

## **Access climate data in ESG from the EGI infrastructure**

Sébastien Denvil (1), Jérôme Raciarezek (2), Monique Petitdidier (3), André Gemünd (4), Horst Schwichtenberg (5)

(1) *sebastien.denvil@ipsl.jussieu.fr, IPSL, CNRS*

(2) *jerome.raciarezek@ipsl.jussieu.fr, IPSL, CNRS*

(3) *monique.petitdidier@latmos.ipsl.fr, LATMOS, IPSL, CNRS*

(4) *andre.gemuend@scai.fraunhofer.de, SCAI, Fraunhofer*

(5) *horst.schwichtenberg@scai.fraunhofer.de, SCAI, Fraunhofer*

### **Overview:**

The work of the Earth Science (ES) activity of EGI-InSPIRE SA3 is devoted to ES data access. The activity is investigating and following the development of specialized data access and management methods in the area of Earth Science. Since Grid Infrastructures have been exploited by ES, methods of how running jobs can access data, stored outside of the infrastructure, have been discussed. ESGF (Earth System Grid Federation) is an international collaboration providing access to state of the art CMIP5 climate data (Coupled Model Intercomparison Project 5). They are using a federated security infrastructure not interoperable with the EGI one. Long term solution may be based on ongoing developments of the ESG community, who is developing mechanisms to manage automatic mapping of credentials, provide data access adapters and call heterogeneous security protocols. We will present this ongoing work between EGI-Inspire and ESG in the context of an extreme events analysis of the CMIP5 simulations.

### **Enjeux scientifiques, besoin de la grille :**

La modélisation numérique est devenue la pierre angulaire des recherches développées en sciences du climat et de l'environnement. Les études des changements climatiques induits par l'activité humaine reposent sur des projections climatiques réalisées à l'aide de modèle de climat de plus en plus complexes. Ces simulations alimentent les études sur la variabilité naturelle, sur les effets des différentes perturbations anthropiques (gaz à effet de serre, aérosols, utilisation des terres...), sur les impacts des changements climatiques sur l'environnement et la société. Au niveau international une coordination s'est mise en place sous l'égide du «World Climate Research Programme» [0]. CMIP5 (Coupled Model Inter comparison Project 5) coordonne la réalisation de nouveaux jeux d'expériences. Cette coordination sert de catalyseur aux études réalisées en amont des différents rapports du GIEC (Groupe d'experts intergouvernemental sur l'évolution du climat, IPCC en anglais). Chacune de ces expériences produit d'importantes quantités de données qui sont stockées sur des centres de calcul haute performance. Ces données sont rendues visibles sur la grille de données « Earth System Grid » (ESG). L'accès à ces données est réglementé par des conditions d'utilisation.

Nous présentons le portage d'une application parallèle et « data intensive » permettant l'étude multi modèles des extrêmes climatiques sur la grille EGI, qui semble adaptée pour cela. Cette application utilise les données CMIP5 et des jeux de données d'observation permettant une correction statistique des biais systématiques des modèles. Ce travail soutient les objectifs de l'activité SA3.6 WP6 du projet EGI-Inspire, visant à permettre l'utilisation de la grille EGI dans le domaine des Sciences de la Planète (ES). Il décrit également une approche permettant d'accéder et d'utiliser les données de la grille ESGF (Earth Science Grid Federation) depuis la grille EGI (European Grid Infrastructure). L'objectif principal est de permettre aux futurs utilisateurs de données CMIP5, pour des études régionales et d'impact, d'utiliser simplement la grille.

### **Développements, déploiement sur la grille :**

L'application, STRDX, est un code parallèle (MPI) écrit en fortran90 qui analyse des données telles que celles de CMIP5. C'est une application de type « data intensive ». Dans le contexte CMIP5 elle cumulera en entrée 10 To de données et en produira 12 To répartie en 600 jobs ; chacun durant environ 2 heures, occupant 16 processeurs et manipulant en moyenne 37 Go de données. L'application utilise le système de compilation automatisé de l'IPSL, MODIPSL, qui a du être modifié pour son utilisation sur la grille. L'application fait appel à des bibliothèques (netCDF,

PnetCDF et IOIPSL) qui ont été installées dans l'espace « logiciel » de l'organisation virtuelle ESR. La figure 1 montre le déroulement de l'exécution de l'application. Les données d'entrées du programme sont transférées sur l'UI, puis copiées dans le système de stockage EGI [3], afin de permettre aux nœuds de travail d'y accéder. L'accès aux données ESGF étant restreint [1] l'utilisateur doit être muni, en plus de l'habituel certificat EGI, d'un second certificat délivré par ESGF. Ce certificat est récupéré depuis le SLCS (Short Live Credential Service) ESGF avec la commande myproxy-logon [2]. A la fin du run, les données de sortie du programme sont transférées dans le système de stockage EGI. Ces données sont ensuite récupérées sur l'UI pour être consultées et analysées avec des outils de visualisation (Ferret, CDAT).

### Outils, difficultés rencontrées :

Les données ESGF sont disponibles via les protocoles HTTP, GridFTP, et OpenDAP. Coté EGI, un client OpenDAP doit être installé si l'on veut utiliser ce protocole. Les métadonnées ESGF peuvent être consultées via des interfaces utilisateurs Web, ou sous forme de fichiers XML appelés "Thredds catalogs" [4]. Les fédérations ESGF et EGI utilisent toutes deux les infrastructures réseaux dédiées à la recherche et à l'éducation (Renater, Geant, Internet2..). La bande passante dépend du site EGI où tournent les jobs et de l'emplacement du nœud de données ESGF.

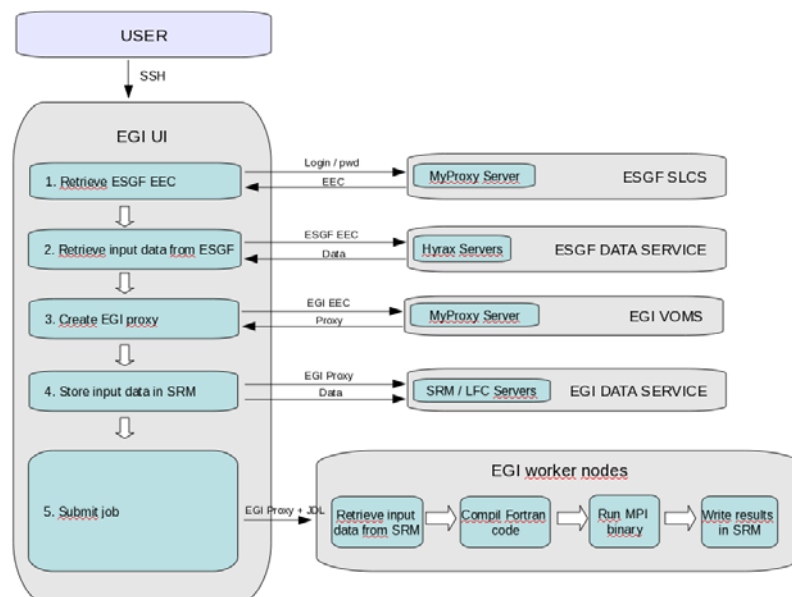


Figure 1 Porting of the STRDX application on the grid / Portage de l'application STRDX sur la grille

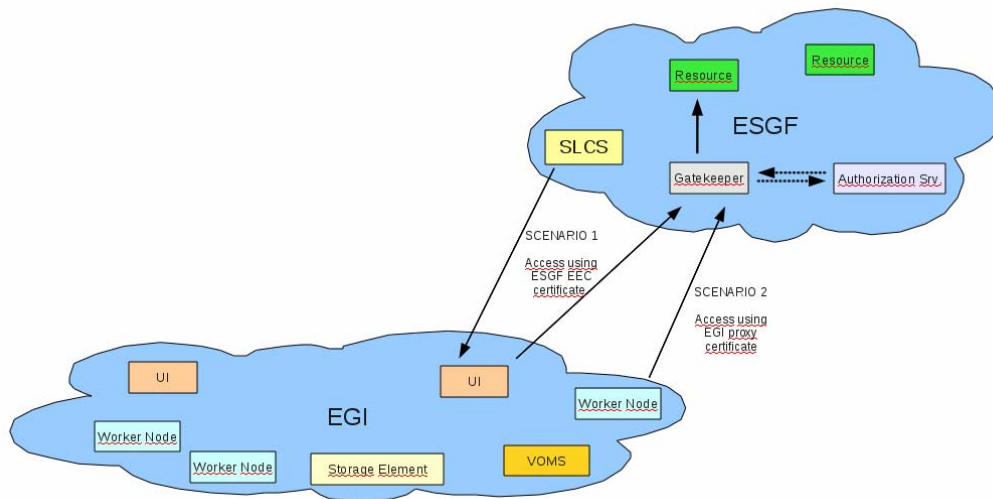
Les tests réalisés entre le site EGI/INSU et le nœud de données Canadien montre un taux de transfert de 32Mo/s pour 160 Go transférés. Le cumul de temps nécessaire au mouvement des 22 To d'E/S sera de 8 jours si ce taux est représentatif.

Un des principaux problèmes rencontrés concerne l'authentification et l'autorisation des utilisateurs dans un contexte d'inter-fédération. Les certificats de type Proxy [5] utilisés côté EGI ne sont pas acceptés dans la fédération ESGF. Les autorités de certification accréditées par EGI [6] ne sont pas les mêmes que celles accréditées par ESGF [7]. En conséquence, l'authentification des utilisateurs EGI dans la fédération ESGF n'est pour l'instant pas possible. De la même manière, l'interopérabilité au niveau des autorisations est limitée, car la fédération ESGF est actuellement la seule à pouvoir attribuer les rôles donnant accès aux données climatiques.

### Résultats scientifiques :

Une des manières de traiter la sensibilité des extrêmes au changement climatique est de considérer des événements dits sévères plutôt qu'extrêmes. Ces événements sévères, s'ils

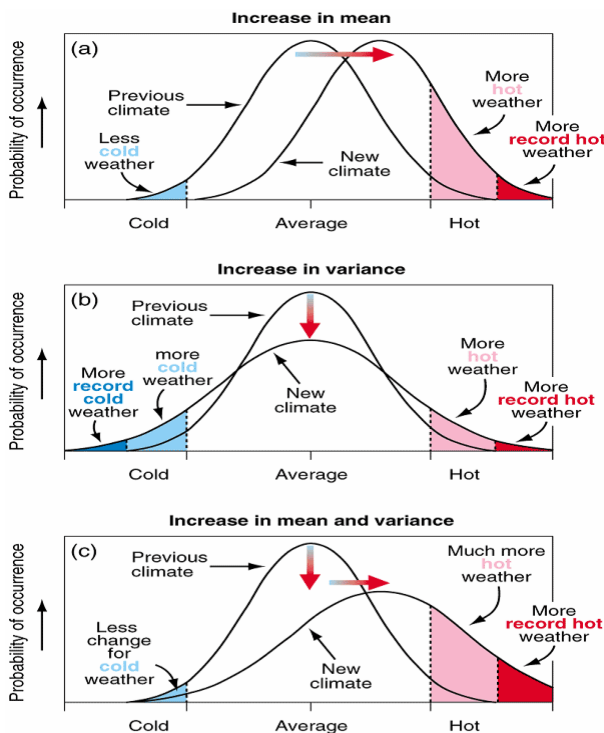
représentent des conditions éloignées du climat moyen, ont l'avantage de se produire plus souvent que les événements extrêmes. Par la suite, le terme extrême désignera les événements sévères du système climatique. C'est le parti pris du projet STARDEX (STATistical and Regional



**Figure 2 Interfederation interoperability scenario / Scénario d'interopérabilité inter-fédération**

dynamical Downscaling of EXtremes for European regions) qui a défini 57 indices associés à des conditions sévères de température et de précipitation. Frich et al (2002) ont documenté le changement de 10 de ces indices au cours de la moitié du 20<sup>e</sup> siècle. Tebaldi et al (2006) ont évalué le signe de leurs tendances actuelles (40 dernières années) et futures (au cours 21<sup>e</sup> siècle) ainsi que les caractéristiques spatiales des modèles de climat ayant participé au dernier exercice CMIP3. Ils se sont concentrés sur la significativité statistique des tendances ; leur accord aux tendances observées ; l'accord entre les modèles.

**Figure 3 Change in mean and in variance / Modification de la moyenne et de la variance**

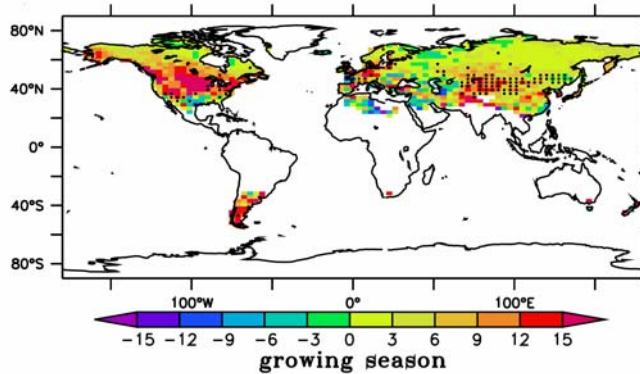


Quel que soit le modèle de climat utilisé, il est une simplification d'une réalité complexe. On ne peut pas s'attendre à ce que les résultats d'un modèle représentent exactement un échantillon de données observées. Comme nous nous intéressons ici à des événements extrêmes, nous nous concentrons sur des données quotidiennes (des températures minimales, maximales, moyennes et des cumuls de précipitation). Déqué (2007) documente cinq techniques de corrections statistiques qui corrigent les biais systématiques des modèles en utilisant des données observées ; en particulier la méthode de corrections de variable quantile-quantile utilisée ici.

Cette technique consiste à construire une table de correction avec les 99 centiles des deux distributions (période commune simulation de référence et observation) pour chaque modèle, variable, saison et point de grille. Une interpolation linéaire est appliquée entre deux centiles. En dehors de la page de la table, une

correction constante est appliquée: si le quantile 99e température maximale d'été est corrigée par  $K + 1$ , toute température au-dessus de ce seuil est corrigée par  $K + 1$  pour cette saison et ce point de grille. La fonction densité de probabilité des données simulées est ainsi modifiée (figure 3).

**Figure 4 Evolution of the growing season length (days) during 20th century from the IPSLCM4 climate model. (99% significance marked)/Evolution de la période de croissance des plantes (jours) au cours du 20<sup>e</sup> siècle d'après le modèle de climat IPSLCM4. (99% de significativité marqué par un point).**



Les indices synthétiques d'événements extrêmes sont calculés pour les simulations non corrigées, et pour les simulations corrigées. Ils sont associés à des valeurs de significativité statistique de leur tendance (figure 4). On peut ainsi vérifier la sensibilité de la significativité des tendances à la correction (avec ou sans correction, fonction du jeu d'observation). Ces simulations corrigées sont utiles pour nourrir les modèles d'impact sensibles aux valeurs de seuil. La correction ne réduit pas, et peut augmenter l'incertitude sur les

projections climatiques. L'utilisation de plusieurs modèles et de plusieurs scénarios est la technique appropriée pour faire face à l'incertitude.

### Perspectives :

Nous avons vu dans ce document que l'utilisation des données ES par la grille EGI est possible, mais cela nécessite actuellement d'utiliser deux identités différentes, et donc deux certificats différents. Cette contrainte complexifie considérablement la tâche pour l'utilisateur final et constitue le frein principal pour une bonne interopérabilité entre les deux grilles. Cela a été mis en évidence avec le portage de l'application STRDX, où les données transitent par l'UI EGI au lieu d'aller directement sur les noeuds de travail (scénario 1 fig. 2). Ce mode de fonctionnement pourrait être simplifié si les noeuds EGI étaient capables d'accéder eux-mêmes à la fédération ESGF pour y chercher les données dont ils ont besoin. Pour rendre cela possible, une solution serait que les noeuds de travail puissent s'authentifier dans la fédération ESGF directement avec le Proxy EGI, sans devoir générer un certificat ESGF (scénario 2 fig. 2). Un certain nombre de modifications sont cependant nécessaires avant de permettre ce nouveau mode de fonctionnement: la fédération EGI doit pouvoir attribuer à ses utilisateurs des rôles équivalents à ceux d'ESGF, la fédération ESGF doit accepter les certificats des utilisateurs EGI, et les utilisateurs EGI doivent être informés des conditions d'utilisation des données ESGF lorsqu'ils rejoignent la VO ESR avec accès aux données d'ESGF.

### References:

- [0] World Climate Research Programme : <http://wcrp.wmo.int/wcrp-index.html>
- [1] CMIP5 Data Access Guide [http://cmip-pcmdi.llnl.gov/cmip5/data\\_getting\\_started.html](http://cmip-pcmdi.llnl.gov/cmip5/data_getting_started.html)
- [2] MyProxy Credential Management Service <http://grid.ncsa.illinois.edu/myproxy/>
- [3] DPM (Disk Pool Manager) <https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm>.
- [4] Thredds Catalog  
<http://www.unidata.ucar.edu/projects/THREDDS/tech/catalog/InvCatalogSpec.html>
- [5] Proxy Certificate <http://www.ietf.org/rfc/rfc3820.txt>
- [6] EGI Trust root [https://wiki.egi.eu/wiki/EGI\\_IGTF\\_Release](https://wiki.egi.eu/wiki/EGI_IGTF_Release)
- [7] ESGF Trust Roots [http://esgf.org/esgf-certs/#ESG\\_Federation\\_Trust\\_Roots](http://esgf.org/esgf-certs/#ESG_Federation_Trust_Roots)
- [8] Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. Klein Tank, and T. Peterson, Global changes in climatic extremes during the 2nd half of the 20th century, *Clim. Res.*, 19, 193–212, 2002.
- [9] Tebaldi et al, 2006: Going to the extremes; An intercomparison of model-simulated historical and future changes in extreme events, *Climatic Change*, 79, 185-211
- [10] Chauvin F., Denvil S. 2007: Changes in severe indices as simulated by two French coupled global climate models. *Global and Planetary Change*, Volume 57, Issue 1-2, p. 96-117. 2007

[11] Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: model results and statistical correction according to observed values. *Global and Planetary Change*, 57, 16-26.