Large-scale data analysis in solid Earth geophysics

Alexandre Fournier & Nikolai Shapiro

Institut de Physique du Globe de Paris, Paris, France

Big data @ USPC, November 30th 2015

Inverse problem in solid-Earth geophysics

Principle

Combine observations & prior information (theory) to better understand and predict the behaviour of a natural system

Data

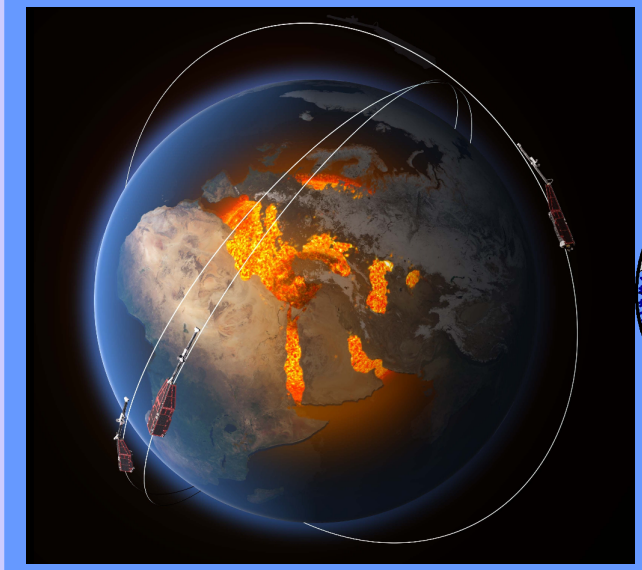
- observations
- simulations

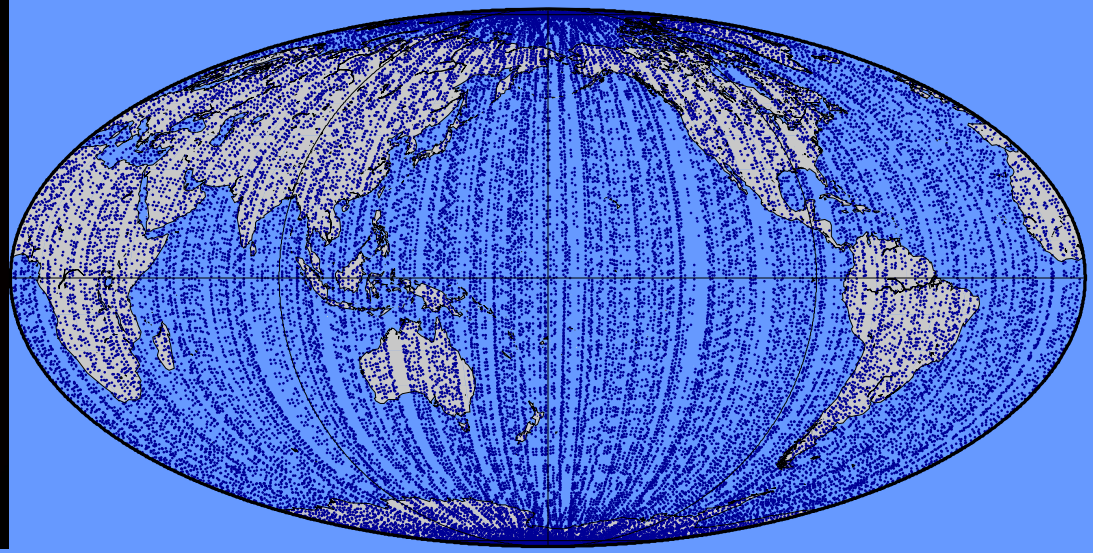
Example 1: Predicting the future geomagnetic field

Example 1: Predicting the future geomagnetic field

Observations

- Measurements of the Earth's magnetic field.
- ESA Swarm mission (launched Nov 22nd, 2013).
  `http://www.ipgp.fr/fr/mission-swarm`





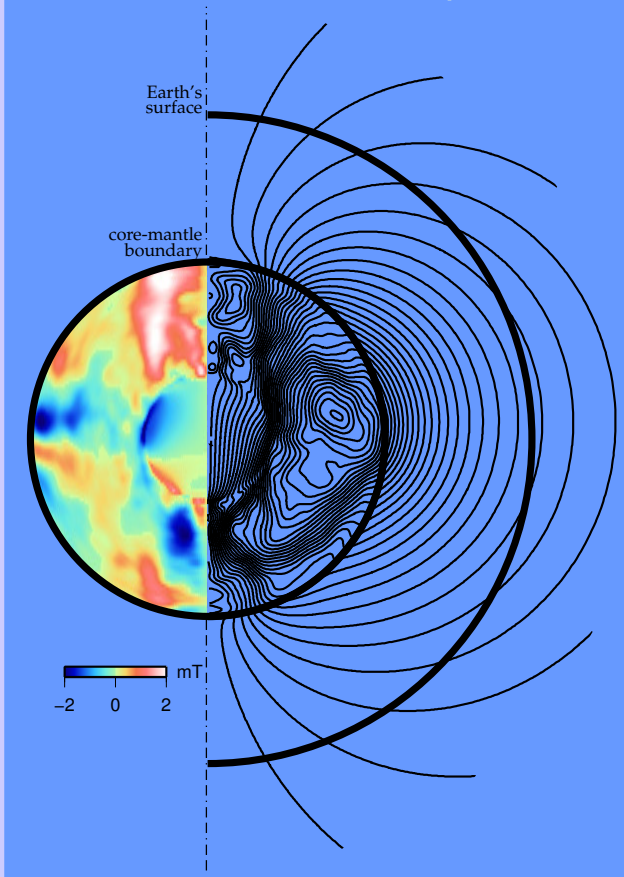
Fournier et al. (Earth, Planets, Space 2015)

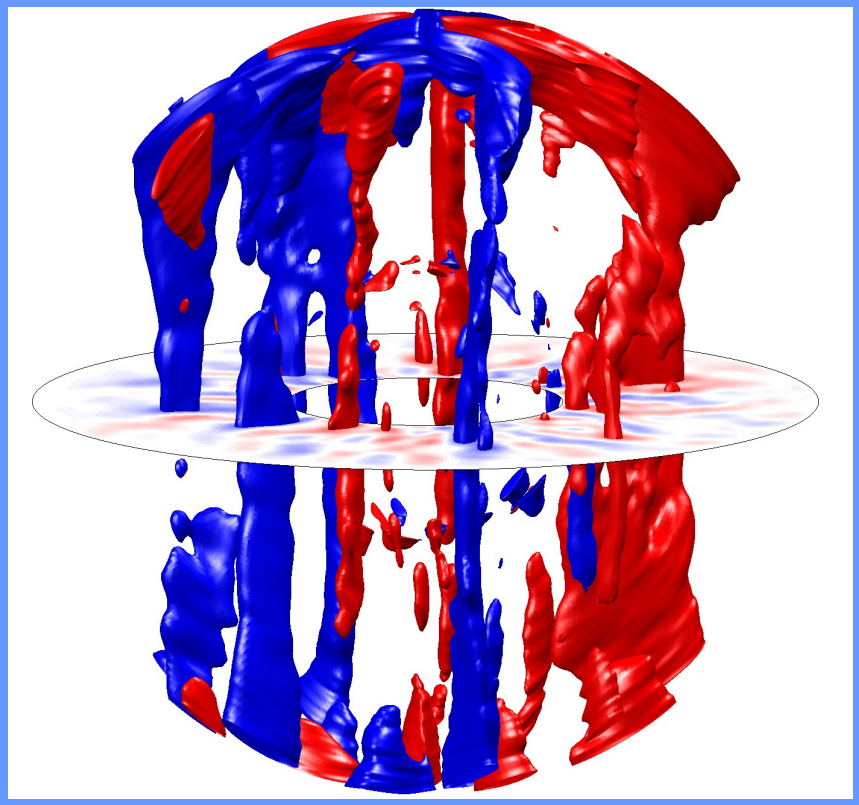
Size of data vector: moderate (few 10^6 data points)

Example 1: Predicting the future geomagnetic field

Theory

Dynamo theory (Navier–Stokes & MHD): numerical model.





Fournier et al. (Earth, Planets, Space 2015)

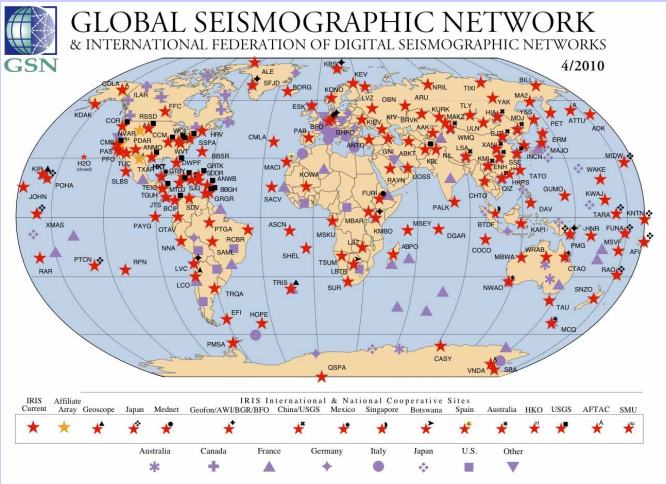
Inverse problem is solved to define initial condition (field, flow, buoyancy) for subsequent integration. Synthetic data (prior statistics): $\sim$1 **Tb**.

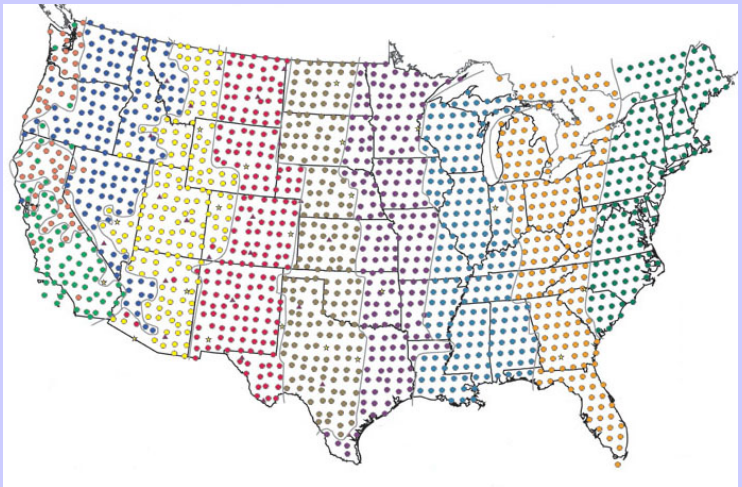
Example 2: Seismic imaging and monitoring

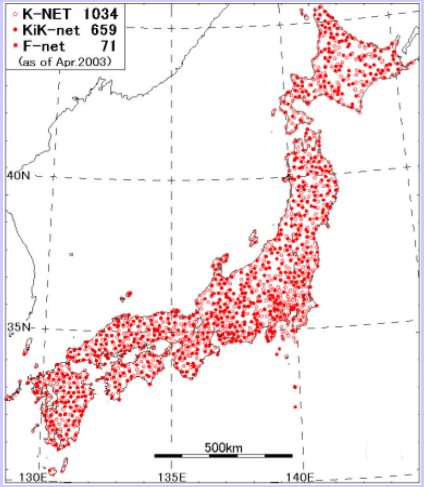
Example 2: Seismic imaging and monitoring

Observations

Broadband 3-component continuous seismograms.





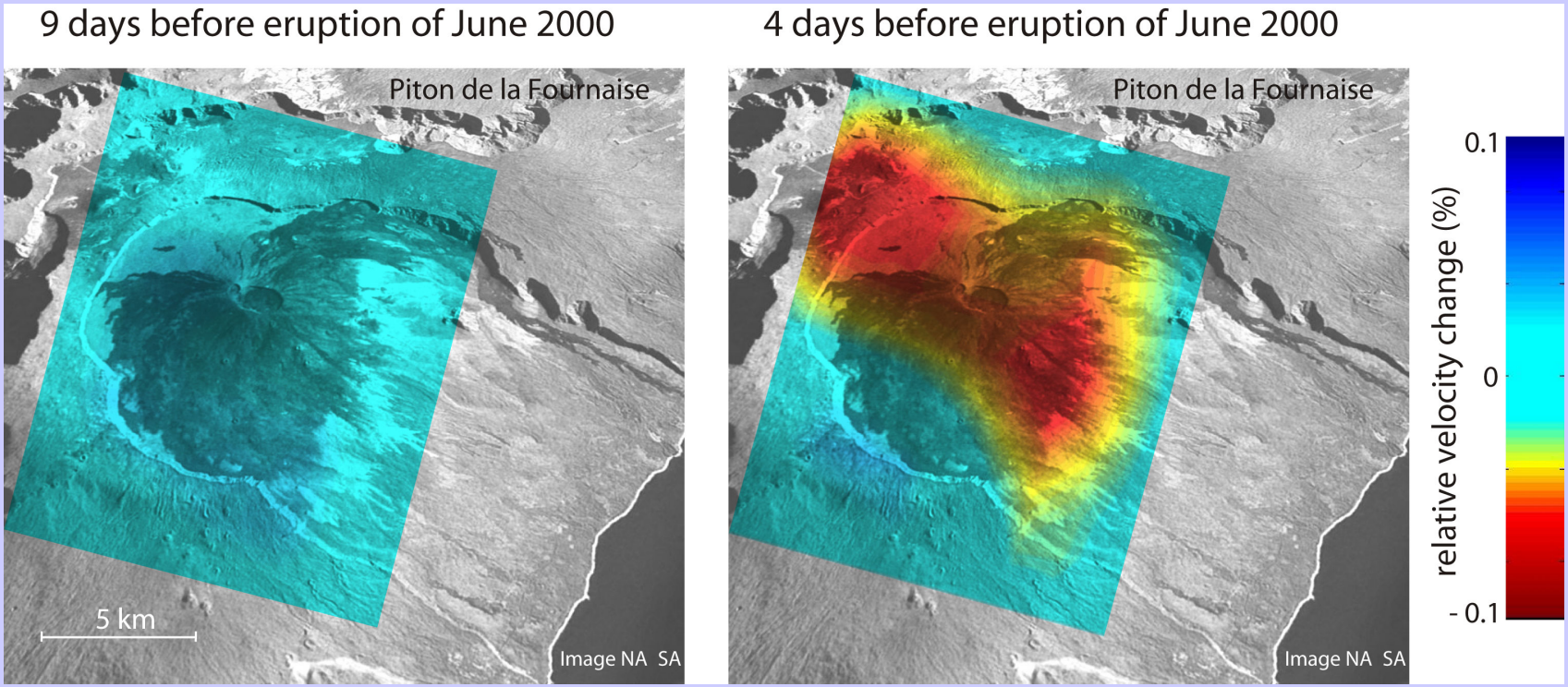


Size of data vector: 10 to 100 Tb.

Example 2: Seismic imaging and monitoring

Theory

- Ambient noise tomography
- Cross-correlations between receiver pairs.
- N seismograms: $N(N-1)/2$ cc (FFT).



Brenguier et al., Nature Geoscience, 2008

Ressources

National (free-run simulations, GENCI) and local ressources well-suited for intensive analysis of large datasets: S-CAPAD
`http://www.ipgp.fr/rech/scp`

IPGP
INSTITUT DE PHYSIQUE DU GLOBE DE PARIS

SORBONNE PARIS CITÉ
RECHERCHE
ENSEIGNEMENT SUPÉRIEUR

Hardware

- **100 nœuds CPU** : PowerEdge C6220*x* (4 nœuds par serveur)
  . 96 C6220 : 2 Intel Xeon E5-2690 2.90 GHz
  . 4 C6220 II : 2 Intel Xeon E5-2650 v2 2.6 GHz
  et pour chaque nœud, 2×8 cœurs, 64 GB RAM, HDD : 2×2 TB
- **28 nœuds "données"** : PowerEdge RX7*x*0d
  . 16 R720xd : 2 Intel Xeon E5-2650 2.00 GHz
  . **12 R730xd : 2 Intel Xeon E5-2630 v3 2.4 GHz**
  et pour chaque nœud, 2×8 cœurs, 128 GB RAM, HDD : 20×600 GB, SSD : 4×200 GB
- **4 nœuds GPGPU** : PowerEdge R720, et pour chaque nœud,
  2×8 cœurs, 2 Xeon E5-2650 2.00 GHz, 64 GB RAM, HDD : 2×2 TB
  2 NVIDIA Tesla K20 PCI-E (2496 cœurs)

soit 2176 cœurs (hors NVIDIA) et plus de 10 To de mémoire, et

- **du stockage parallèle**
  . nœuds GPFS : 2 PowerEdge R720
  . stockage : 4 PowerVault MD3260 (576 To) + **4 PowerVault MD3060e (123 To)**
  $\Rightarrow$ 699 To utiles
- **un réseau à haut débit et faible latence**
  14 switchs Intel True Scale 12200-BS01 36-port Infiniband QDR

Puissance théorique crête totale $\simeq$ 47 Tflops (hors NVIDIA).

(financé par les crédits USPC)

Enjeux

- Higher resolution (geodynamo: better representation of physical processes)
- Larger datasets (seismology: longer timeseries imply better images)
- Emerging fields, eg surface tectonics (deformation w/ interferometry, . . .)
- Ressources