Data analysis tools and data mining in ensemble of ocean re-analysis and climate models

Guillaume Maze and Nicolas Kolodziejczyk
LOPS, Brest

Acknowledgement: C. Kermabon, P. Le Bot, F. Paul, F. Gaillard ,
axe transverse Data / LOPS, SO Argo-France, Brest

Journée de rencontre des utilisateurs du PCIM
"De CAPARMOR vers DATARMOR "
vendredi 30 septembre 2016, Ifremer

Introduction

Context

→ Growing global data base (15 years of Argo, satellite …)
→ Growing resolution of oceanic and climate models
→ Ensemble approach
→ **Complexify data flow for analyzis**

What do we do?

→ Explore multi-dimensional non-local diagnostics
→ Explore large ensemble of diagnostics
→ Inter-compare products

What do we need ?

→ We need efficient numerical libraries and work flows
e.g. standard matrix manipulation stats, dimensionality reduction, inversion, covariance, interpolation ...
→ for more complex methods
e.g. regression, classification, neural networks, support vector machine, deep learning, etc…

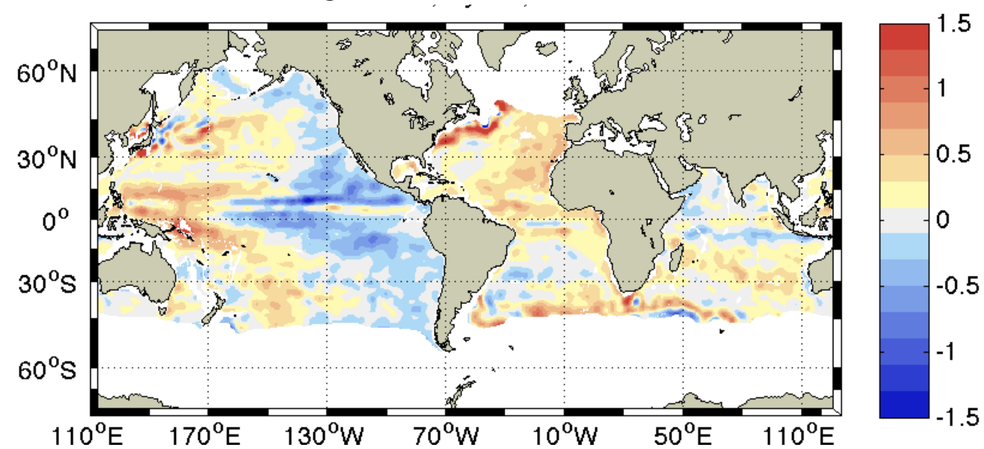
ISAS tools

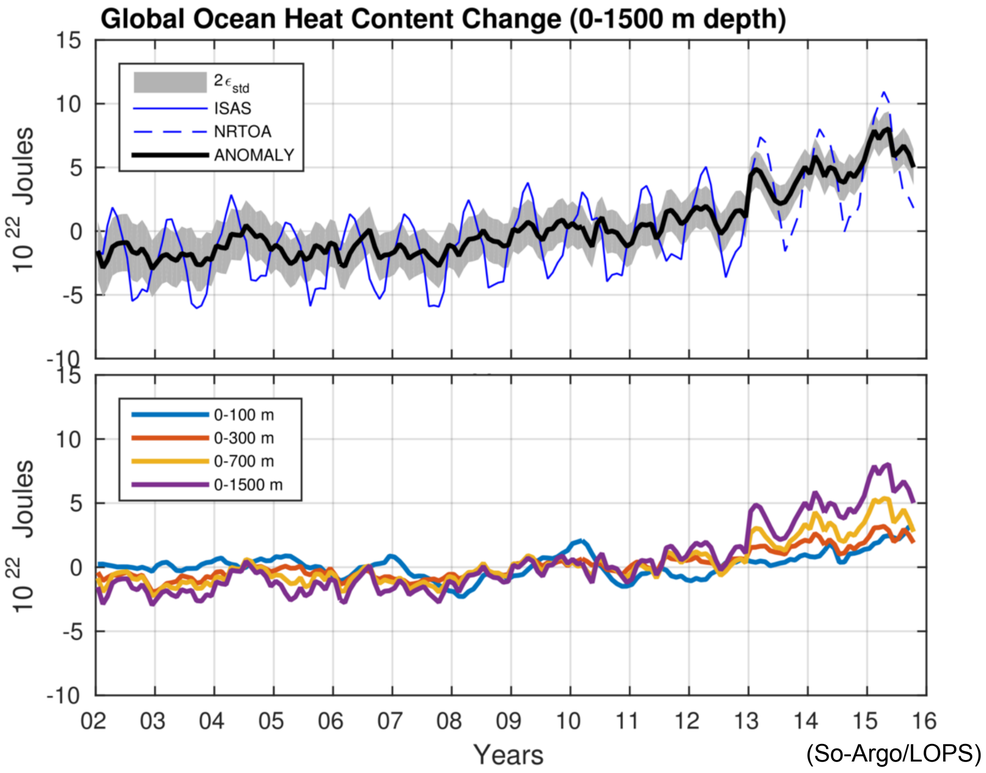
OI Analysis tools (Kalman filter)
→ Gridding global scalar field (e.g.: T, S) from in situ measurements
→ Configuration: Global ocean
→ Résolution : dx = 0.5° ; dy = 5 to 20 m
(152 levels from 0 to 2000m depth)
→ Monthly fields : 30 days and ~300km covariance scales
→ Data: Argo, CTD, mouillages, Memo, … BUT NO XBT

Diagnostics
From T and S fields many derived quantities :
→ TEOS-10 : Potential temperature and density, Absolute salinity, spiciness …
→ Integral quantities : Heat and Fresh Water Content...
→ Second order quantities : stratification, potential vorticity, …

Products
→ Scientific products and tools for community
→ Periodic releases (~ yearly) and growing dataset
→ Develop tools for operational

T and S averaged between surface-σ_0 < *26.5*
Diagnosed from ISAS tools





ISAS tools

- Need to implement the ISAS system on dedicated platform

→ Datarmor ? Both scientific and operational aims

- Optimize generation of gridded products

→ Growing dataset to analyses, parallelization…

- Optimize diagnostics and storage of 4D field

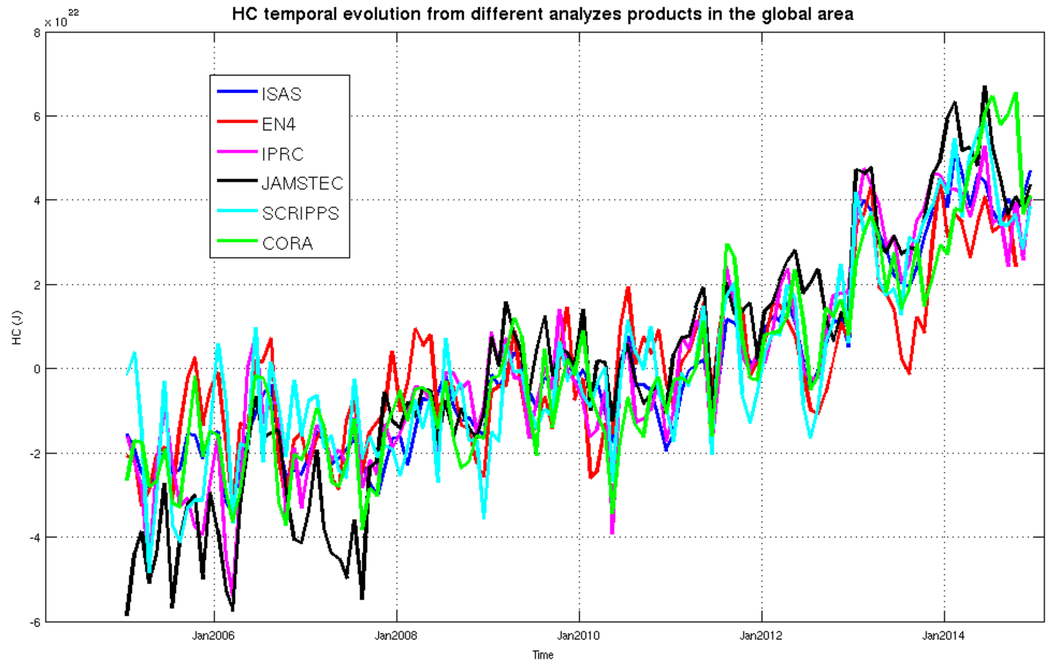
→ On demand, storage (~3T), access to the platform ?

- Optimize statistics analysis of 4D field

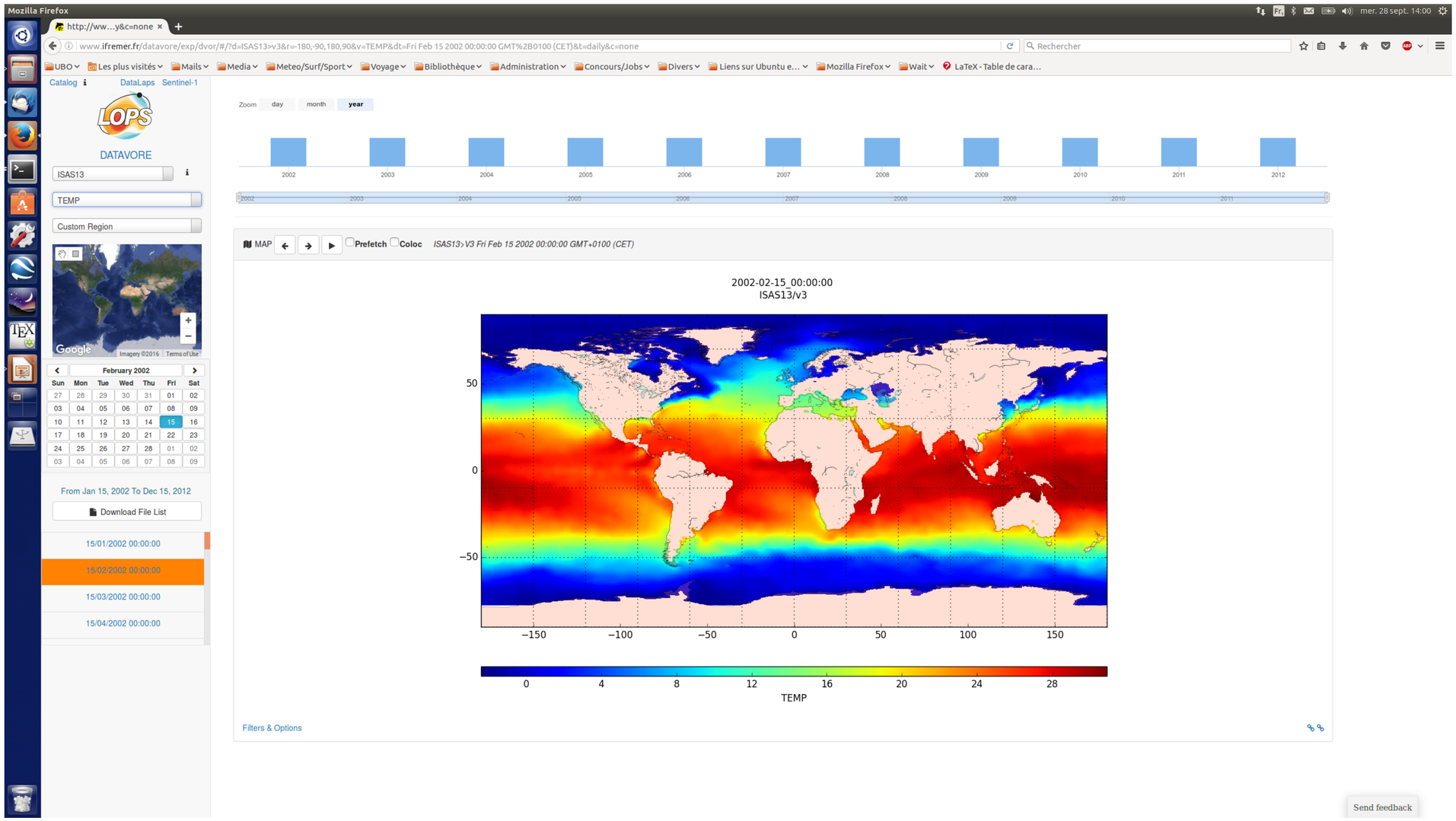
→ standard matrix manipulation stats, more complex methods

- Intercomparaison in situ data products + diagnostics

→ x 10-20 size from other products, preprocessing, ...



Quick data visualization and access



User friendly interface for data visualization and extraction

- Visualisation tools base on quick extraction
- Quick diagnostics
- Download data selection

→ **Backed up by DATAVOR cluster/cloud at LOPS → Datarmor**

http://www.ifremer.fr/datavore/exp/dvor/

http://www.umr-lops.fr/SO-Argo

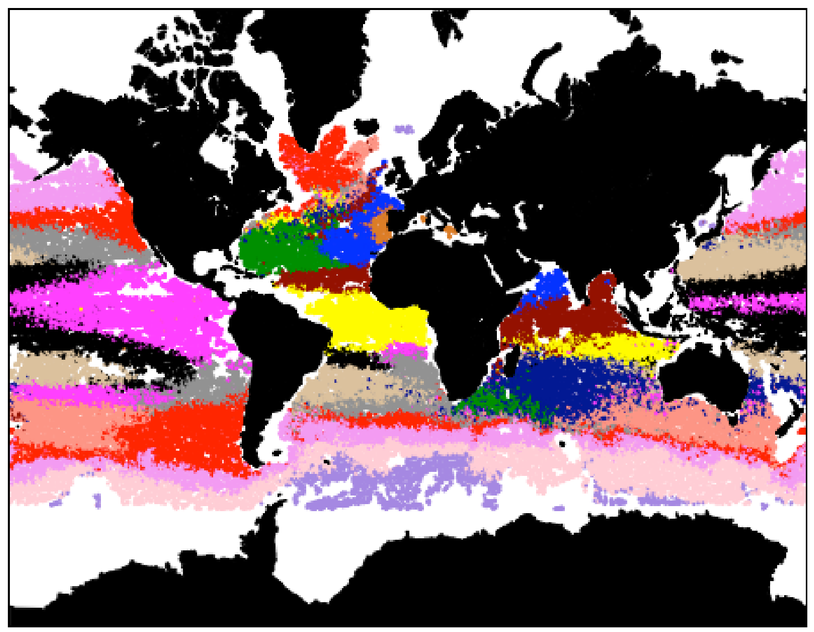
Inter-comparison and ensemble approach

We explore :

→ Multi-dimensional non-local diagnostics

→ large ensemble of diagnostics

→ Inter-compare products and ensemble



here: 15 years, North Atlantic: 0.1×10^6 profiles

x10

All **Argo**: 15 years, global: 1.5×10^6

x10

ORA-S4: 50 years, monthly, global 1/1 gridded: 26×10^6

ISAS13+nrt: 13 years, monthly, global 1/2 gridded: 43×10^6

x10

HadGEM: 140 years, monthly, global 1/1 gridded: 92×10^6

x10

ORCA025: 40 years, weekly, global 1/4 gridded: $1\,400 \times 10^6$

CMIP5: 50 years, monthly, global 1/1 gridded, 50 runs: $1\,500 \times 10^6$

DRAKKAR12: 20 years, weekly, global 1/12 gridded: $6\,400 \times 10^6$

x10

OCCIPUT: 50 years, weekly, global 1/4 gridded, **50 runs** $8\,900 \times 10^6$

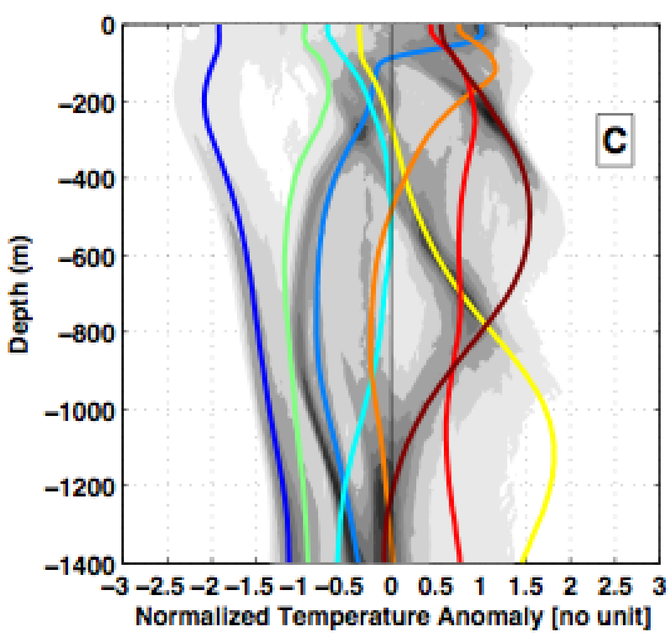
desktop computer

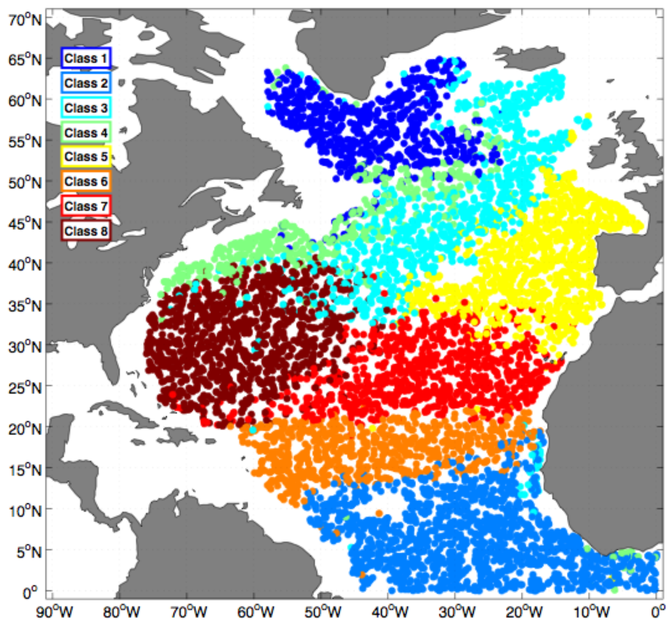
what can we do on Datarmor ?

Data mining

Example: Determine a Profile Classification Model:

- interpolate profiles on standard depth levels
- extract 2D plain matrix from 4D products (time series of 3D fields)
- center/standardise
- compute eigenvectors and singular values
- train Gaussian Mixture Model (computation and inversion of multiple covariance matrices)
- compute weighted statistics of profiles





Maze et al, 2016

Data mining

We are testing Spark on the LOPS data cluster

it’s a “fast and general engine for large-scale data processing and machine learning”

- we found bugs and inconsistencies in the machine learning library
- environmental/ocean data were not in the mind of the developers
- we found them very responsive with updating/patching from our suggestions
- we need access to the library source codes to fix it if required
- we need a permanently updated library because it’s going fast
- we need monitoring tools

Many questions about DATARMOR:

How to access diagnostics for analysis methods ?

→ On demand from raw data/fields
vs
→ On disk from pre-processing ?

Require rapid access to large memory
Require rapid disk reading

→ Research needs both !

Access and storage policies ?