

Gestion et manipulation de données en Climatologie

Benjamin POHL & Thierry CASTEL

*CRC / Biogéosciences,
CNRS / uB, Dijon, France*

benjamin.pohl@u-bourgogne.fr
thierry.castel@u-bourgogne.fr

21 Mars 2013



BIOGÉOSCIENCES



La production de données au CRC : les modèles de climat

HISTORIQUE RAPIDE

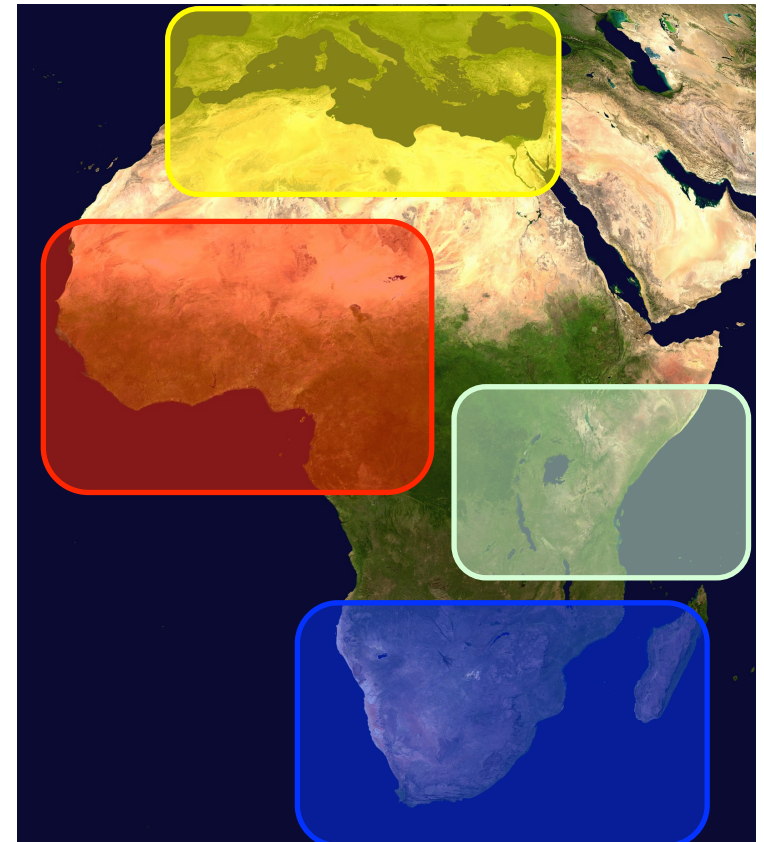
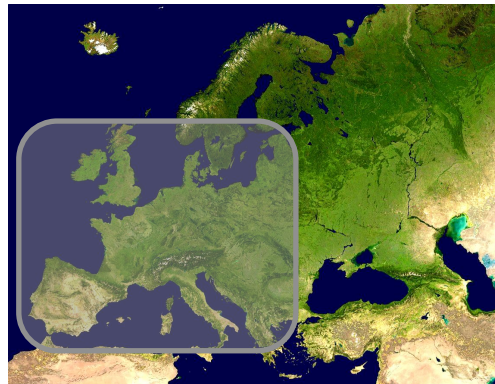
2004 : MM5 implémenté au CCuB (Sijikumar & Roucou)

2007 : WRF (Castel & Roucou)

2010 : Arpege-Climat (Al-Sayed-Ali, Castel & Pohl)

2012 : fusion CRC – Biogéosciences, création du "Pôle de Simulations Climatiques"

FOCUS : essentiellement (mais pas uniquement) sur le fuseau eurafricain & l'océan Indien



PRINCIPAUX UTILISATEURS

Permanents : 1 uB, 1 AgroSup, 1 CNRS (+1 ingé calcul contractuel CNRS)

Thèsard(e)s : 5 (dont 2 dont c'est le but exclusif)

Post-docs : 3

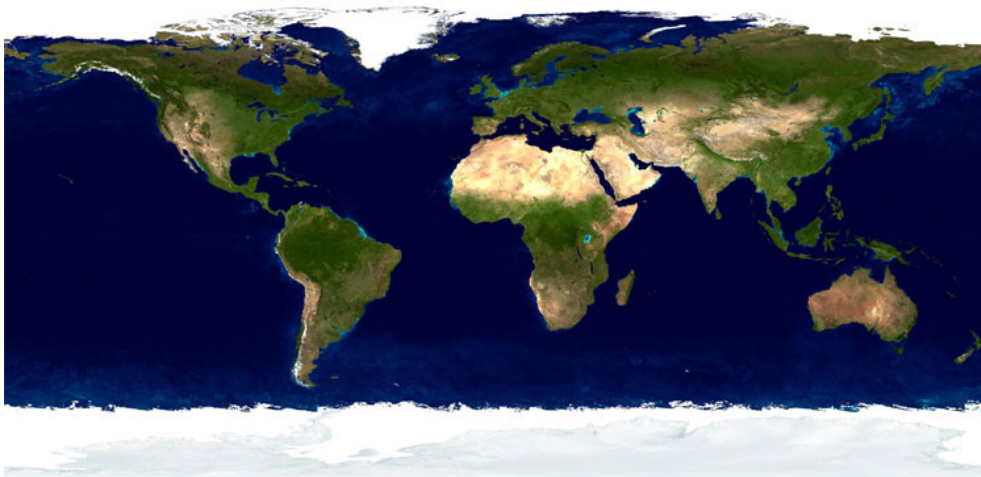
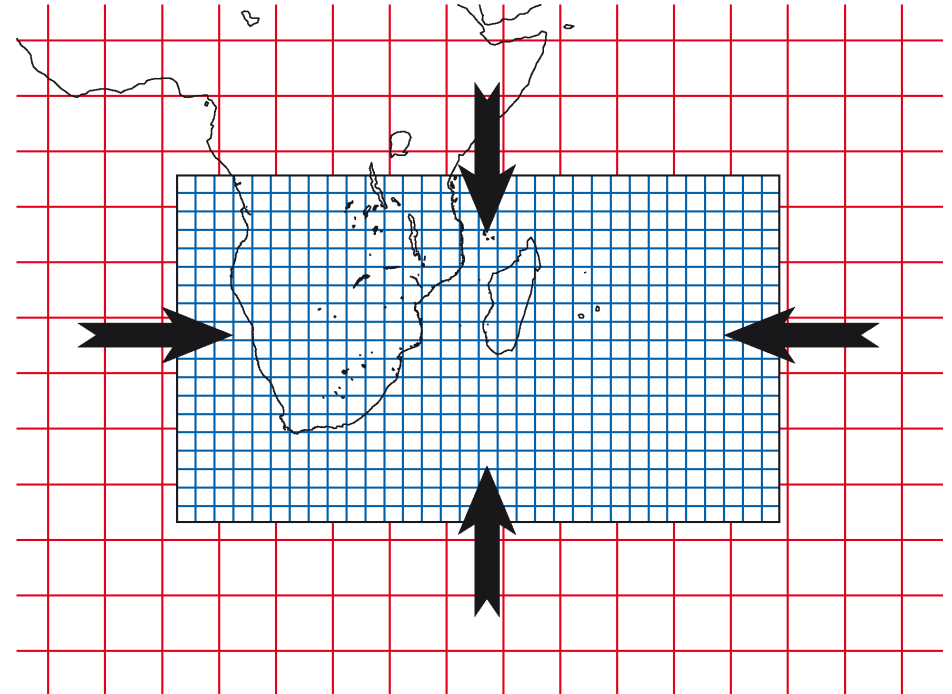
Extérieurs : 1 doctorant univ. Tunis, 1 MCF univ. Réunion, 1 doctorante UPMC

Etudiants Master II Géobiosphère (une journée de formation chaque année)

La production de données au CRC : les modèles de climat

MM5, WRF : modèles à aire limitée haute résolution (forcés latéralement par des modèles globaux basse résolution)

- pas de limite *théorique* à la résolution spatio-temporelle
- coûteux en ressources CPU

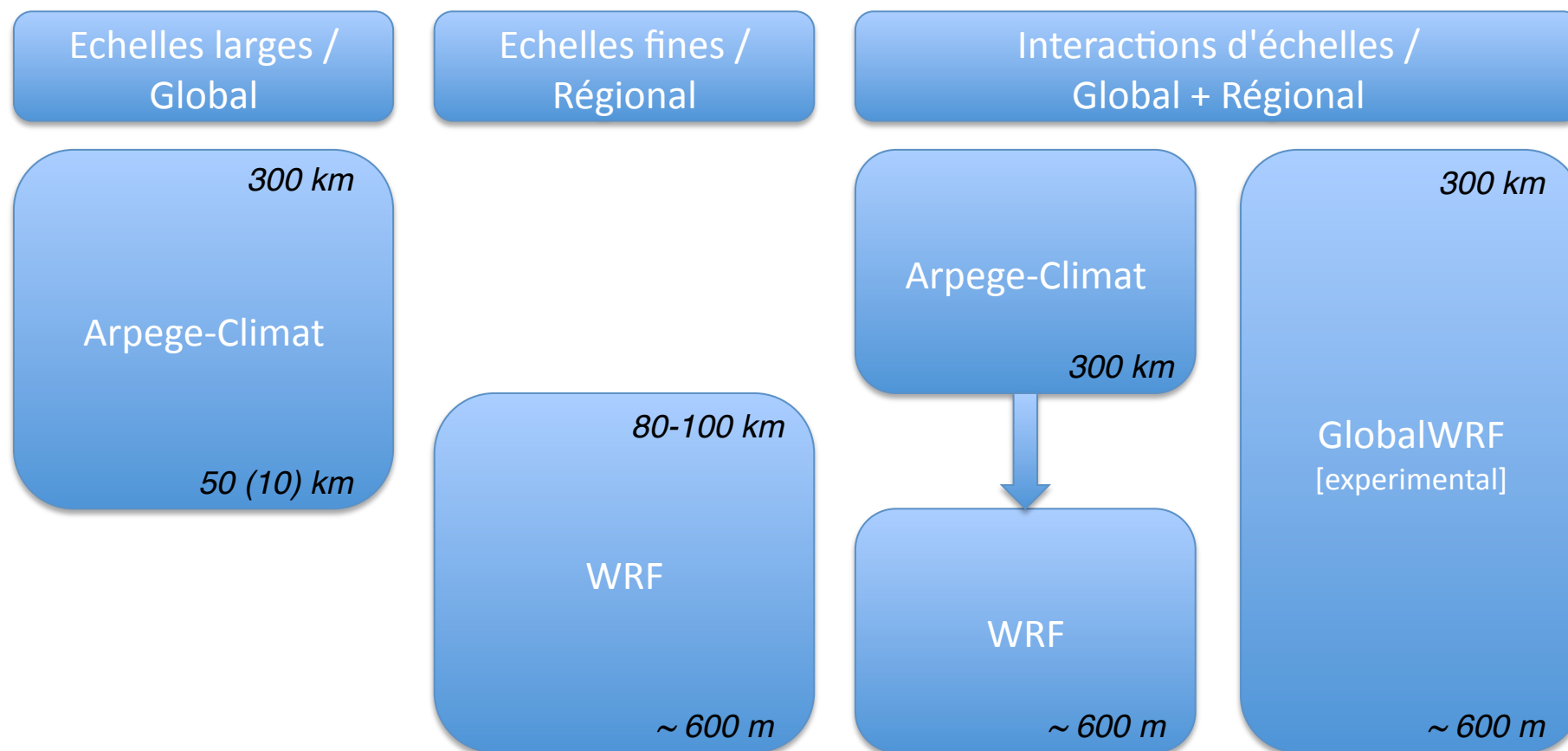


Arpege-Climat : modèle de climat global basse résolution (< 10km)

- code moins gourmand
- limites dues à des approximations valables sur les seules grandes échelles

La production de données au CRC : les modèles de climat

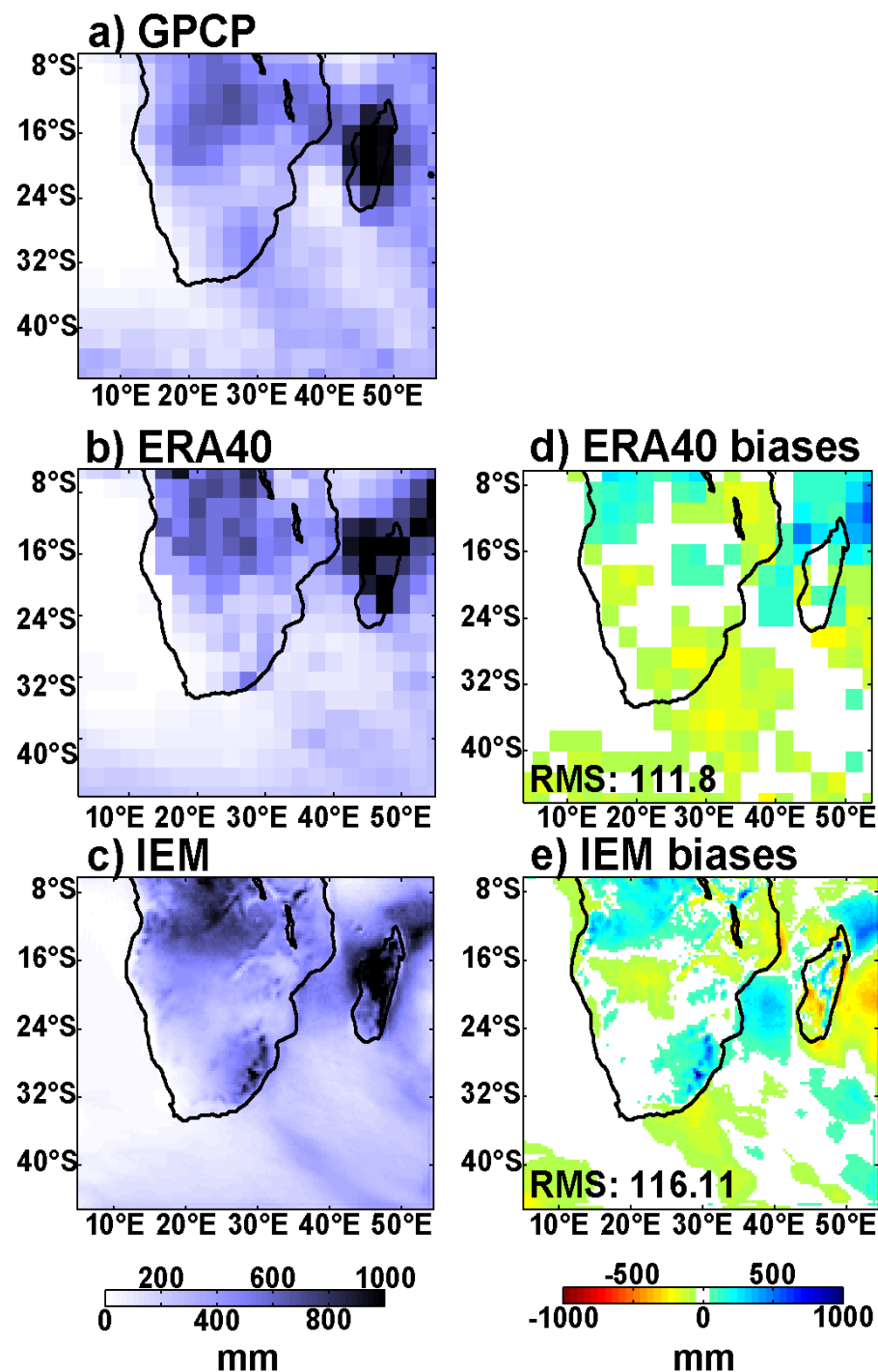
Chaine de modélisation numérique du climat au CRC



Quelques résultats (qui font plaisir)

Spatialisation du champ de pluie en Afrique Australe...

- pluies tropicales : pas facile !
- biais du même ordre de grandeur que les meilleurs modèles...
- ... la résolution en plus !



Quelques résultats (qui font plaisir)

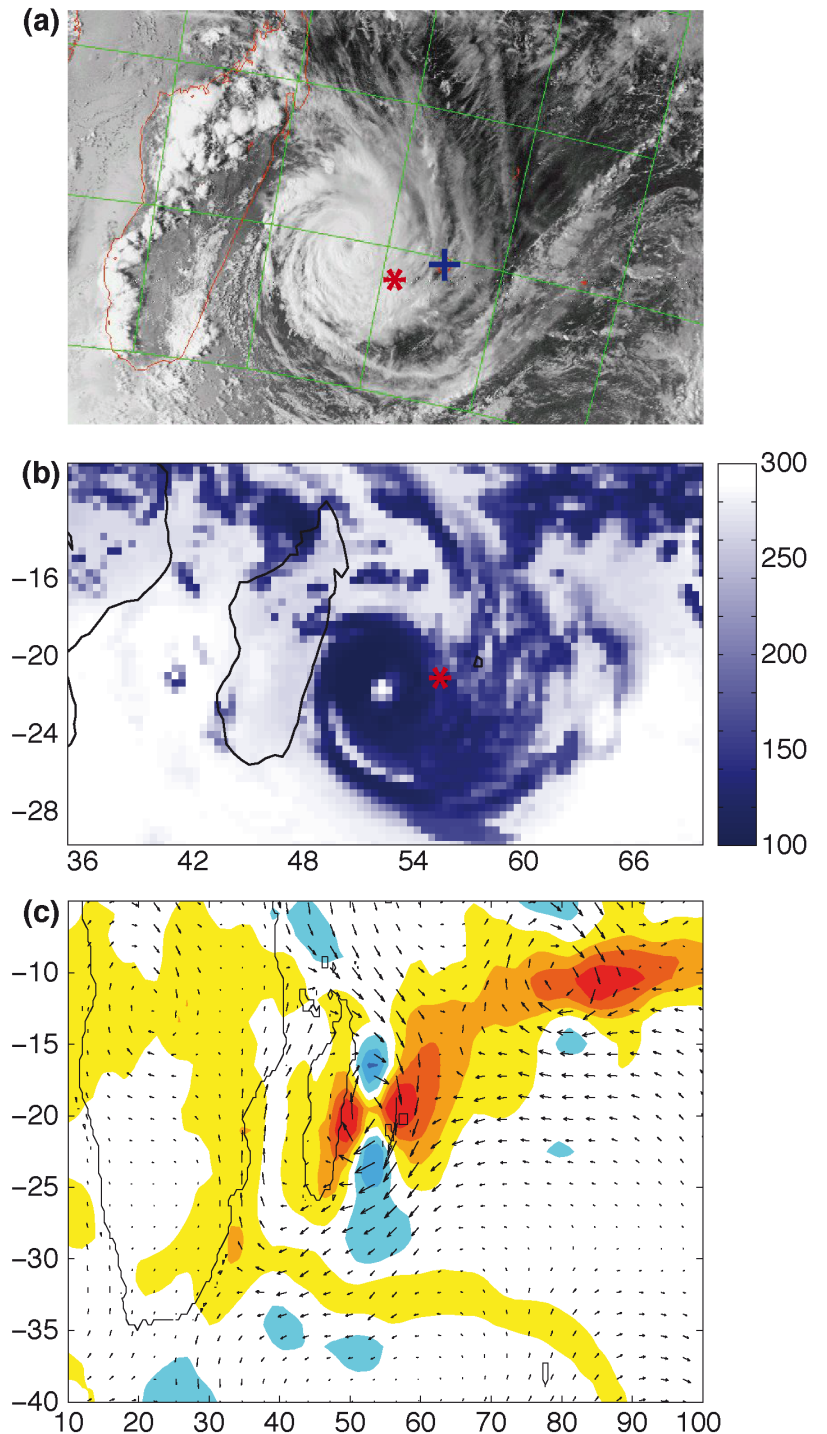
2-11 janvier 2001: TC Ando

- TC Catégorie 5
- 922hPa, rafales 270km/h, passé à 215km de la Réunion
- dans WRF le nudging appliqué sur le domaine #1 assure un timing et une trajectoire réalistes

(a) NOAA-14, canal visible, 06/01/01 17h04 UTC

(b) OLR WRF, 07/01/01 0h00 UTC

(c) Flux d'humidité intégrés 1000-700hPa : Ando et son petit frère Bindu

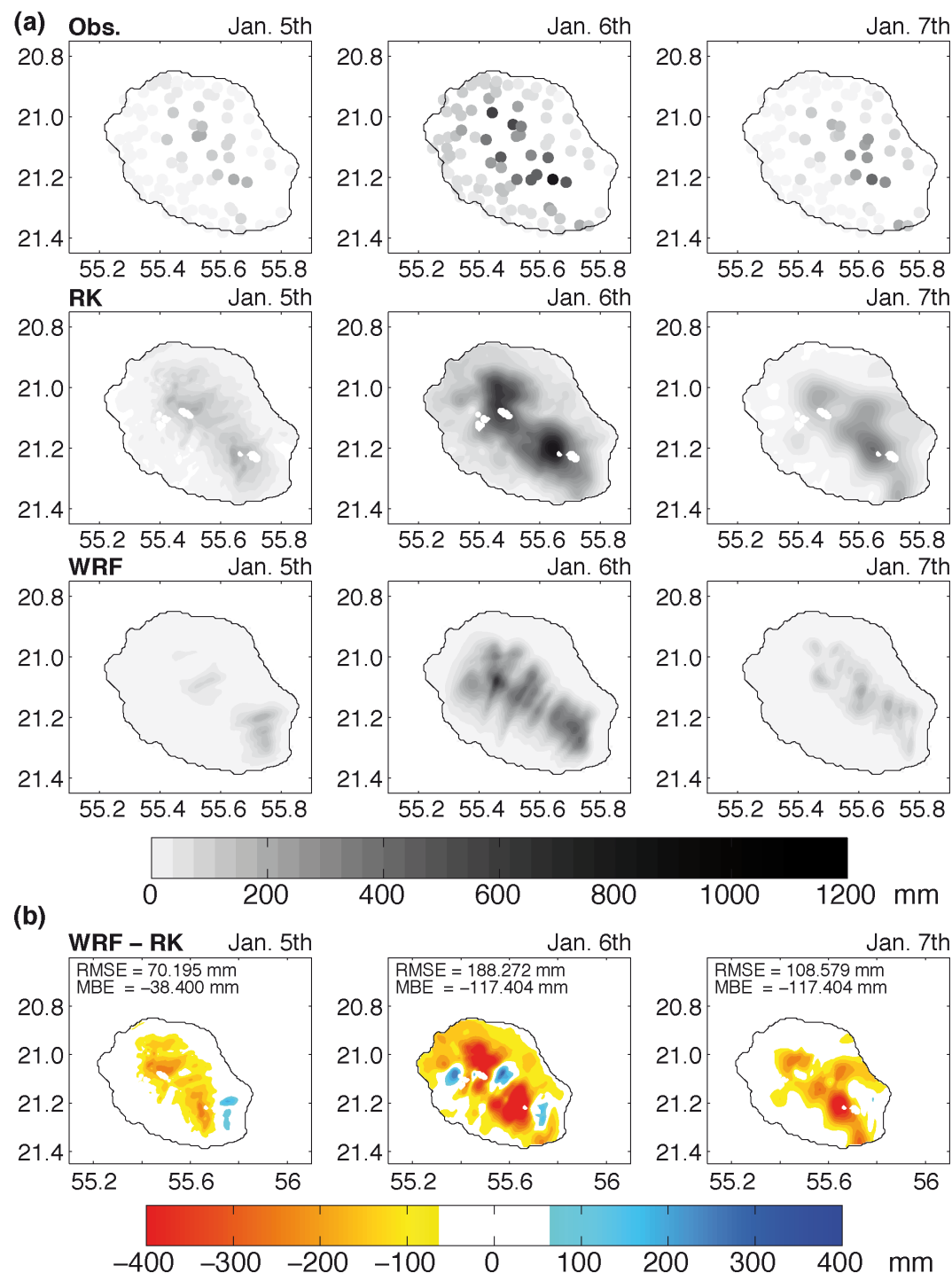


Quelques résultats (qui font plaisir)

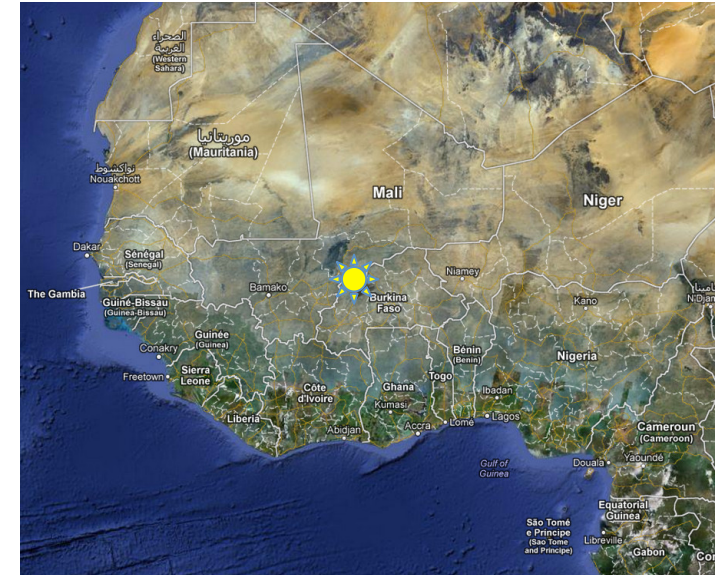
2-11 janvier 2001: TC Ando

Effets sur la Réunion: 6 janvier

**"Les deux méthodes de downscaling
semblent capables de combler le
saut d'échelle terrain / situation
synoptique".**



La production de données au CRC : les réseaux d'observation



Depuis mars 2012: **mesure d'aérosols à Dédougou (Burkina)**

- une mesure de concentrations de PM10 toutes les 5' (aérosols désertiques), 24/24, 365 j/an
- liens aérosols / climat / méningites

Piloté par le CRC dans le cadre d'AMMA



La production de données au CRC : les réseaux d'observation

[illegible]

24H de données : ça marche ! ...

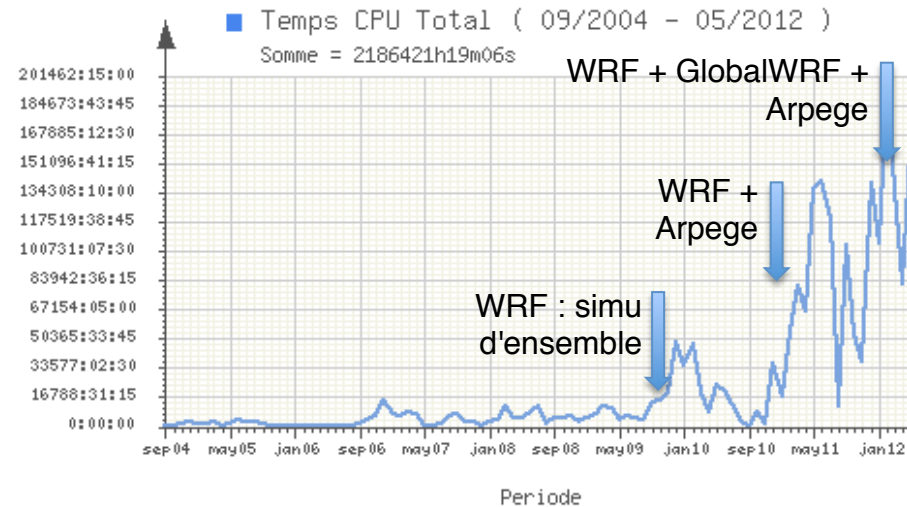
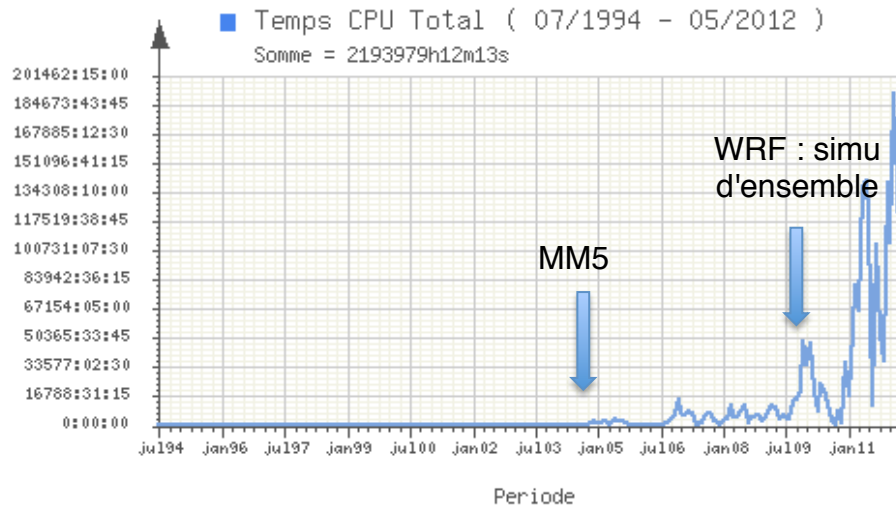
La production de données au CRC : les réseaux d'observation



24H de données : ça marche ! ...

... non sans quelques difficultés parfois ...

Stockage et calcul



Consommation calcul au CCuB

1994 – 2012 : $\sim 2 \times 10^6$ heures CPU au CCuB (+ ressources extérieures : IDRIS, CRIHAN)

1994 – 2011 : 10^6 heures en 17 ans

2011 – 2012 : 2^{ème} million atteint en moins de 18 mois, 3^{ème} en 5 mois, 4^{ème} en 3 mois

Mars 2013 : environ 4.5×10^6 h CPU maintenant



Contribution au CCuB

Nov. 2004 : une machine de calcul de 2 cœurs

Nov. 2009 : un serveur Sun de 24To brut

Juin 2011 : 96 cœurs de calcul (2 Dell C6100)

Juin 2012 : un serveur Dell R510 de 36To brut

Décembre 2012 : 64 cœurs de calcul (1 Dell C6220)

Avril 2013? : un serveur Dell MD1200 de 48To bruts

... to be continued ...

Gestion de la donnée : le format NetCDF

.nc



.grb

```
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared]
$ ls
20thRean  ECHAM45      GPCC          MODIS_VGT  NCEP_NCAR  OLR  RESULT_SarraH  SRTM30_V2_AFRICA  WRF331_WestAfrica
CMIP5     ERA_INTERIM  Index_climato monde.mat  NDVI_ADDS  RCP26  RESULT_SarraH2  sst.mnmean.nc    WRF_BOURGOGNE
DUST      GlobalWRF331 master       NCEP2     OAFIux    RCP85  SLP_KNMI_1950_2100_ECHAM  WRF321
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared]
$ cd WRF321/AfrS
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS]
$ ls
1971 1973 1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 AExperiments
1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 README_WRF321_AFRS
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS]
$ cd 1993
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS/1993]
$ ls
B93_23 E93_02 E93_05 E93_08 E93_11 E93_14 E93_17 E93_20 E93_23 E93_26 E93_29 F93_23 P93_23
C93_23 E93_03 E93_06 E93_09 E93_12 E93_15 E93_18 E93_21 E93_24 E93_27 E93_30 K93_23 S93_23
E93_01 E93_04 E93_07 E93_10 E93_13 E93_16 E93_19 E93_22 E93_25 E93_28 EI_93  M93_23 YR9394
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS/1993]
$ cd EI_93
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS/1993/EI_93]
$ ls
EI_93.GLW.dly.nc EI_93.LH.dly.nc EI_93.PSFC.dly.nc EI_93.SFROFF.dly.nc EI_93.tc.dly.nc EI_93.V10.dly.nc
EI_93.GRDFLX.dly.nc EI_93.OLR.dly.nc EI_93.QFX.dly.nc EI_93.SST.dly.nc EI_93.TSK.dly.nc EI_93.va.dly.nc
EI_93.HFX.dly.nc EI_93.PBLH.dly.nc EI_93.QVAPOR.dly.nc EI_93.SWDOWN.dly.nc EI_93.U10.dly.nc EI_93.wa.dly.nc
EI_93.HGT.dly.nc EI_93.POTEVP.dly.nc EI_93.rain.dly.nc EI_93.T2.dly.nc EI_93.ua.dly.nc EI_93.z.dly.nc
be0033po@krenek01 CentOS_Intel64_5.2 [/archive/crct/shared/WRF321/AfrS/1993/EI_93]
$
```


Gestion de la donnée : le format NetCDF

```
netcdf EI_93.wa.dly {
dimensions:
    lon = 243 ;
    lat = 149 ;
    lev = 10 ;
    time = UNLIMITED ; // (182 currently)
variables:
    float lon(lon) ;
        lon:long_name = "longitude" ;
        lon:units = "degree_east" ;
        lon:standard_name = "longitude" ;
        lon:axis = "X" ;
    float lat(lat) ;
        lat:long_name = "latitude" ;
        lat:units = "degree_north" ;
        lat:standard_name = "latitude" ;
        lat:axis = "Y" ;
    float lev(lev) ;
        lev:long_name = "Pressure Levels" ;
        lev:units = "hPa" ;
        lev:axis = "Z" ;
    double time(time) ;
        time:units = "hours since 1993-10-01 00:00:00" ;
        time:calendar = "standard" ;
    float wa(time, lev, lat, lon) ;
        wa:units = "m/s" ;
        wa:FillValue = -999999.f ;
        wa:description = "Vertical Velocity w" ;

// global attributes:
    :CDI = "Climate Data Interface version 1.4.7 (http://code.zmaw.de/projects/cdi)" ;
    :Conventions = "None" ;
    :history = "Wed Mar 21 13:00:50 2012: cdo -r daymean /work/crcr/db124299/WRF321/EI_93/EI_93.wa.nc /work/crcr/db124299/WRF321/EI_93/EI_93.wa.dly.nc" ;
    :creation_date = "mer mar 21 13:00:36 CET 2012" ;
    :source_file = "/work/crcr/db124299/WRF321/EI_93/wrfout_d01.nc" ;
    :title = "WRF v32 output - (c) CRC 2011" ;
    :CDO = "Climate Data Operators version 1.4.7 (http://code.zmaw.de/projects/cdo)" ;

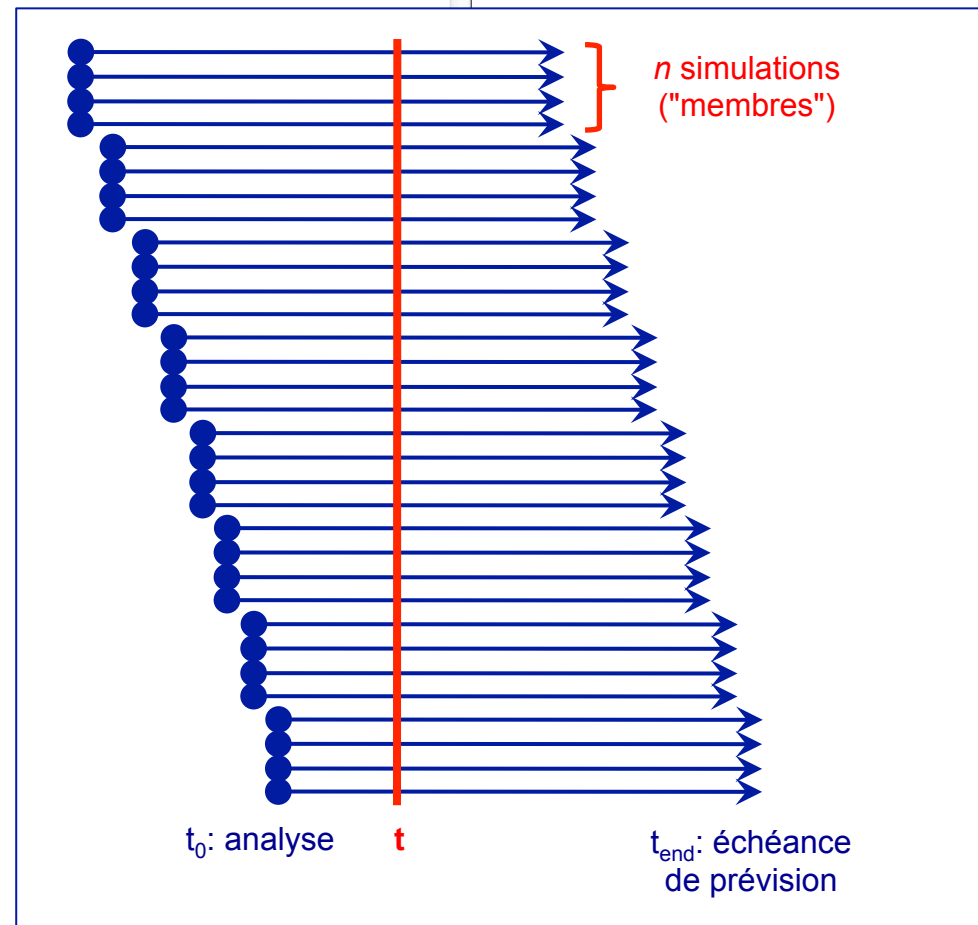
data:
lon = 0.4826355, 0.8092041, 1.135803, 1.462402, 1.788971, 2.11557, 2.442169,
2.768768, 3.095337, 3.421936, 3.748535, 4.075104, 4.401703, 4.728302,
5.054871, 5.38147, 5.708069, 6.034668, 6.361237, 6.687836, 7.014435,
7.341003, 7.667603, 7.994202, 8.32077, 8.647369, 8.973969, 9.300568,
9.627136, 9.953735, 10.28033, 10.6069, 10.9335, 11.2601, 11.58667,
11.91327, 12.23987, 12.56647, 12.89304, 13.21964, 13.54623, 13.8728,
14.1994, 14.526, 14.8526, 15.17917, 15.50577, 15.83237, 16.15894,
16.48553, 16.81213, 17.1387, 17.4653, 17.7919, 18.1185, 18.44507,
18.77167, 19.09827, 19.42484, 19.75143, 20.07803, 20.4046, 20.7312,
```

```
$ ncdump EI_93.wa.dly.nc | more
```

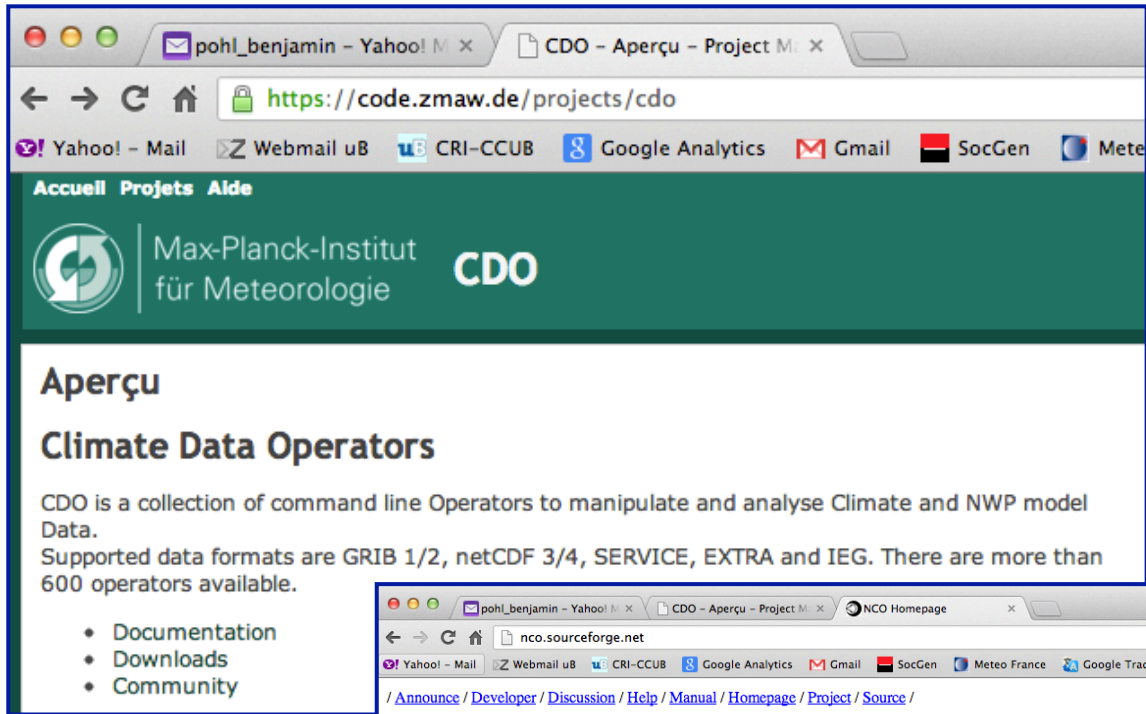
Gestion de la donnée : le format NetCDF

```
netcdf u_20001229_20010112 {
dimensions:
    time = UNLIMITED ; // (15 currently)
    lat = 31 ;
    lon = 41 ;
    ens = 11 ;
    fhour = 23 ;
    pressure = 11 ;
variables:
    double time(time) ;
        time:long_name = "Time" ;
        time:units = "hours since 1800-01-01 00:00:00" ;
        time:axis = "T" ;
    int intTime(time) ;
        intTime:long_name = "time as an integer (YYYYMMDDHH)" ;
    float lat(lat) ;
        lat:long_name = "Latitude" ;
        lat:standard_name = "latitude" ;
        lat:units = "degrees_north" ;
        lat:actual_range = -35.f, -5.f ;
        lat:axis = "Y" ;
    float lon(lon) ;
        lon:long_name = "Longitude" ;
        lon:standard_name = "longitude" ;
        lon:units = "degrees_east" ;
        lon:actual_range = 40.f, 80.f ;
        lon:axis = "X" ;
    short ens(ens) ;
        ens:long_name = "ensemble" ;
        ens:standard_name = "ensemble" ;
        ens:axis = "E" ;
        ens:description = "0 is control, other values are perturbation numbers" ;
    int fhour(fhour) ;
        fhour:long_name = "forecast hour" ;
        fhour:units = "hours" ;
    int intValidTime(time, fhour) ;
        intValidTime:long_name = "valid time as an integer (YYYYDDMMHH)" ;
    float pressure(pressure) ;
        pressure:long_name = "Isobaric surface" ;
        pressure:units = "hPa" ;
        pressure:positive = "down" ;
        pressure:GRIB_level_type = "100" ;
    float U-component_of_wind(time, ens, fhour, pressure, lat, lon) ;
        U-component_of_wind:FillValue = 9999.f ;
        U-component_of_wind:units = "m s-1" ;
        U-component_of_wind:long_name = "U-component_of_wind @ pressure" ;
        U-component_of_wind:GRIB_param_discipline = "Meteorological_products" ;
        U-component_of_wind:GRIB_param_category = "Momentum" ;
        U-component_of_wind:GRIB_param_name = "U-component_of_wind" ;
        U-component_of_wind:GRIB_generating_process_type = "Forecast" ;
        U-component_of_wind:GRIB_param_id = 2, 0, 2, 2 ;
        U-component_of_wind:GRIB_product_definition_template = 0 ;
        U-component_of_wind:GRIB_product_definition_template_desc = "Analysis/forecast at horizontal level/layer at a point in time" ;
        U-component_of_wind:GRIB_level_type = 100 ;
        U-component_of_wind:GRIB_level_type_name = "pressure" ;
        U-component_of_wind:GRIB_VectorComponentFlag = "easterlyNortherlyRelative" ;

// global attributes:
    :Conventions = "CF-1.0" ;
    :title = "Subset of data from 2nd-generation multi-decadal ensemble reforecast generated from the NCEP Global Ensemble Forecast System, mim-
    icking version operational at NCEP/EMC circa mid-2012." ;
    :institution = "NOAA Earth System Research Laboratory (ESRL)" ;
    :source = "NCEP GFS v 9.01, T254L42. Control initial conditions from CFSRR. Perturbed initial conditions from ETR. Model error simulated
    with STTP." ;
    :references = "http://www.esrl.noaa.gov/psd/forecasts/reforecast2/index.html" ;
    :history = "Subset created 2013-03-08 15:15:07 UTC" ;
}
```



Gestion de la donnée : le format NetCDF



A screenshot of a web browser showing the CDO (Climate Data Operators) website. The browser tabs include 'pohl_benjamin - Yahoo! Mail' and 'CDO - Aperçu - Project M...'. The address bar shows 'https://code.zmaw.de/projects/cdo'. The website header features the Max-Planck-Institut für Meteorologie logo and the text 'CDO'. Below the header, the page title is 'Aperçu' and the main heading is 'Climate Data Operators'. The text describes CDO as a collection of command line operators for manipulating and analyzing climate and NWP model data, listing supported formats: GRIB 1/2, netCDF 3/4, SERVICE, EXTRA, and IEG. A list of links is provided: Documentation, Downloads, and Community.

Accueil Projets Aide

Max-Planck-Institut für Meteorologie CDO

Aperçu

Climate Data Operators

CDO is a collection of command line Operators to manipulate and analyse Climate and NWP model Data.
Supported data formats are GRIB 1/2, netCDF 3/4, SERVICE, EXTRA and IEG. There are more than 600 operators available.

- Documentation
- Downloads
- Community



A screenshot of the netCDF Operator (NCO) Homepage. The browser tabs include 'pohl_benjamin - Yahoo! Mail', 'CDO - Aperçu - Project M...', and 'NCO Homepage'. The address bar shows 'nco.sourceforge.net'. The page title is 'Welcome to the netCDF Operator (NCO) Homepage'. The text states the current NCO version is 4.2.5, last updated on Tuesday, 05-Mar-2013 19:37:11 UTC. It describes NCO as a tool for manipulating data stored in netCDF format, exploiting geophysical expressivity of CF (Climate & Forecast) metadata conventions, flexible description of physical dimensions translated by UDUnits, network transparency of OPeNDAP, storage features (e.g., compression, chunking, groups) of HDF (the Hierarchical Data Format), and powerful mathematical and statistical algorithms of GSL (the GNU Scientific Library). A list of links is provided: News & Milestones, What is NCO?, Contributing, NCO→SDO Project, Publications, Version Summary, Binaries, Documentation, FAQ, README/ChangeLog/TODO, Source Code, Compiling, UCI and NCAR, Benchmarks, Known Problems, People, and Contact us. The page features a large graphic with the text 'NCO = Σ' and logos for netCDF, CF, OPeNDAP, HDF, and GSL.

Welcome to the netCDF Operator (NCO) Homepage

Current NCO version is 4.2.5 last updated on Tuesday, 05-Mar-2013 19:37:11 UTC

NCO manipulates data stored in [netCDF](#) format. It also exploits the geophysical expressivity of many [CF](#) (Climate & Forecast) metadata conventions, the flexible description of physical dimensions translated by [UDUnits](#), the network transparency of [OPeNDAP](#), the storage features (e.g., compression, chunking, groups) of [HDF](#) (the Hierarchical Data Format), and many powerful mathematical and statistical algorithms of [GSL](#) (the GNU Scientific Library). NCO is [fast](#), [powerful](#), and [free](#).

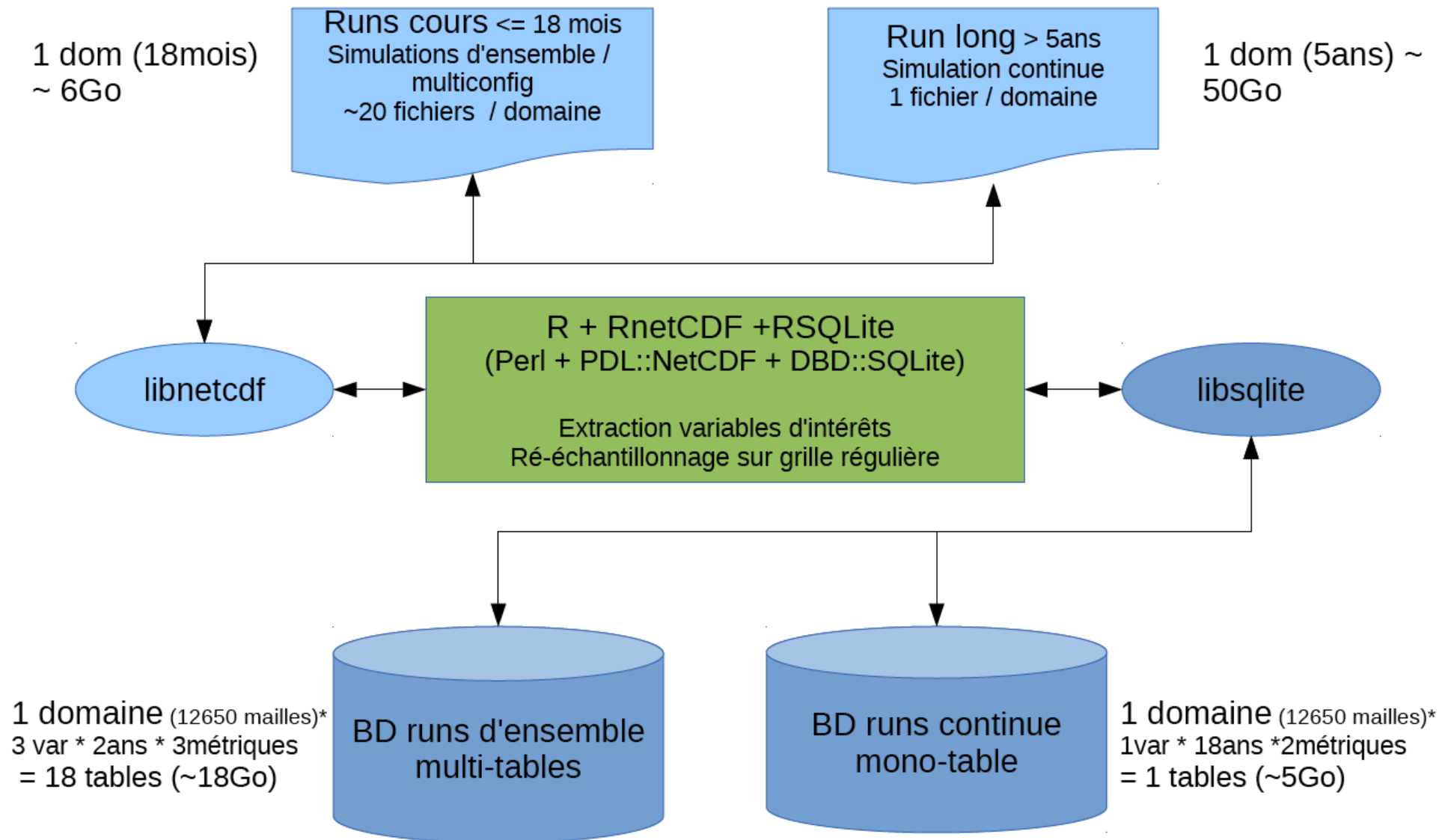
- [News & Milestones](#)
- [What is NCO?](#)
- [Contributing](#)
- [NCO→SDO Project](#)
- [Publications](#)
- [Version Summary](#)
- [Binaries](#)
- [Documentation](#)
- [FAQ](#)
- [README/ChangeLog/TODO](#)
- [Source Code](#)
- [Compiling](#)
- [UCI and NCAR](#)
- [Benchmarks](#)
- [Known Problems](#)
- [People](#)
- [Contact us](#)

NCO = Σ

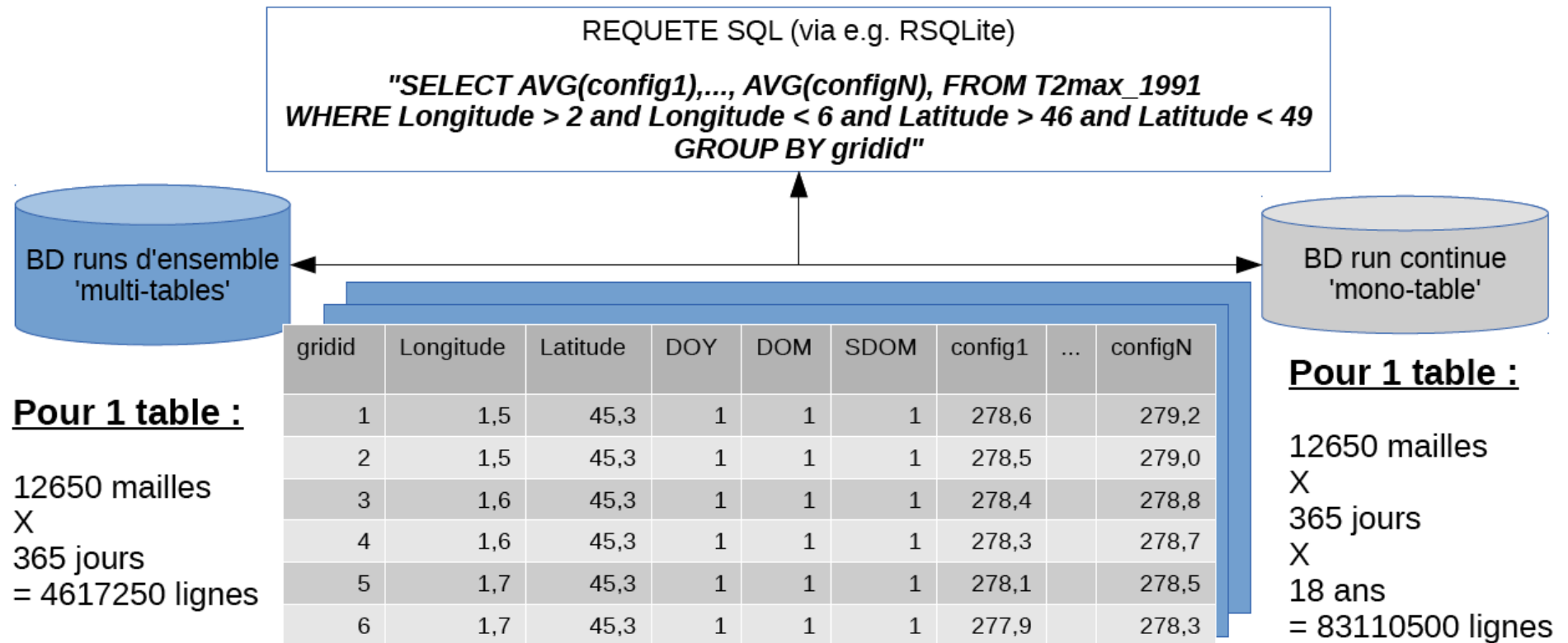
netCDF CF OPeNDAP HDF GSL

netCDF Operators (NCO) Software Stack

Du .nc aux BD : SQLite



Du .nc aux BD : SQLite



- + Format générique et accessible facilement via divers logiciels (stat, SIG etc.)
- + Langage de requête puissant, normalisé avec support et extensions possibles
- + Evite le 'out of memory', solution adaptée au langage de script (R, Matlab, etc.)
- Gestion des données multi-dimensionnelles → pas optimale (nécessite 'reshape' vers 2D)
- Pour table avec nombre de lignes > 20×10^6 → calculs lents (problème d'indexation ?)
- Manque de connaissance sur SGBD
- ? Amélioration temps calcul pour grosse table par indexation avec R*TREE ?
- ? Prise en compte du spatial → spatialite/rasterlite utiles ?
- ? Possibilité de mieux gérer le multidimensionnel ?

Conclusions

Besoins :

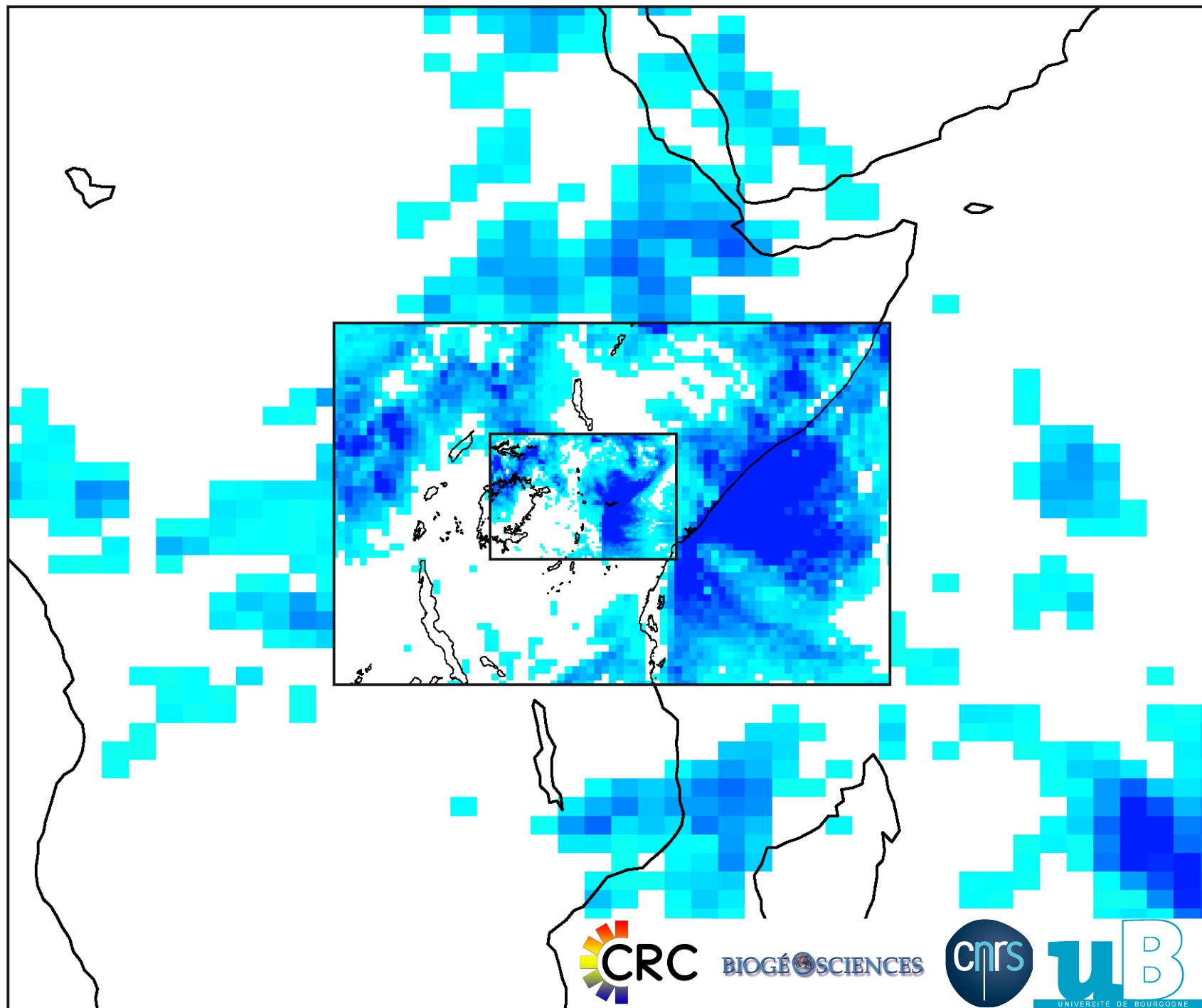
- gérer des matrices multi-dimensionnelles
- édition des données (concaténation dans le temps, ...)
- accès à ces données (import dans R, Matlab, ...)

Limites actuelles :

- soft : SGBD ?
- hard : serveurs d'archivage grande capacité faibles performances (or besoin d'accès...)
- en l'état actuel, dur de partager la donnée (hors uB)

Questionnement (cadre OSU) :

- intérêt pour ces données : archivage, pérennisation des mesures terrain, ...



BIOGÉOSCIENCES

