Project 7 (TWP1): Facilitating the distribution and analysis of CMIP5 and related projects

Project lead: Sebastien Denvil
Post-doctoral researcher: Guillaume Levavasseur
Project Start/End: October 2013 – September 2014

Position offer:

Intitulé du poste : Soutien à la diffusion et à l'utilisation des résultats des simulations climatiques du projet CMIP5

CDD, 12 mois, Ingénieur d'étude ou Ingénieur de recherche

Lieu de travail : IPSL, Paris, Université Pierre et Marie Curie (Jussieu)

Rémunération brute: 2000 à 2600 € (selon diplôme et expérience)

Contexte : Le projet CMIP5 coordonne au niveau international la réalisation de simulations climatiques et la distribution de leurs résultats. Ces simulations servent de supports à de très nombreux travaux d'analyses scientifiques et alimenteront notamment le prochain rapport du GIEC.

Le Pôle de Modélisation de l'Institut Pierre Simon Laplace (IPSL) contribue au projet CMIP5 en réalisant des simulations climatiques avec trois versions du modèle climatique qu'il développe. Ces données, ainsi que celles de la trentaine de modèles participants à ce jour au projet CMIP5, sont stockées sur un important espace disque (450 To) accolé à un cluster de calcul afin de faciliter leur analyses par les personnels de l'IPSL. Ce travail est réalisé en liaison avec les autres activités de l'IPSL concernant l'analyse, l'archivage, et la distribution de données scientifiques (projet ESPRI, Ensemble de Services Pour la Recherche à l'IPSL) et est supporté par le labex L-IPSL.

La précédente phase du projet CMIP (CMIP3) a connu un très grand succès: les données ont été et sont toujours très utilisées, par des acteurs très variés (des scientifiques aux bureaux d'études) et ont alimenté un nombre considérable d'articles scientifiques (supérieur au millier). On anticipe un succès encore plus important pour le projet CMIP5, pour lequel le nombre de variables et de simulations est beaucoup plus élevé, avec des fréquences de sorties et des résolution spatiales également plus élevées. Au sein de l'IPSL, plusieurs dizaines de personnes analysent déjà les données CMIP5.

Mission :

- contribuer à la mise en forme des données CMIP5, à leur transfert sur l'espace de stockage, à leur publication sur le système ESGF (Earth System Grid Federation) et à leur documentation
- recenser les erreurs identifiées dans ces données, contribuer à leur corrections, tenir le journal d'évolution des versions des fichiers au fur et à mesure de ces corrections
- aider les utilisateurs pour l'utilisation des données CMIP5, notamment via les outils et les moyens mis à disposition par l'IPSL
- compléter le site icmc.ipsl.fr avec les informations sur les simulations, sur les variables, sur leur accès et leur utilisation
- constituer une FAQ à partir des questions-réponses déjà existantes

- faire évoluer cette FAQ avec les nouvelles questions et les réponses obtenues auprès des experts de l'IPSL

- orienter les utilisateurs dans l'utilisation des résultats de simulations en fonction de leurs besoins

Profils et Compétences:
- Unix/Linux ; scripts (bash/python)
- format de fichier netcdf et utilitaires associés (nco, cdo...),
- visualisation et analyses statistiques simples
- excellent relationnel, capacités organisationnelles et autonomie
- bonne pratique de l'anglais indispensable
- connaissances de base en climatologie appréciées
- le site icmc est sous joomla : apprentissage prévu si besoin

Contact:
Sébastien Denvil: tel: 01 44 27 21 10 - courriel: sebastien.denvil@ipsl.jussieu.fr

Preliminary results

IPSL researchers have different knowledge of existing databases (e.g. CMIP5) or computer languages. Consequently, data analysis quickly becomes complex and time-consuming, especially for beginners or doctoral students. It appears necessary to clarify the data access and to facilitate analysis without resorting to more training. Task 5 of the TWP1 of Labex-IPSL aims to improve support for CMIP5 analysis at IPSL.

I - Clarify and facilitate data access

Initially CMIP5 IPSL files were divided into 3 storage areas: DMF and STORE at TGCC-CCRT and CICLAD filesystems leading to an arduous search files. This was due to the tape archive system replacement at TGCC. In order to make easier data access, we first merged CMIP5 IPSL files into the local work disk at TGCC-CCRT. This migration process has been performed for all IPSL results. This procedure involves several steps like matching files between the 3 filesystems, copying datasets if necessary, cleaning, updating and assigning version to each CMIP5 datasets and finally deleting orphaned files.

To make our merged CMIP5 data available on ESGF from TGCC, we installed and configured an ESGF data node at TGCC accessible through any ESGF front-end and in particular from the IPSL ESGF front-end. Consequently, CICLAD cluster will thus partly be dedicated to the use and the development of CMIP5 analysis. CMIP5 data distribution of IPSL results per se becoming the role of the TGCC ESGF data node.

II - Improving support through errata and documentation

In order to clearly identify the produced data and their access, a huge documentation work was conducted to list all known errors in IPSL CMIP5 data. Today, 20 of 40 identified issues have been corrected. These files modifications concern about 17% of the IPSL results (about 27 000 files among 140 000) and leads to 25 version number. Each file version has been assign to a possible problem to built a fully documented errata. This errata is stored in an SQL database that can be queried through a web interface on the ICMC website. Users can now easily find information about (i) all issues as short and complete descriptions, the affected files, the corresponding version numbers, graphics and/or maps and the issue status (corrected, in progress or to do) ; or (ii) send a request about downloaded data to know if they have the latest file version, if modifications occurs and their history and the links to the corresponding description (cf. schema below).

This web-module is included in a fully redesign and documented page called «IPSL contribution to CMIP5». The user can find in the same place of the errata, a lot of useful information for their analysis

as: details about IPSL-CM5 model versions, all related documentation and references, the forcing files, details and graphics about CMIP5 requirements, vocabulary and tree, links to European ESGF data node and IPSL services, and answers to the most frequently asked question.

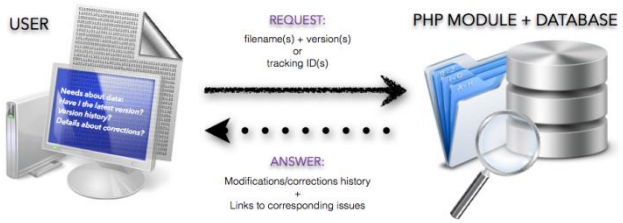
Also, we use the IPSL-CM5 errata as a proof of concept to promote such a tool into the ESGF platform and to design a more controlled versioning system during the publication process. A large portion of requests from CMIP5 users deal with files errors. To establish a controlled errata mechanism inside the ESGF platform clearly appears as a major support tool. To that end we wanted to strengthen our investment in the ESGF development. In relation to that goal we are now part of the ESGF Publication Working Team where we represent the French ESGF users community, especially located on computing center and partner institutes (TGCC, IDRIS, CINES, CERFACS, CNRM). We are the key French point of contact for publication on ESGF and we provide an essential support to those French actors.

III - A CMIP5 analysis software stack for IPSL community

Another goal specifically covered by this task was to help IPSL researchers with different computer languages skills to easily perform a reliable analysis and save time. We decided to develop an ecosystem of CMIP5 analysis software stack that will be easily usable by IPSL community:

1. The CMIP5 files are fictitiously concatenated along time dimension through OpenDAP URL. These aggregations avoid dealing with the files splitting depending on model and frequency over a time period. We developed a Python command-line tool allowing the users to find and list the available CMIP5 aggregations at IPSL in a fast and researcher-friendly way. The user just has to fill a template with the required variables, experiments and ensembles. The script directly returns available models that match ALL requirements with the corresponding aggregations list.

2. Aggregations can exist without a correct time axis. Nevertheless, time axis often is mistaken in CMIP5 files and leads to flawed studies or unused data. We developed a Python program to check and rebuild a proper time axis if necessary. About 107 000 files (i.e. 23% of the total files, see statistics below) more aggregations will be thus available. This script will be included in synchro-data downloading workflow to deliver proper data to users.



There tools are currently test following a study on Pattern Scaling. Our goal is to identify the respective role of forcings and models in the characteristics of the climate changes patterns. Pattern scaling allows to describe, at first order, the general pattern of temperature and precipitation changes. Previous tools will be part of standard analysis stack for CMIP5 file allowing to easily compare simulations of different scenarios. In just one month we bring the following conclusion : (i) pattern differences between models are much larger than pattern differences between scenarios ; (ii) for scenarios with small radiative forcings (e.g. RCP2.6), the internal variability has a significant contribution to the spread of the pattern scaling for temperature (mostly at high latitudes) and precipitation (mostly in the tropics) and (iii) stabilization of temperature, and therefore of forcings,

References:

http://icmc.ipsl.fr/

Statistics about times axis from /prodigs/esg/CMIP5

Institute	Nb files in latest	Nb of mistaken time axis	% of mistaken
BCC	8194	5416	66,10 %
BNU	837	0	0,12 %
CCCma	6478	0	0,06 %
CMCC	28345	27072	95,51 %
CNRM-CERFACS	36185	3891	18,75 %
CSIRO-BOM	4967	25	0,52 %
CSIRO-QCCCE	4966	150	3,02 %
DOE-COLA-CNMAP-GMU	179	0	0,00 %
FIO	2380	1634	68,66 %
ICHEC	2130	949	44,55 %
INM	3905	0	0,00 %
INPE	302	0	0,00 %
IPSL	106345	34	0,28 %
LASG-CESS	32199	2	5,60 %
LASG-IAP	10351	9058	87,51 %
MIROC	49535	0	0,00 %
MOHC	34216	16	7,40 %
MPI-M	65854	5822	8,85 %
MRI	21905	0	0,00 %
NASA-GISS	53180	47846	89,97 %
NASA-GMAO	2	0	100,00 %
NCAR	9862	2106	40,29 %
NCC	12267	480	3,91 %
NICAM	39	0	0,00 %
NIMR-KMA	521	212	41,46 %
NOAA-GFDL	291206	195	0,58 %
NSF-DOE-NCAR	6762	1706	25,23 %
UNSW	192	0	0,00 %
Total	**793304**	**106614**	**23,60 %**