Identifying the product

Martin Juckes, Karl Taylor,

Version 0.5, Oct 6th.

0 Background

PCMDI has issued a request for data in the form of a "standard output" specification in a spreadsheet. Some modelling groups may wish to archive more data than has been requested. The federated archive infrastructure gives them the flexibility to archive and publish as much as they wish. However, a significant portion of data will be collected to central archives (PCMDI, BADC, DKRZ and others) to facilitate quality control and data processing and to ensure efficient access by providing multiple distribution points for the data which is expected to be accessed most often. The transfer of data to central archives will preserve "publication units" – that is, each block of data which is published on the data nodes will be either copied in full or not at all. In order to ensure that there is sufficient flexibility to copy the data which is wanted without having to transfer large amounts of unwanted data, the algorithm described below will split up publication units where necessary.

1 Information required about the file.

From file name (assuming a CMOR2 compliant file name):

- Identify DRS components, particularly variable_id, and hence MIP table and CMOR variable name, and the temporal coverage.

From configuration file:

- for each model: time at which historical and 1pctCo2 fork from piControl, start times of the latter two. Whether the model is used for the 'decadal', 'centennial', 'atmos-only' suite of runs. This information is used to prioritise runs for replication in a consistent way.

2 Issues

2.1 Policy

In order to simplify the process and allow future flexibility, the decision as to whether a dataset should be replicated or not will not be made at the publication stage. Rather, at the publication stage datasets which which may be too large for replication as a whole will be split into high and low priority sections. In identifying the high priority section, the algorithm will err on the generous side, taking more years than explicitly requested. All submitted years will be treated as high priority except in those tables and experiments where isolated slices have been requested.

2.2 Decadal vs. centennial

For piControl and 1pctCo2 experiments, there are different time slices requested for the short term and long term experiments. The correspondence between the piControl output and historical output for the long term (centennial) experiments is important, because of shared initial conditions. This

does not apply to the short term runs. In order to identify whether piControl data is to be judged by the short term or long term request, a configuration file must specify which category the model belongs in.

3 Algorithm steps:

3.1 Decisions at atomic dataset level

For the majority of atomic datasets, the decision can be taken solely on the basis of: variable, MIP table and experiment.

1. If the variable is not in the request list of the MIP table: → output2.
2. All MIP tables not in [Oyr, Omon, aero, day, 6hrPlev, 3hr, cfMon, cfOff]: → output1.
3. If no output is requested for MIP table/experiment combination → output1.
4. Otherwise:
   4.1. For Oyr: decision based on variable priority.
   4.2. For table Omon: decision based on variable priority and dimensions.
   4.3. For table aero: If "alev1" is not in the dimensions → output1; otherwise go to file level decision.
   4.4. For table day: If the variable is in the first 10 of the table, → output1, or if all or no output requested, assign to output1; otherwise go to file level decision.
   4.5. For tables 6hrPlev, 3hr, cfMon, cfOff: If all or no output requested, assign to output1; otherwise go to file level decision.

3.2 Decisions at file level

If the number of years submitted is less than the number requested plus 5, there is an option to assign all to output1. At present this option is only applied to data requests specified in terms of relative dates (see below). With this option there are potential problems when published datasets are updated by adding new files to the existing collection. If decisions are taken on the basis of all files, cases can arise in which the addition of new files causes the product designated for previously published files to change. The risk of this happening cannot be avoided completely (and will result in an "ERR007" return code in version 0.4 of the code). The advice to the data node manager in these cases is to collect all the data required for the new version and do a new publishing step. In order to reduce the frequency with which this happens

3.2.1 Absolute dating

The requests come in the form of time slices (e.g. 2026-2045, 2081-2100) and lists of years (e.g. [2010, 2020, 2040, 2060, 2080, 2100]).

3.2.2 Relative dating

See section 4.

4 Relative time

Some data is requested in time slices which are expressed in time relative to some point which cannot be identified from the DRS elements. In all cases, a check will first be run to see if the number of years submitted is less than or equal to the number of years requested plus 5. If this is the case, all years will be treated as requested.

4.1 Decadalxxxx

The decadal hindcast and projection run requests for table aero are relative to the start, but the start year can be determined from the experiment name, so there is no ambiguity here. Years 10,20,30 will be taken for 1960, 1980, 2005, and year 10 for 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2001, 2002, 2003, 2004, 2006, 2007, 2008, 2009, 2010.

4.2 Relative to historical

In the piControl (centennial) experiment, tables aero, day and 6hrPlev, the time slices requested should correspond to the time slices of data saved for historical/RCP runs and for EsmPiControl, tables aero and day, time slices corresponding to EsmHistorical. The historical run is started from some year, “y_pic2h”, of the piControl, and this year corresponds to the start of the historical run, “y_h0”. For data in these categories, the request time in the piControl years is: “y_req-y_pic2h+y_h0”. Getting this correspondence right is important and the program will fail if it gets this far (i.e. if there are more years submitted than requested) and the necessary parameters are not given.

4.3 Relative to 1pctCo2

Corresponding to years of 1pctCo2: in piControl (centennial and decadal), table 3hr and cfMon, the time slice is relative to the year used to start the 1pctCo2 run, “y_pci21pc”. If this is not available:

- cfMon, column 2: take first 25 years submitted
- cfMon, column 4: take first 25 years and years 120-145 or last 25.
- decadal piControl, 3hr: take last 35 years submitted.
- centennial piControl, 3hr: take last 35 years submitted.

4.4 Last years

Last <n> years: in table 6hrPlev, mid-Holocene and last-glacial maximum: need last year of experiment in principle, but in this case the last <n+5> years submitted will be taken. In table 3hr, 1pctCo2 (decadal and centennial), we should ideally have the last 30 years of the requested period and some groups may extend the runs. Then require the start year of the experiment: “y_1pc0”.

4.5 First and last

First <m> and last <n>: in table 3hr, abrupt4xco2: ideally, the last <n> should be consistent across modelling groups (because the response is transient), so “y_4x0” is needed. The requested experiment length is 150 years of abrupt4xco2 and 140 years for 1pctCo2. If not supplied, or if inconsistent with data submitted, the first <m+5> and last <n+5> will be taken.

4.6 Specific years

Specific years in decadal experiments (excepting hindcats and projections dealt with above):

piControl: 5 individual years of "aero" output are requested: ideally need "y_pic0", but take all years if less than 10 are submitted. If more than 10 are submitted, try to take relative to 1st year submitted. 1 year is requested for the 3hr data: take year submitted, or closest to 30 years from start. For the 2010 volcano and initialisation alternatives (year 10 requested): follow the procedure just specified if start dates ("y_2010v0" and "y_initalt0") are not given. The possibility that there may be multiple initialisation alternatives with different start years is not going to be supported in the initial release.

4.7 cfMon

There are many experiments for which years 1 to <n> are requested, with n=5,20,30. Some also have years 121-140 requested, and 3 only have 121-140 requested.

- For 6.2a, 6.2b, 6.3-E, 6.4a, 6.4b, 6.7a, 6.7b, 6.7c: take first <n+5> years submitted.
- 3.1 (piControl): see above (section 4.3).
- 5.4-l, 5.5-l, 6.1 (1%..), 6.3 (abrupt4xco2): use "y_1pct0" and "y_4x0" if available, otherwise take all years if less than 25, years 121-145 if more than 145 submitted, or last 25 years submitted.

These 13 experiments cover all the relative time references in cfMon.

5 Appendix 1: Experiment dependencies.

The "Child" experiments in the table are initialised from the "primary" experiments. In these cases there is an important connection between ensemble members of primary and children which needs to be taken into account in archive management.

"Primary" experiment	**"Child" experiments**
historical	rcp26, rcp45, rcp60, rcp85
piControl	1pctCo2, abrupt4xco2
esmHistorical	esmRcp85
esmControl	