Data formats and requirements in CMIP6: the climate-prediction case

Pierre-Antoine Bretonnière

I – Experience from previous Model Intercomparison Projects

- CMIP5
- SPECS

II – How to comply with strict metadata and format conventions

- A CMOR history
- Plans for CMIP6

Experience from CMIP5

- 1.8PB for 59000 data sets stored in 4.3Mio Files in 23 data nodes, 116 experiments published. (=CMIP3x50)

- NetCDF3 format

tas_Amon_EC-Earth_historical_r1i1p1_185001-200512.nc

- Triple data quality control:

    ESGF publisher conformance checks,
    Data consistency checks,
    Double and cross-checks of data and metadata and DataCite
    data publication

- Experience and model description shared before data publication (ES-DOC database)

<u>Lessons and future requirements:</u>

- Usability of ESGF data access interface

- Automated data replication between ESGF data nodes

- More powerful, more stable and scalable wide area data networks (service level agreements)

- Detailed information of initialization, physics, etc should be more easily accessible (were in the model documentation)

Experience from SPECS

- 80 TB of data, being stored at the BADC and published on the ESGF nodes

- **NetCDF4** format (1.5 to 2x space saving)

- **Double time axis** to encode seasonal to decadal predictions

- Add **start date** in the name of the file

IPSL-CM5A-LR/decadal/S19890101/mon/ocean/tos/r3i1p1/tos_Omon_IPSL-CM5A-LR_decadal_S19890101_r3i1p1_198901-199812.nc

- New attributes:
  - **initialization** and **physics description**
  - associated experiment

- Creation of "deposit receipts" when data is published

CMOR history

- CMOR is a library of C functions (with Fortran90 and Python interfaces) which facilitate/enforce **compliance with MIP requirements**.

- CMOR was designed to be adapted to the **different metadata requirements** of each "model inter-comparison project"

- CMOR2 has been used in CMIP5, and a SPECS patch to encode new requirements and project specific conventions was developed.

CMOR history

- Ensures compliance with:
  - NetCDF – (www.unidata.ucar.edu/software/netcdf/)
  - CF Conventions (provides standardized description of data contained in a file (cf-convention.github.io)

- Data Reference Syntax (DRS) – defines vocabulary used in uniquely identifying MIP datasets and specifying file and directory names (cmip-pcmdi.llnl.gov/cmip5/output_req.html). **Project specific**

Plans for CMIP6

- **The CMOR library will be used to ensure that all the data produced by the different partners have the same standardized format**

- **CMOR3 is under development to:**

    - Better handle a wide range of models and observational data

    - Modularize CMOR input tables

    - Integrate CMIP6 format conventions changes (including SPECS standards)

- **New format and DRS under development**

- **Data quality control** handled at several levels (improved in the CMORization process itself: e.g., valid max and min for each field)

EC-Earth meeting, Reading, May 2015

Plans for CMIP6

- SPECS contribution to CMOR will ensure that the needs of the Decadal Climate Prediction Project are taken into account

- For EC-Earth users, unifying the data formatting (using tools like ece2cmor) in CMIP6 will ensure better efficiency in the CMORization

- Use of XIOS to facilitate the formatting of the data
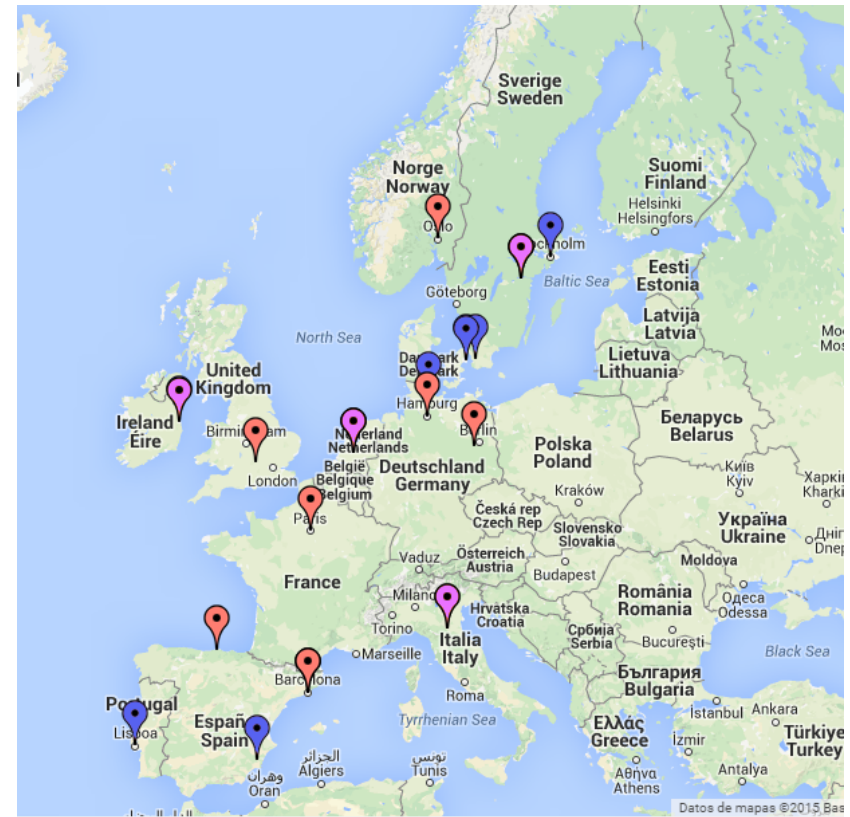
ESGF nodes and EC-Earth users map

Thanks for your attention!

Questions?