



Big Data for the study of climate and air quality

Francisco J. Doblas-Reyes
BSC Earth Sciences Department

What

Environmental modelling and forecasting

Why

Our strength ...

- ... research ...
- ... operations ...
- ... services ...
- ... high resolution ...

How

Develop a capability to model air quality processes from urban to global and the impacts on weather, health and ecosystems

Implement climate prediction system for subseasonal-to-decadal climate prediction

Develop user-oriented services that favour both technology transfer and adaptation

Use cutting-edge HPC and Big Data technologies for the efficiency and user-friendliness of Earth system models

Earth system
services

Climate
prediction

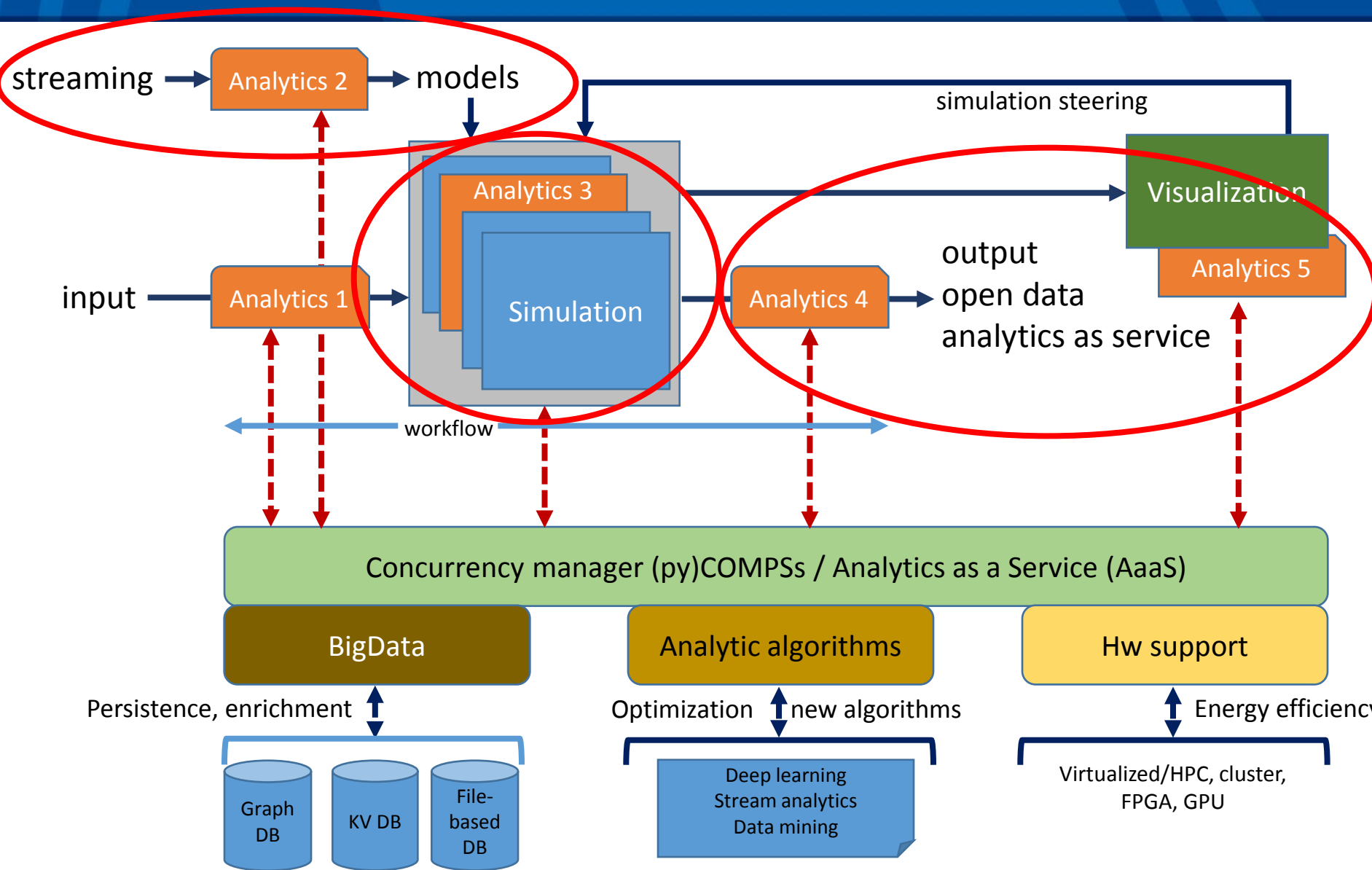
Atmospheric
composition

Computational
Earth sciences

- There are problems involving large, complex datasets: operational weather and air quality prediction.
- There are large problems involving data: simulation of anthropogenic climate change.
- And there are Big Data problems: dealing with heterogeneous data sources to produce end-user information with a weather, climate and air quality component.

I will not address the issue of open data and assume that access to data is not a difficulty.

A conceptual model from CS





Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

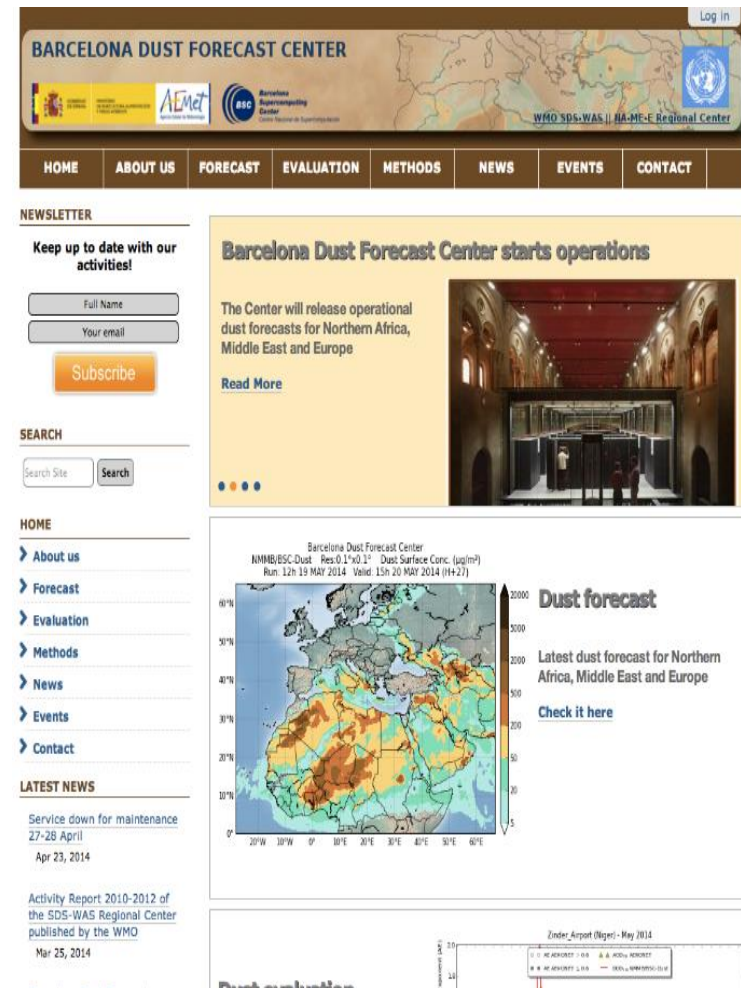
Case 3: Analytics as a service

Barcelona Dust Forecast Center (BDFC) and Sand and Dust Storm-Warning and Advisory System (SDS-WAS) for North Africa, Middle East and Europe, both operated jointly by BSC-CNS and AEMET:

- BDFC is the first specialized WMO centre for mineral dust prediction, and provides forecasts to WMO GTS, EumetCAST and AEMET.
- SDS-WAS NAMEE is a research and development multi-model system gathering model output, AERONET observations and satellite data.

<http://dust.aemet.es>

<http://sds-was.aemet.es>



BARCELONA DUST FORECAST CENTER

HOME ABOUT US FORECAST EVALUATION METHODS NEWS EVENTS CONTACT

NEWSLETTER

Keep up to date with our activities!

Full Name

Your email

Subscribe

SEARCH

Search Site Search

HOME

- ▶ About us
- ▶ Forecast
- ▶ Evaluation
- ▶ Methods
- ▶ News
- ▶ Events
- ▶ Contact

LATEST NEWS

Service down for maintenance
27-28 April
Apr 23, 2014

Activity Report 2010-2012 of the SDS-WAS Regional Center published by the WMO
Mar 25, 2014

Barcelona Dust Forecast Center starts operations

The Center will release operational dust forecasts for Northern Africa, Middle East and Europe

Read More

Dust forecast

Barcelona Dust Forecast Center
MNM-BSC-Dust - Res: 0.1°x0.1° - Dust Surface Conc. (µg/m³)
Run: 12h 19 MAY 2014 Valid: 15h 20 MAY 2014 (H+2)

Latest dust forecast for Northern Africa, Middle East and Europe

Check it here

Latest News

Zinder Airport (Niger) - May 2014

CALIOPE air quality operational forecasts

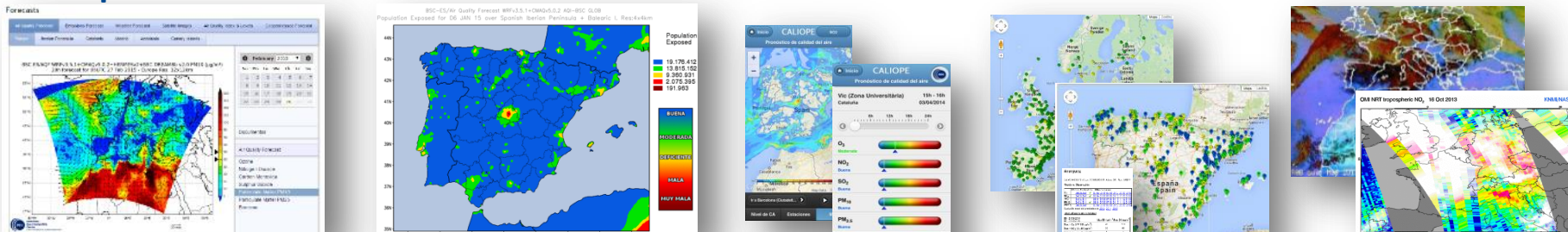


Barcelona Supercomputing Center
Centro Nacional de Supercomputación



AQF CALIOPE system: daily forecast and evaluation
Forecast products

Daily forecast for **meteorology, emissions and air quality**: Europe (12km), Iberian Peninsula (4km), Andalusia, Catalonia and Madrid (1km), since 2007



Search



Categories ▾ Home Top Charts New Releases

My apps

Shop

Games

Family

Editors' Choice



CALIOPE: Air Quality

Barcelona Supercomputing Center Health & Fitness

★★★★★ 164

PEGI 3

This app is compatible with your device.

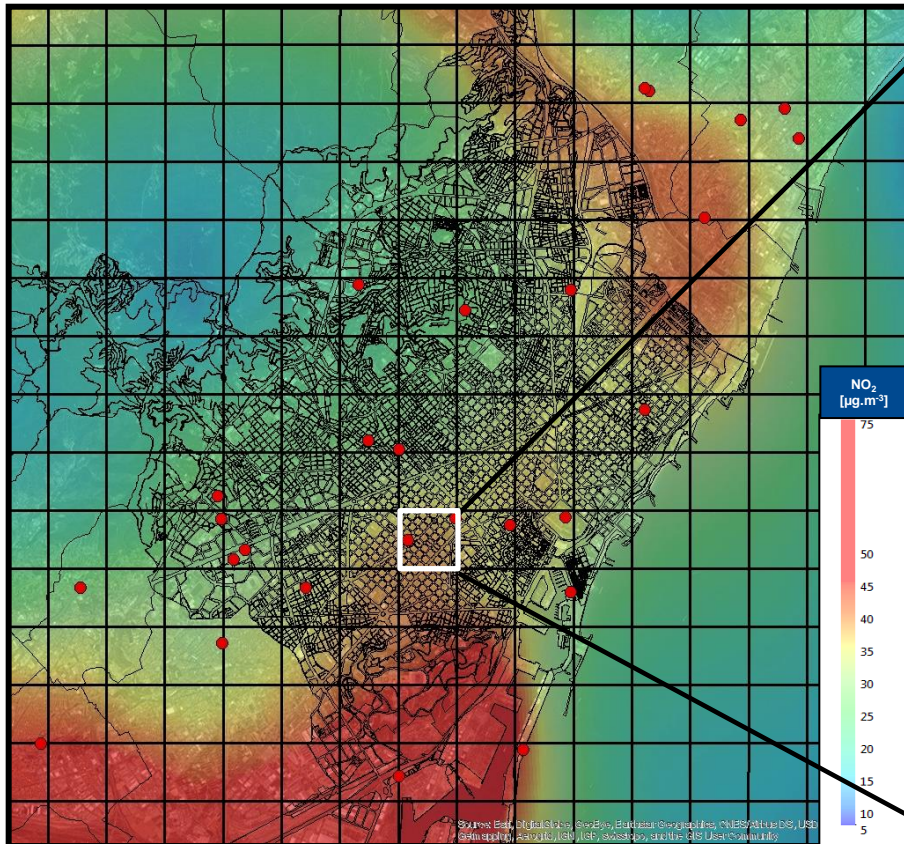
Add to wishlist

Install

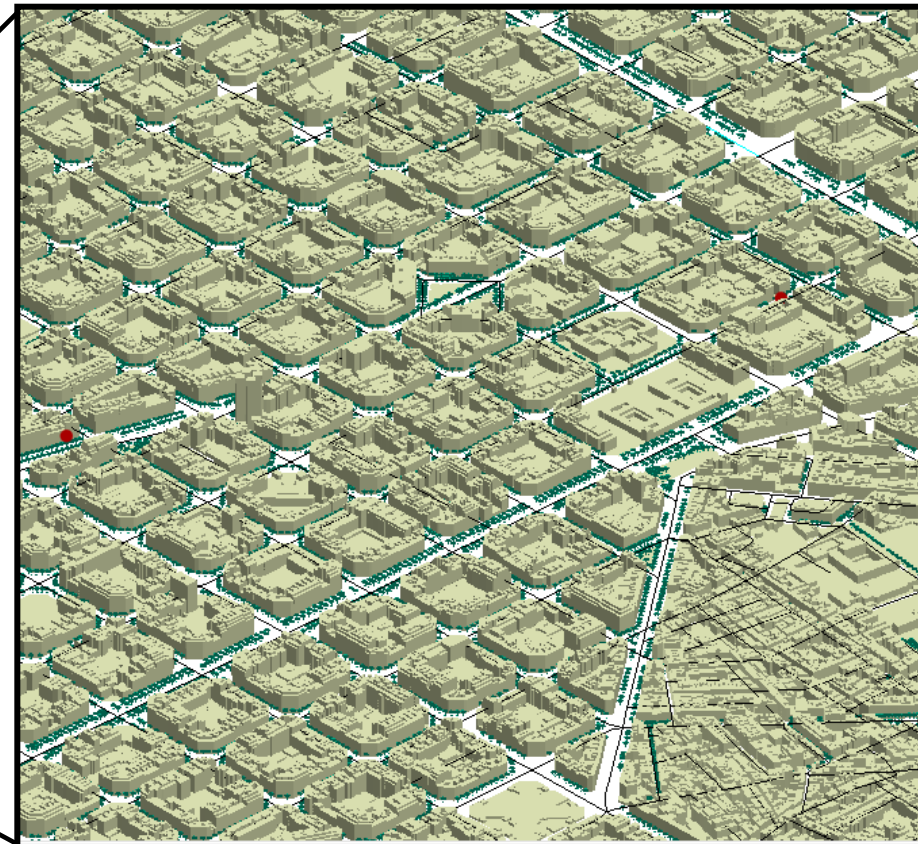
PM10_KF annual average skill evolution (2011-2014)



Where we are now



Where we want to be

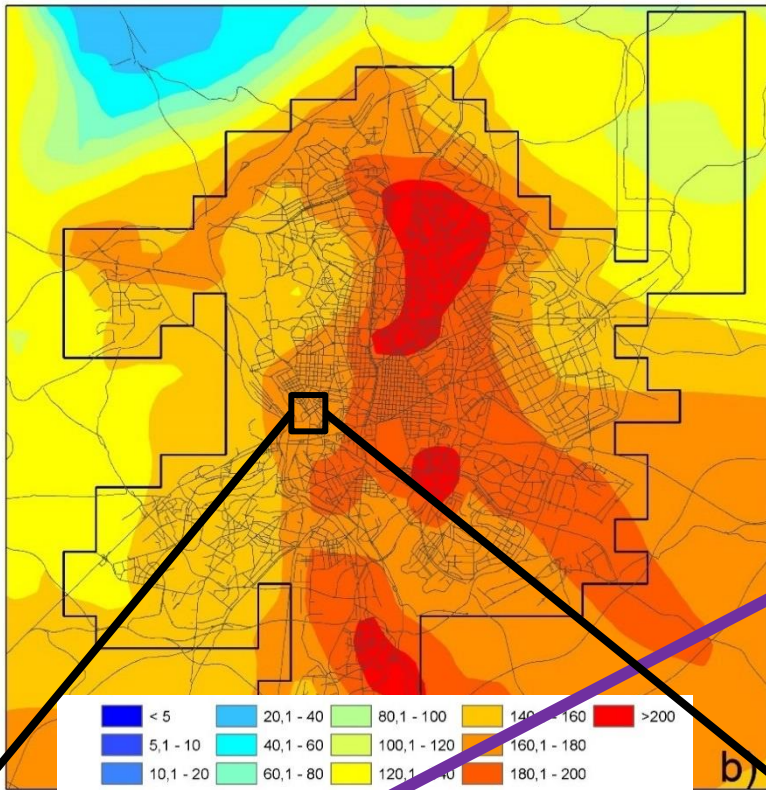


Objective

To develop an air quality model at city scale, enhanced by the use of Big Data technologies, to assess urban air quality.

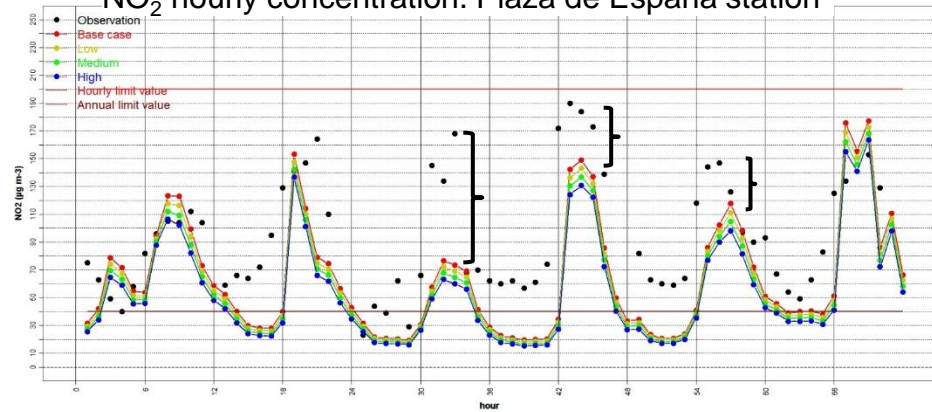
Need of a more detailed view

NO₂ (ug m⁻³) Max h
Base case; Madrid

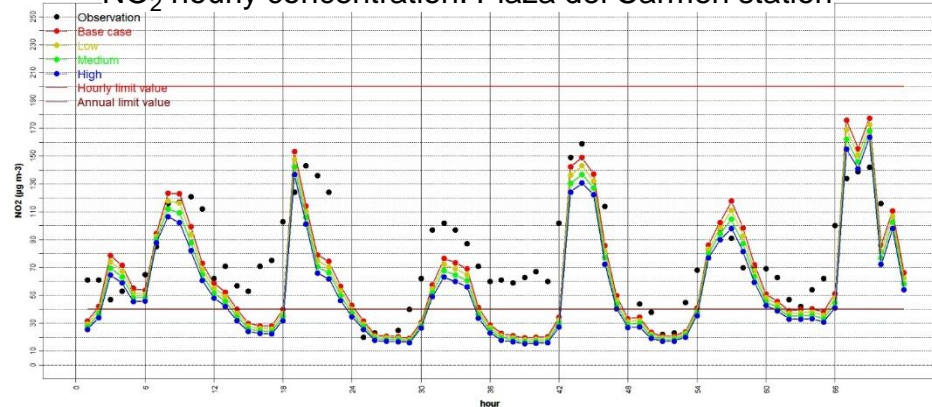


- 1) Mesoscale models (current solutions) show a satisfactory performance to simulate background concentrations (e.g. Plaza del Carmen).
- 2) However, they are not able to reproduce street-level strong concentration gradients (e.g. Plaza de España located close to Gran Via); purpose-built tools are needed.

NO₂ hourly concentration. Plaza de España station



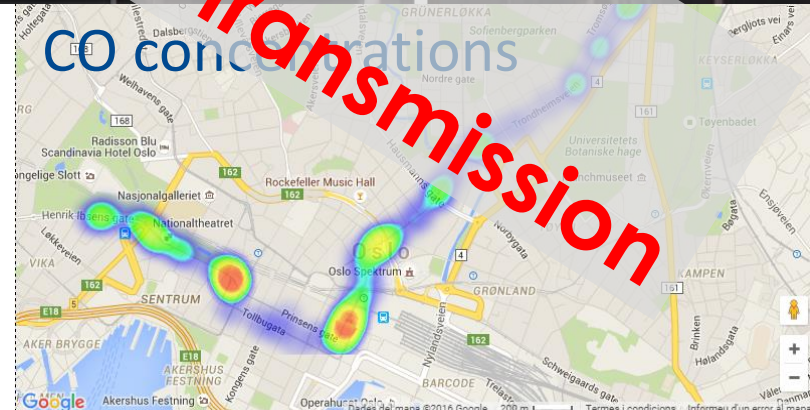
NO₂ hourly concentration. Plaza del Carmen station



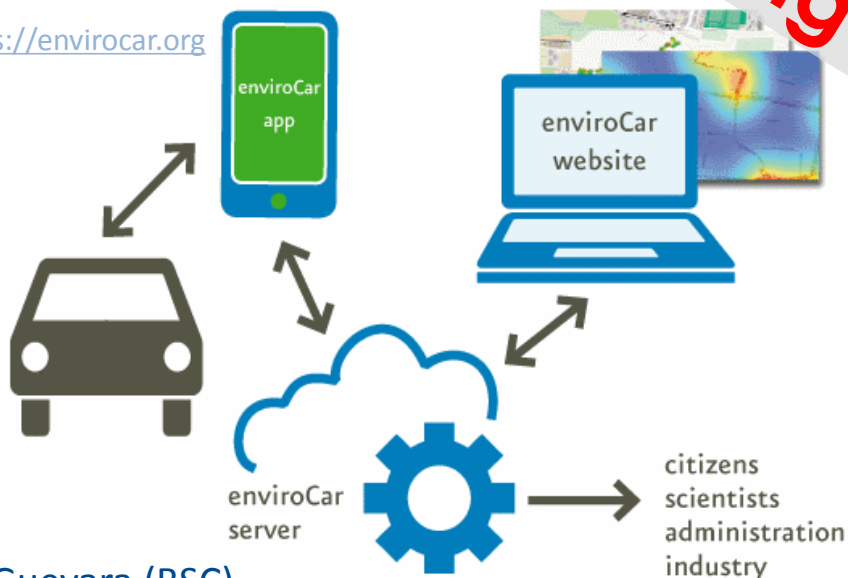
High-resolution air quality modelling requires appropriate emissions

- Collection and processing of sensor-generated data to feed a real-time emission model (and to validate air quality predictions)
- Providing the sensors with the adequate technology is a challenge
- Managing large volumes is another one: sampling 10 Hz, ten variables ~ 30 MB/day, city-wide ~ 300 GB/day (10,000 vehicles)

Benefits of Many more sensors possible before transmission



<https://envirocar.org>



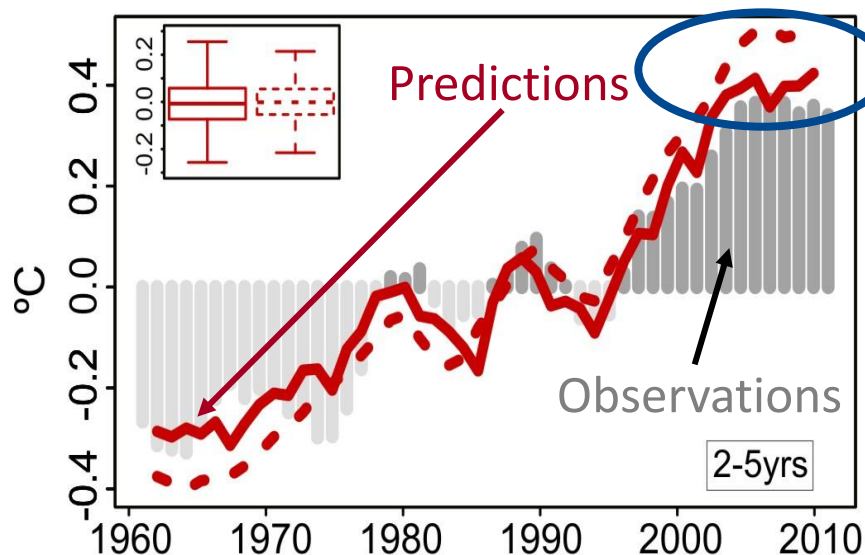
Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

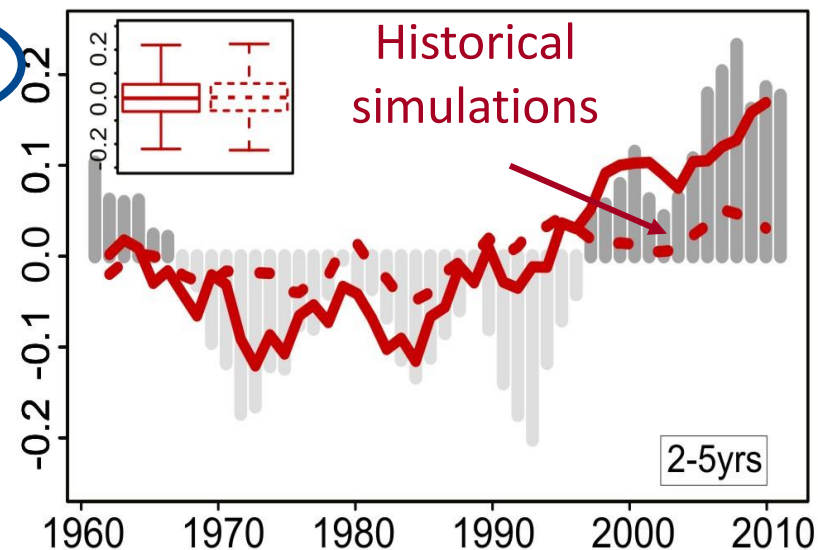
Case 3: Analytics as a service

Global-mean near-surface air temperature and AMV against
GHCN/ERSST3b for forecast years 2-5.

Global mean surface air
temperature (GMST)

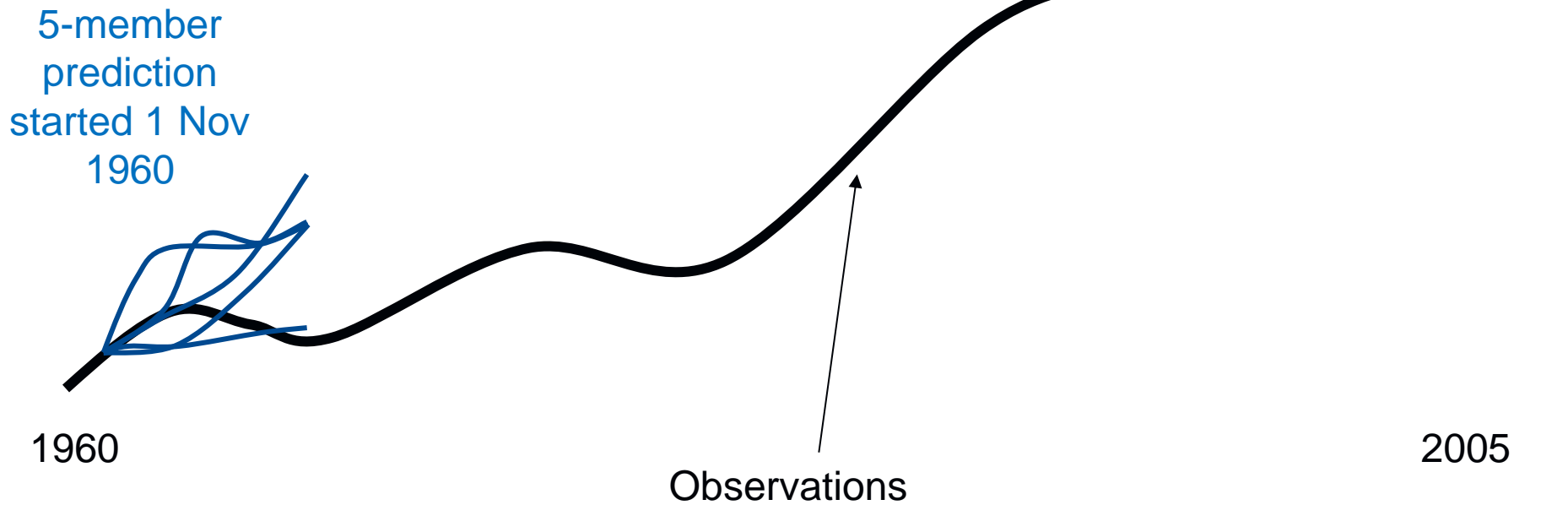


Atlantic multidecadal variability
(AMV)

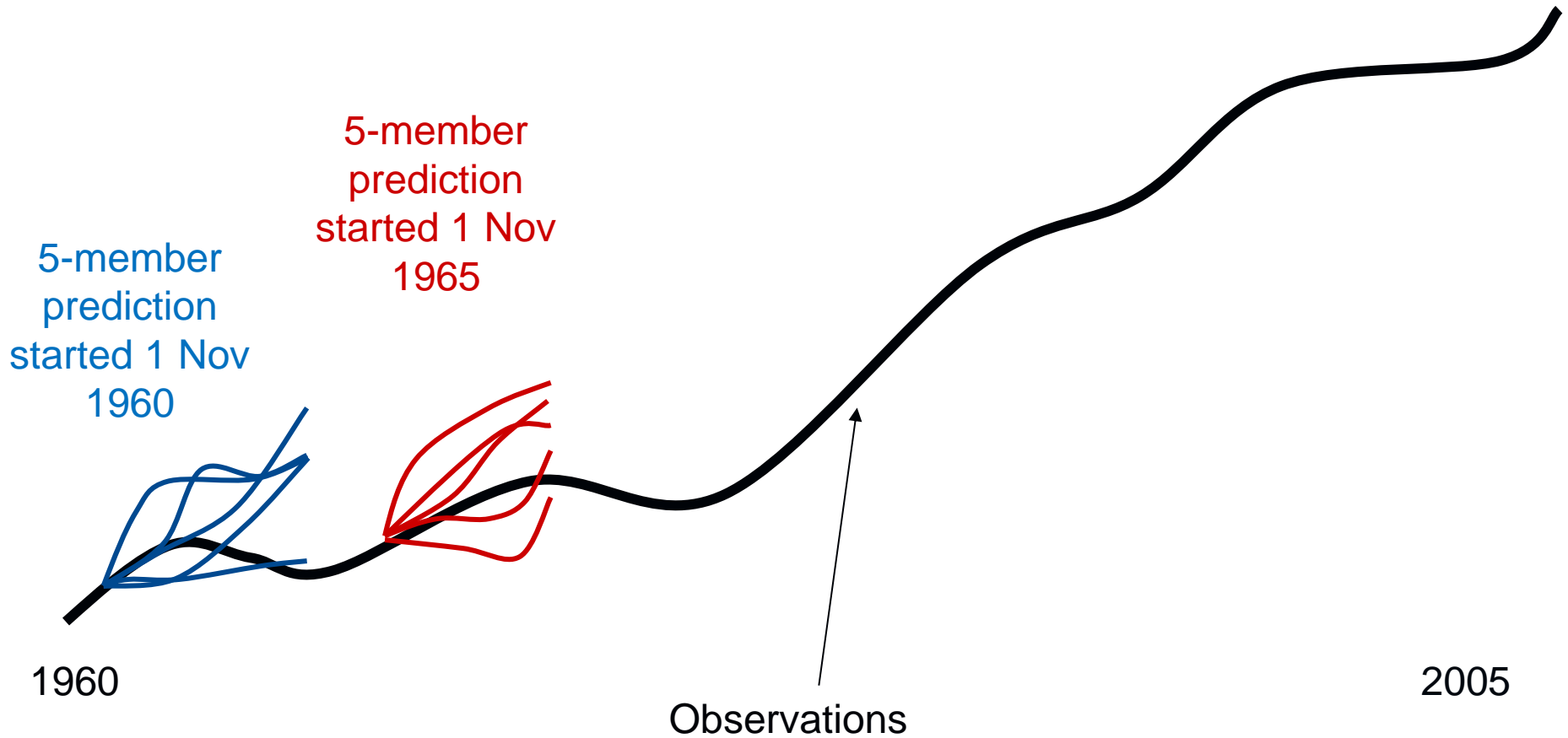


Initialised simulations reproduce the global temperature and some of the AMV tendencies and suggest that initialization corrects the forced model response **and** phases in internal variability.

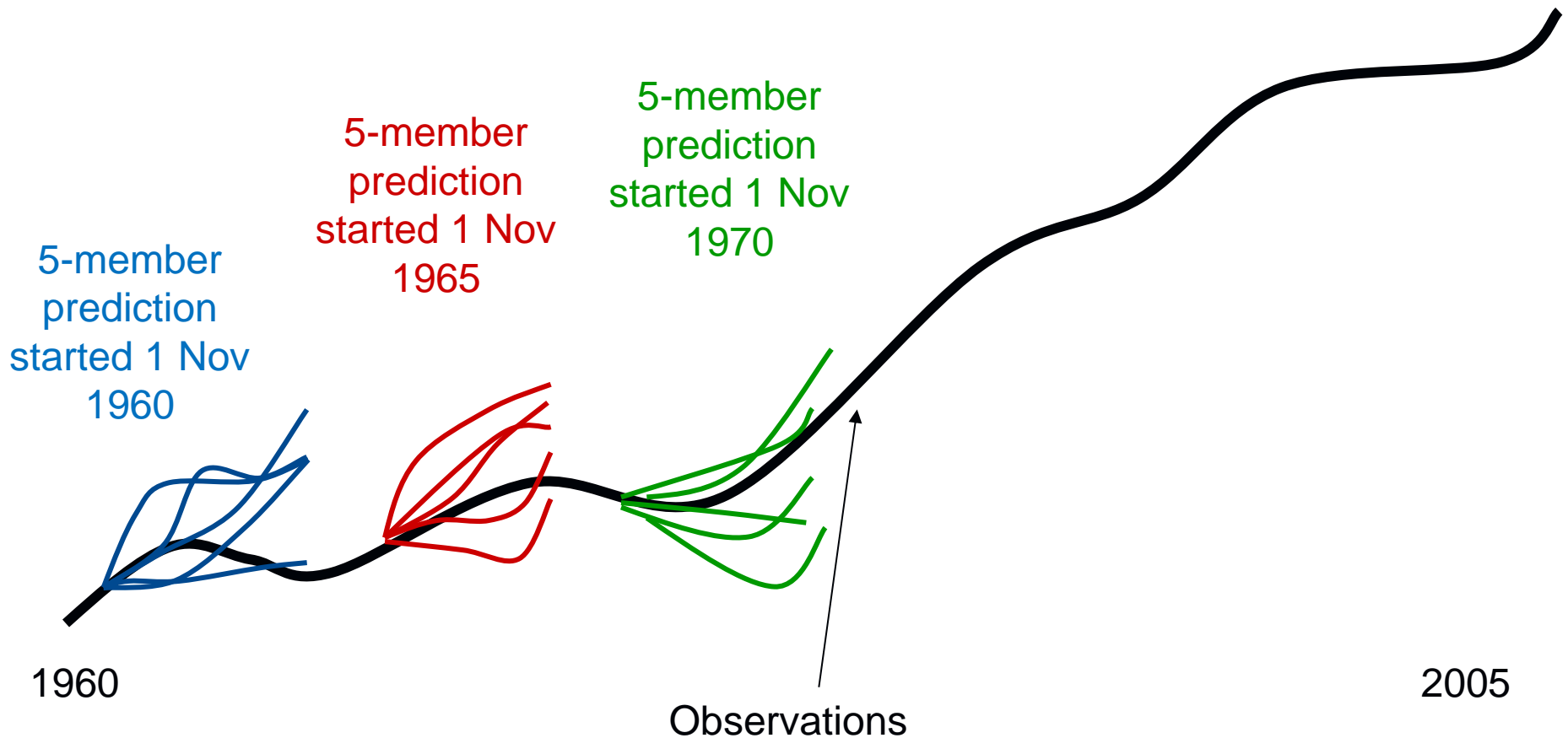
Climate predictions



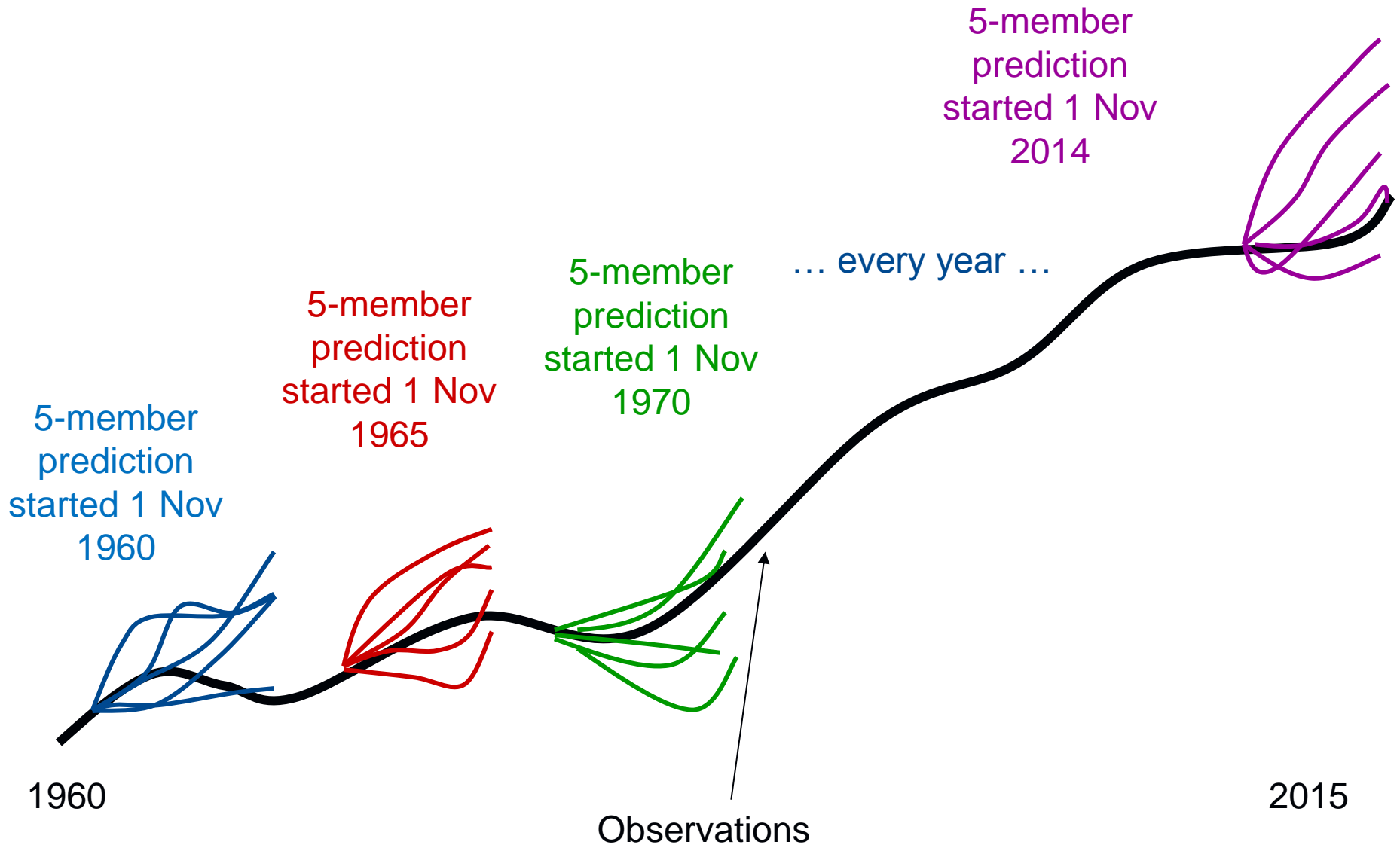
Climate predictions



Climate predictions



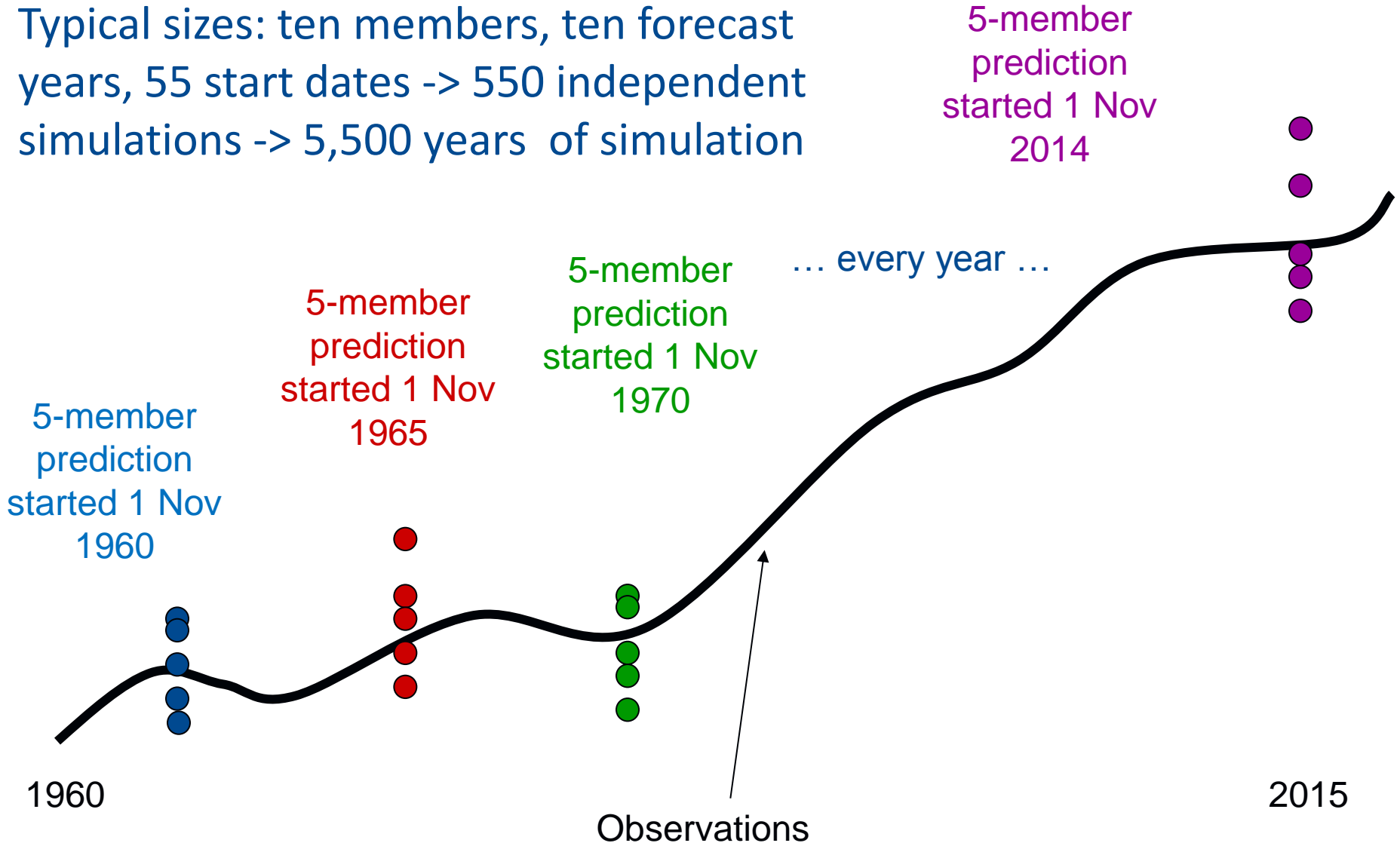
Climate predictions



Climate predictions



Typical sizes: ten members, ten forecast years, 55 start dates -> 550 independent simulations -> 5,500 years of simulation



Running global climate predictions



Climate prediction allows running jobs independently and simultaneously by wrapping together ensemble members for different start dates.

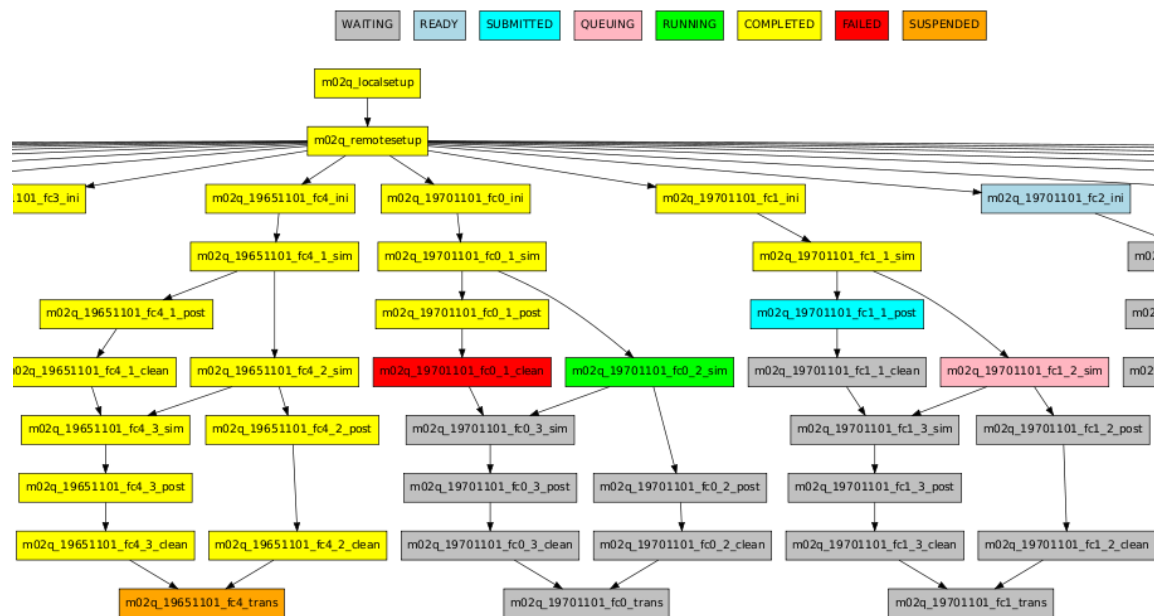
A workflow manager is required.

5,500 simulated years -> 1.4 PB
CMIP6 DCPP

EC-Earth3 at Lindgren, PDC

		5	10	10	20
Number of Start Dates		1	5	10	20
Number of Members		1	5	10	10
Number of Independent Simulations		1	25	50	200
T159-ORCA1	Cores	144	3600	7200	28800
	Wall-clock Time (Hours) / year	5	5	5	5
	CPU Time (Hours) / year	1800	18000	36000	144000
	Output Size (GB) / year	10,80	480	960	3840
T255-ORCA1	Cores	360	9000	18000	72000
	Wall-clock Time (Hours) / year	5	5	5	5
	CPU Time (Hours) / year	1800	45000	90000	360000
	Output Size (GB) / year	19,20	5184	10368	41472
T799-ORCA025	Cores	1104	27600	55200	220800
	Wall-clock Time (Hours) / year	40	40	40	40
	CPU Time (Hours) / year	44160	1104000	2208000	8832000
	Output Size (GB) / year	256,80	6420	12840	51360

- Automatisation: Preparing and running, postprocessing and output transfer, all managed by Autosubmit. No user intervention needed.
- Provenance: Assigns unique identifiers to each experiment and stores information about model version, configuration options, etc
- Failure tolerance: Automatic retrials and ability to repeat tasks in case of corrupted or missing data.
- Versatility: Currently runs EC-Earth, NEMO and NMMB models on several platforms.



Workflow of an experiment monitored with Autosubmit (yellow = completed, green = running, red = failed, ...)

Predicting extremes



JJA near-surface temperature correlation of the ensemble mean from experiments with a climatological (top) and difference with one with realistic (bottom) land-surface initialisation. Results for EC-Earth2.3 started in May over 1979-2010.

Two ways for the analysis: reducing data traffic online or offline

a) q90 of Tx

b) nb of warm days

c) q90 of Tn

d) nb of warm nights

e) q10 of Tn

f) nb of cold nights

g) q90 of Tx

h) nb of warm days

i) q90 of Tn

j) nb of warm nights

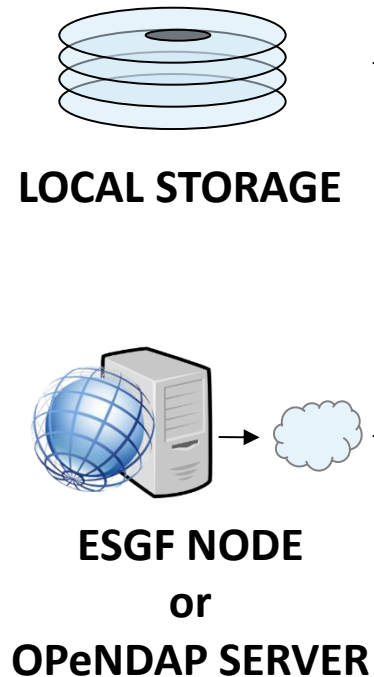
k) q10 of Tn

l) nb of cold nights



-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1

S2dverification is an R package to verify seasonal-to-decadal forecasts by comparing simulations with observational data. It allows analysing data available either locally or remotely. **It can also be used online as the model runs.**



- Supports datasets stored locally or in ESGF (OPeNDAP) servers.
- Exploits multi-core capabilities
- API available
- Collects observational and experimental datasets stored in multiple conventions:
 - NetCDF3, NetCDF4
 - Supports specific folder and file naming conventions.

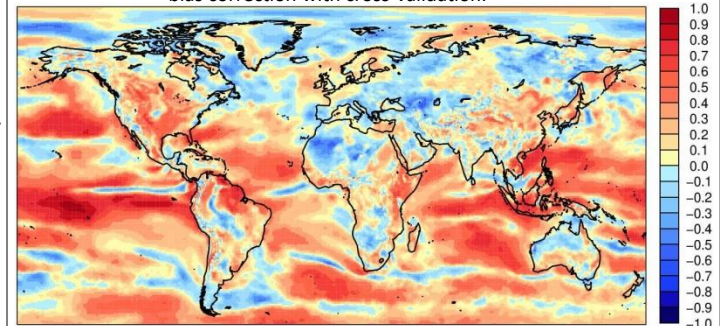
s2dverification package

BASIC STATISTICS

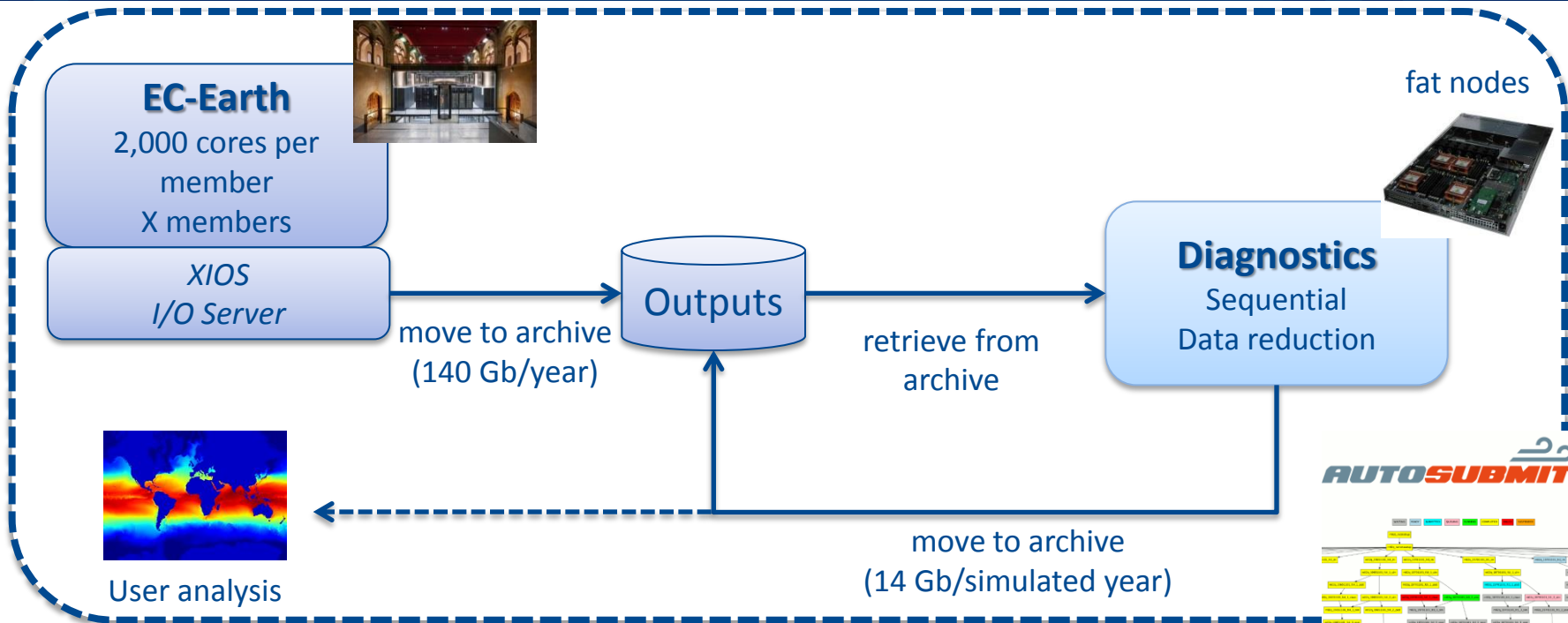
SCORES EXTREMES

PLOTS

Anomaly Correlation Coefficient. 10M Wind Speed ECMWF S4 1 month lead with start dates once a year on first of November and Era-Interim in DJF from 1981 to 2011. Simple bias correction with cross-validation.



Current workflow for diagnostics



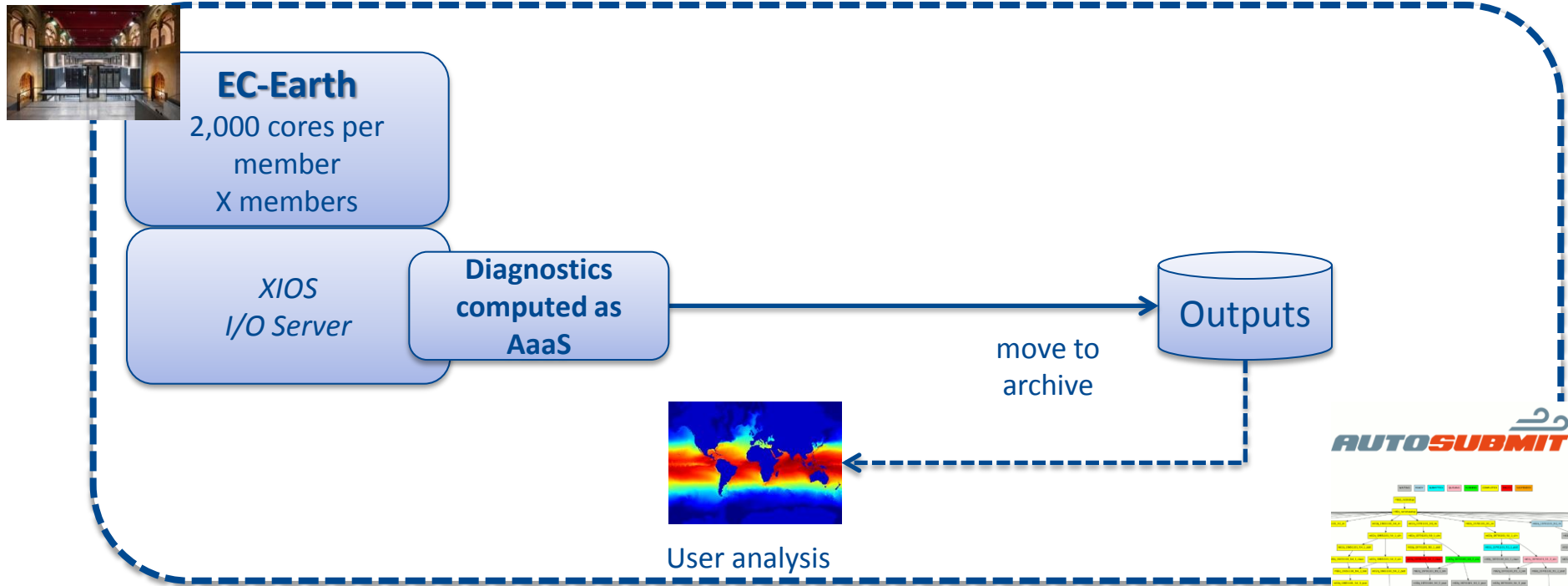
XIOS

- XIOS is an open source C++ I/O server widely used by the climate community
- XIOS is already integrated in NEMO and will be integrated in OpenIFS
- The diagnostics should be computed at the XIOS level
- Unfortunately, XIOS does not compute diagnostics yet

Drawbacks

- Diagnostics only computed offline (after model runs)
- High level of data traffic
- Fat nodes are required
- Delays on making significant data to the user

Proposed workflow for diagnostics



XIOS could be modified to add a layer of Analytics as a Service (based in PyCOMPSs/COMPSs)

- Diagnostics online (during model run)
- Reduced data traffic
- Diagnostics possible on the computing nodes
- New diagnostics (data mining of extremes) possible
- The user gets the results faster

- XIOS will require a substantial modification to run efficiently with high-resolution configurations (computer science expertise will be needed)
- Diagnostics need to be implemented from scratch, with parallelization of diagnostics in a multi-node XIOS configuration
- Coordinated effort with other institutions using the same applications, or dealing with similar problems, e.g.:
 - RIKEN: research to reduce the time to compute diagnostics in high-resolution simulation (200 days to post-process 960 m resolution simulations)
 - ARGONNE: confronted with the same problems in the development of the ACME climate model for DoE



Case 1: Data streaming for air quality forecasting

Case 2: Simultaneous analytics and HPC in climate prediction

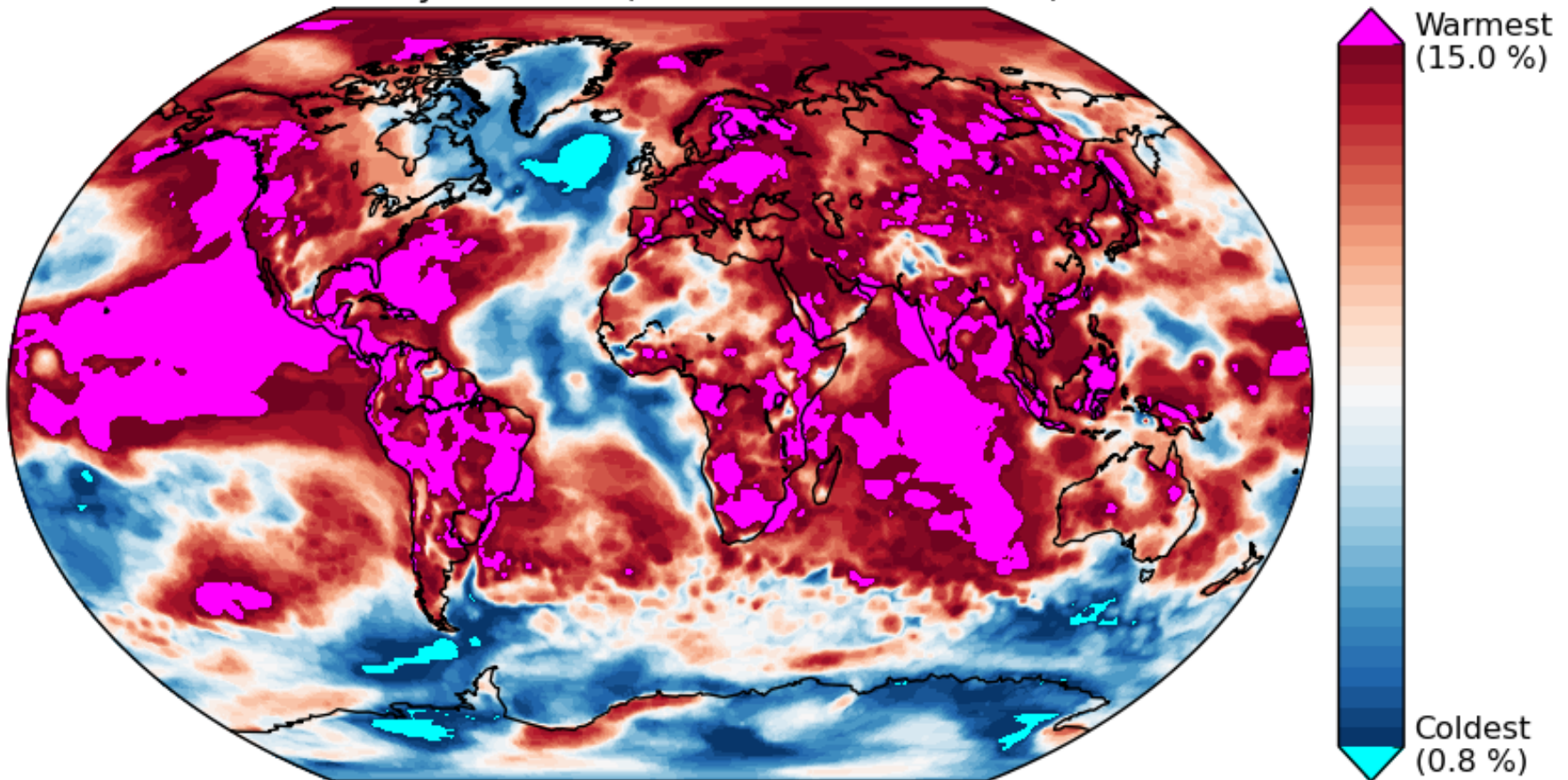
Case 3: Analytics as a service

Climate change is taking place



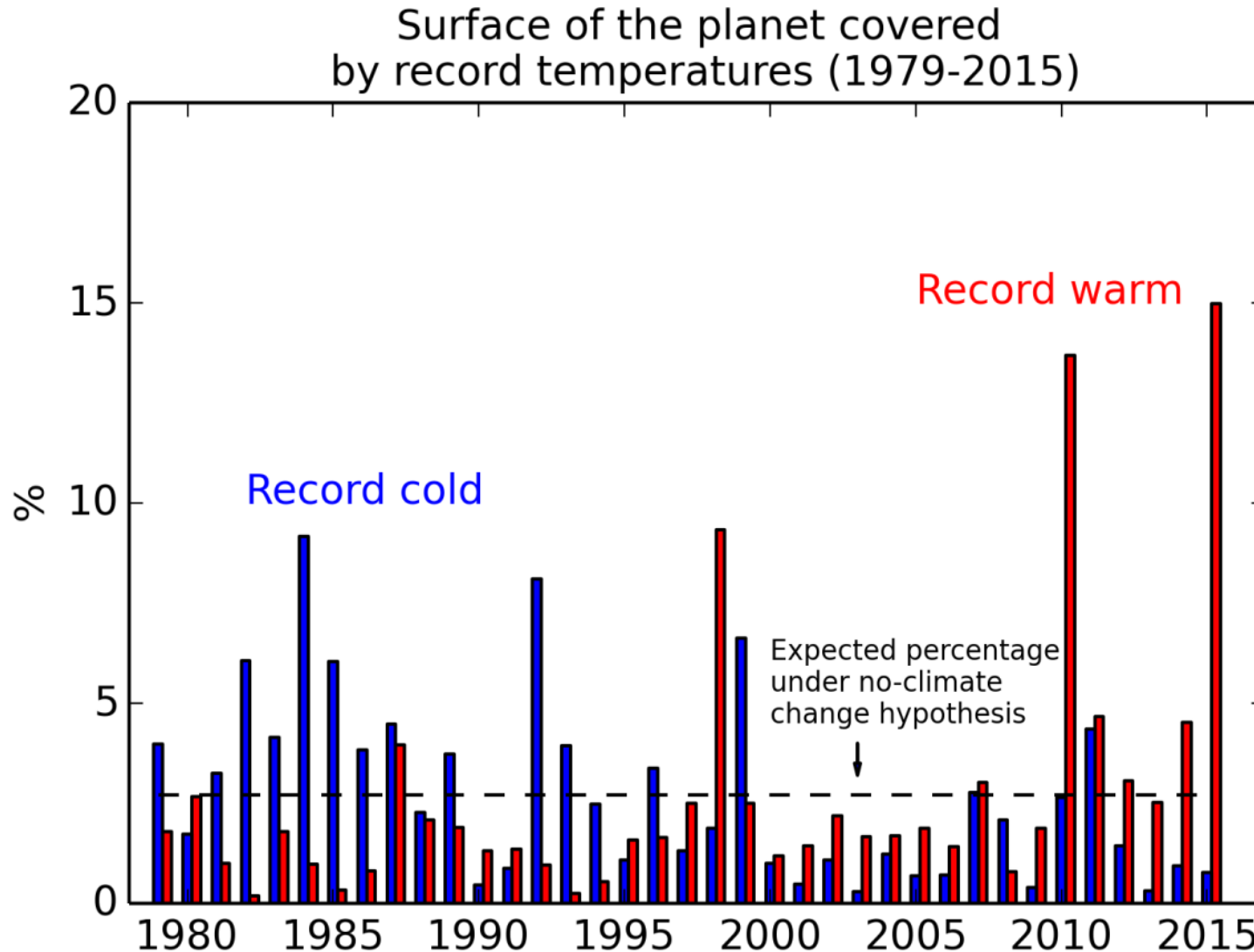
Rank of the 2015 annual mean temperature over the last 37 years from ERA Interim.

Annual mean 2m temperature
Rank of year 2015 (reference: 1979-2015)

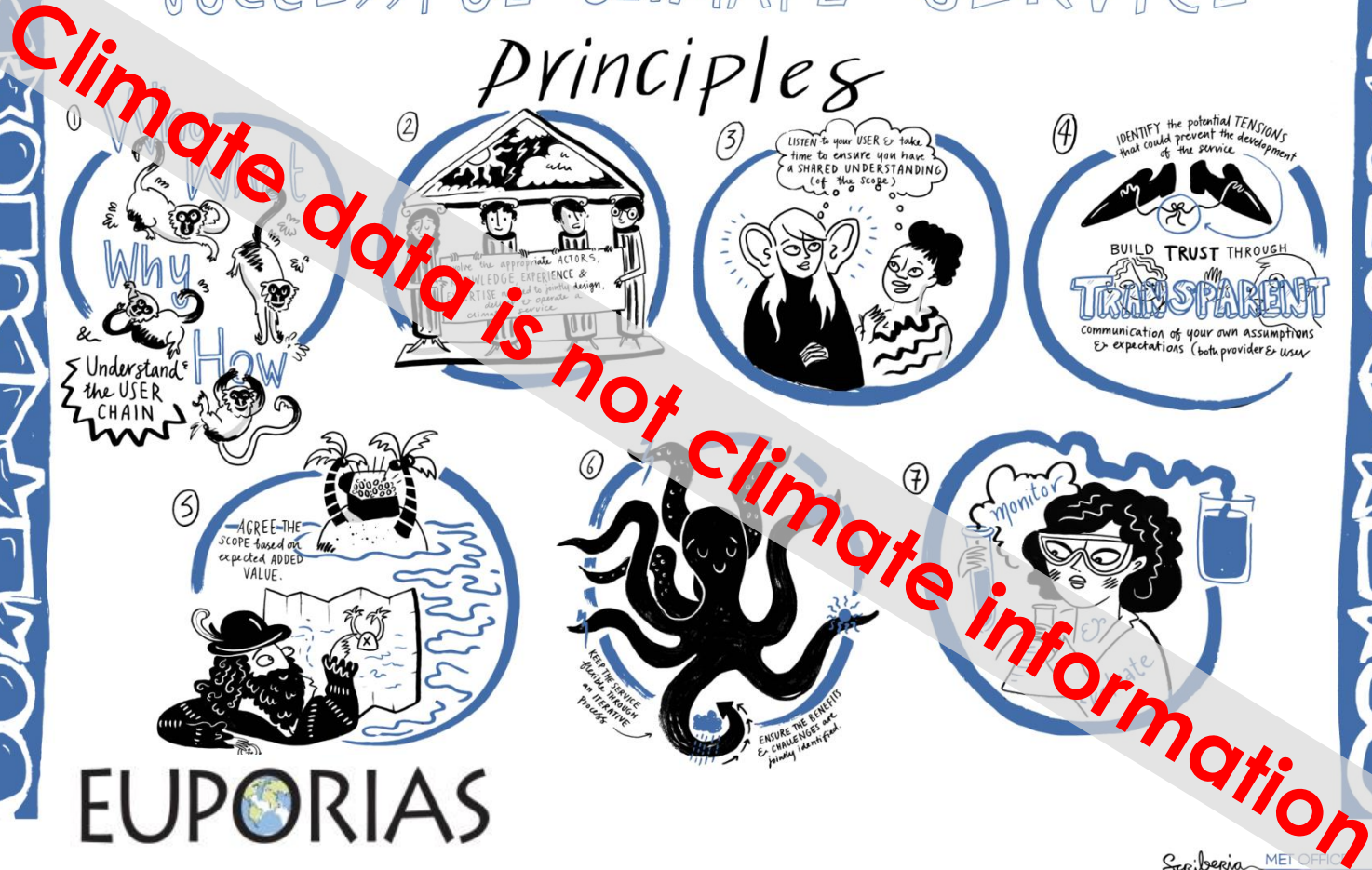


Data: ERA-Interim. Figure: F. Massonnet - BSC

Climate change is taking place

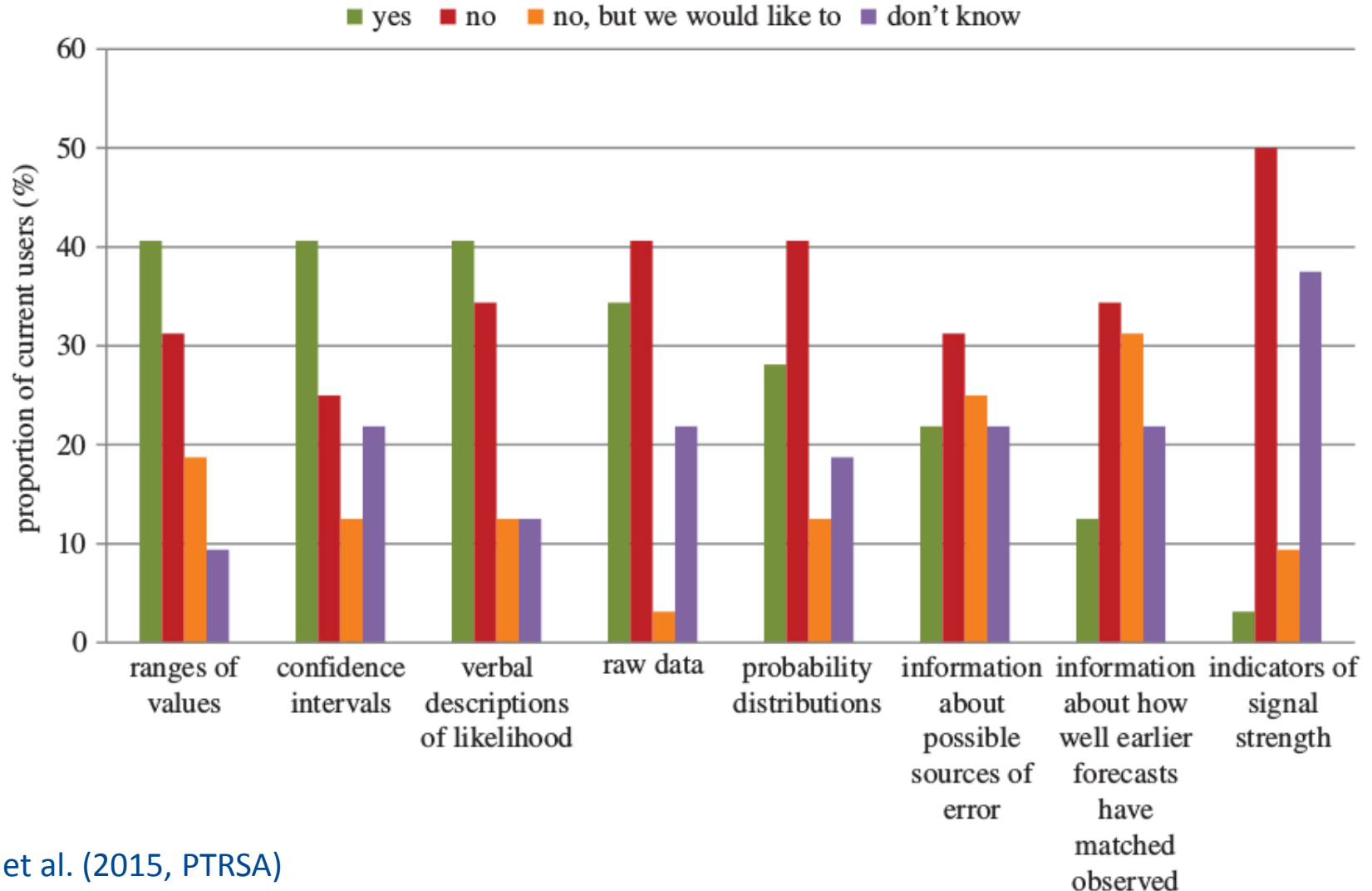


SUCCESSFUL CLIMATE SERVICE Principles



Ethical Framework for Climate Services four core elements: integrity, transparency, humility and collaboration.

Proportion of users of seasonal forecasts (n = 32) indicating whether they received different forms of information about uncertainty.

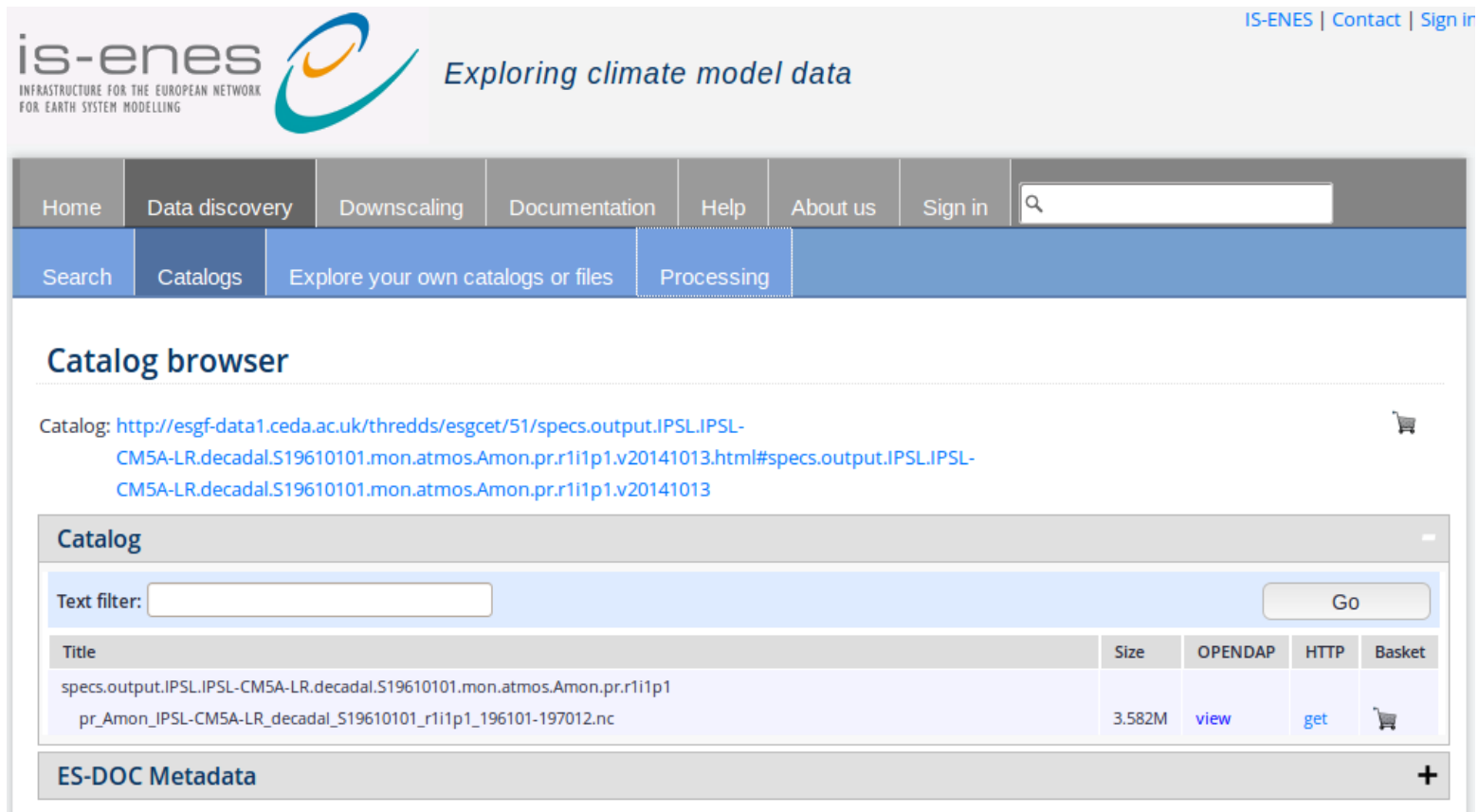


Users need services to respond to their questions.

An example of downstream service <http://climate4impact.eu>.

This is a key aspect for the success of the community's research.

Based on access to **ESGF**.



The screenshot shows the 'is-enes' catalog browser interface. The header includes the logo and tagline 'Exploring climate model data'. A navigation bar contains links for Home, Data discovery, Downscaling, Documentation, Help, About us, and Sign in, along with a search input field. Below this is a secondary navigation bar with buttons for Search, Catalogs, Explore your own catalogs or files, and Processing. The main content area is titled 'Catalog browser' and displays a search result for a specific climate model dataset. The result is shown in a table with columns for Title, Size, OPENDAP, HTTP, and Basket. Below the table is a section for 'ES-DOC Metadata'.

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK
FOR EARTH SYSTEM MODELLING

Exploring climate model data


IS-ENES | Contact | Sign in


Home Data discovery Downscaling Documentation Help About us Sign in

Search Catalogs Explore your own catalogs or files Processing

Catalog browser

Catalog: <http://esgf-data1.ceda.ac.uk/thredds/esgcat/51/specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1.v20141013.html#specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1.v20141013>

Title	Size	OPENDAP	HTTP	Basket
specs.output.IPSL.IPSL-CM5A-LR.decadal.S19610101.mon.atmos.Amon.pr.r1i1p1 pr_Amon_IPSL-CM5A-LR_decadal_S19610101_r1i1p1_196101-197012.nc	3.582M	view	get	

ES-DOC Metadata 

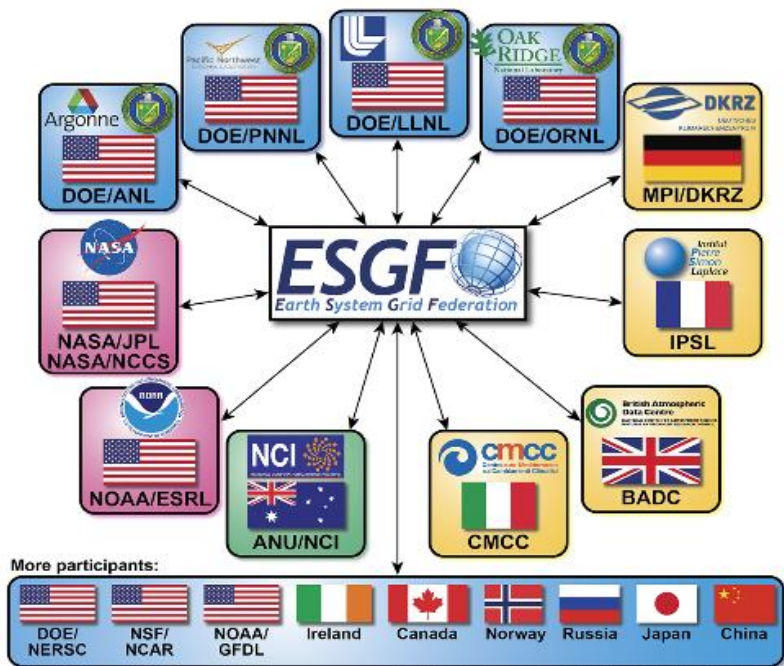
The Earth System Grid Federation is an open source effort providing a robust, distributed data platform, enabling worldwide access to peta/exascale weather and climate data.

ESGF promotes common formatting, has discovery tools and data indexing.

EUDAT, a pan-European network, is working on general-purpose additional solutions.

ESGF is organized as a framework of worldwide **distributed nodes**

- **Data nodes** (THREDDS server, gridFTP server, ...)
- **Index nodes** (database description)
- **Identity nodes** (authentication)
- **Compute nodes** (data analysis and visualization)



*Additional participants could not be illustrated in this figure.

Need to compute where the data are



ESGF distributes mainly CMIP (Coupled Model Intercomparison Project) data. CMIP is one of the bases of the assessment reports of the Intergovernmental Panel on Climate Change (IPCC).

	CMIP1 (1996)	CMIP2 (1997)	CMIP3 (2005)	CMIP5 (2012)
Number of experiments	1	2	12	110
Institutions participating	16	18	16	24
Number of models	19	24	21	5
Total dataset size	1 GB	540 GB	36 TB	3.3 PB

Derived variables require either huge data traffic or large computing power and memory



100 times increase from one CMIP and the next one

- Spanish Network for weather, climate and air quality data.
 - Initiative started with a workshop in Madrid the 13th of November 2015.
 - Foster the discussion about solutions to issues linked to Big Data in our sectors among the research, operational and industrial Spanish community
 - <http://www.bsc.es/projects/earthscience/bigdata>
- Research Data Alliance (international initiative with the objective to promote data sharing among all kinds of research communities) **Interest Group in weather, climate and air quality** to address a strong **pressure from users**
 - Workshop held in Barcelona the 11th of February 2016
 - earthsciences-rda-ig@bsc.es, <http://rd-alliance.org>

- **Education:** in the era of open data, the community should take advantage of the open education opportunities.
- **Heterogeneity:** how to link to and merge our data with those from communities with larger impact (urban development, arts, social responsibility)?
- **Technology:** how can we make the most of a rapidly evolving technology (heterogeneous nodes, data software, mobile data capture, storage/compression, computing and storage outsourcing)?
- **Awareness:** what are the priority issues for our communities and how can we work better together?
- **Industry engagement:** how can we solve the problem of involving the private sector?