, punctuated by "." and "_". Thus, none of these characters can be permitted within the component values. Other characters will also be excluded at this time, so the permitted characters will be: a-z, A-Z, 0-9, and "-".

3. Using the DRS Syntax

The DRS component vocabularies are used in various places within the CMIP5 archive to identify digital objects. In each case there are slight variations in the encoding syntax and subset of DRS components used, reflecting the practicalities of mapping DRS concepts to different applications. Here are three use cases for the DRS syntax: in URLs, for a directory layout, and in filenames.

3.1 CMOR directory structure

The standard CMIP5 output tool CMOR2[3] optionally writes output files to a directory structure mapping DRS components to directory names as:

<activity>/<product>/<institute>/<model>/<experiment>/<frequency>/
<modeling realm>/<variable name>/<ensemble member>/

For example

/CMIP5/output/MOHC/HadCM3/decadal1990/day/atmos/tas/r3i2p1/

or

/CMIP5/output/MOHC/HadCM3/rcp45/mon/ocean/uo/r1i1p1/

This structure, based on a previous version of the DRS, is incompatible with the recommended current DRS directory structure (see below). However it remains relevant as a possible structure for model output prior to transforming into the DRS directory structure.

[3] See the Climate Model Output Rewriter: http://www2-pcmdi.llnl.gov/cmor/documentation/

Deleted: Here are three use cases for the DRS syntax: in URLs, for a directory layout, and in filenames. ¶
3.1 URL syntax.¶
When the DRS is used in a URL, we would expect the URL to comprise a hostname, the atomic dataset name, possibly an extended path name, and possibly a service endpoint name. That is, we would expect to see usage like:¶
http://<hostname>/<activity>/<product>/<institute>/<model>/<experiment>/<frequency> /<modeling realm>/<variable identifier>/<ensemble member>/<version>/ [<endpoint>],¶
where square brackets enclose optional elements (in this case, only the service endpoint).¶
Where no service endpoint appears, it should be expected that an HTTP GET on the URL would return the netCDF data. (Currently there is no CMIP5 controlled vocabulary for endpoints, when one appears it will have values which encompass services such as OPeNDAP and WCS etc.)¶
Note that ensemble member and version numbers are mandatory, to ensure that if subsequent versions or ensemble members appear, there is no possibility of ambiguity as to what data is referenced at a given URL.¶
¶
Should replace the following with "real" examples¶
http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/ varname/r1/v1/¶
or¶
http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/
varname/r1/v1/extended_path/ ¶
or¶
http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/
varname/r1/v1/extended_path/service_endpoint¶
or¶
http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/
varname/r1/v1/service_endpoint ¶
Controlling the vocabulary for service endpoints is beyond the scope of this document, but will be a necessary part of the distributed URL design, and impact on what appears in catalogues.¶
However, we might expect that without a service endpoint, dereferencing these URLs will return either netCDF data, or catalogue entries. (Exa ... [1]

Deleted: /CMIP5/output/UKMO/HADCM3/decadal1990/day/atmos/tas/r3i2p1/v1/¶
or¶
/CMIP5/output/UKMO/HADCM3/rcp45/mon/ocean/uo/r1i1p1/v3/¶
Links to the latest version of all files constituting the dataset will appear directly in the *<ensemble member>* directory. The files themselves will be found a level below this in a particular *<version_number>* subdirectory. Note that for CMIP5 the second of the two strings that identify the variable (i.e., the "variable table" discussed in the section on "Controlled Components") does not appear in the directory structure, since the "activity", the "frequency" and the "modeling realm", which already do appear in the directory path, together unambiguously imply a certain table.¶
3.3 Filenames¶

3.2 CMIP5 filename encoding

Because users will download data into a file system that will usually differ from the archival directory structure (and because in some cases it aids in archive management), the filename structure should include some DRS content. For CMIP5 the filename will be constructed as follows:

filename = <*variable name*>_<*MIP table*>_<*model*>_<*experiment*>_
<*ensemble member*>[_<*temporal subset*>].nc

where:

- <*variable name*>, <*MIP table*>, <*model*>, <*experiment*>, and <*ensemble member*> are DRS components,
- The < *temporal subset*> is omitted for variables that are time-independent.

Example:

tas_Amon_HADCM3_ historical_r1i1p1_185001-200512.nc

In CMIP5 there is a single exception to use of the above template. For so-called gridspec files, which describe the grids used in a model, the filename should be constructed as follows:

gridspec filename = <*variable name*>_<*modeling realm*>_<*MIP table*>_<*model*>_
<*experiment*>_<*ensemble member*>.nc

where the elements are the same as above, except the <*modeling realm*> is now included (and the <temporal subset> is omitted because gridspec information is time-independent).

Example:

- gridspec_atmos_fx_IPSL-CM5_historical_r0i0p0.nc

3.3 ESGF data node directory structure

It is recommended that ESGF data nodes should layout datasets on disk mapping DRS components to directories as:

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

Example:

/CMIP5/output1/UKMO/HadCM3/decadal1990/day/atmos/day/r3i2p1/v20100105/tas/
tas_day_HADCM3_ decadal1990_r3i2p1_199001-199012.nc

Deleted: variable

Formatted: Font: Times New Roman, 12 pt

Deleted: >_

Deleted: *variable*

Deleted: from the atomic dataset definition

Deleted: Note that together, <*variable name*> and <*variable table*> constitute the "variable identifier", which uniquely defines the variable.¶

Formatted: Font: Not Italic

3.4 Publication-level *dataset_id encoding*

Formatted: Keep with next

Publication-level datasets are assigned an identifier *dataset_id* within THREDDS catalogs on ESGF data nodes. The CMIP5 best practices document[4] defines a Publication-level *dataset_id* as:

<activity>.<product>.<institute>.<model>.<experiment>.<frequency>.<modeling realm>.<MIP table>.<ensemble member>

Each publication-level dataset version will have the THREDDS id:

<activity>.<product>.<institute>.<model>.<experiment>.<frequency>.<modeling realm>.<MIP table>.<ensemble member>.<version>

Note that the version number assigned to the dataset by ESG is supposed to reflect the date of ESG publication, but the version will usually be assigned by the user so this cannot generally be guaranteed. The user will be instructed to provide ESG with the date that appears in the ESGF data node directory structure for the dataset being published. In many cases the directory structure will be generated some days prior to publication, so the date will not in fact reflect the date of publication, but the date that the directory structure was created.

3.5 URL syntax

URLs referencing the data files will have a site dependent prefix followed by the DRS directory structure.

[4] See CMIP5 Best Practices for Data Publication: http://esg-pcmdi.llnl.gov/internal/esg-data-node-documentation/cmip5-best-practices

Appendix: Controlled Vocabularies

Appendix 1.1 Experiment Controlled Vocabulary

Coupled Model "Decadal" Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
1.1, 1.2 & 1.5	decadalXXXX*	10- or 30-year run initialized in year XXXX*	decadal hindcasts/predictions, some extended to 30 years	10-30	≥3*≥10
1.3	noVolcXXXX*	volcano-free hindcasts	hindcasts but without volcanoes	10-30	≥3
1.4	volcIn2010	prediction with 2010 volcano	Pinatubo-like eruption imposed in year 2010	10-30	≥3
1.6	**	chemistry-focused runs	near-term runs with enhanced chemistry/aerosol models	10-30	1

* Replace 'XXXX' with the year in which the decadal prediction was initiated (typically near the end of year XXXX). As an example, a simulation focusing on the 10-year period from January 1966 through December 1975 will typically be initiated sometime between September 1, 1965 and January 1, 1966. All such simjulations would be labeled decadal1965.

** These runs will be placed in the appropriate directory (defined by the experiment; 1.1-1.5); Experiment 1.6 differs from the others only because a different model is used, which will be indicated by a unique model name, so there is no need for a new directory for these runs.

Deleted: If

Deleted: run is initiated

Deleted: , then XXXX should

Deleted: the year immediately preceding the date of initialization

Coupled Model Long-Term Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
3.1	piControl	pre-industrial control	coupled atmosphere/ocean pre-industrial control run	≥500	1
3.2	historical	historical	simulation of recent past (1850-2005)	156	≥1
3.4	midHolocene	mid-Holocene	consistent with PMIP, impose Mid-Holocene conditions	100	1
3.5	lgm	last glacial maximum	consistent with PMIP, impose last glacial maximum conditions	100	1
3.6	past1000	last millennium	consistent with PMIP, impose forcing for 850-1850	1000	1
4.1	rcp45	RCP4.5	future projection (2006-2300) forced by RCP4.5	95-295	1
4.2	rcp85	RCP8.5	future projection (2006-2300) forced by RCP8.5	95-295	1
4.3	rcp26	RCP2.6	future projection (2006-2300) forced by RCP2.6	95-295	1
4.4	rcp60	RCP6	future projection (2006-2100) forced by RCP6	95	1
5.1	esmControl	ESM pre-industrial control	as in experiment 3.1, but emissions-forced (with atmospheric CO2 determined by model)	250	1
5.2	esmHistorical	ESM historical	as in experiment 3.2, but emissions-forced (with atmospheric CO2 determined by model)	156	1
5.3	esmrcp85	ESM RCP8.5	as in experiment 4.2, but emissions-forced (with atmospheric CO2 determined by model)	95	1

5.4-1	esmFixClim1	ESM fixed climate 1	radiation code "sees" control CO2, but carbon cycle sees 1%/yr rise	140	1
5.4-2	esmFixClim2	ESM fixed climate 2	radiation code "sees" control CO2, but carbon cycle sees historical followed by RCP4.5 rise in CO2	251	1
5.5-1	esmFdbk1	ESM feedback 1	carbon cycle "sees" control CO2, but radiation sees 1%/yr rise	140	1
5.5-2	esmFdbk2	ESM feedback 2	carbon cycle "sees" control CO2, but radiation sees historical followed by RCP4.5 rise in CO2	251	1
6.1	1pctCO2	1 percent per year CO2	imposed 1%/yr increase in CO2 to quadrupling	140	1
6.3	abrupt4xCO2	abrupt 4XCO2	impose an instantaneous quadrupling of CO2, then hold fixed	150	≥1
7.1	historicalNat	natural-only	historical simulation but with natural forcing only	156	≥1
7.2	historicalGHG	GHG-only	historical simulation but with greenhouse gas forcing only	156	≥1
7.3	historicalMisc	other-only	historical simulation but with other individual forcing agents or combinations of forcings.	156	≥1

* The forcing in these runs will be specified in a global attribute in the file, relying in so far as possible on the controlled vocabulary abbreviations defined in Appendix 1.2. When more than one run of this type is done, the runs will be distinguishable by being assigned different "perturbed physics" numbers, p<L>, as described under "ensemble member" in this document.

Formatted: Space Before: 12 pt

Deleted: ?**

Deleted: ?*

Deleted: "?**" should

Deleted: replaced with

Deleted: two- or three-letter character string, which will uniquely identify

Deleted: individual forcing agent that is active. Choose strings from

Deleted: given

Atmosphere-Only Simulations

Experiment number	Short Name of Experiment	Experiment Name	Experiment Description	Years requested per run	Ensemble size requested
3.3	amip	AMIP	AMIP (1979- at least 2008)	≥30	≥1
2.1	sst2030	2030 time-slice	conditions for 2026-2035 imposed	10	≥1
6.2a	sstClim	control SST climatology	control run climatological SSTs & sea ice imposed	30	1
6.2b	sstClim4xCO2	CO2 forcing	as in experiment 6.2a, but with 4XCO2 imposed	30	1
6.4a	sstClimAerosol	all aerosol forcing	as in experiment 6.2a, but with aerosols from year 2000 of experiment 3.2	30	1
6.4b	sstClimSulfate	sulfate aerosol forcing	as in experiment 6.2a, but with sulfate aerosols from year 2000 of experiment 3.2	30	1
6.5	amip4xCO2	4xCO2 AMIP	AMIP (1979-2008) conditions (experiment 3.3) but with 4xCO2	30	1
6.6	amipFuture	AMIP plus patterned anomaly	consistent with CFMIP, patterned SST anomalies added to AMIP conditions (experiment 3.3)	30	1
6.7a	aquaControl	aqua planet control	consistent with CFMIP, zonally uniform SSTs for ocean-covered earth	5	1
6.7b	aqua4xCO2	4xCO2 aqua planet	as in experiment 6.7a, but with 4XCO2	5	1
6.7c	aqua4K	aqua planet plus 4K anomaly	as in experiment 6.7a, but with a uniform 4K increase in SST	5	1
6.8	amip4K	AMIP plus 4K anomaly	as in experiment 3.3, but with a uniform 4K increase in SST	30	1

Appendix 1.2 Controlled Vocabulary for Abbreviated "Forcing" Descriptors

The abbreviations in this table can be used to describe the different externally imposed forcing agents that are active in a given simulation. A forcing agent will show some secular variation due to prescribed changes in concentration or emissions (or in the case of land-use, and change in prescription of surface conditions). Sometimes the change will be due to emissions of a precursor species that relatively quickly becomes transformed into the forcing agent itself (e.g., transformation of SO2 emissions to sulfate aerosols).

Abbrev.	Forcing Description	Abbrev.	Forcing Description
Nat	natural forcing (a combination, not explicitly defined here, that might include, for example, solar and volcanic)	LU	land-use change
Ant	anthropogenic forcing (a mixture, not explicitly defined here, that might include, for example, well-mixed greenhouse gases, aerosols, ozone, and land-use changes).	Sl	solar irradiance (note: Sl is "S" followed by a lower case "L", not an upper case "I")
GHG	well-mixed greenhouse gases (a mixture, not explicitly defined here)	Vl	volcanic aerosol (note: Vl is "V" followed by a lower case "L", not an upper case "I")
SD	anthropogenic sulfate aerosol, accounting only for direct effects	SS	sea salt
SI	anthropogenic sulfate aerosol, accounting only for indirect effects	Ds	Dust
SA (= SD + SI)	anthropogenic sulfate aerosol direct and indirect effects	BC	black carbon
TO	tropospheric ozone	MD	mineral dust
SO	stratospheric ozone	OC	organic carbon
Oz (= TO + SO)	ozone (= tropospheric and stratospheric ozone)	AA	anthropogenic aerosols (a mixture of aerosols, not explicitly defined here)

Here are three use cases for the DRS syntax: in URLs, for a directory layout, and in filenames.

3.1 URL syntax.

When the DRS is used in a URL, we would expect the URL to comprise a hostname, the atomic dataset name, possibly an extended path name, and possibly a service endpoint name. That is, we would expect to see usage like:

http://<hostname>/<activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable identifier>/<ensemble member>/<version>/ [<endpoint>],

where square brackets enclose optional elements (in this case, only the service endpoint).

Where no service endpoint appears, it should be expected that an HTTP GET on the URL would return the netCDF data. (Currently there is no CMIP5 controlled vocabulary for endpoints, when one appears it will have values which encompass services such as OPeNDAP and WCS etc.)

Note that ensemble member and version numbers are mandatory, to ensure that if subsequent versions or ensemble members appear, there is no possibility of ambiguity as to what data is referenced at a given URL.

Should replace the following with "real" examples

http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/varname/r1/v1/

or

http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/varname/r1/v1/extended_path/

or

http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/varname/r1/v1/extended_path/service_endpoint

or

http://badc.nerc.ac.uk/activity/product/institute/model/experiment/frequency/realm/varname/r1/v1/service_endpoint

Controlling the vocabulary for service endpoints is beyond the scope of this document, but will be a necessary part of the distributed URL design, and impact on what appears in catalogues.

However, we might expect that without a service endpoint, dereferencing these URLs will return either netCDF data, or catalogue entries. (Examples of service endpoints, might be: las, opendap, wcs, wms, wfs etc).

(Note that "hostnames" will probably be intuitional virtual hostnames, rather than individual system names, but either way, will need to be present in catalogues).

BNL Note: actually, once one starts considering service endpoints there is a strong argument that the variable identifier should be after the realization and version numbers, allowing one to construct service endpoints which serve multiple variables.

3.2 Directory Layout

For CMIP5, certain software will assume a directory layout as follows:

<activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable name>/<ensemble member>/<version_number>/