Chemistry-Climate Model Initiative (CCMI) Model Output Requirements and Data Reference Syntax

David Plummer, Jean-Francois Lamarque and Michaela Hegglin

Version 2.2a – March 28, 2014

Overview

The two phases of the Stratosphere-troposphere Processes And their Role in Climate (SPARC) Chemistry-Climate Model Validation (CCMVal) activity showed the tremendous value of coordinated model experiments and analysis for stratospheric chemistry-climate modelling. The goal of the first phase of the Chemistry-Climate Model Initiative (CCMI) is to further advance these efforts with a greater focus on tropospheric processes and the links between the troposphere and stratosphere. CCMI is planning to participate as a formally approved Model Intercomparison Project (MIP) under the umbrella of CMIP6 – a so-called 'sub-MIP'. In preparation for this greater integration with the larger efforts of the coupled-model community, data collection and archiving is moving towards the infrastructure used for CMIP5. Model output will be in netCDF files that conform to the 'Climate and Forecast' (CF) convention. In addition, filenames and directory structures will follow that used for CMIP5.

To facilitate the generation of the output, the organizers strongly suggest that individual modelling groups use CMOR2 to re-write model output from the native format used at their modelling centre to the CF-compliant netCDF format. A set of 'MIP tables' have been created that provide a large variety of the required metadata, tailored to the CCMI data request, which, when used with CMOR2 should greatly assist the generation of properly formatted netCDF files from the model output.

The **Model Output Requirements** (required metadata written into the netCDF 'global_attributes', as well as requirements for coordinate and output variables) is described below. The requirements closely follow that used for CMIP5 and the CCMI document closely follows the document describing the CMIP5 requirements written by Taylor and Doutriaux 'CMIP5 Model Output Requirements: File Contents and Format, Data Structure and Metadata' available at

http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf

Further below, we give an overview of the **Data Reference Syntax** (DRS) used to construct filenames and the data archive structure from strictly defined components. The DRS description adapts the CMIP5 DRS for use by the Chemistry-Climate Model Initiative (CCMI) project. We borrow heavily from the original CMIP5 DRS document by Karl Taylor et al. available at

http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

We have provided a brief overview of the required information. For additional information or clarifications, please see the original CMIP5 documents.

Note that if using CMOR2 to re-write model output, there is a link between the netCDF global_attributes, the variable attributes taken from the MIP tables and the DRS components used to construct the filenames and directory structure. We have attempted to outline these connections below.

Data will be uploaded to a holding area at the BADC and a series of tests will be performed on the construction of the filenames and the metadata in the netCDF file to ensure the uploaded files conform, at a minimum level, to the requirements outlined here. If the newly uploaded files pass these tests, they

will be 'published' (made available for download) through the Earth System Grid Federation (ESGF).

A) CCMI Model Output Requirements

Data format, data structure and file content requirements

Data must be written in netCDF-4 format using compression and conform to the CF metadata standards. To take advantage of data compression available under netCDF-4, while also ensuring the greatest degree of compatibility of the resulting files with existing software, the netCDF-4 files must be written in 'classic model' format, so called 'netCDF-4 classic'. This is achieved by setting the creation mode flag to 'NC_CLASSIC_MODEL'.

Each file must contain only a single output field from a single simulation. In addition, each file must contain coordinate variables, attributes and other metadata as specified below. Data representing a long time-series will usually be split into several files, which should neither be too large (to be unwieldy) not too small (as to create vexing I/O performance issues). It is recommended that the same size chunks be used for all variables found in the same table having the same spatial extent and sampling frequency. For example, monthly-average data requested for three-dimensional data may be put into multi-decadal chunks, while monthly-average 2-D fields (either lat-lon or zonal-average on constant pressure surfaces) may be put into individual files with much longer time spans.

Required Global Attributes

The netCDF files contains a significant amount of metadata that allows each file to be completely self-describing. Given below are the required global attributes that must be written to each netCDF file. Given further below is a description of the Data Reference Syntax (DRS) that provides the rules for constructing filenames and the structure of the data repository. If using CMOR to create the netCDF files, there is a direct link between many of the global attributes and the DRS components used to construct filenames and directory structures and this link is given, along with the complete list of required global_attributes in Appendix 2.

branch_time = time in parent experiment when this simulation started. Applicable for models coupled to an ocean that branch the transient simulation from a long control run. Should also be used for the case of an RCP projection simulation that started from a different historical simulation. The branch_time should be relative to the basetime of the parent. If the run did not start from another run, from initialization for a run using specified SSTs/sea-ice, for example, the branch_time should be set to 0.0.

contact = name and contact information (e.g. e-mail address, phone number) of person who should be contacted for information about the data.

Conventions = CF-1.6, taken from the MIP table.

creation_date = a string representation of the date when the file was created in the format: '*YYYY-MM-DD*-T*HH*:*MM*:*SS*Z'. The 'T' and 'Z' are not modified, but the other variables are replaced with the correct time stamp. If using CMOR, this is generated automatically.

experiment = a string specifying the long title of the experiment and found in the headers of the MIP tables, such as 'Projection' or 'Projection Scenario – RCP8.5'. See Appendix 1 below for a list.

experiment_id = the short name of the experiment found in the headers of the MIP tables, such as 'refC2' or 'senC2rcp85'. See Appendix 1 below for a list.

forcing = a string containing the list of the 'forcing' agents that should cause the climate to change.

This follows CMIP5, so if running a coupled-model include the forcings following the CMIP5 specification. Otherwise, set to 'N/A'.

frequency = a string indicating the interval between individual time samples. These are set by CMOR from the MIP tables depending on the variable requested and may take on the following values: 'yr', 'mon', 'day', 'subhr' or 'fx' (the last for time-independent fields).

initialization_method = an integer (≥ 1) referring to the initialization method or observational dataset used to initialize the simulation. Typically if only a single method was used, use a value of 1. The initialization_method is used to assign 'M' of 'r<N>i<M>p<L>' of the experiment rip and the DRS component 'ensemble member'. For invariant fields ('fx ') use a value of 0. See the discussion of the DRS component 'ensemble member' below for more information.

initialization_description = a string containing the details of the particular initialization method used for the experiment. This is only required if groups are submitting runs with more than one initialization method.

institute_id = a short acronym describing the institution (e.g., 'GFDL'). The 'institute_id' is used by CMOR to set the 'Institute' DRS component and is, therefore, subject to the DRS character restrictions. The 'institute_id' must be agreed with the CCMI organizers. A list of the accepted institute_id strings can be found in the **CCMI_codes.txt** file available at http://www.met.reading.ac.uk/ccmi/?page_id=245.

institution = a more free-form attribute to identify the institution that generated the data. e.g. 'GFDL (Geophysical Fluid Dynamics Laboratory), Princeton, NJ, USA'

model_id = like the institute_id, a short acronym identifying the model used to generate the output that must be agreed upon with the CCMI organizers. The 'model_id' is used to set the 'Model' component of the DRS and is, therefore, subject to the DRS restrictions on which characters may be used. A list of the accepted model_id strings can be found in the **CCMI_codes.txt** file available at http://www.met.reading.ac.uk/ccmi/?page_id=245.

modeling_realm = denotes which high-level modeling component is of particular relevence for the dataset. Following CMIP5, permitted values are 'atmos', 'ocean', 'land', 'landIce', 'seaIce', 'aerosol' 'atmosChem' and 'ocnBgchem'. These are specified for each variable in the MIP table and the data request spreadsheet. For CCMI1, we have kept chemical fields as 'atmos' to simplify the resulting file structure of the database. If using CMOR, this attribute is read from the MIP table and used to set the 'modeling realm' DRS component.

parent_experiment_id = indicates which experiment this simulation branched from. This should match the experiment_id of the parent experiment, unless the parent is irrelevant, in which case the parent_experiment_id should be set to 'N/A'. For example, if branching an RCP8.5 simulation (senC2rcp85) from a preexisting REF-C2 simulation, parent_experiment_id='refC2'.

parent_experiment_rip = identifies which member of an ensemble of parent simulations, this simulation branched from. If parent_experiment_id is 'N/A' then parent_experiment_rip should also be set to 'N/A'. The 'rip' is the same series of numbers as used for the DRS component 'Ensemble member' (see below) and takes the form 'r<N>i<M>p<L>', where these numbers denote 'realization', 'initialization method' and 'physics version' respectively. When possible, and not inappropriate, the child experiment should inherit the 'rip' value from the parent.

physics_version = an integer (≥ 1) referring to the physics version used by the model for this run. The physics_version is used to assign 'L' of 'r<N>i<M>p<L>' of the experiment rip and the DRS component 'ensemble member'. For invarient fields ('fx') use physics_version=0. See the discussion of the DRS component 'ensemble member' below for more information.

physics_description = a string describing the particular variations in the model that produced the

current physics versions referred to by the 'physics_version' attribute. Only required if groups are submitting results from more than one physics version.

product = 'output', which indicates that the data you are writing is model output.

project_id = 'CCMI1'

realization = an integer (≥ 1) distinguishing among the members of an ensemble of simulations (e.g. 1, 2, 3,...). If only a single simulation was performed, then it is recommended that realization=1. The physics_version is used to assign 'N' of 'r<N>i<M>p<L>' of the experiment rip and the DRS component 'ensemble member'. For invarient fields ('fx') use realization=0. See the description for the DRS component 'ensemble member' below for more information.

source = character string fully identifying the model and version used to generate the output. The first portion of the string should be a copy of the global attribute 'model_id'. Additionally, this attribute must include year (i.e. model vintage) when this model version was first used in a scientific application. Finally, it should include information concerning the component models. The following template should be followed in constructing this string: '<model id> <year> atmosphere: <base_model_name> (<technical name>, <resolution_and_levels>); ocean: <model name> (<technical name>, <resolution_and_levels>); sea ice: <model_name> (<technical_name>); land: <model_name> (<technical_name>)'. Depending on the scope of the model, not all of these components are necessarily applicable. An example might be

'CMAM 2008 atmosphere: CGCM3.1 (GCM13E, T31L71) ocean: NCOM3.1 land: CLASS2.7'

table_id = should be assigned a character string that identifies the CCMI MIP table where this variable appears. The string is of the form 'Table <table name> (<date of table>). If using CMOR, the values of these variables will be pulled from the MIP table. An example would be 'Table monthly (1 February 2014)'

tracking_id = a string that is almost certainly unique to this file and must be generated using the OSSP utility which supports a number of different DCE 1.1 variant UUID options. For CCMI, as for CMIP5, version 4 (random number based) is required. The software can be downloaded from http://www.ossp.org/pkg/lib/uuid/.

Optional global attributes:

comment = a character string containing additional information about the data or the methods used to produce it.

history = a character string containing an audit trail of modifications to the original data. Each modification is typically proceeded by a timestamp. The history attribute provided here will be a global one that should not depend on wich variable is contained in the file.

references = a character string containing a list of published or web-based references that describe the data or the methods used to produce it. Typically, the user should provide references describing the model formalation here.

title = a character string of the form '<model_id> model output prepared for CCMI1 <experiment>'.

Requirements for coordinate variables

All coordinate variables must be written as double precision floating point (netCDF type NC_DOUBLE).

The time specification mandated by the CF standards is 'days since <basetime>'. The <basetime> variable must be specified by the user, typically in the form year-month-day as in 'days since 1950-01-01'.

- the same <basetime> should apply to all time samples in a single simulation
- for simulations meant to represent a particular historical period, set the <basetime> to the time at the beginning of the simulation. If starting a REF-C2 run on January 1st, 1950, set <basetime> to '1950-01-01', for example
- for future scenarios that branch from the CCMI historical run, retain the same <basetime> as used in the historical part of the run. For example, if starting a RCP8.5 scenario January 1st, 2005 from an existing REF-C2 run, keep the <basetime> from the REF-C2 run.

In addition to the required, CF-compliant time specification we ask groups to include a 'human readable' time as an 'auxiliary coordinate variable' in the netCDF files. This will be represented by five one-dimensional integer arrays with the same length as the time coordinate. The five arrays are for year (ccmi_year), month (ccmi_month), day of month (ccmi_day), hour (ccmi_hour) and minute (ccmi_minute) according to the Gregorian calendar at GMT. In conjunction with the five additional one-dimensional fields, the 'coordinates' attribute of the output variable must be set to

coordinates="ccmi_year ccmi_month ccmi_day ccmi_hour ccmi_minute"

For time-mean data, the time coordinate value and the CCMI timestamp should be defined as the mid-point of the interval over which the average is defined.

Much of the three-dimensional data is requested on the model atmospheric levels, and users must include in the file all the information needed to positively and uniquely indicate the vertical location of the data. Usually in this case, the 'formula_terms' attribute must be defined and additional variables or parameters will need to be stored in the file. For example, for an atmospheric-sigma coordinate system the surface pressure field and the pressure at the top of the model would need to be included. For monthly-average fields on model levels, in the case of an atmospheric-sigma coordinate, monthly-average surface pressure would need to be included. See Appendix D of the CF-conventions document at:

http://cf-pcmdi.llnl.gov/documents/cf-conventions/latest-cf-conventions-document-1

Several fields are requested that have a coordinate that is single-valued. These include daily surface air temperature at 2 metres, the monthly-average eddy heat flux at 100 hPa or the aerosol optical thickness at 550 nm. These fields have an associated scalar coordinate assigned in the MIP table and must have the 'coordinates' attribute of the output field defined to include the name of the scalar coordinate (e.g. 'height2m' or 'wv550nm').

Additional requirements for coordinates follow the CMIP5 specifications including:

axis – set to 'X, 'Y', 'Z' or 'T', as appropriate, for the coordinate variables that describe these dimensions

bounds – a character string containing the name of the variable where the cell bounds are stored, if required for a particular coordinate variable.

calendar – for the time coordinate only, following one of the options described in Section 4.4.1 of the CF-conventions

formula_terms – for dimensionless vertical coordinates only, a character string as described in Section 4.3.2 of the CF-conventions.

positive – set to 'down' or 'up' as appropriate

standard_name – a character string containing the standard name of the dimension variables. Note that many of the requested fields do not have accepted CF standard names, in which case the

standard_name attribute must be omitted.

units – a character string containing the units of the dimension variable. For dimensionless vertical coordinates, this attribute may be omitted or set to '1'.

Requirements for output variables

Almost all variables will be written as single precision floating point (netCDF type NC_FLOAT).

The variable names must be as given under 'Output variable name' in the data request spreadsheet or, identically, the 'variable_entry' in the MIP tables.

The units must be as given under 'Units' in the data request spreadsheet, or the 'units' entry in the MIP table.

The positive direction of vertical fluxes must be consistent with that given under 'Positive Flux Dir.' in the data request spreadsheet, if applicable, or the 'positive' entry in the MIP table.

For Cartesian latitude-longitude grids, data must be ordered with longitude increasing from west to east, starting with the first grid point greater than or equal to 0 degrees east. Latitude must be ordered with latitude increasing from south to north.

If there is a vertical coordinate, data must be stored with data starting with the grid point nearest the surface.

If there is a time dimension, data must be stored with time increasing.

All missing data must be assigned the single precision value of 1.0E+20.

Additional requirements for output variable attributes, largely following CMIP5:

associated_files – a string listing the base URL for CCMI, the location of the models gridspec file and, as appropriate, the name of the file containing the grid cell areas. For CCMI, this string is given by 'baseURL:http://www.met.reading.ac.uk/ccmi/ gridspecFile:<gridspec file name> [areacella: <atmospheric cell area file name>]. With the exception of the monthly-mean 1-D and annual-mean 0-D fields, all variables will require an 'areacella' specification. These cell areas should be defined such that the exact global integrals of a chemical quantity, such as the global burden or an emission or deposition flux, could be calculated.

cell_measures – with the exception of the monthly-mean 1-D and annual-mean 0-D fields, this will be given a string given by 'area: areacella'.

cell_methods – the string that appears under the 'Cell Method' column of the data request spreadsheet or the 'cell_methods' entry in the MIP table.

coordinates – a blank-separated list of auxiliary coordinate or coordinate variables. See the discussion above of 'coordinates' for specifying the auxiliary CCMI time coordinate and the case of variables requested at a scalar coordinate value.

_FillValue and missing_value – assigned a value of 1.0E+20 for grid points with 'missing' data. Note that defining both _FillValue and missing_value are required since different software is expecting one or the other of these to be set.

long_name – a character string for each variable given by the 'Long_name' column of the data request spreadsheet or the 'long_name' entry in the MIP tables. This should always be assigned, in particular because many of the variables requested do not have accepted CF-compliant standard names, in which case the long_name is used in place.

standard_name – the CF standard name as accepted under version 26 of the CF Standard Names

table, if a name has been accepted.

units – a character string denoting the units for the variable, given under the 'Units' column of the data request spreadsheet or the 'units' entry of the MIP tables.

B) CCMI Data Reference Syntax (DRS) and Controlled Vocabularies

This document adapts the CMIP5 DRS for use by the Chemistry-Climate Model Initiative (CCMI) project. We borrow heavily from the original CMIP5 DRS document by Karl Taylor et al. available at

http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

We have provided a brief overview of the required information. For additional information or clarifications, please see the original CMIP5 document.

Purpose

The Data Reference Syntax (DRS) provides the conventions to create directory structures and file names within the data archive, providing users and software with a predictable way of finding data.

Definitions:

Atomic dataset: 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, modelling realm, variable name, MIP table, ensemble member and version number.

Publication-level dataset: For CMIP5, a publication-level dataset was defined as a set of atomic datasets differing only by variable name. In a departure from CMIP5, we define a publication-level dataset identically to that of an atomic dataset. This has implications for the assignment of a version number to submitted data (see below), as version control is applied on publication-level datasets.

Component Definitions and Controlled Vocabularies

These are the individual components that are combined to produce file names and directory structures. Many of these descend directly from the netCDF 'global_attributes' when using CMOR.

Activity identifies the model intercomparison activity or other data collection activity. For CMIP5 this is 'CMIP5', here we use 'CCMI1'.

Product, following CMIP5, currently has four options: 'output', 'output1', 'output2' and 'unsolicited'. Files will be initially designated as 'output' or 'unsolicited'. Subsequently, data from the requested variable list will be assigned a version (see below) and placed in either 'output1' or 'output2'. Variables not specifically requested by CCMI will remain in 'unsolicited'. In some cases a continuous sequence of model data will be split between 'output1' and 'output2' to facilitate archive management. Note that although output of some variables is requested only for limited time-periods, if output of those variables is made available for other time periods, it will be placed into 'output2', not 'unsolicited'.

Institute identifies the institute responsible for the model results (e.g. UKMO), and it should be as short as possible. The DRS 'institute' component should be identical to the netCDF global attribute 'institute_id'. If using CMOR, this will be automatic. There are restrictions on which characters are

included in the DRS components, so both 'institute' and the netCDF global attribute 'institute_id' are subject to these restrictions (see Permitted Characters below).

Model identifies the model used (e.g. MOCAGE or UMSLIMCAT). In a similar manner to 'institute', the DRS component 'model' will be set from the corresponding netCDF global attribute 'model_id' by CMOR. If not using CMOR, ensure these two variables are identical. 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 CCM coupled to an ocean performs the REF-C1 simulation with specified SSTs/sea-ice then the name may not necessarily change.

> NOTE: The 'institute' and 'model' (along with the corresponding netCDF global attributes 'institute_id' and 'model_id') must be the accepted forms agreed to by all within CCMI. The full list is available at:
>
> http://www.met.reading.ac.uk/ccmi/?page_id=245

Experiment identifies the assigned experiment name and, due to restrictions on characters and length, are not necessarily identical to the names given in the CCMI Community Simulations document. The list of experiment names is given in Appendix 1 and can also be found in the headers of the CCMI CMOR tables.

Frequency indicates the interval between individual time samples. For CCMI the following are the only options: 'yr', 'mon', 'day', 'hr', 'subhr' and 'fx' (fixed, i.e. time-independent fields like topography). For each variable requested there is an associated 'frequency' and these are grouped together in the CMOR tables.

Modeling Realm indicates which high-level modeling component is of particular relevance for the dataset. Following CMIP5, permitted values are 'atmos', 'ocean', 'land', 'landIce', 'seaIce', 'aerosol' 'atmosChem' and 'ocnBgchem'. These are specified for each variable in the CMOR table and the data request spreadsheet. For CCMI, we have kept chemical fields as 'atmos' to simplify the resulting file structure of the database.

Variable name in the DRS comes from the 'output variable name' in the CCMI data request and also corresponds to the 'variable_entry' assigned in the MIP tables.

MIP table is the name assigned to the individual MIP tables, with each MIP table corresponding to fields sampled at a single frequency or sampled in a particular fashion (satellite sampling) and traces back to the different sheets in the CCMI data request spreadsheet.

> NOTE: The combination of 'variable name' and 'MIP table' will uniquely identify the physical quantity and often implies something about the sampling frequency and modeling realm. For example, the three-dimensional ozone volume mixing ratio is given the same 'variable name' (vmro3) across multiple MIP tables, but will differ in sampling frequency and averaging between the 'daily' and 'monthly' MIP tables.

Ensemble member differentiates closely related simulations by a single model and is required even if only a single simulation is performed. The 'ensemble member' is coded as a triad of integers (N, M and L) in the form 'r<N>i<M>p<L>'.

It is recommended that all three numbers are assigned sequentially starting with 1. Time independent variables (.e. those with frequency='fx') are not expected to differ across ensemble members, so for these fields the three numbers should be assigned the value zero (ie. 'r0i0p0').
The first of the three numbers is the 'realization' number (a positive integer value of 'N') and is used to distinguish among members of an ensemble typically generated by initializing a set of runs with different, but equally realistic, initial conditions.

Models used for forecasts that depend on 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>). Since CCMI does not specify experiments to look at predictability, it is expected that 'i1' will be used across all atomic datasets.

The last of the three numbers is the 'perturbed physics' number (p<L>) and has been implemented to distinguish between closely related model versions that run a perturbed physics ensemble (e.g. climateprediction.net). CMIP5 also used the 'perturbed physics' number to differentiate between different sets of 'forcings' in Experiment 7.3 - Other individual forcing runs, where groups were given latitude to decide which sets of forcings they used for attribution studies. For CCMI, there are no perturbed physics experiments specified and forcings are generally uniquely defined by the experiment, therefore 'p1' will be generally used. One possible exception may be groups that run more than one version of the SEN-C1-Emis experiment using more than one alternate set of emissions. It is also possible groups would want to differentiate between two sets of simulations using a different version of physics or chemical mechanisms, for example. In this case, these sets of simulations could be distinguished by either a distinct model name, if warranted, or a different value of L. If there are instances of L>1, then a table should be made available that associates each value of L with a particular model setup, detailing the differences in physical and chemical parameterizations, parameter values or forcings.

Version number is given by 'v' followed by an integer (i.e. vN) and is used to uniquely identify a particular version of a publication-level dataset, for example distinguishing between an original version of output that was later found to be flawed and replaced with a corrected version. For CMIP5 the version number was based on the date of publication (e.g. v20100105 for a version provided on January 5th, 2010). For CCMI we propose to use an identical version number assignment, with the difference that the CCMI definition of a publication level dataset is identical to that of an atomic dataset.

The implication for users is that if a single file must be replaced due to errors, then all files with the same DRS components activity, product, institute, model, experiment, data sampling frequency, modelling realm, variable name, MIP table and ensemble member must also be given the updated version number.

Extended Path

Note that thus far we have not considered datasets that contain spatio-temporal subsets or means. For example, it may not be practical to include the entire time series record for a single variable from a particular experiment (a single atomic dataset) in a single netCDF file. The DRS supports the specification of such subsets or means, however these represent or are derived from only 'parts' of an atomic dataset and hence they were not included in the definition of an atomic dataset above.

Temporal subsets or means: Time instants and periods (N1-N2)

Time instants or periods will be represented by a construction of the form 'N1-N2', where N1 and N2 are of the form 'yyyy[MM[dd[hh[mm[ss]]]]-suffix], where 'yyyy', 'MM', 'dd', 'hh', 'mm' and 'ss' are integers specifying year, month, day, hour, minute and second, respectively. Everything after 'yyyy' is optional, but must include a minimum of precision to unambiguously resolve the interval between time samples contained in the file. For example, monthly mean data would normally include 'yyyy' and 'MM', but not 'dd', 'hh', 'mm' or 'ss'. If a single time instant is included in the dataset, CCMI requires, as did CMIP5, that both N1 and N2 are specified and assigned the same value.

The optional '-suffix' can be included to indicate that the netCDF file contains a climatology (suffix = '-clim') or a single time mean, for example, over multiple years (suffix = '-avg'). For example, a file with sampling frequency 'mon' and the time designation 196001-198912-clim represents the mean monthly climatology (12 time values) computed for the period extending from 1960-01 to 1989-12.

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 puntuated 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 '-'. In constructing the 'variable name' component of the DRS it is recommended that the '-' be avoided since hyphens cannot be imbedded in Fortran and IDL variable names and some users would like to maintain consistency between the DRS name and the name appearing in their code.

Using the DRS Syntax

The DRS syntax described above is important because it provides the ingredients to construct the structure of the repository and the netCDF filenames. These components must be strictly constructed and predictable to allow data to be found, and avoid characters that are ambiguous for system commands. We hope to have CCMI1 data available over the Earth System Grid Federation (ESGF), therefore the strict adherence to the DRS is required for use by the ESGF infrastructure.

CMOR directory structure

The output tool CMOR2, recommended for use in converting model output to the CF-compliant netCDF files for the CCMI data archive, can optionally write output files to a directory structure derived from DRS components as:

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

For example:

/CCMI-1/output/ETH-PMOD/SOCOL3/refC2/mon/atmos/vmro3/r1i1p1/

This structure is incompatible with the DRS directory structure used on the ESGF data nodes for CMIP5, however it remains relevant as a possible structure for model output for individual groups when creating their data prior to transferring it to BADC.

ESGF data node directory structure

We propose to follow the relationship between DRS components and the directory structure used for

CMIP5 on the ESGF data nodes for the CCMI data archive at BADC:

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

For example:

/CCMI-1/output1/ETH-PMOD/SOCOL3/refC2/mon/atmos/monthly/r1i1p1/v1/vmro3/
vmro3_monthly_SOCOL3_refC2_r1i1p1_200001-201012.nc

Filename encoding

Following the construction of filenames used in CMIP5, the filename will be constructed as:

<variable name>_<MIP table>_<model>_<experiment>_
<ensemble member>[_<temporal subset>][_<geographical info>].nc

The <temporal subset> (along with the preceding underscore) is omitted for variables that are time-independent and the geographical information (along with the preceding underscore) is included only when needed. Note that <geographical_info> is not used for CCMI.

For example:

vmro3_monthly_SOCOL3_refC2_r1i1p1_196001-200912.nc

Following CMIP5, there is a single exception to the above. For 'gridspec' files, which describe the grids used in a model, the filename should be constructed as follows:

gridspec_<modeling realm>_fx_<model>_<experiment>_r0i0p0.nc

For example:

gridspec_atmos_fx_SOCOL3_refC2_r0i0p0.nc

Appendix 1 – Experiment Names

Simulation name used in CCMI Community Simulations Document	Experiment name recorded in the netCDF global attribute 'experiment'	Short name of experiment used in the DRS and recorded in the netCDF global attibute 'experiment_id'
REF-C1	Hindcast	refC1
REF-C1SD	Hindcast – Specified Dynamics	refC1SD
REF-C2	Projection	refC2
SEN-C1-Emis	Hindcast Scenario - altEmis	senC1Emis
SEN-C1SD-Emis	HindcastSD Scenario - atlEmis	senC1SDEmis
SEN-C1-fEmis	Hindcast Scenario- fxdEmis	senC1fEmis
SEN-C1SD-fEmis	HindcastSD Scenario - fxdEmis	senC1SDfEmis
SEN-C1-SSI	Hindcast Scenario - altSSI	senC1SSI
SEN-C2-RCP	Projection Scenario - RCP2.6	senC2rcp26
SEN-C2-RCP	Projection Scenario - RCP4.5	senC2rcp45
SEN-C2-RCP	Projection Scenario - RCP8.5	senC2rcp85
SEN-C2-fODS	Projection Scenario - fxdODS	senC2fODS
SEN-C2-fODS2000	Projection Scenario - fxdODS2000	senC2fODS2000
SEN-C2-fGHG	Projection Scenario - fxdGHG	senC2fGHG
SEN-C2-fEmis	Projection Scenario - fxdEmis	senC2fEmis
SEN-C2-GeoMIP	Projection Scenario - GeoMIP-G1	senC2GeoMIPG1
SEN-C2-GeoMIP	Projection Scenario - GeoMIP-G2	senC2GeoMIPG2
SEN-C2-GeoMIP	Projection Scenario - GeoMIP-G3	senC2GeoMIPG3
SEN-C2-GeoMIP	Projection Scenario - GeoMIP-G4	senC2GeoMIPG4
SEN-C2-SolarTrend	Projection Scenario - SolarTrend	senC2SlrTrnd

Appendix 2 – netCDF 'global_attributes' required from Modelling Groups

Following CMIP5, we require the specification of the 'global_attributes' within each netCDF file. Some of these will be set internally by CMOR from the information contained in the MIP table, while many of the others must be set by the users. Note that if using CMOR to convert model data, the DRS components required to construct filenames and directory structures will be largely constructed from the 'global_attributes', as shown below.

The institute_id and model_id, being identical to the corresponding DRS components, are also subject to the DRS restrictions on which characters may be used. We ask groups to seek agreement with the CCMI organizers for institute_id and model_id to avoid problems. The most up-to-date list of agreed strings for model_id and institute_id can be found in the **CCMI_codes.txt** file available at http://www.met.reading.ac.uk/ccmi/?page_id=245.

Required netCDF global_attribute	Corresponding DRS Component
project_id	Activity
product	Product
institute_id	Institute
model_id	Model
experiment_id	Experiment
frequency	Frequency
modeling_realm	Modeling realm
table_id	MIP table
realization	Ensemble member (r<N>)
initialization_method	Ensemble member (i<M>)
physics_version	Ensemble member (p<L>)
branch_time	
contact	
Conventions	
creation_date	
experiment	
forcing	
initialization_description	
institution	
parent_experiment_id	
parent_experiment_rip	
physics_description	
source	
tracking_id	