Bertrand Chapron

Initiallement par Jean-François Piollé

(Ifremer)

"Nephelae"

26 Novembre 2013, Ifremer, Brest

Ifremer

Lab-STICC

http://wwz.ifremer.fr/bigdata

Nephelae : a platform for data intensive science – an application to ocean

Jean-François Piollé, Frédéric Paul, Olivier Archer, Bertrand Chapron
Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer), Brest, France

Des données, des données, des données

Pourquoi :

Défi technologique ? Conquête de l'Espace et démonstration politique ?

Espoir de résoudre (aider à) des questions scientifiques majeures et/ou nouvelles?

Inadéquation entre les demandes (recherche, services et applications) et modèles (encore) trop simplifiés ? Augmenter (toujours) l'échantillonnage spatio-temporel des observations

IFREMER / CERSAT

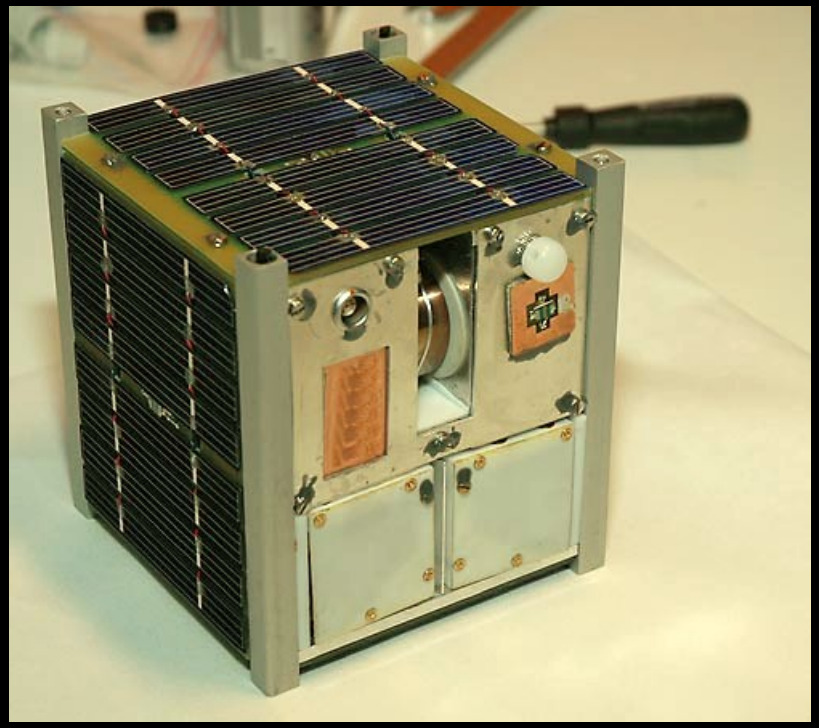
CERSAT is the satellite data centre of Ifremer, addressing international user community

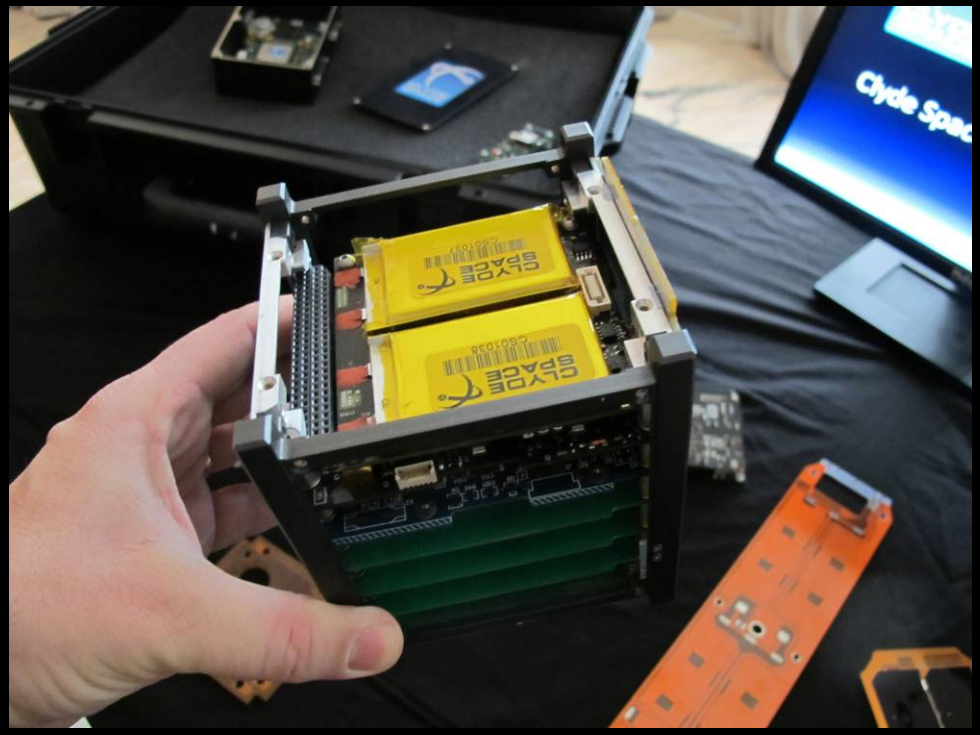
- Focus on surface parameters [**wind, waves, fluxes, sea surface temperature, sea ice, salinity,...**], data from radars and passive radiometers
- Unique archive for some missions (ERS-1 & ERS-2), more than 25 satellite missions & 300 data collections
- Include **satellite** but also **model** and **in situ** data, original datasets or mirrored datasets
- Associate engineering team for data management & processing and research team for processing algorithm & validation, geophysical applications
- Focus on cross sensor comparison and synergy : cross colocation, combination of multiple sources, multi variate data analysis and processing, tools for online data analysis

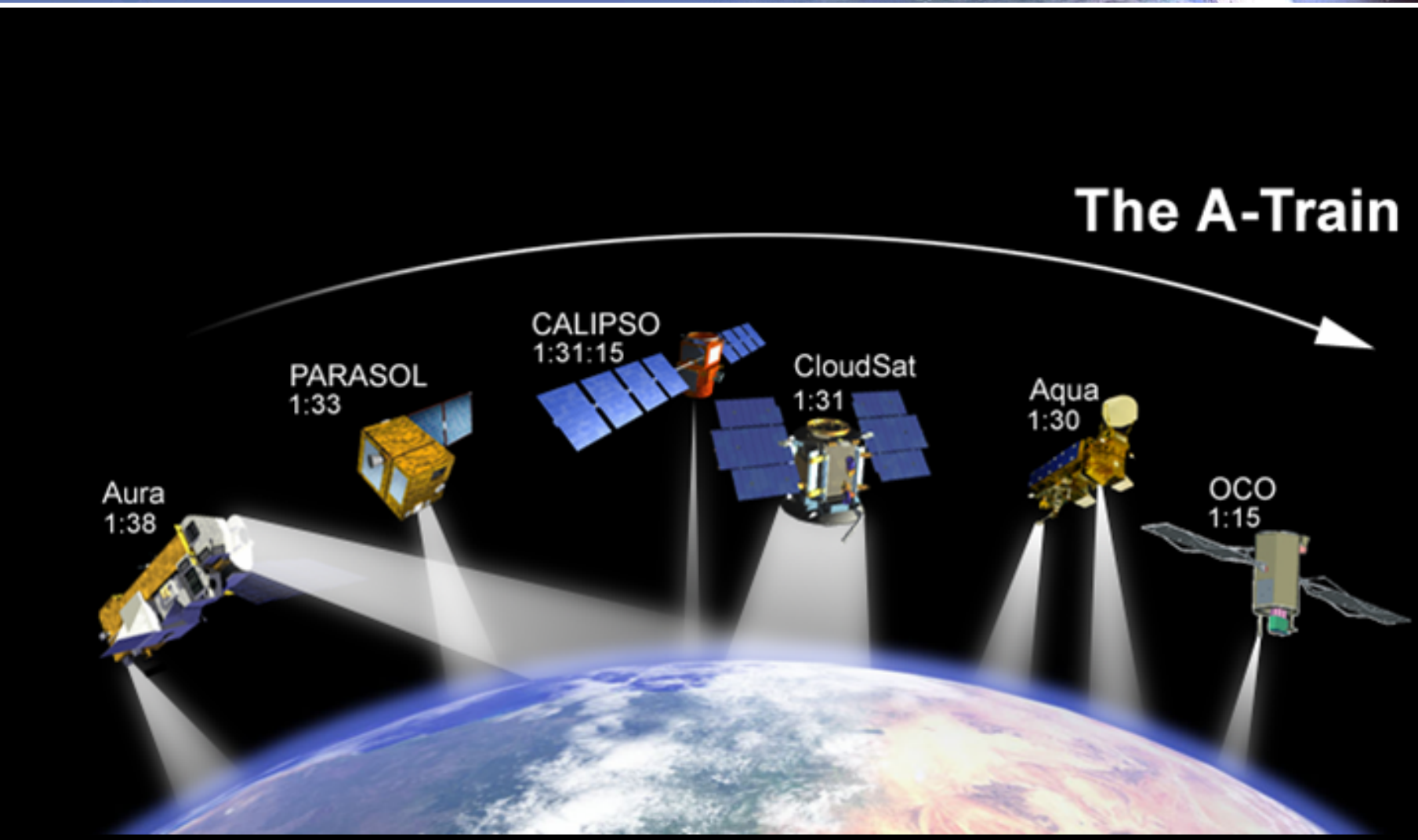
Large spectrum of data, strong data management issues, major interest and focus on extracting knowledge from this large ecosystem of data for research and decision making

New Era - Nanosatellites - CubeSat

A **CubeSat** is a type of miniaturized satellite for space research that usually has a volume of exactly one liter (10 cm cube) mass of no more than 1.33 kilograms.






The A-Train
CALIPSO
1:31:15
PARASOL
1:33
CloudSat
1:31
Aqua
1:30
Aura
1:38
OCO
1:15

En quelques chiffres

Environ 25000 tâches lancées quotidiennement

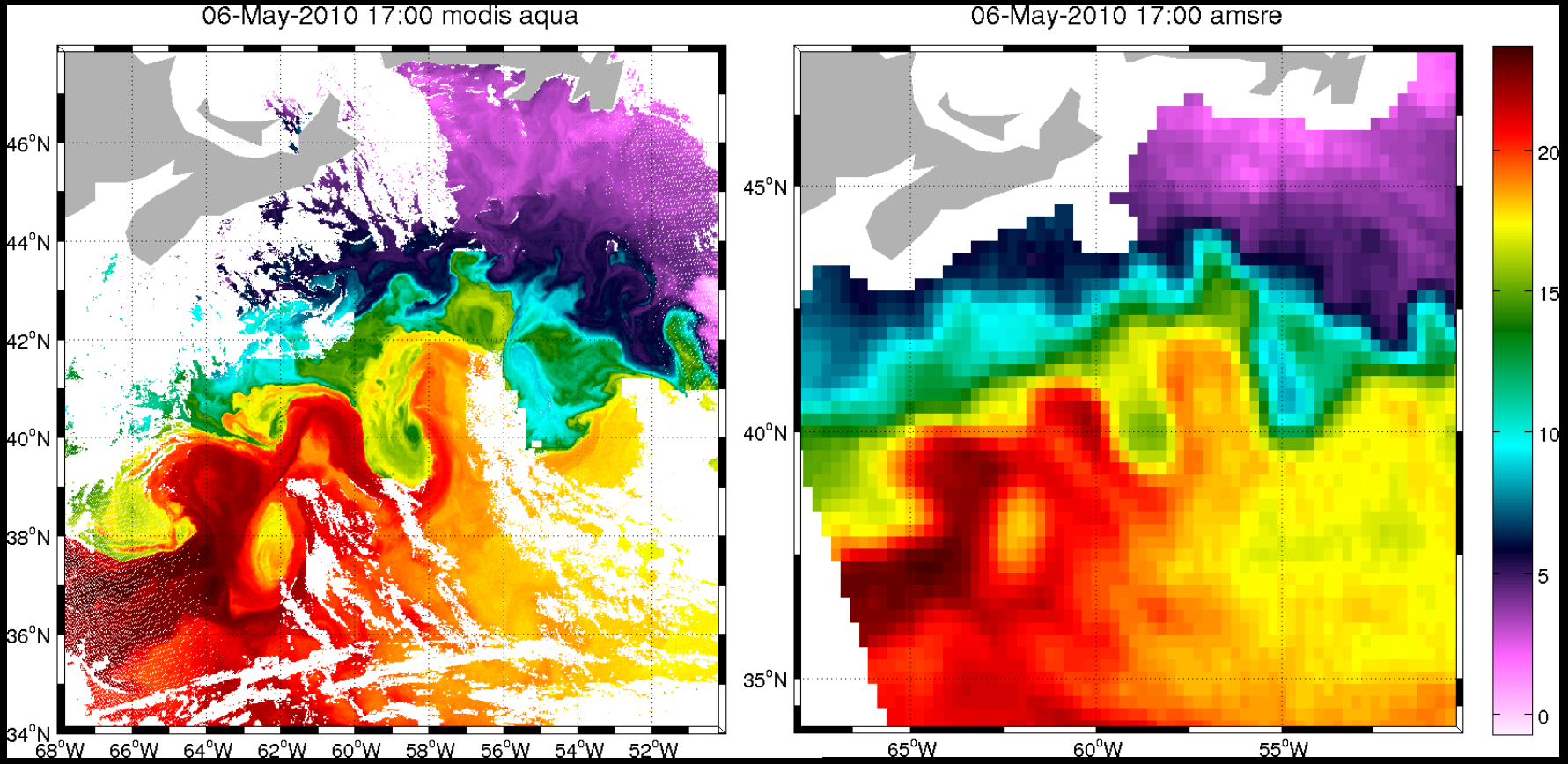
80 h de calcul quotidien (parallélisation non prise en compte)

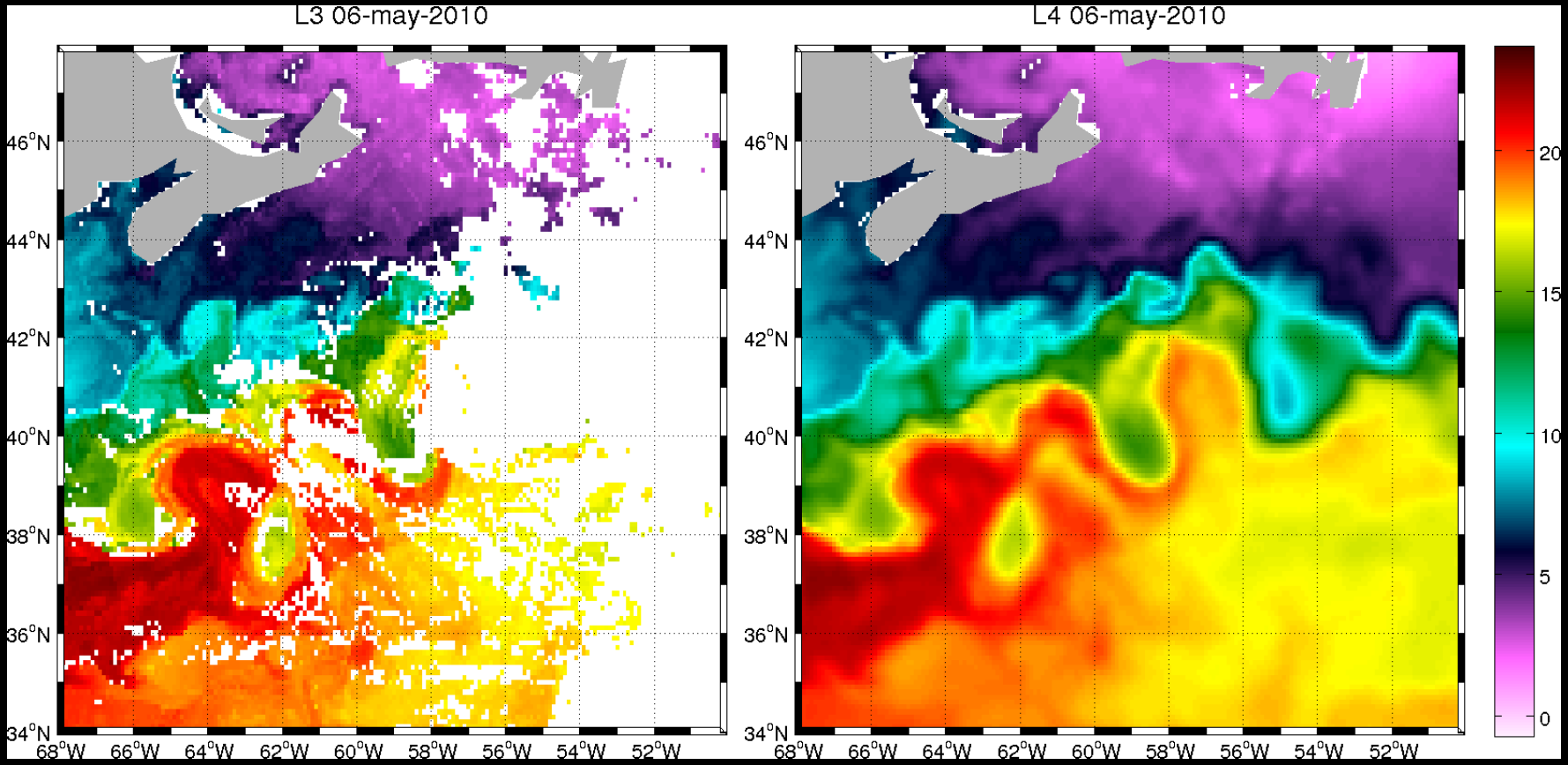
Environ 300 Go de données réceptionnées par jour

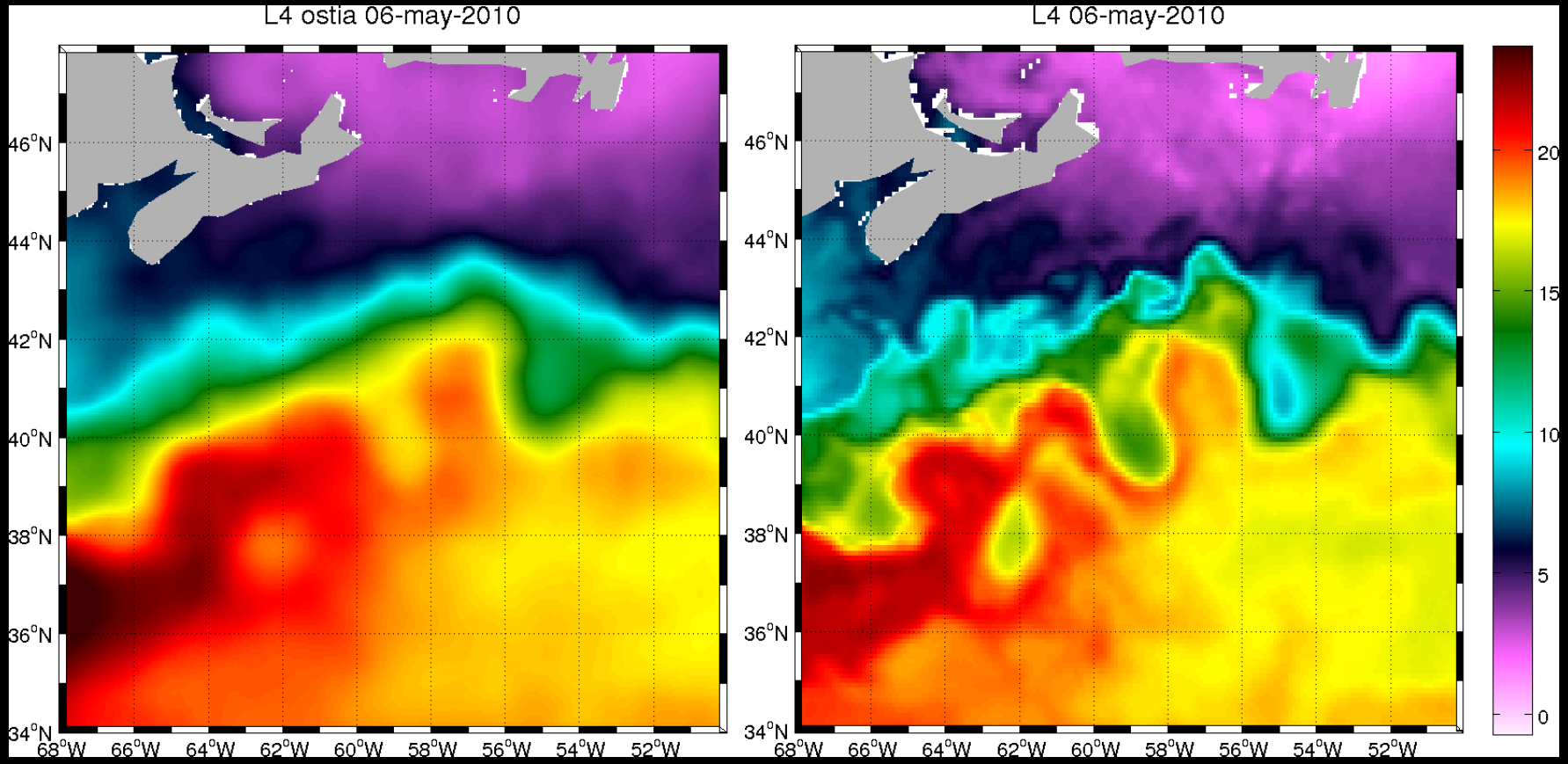
200 To d'archive sur bande

600 To de données sur disques

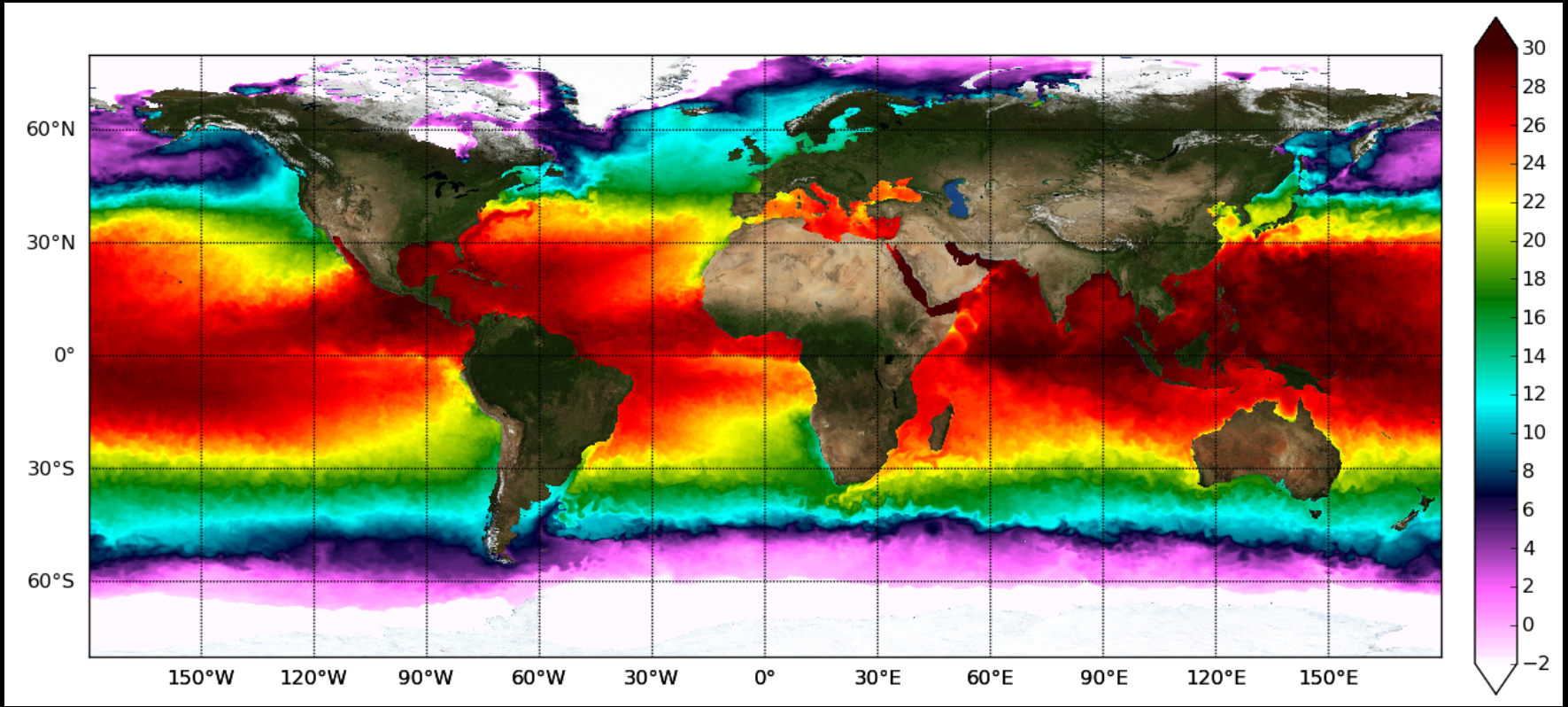
Missions futures : Sentinel-1 (1 Po/an), Sentinel-3 (1 Po/an)


06-May-2010 17:00 modis aqua
06-May-2010 17:00 amsre
46°N
44°N
42°N
40°N
38°N
36°N
34°N
68°W
66°W
64°W
62°W
60°W
58°W
56°W
54°W
52°W
45°N
40°N
35°N
65°W
60°W
55°W
20
15
10
5
0


L3 06-may-2010
L4 06-may-2010
46°N
44°N
42°N
40°N
38°N
36°N
34°N
68°W
66°W
64°W
62°W
60°W
58°W
56°W
54°W
52°W
20
15
10
5
0


L4 ostia 06-may-2010
L4 06-may-2010
46°N
44°N
42°N
40°N
38°N
36°N
34°N
68°W
66°W
64°W
62°W
60°W
58°W
56°W
54°W
52°W
20
15
10
5
0

Global SST



Global reanalysis 2006-present
at 10 km resolution

medspiration
EUROPEAN NODE FOR SEA SURFACE TEMPERATURE

esa

Activités autour de la gestion de données

Traitement et valorisation

- On demand / Interactive processing
- Produits évolués
- Produits géophysiques
- Contrôle qualité validation
- consolidation
- reprocessing

Outils d'indexation, fouille et analyse

- Indexation / caracterisation de données
- Outils de recherche
- Colocalisation multiple
- Outils de visualisation
- cross-references graph database,etc...

Moyens de traitement/archivage de masse

- Map/reduce hadoop
- Cloud computing
- Outils de batch
- Outils de monitoring
- Big data

Architectures data intensive
Plateforme Nephelae

Données

- Données satellites
- Modèles / simulations
- Données de validation
- Données ancillaires

Opening the Pandora's box ?

Archiving data leads to very large heterogeneous and multimodal databases

Data assimilation is growing in response to the growth of data collected, but (personal opinion) tremendous amounts of information still remain hidden in data archives.

Knowledge trees and complex algorithms are essential to avoid the Google's principle, i.e. pertinence = popularity

Research efforts to be concerned with the definition of adequate exploratory processes to detect relevant patterns in large, heterogeneous, multidimensional observation data sets with different resolutions to better approach complex spatial and/or temporal dynamics of the ocean system.

Problématiques

- Caractériser les données : homogénéité, formes, ...
- Suivi de structures
- Reconstruction d'information
- Compression de l'information
- Indexation de l'information, lier les observations
  - Bases de données
- Recherche de données :
  - Description
  - Sémantique
  - Navigation
- Datamining : analyse croisée de différentes sources / paramètres

Accumulation

Archivage long-terme : Nécessaire (on ne sait jamais, besoin pour les re-traitements)

Archivage + dynamique : Décomposition/Réduction d'échelles et compression

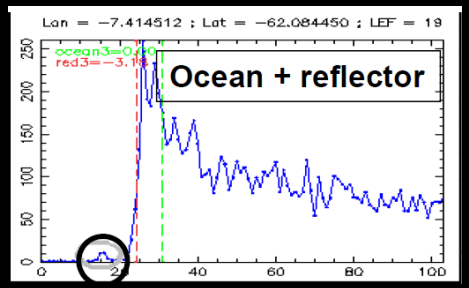
-> Ne pas (plus) rater/éliminer les évènements rares ou encore des associations encore non-anticipées

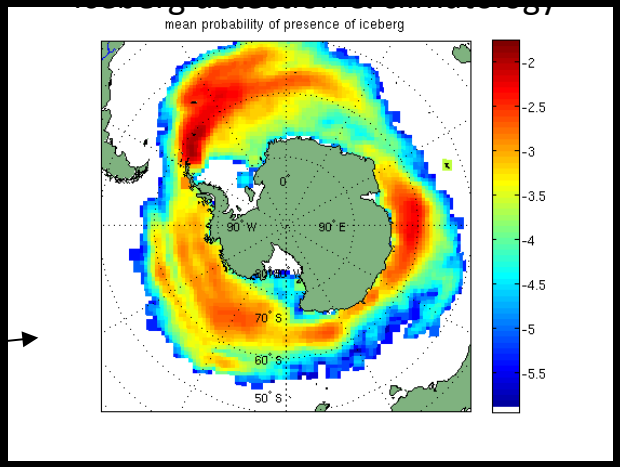
-> Favoriser la détection des signaux faibles et identification des seuils critiques

Extracting new knowledge

Analysis of altimeter wave forms :
ERS 1 & 2, Envisat, Jason 1 & 2, Cryosat, AltiKa (12 TB)

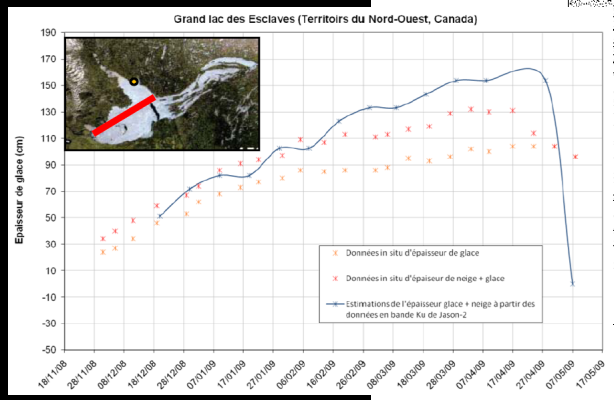
Disposing of a sandbox with permanent access to all data and processing power greatly ease bridging the gap between initial idea and full demonstration / long term assessment

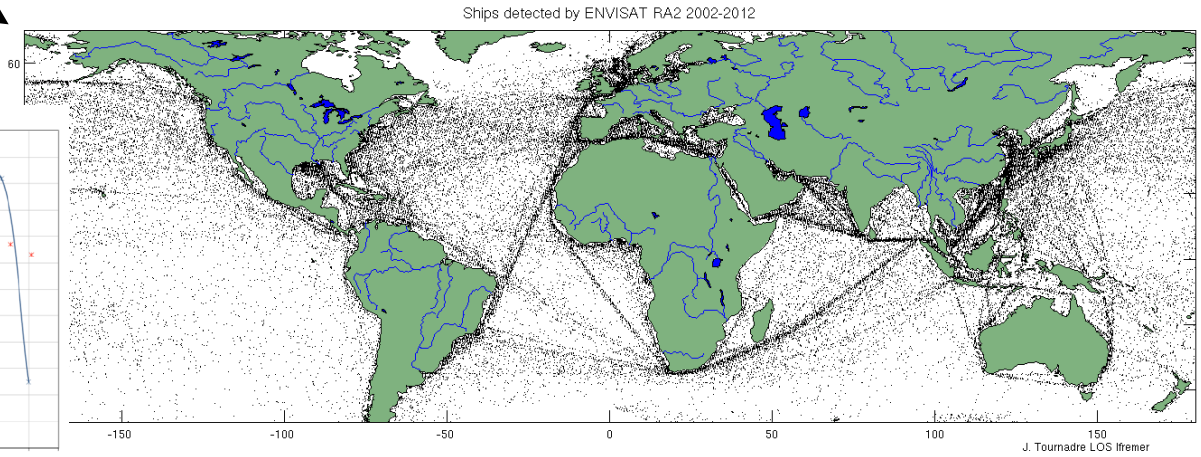




Ship detection

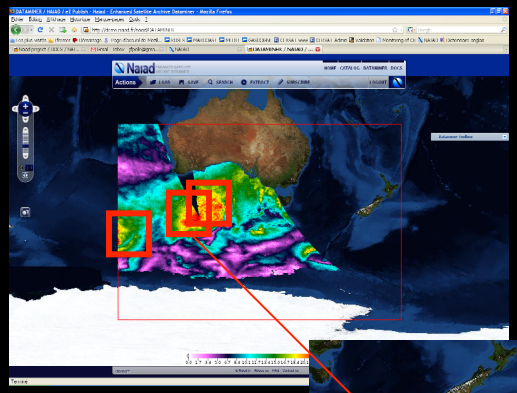
Lake ice thickness



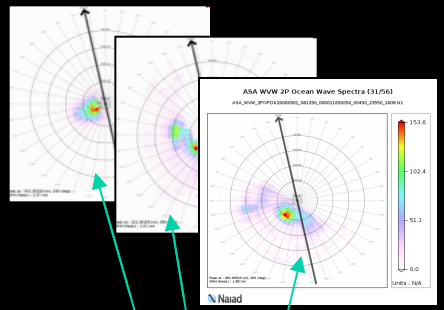


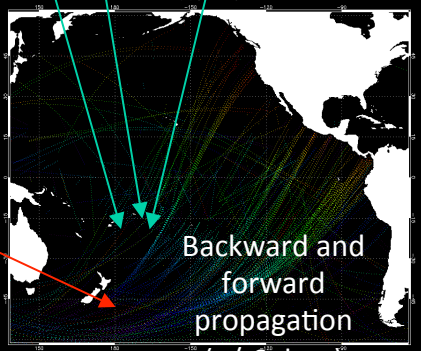
Big questions….

Are storms more numerous and intensifying with climate

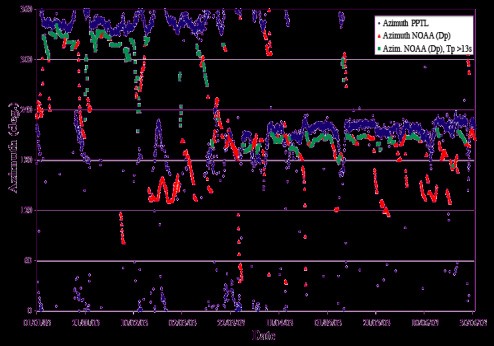


Scatterometer and SAR (20 years)



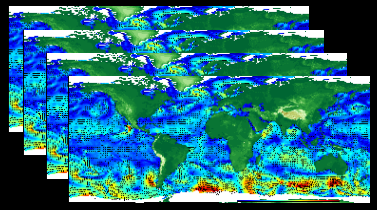


Backward and forward propagation (+/-6 days)



Seismic noise (50 years)

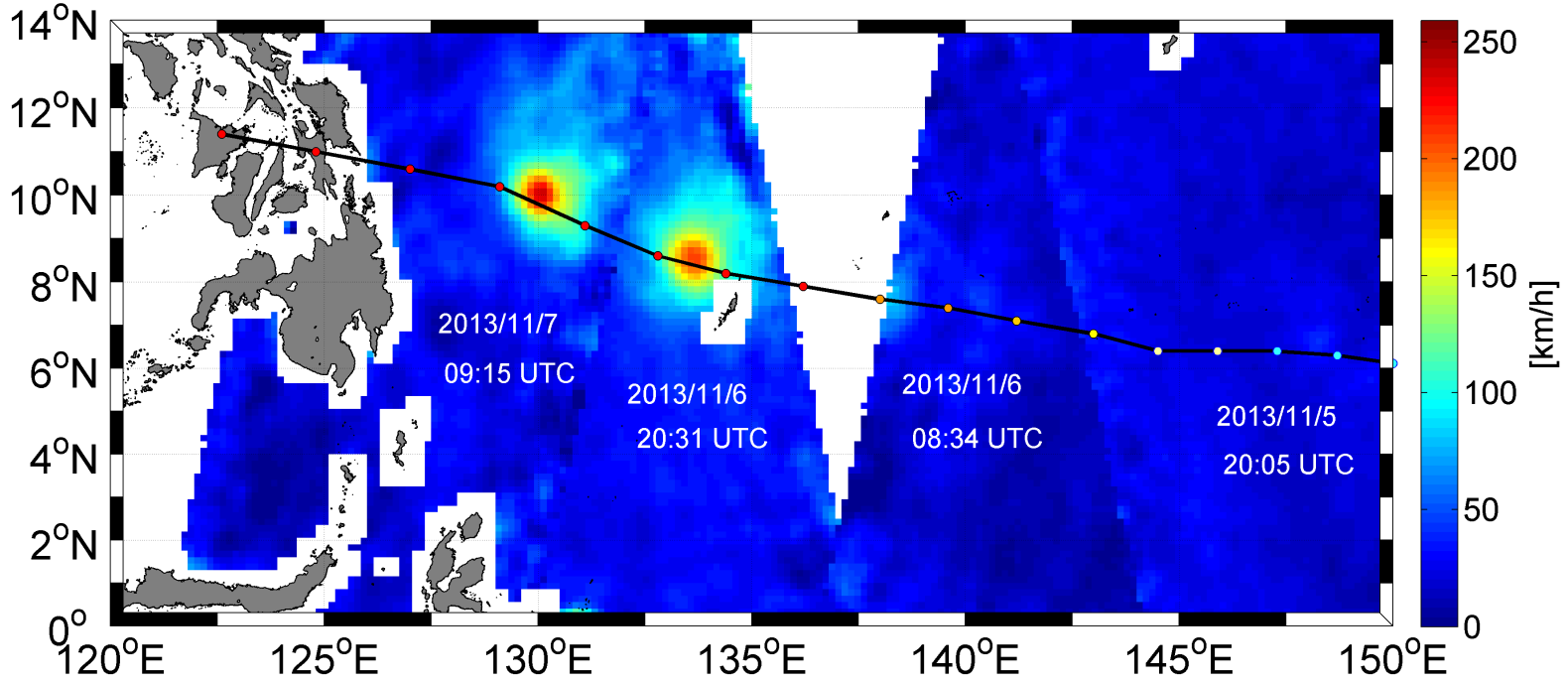
Buoys (30 years)

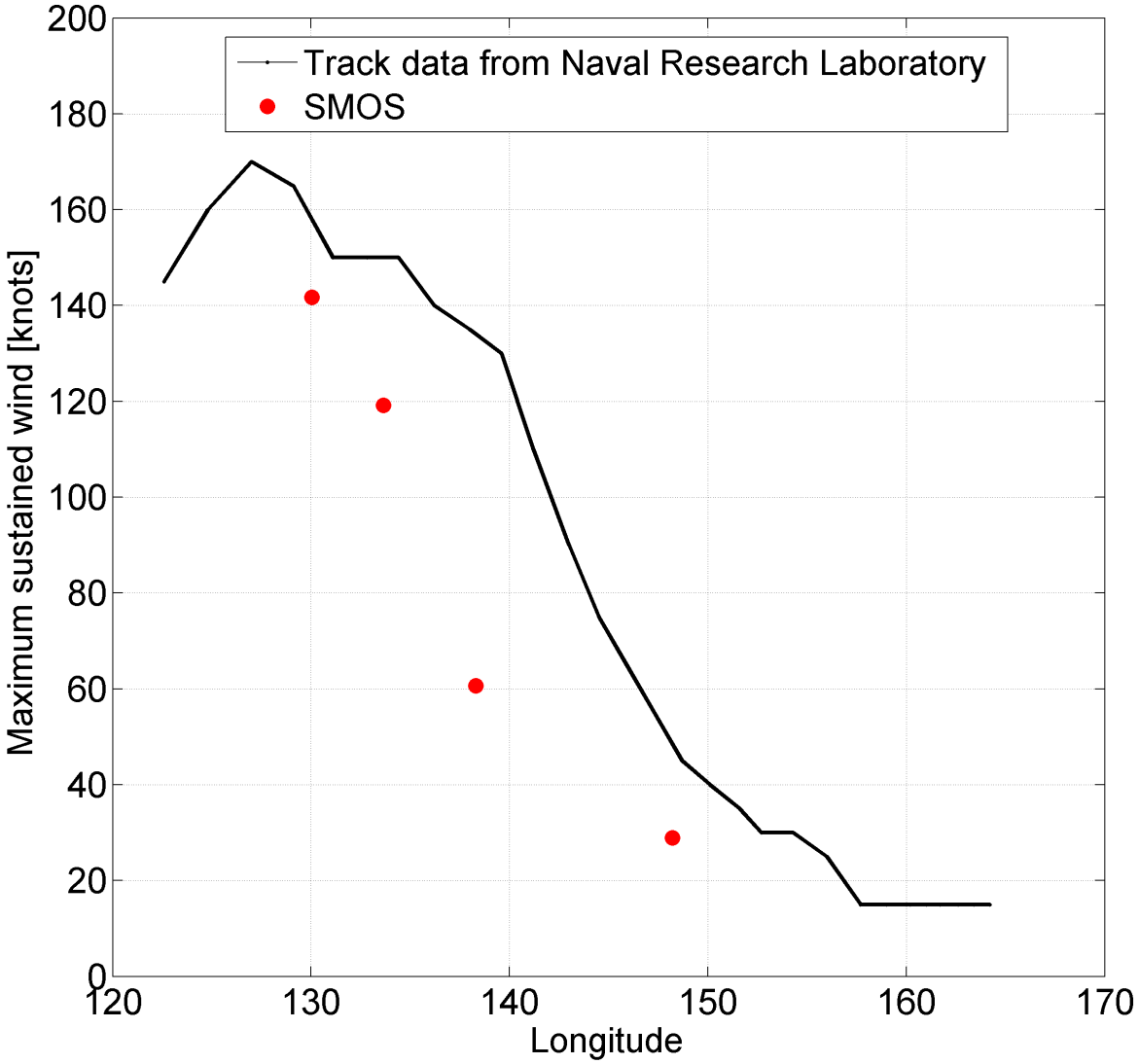


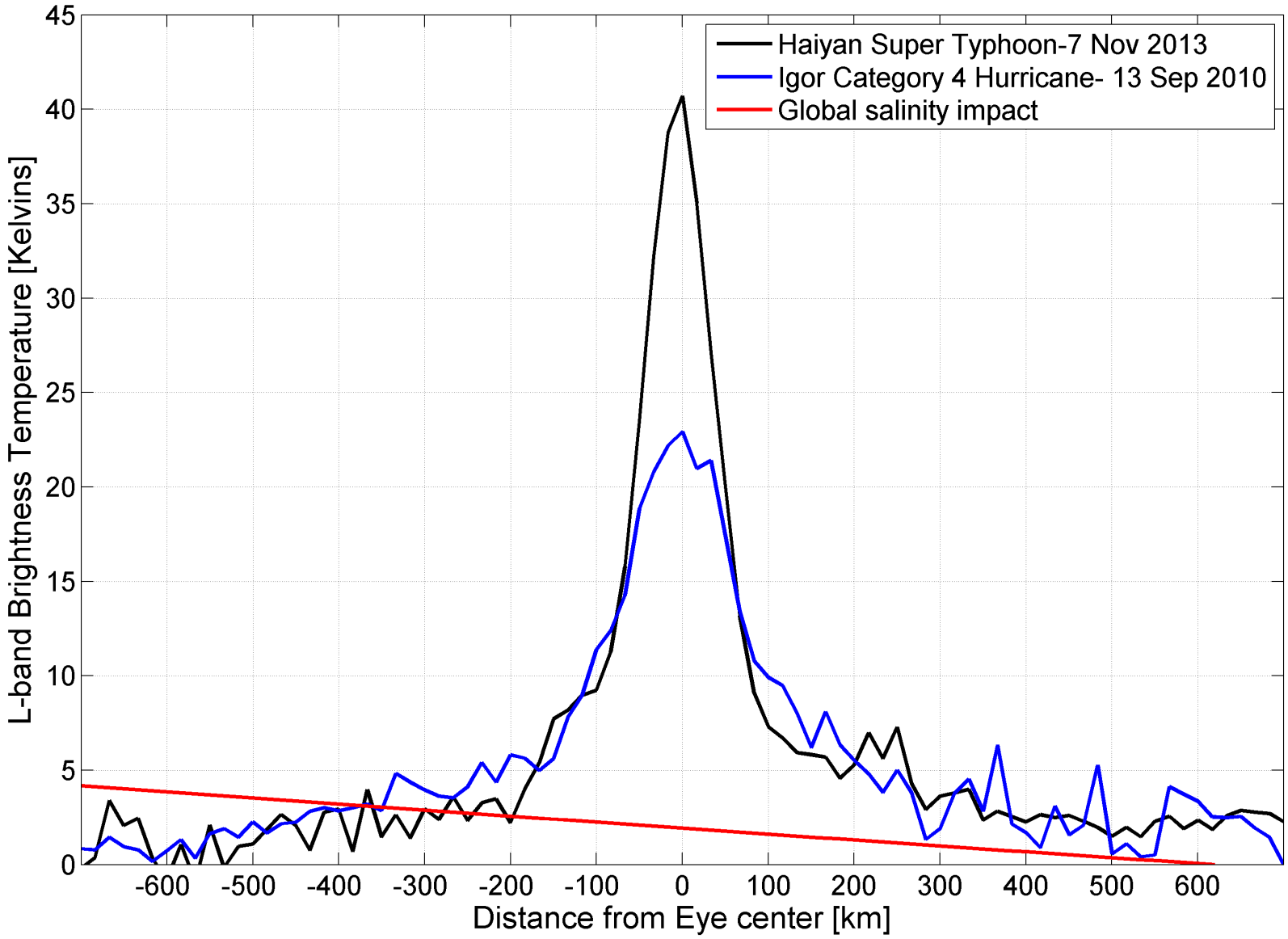
Weather model (25 years)

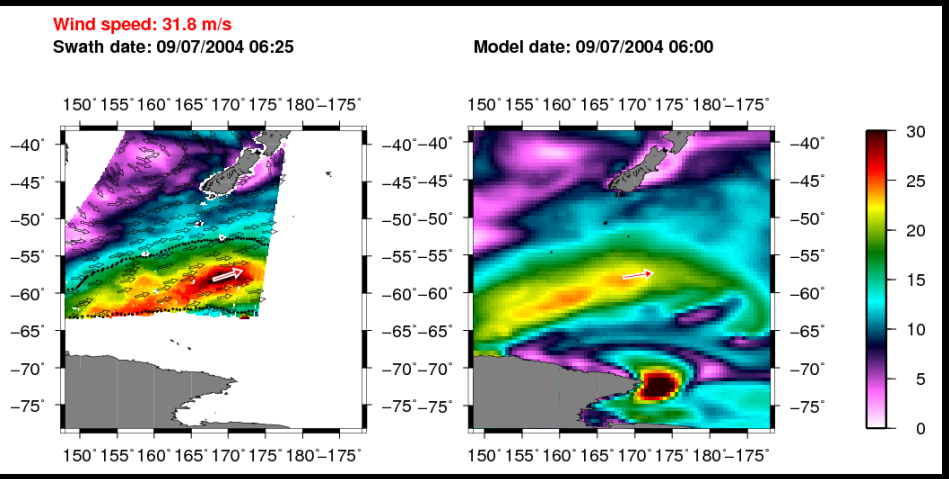
Feature and tracks extraction

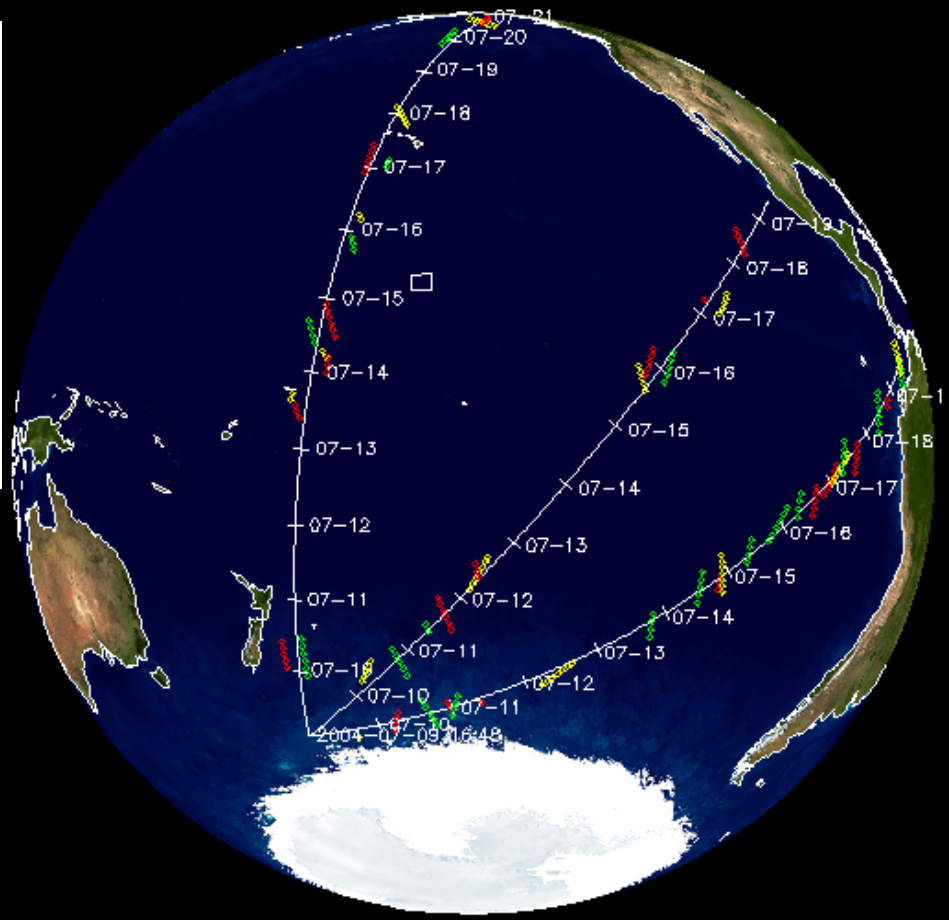
Extracting new knowledge: SMOS Extreme Typhoon




Track data from Naval Research Laboratory
SMOS
Maximum sustained wind [knots]
Longitude
200
180
160
140
120
100
80
60
40
20
0
120
130
140
150
160
170

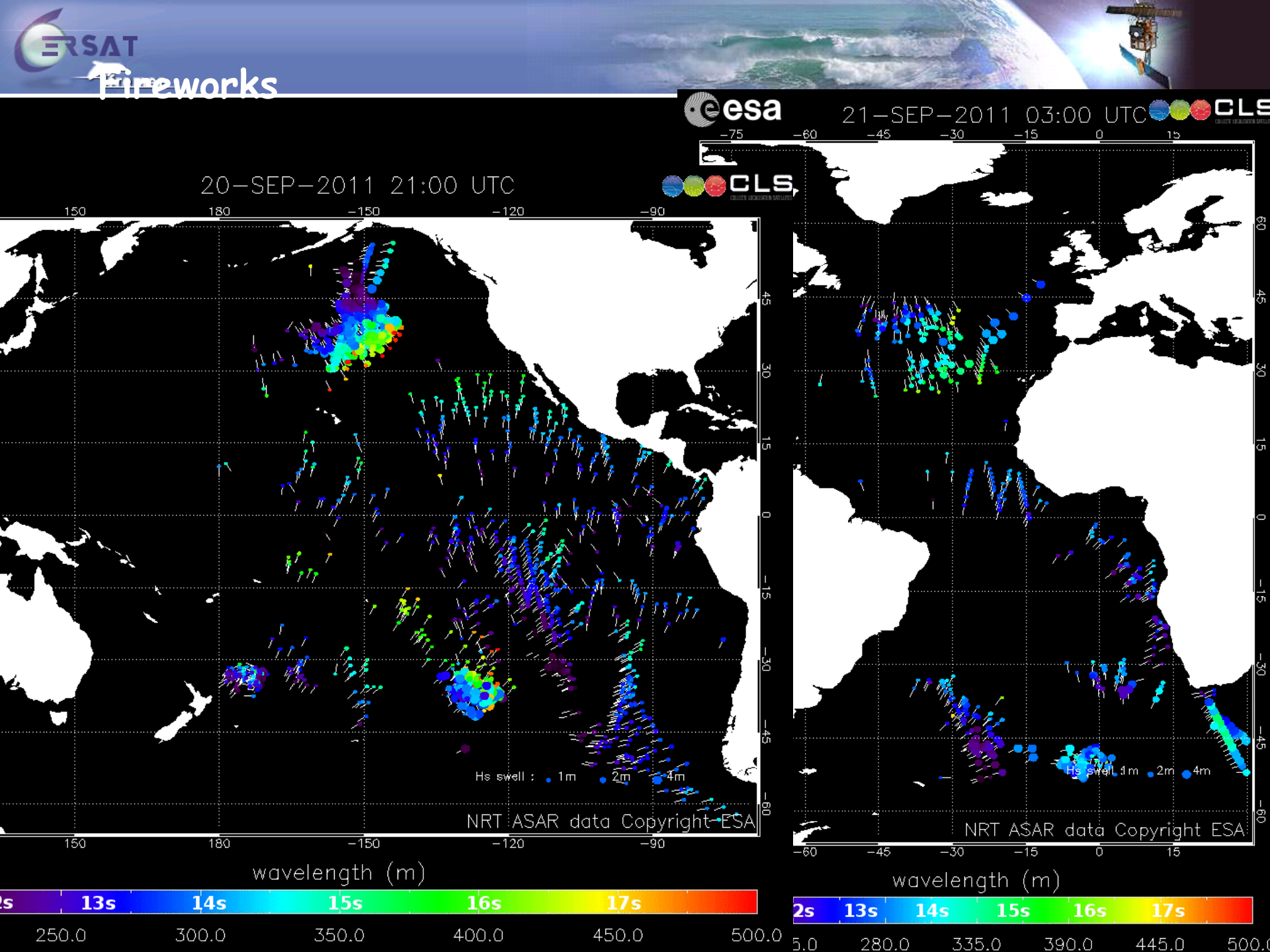

Haiyan Super Typhoon-7 Nov 2013
Igor Category 4 Hurricane- 13 Sep 2010
Global salinity impact
L-band Brightness Temperature [Kelvins]
Distance from Eye center [km]
0
5
10
15
20
25
30
35
40
45
-600
-500
-400
-300
-200
-100
0
100
200
300
400
500
600





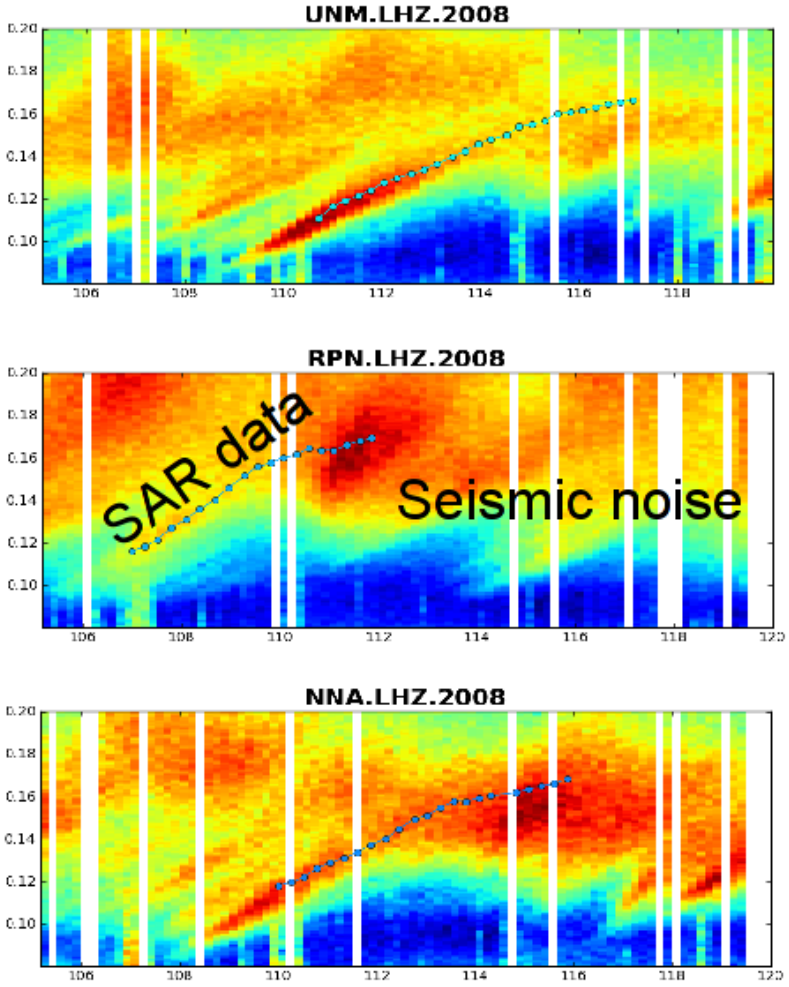
RED : ENVISAT ASAR
GREEN : ENVISAT RA2
YELLOW : JASON ALTIMETER

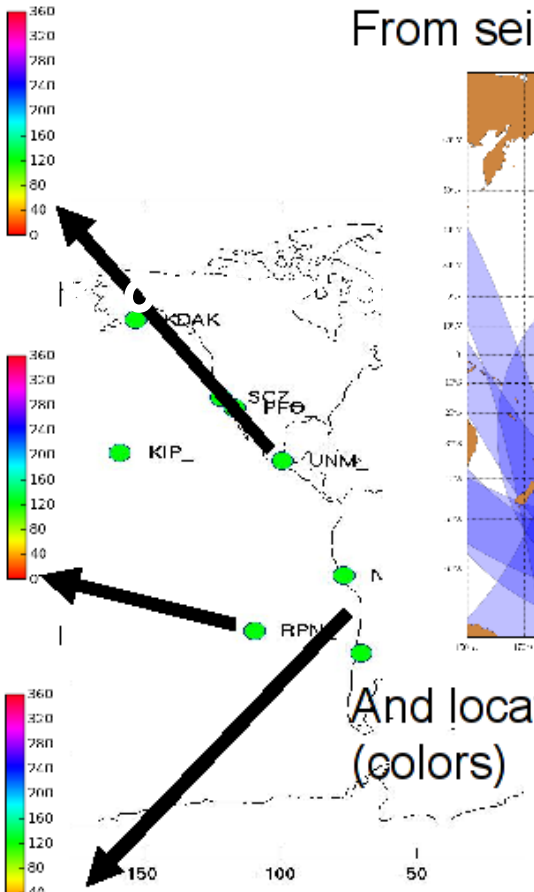
Fireworks


20-SEP-2011 21:00 UTC
21-SEP-2011 03:00 UTC
esa
CLS
Hs swell : 1m 2m 4m
NRT ASAR data Copyright ESA
wavelength (m)
13s 14s 15s 16s 17s
250.0 300.0 350.0 400.0 450.0 500.0
280.0 335.0 390.0 445.0

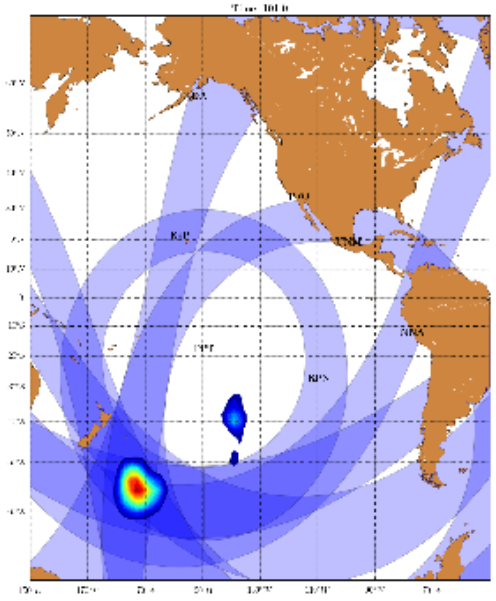
Consistency and storm severity

Example of seismic - SAR synergy





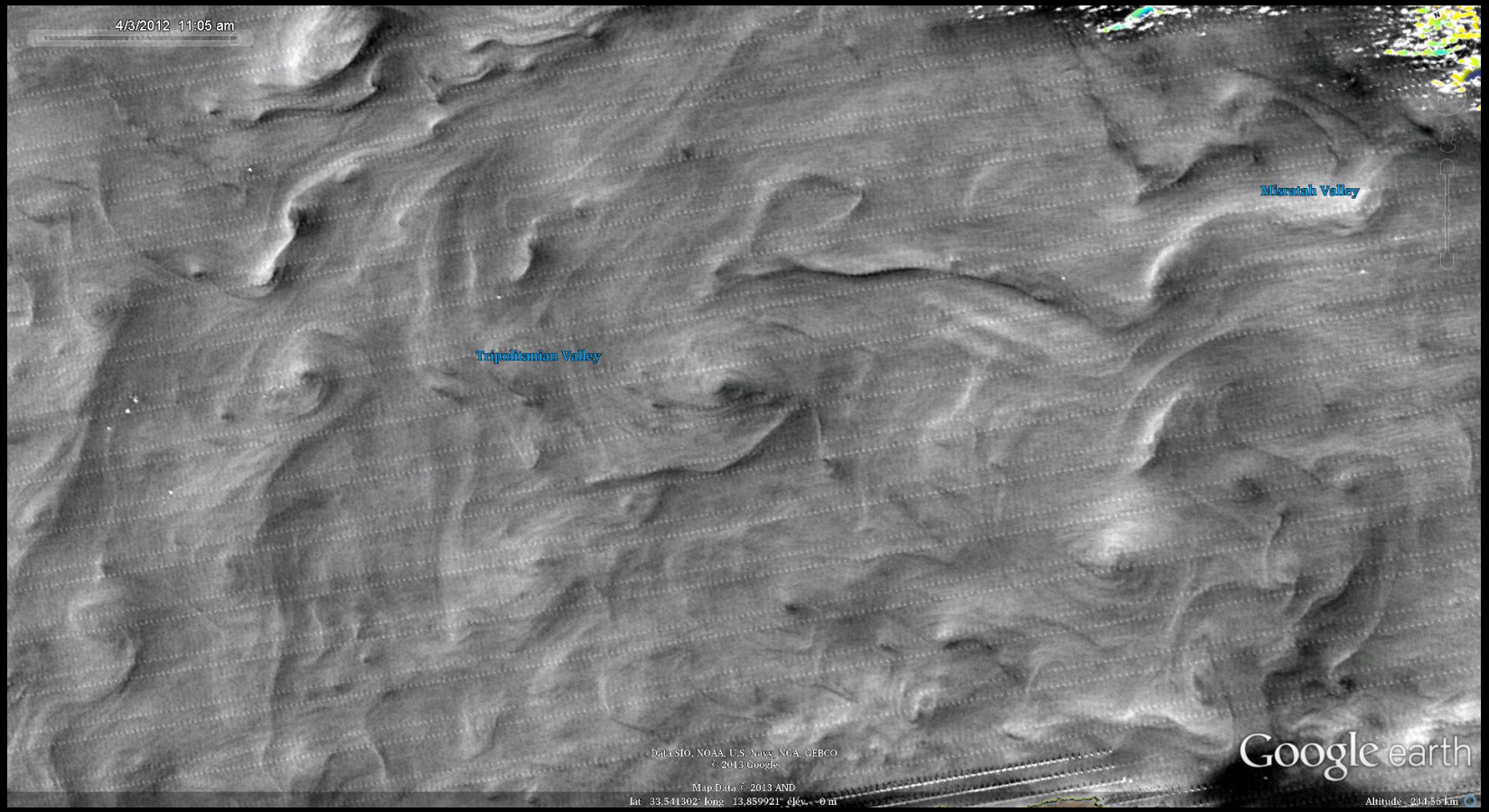
Storm location
From seismic (blue bands)



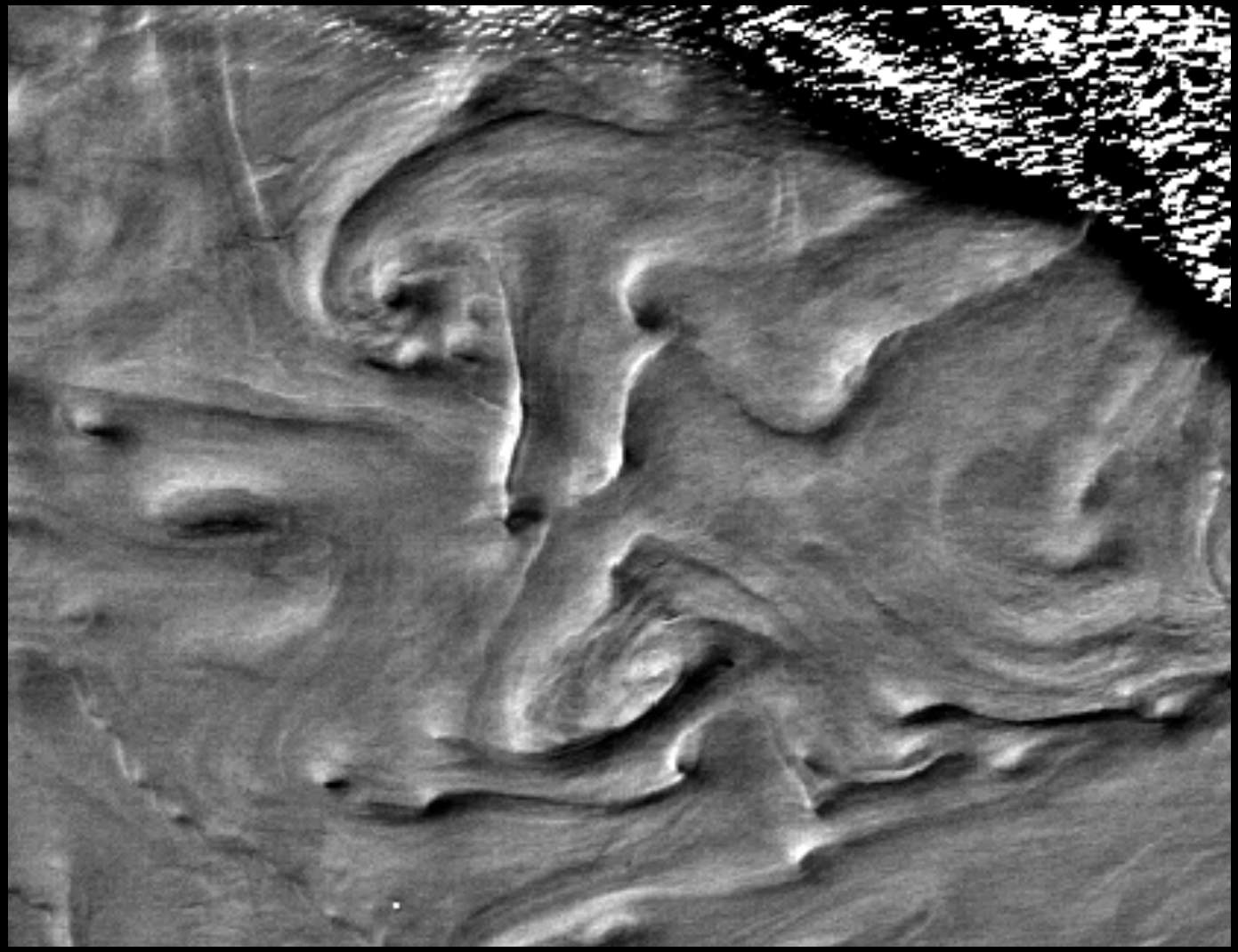
And location from SAR (colors)

Révéler l'enfoui: MERIS Glitter

MODIS Terra Glitter

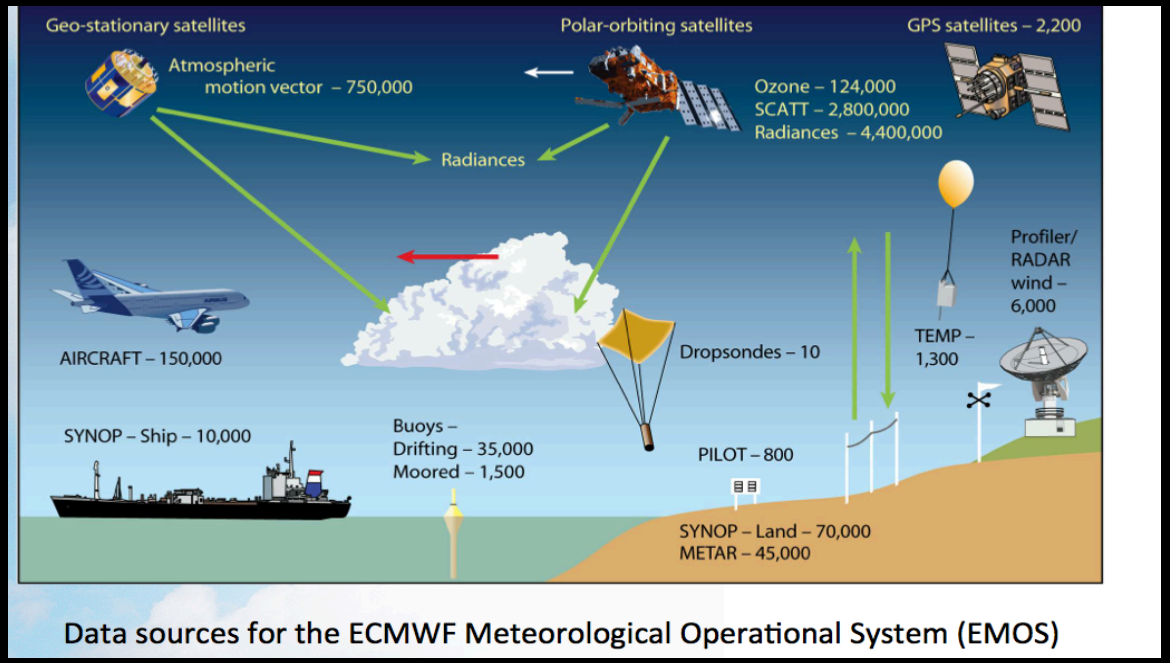


Soupe Turbulente



Ecosystème de données

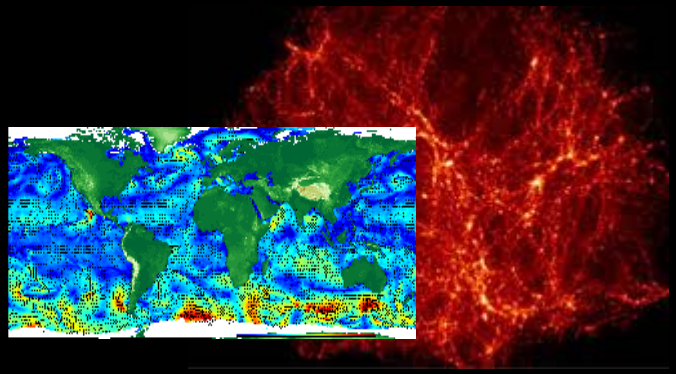
In situ



Data sources for the ECMWF Meteorological Operational System (EMOS)

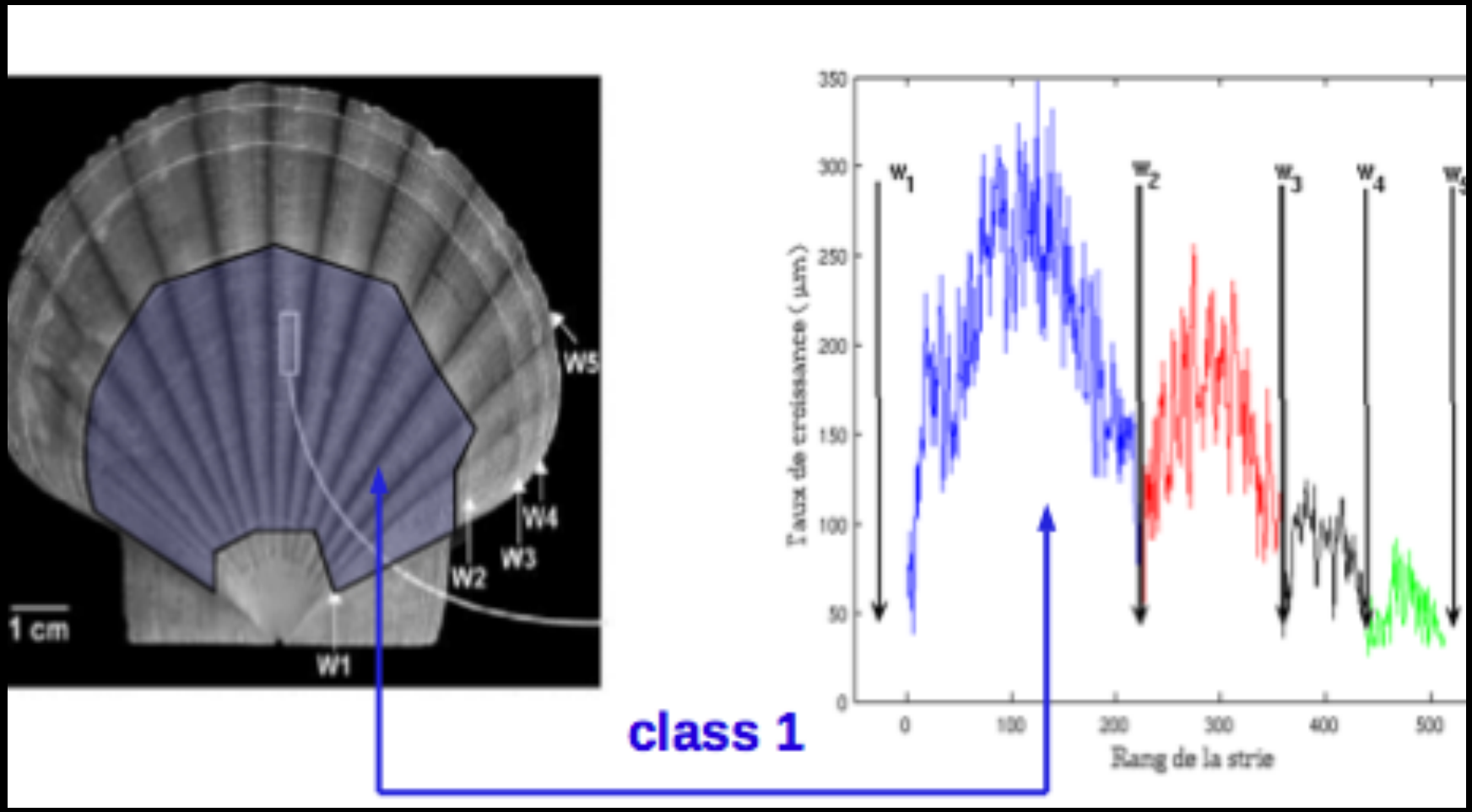
In vivo

Données sociétales et économiques



In silico

Accumulation: croisement avec autres bases de données



« Data intensive science : le 4ème paradigme »

Comment faire face au déluge de données d'observation, de modélisation, de simulation, économiques ou sociales ?

Les nouveaux défis des centres de données

- Permettre l'accès rapide à des archives complètes de données
- Supprimer les transferts de données et la duplication
- Minimiser l'effort de l'idée à l'implémentation
- Permettre le traitement de masse (« embarassingly parallel »)

- Réduire/synthétiser les masses de données
- Indexer et rechercher les informations pertinentes

- Comment exploiter ou adapter les nouvelles technologies de l'ère numérique (cloud computing, big data) ?
- Coût réduit
- De nouveaux modèles économiques

Finalement … (il y a déjà 4 ans)

An ideal instrument … (cloud-free, wide-swath, high-resolution, topography, roughness, Doppler, emissivity, reflectance, …) = the combined use of observations

Improved technologies (instruments, resolution, computer capabilities, storage, dissemination) all contribute to improved combined analysis

Theoretical and dynamical frameworks must be used to assess the quiddity, causes, contexts and essences of the different observations (including sensor physics, observability conditions and instrument capabilities)

Development of future observing systems (including in situ) to capitalize on such a wealth: new analyzing tools and improved dynamical frameworks shall complement the definitions of new sensors.

Et encore …

- Thematically-driven Mining applications shall rapidly emerge to avoid the data deluge, and to emphasize the synergy between observations (in situ and satellite), numerical simulations and theoretical developments
- 'collaborative' efforts to promote future developments to avoid (limit) computation burden and/or (redundant) archive volume growth.
- Data on an EO-'cloud' and software utilities/ applications more efficiently developed to search, process, visualize, analyze the data in a common approach.
- Usual discussions – the need for standard data formats, metadata conventions, open access etc.

Analyse de très grands ensembles de données satellitales ; détection de signaux guidée par la formalisation des processus dynamiques ?