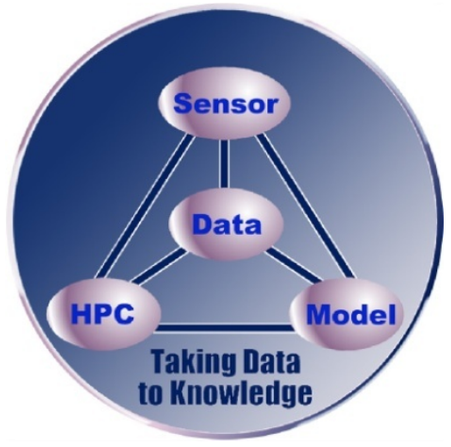
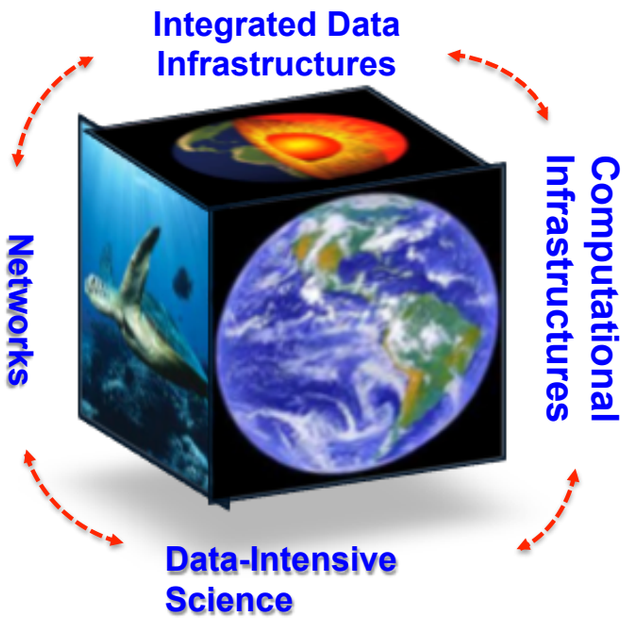
Depuis 80 ans, nos connaissances bâtissent de nouveaux mondes



Multi-source Data Analysis and modelling Challenges in Earth Systems and Universe Sciences

Jean-Pierre Vilotte

Institut des Sciences de l'Univers (INSU) - CNRS (France)



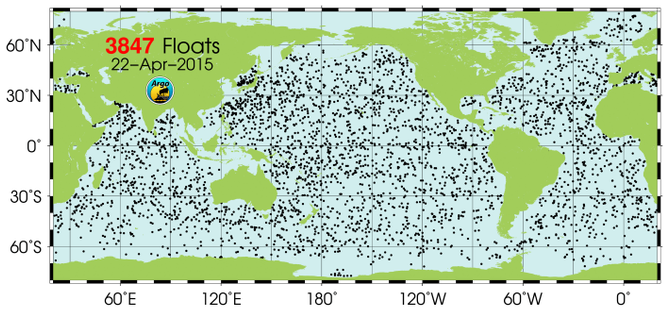
BDEC meeting, July 10, 2020

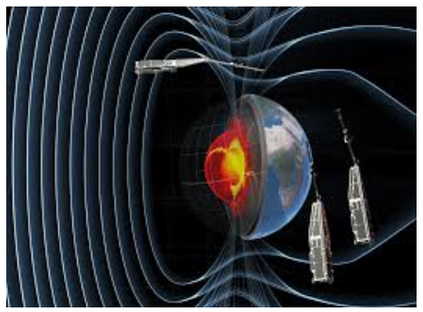
Data flux explosion and diversity

Ubiquity and explosion of data

NenuFAR/SKA

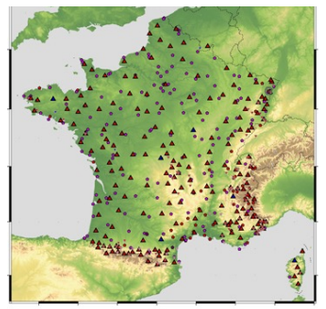
ARGO





Swarm mission

SVOM

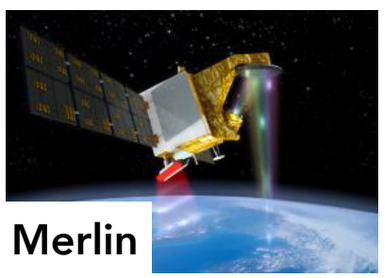


Seismic/geodesy

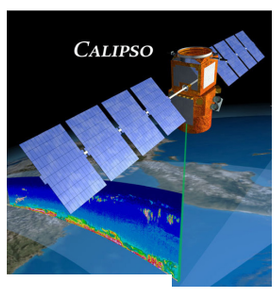
CFHT

Hayabusa2-Mascot module

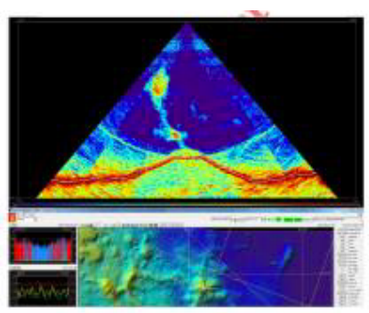
Copernicus



Merlin

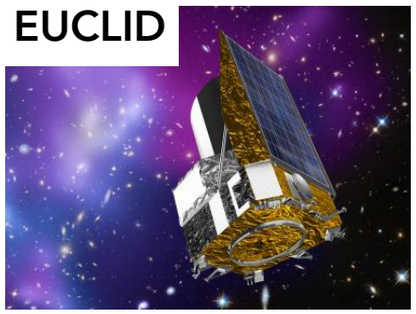


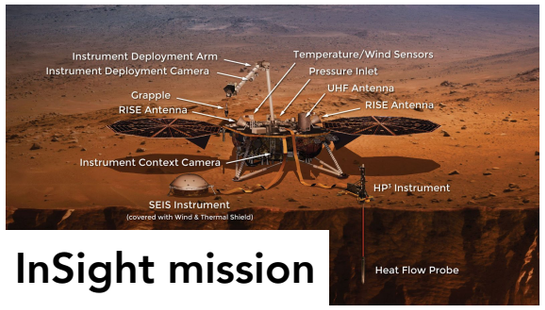
Calipso



Active volcanos

EUCLID





InSight mission

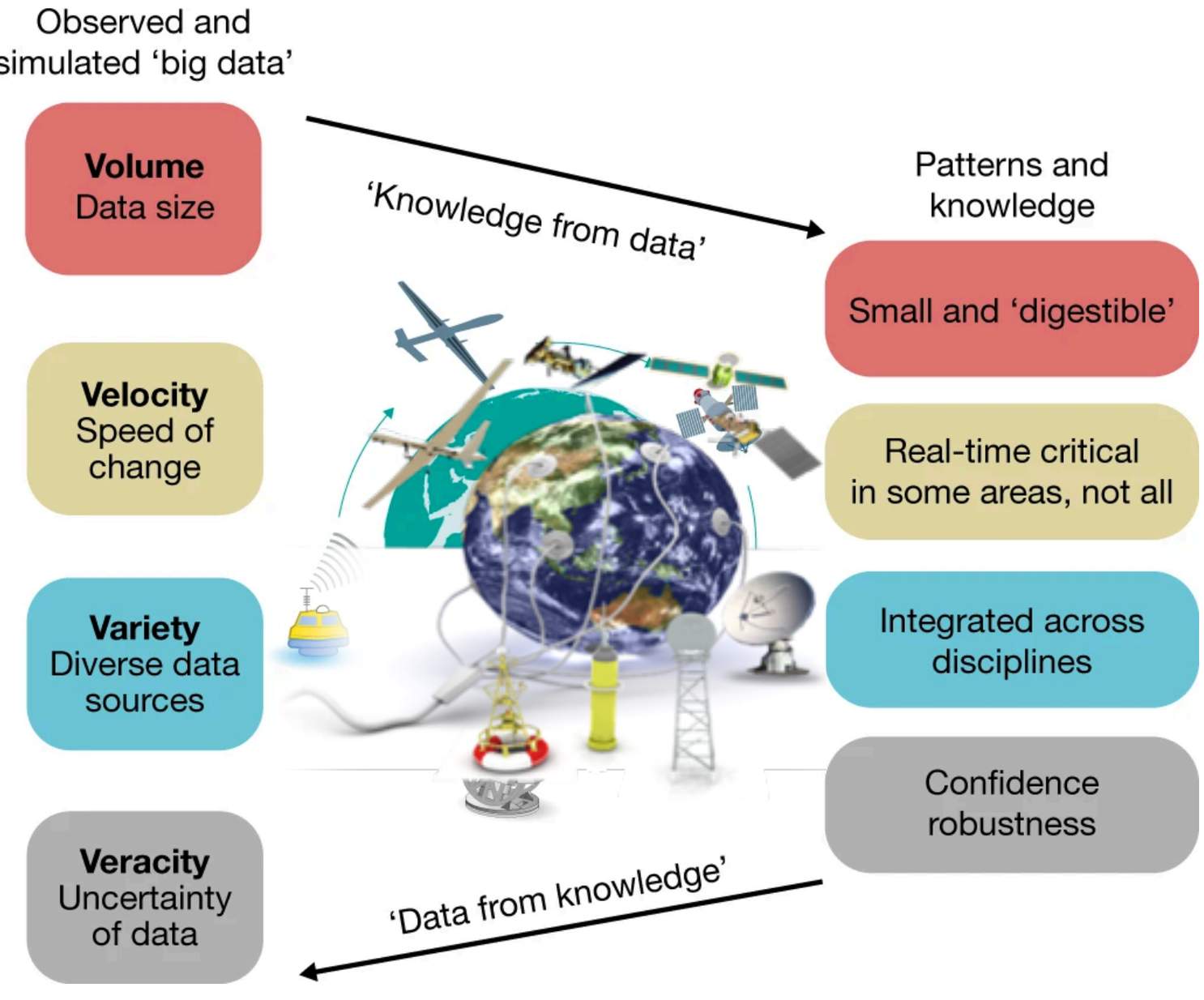
Data explosion (rate, volume, diversity)

- **Edge environments**: in-situ (land, sea), air and space observation
- **Centralised environments (**Cloud and HPC): large ensemble simulations

Challenges:

- **Data logistics**: *data stream (processing/reduction/compression/transfer)*
- **HDA**: *multi-source data statistical analysis, ML*
- **HPC**: *ensemble of multi-physics/multi-scale simulations, statistical data assimilation, ML*
- **Data Management**: *long-term archiving & curation*

Big Data Challenges



BigData Challenges

- Flux rate, volume, diversity
- Multi-source, multi wavelength
- Reprocessing and versioning
- Large ensemble simulations

Data Policy and management

- Open Data by default, FAIR data services
- Long-term archiving and curation
- Data veracity, certified repositories
- Software management and certification

Statistical challenges

- Multi-temporal, multi-angular, multi-source
- Non-linear and non-stationary (non Gaussian)
- Data and systemic uncertainties,
- Extreme events

Machine learning challenges

- Few supervised information available
- Computationally intensive and timeliness
- Consistency, learning and interpretability
- Multi source uncertainty propagation

Data-intensive astronomy/astrophysics

Data Intensive Astronomy

Era of big surveys: LOFAR, LSST, CTA, SKA

Exponential Growth of Data Volumes

…and Complexity

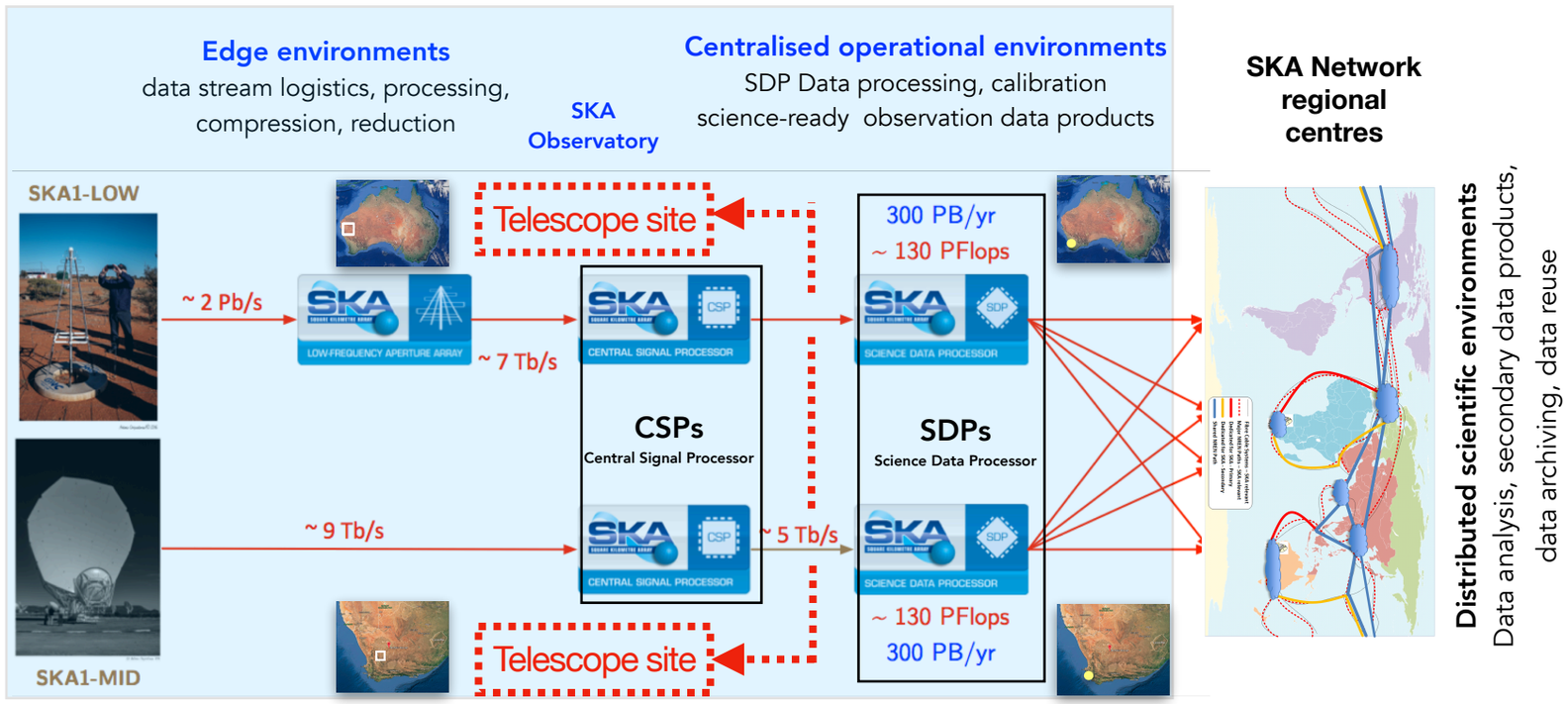
User interaction with the data has become the bottleneck in research!

- **From data poverty to data glut**
- **From data sets to data streams**
- **From static to dynamic, evolving data**
- **From offline to real-time analysis**
- **From centralized to distributed resources**

- Science increasingly driven by large data sets; massive multi-source, multi-wavelength data
- Large interdisciplinary scientific collaboration
- Science extraction: FAIR multi-source data services (multi-messager)
- Increasing use of ML/DL: data analysis and HPC simulations

SKAO pathfinder: SKA shaping-strategy & End-to-End partnership

Continuous observation dependent data stream processing/reduction



SRC Network: science-driven data analysis and modelling

SKA KSP, PIs, user community

SKA
Regional Centres Network

multi-source data analysis, advanced science data products
data archiving and reuse

Scientific platform of distributed services
Storage, Computing (HDA, HPC, AI), Archiving

NRENs/transnational Data Logistics

SRC1 SRC2 SRC3 SRCn

Existing Shared centralised Infrastructure providers (HPC, Cloud, Storage)

SKA Science Archive
600 PB/yr

SKAO science users

Broader science community

Shared with other communities: Space & Earth Systems Observation

SKA observatory

From edge -> centralised infrastructures

- Observation dependent continuous data stream logistics (stateful services)
- Edge computing: numerical beam forming of signals, removal of radio-frequency interference
- Data loss-compression and reduction

Centralised HPC/HDA operational infrastructures

- Storage and computing capabilities/capacities
- High-rate data processing
- Complex HDA workflows (processing & calibration)

Observation products (events, images, cubes)

- Data models (standards, metadata, provenance)
- Archiving and distribution (data placement)

- > **Machine Learning moving to the edge**

SKA Regional Centres (SRCs)

New organisational, operational, business model

- SKA-driven shaping strategy (providers, science users)

Scientific software platform

- Services across distributed infrastructures
- Multi providers (Cloud, HPC, Storage), Federated AAI
- Application-dependent shared resource efficiency

Application workflows

- Diverse and complex workflows (HDA, HPC, AI)
- Data logistics in multi-provider context, provenance

Data archiving, curation and reuse

- Primary and secondary scientific data products
- FAIR multi-source data/software services

Scientific Users

- Key SKA Projects and PI granted observation projects
- Reuse of SKA data products: multi-messenger

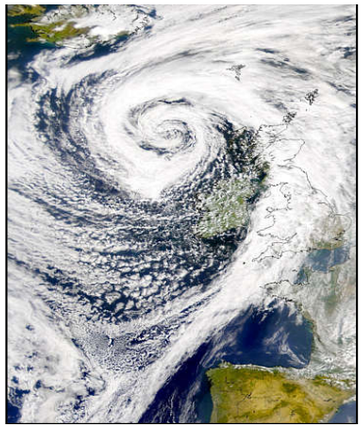
-> **Distributed infrastructures (HPC, HDA)**

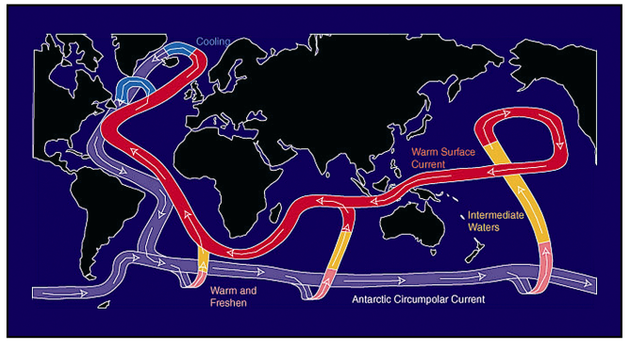
Climate system: a scientific and societal chalenge

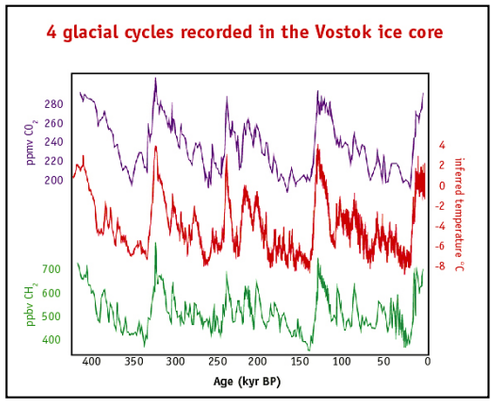
Several complex and multi-physics processes to be simulated

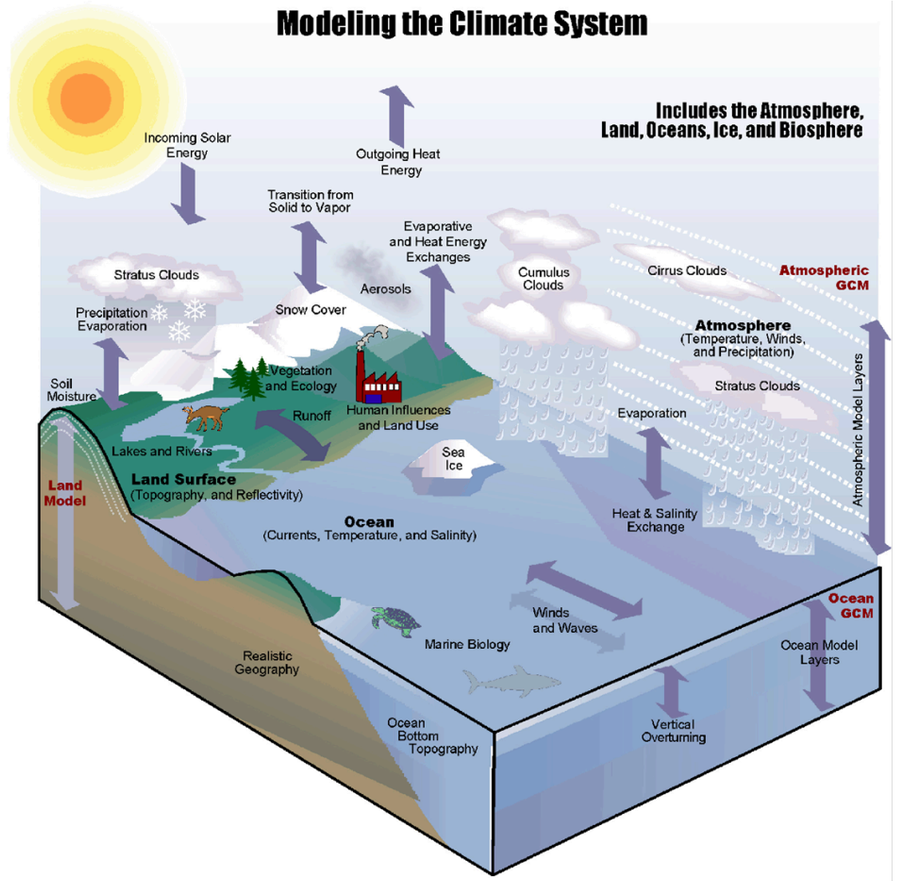
Several interacting processes

Large range of time scales: from days to months, years, decades and millennia

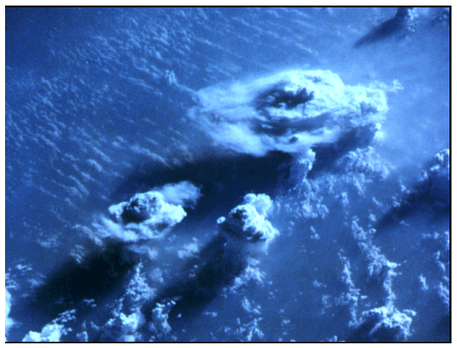


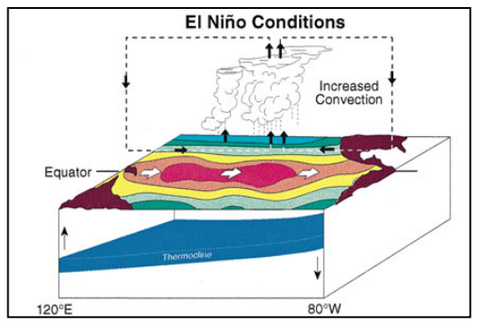


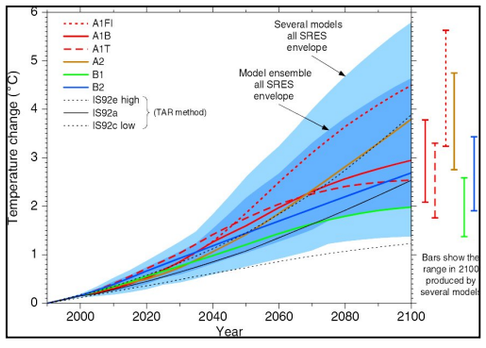




Large range of space scales: from local to regional, continental and global







Inherently non-linear dynamical systems

Capacity/Capability demanding <-> large volume of data

A number of models:

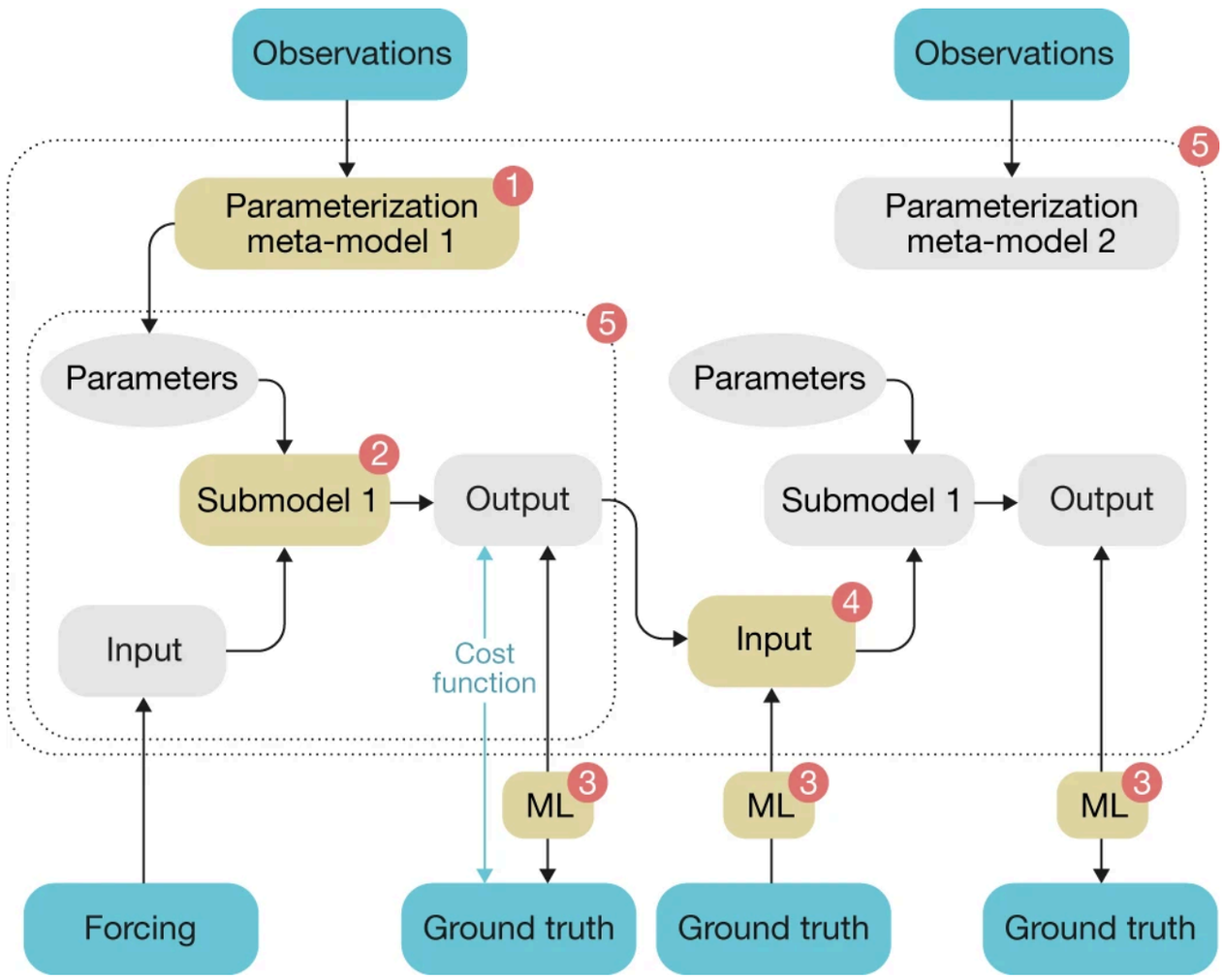
- configuration (parameterisation)
- experiences (scenarios)
- ensemble of realisations (uncertainty)

Detection, attribution and prediction of extreme events and modes of climate variability

Climate science, impacts and societal services

ML & physical modelling

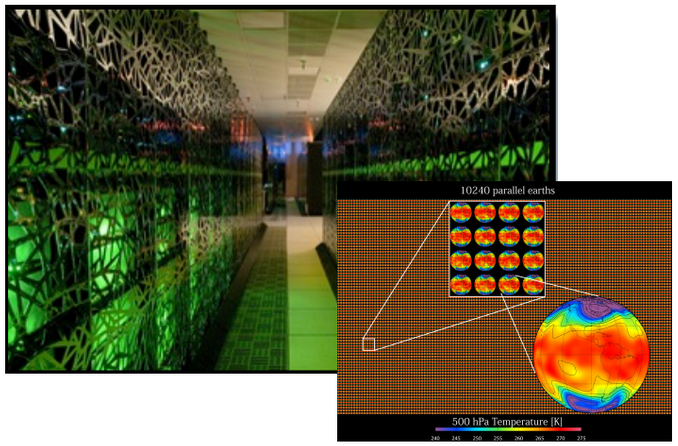
Reichstein et al, 2019

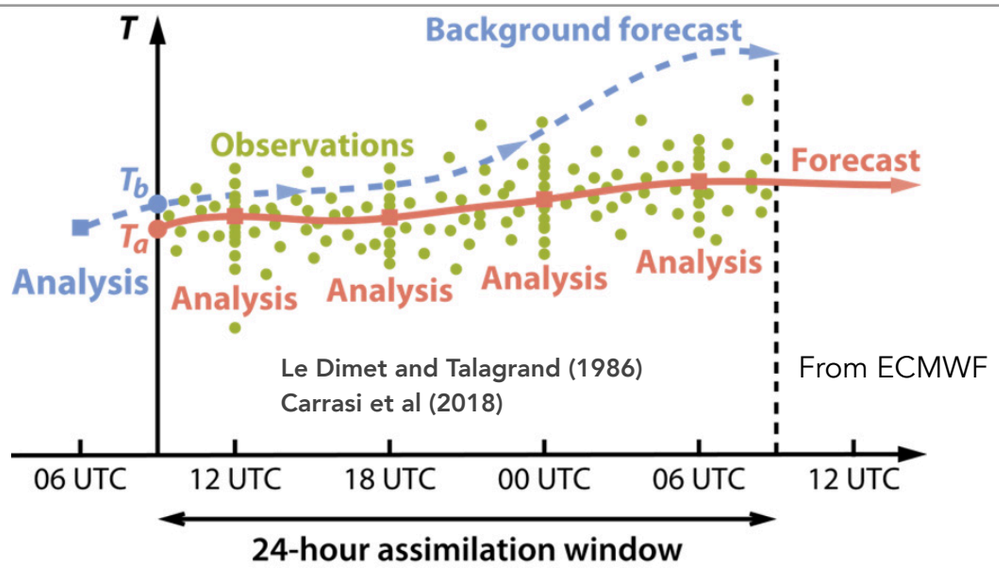


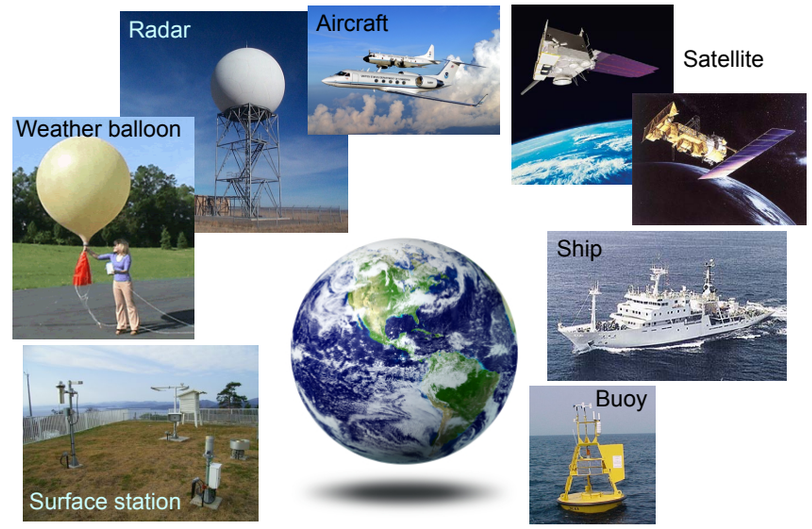
1. Improving parameterisations (global atmospheric modelling)
2. Physical sub-models -> ML models
3. Analysis Model-Observation mismatch
4. Constraining sub-models (from ML)
5. Surrogate modelling or emulations (ML emulators)

- **Interpretability, Physical consistency**
- **Data complexity, uncertainty and noise**
- **Limited available labelled data sets**
- **Extrapolation versus prediction**
- **Computational cost & time: transfer learning**

Data assimilation: numerical weather prediction





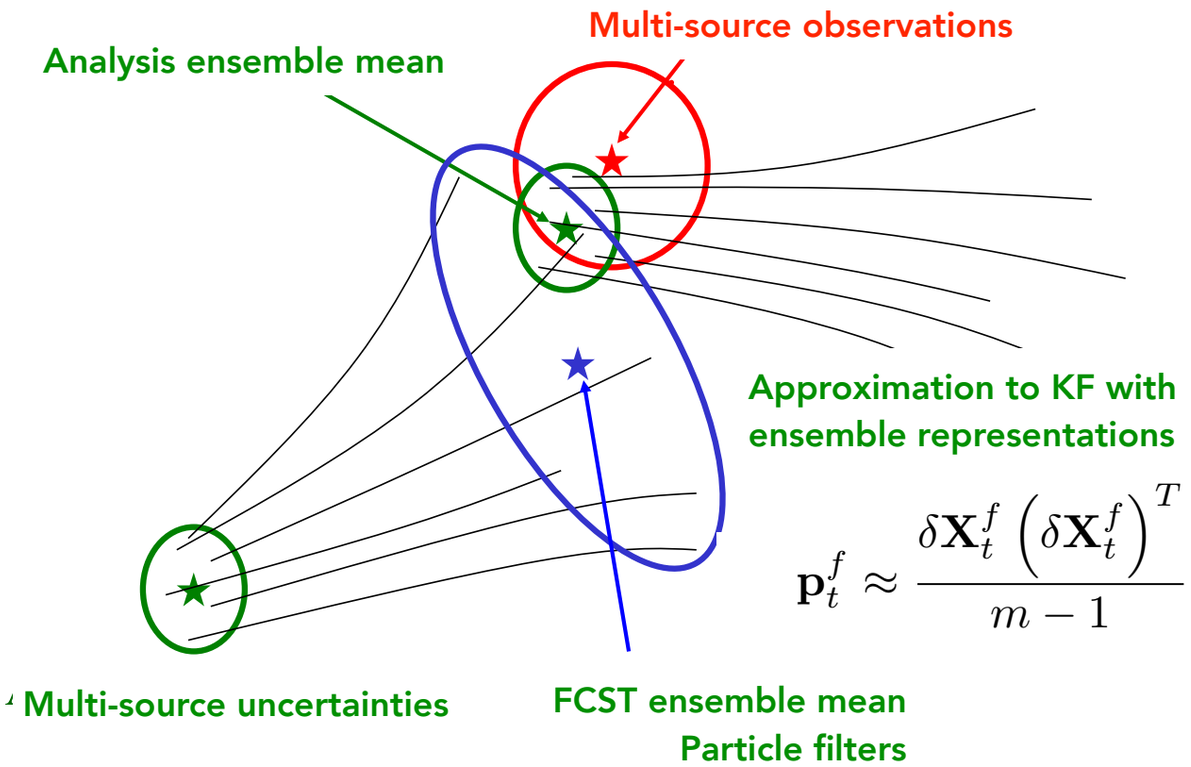


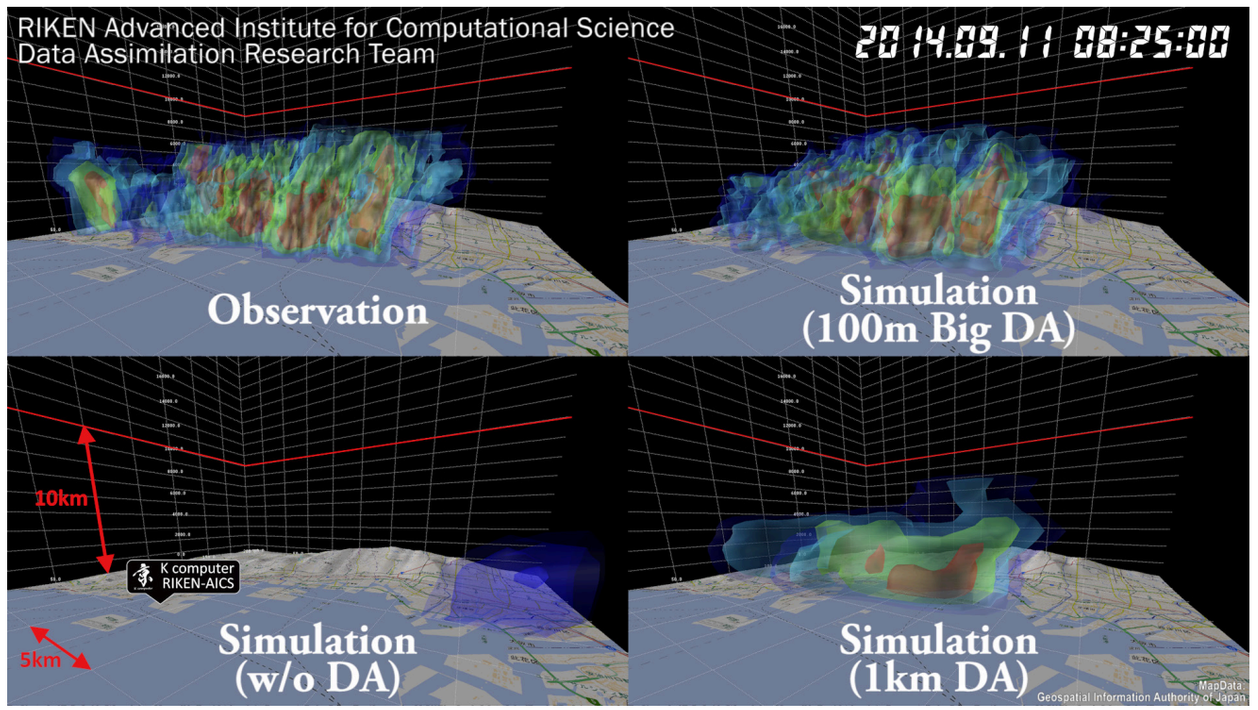
Multi-source and multi-scale data

From edge: streaming data processing and reduction

to centralised infrastructures (HPC, Cloud):

large ensemble simulations & Bayesian inference





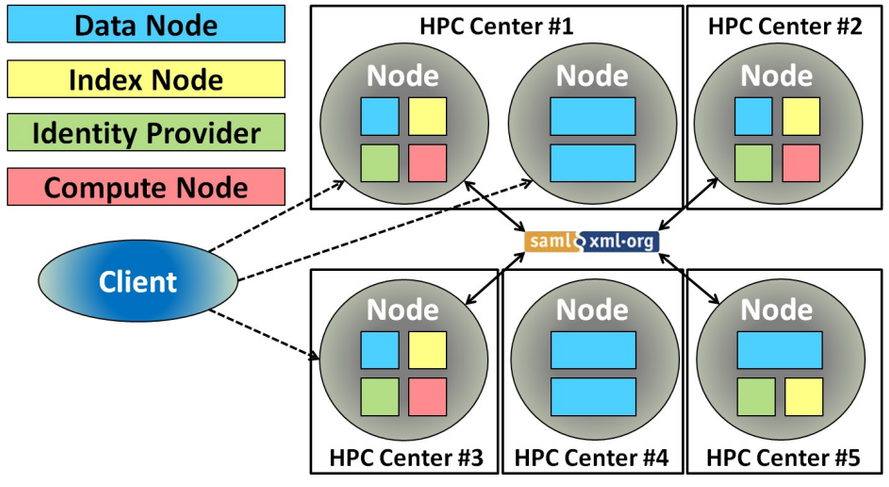
30 minutes forecasting

T. Miyoshi, Riken aics

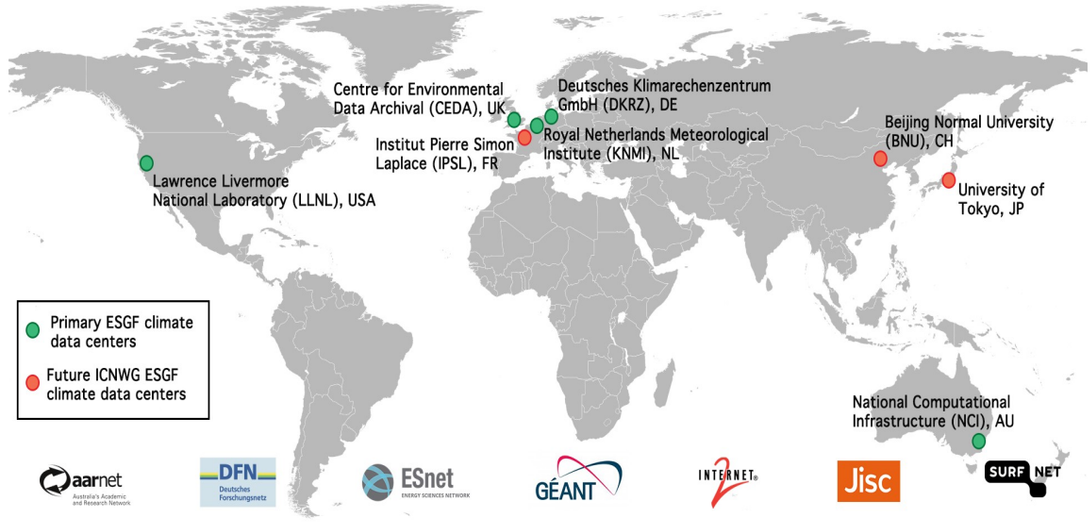
- Data assimilation **is** equivalent **to a** machine learning problem (Abarbanel et al (2018), Bocquet et al (2018)
- Artificial Intelligence: a natural framework **to take up challenges of Earth Observation and Modelling**

Numerical laboratory: Earth System Grid Federation

ESGF Test Infrastructure based on virtual machines



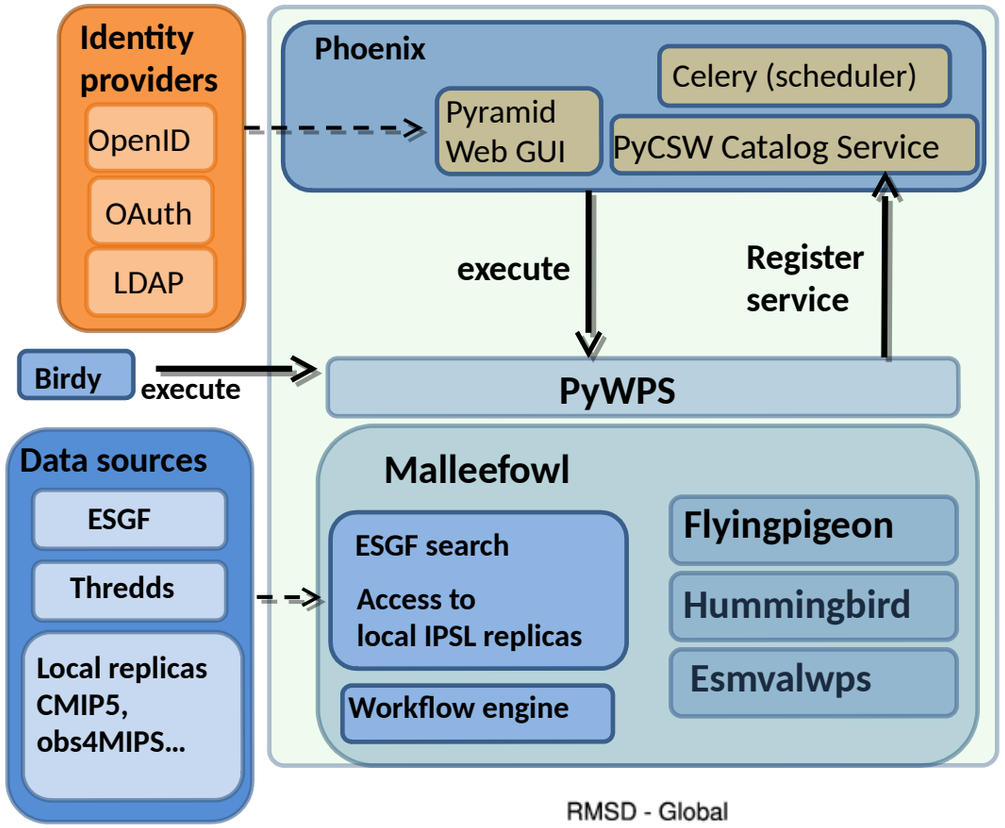
~ PBs scale



International Climate Networking Group

Climate Model Assessment Framework (CLiMAF)

Web processing services (WPS)



RMSD - Global

from S. Denvil

- Access to models, simulations and observations
- Share data analytic methods and tools
- Advanced management and documentation of models, simulations (indexation, metadata, provenance)
- Induction of a broad research and user community
- Data analysis platforms and web services
- Pervasive provenance system

Paleoclimate Modelling

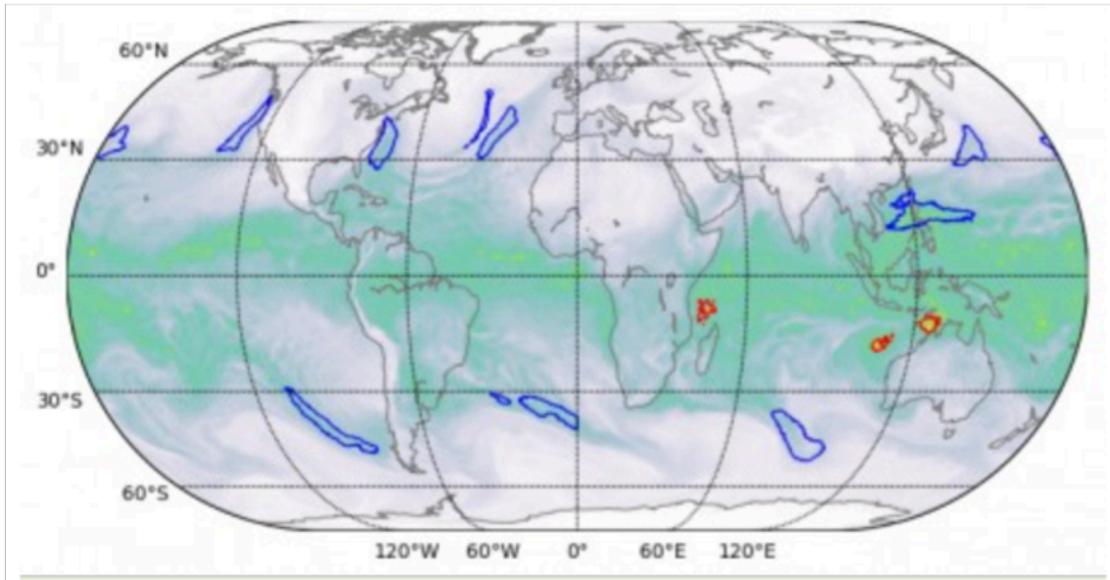
PMIP3

Intercomparison Project

Machine learning - Earth Systems Science

Analytical task	Scientific task	Conventional approaches	Limitations of conventional approaches	Emergent or potential approaches
Classification and anomaly detection				
	Finding extreme weather patterns	Multivariate, threshold-based detection	Heuristic approach, ad hoc criteria used	Supervised and semi-supervised convolutional neural networks[41,42]
	Land-use and change detection	Pixel-by-pixel spectral classification	Shallow spatial context used, or none	Convolutional neural networks[43]
Regression				
	Predict fluxes from atmospheric conditions	Random forests, kernel methods, feedforward neural networks	Memory and lag effects not considered	Recurrent neural networks, long-short-term-memories (LSTMs)[89,99,100]
	Predict vegetation properties from atmospheric conditions	Semi-empirical algorithms (temperature sums, water deficits)	Prescriptive in terms of functional forms and dynamic assumptions	Recurrent neural networks[90], possibly with spatial context
	Predict river runoff in ungauged catchments	Process models or statistical models with hand-designed topographic features[91]	Consideration of spatial context limited to hand-designed features	Combination of convolutional neural network with recurrent networks
State prediction				
	Precipitation nowcasting	Physical modelling with data assimilation	Computational limits due to resolution, data used only to update states	Convolutional–LSTM nets short-range spatial context[92]
	Downscaling and bias-correcting forecasts	Dynamic modelling and statistical approaches	Computational limits, subjective feature selection	Convolutional nets[72], conditional generative adversarial networks (cGANs)[53,93,101]
	Seasonal forecasts	Physical modelling with initial conditions from data	Fully dependent on physical model, current skill relatively weak	Convolutional–LSTM nets with long-range spatial context
	Transport modelling	Physical modelling of transport	Fully dependent on physical model, computational limits	Hybrid physical–convolutional network models[68,94]

Deep-Learning Methods to Understand Weather Patterns (LBL), 2018 Gordon Bell Prize (https://bit.ly/2X42Vur)



High-quality segmentation results produced by deep learning on climate datasets.

Reichstein et al, 2019

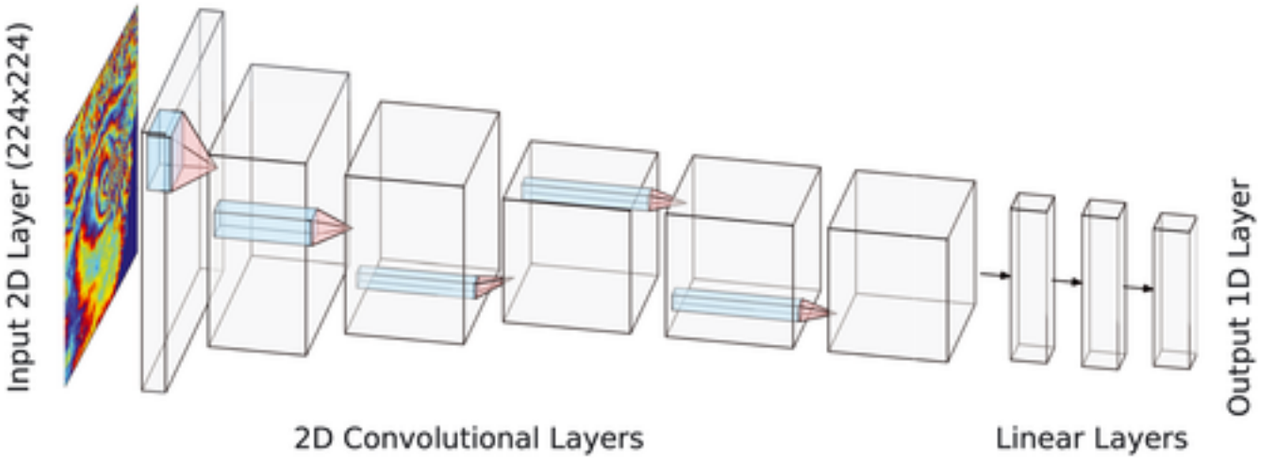
ML classification of volcanic deformation: InSAR data

- **Earth Observation (routinely)**
  - Volcanoes in remote regions
- **InSAR satellite remote sensing**
  - High-resolution deformation signal
  - Large geographic area & coverage
  - Statistical link to eruption
- **Increasingly large data sets**
  - Sentinel-1 (A and B) 6-day repeat cycle
  - More than 10-TB/day, 5 PB (2014-2020)
  - Challenge manual inspection
  - Timely dissemination of information
- **ML & satellite-based volcano geodesy**

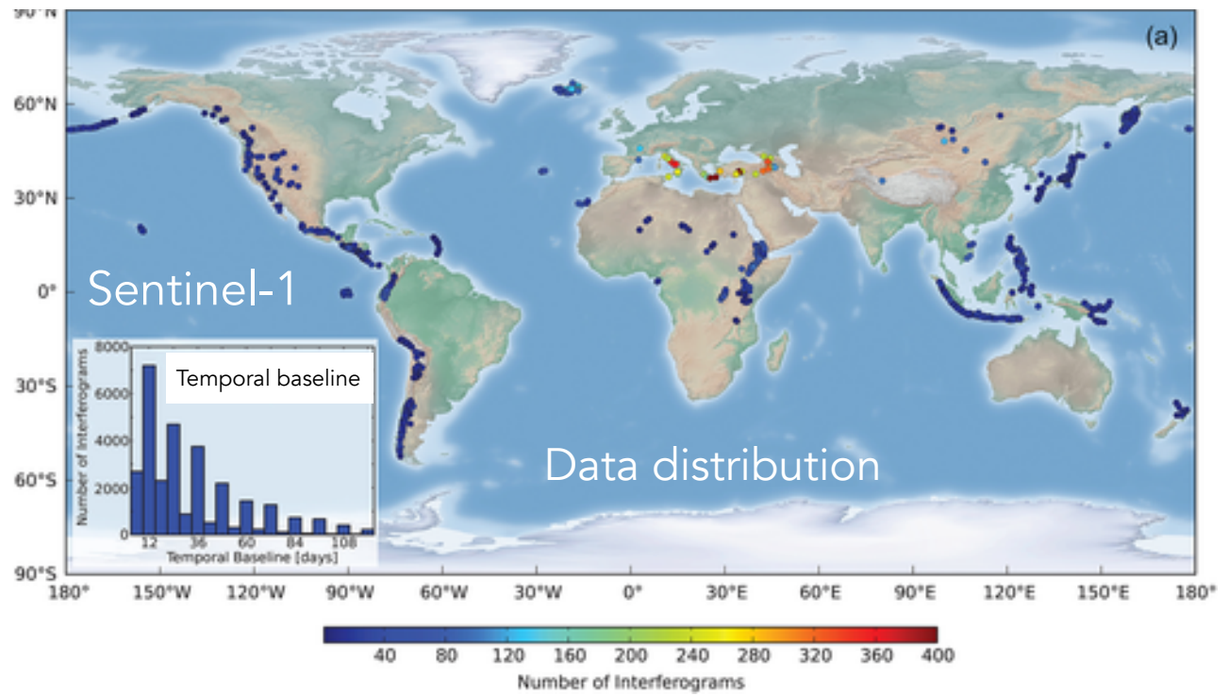
  Detection: deformation patterns

  Classification: interferometric fringes in wrapped interferograms (no atmospheric corrections)

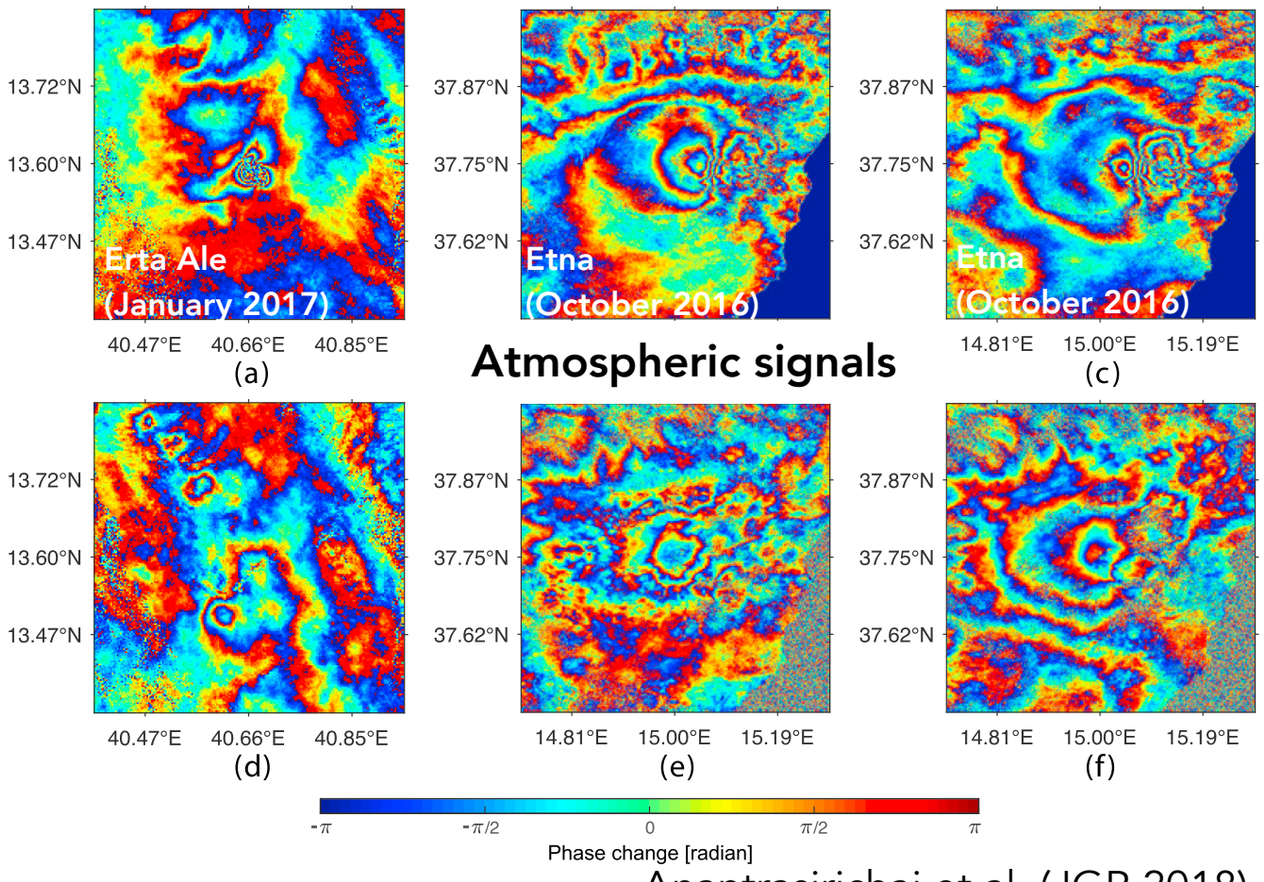
  Transfer learning with pre-trained networks (AlexNet)



> 30,000 ST interferograms over 900 volcanoes (2016-2017)



Volcanic ground deformation Sentinel-1 interferograms

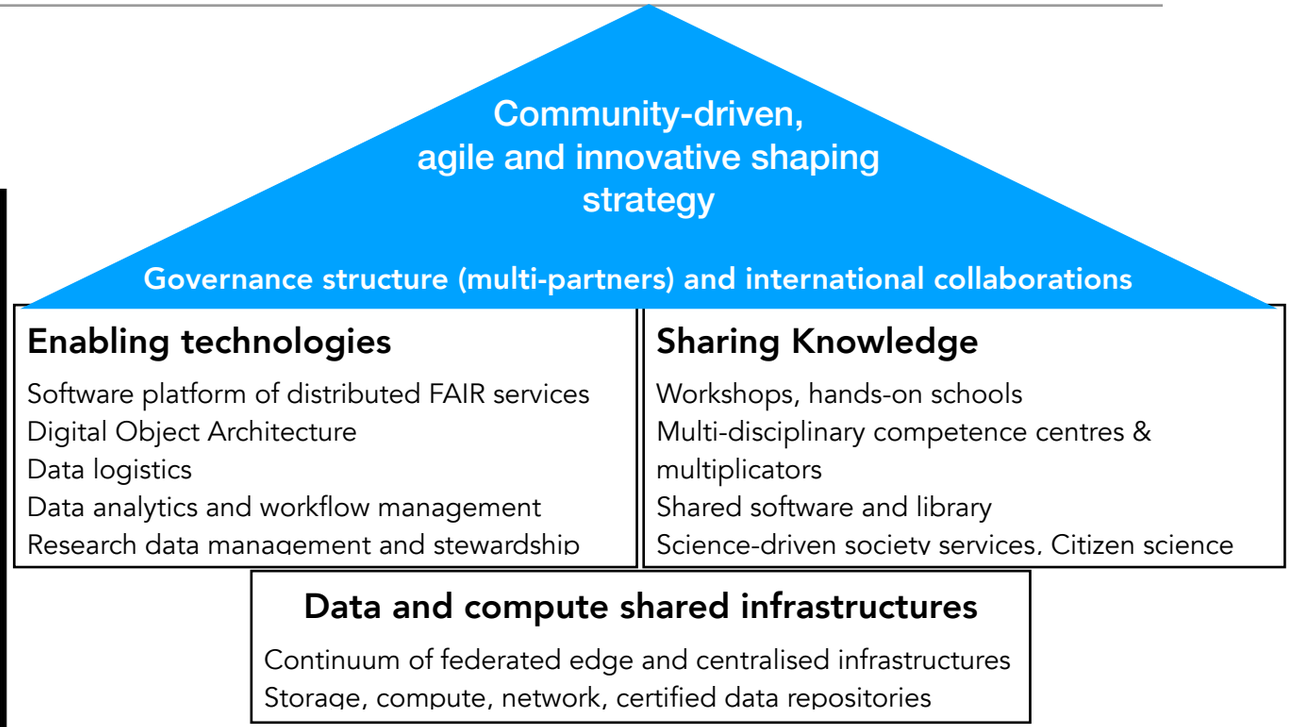


Anantrasirichai et al. (JGR,2018)

A Digital Object Architecture with a spanning Layer

Software Platform of services

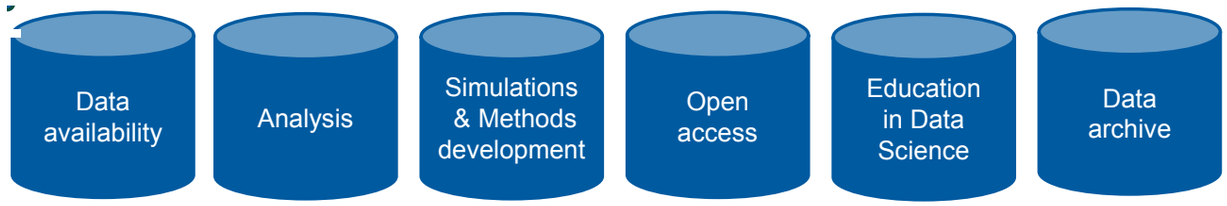
- across edge and centralised Data and computing (HPC, Cloud) Infrastructures
  - ✳ Persistent/transient storage (variable data life cycles)
  - ✳ Data model and storage abstraction layers
- End to End data logistics and data reduction
- Flexible services (storage, compute, communications)
- Rendering services (visualise, analyse)

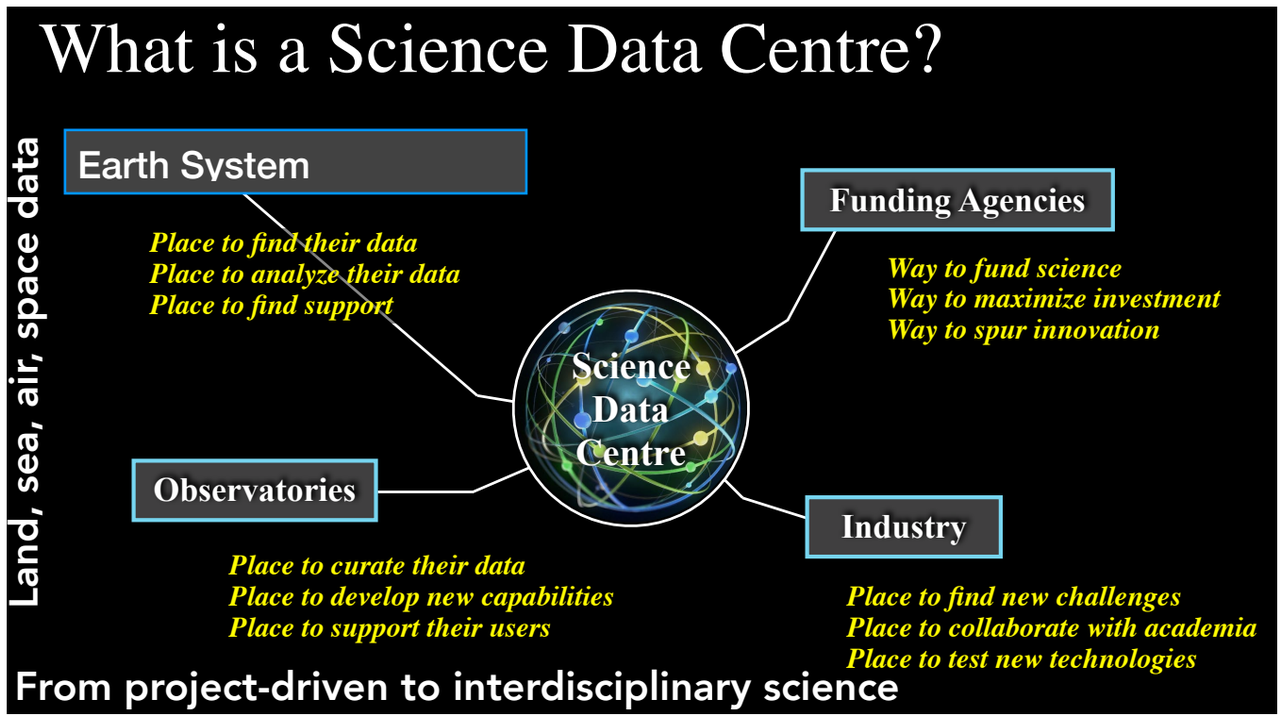


Digital Object Architecture and software services

Centralised Environments (HPC, Cloud)

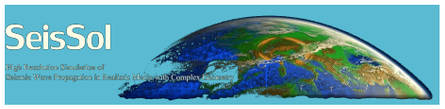
- Multiple research communities
- Convergence between HPC and HDA
  - ✳In-situ Data processing and reduction
  - ✳ Batch and streaming execution models
  - ✳ Containers technology (Kubernetes, Singularity, beyond)
  - ✳ Integrate different programming models
  - ✳Provenance systems
  - ✳HPC/HDA workflows including machine learning
  - ✳Leveraged HPC libraries for HDA and AI
- Collaborative, flexible and resilient environments





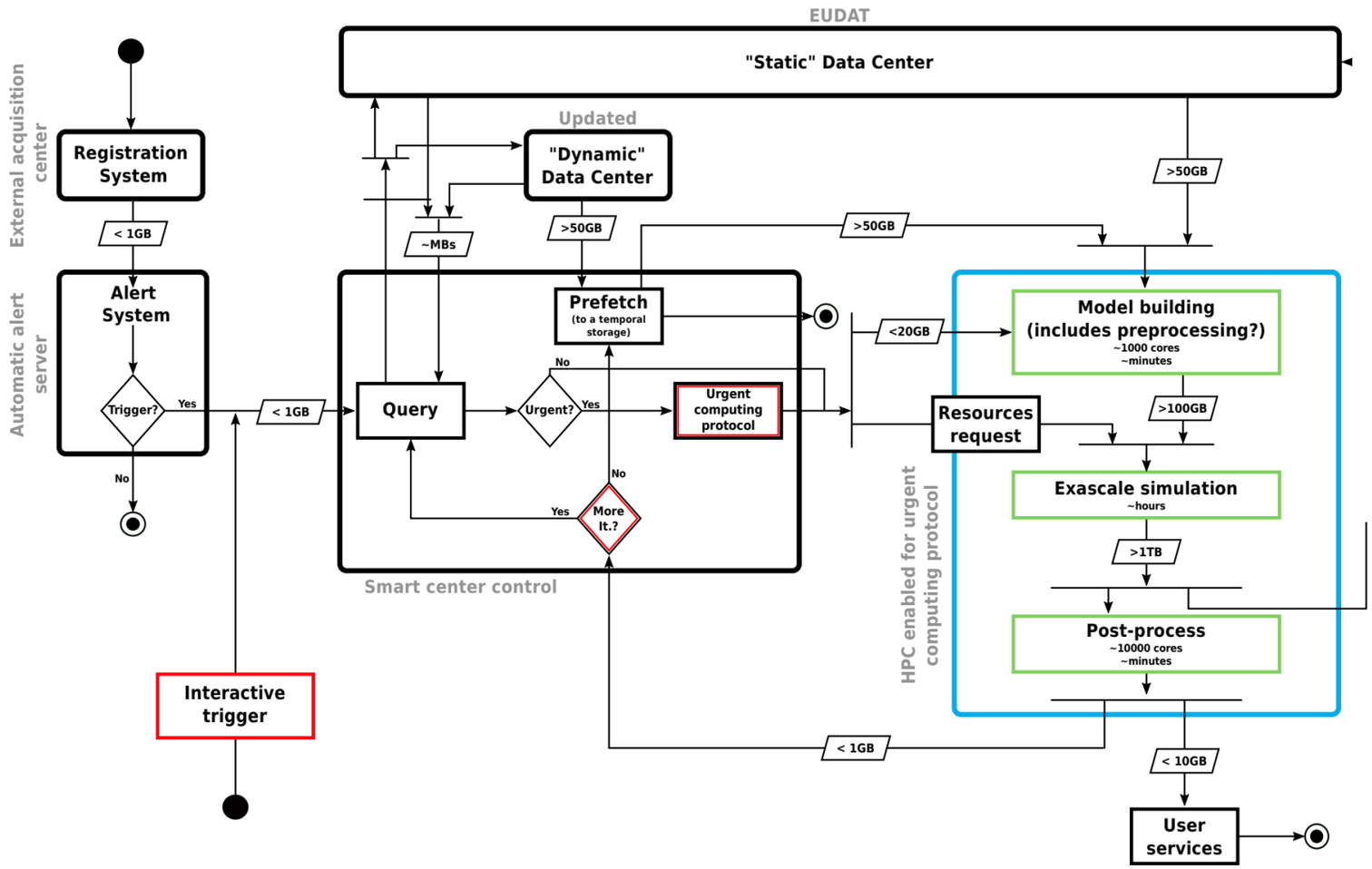
Urgent Computing: environmental risk and resilience

ChEESE



SALVUS
IMAGE EVERYTHING

ExaHyPE
An Exascale Hyperbolic PDE Engine



PRACE **Distributed Supercomputing Infrastructure**

26 members, including

5 Hosting Members

(Switzerland, France, Germany, Italy and Spain)

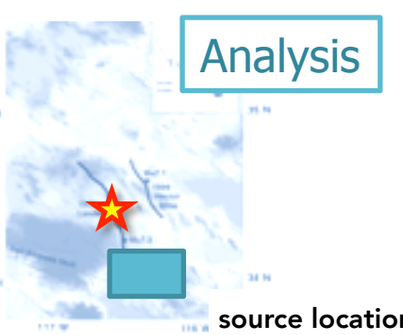
652 scientific projects **enabled**

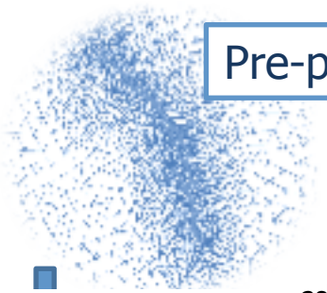
110 PFlops/s of peak performance on **7 world-class systems**

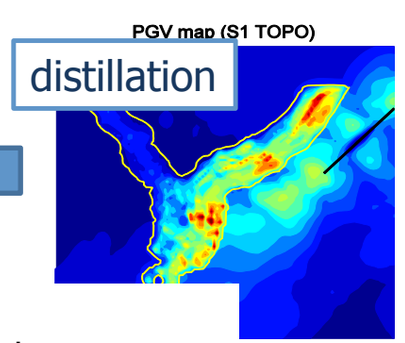
>12.000 people trained by **6 PRACE Advanced Training Centers** and others events

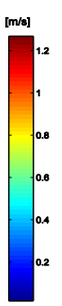
Access prace-ri.eu/hpc-acces

Barcelona Supercomputing Center
Centro Nacional de Supercomputación









Wide area workflows

Data logistics

HPC services/capacity

Stakeholders