CERFACS

CENTRE EUROPÉEN DE RECHERCHE ET DE FORMATION AVANCÉE EN CALCUL SCIENTIFIQUE

The stakes and prospects of “Data Driven Modeling” at CERFACS

O. Thual, E. Emanuele, C. Pagé, S. Ricci, E. Sanchez,
M. Rochoux, O. Guillet, A. Weaver, C. Lambert…

Outline

CERFACS has a long-established record of excellence in **environmental and industrial Computational Fluid Dynamics** for complex flow simulation on high-resolution grid enhanced by continuous developments in numerical models and in High Performance Computing.

Data Assimilation of satellite data for ocean, atmospheric chemistry or hydraulics modeling is also one of its strong expertise domains.

Uncertainty Quantification has become a developing field based on ensemble approaches and model-reduction objectives.

Based on these expertise domains, a new challenge for CERFACS is to develop a Data Driven Modeling axis combining **Data Science**, Uncertainty Quantification and Data Assimilation.

CERFACS expertise

CERFACS has a long-established record of excellence in **environmental and industrial Computational Fluid Dynamics** for complex flow simulation on high-resolution grid enhanced by continuous developments in numerical models and in High Performance Computing.

Data Assimilation of satellite data for ocean, atmospheric chemistry or hydraulics modeling is also one of its strong expertise domains.

Uncertainty Quantification has become a developing field based on ensemble approaches and model-reduction objectives.

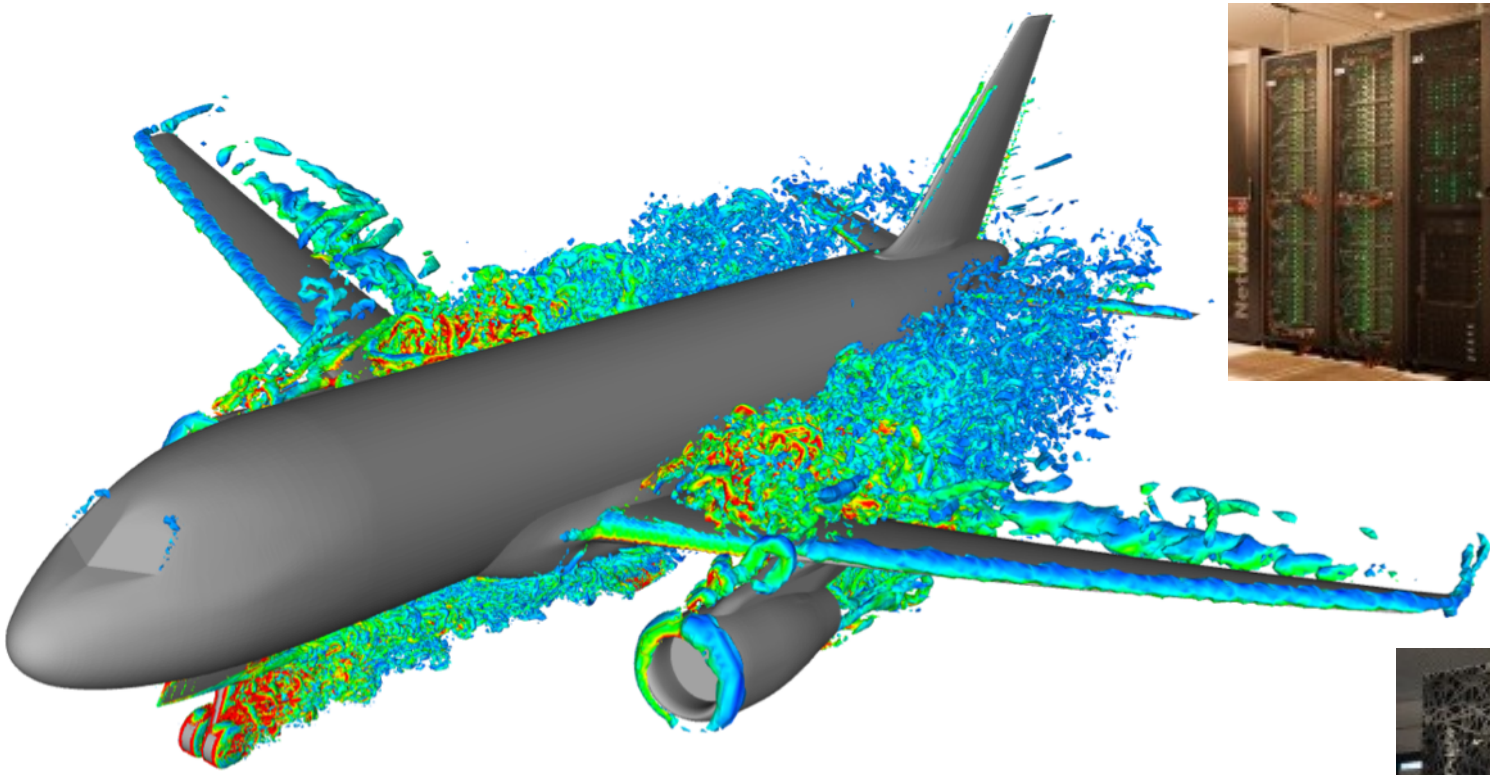
Based on these expertise domains, a new challenge for CERFACS is to develop a Data Driven Modeling axis combining Data Science, Uncertainty Quantification and Data Assimilation.

CERFACS: European Centre of Research and Advanced Training in Scientific Computing

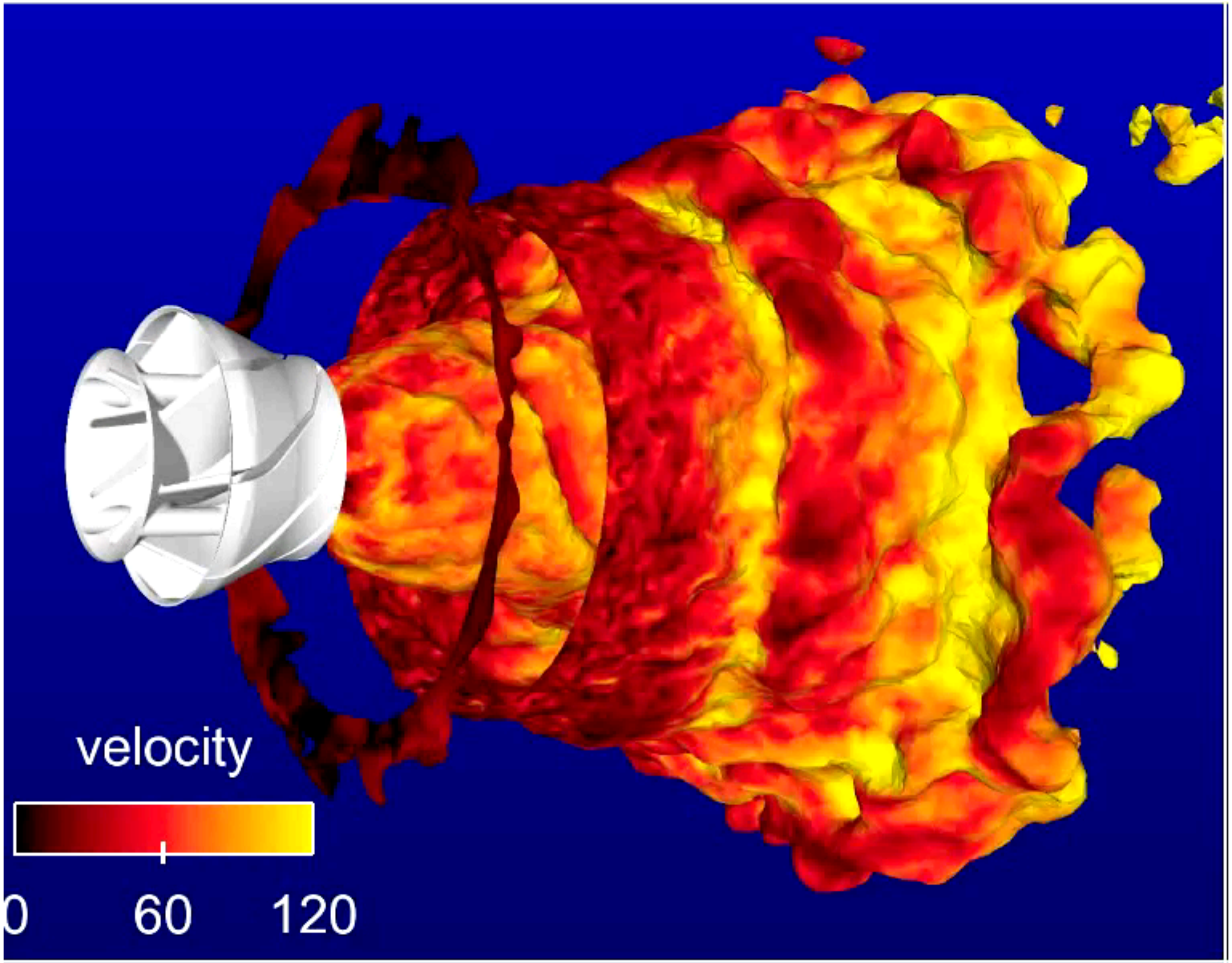
- Scientific and technical researches in order to improve advanced computing methods
- Transfer of scientific knowledge and technical methods for industrial applications

AIRBUS GROUP · cnes CENTRE NATIONAL D'ÉTUDES SPATIALES · eDF · METEO FRANCE · ONERA THE FRENCH AEROSPACE LAB · SAFRAN AEROSPACE · DEFENCE · SECURITY · TOTAL

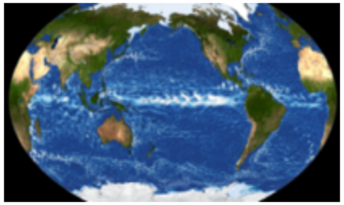
High Performance Computing for aerodynamics



High Performance Computing for combustion

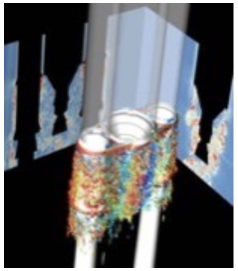


CERFACS Strategic Research Plan

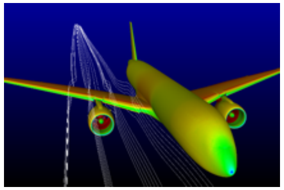


Climate variability and predictability: from ocean to continental impacts

Methane-Lox engine simulation



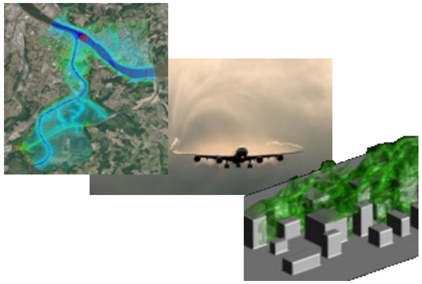
Full aircraft simulation



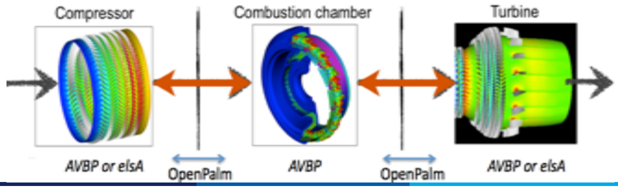
Linear algebra	Numerical methods for PDE
Coupling	Data driven modelling

Exascale

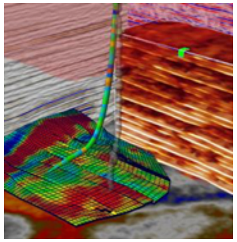
Modelling for environment and safety



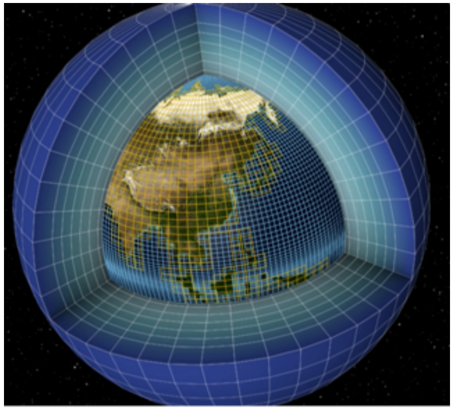
Full gaz turbine simulation



Physics of oil reservoirs (including history matching)



High Performance Computing for climate

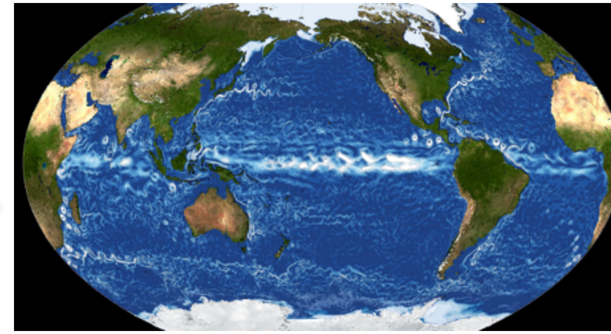


$$\frac{D\rho}{Dt} = -\rho\vec{\nabla}\cdot\vec{V}$$

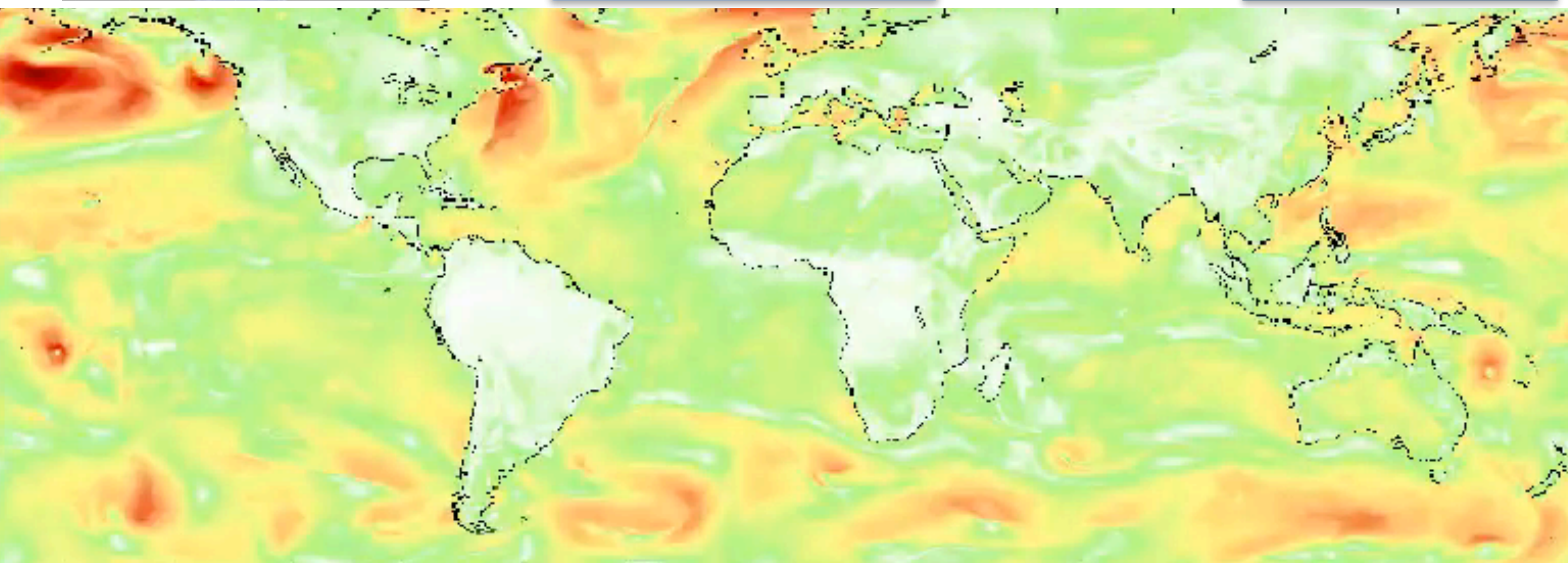
$$\frac{D\vec{V}}{Dt} = \frac{-\vec{\nabla}p}{\rho} - 2\vec{\Omega}\wedge\vec{V} + \vec{g} + \vec{F}$$

$$C_p\frac{DT}{Dt} = \frac{RT}{p}\frac{Dp}{Dt} + Q$$

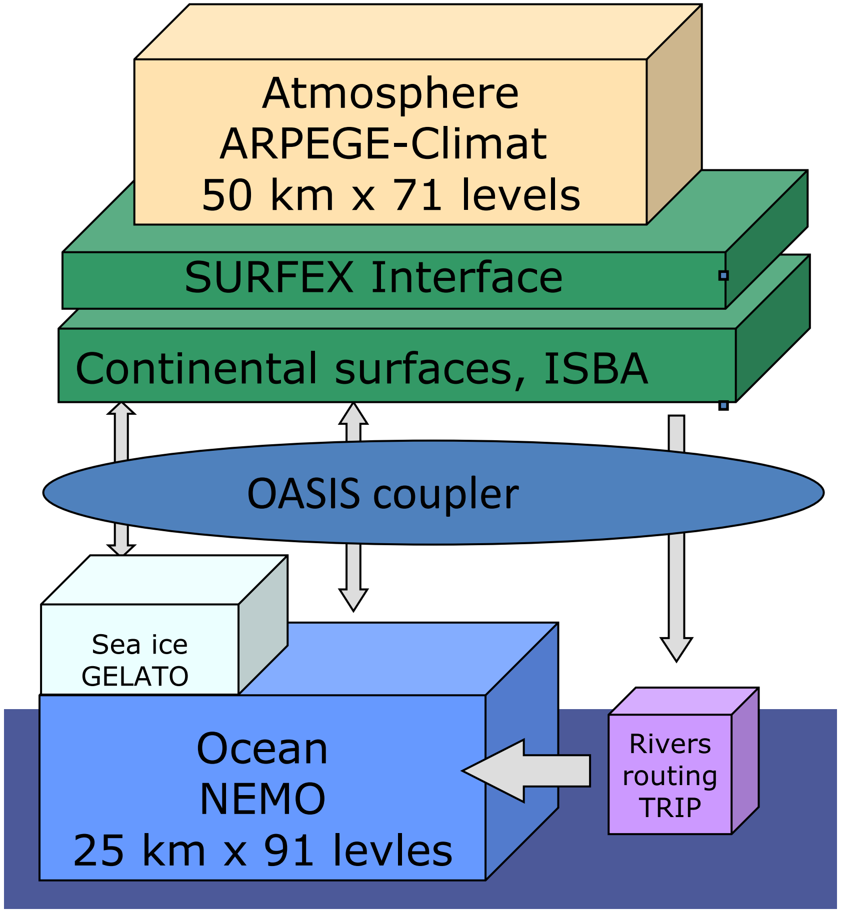
$$\frac{Dq}{Dt} = Q'$$

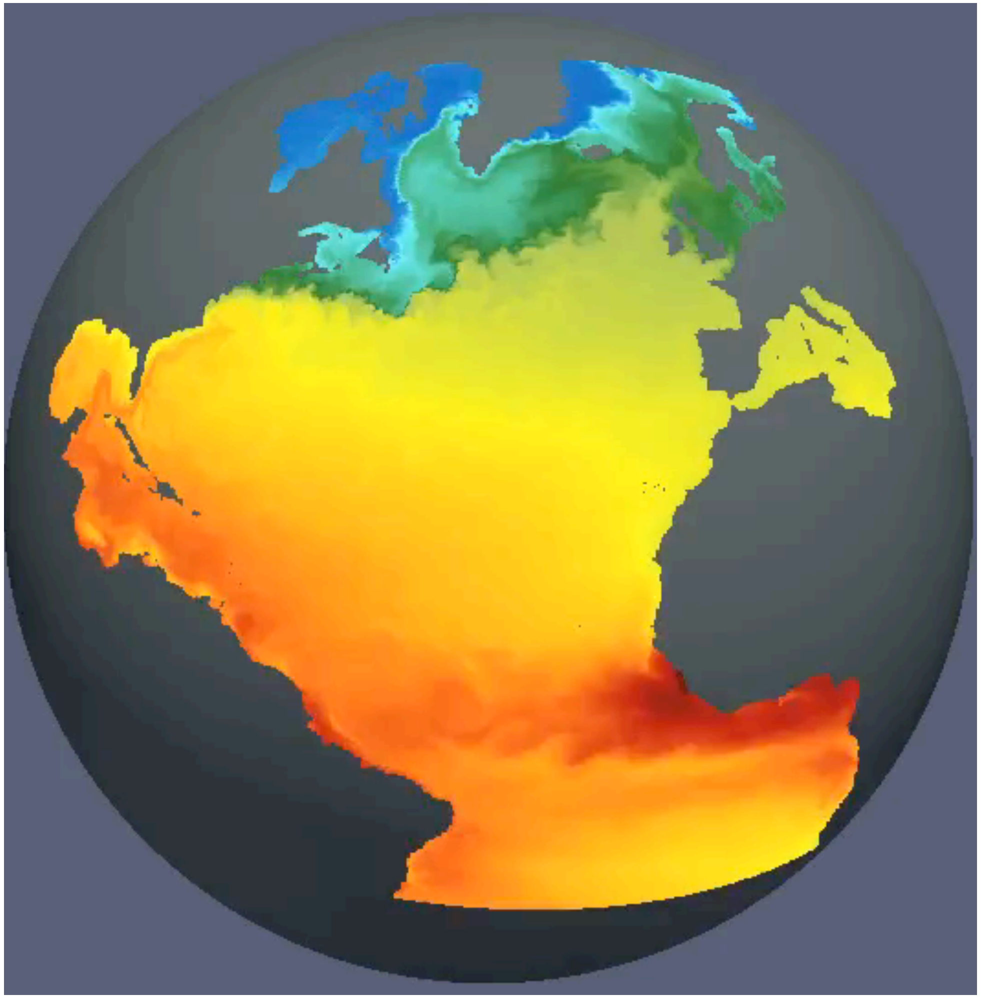


10 m wind, 1994-2003



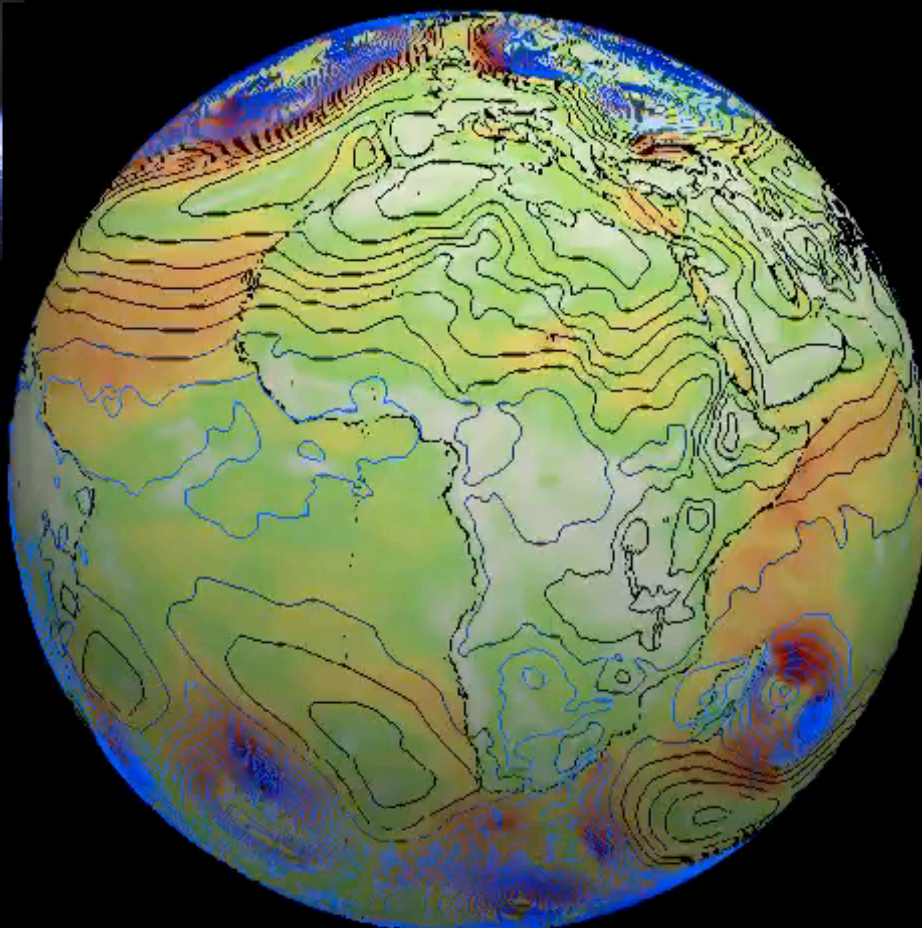
Coupled climate model



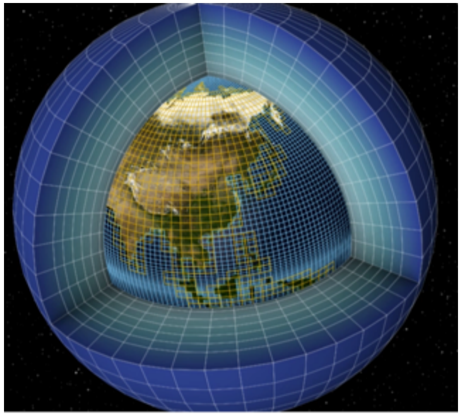


Use of satellite data for climate modelling

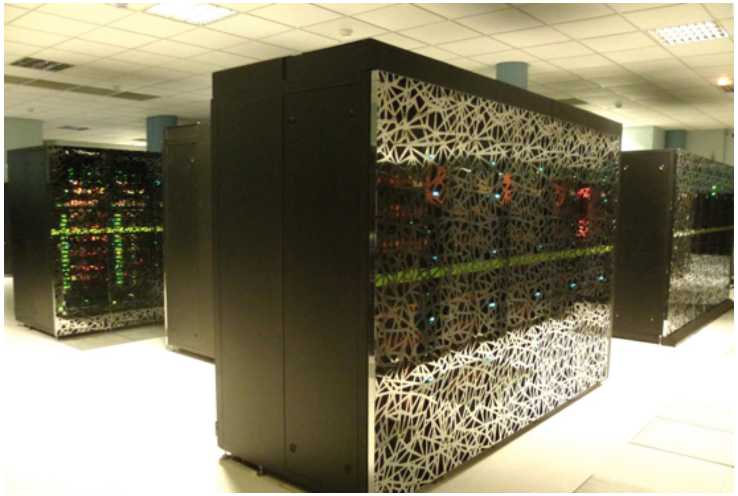
Sea level pressure and wind at 10 m, 2004-2013

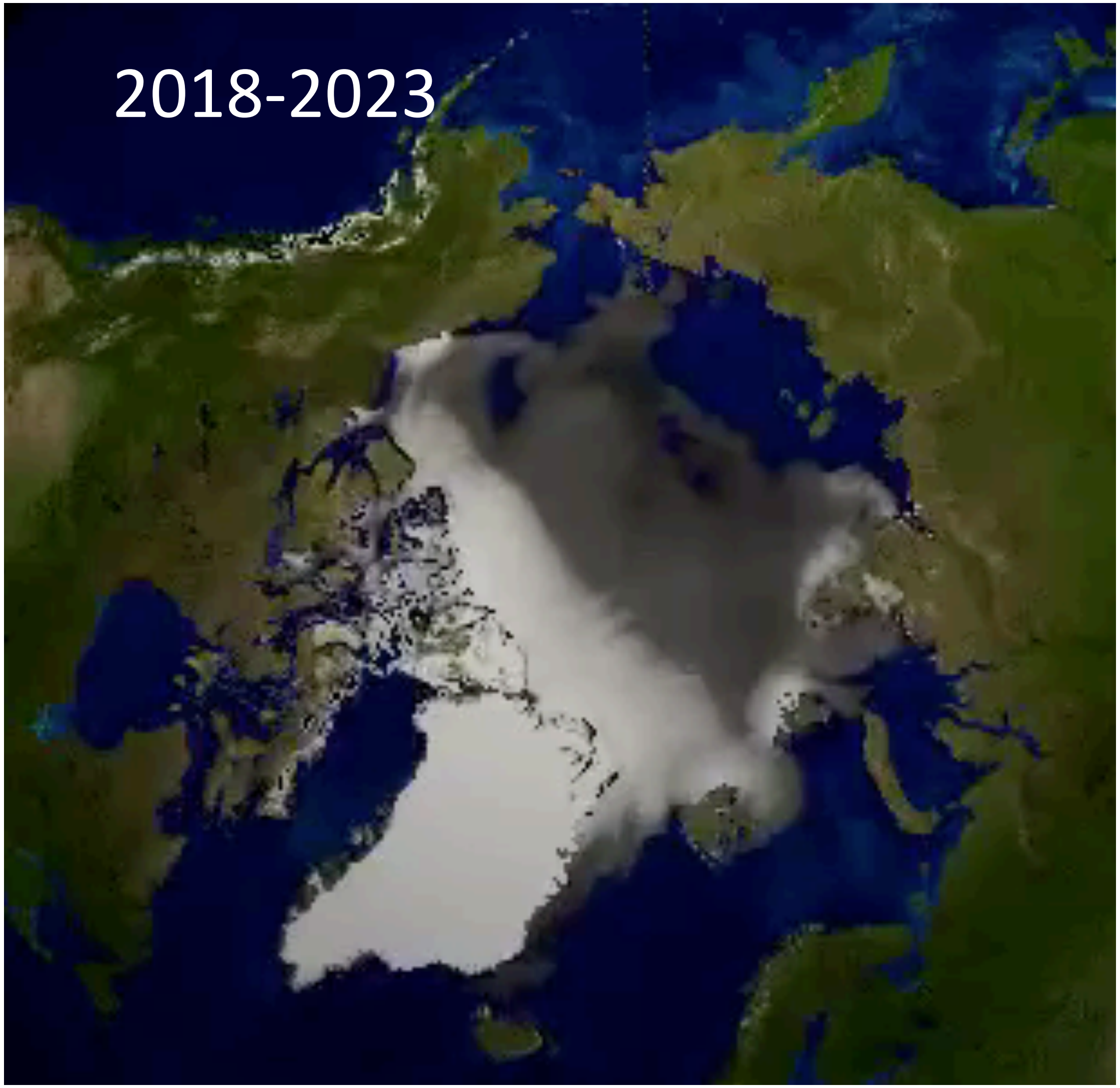


Arctic sea ice modelling

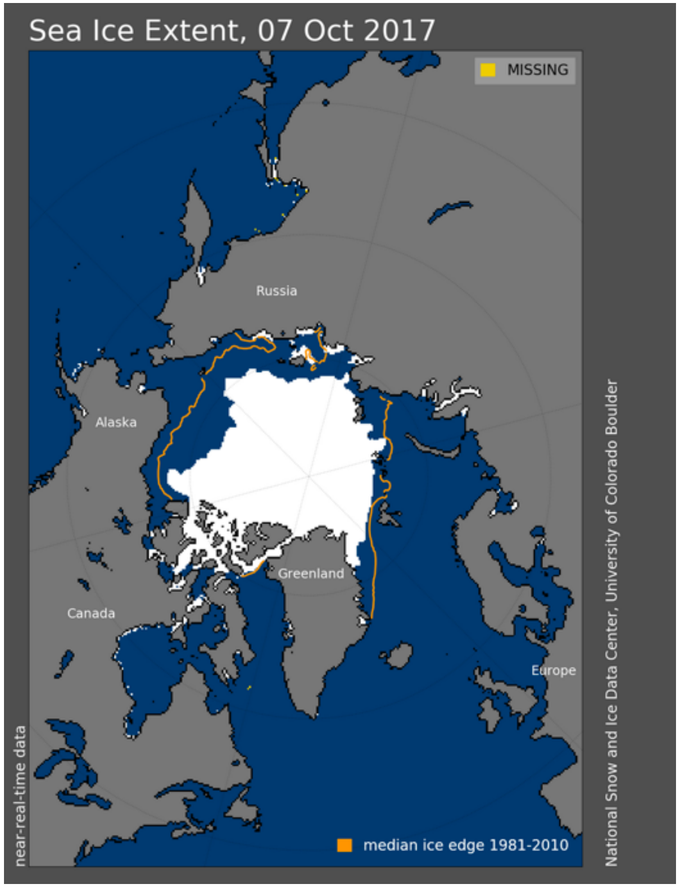


Numerical simulation

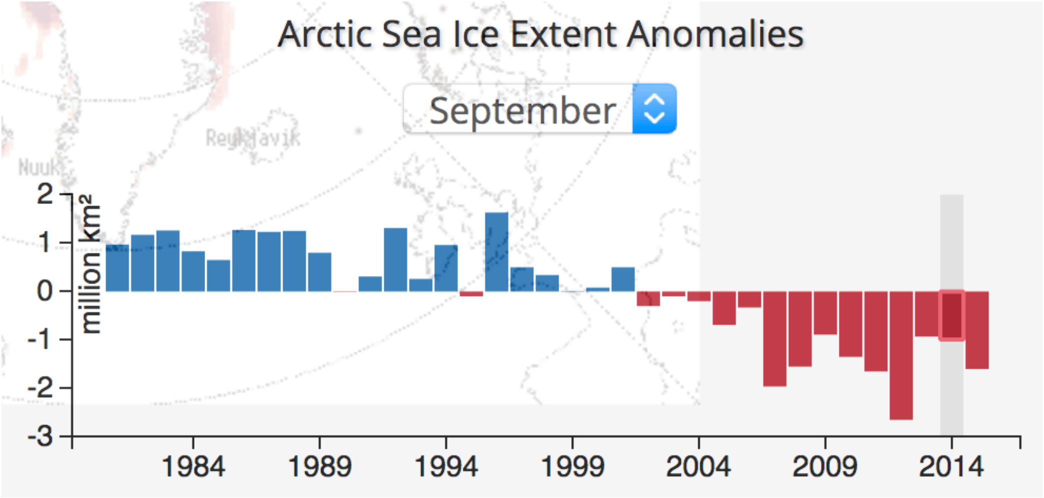




Impact of Artic sea ice changes on the climate system



From NSIDC satellite data (NASA)



Arctic sea-ice is declining and **projected to dramatically decrease** in summer by mid-to-late 21st century in response to greenhouse gases.

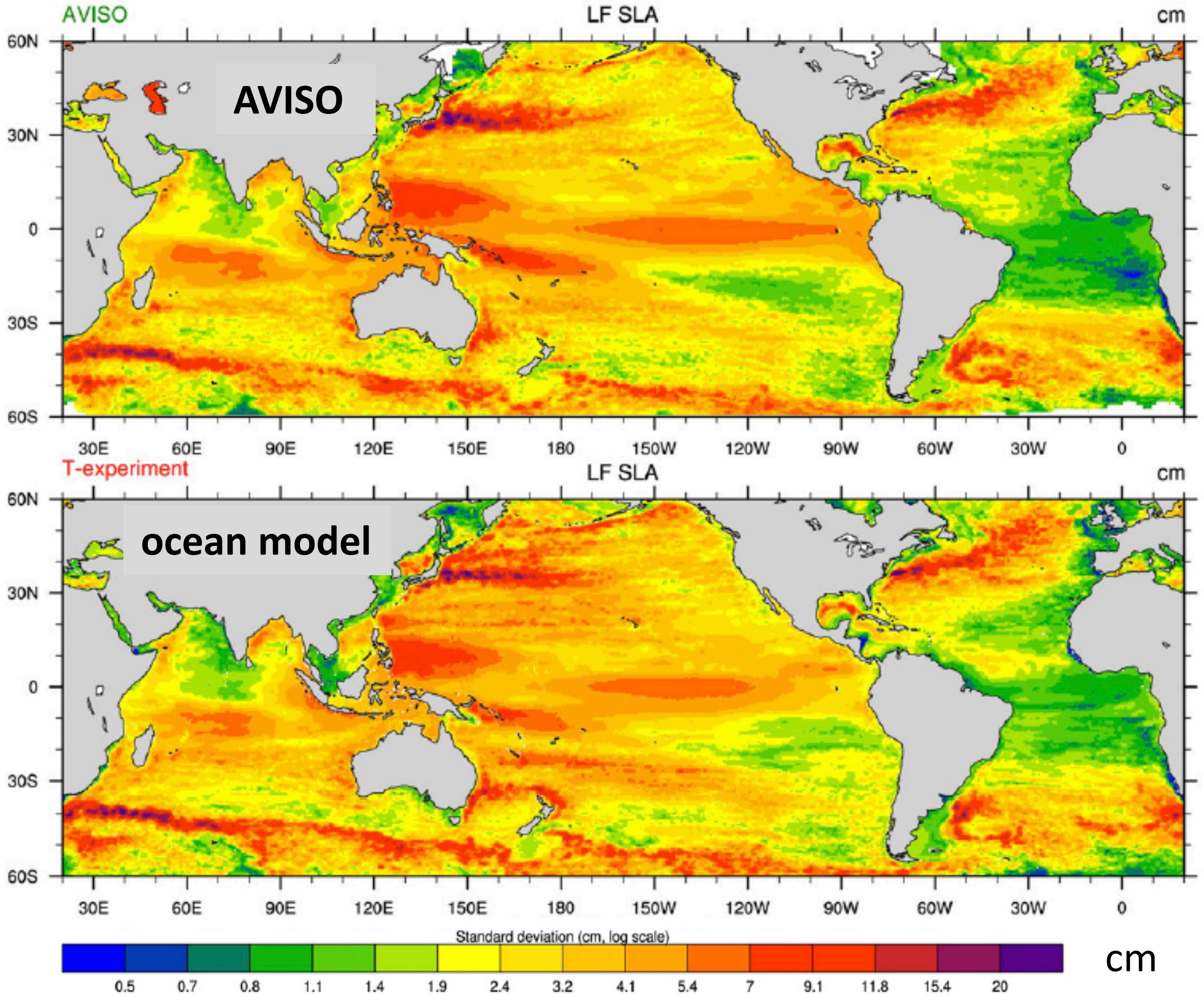
- **Impact on the climate system?**

CERFACS expertise

Ocean variability: data and model

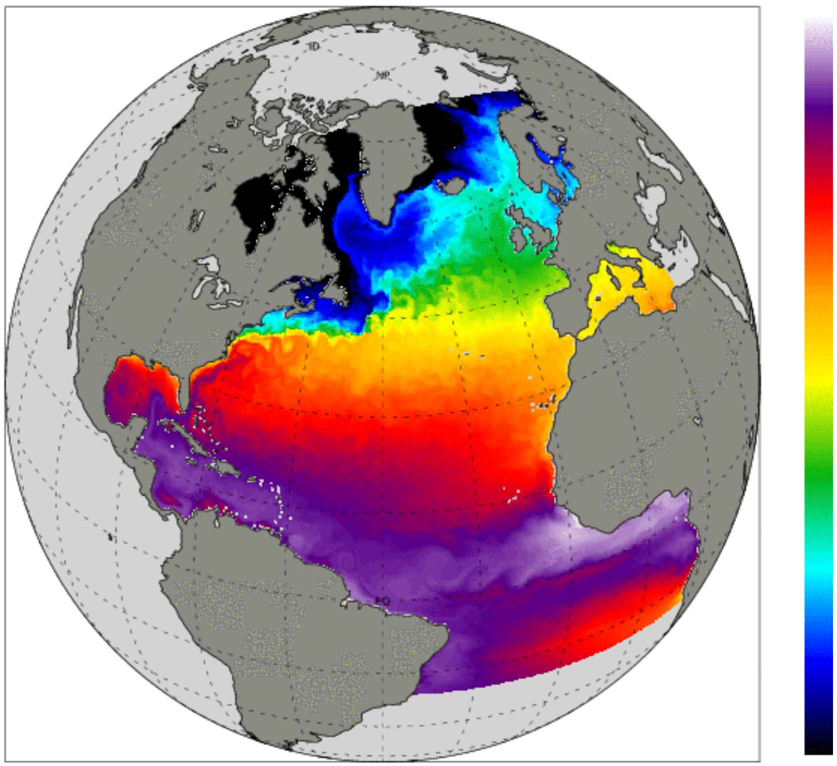
Reference: SLA from AVISO (CNES/CLS/LEGOS)

Sea Level Anomaly (SLA) low frequency variability

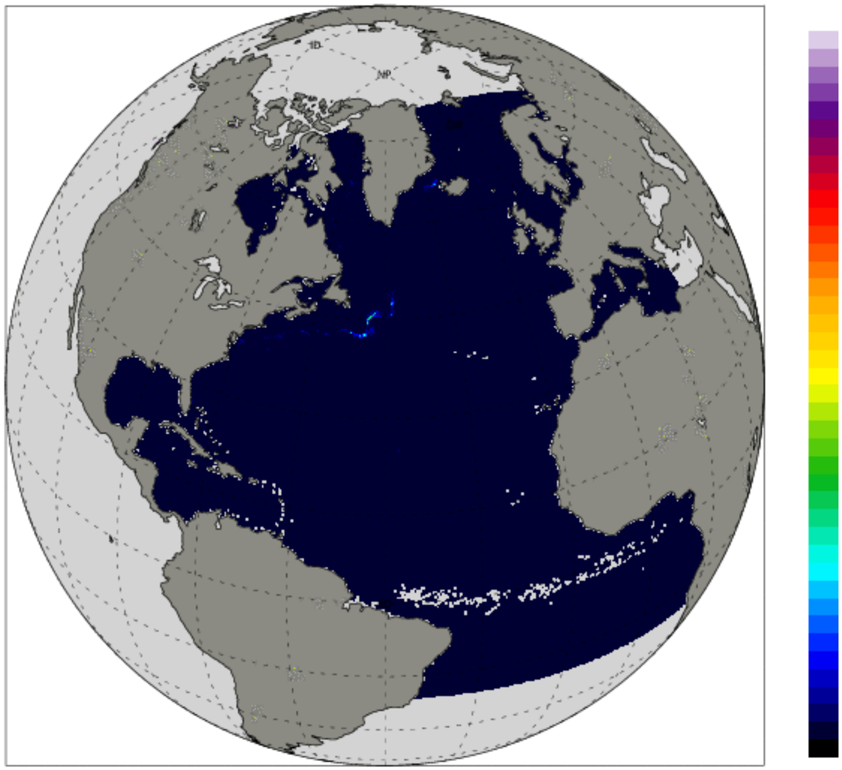


Oceanic intrinsic variability with ensemble runs

Year 1993, Month 1, Day 3



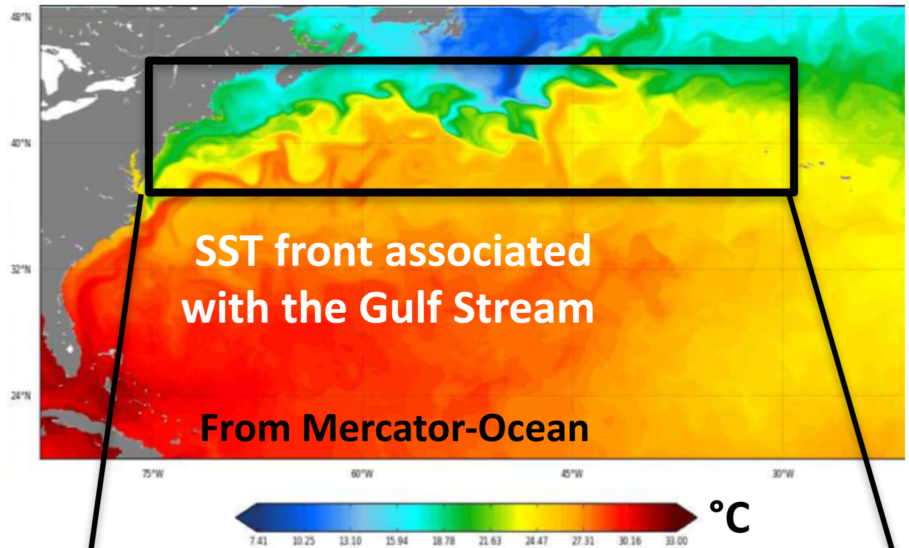
Year 1993, Month 1, Day 3



PRACE

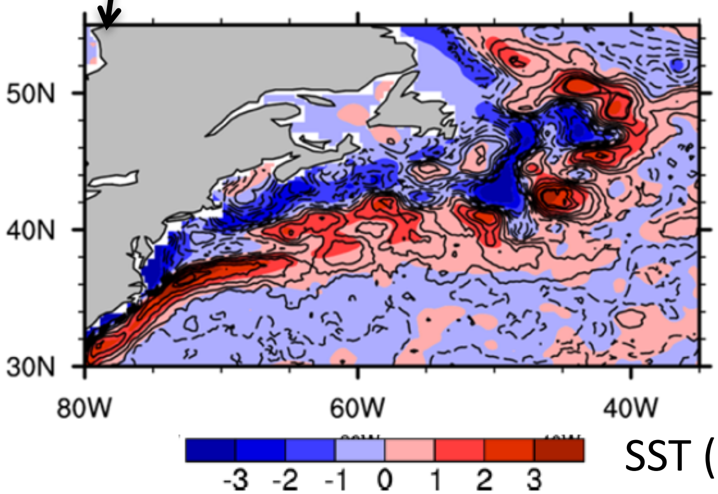
Root mean square of the sea surface temperature for an ensemble simulation with 50 members during 50 years

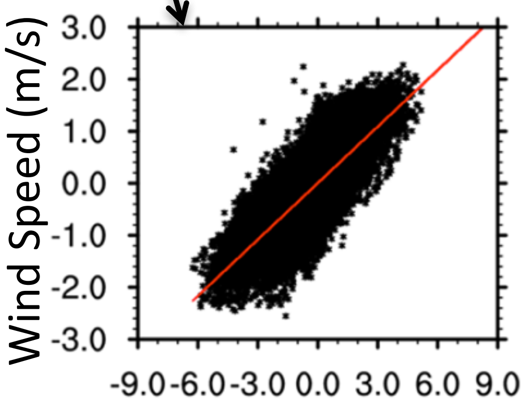
Air-sea coupling at fine scale over SST fronts



High resolution satellite data has allowed to characterise the mesoscale air-sea coupling over the Sea Surface Temperature (SST) frontal regions.

- ❑ **How this coupling is represented in high resolution climate models?**
- ❑ **Impact on simulated climate?**





AMSE-R SST (colors) and QuickSCAT surface winds (contours)

Data Assimilation

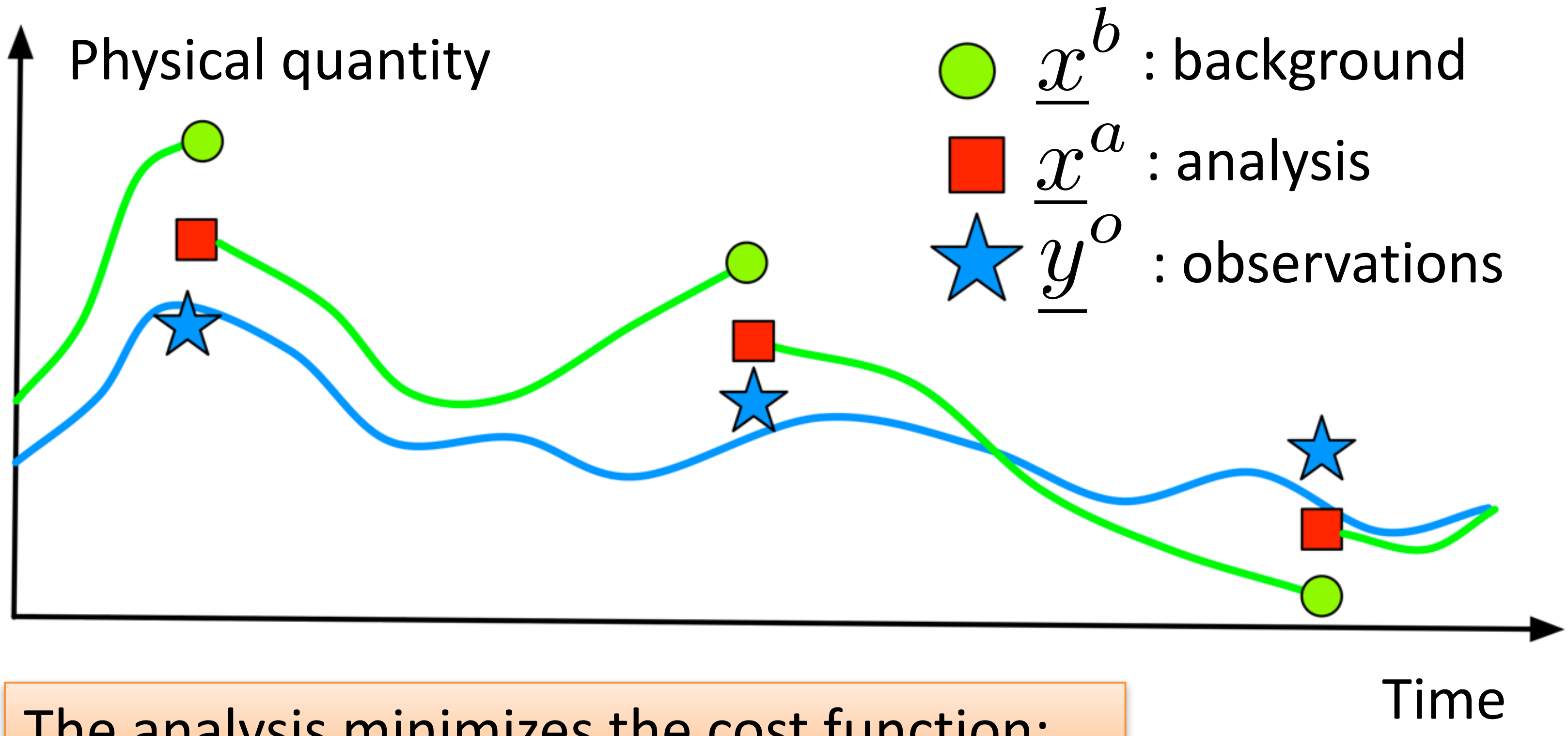
CERFACS has a long-established record of excellence in environmental and industrial Computational Fluid Dynamics for complex flow simulation on high-resolution grid enhanced by continuous developments in numerical models and in High Performance Computing.

Data Assimilation of satellite data for ocean, atmospheric chemistry or hydraulics modeling is also one of its strong expertise domains.

Uncertainty Quantification has become a developing field based on ensemble approaches and model-reduction objectives.

Based on these expertise domains, a new challenge for CERFACS is to develop a Data Driven Modeling axis combining Data Science, Uncertainty Quantification and Data Assimilation.

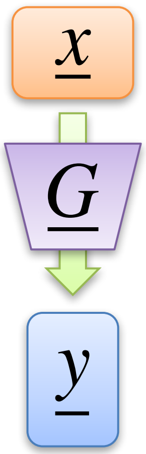
Principles of data assimilation



The analysis minimizes the cost function:

$$2\,J(\underline{x}) = \|\underline{x} - \underline{x}^b\|^2_{\underline{\underline{B}}^{-1}} + \|\underline{y}^o - \underline{G}(\underline{x})\|^2_{\underline{\underline{R}}^{-1}}$$

Gaussian data assimilation



$\underline{x}^b$: background

$\underline{x}^a$: analysis

$\underline{G}$: observation operator

$\underline{y}^o$: observations

$$J(\underline{x}^a) = \mathrm{Min}\, J(\underline{x})$$

Informations are weighted with respect to their uncertainties

$$2\,J(\underline{x}) = (\underline{x} - \underline{x}^b)^T\, \underline{\underline{B}}^{-1}\, (\underline{x} - \underline{x}^b) + [\underline{y}^o - \underline{G}(\underline{x})]^T\, \underline{\underline{R}}^{-1}\, [\underline{y}^o - \underline{G}(\underline{x})]$$

$\underline{\underline{B}}$: background error covariance matrix

$\underline{\underline{R}}$: observation error covariance matrix

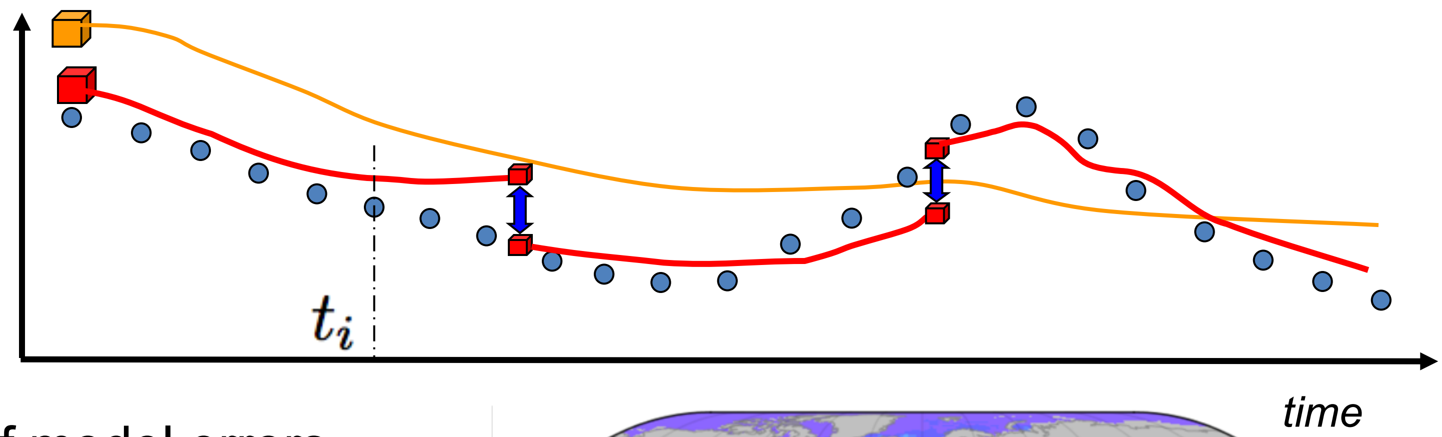
The 4D-Var chain with model error

: background

: analysis

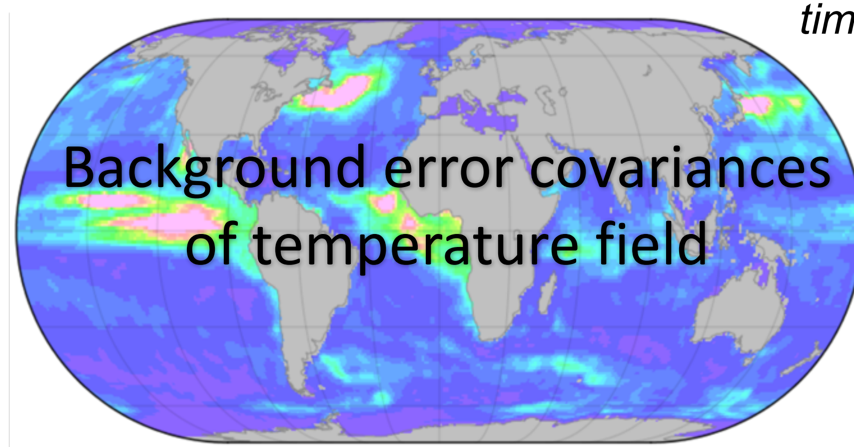
: observations

: trajectories



: corrections of model errors

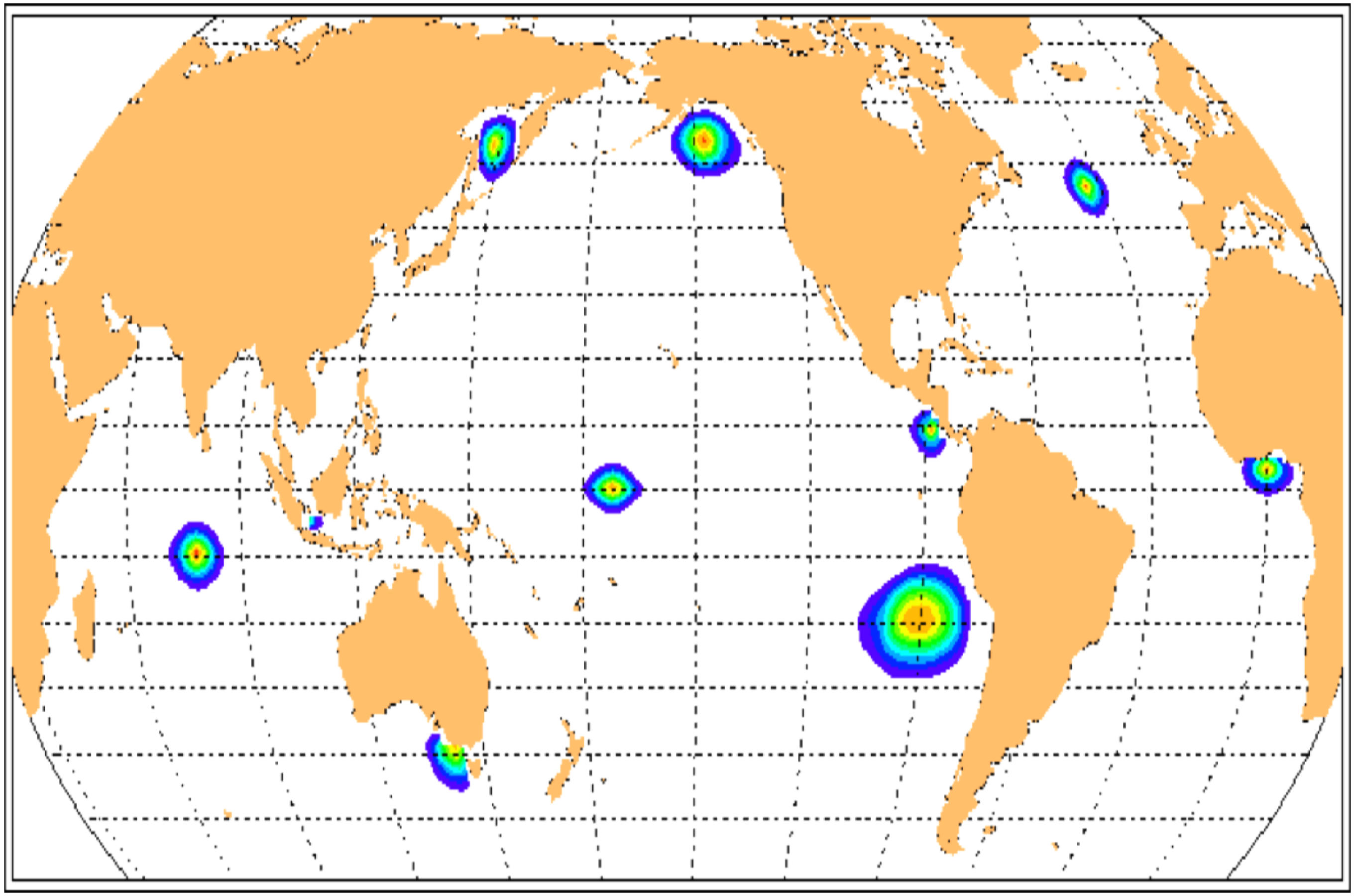
$$J(\underline{x}^a, \underline{q}^a) = \mathrm{Min}\ J(\underline{x}, \underline{q})$$



$$2\,J(\underline{x}, \underline{q}) = \|\underline{x} - \underline{x}^b\|^2_{\underline{\underline{B}}^{-1}} + \sum_i \left\| \underline{y}_i^o - \underline{H}_i \left[\mathcal{M}_i(\underline{x}, \underline{q})\right] \right\|^2_{\underline{\underline{R}}^{-1}} + \|\underline{q}\|^2_{\underline{\underline{Q}}^{-1}}$$

Background error covariance matrix B

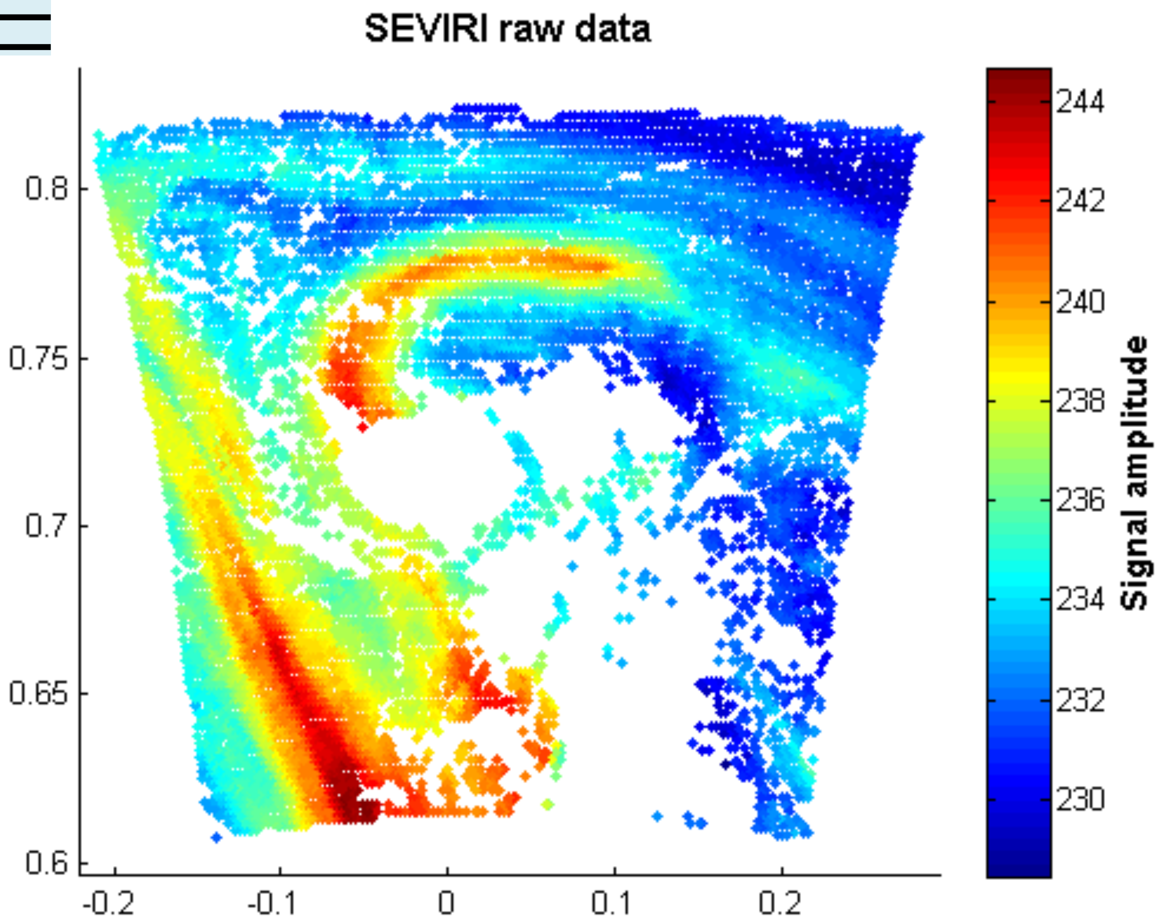
$\underline{\underline{B}}$

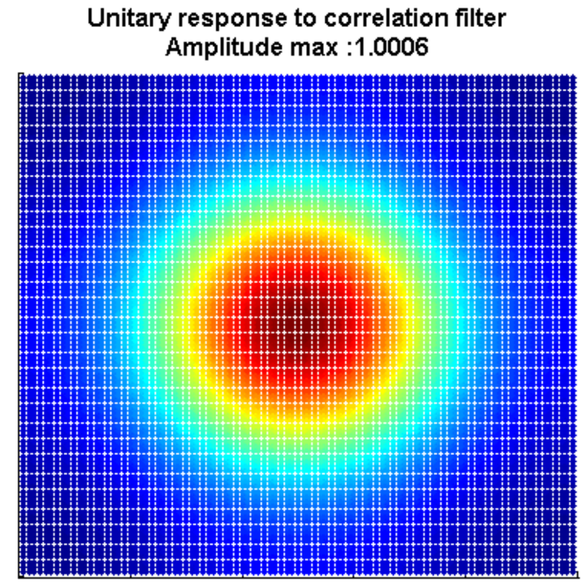


Correlation functions around specific points
Modelling of the B operator with a anisotropic diffusion equation

Modelling the observation error covariance matrix R

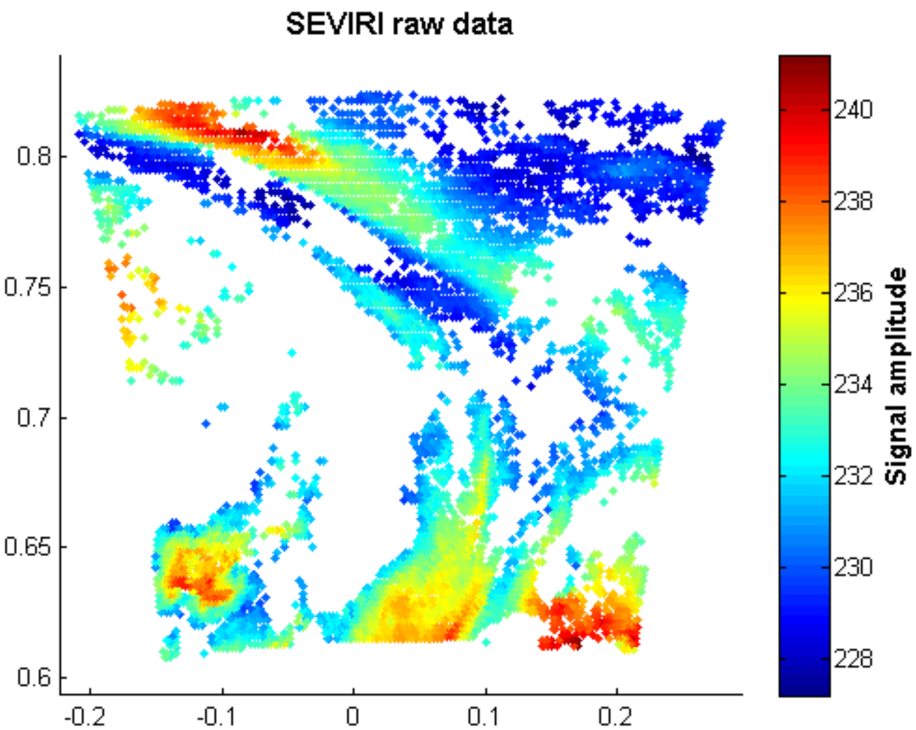
$\underline{\underline{R}}$

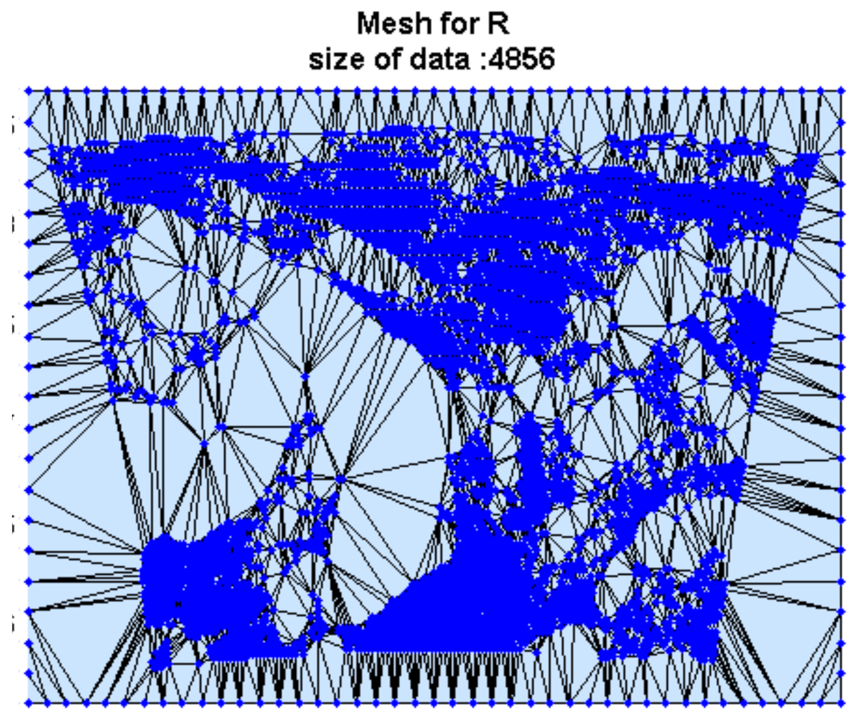




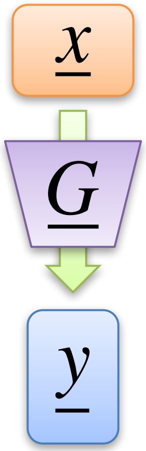
Modelling spatial observation error correlations

$\underline{\underline{R}}$



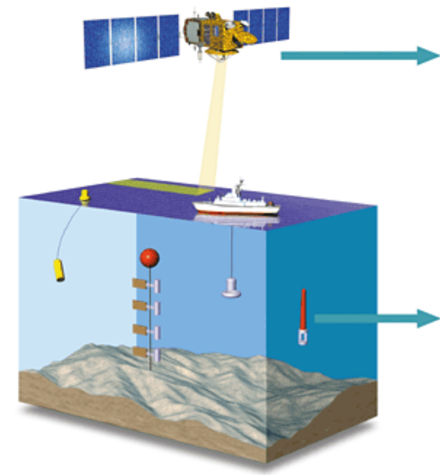


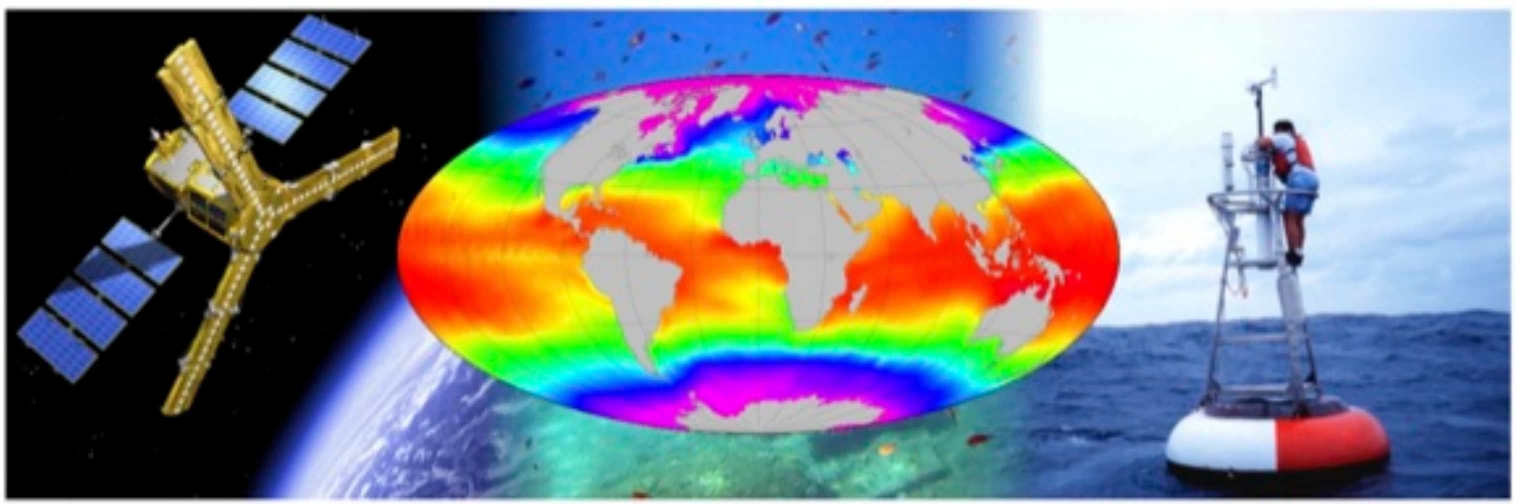
Data assimilation for oceanography



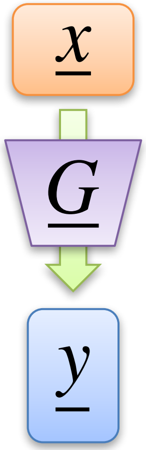
Control space:
- Temperature, salinity and currents (3D)
- Sea Surface Height (2D)
- $O(10^7)$ grid points

Observation space:
- Satellite and in situ data
- $O(10^5)$ measurements



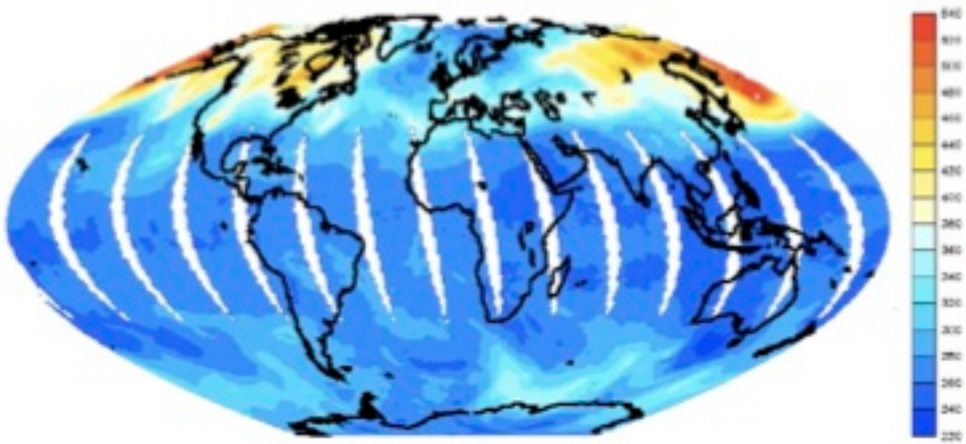


Data assimilation for atmospheric chemistry

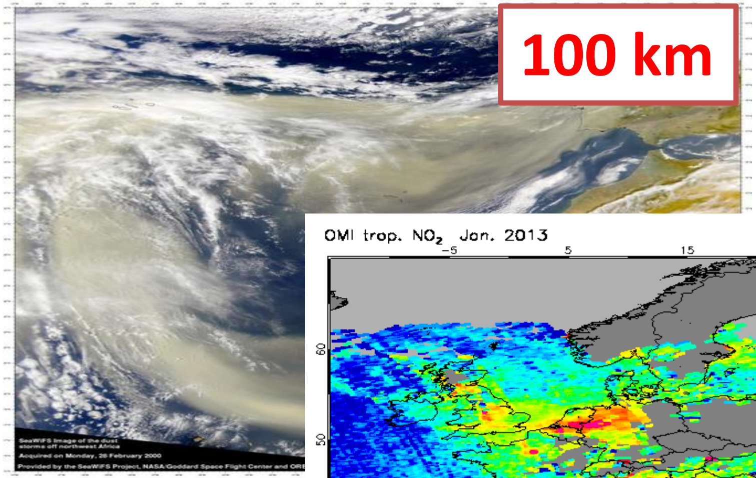


Control space:
- O_3 , NO_2 and other species fields (3D)
- Atmospheric fields (3D)
- O(10^7) grid points

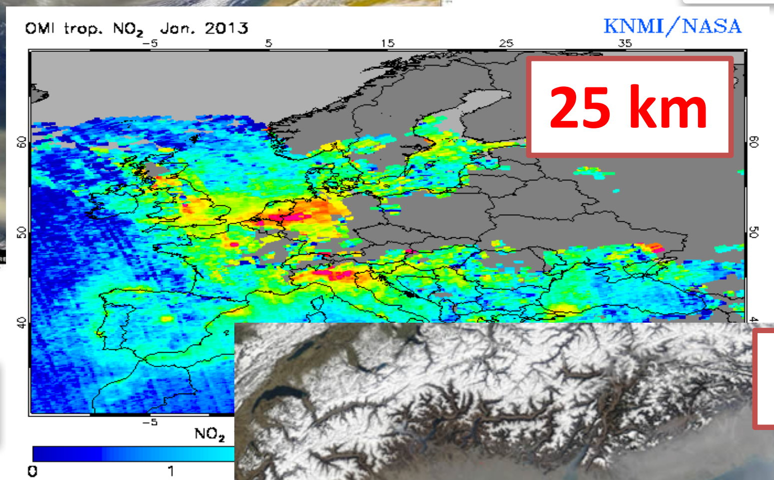
Observation space:
- Satellite data
- In situ-data
- O(10^5) measurments



Scales and use of Earth Observation (EO) for atmospheric data assimilation at CERFACS



100 km



25 km

10 km

1 km

Global scale

Continental scale

Regional scale

Urban scale

MLS (Aura, NASA)

GOME2 (MetOp, EUMETSAT)

OMI (Aura, NASA)

IASI (MetOp, EUMETSAT)

SEVIRI (MSG, EUMETSAT)

MODIS (Terra, NASA)

Upcoming SENTINEL 4, 5 (EUMETSAT)

.....?

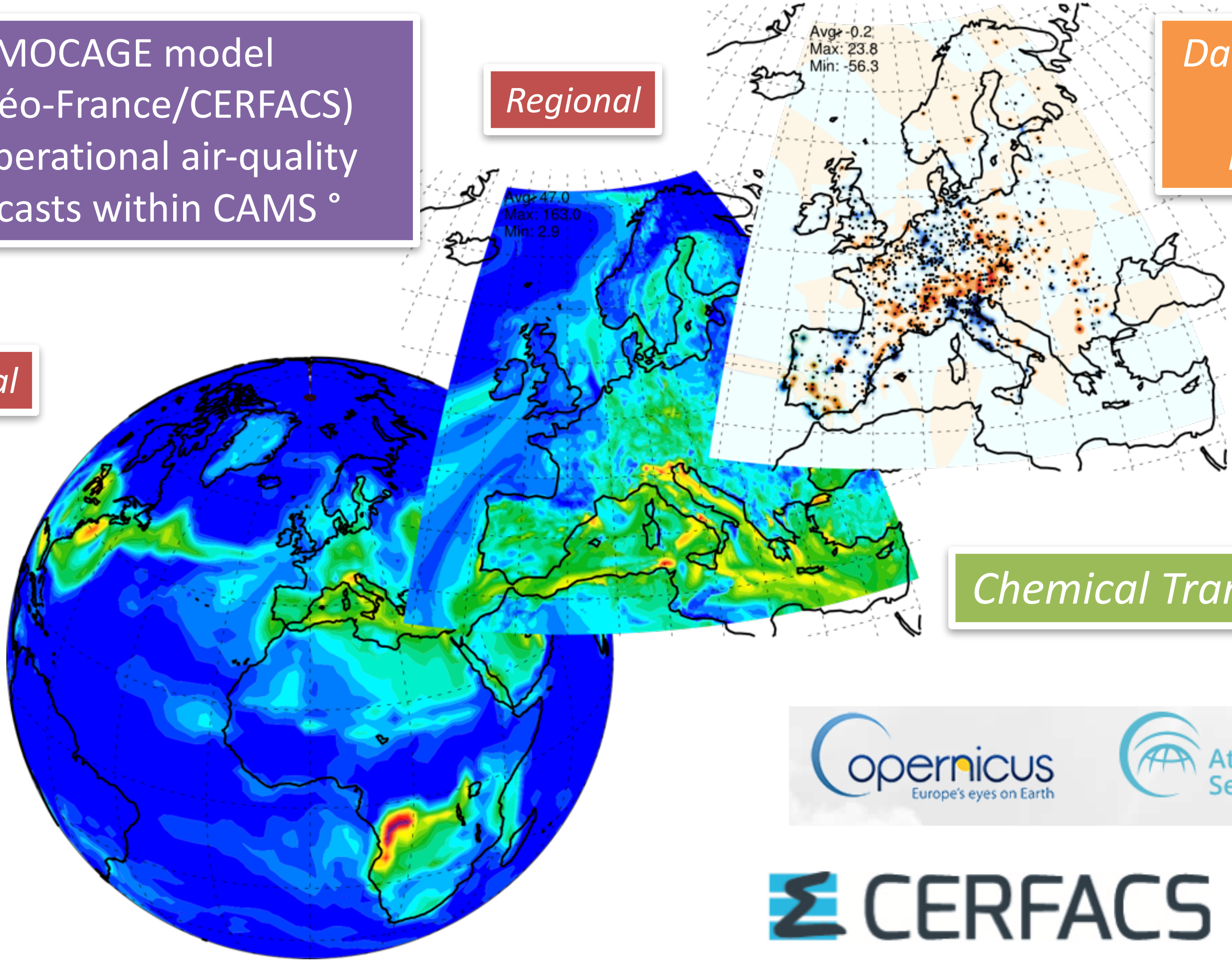
A hierarchy of models of increasing spatial resolution
Copernicus air-quality services

MOCAGE model (Météo-France/CERFACS) for operational air-quality forecasts within CAMS °

Regional

Data assimilation of *in-situ ozone measurements*

Global



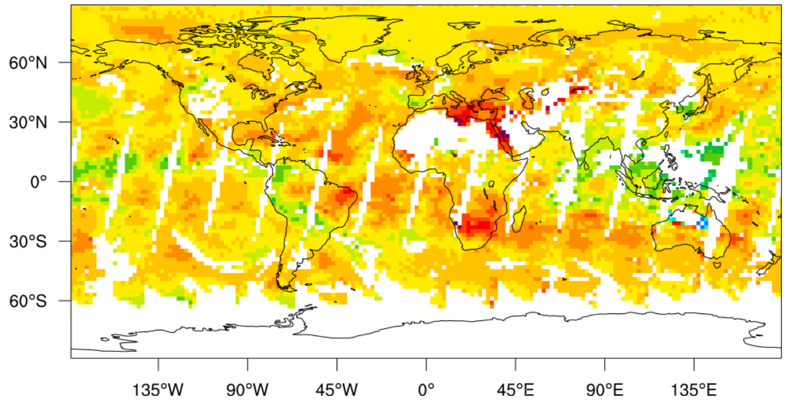
Chemical Transport Model

Copernicus Europe's eyes on Earth — Atmosphere Monitoring Service

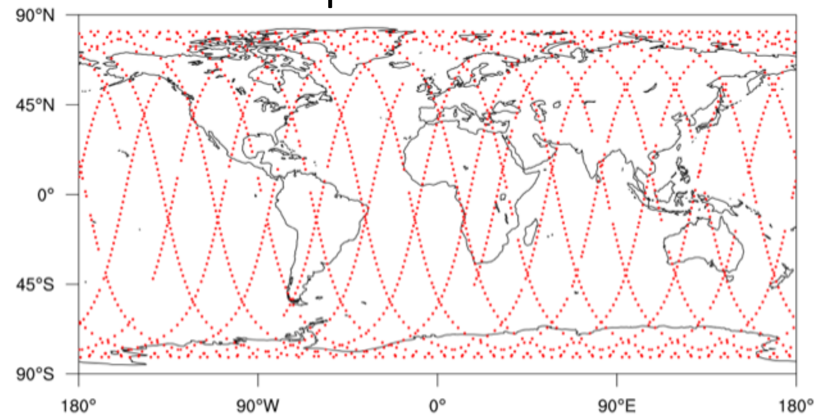
CERFACS — METEO FRANCE Toujours un temps d'avance

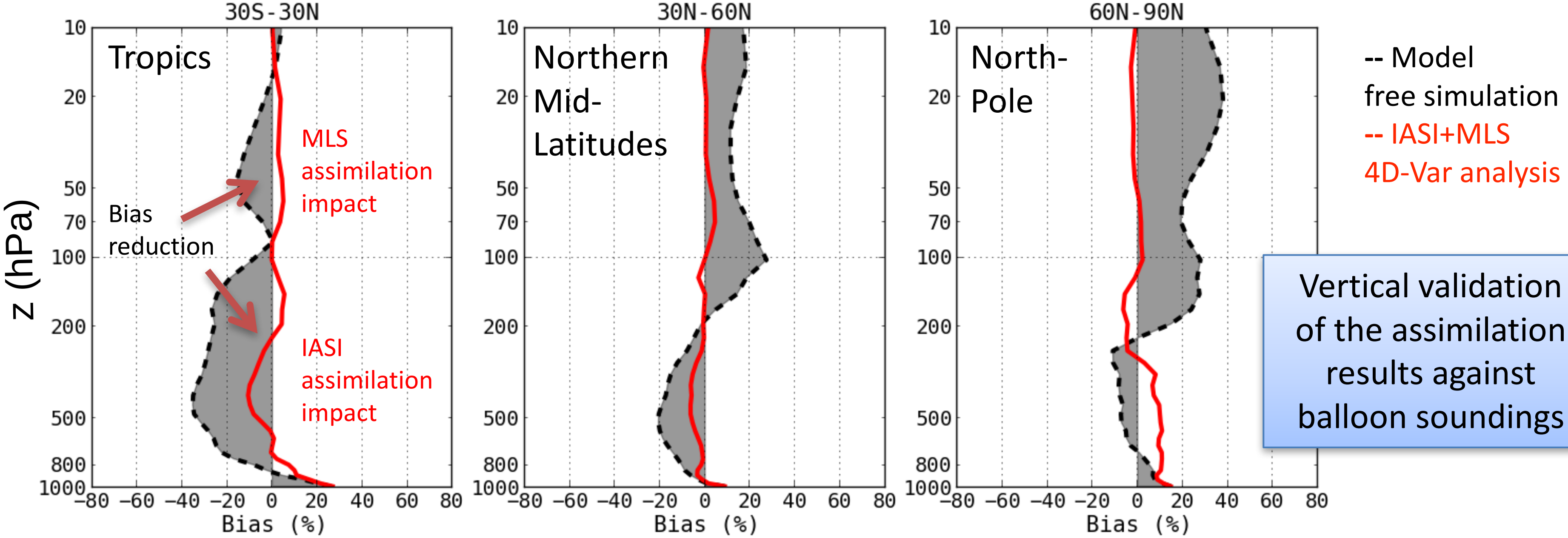
Global assimilation of IASI (MetOp) and MLS (Aura) Level 2 Ozone products

IASI L2 Tropospheric Ozone



MLS footprint on 1-8-2008

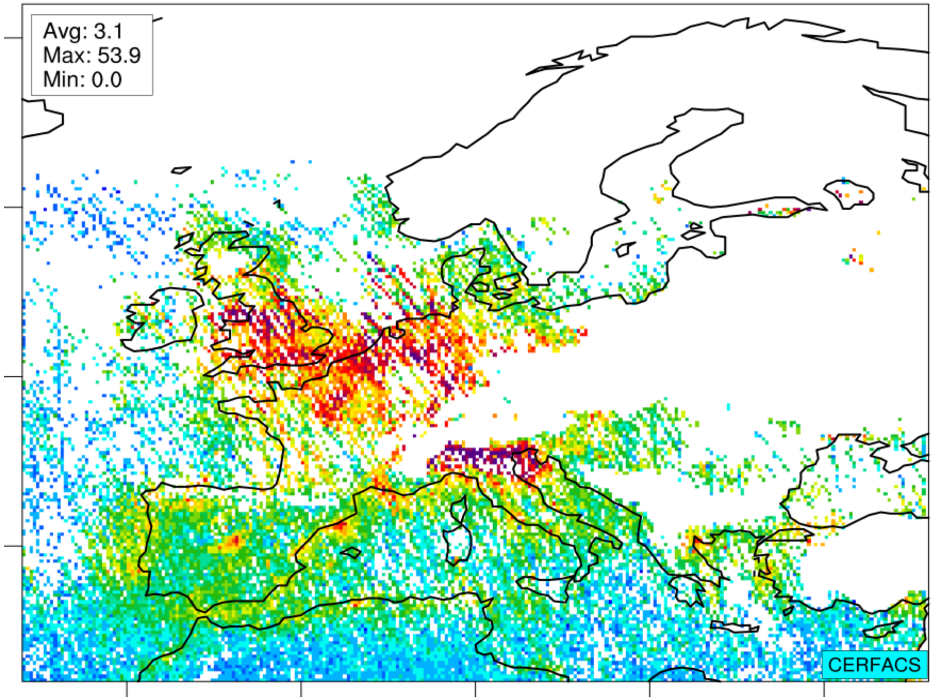




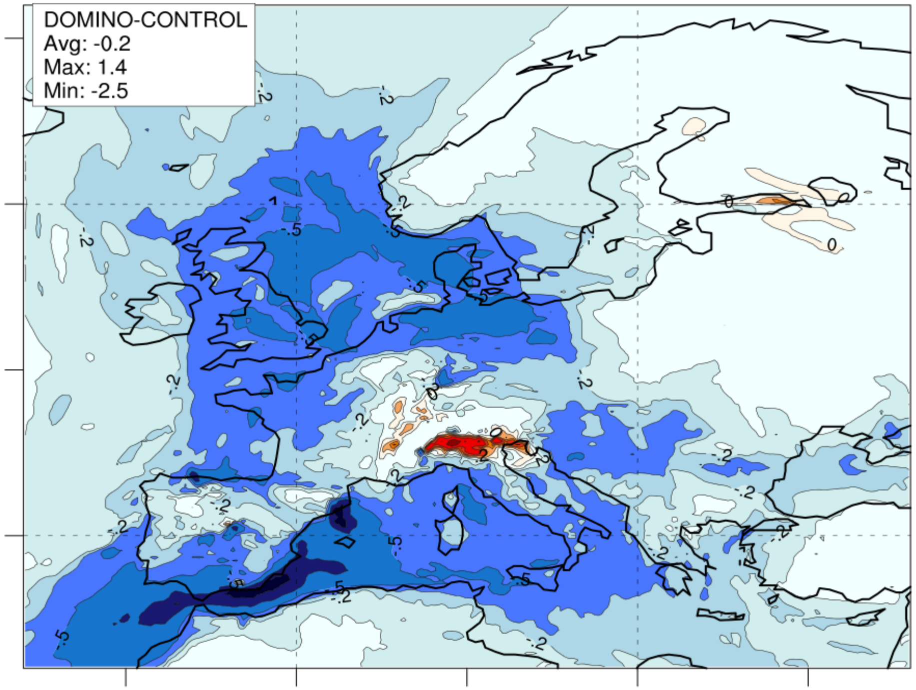
Vertical validation of the assimilation results against balloon soundings

Assimilation of OMI (Aura) NO_2 columns over Europe

DOMINO (KNMI) NO_2 TROP. COLUMN

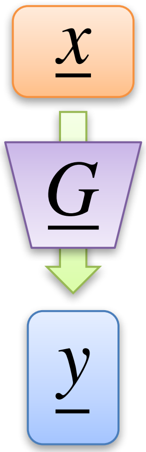


ANALYSIS minus FREE RUN surface NO_2 (ppbv)



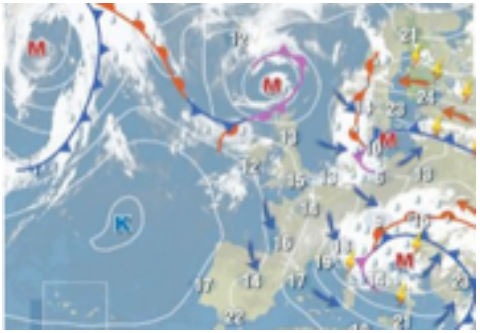
NO_2 short life-time, OMI passes once per day, clouds.

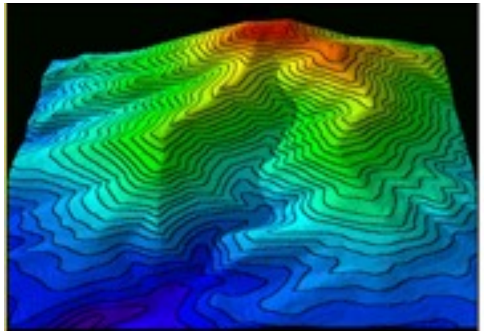
Data assimilation for forest fires



Control space:
- Velocity fields (3D),
- Flamme front (1D)
- Model parameters
- $O(10^5)$ grid points

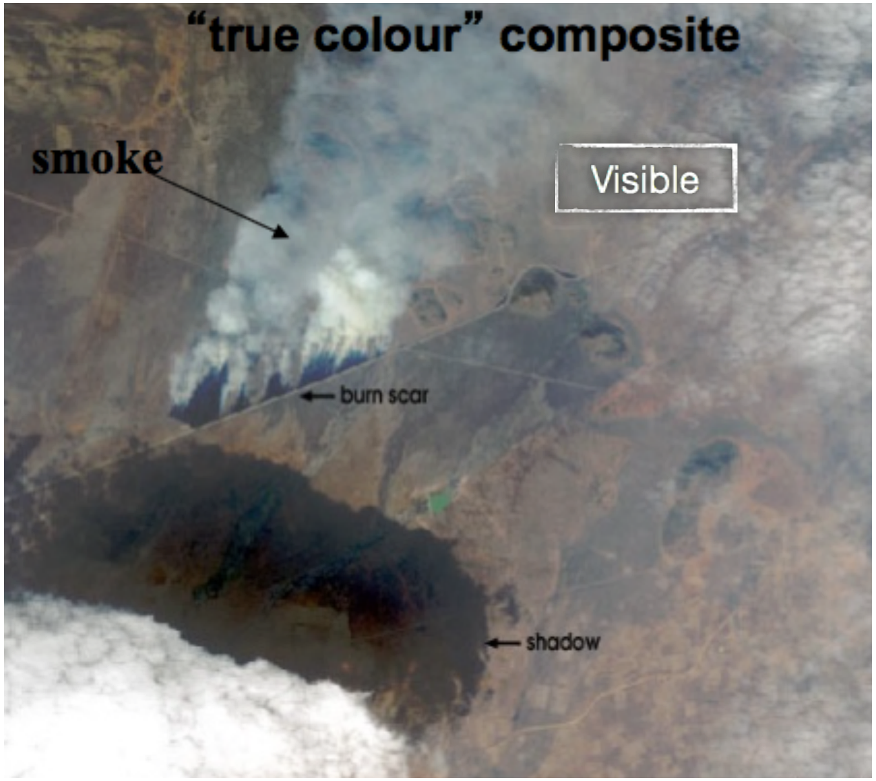
Observation space:
- Satellite images
- Temperatures
- $O(10^4)$ measurements

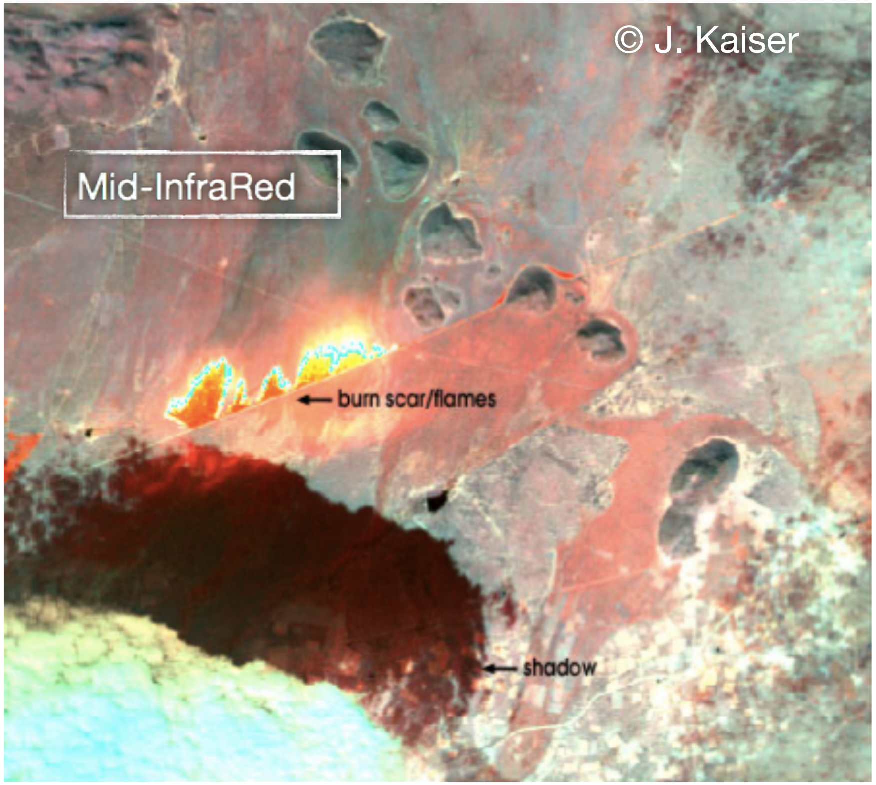

Meteorology

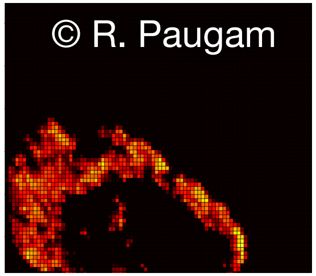

Topography


Vegetation

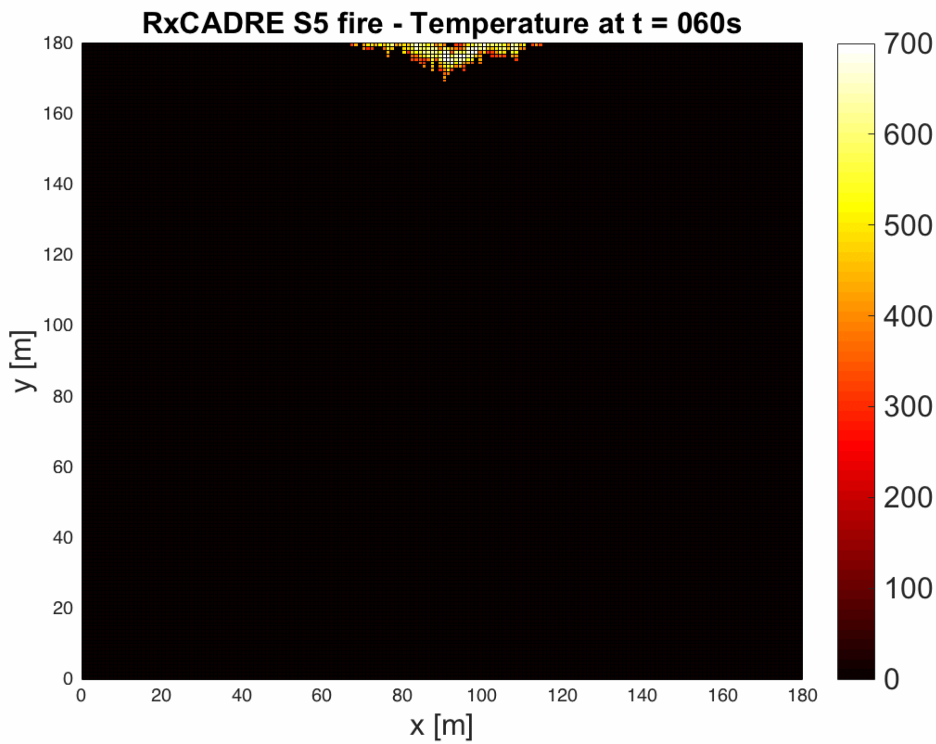
Mid-InfraRed (MIR) imagery

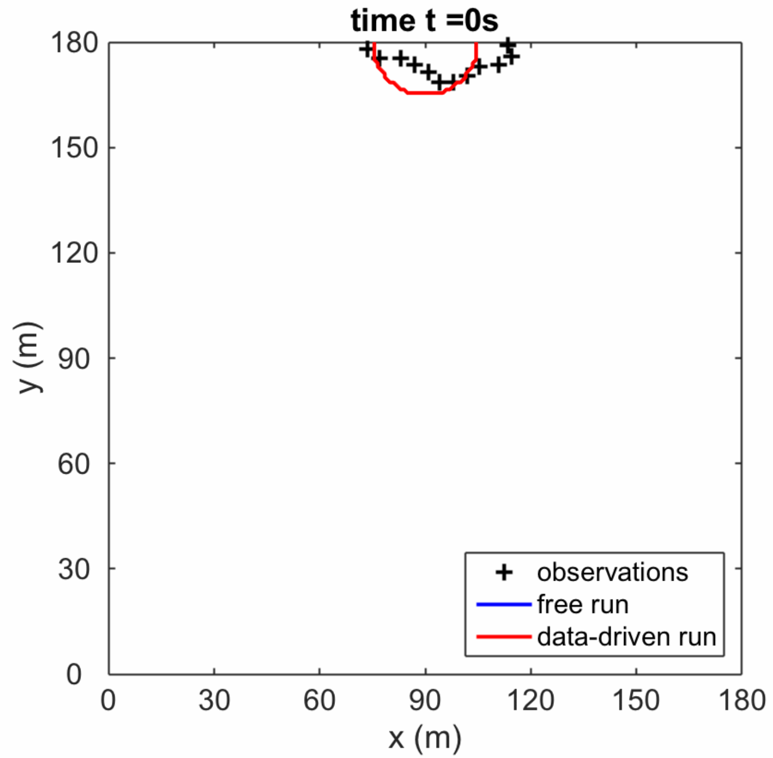


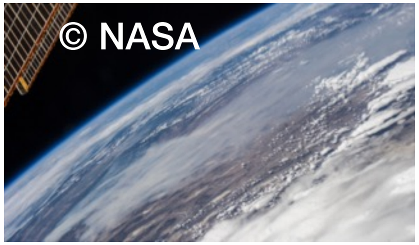




Data assimilation of images

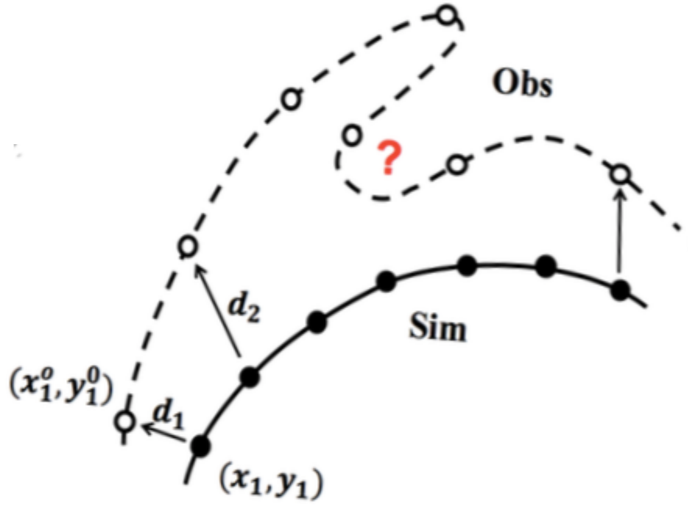


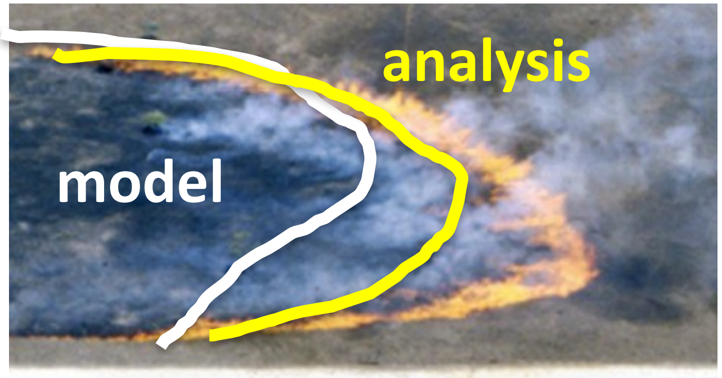




Front shape similarity measure

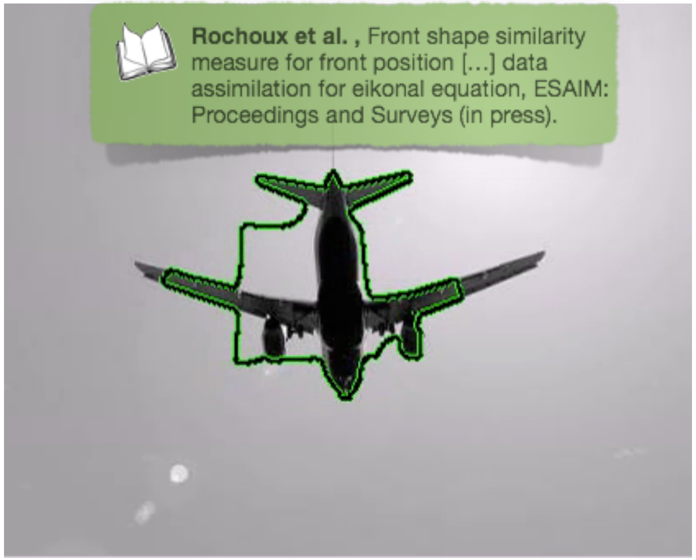
How to properly address shape and position errors for complex fire front topology?





- Non-Euclidean operator to represent shape and topological discrepancies
- Direct assimilation of image data

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}\boxed{\mathcal{D}\left(\mathbf{y}^o, \mathcal{G}(\mathbf{x}^b)\right)}$$



Uncertainty Quantification

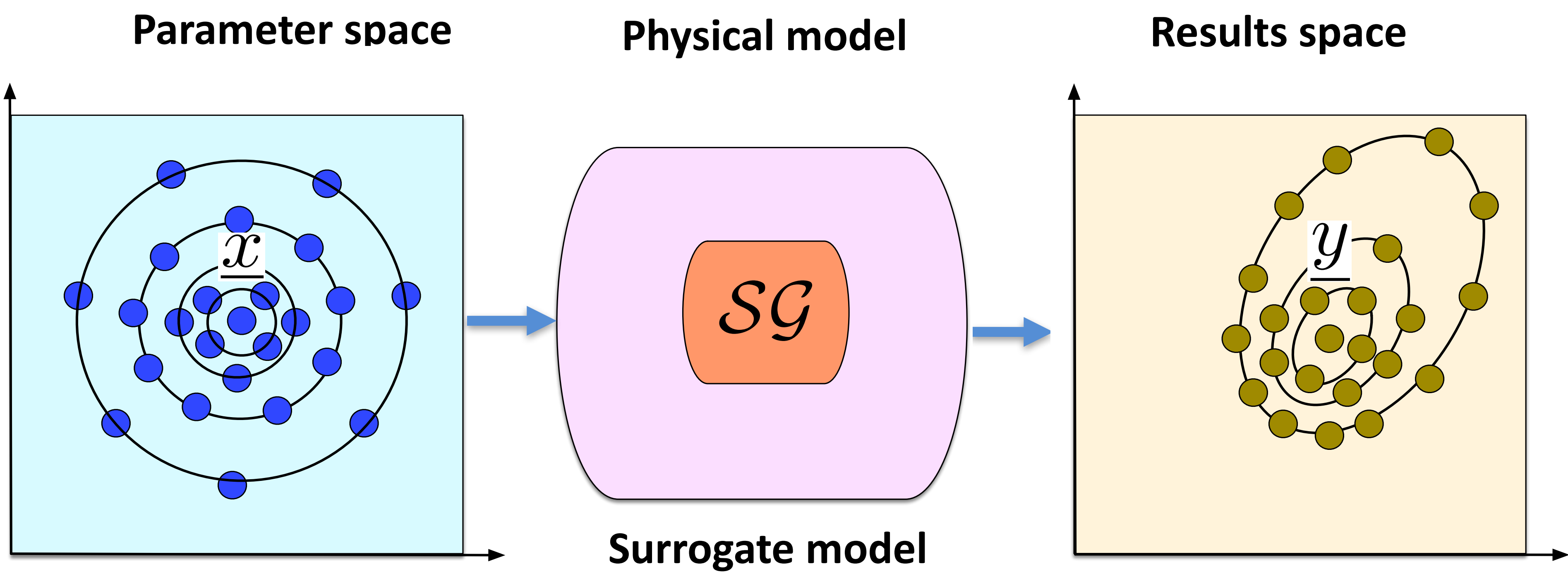
CERFACS has a long-established record of excellence in environmental and industrial Computational Fluid Dynamics for complex flow simulation on high-resolution grid enhanced by continuous developments in numerical models and in High Performance Computing.

Data Assimilation of satellite data for ocean, atmospheric chemistry or hydraulics modeling is also one of its strong expertise domains.

Uncertainty Quantification has become a developing field based on ensemble approaches and model-reduction objectives.

Based on these expertise domains, a new challenge for CERFACS is to develop a Data Driven Modeling axis combining Data Science, Uncertainty Quantification and Data Assimilation.

Uncertainty quantification and reduced models

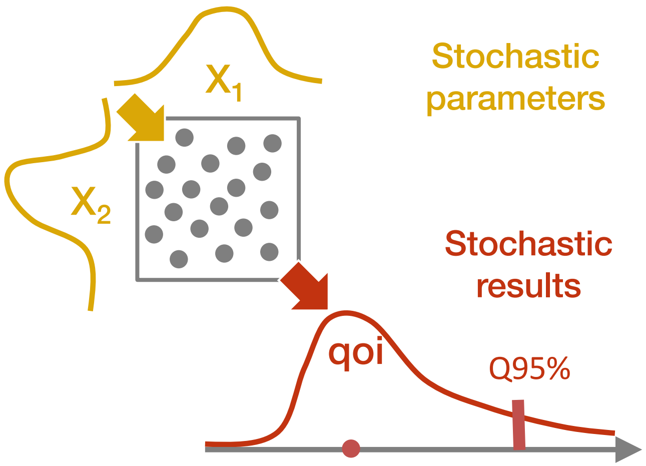


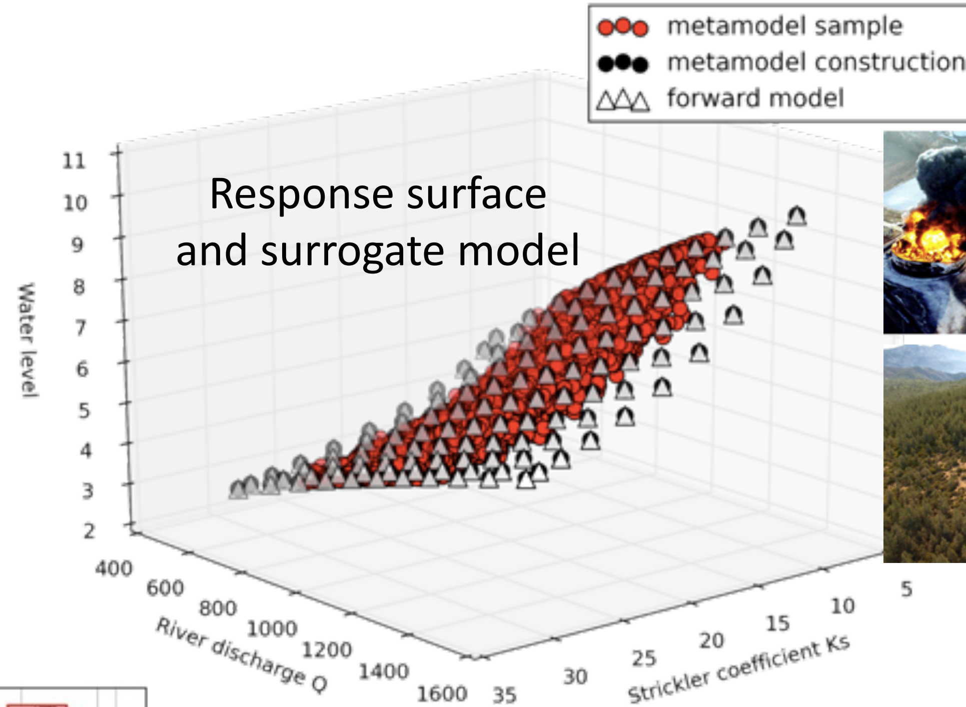
- Model constants
- Geometry
- Initial conditions
- Boundary conditions
- Etc.

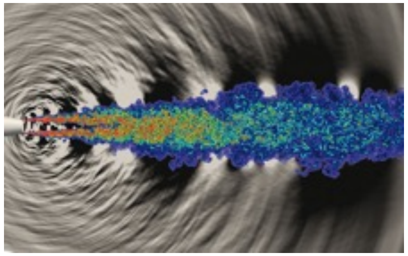
- Equations
- Spatial grid
- Numerical scheme
- Temporal evolution
- Etc.

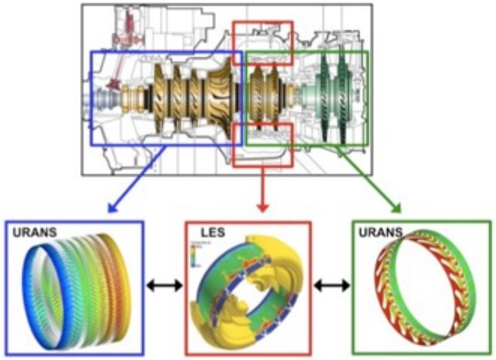
- 0D to 3D fields
- Temporal evolutions
- Integrated quantities
- Extreme events
- Etc.

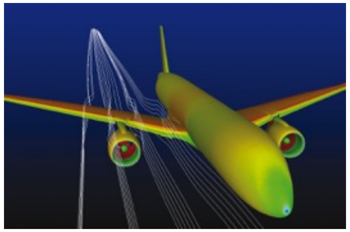
Uncertainty quantification for modelling



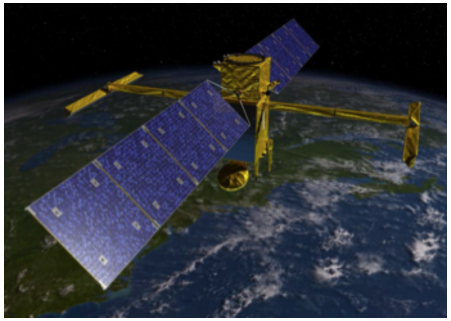






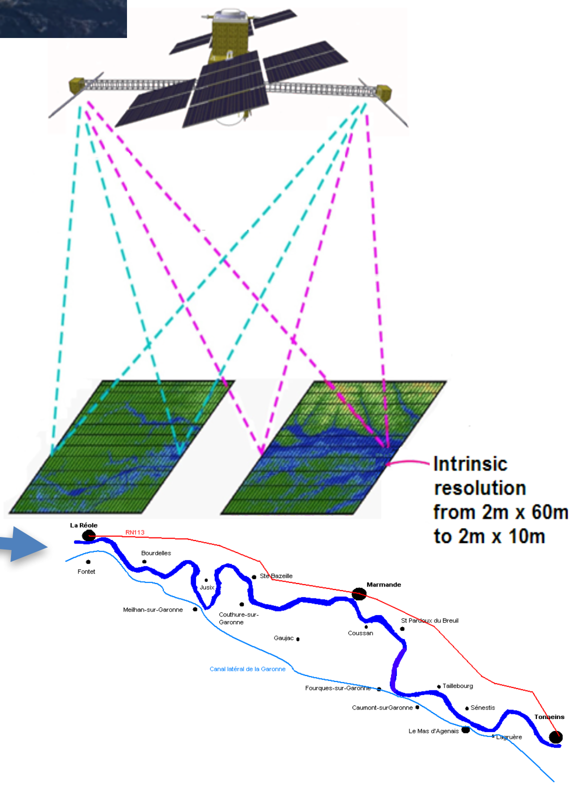


Surface Water Ocean Topography (SWOT) mission

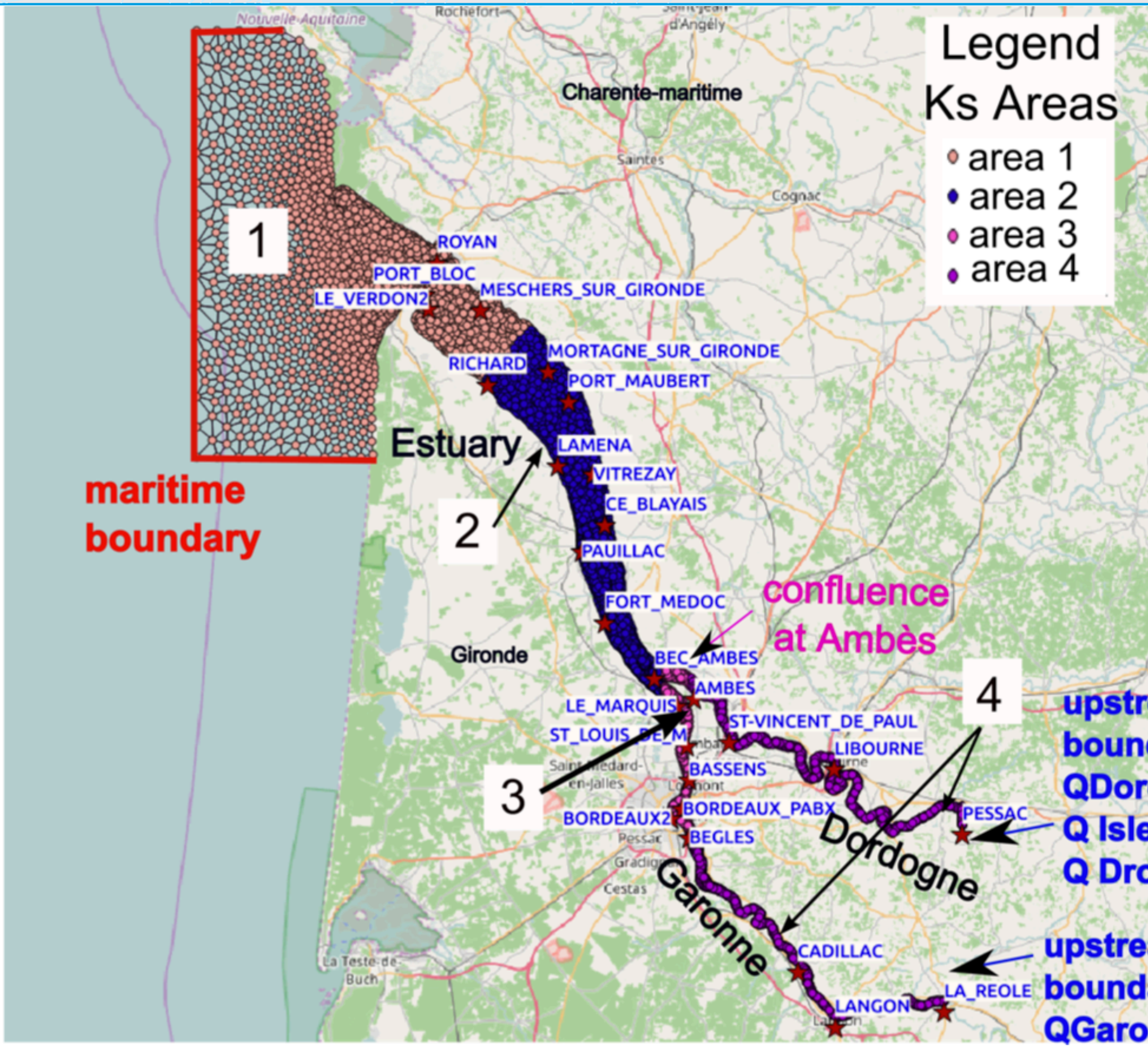


Potential applications

- Transboundary rivers management (international & inter-regional)
- Clear water management for urban, industrial and agricultural needs
- Hydroelectricity production management
- Prevention of the propagation of epidemics
- Fluvial navigation support
- Integrated management for estuaries
- **A better modelling of floods**

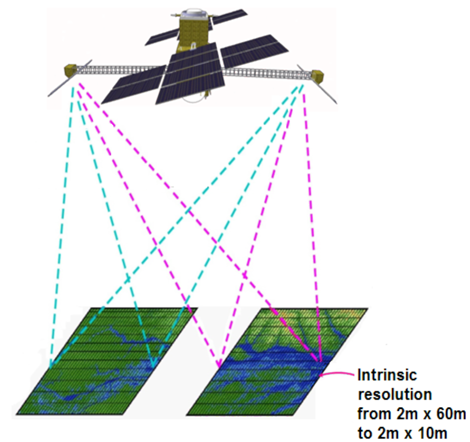


2D simulation of the Gironde estuary



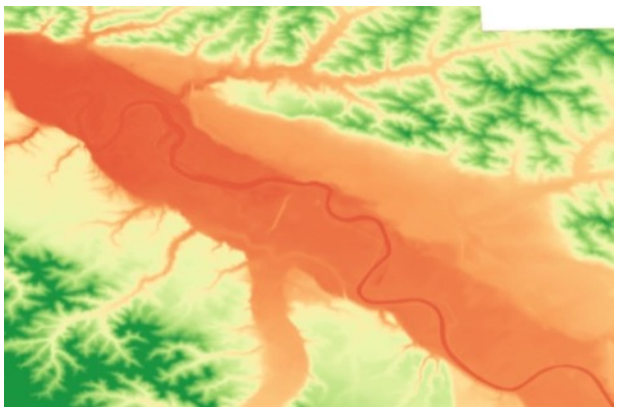
Identify which forcing data parameters are relevant for data assimilation.

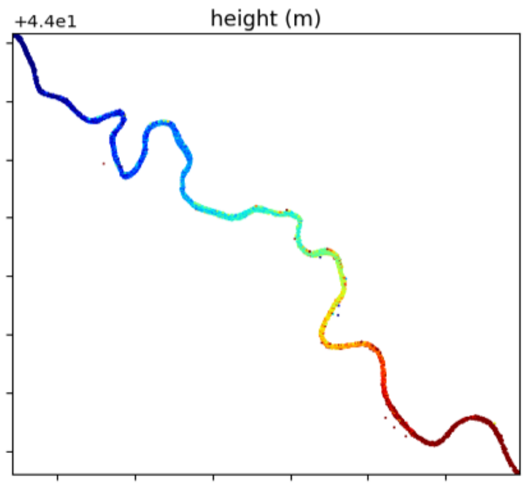
Reduce the uncertainty of the variables that reduce the most the uncertainty of the water level outputs.

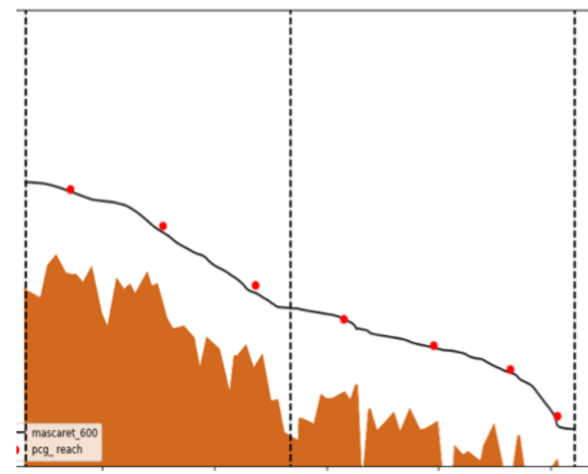


Artificial SWOT data et results production chain

Input of the simulator → Artificial SWOT data → Assessments and uses



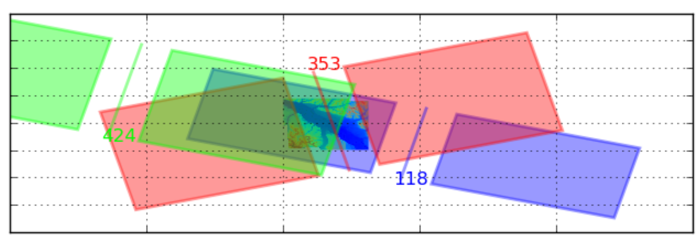




Outputs of models :

- MASCARET (1D)
- TELEMAC 2D

Taking into account satellite orbits



Ensemble Kalman Filter

$$\underline{\delta x}^a = \mathbf{K}\left[\underline{y}^o - \mathcal{G}(\underline{x}^b)\right]$$

Surrogate model in hydraulics – Polynomial Chaos

Context
- Water resources management at EDF
- Flood forecasting at SCHAPI

Sources of uncertainty
- Epistemic errors: friction K_s
- Random errors: upstream forcing Q

Quantity of interest
- Water level
- Discharge flux

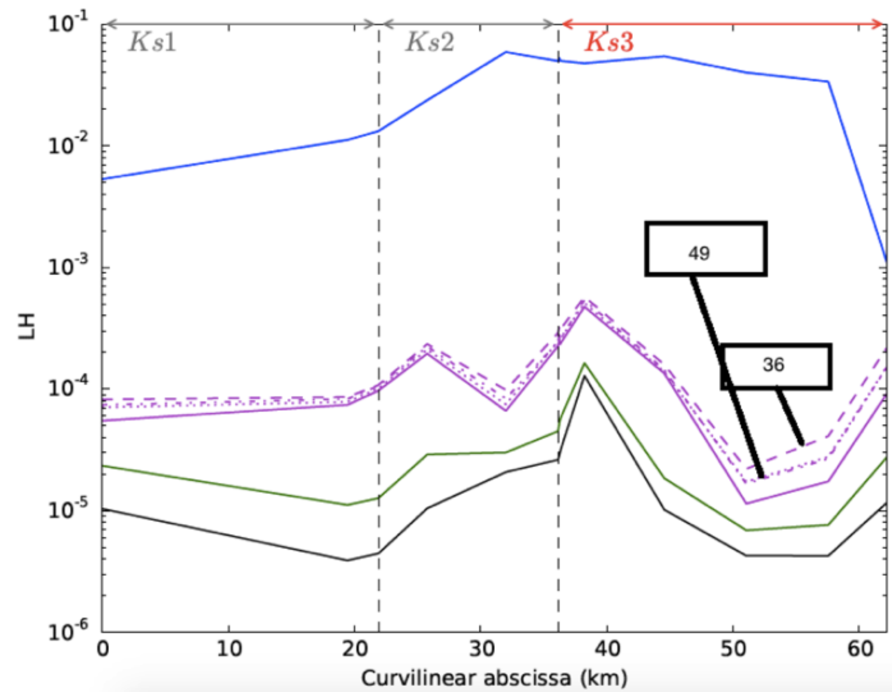
Motivation
- Low cost estimation of statistical moments and pdfs
- Reduced-cost EnKF (Ensemble Kalman Filter)
- Description of water level covariance matrix for EnKF

Non-intrusive PC surrogate model

$$h(K_s, Q; a) = \sum_{i=1}^{N} \widehat{h}_i \, \Psi_i(K_s, Q; a)$$

Water level *h(a)* is expressed as a truncated sum of polynoms that form an orthogonal basis for the probability density functions of the uncertain input random variables (K_s, Q):

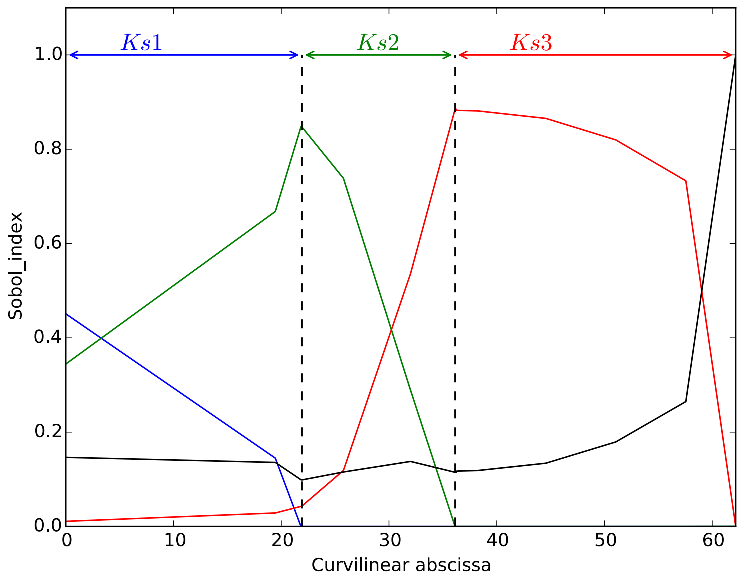
with $N = \dfrac{(n+P)!}{n!\,P!}$ and $n = 2$



Validation of the PC surrogate over the Garonne River. L2-error through the channel for P = 1, 6, 10, 15 compared to 100 000 samples Monte Carlo experiment.

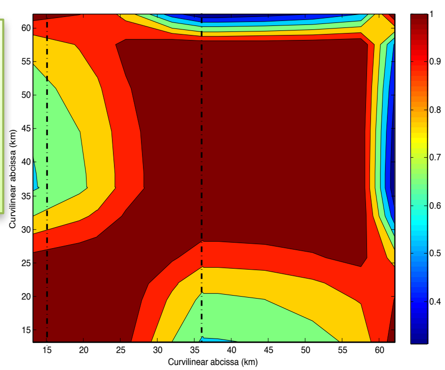
Sensitivity Analysis and Data Assimilation with surrogate models
The Garonne river

Sobol indices (sensitivity)



Background covariance error matrix

$$\underline{\underline{B}}$$



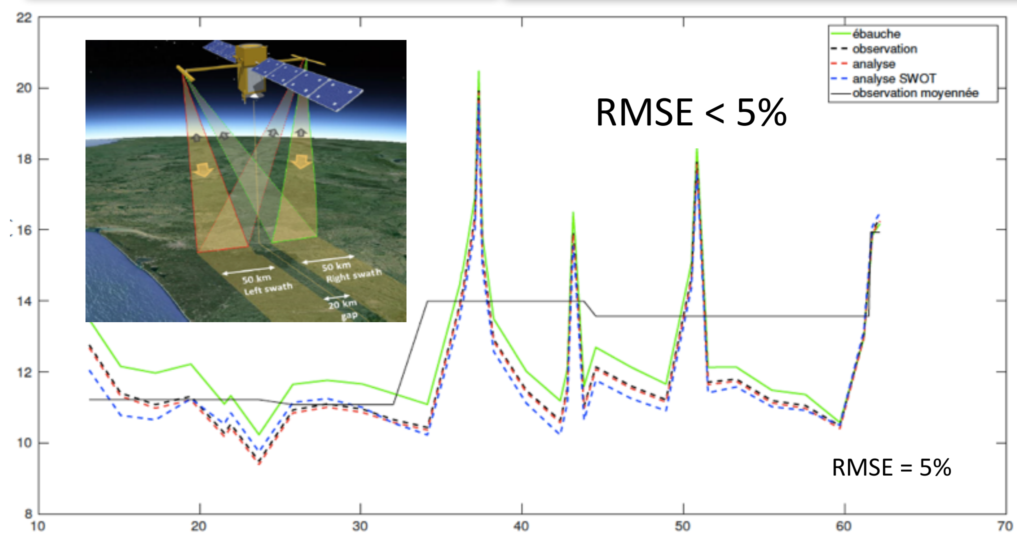
Ensemble Kalman Filter with altimetric observations (obs)

$$\underline{\delta x}^a = \mathbf{K}\left[\underline{y}^o - \mathcal{G}(\underline{x}^b)\right]$$

P-7 PC-EnKF, 2401 samples

EnKF, 2500 samples

One obs per grid point | **SWOT-like obs, 10 km averaged**



Data Science

CERFACS has a long-established record of excellence in environmental and industrial Computational Fluid Dynamics for complex flow simulation on high-resolution grid enhanced by continuous developments in numerical models and in High Performance Computing.

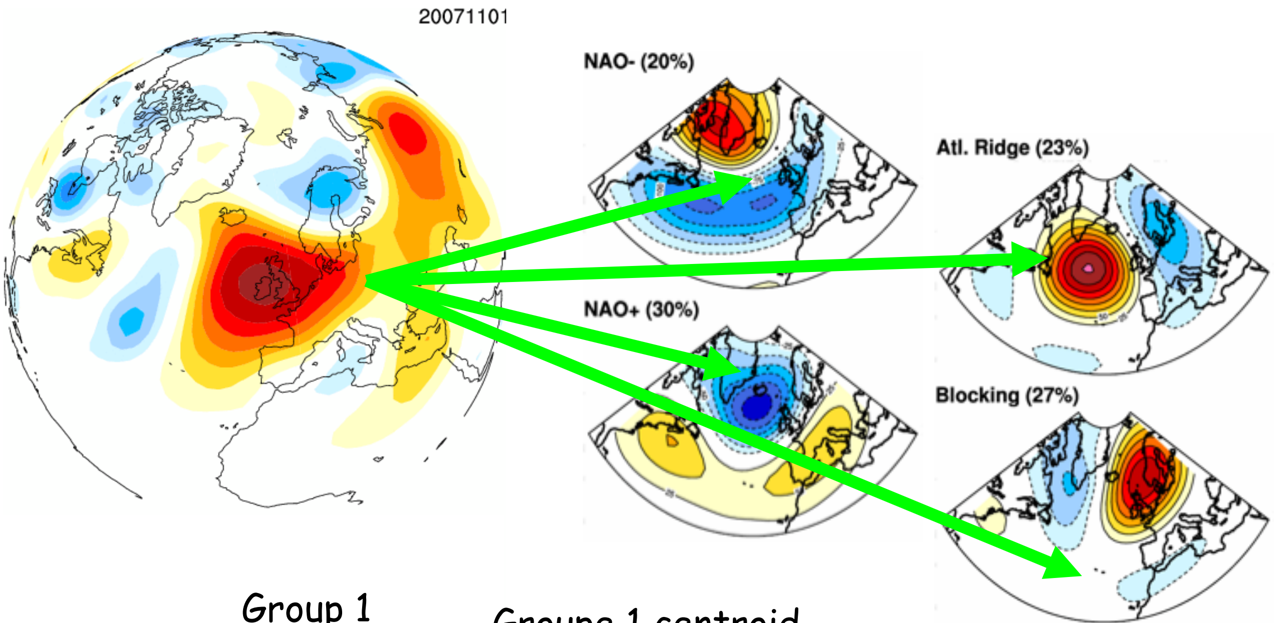
Data Assimilation of satellite data for ocean, atmospheric chemistry or hydraulics modeling is also one of its strong expertise domains.

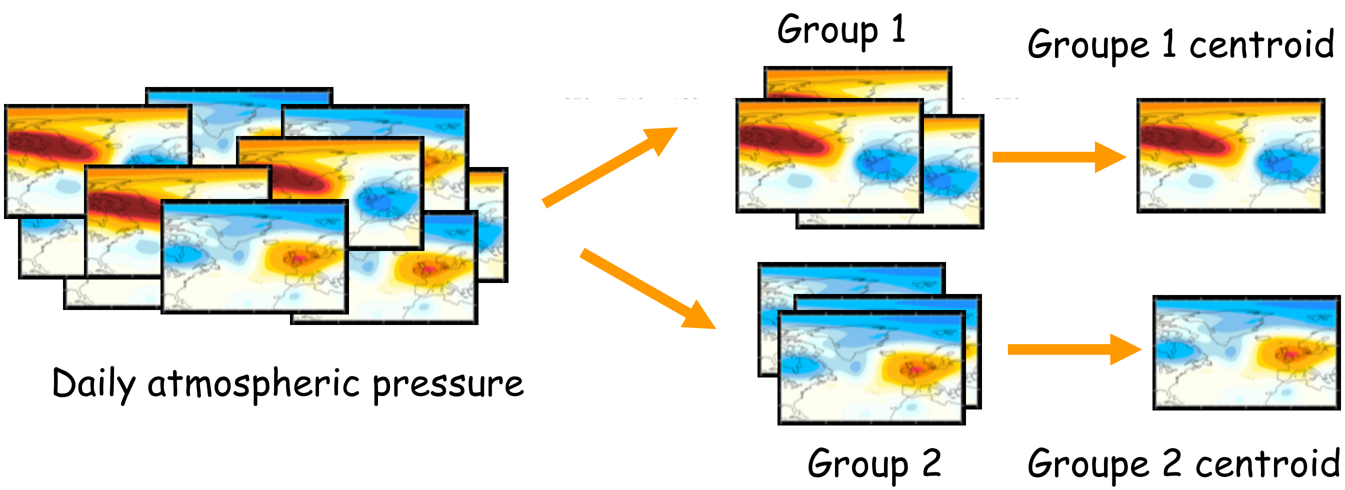
Uncertainty Quantification has become a developing field based on ensemble approaches and model-reduction objectives.

Based on these expertise domains, a new challenge for CERFACS is to develop a Data Driven Modeling axis combining **Data Science**, Uncertainty Quantification and Data Assimilation.

Practical example of data mining: classification and downscaling

Linear classication of weather types (EOF)

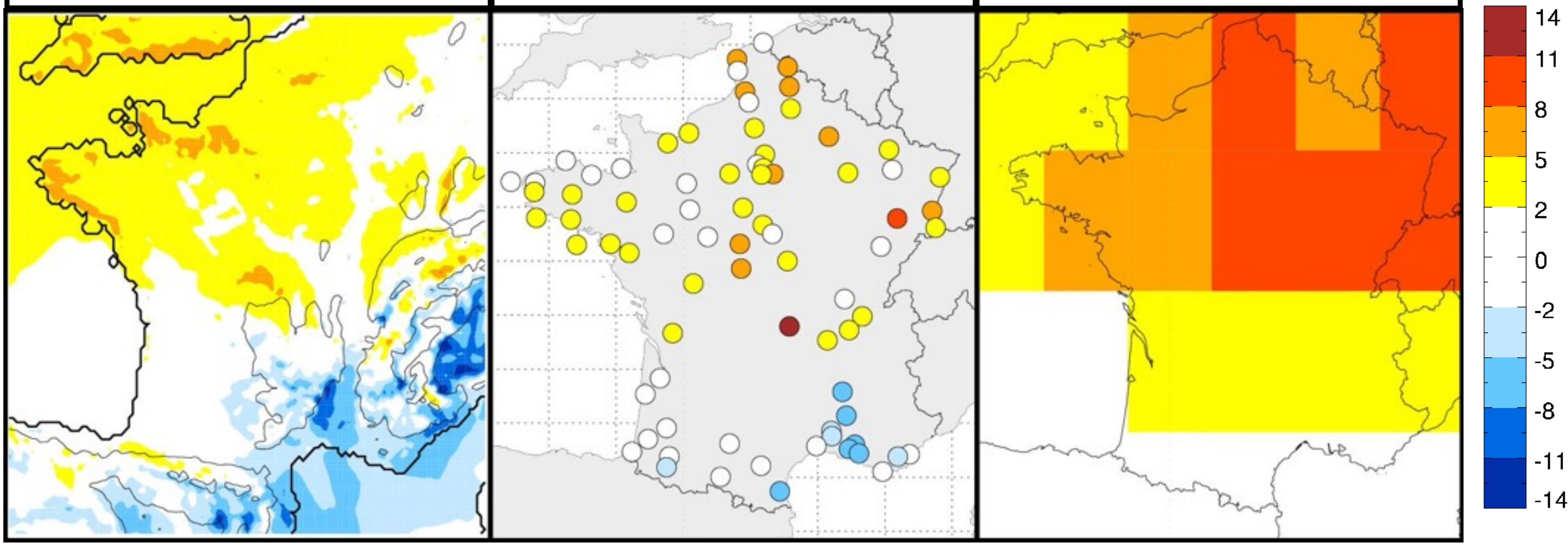




Nonlinear classification methods

Statistical and hybrid downscaling for wind turbine potential

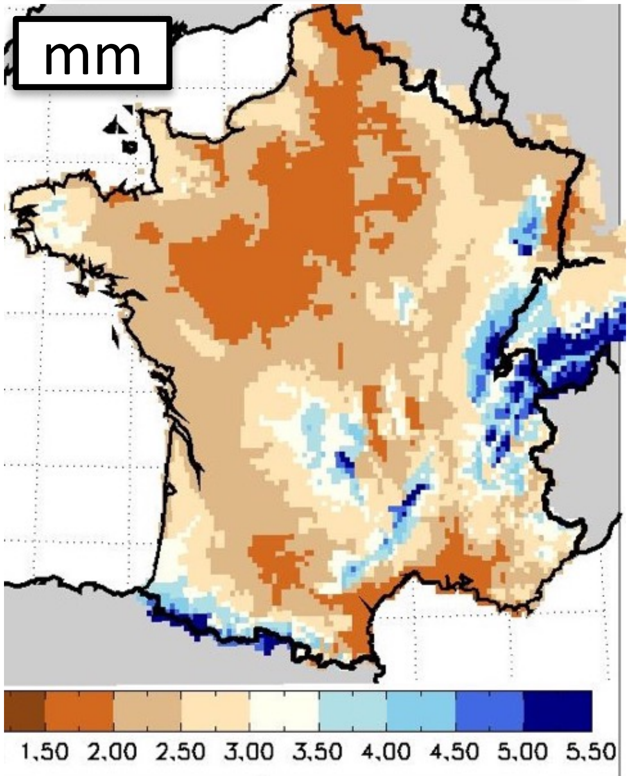
Hybrid downscaling with mesoscale simulations	Statistical downscaling with weather classification	Large scale climate simulations (GIEC)



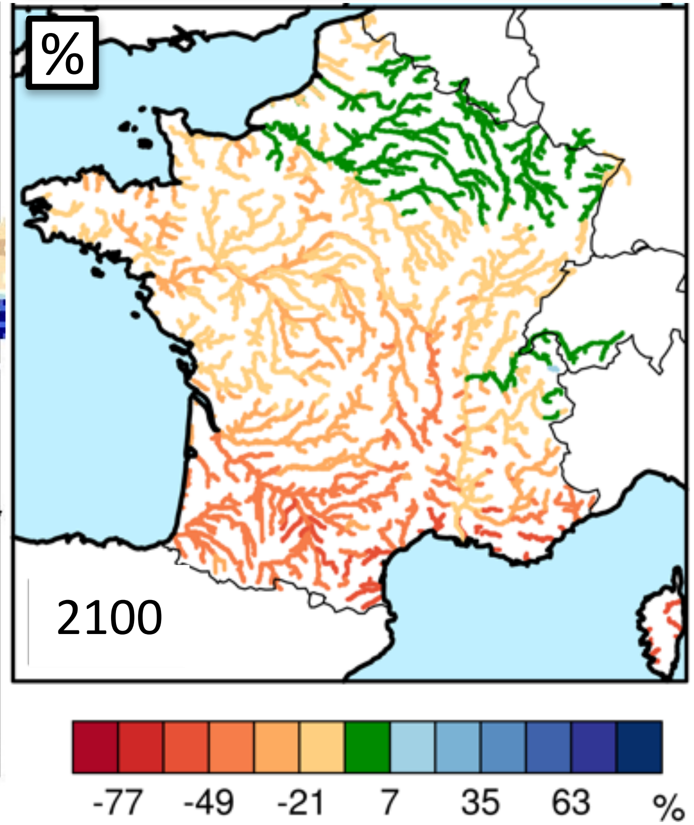
Trend of the winter wind turbine potential for 2050

Statistical and hybrid downscaling for river discharge fluxes

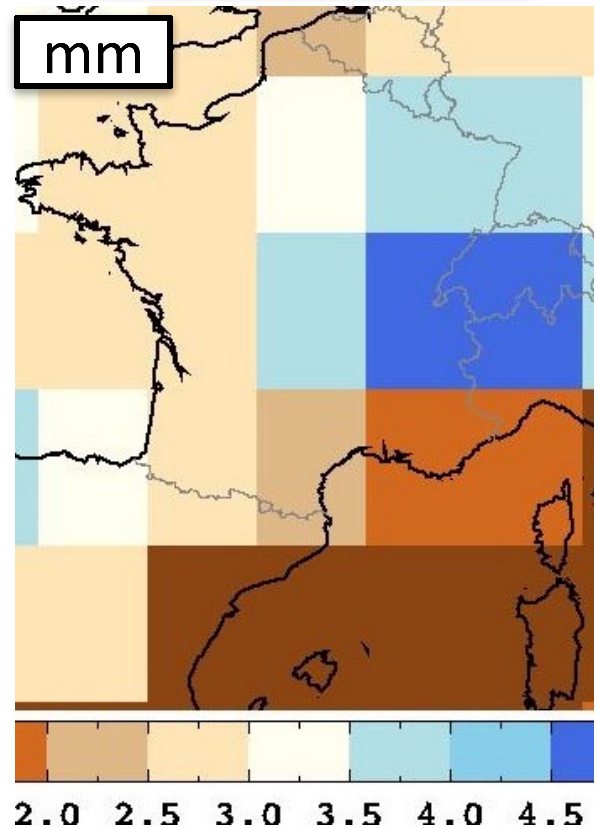
Downscaled rainfall



River discharge 2100 trend



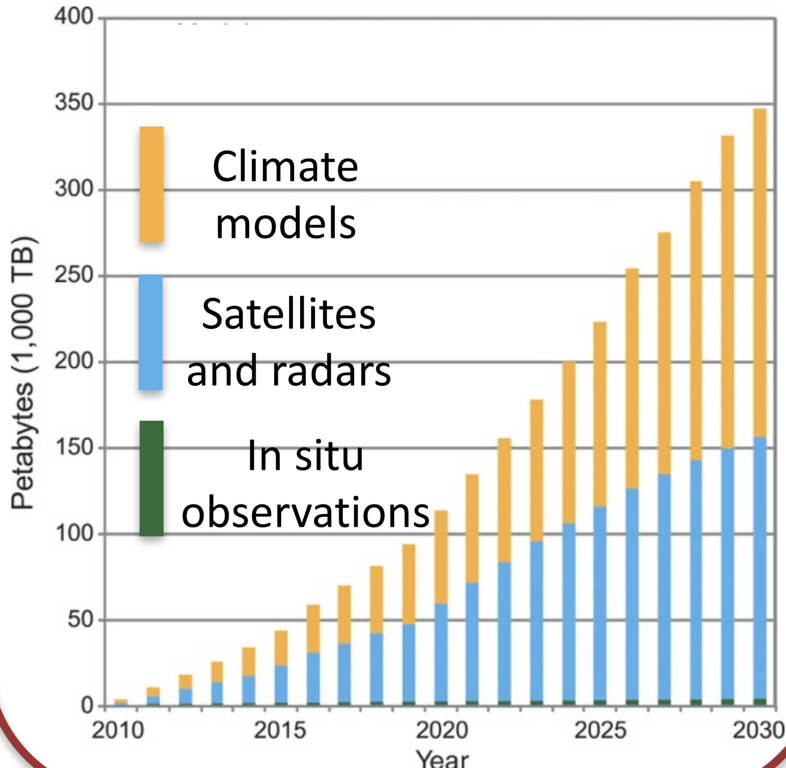
Large scale rainfall

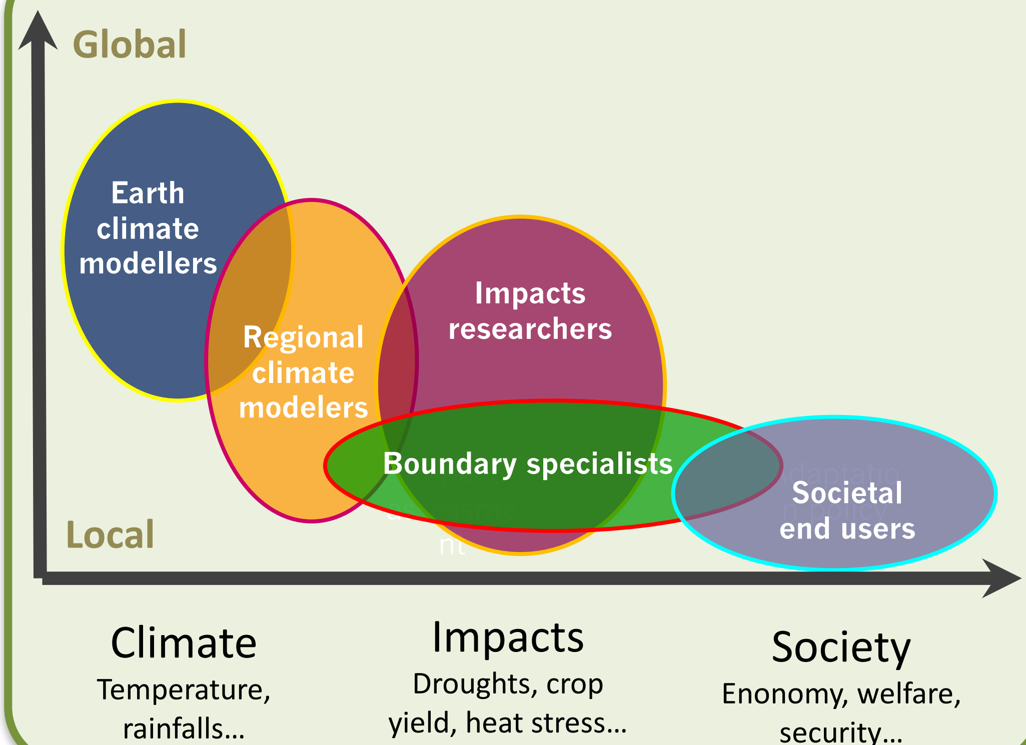


Evolution of river discharge fluxes for 2100

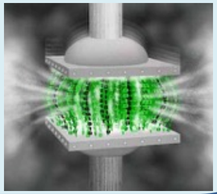
« Big Data » and climate

Climate data increase
(1 Petabyte = 10^3 To)

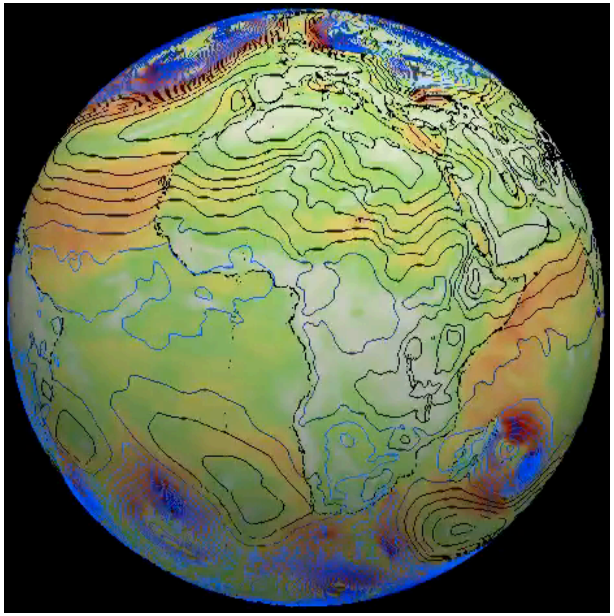


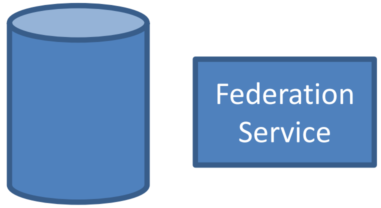


Data and transfer reduction



Practical examples of climate study

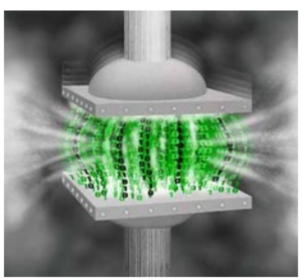




TOTAL to download: 600 Gb

Data required to the study:
- Temperature at 850 hPa field
- 10 climate models
- 60 years = 21 915 days
- Daily fields = 1 field per day
- Global scale 100 km resolution

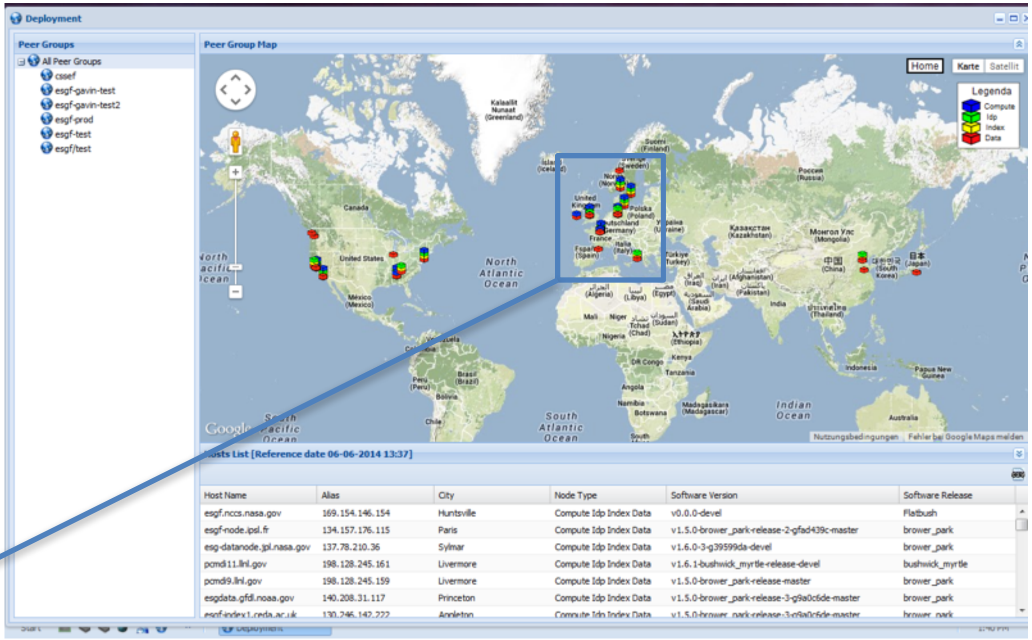
Data produced through the post-processing:
- Anomaly of the average of two periods
- Over a specific country for each climate model
- 10 times 2D fields over a small domain



Need of data reduction before the download

Estimated datasize after post-processing: **1 Mb**

Climate Data Distribution: ESGF



ESGF Data Nodes 2015:

- 40 worldwide
- 18 in Europe (coordinated in IS-ENES)

IS-ENES ESGF Portals

- BADC (UK)
- DKRZ (Germany)
- IPSL (France)
- SMHI (Sweden)
- CMCC (Italy)
- DMI (Denmark)

IS-ENES climate4impact Portal

- KNMI (Netherlands)
- Interlinked with Uni. Cantabria downscaling portal (Spain)

CLIPC Portal

- Climate Information Portal for Copernicus

Ack: Michael Lautenschlager, DKRZ

Data distribution platforms
Scientific, Technical and Societal motivations

Scientific

- **Efficient Data Analysis:** ensemble of scenarios, uncertainties range estimation, higher resolution, easy share of results…
- **Robust and flexible Data Life Cycle:** more robust experiments setup, several configurations, reproducible experiments

Technical

- **Process large data volumes, near the data storage**: Data Analytics, Data Life Cycle, streamline the data processing workflow..
- **Interconnect e-infrastructures and research infrastructures:** metadata description of the data, track provenance…

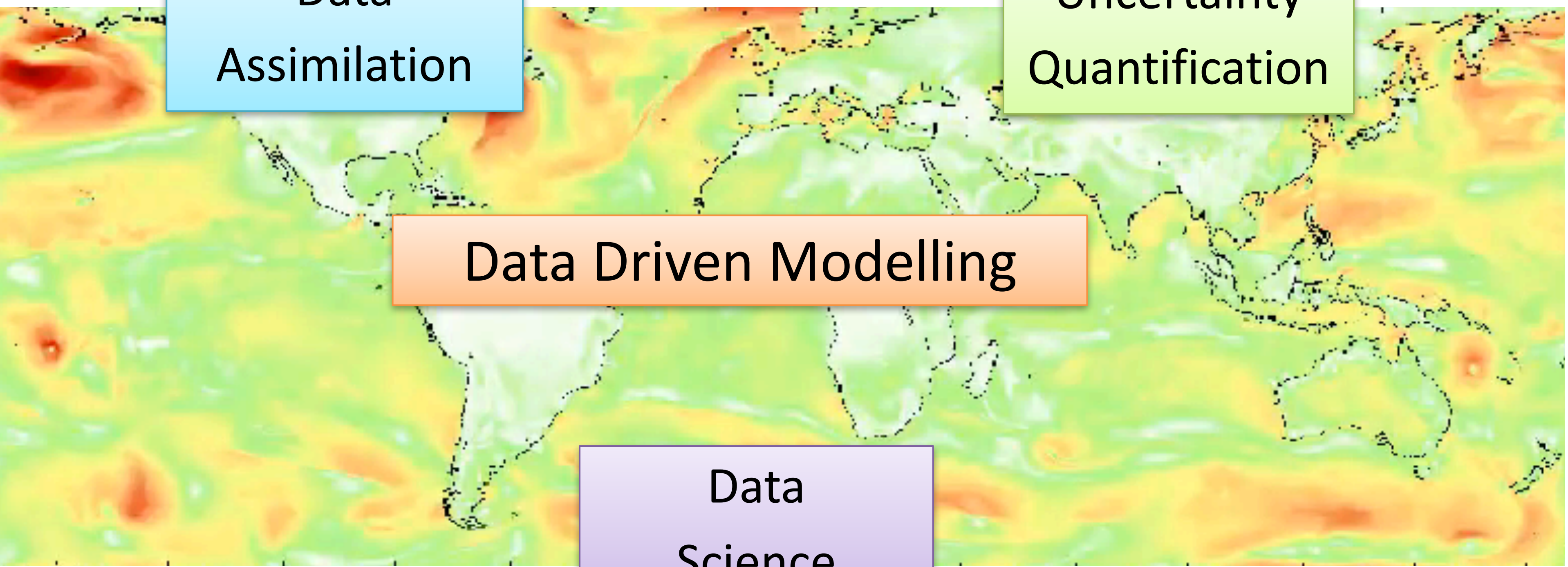
Societal

- **Provide climate projections data:** impact researchers, facilitators, practitioners…
- **Ease access with better intuitive interfaces:** tailored products from data processing workflows…

Prospects of big data for climate

- Infrastructure to access relevant climate data (climate models, satellite observations…)
- Community Services with standard interfaces (on-demand services and calculations…)
- Bridge e-infrastructures and research (ease data sharing, provide support…)
- Big Data Techniques (data mining for geophysical data, neural networks…)

Conclusion : Data Driven Modelling at CERFACS



Data Assimilation

Uncertainty Quantification

Data Driven Modelling

Data Science

To combine data and physical models
in a framework of high performance computing