Implementation notes for the DRS

11/07/10
Martin Juckes

1 Introduction

Some aspects of the definition of DRS and it implementation with respect to replication need to be clarified. Given the specification of the replicated portion of the data in the "archive_size.xls" spreadsheet, I suggest we activate the '[WILL POSSIBLY MODIFY THE ABOVE IF WE DON'T NEED TO KNOW ABOUT "REQUESTED"] ' option referred to in the paragraph defining the "product" component of the DRS in version 27.

2 Review of selection choices in *archive_size.xls*

2.1 Background

Output is requested from around 17 tables of variables – each table has a common frequency, but may cover more than one realm (e.g. "3hr"). Most of the 17 tables are further divided into sections with different spatial dimensions. The variables within each table are assigned priorities 1 (high), 2 (medium) or 3 (low). There are around 60 experiments. In some cases, output requested for the first, or first 3, ensemble member(s) is more extensive than that requested for subsequent ensemble members. In such cases, the experiment is listed twice in the "experiment" column of the archive size estimate, though it will, of course, only appear once in the DRS.

For some tables (specifically, for higher frequency output), output is only requested for selected time slices of the experiments.

Independent assessment from DKRZ suggests that supplying data for all ensemble members and for the complete time periods would increase the archive size 10-fold.

As detailed below, it appears difficult to reconcile the definition of "requested" data implicit in the CMIP5 spreadsheet distributed to data suppliers with that implied by the DRS (version 27). In order to be clear about which "requested" is being referred to, I will use "CMIP5:requested" to refer to the request sent out to modelling groups and use "DRS:requested" or "product=requested" to refer to its usage in the DRS.

3 Archive size: replicated versus cmip5:requested

3.1 Variables omitted

The following collections of variables in the cmip5:requested output are omitted from the replicated subset:

3.1.1 Annual ocean data

frequency:annual/realm:ocean/variables:{low priority}

This reduces the number of ocean annual variables (all 3d) from 57 to 16.

3.1.2 Monthly ocean data

frequency:monthly/realm:ocean/variables:{3d, low and medium priority}

This reduces the number of 3d variables ocean monthly variables from 19 to 9.

3.1.3 6 hourly atmospheric data

expt:{historical, AMIP,RCP4.5, RCP8.5, AMIP-atmos}/frequency:6hr/realm:atmos/ variables:{all of table 6hrLev}

Note that 6 hourly pressure level data for these experiments is still to be replicated, so this partition cannot be represented using the DRS frequency category.

3.2 Reduced time coverage

The following collections of variables will have reduced time coverage: realm:atmos/frequency:3hr/expt:{emission-historical}/variables:{in table 3hr}: reduced from 56 to 46 years.
This reduction in time coverage reduces the estimated size of the replicated portion by 1Tb. It would be a useful simplification to remove this change in time coverage.

3.3 Discussion

When the DRS version 27 was agreed on, there was agreement that replication should take place at the same level of granularity as publication on the ESG data node. Any finer granularity would be impossible to track in the catalogue. It was intended that the "product=requested" data collection would serve this purpose, and that the replicated data would be made up of a subset of the published units in the "product=requested" collection. It has been decided that data will be published at the "realm" level, meaning that a ESG publication unit (EPU) will be all data with a given: "<activity>/<product>/<institute>/<model>/<experiment>/<frequency> /<modeling realm> " and version. Given this definition of a publication unit, the replicated data for monthly/ocean data will be a subsets of the CMIP5:requested data EPU.

I believe the decision to proceed with two product categories was based on a misunderstanding of the criteria specifying the replicated subset. In order to fit replication into subset of EPUs in the currently defined "product=requested" data collection, we would have to omit high priority and 2d ocean variables, as well as 6 hourly pressure level atmospheric data: this option appears too drastic. Therefore, we need a "product=replicated" element of the DRS.

The "product=requested" branch does not appear to be required, so implementation can be given a lower priority.

4 Other DRS issues

4.1 Data beyond the CMIP5 request

As noted above, some modelling centres expect to archive (and process into CMOR2 compliant NetCDF) more data than specified in the CMIP5 request. This section raises a few questions about just how much additional data will be supported by level 1 QC process which data must pass to get into the ESGF CMIP5 archive. The most likely form of additional data will be extended time coverage and increased ensemble numbers, which will clearly be supported. Two other options are

listed below.

4.1.1 Experiments and frequencies

Clear definitions of experiments and frequencies should be given, if this option is required.

4.1.2 Additional variables

Additional variables could be added to the output if they have valid CF standard name and a MIP table name.

4.2 The DRS "variable identifier"

As noted above, the DRS states that the "variable identifier" is composed from the variable short name and the name of the table in which it occurs. The DRS does not, however, say how the variable short name and table name will be combined, though it does imply that neither hyphen nor underscore will be used (the first is ruled out for variables, the second for all components). The hyphen is ruled out because of a desire to have variable names which can be used as Fortran variable names – this objective also rules out ":".

A simple modification to the DRS document will make it consistent with usage in CMOR: let "variable name" ('tas' etc) and "table id" ('Amon' etc) be DRS components. Then the file name would be made up of DRS components joined by "_" and the directory names would also be DRS components.

5 Replicated as a subset of total output

Since the "requested" subset has minor importance from an archive management perspective, it is useful to document how the "replicated" subset relates to all possible output. The following selection rules are designed to be implemented sequentially.

5.1 Experiments, frequencies and MIP tables

Only the experiments and frequencies listed in the data request will be replicated.

5.2 Selection by Variables priority

Only variables listed in the request in each MIP table will be replicated. In the request, each variable in each table is assigned a priority and belongs to a section. In the following two tables, only a subset of the requested variables will be replicated:

- Oyr: only priorities 1 and 2.
- Omon: lat x lon x olev: only priority 1.

5.3 Selection by table-section, experiment and ensemble number

The final selection rules are based on tables in the "template" and "replicated subset" pages of "CMIP5_archive_size.xls". On these pages, and throughout the spreadsheet, rows correspond to experiments OR to subsets of ensemble members from experiments. In the latter case, there is one row for the first 1-3 ensemble members and a second row for the rest. This allows the output request to be more restrictive for latter ensemble members.

The is a requested ensemble size column in the template sheet, in some cases this is e.g. ">=3", indicating no upper limit [the calculation in the spreadsheet does not impose the upper limit, but I have not found any cases where the modelling groups expect to submit more than requested when there is a limit]. There is a length of run (in years) for each experiment, indicating the length of time which is considered part of the request.

Other omissions can be read from the "replicated subset" page of CMIP5_archive_size.xls", rows 40 onwards for years included and 114 onwards for ensemble members included. Blank indicates no data for that table section and experiment, 1000 indicates all requested years/ensemble members, and a number less than 1000 indicates a subset of years/ensemble members of the indicated length. For years, the definition of the time slice is given in the relevant sheet of "standard_output.xls". For ensemble members, if "n" are specified, the first "n" will be taken. The requested number of years in an experiment and number of ensemble members is listed in the "template" sheet of "CMIP5_archive_size.xls".