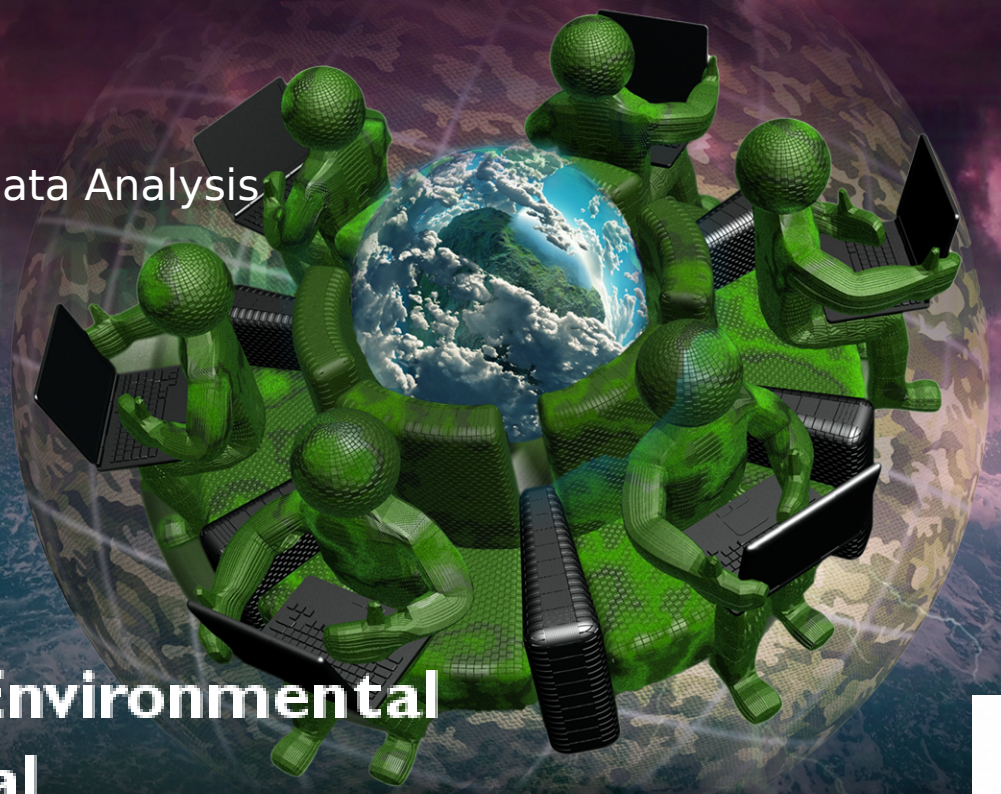
8th Annual Earth System Grid Federation

Status and Outlook for the CMIP Data Request

Martin Juckes
Centre for Environmental Data Analysis



December 4-7, 2018

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

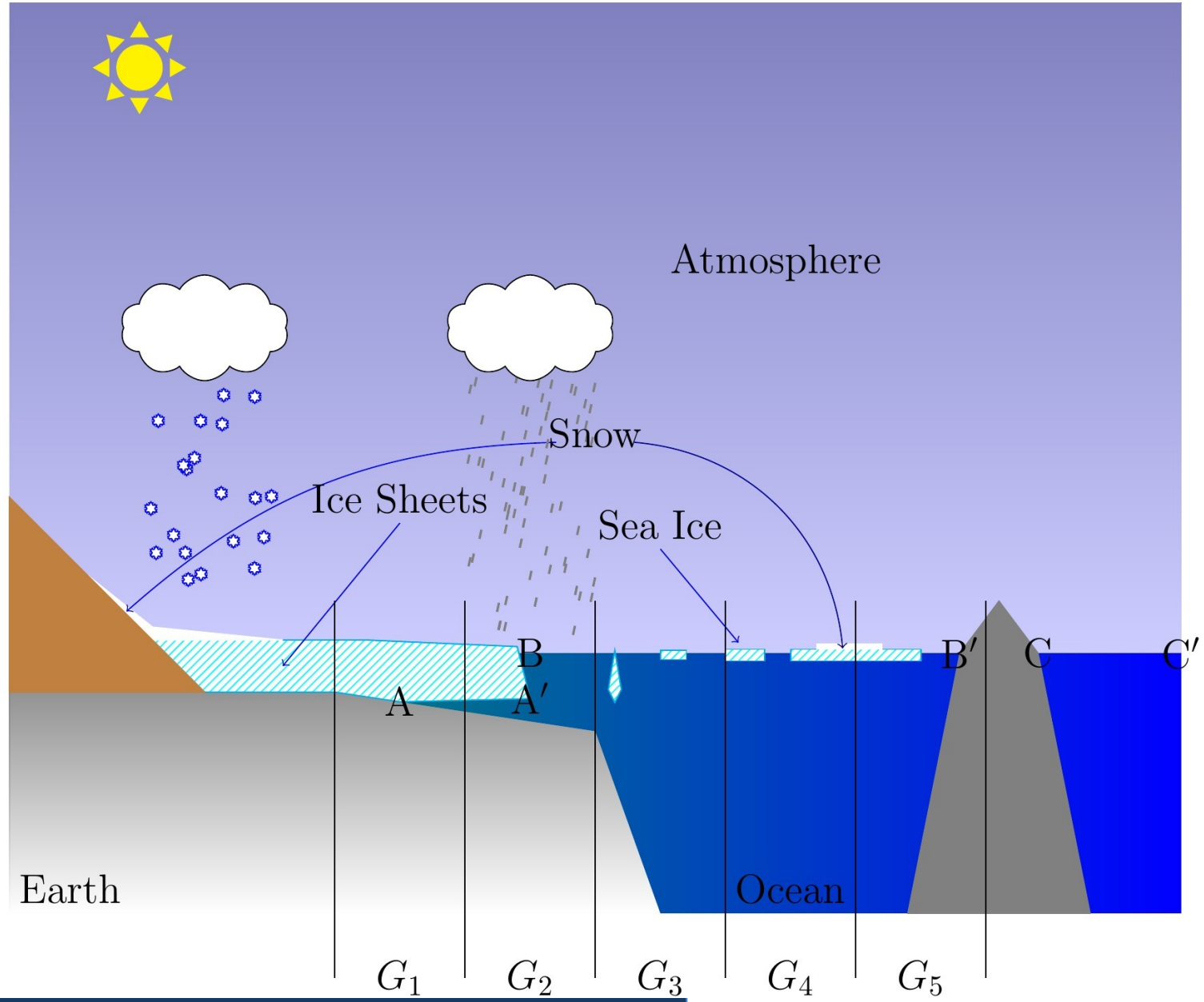
- Martin Juckes
- Karl Taylor
- Matthew Mizielinski
- Alison Pamment
- Sébastien Denvil
- Stéphane Sénési

Outline

- CMIP6 Data Request
  - What is the Data Request? [2]
  - Information Content [5]
  - Components [4]
  - Work-flows [1]
- CMIP7
  - Initial lessons [2]
  - Outlook [2]

- Consolidated specifications of output variables and requirements for 25 MIPs across 287 experiments;
- Reference document, web site and python package;

http://w3id.org/cmip6dr



CMIP6 Data Request

- Defining the diagnostics to be archived;
- Many new surface categories, especially in the cryosphere (e.g. floating_ice_shelves,

The Role of the Data Request

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

The data request supports work-flow configuration and CMORisation.

MIPs develop scientific objectives

Simulation workflows

Data Request

ESDOC

CMOR

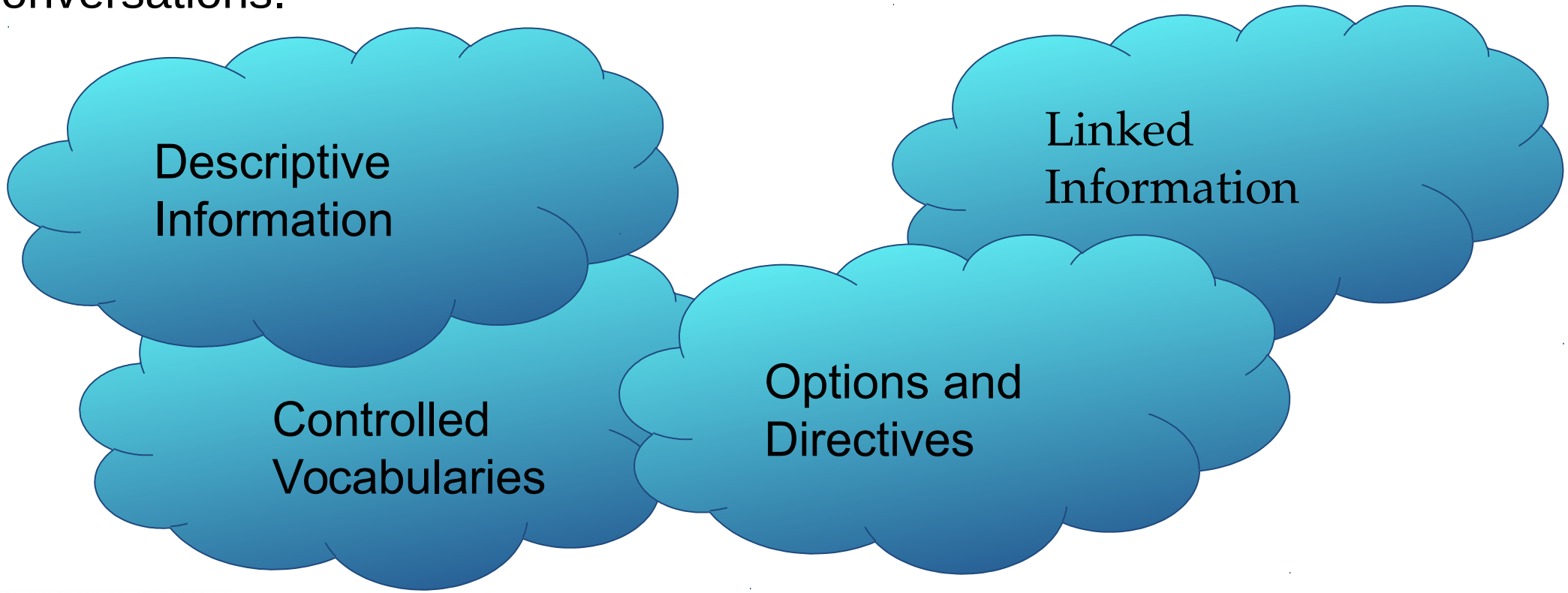
Earth System Grid Federation
ESGF

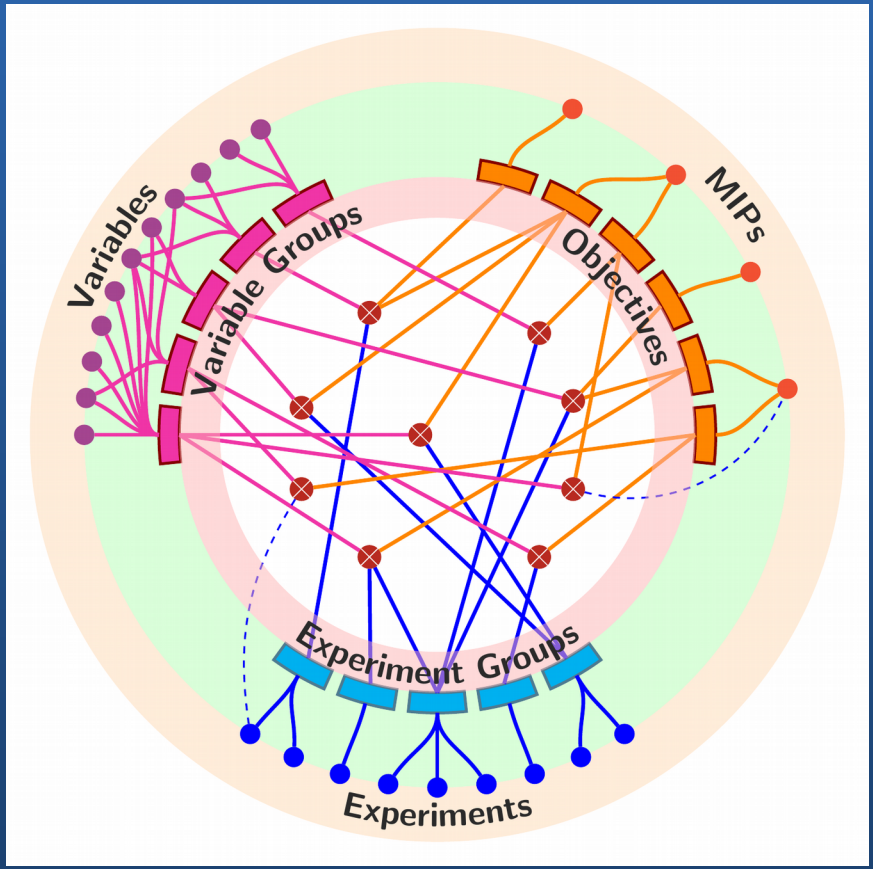
Scientific Analysis

CVs link all components

Information Content of the CMIP6 Data Request

The information in the Data Request comes in many forms, each requiring a different range of conversations.





Linked Content

- The **variable group** DCPP-Amon (36 variables) is requested by DCPP from the historical and DCPP-A **experiments** to support the “hindcasts” **objective** (i.e. “To assess the decadal predictability and forecast skill of forced and internally generated climate”).

Descriptive Content

Example

- **Label:** fNproduct

- **Title:** Deforested or Harvested Biomass as a Result of Anthropogenic Land Use or Change

- **Description:** When land use change results in deforestation of natural vegetation (trees or grasslands) then natural biomass is removed. The treatment of deforested biomass differs significantly across models, but it should be straight-forward to compare deforested biomass across models.

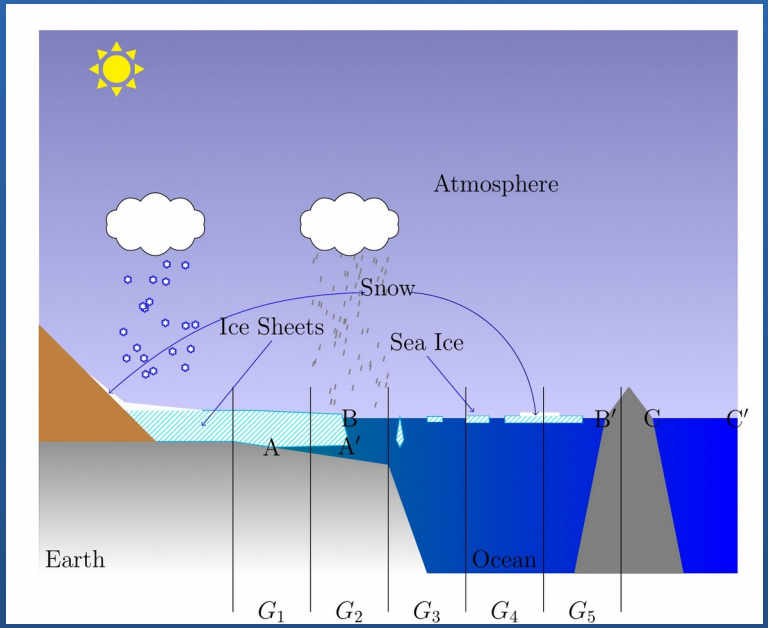
- Descriptive information contains “plain language” (sometimes with specialised scientific terminology) describing, for example, a diagnostic

Controlled Vocabularies

- Activity
- Experiment
- Frequency
- Grid Label
- Institution
- Nominal Resolution
- Realm
- Source (model)
- Source Type
- Sub-experiment
- Table

- The request relies heavily on external vocabularies, such as CF standard names, CMIP Controlled Vocabularies, CMOR directives.

- Following the evolutions of these vocabularies while consolidating input from multiple sources is a significant challenge.



Options and Directives

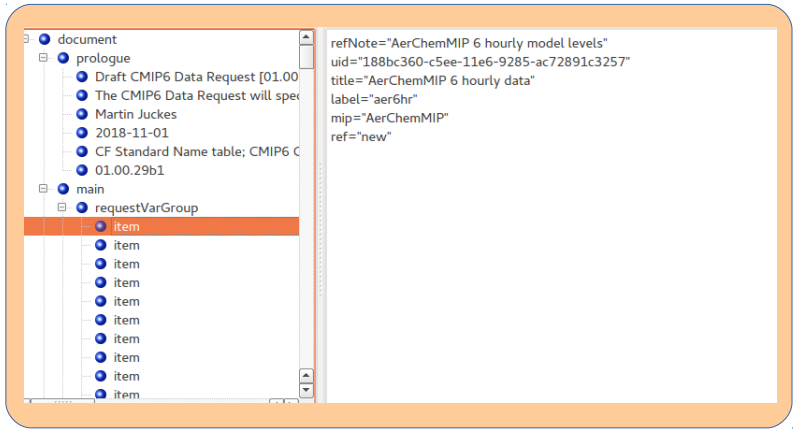
Some attributes represent values that need to be inserted into file metadata, other indicate options or instructions.

- cell_measures = “--MODEL” --> data should be archived on cell position used in model (e.g. vertex, edge or centre) rather than being interpolated to centre.
- dimensions = “latitude, longitude” --> data should be archived on a spatial grid appropriate to the parameter in question.

Accessing the data request

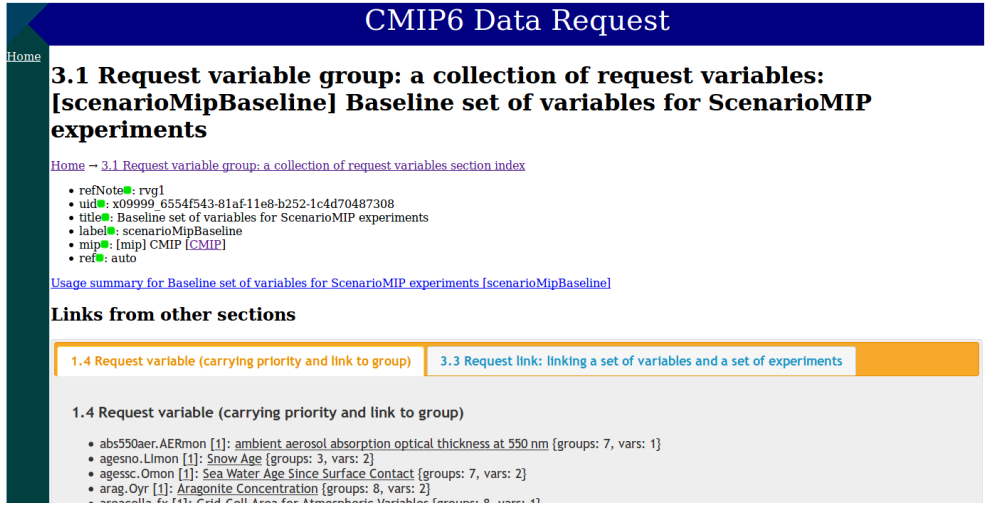
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

- The request is delivered through a range of formats
- The primary reference XML document is around 20,000 records
- Python library with command line and API
- Web pages



Reference Document

Web pages



Modeling centres have their own code to extract the information they need from the request, using the XML document directly, the python library, or parsing the output of the command line tool.

XML reference document

- The full request is provided in a single structured document, ensuring portability and enabling use in data production environments.
- Flat “table” structure: 35 sections, each consisting of a list of “item” records;
- Constrained to enforce typing of data values (e.g. float, integer, integer list, etc.
- Schema largely fixed since early 2016.

Web Pages

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Web pages support searching and browsing of content.

The web site provides access to tables of variables for each MIP and experiment.

The search page allows variables to be filtered on variable name, long name, standard name, units and description.

Units: [kelvin] Kelvin

Home → Units section index

- uid[i]: fd70554e-3468-11e6-ba71-5404a60d96b5
- title[i]: Kelvin
- text[i]: K
- label[i]: kelvin

Links from other sections

1.2 MIP Variable

1.2 MIP Variable

Variable	Standard name	Long name	Units	Description	uid
$t		flux			

tau , *magnitude_of_surface_downward_stress* [Momentum flux] (N m-2) {module of the momentum lost by the atmosphere to the surface.} <590e6e4c-9e49-11e5-803c-0d0b866b59f3>

tendacabf , *tendency_of_land_ice_mass_due_to_surface_mass_balance* [Total surface mass balance flux] (kg s-1) {The total surface mass balance flux over land ice is a spatial integration of the
balance flux} <4146943a-4f40-11e6-a814-ac72891c3257>

tendlibmassbf , *tendency_of_land_ice_mass_due_to_basal_mass_balance* [Total basal mass balance flux] (kg s-1) {The total basal mass balance flux over land ice is a spatial integration of the ba
balance flux} <414699d0-4f40-11e6-a814-ac72891c3257>

tendlicalvf , *tendency_of_land_ice_mass_due_to_calving* [Total calving flux] (kg s-1) {The total calving flux over land ice is a spatial integration of the calving flux} <41469f3e-4f40-11e6-a814-ac728

Python Library

Available in svn or Python Package Index: pip install dreqPy

Use in python code:

```
from dreqPy import scope
sc = scope.dreqQuery()
v1 = sc.volByMip2( 'C4MIP', pmax=2 )
v2 = sc.volByMip2( {'C4MIP', 'LUMIP'}, pmax=2 )
```

From the LINUX command line:

```
drq -m C4MIP,LUMIP -p 1 -t 1                  →  4.20Tb
drq -m HighResMIP -p 1 -t 1                   → 29.0Tb
drq -m HighResMIP:DiurnalCycle -p 1 -t 1
                                              → 3.9Tb
```

Workflows

Input from:

- MIPs;
- Modelling groups;
- Teams working on CMOR, ES-DOC

Communication through:

- spreadsheets (excel and google);
- Emails;
- Discussion forum (years 1 & 2);
- Github issues;

Since June 2018:

- Beta releases and reviews to improve stability of new releases.

Successes

- Schema allowing consolidation of multiple MIP requirements;
  - Variable definitions re-used (927 standard names, 1272 MIP variables, 2063 CMOR variables) enbling robust definitions;
  - Permits selection of variables based in scientific objectives, priorities, and experiments;

- Re-usable meta-data structures, e.g. cell methods, units, spatial structures;
  - robust validation of syntax;

- Multiple viewing formats giving flexibility to users;
  - e.g. web site for search and browsing, XML for embedding in work flows, python API for systematic exploration;

Problems

1. Lack of stability early on;

2. Slow trickle of requirements;
   -- e.g. revisions to reference vocabularies;

3. Too much complexity;

4. Divergent approaches from MIPs;
   -- developing the data request at the same time as the new "endorsed MIP" process was being introduced added confusion and stress.

5. Time to solution too long;

File names

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Sub-expt. (optional)

For CMIP6:
tas_Amon_BCC-CSM2-HR_1pctCO2_r1i1p1f1_gn_198001-198412.nc

Variable | Table | Source (model) | Experiment | Member | Grid id. | Time range

The file naming scheme ensures unique file names for each data product. The name can be split into 3 categories of information: (1) the parameter (green), (2) the experiment (blue) and (3) implementation details (red).

Variable names

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

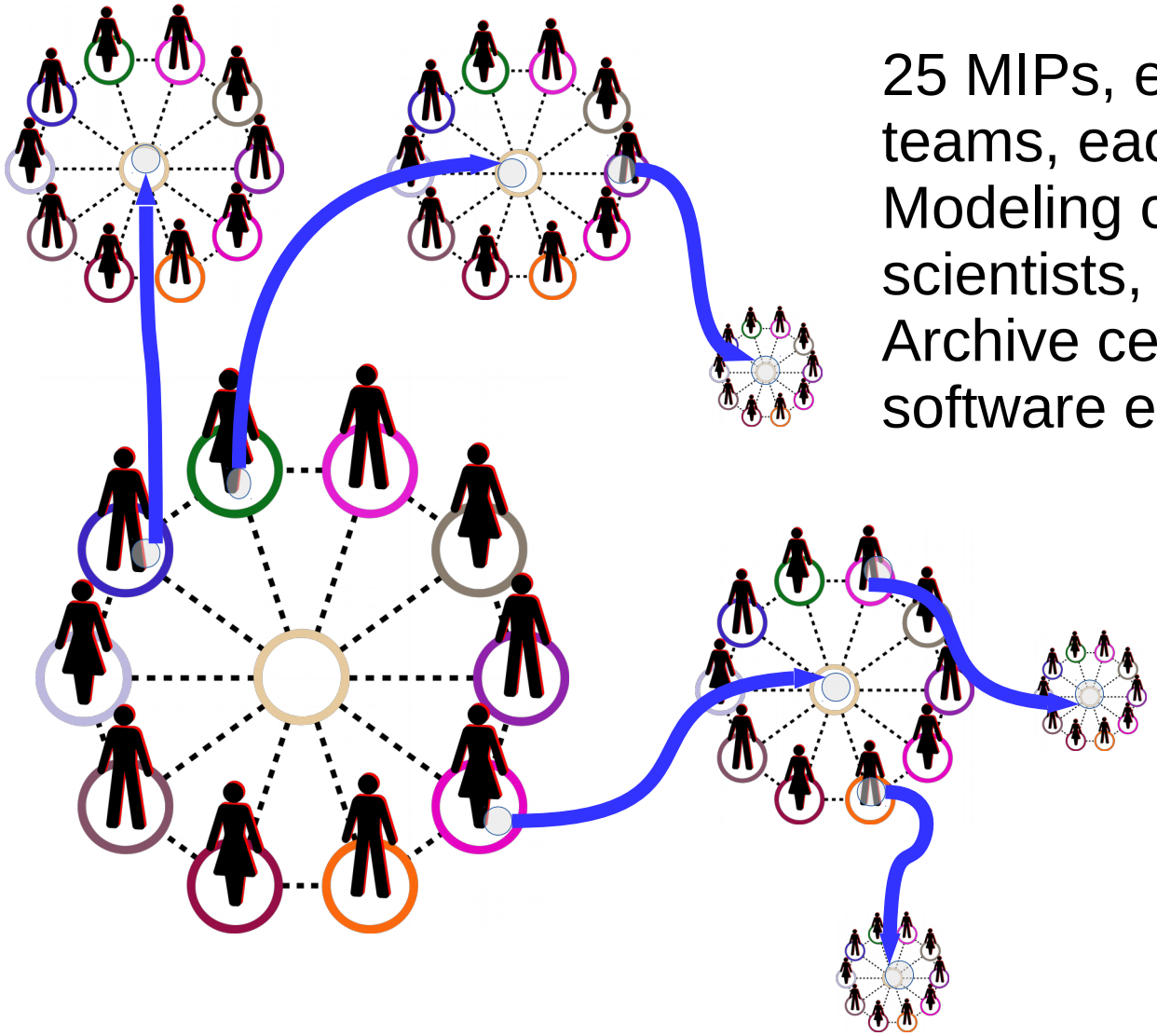
Variable names must be consistent with unique file names;
Precise rules for variables depend on the file naming convention;
Discussion of variable names starts early .... file naming convention is finalised later in the process;

However:
Rules for variables only depend on part of the file naming convention (the “variable name + table” pair) which could be finalised earlier .....

tas_Amon_BCC-CSM2-HR_1pctCO2_r1i1p1f1_gn_198001-198412.nc

Communication: the challenge

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL



25 MIPs, each involving multiple research teams, each team with multiple specialists; Modeling centres: managers, environmental scientists, data scientists, software engineers; Archive centres: standards, data scientists, software engineers.

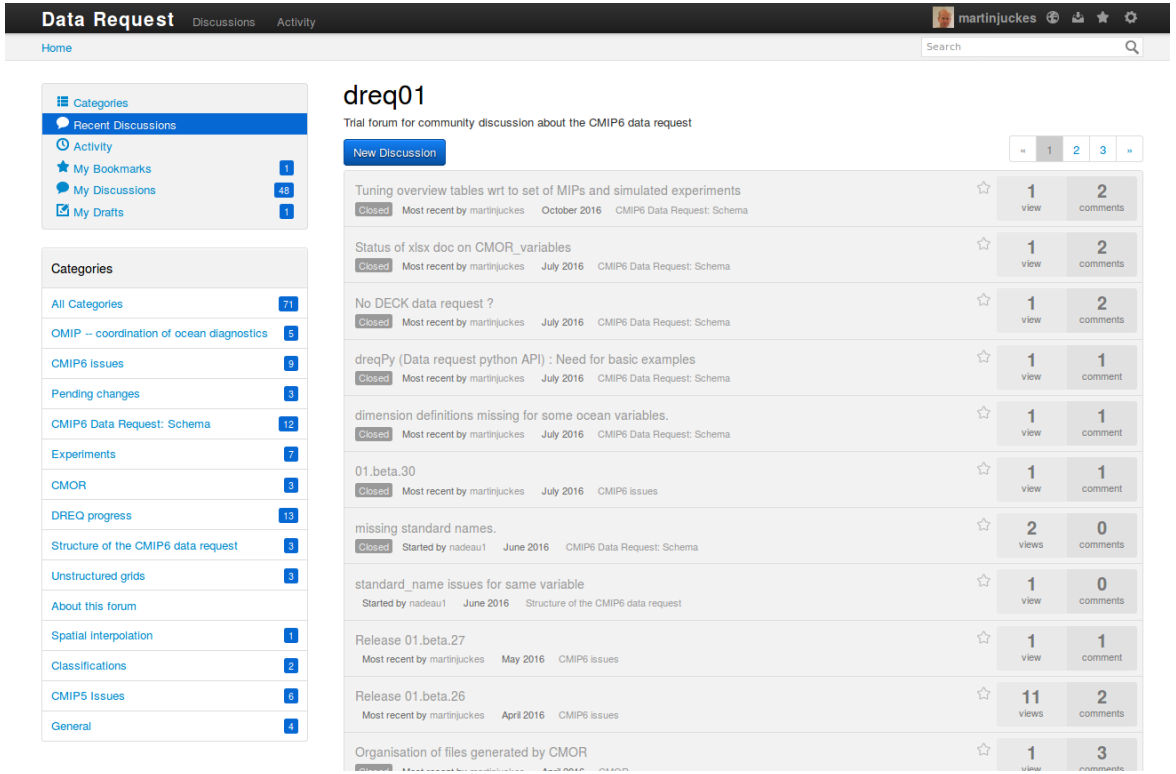
Different communication strategy needed for different user groups.

Communication: approaches

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Early CMIP6

Emails (lists for distribution: individual for queries);
Forum site for discussion;
Spreadsheets;



Late CMIP6

Emails (lists for distribution: individual for queries);
Github for issues;
Spreadsheets + google sheets;

CMIP7 ideas

Emails (lists: general for distribution + **MIP specific for queries**);
Github for issues;
Google sheets? (*with python integration?*);

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Data Request Working Group

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Initial meeting in June 2018 (see http://bit.ly/dreq_rev2018);
Discussed finalisation of CMIP6 request and outlook for the future.

- Rationalisation: e.g. fewer grid specification options;
- Clearer schema;
- Clearer interfaces to other components;

Funding for a further meeting is allocated in IS-ENES3;

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Recommendations

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Data Request Decisions

- Rationalise;
- Streamline interactions with CF;
- Improve tracking of discussions (all discussions in a tracked environment);

Request to WIP

- Agree framework for parameter component of file names for CMIP7 early (i.e. 2019), so that variable naming rules are clear;
- Agree a communication framework (e.g. github, forum, email lists: thisMIP@dreq.org) and insist on registration by interested MIPs;
- Establish meaningful metrics of progress;

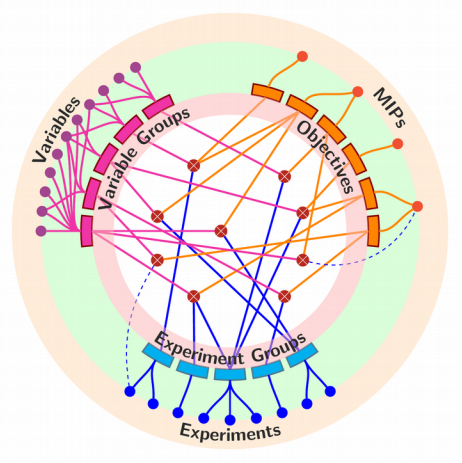
Request to WGCM/CMIP panel

- Clarity around responsibilities of MIPs;
- Encourage coordination among related MIPs (perhaps through core projects);

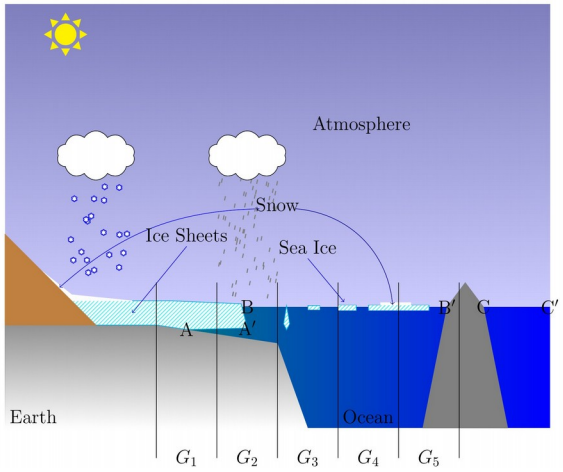
Outlook

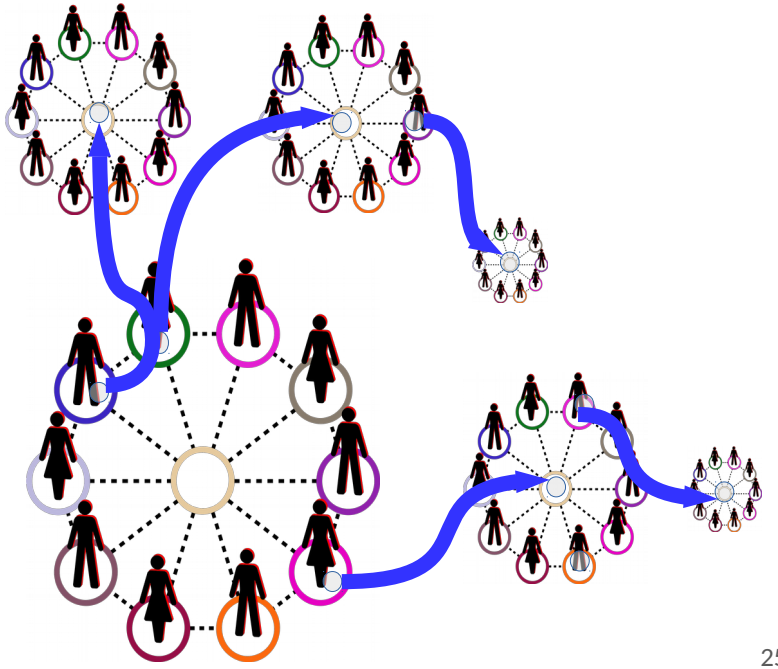
- Funding for further work through IS-ENES3;
- Plan for more MIPs in CMIP7 (e.g. hydrology, fire);
- Explore options for supporting more projects that want to use

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING


Variables
Variable Groups
Objectives
MIPs
Experiment Groups
Experiments

The End


Atmosphere
Snow
Ice Sheets
Sea Ice
A
A′
B
B′
C
C′
Earth
Ocean
G_1
G_2
G_3
G_4
G_5



Successes

- Schema allowing consolidation of multiple MIP requirements;
- Re-usable meta-data structures;
- Multiple viewing formats;
- Automated tests of units, cell methods, attribute types and much more;

Problems

1. **Lack of stability early on;**
2. **Slow trickle of requirements;**
3. **Too much complexity;**
4. **Divergent approaches from MIPs;**
5. **Revisions to reference vocabularies;**
6. **Time to solution too long;**

Status and Outlook for the CMIP Data Request

Martin Juckes, Center for Environmental Data Archival

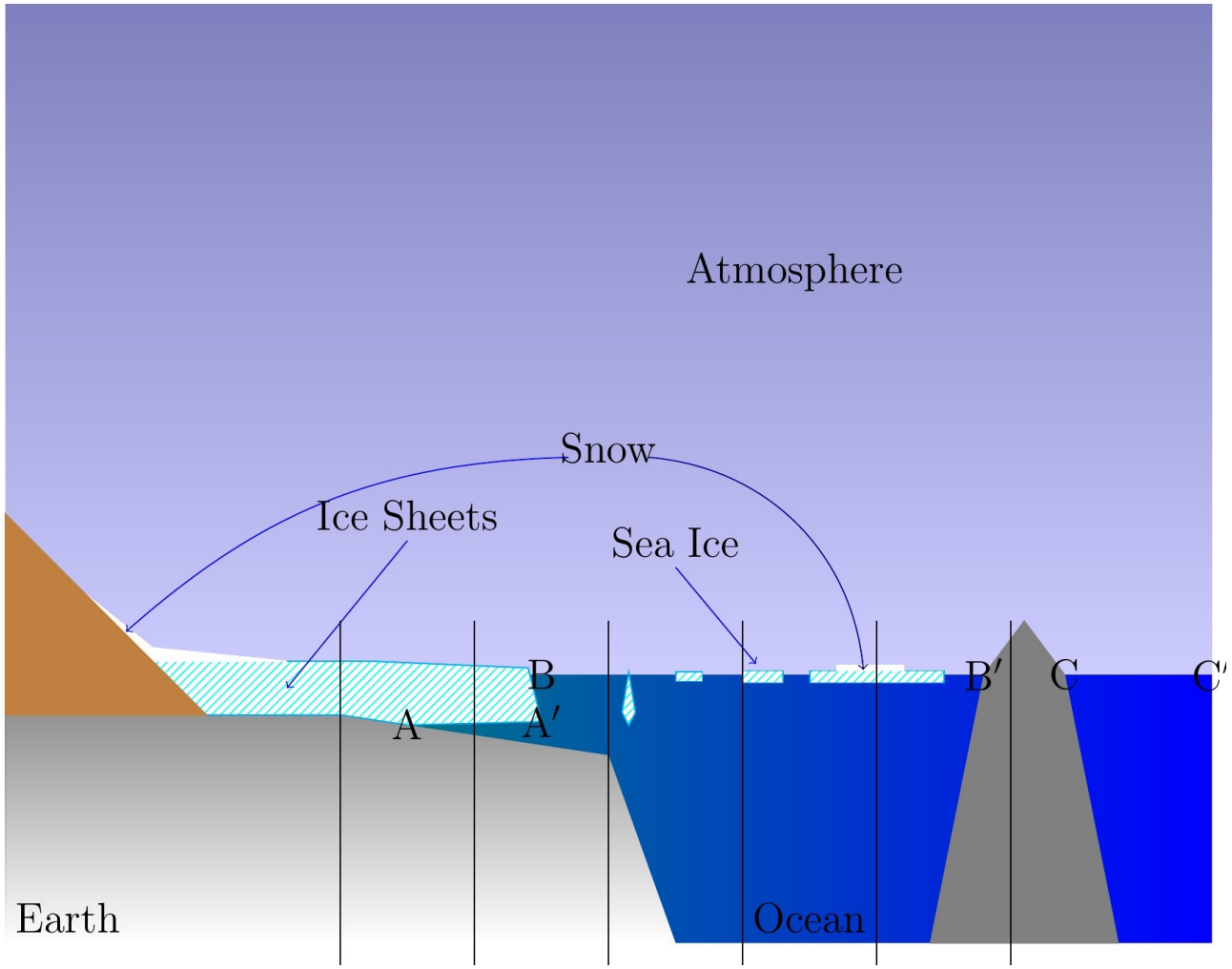
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

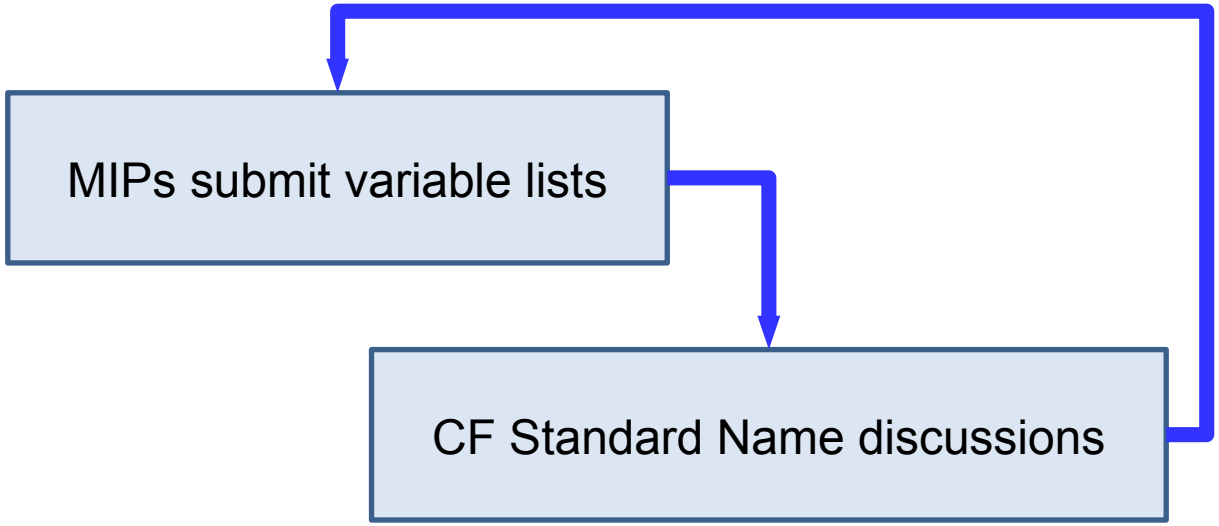
http://w3id.org/cmip6dr

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

New surface types to consider, e.g.:
floating ice shelves,
grounded ice sheets
snow on sea ice



Workflows



Revisions to structure

Feedback from data archives and service providers

Feedback from modelling centres

Aerosol	177	153	129
Atmosphere	753	325	256
Atmospheric Chemistry	18	13	11
Land	365	291	191
Landice	174	78	57
Ocean	260	165	148
Ocean Biogeochemistry	287	214	151
Sea Ice	104	92	79

Lessons ...

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Lessons:
Streamline interactions with CF;
MIPs need a technical point of contact who understands the relevant standards;
Need a clear framework for organising content early on;
For many scientists, email is the most effective means of communication;

Lessons:
Streamline interactions with CF;
MIPs need a technical point of contact who understands the relevant standards;
Need a clear framework for organising content early on;
For many scientists, email is the most effective means of communication;

Scope of CMIP7

It is usual for the climate modelling community to start talking about limiting the scope of CMIP, especially when the stress of dealing with the current phase is high.

But .... the world is warming ... CMIP5 provided scenario runs with coupled carbon cycles, CMIP6 adds nitrogen cycles, and greater variability in the biosphere and cryosphere.

CMIP7: expect greater complexity in models, e.g. resolving food productivity, and moving from science questions (what is the environmental impact of forcing X?) to policy questions (what is societal impact of policy decision Y?)

Lessons:

- Need to simplify the request: remove complex grid preferences; clarify linkages;
- Variable definitions can be re-used by other MIPs;