Science and
Technology
Facilities Council

Natural
Environment
Research Council

Earth System Grid Federation Future Architecture, Copernicus, Cloud and ESA

ClimateData.ca Meeting, 23 November 2021

Philip Kershaw, Technical Manager
Centre for Environmental Data Analysis

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

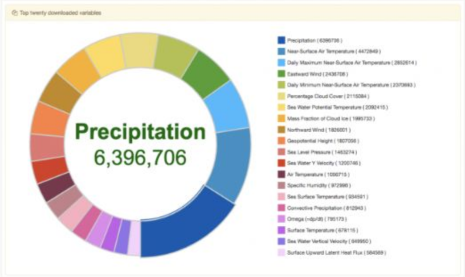
Earth System Grid Federation: a globally distributed data archive for climate data

31 Projects	170 Countries	24 Data nodes	12,870,551 published datasets	722,400,940 files downloaded	25.24 PB downloaded

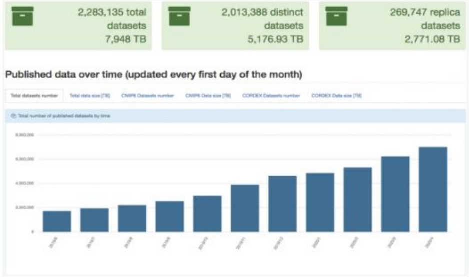
ESGF Dashboard: http://esgf-ui.cmcc.it

ESGF Federation

Data usage


Top twenty downloaded variables
Precipitation
6,396,706
Precipitation (6396706)
Near-Surface Air Temperature (4472849)
Daily Maximum Near-Surface Air Temperature (2852614)
Eastward Wind (2436708)
Daily Minimum Near-Surface Air Temperature (2370693)
Percentage Cloud Cover (2115084)
Sea Water Potential Temperature (2092415)
Mass Fraction of Cloud Ice (1995733)
Northward Wind (1826001)
Geopotential Height (1807056)
Sea Level Pressure (1463274)
Sea Water Y Velocity (1200746)
Air Temperature (1050715)
Specific Humidity (972996)
Sea Surface Temperature (934591)
Convective Precipitation (812943)
Omega (=dp/dt) (795173)
Surface Temperature (678115)
Sea Water Vertical Velocity (649950)
Surface Upward Latent Heat Flux (584589)

Data publication



2,283,135 total datasets 7,948 TB

2,013,388 distinct datasets 5,176.93 TB

269,747 replica datasets 2,771.08 TB

Published data over time (updated every first day of the month)

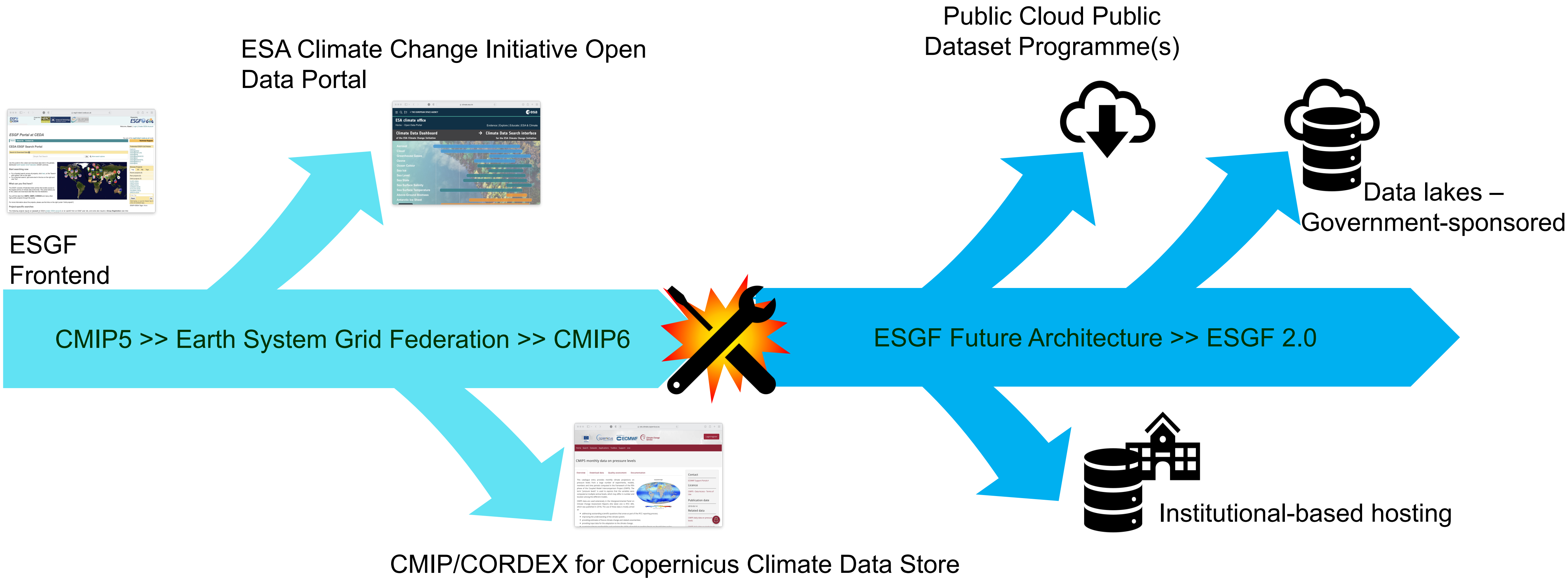
UKRI Science and Technology Facilities Council Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

ESGF – Application and Evolution



ESA Climate Change Initiative Open Data Portal

Public Cloud Public Dataset Programme(s)

ESGF Frontend

Data lakes – Government-sponsored

CMIP5 >> Earth System Grid Federation >> CMIP6

ESGF Future Architecture >> ESGF 2.0

Institutional-based hosting

CMIP/CORDEX for Copernicus Climate Data Store

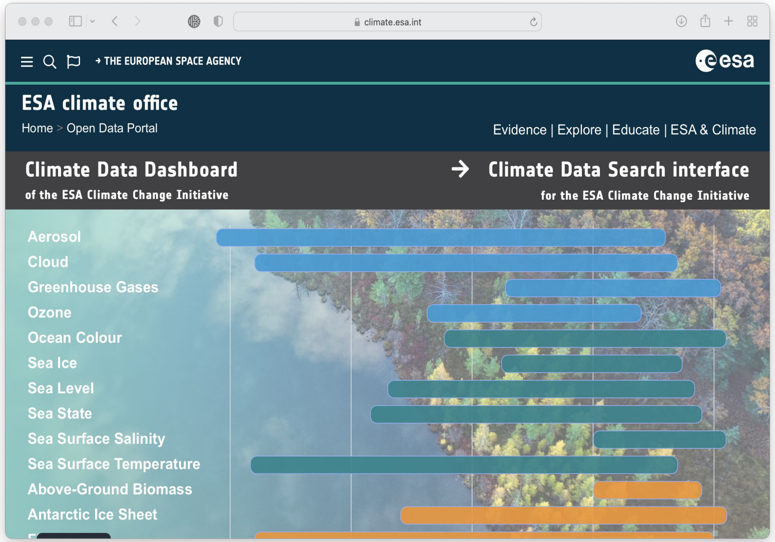
UKRI Science and Technology Facilities Council Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

ESA Climate Change Initiative Open Data Portal



2 Phases:

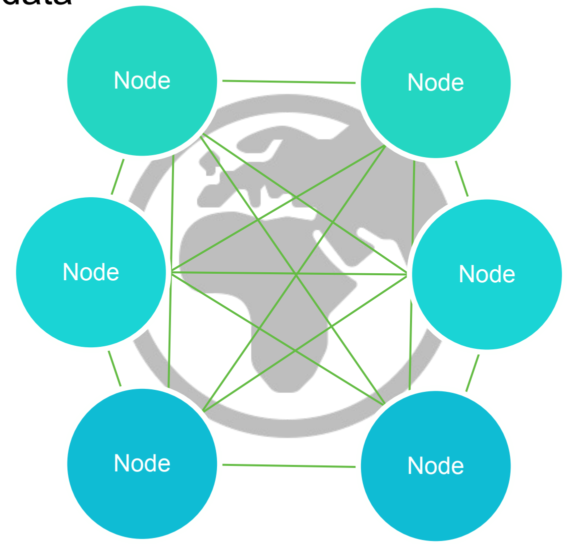
1) Leveraged ESGF
   1) Quick win with search and download
   2) Bespoke search API incompatible with other community standards - OGC CSW
   3) THREDDS Data Server couldn't scale to our needs

2) Redesigned to address issues
   1) OpenSearch API replaced ESG Search
   2) Scalable data service with Kubernetes
   3) Zarr format cache of netCDF data on object store for performance

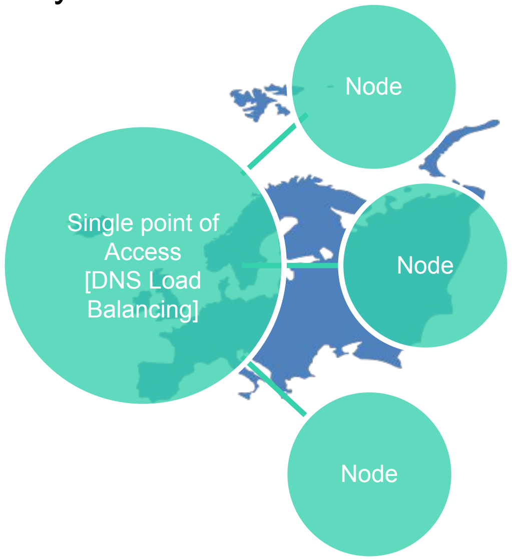
C3S 34[a-f] Projects for the CDS

- Architected a system for delivering resilient CMIP5 and CORDEX data access for the CDS by creative application of federated architecture for Earth System Grid Federation

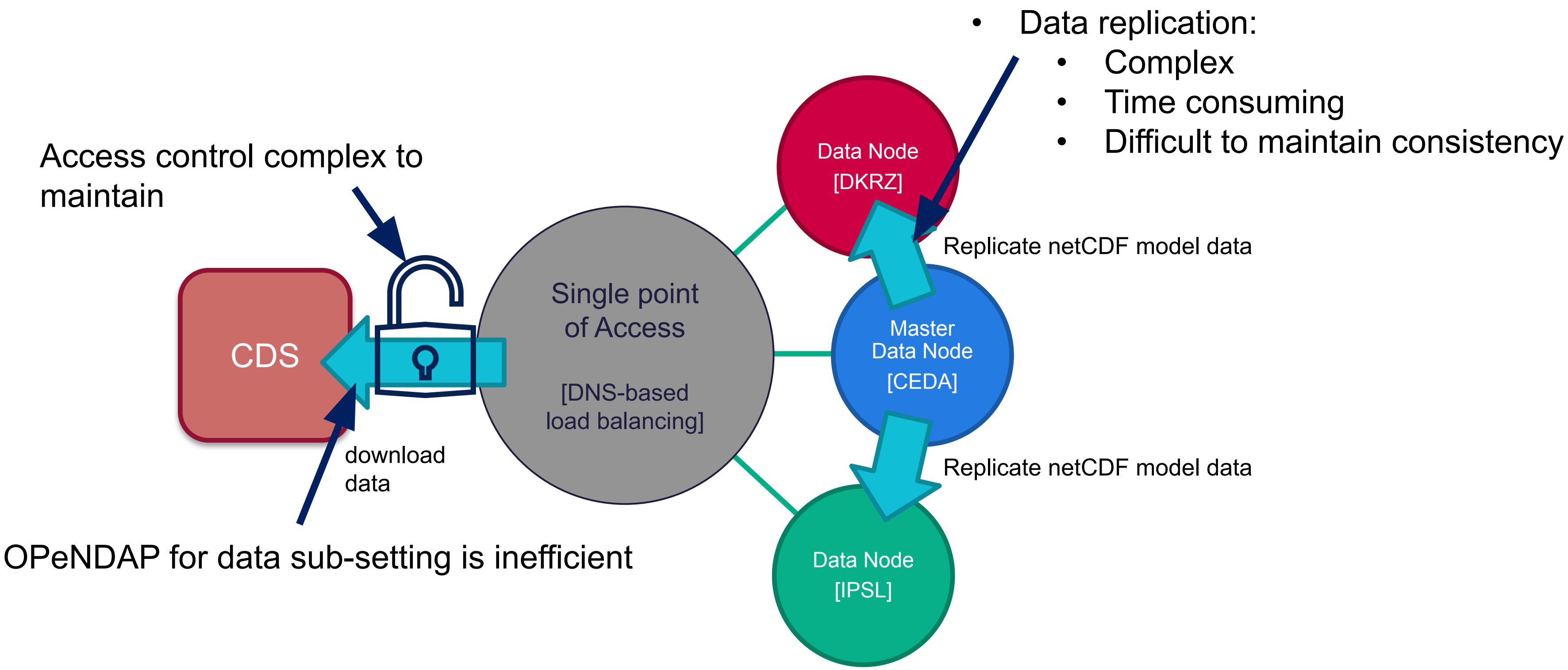
ESGF: an international federation of nodes providing a network of access points to model data



C3S 34a/b system: a single resilient point of access to data delivered through replication and redundancy



C3S Resilient CMIP and CORDEX Data Access



- Data replication:
  - Complex
  - Time consuming
  - Difficult to maintain consistency

Access control complex to maintain

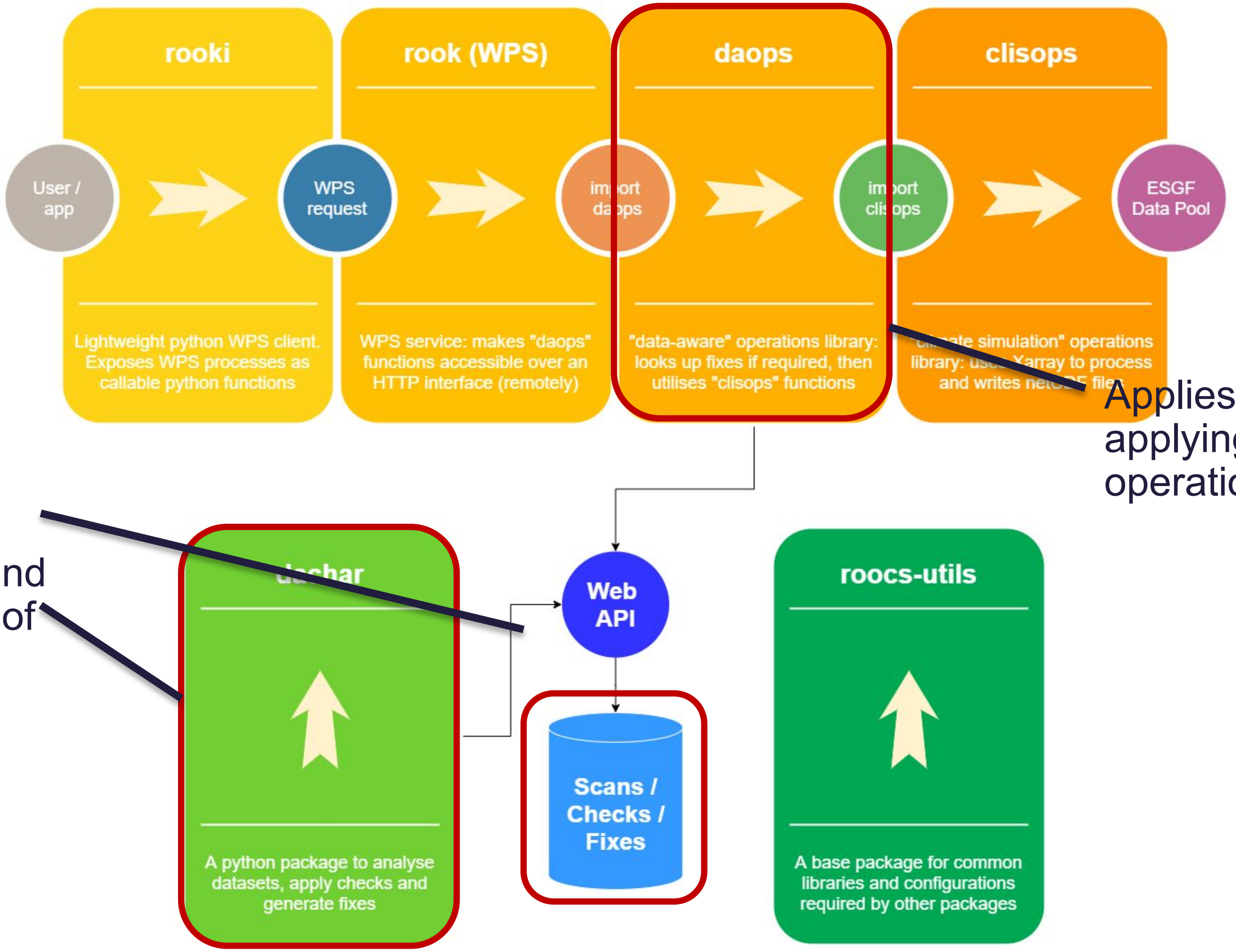
OPeNDAP for data sub-setting is inefficient

Sub-setting Services for C3S 34e Project

Climate Change Service

climate.copernicus.eu

C3S 34e Project


rooki
User / app
Lightweight python WPS client. Exposes WPS processes as callable python functions
rook (WPS)
WPS request
WPS service: makes "daops" functions accessible over an HTTP interface (remotely)
daops
import daops
"data-aware" operations library: looks up fixes if required, then utilises "clisops" functions
clisops
import clisops
"climate simulation" operations library: uses Xarray to process and writes netCDF files
ESGF Data Pool
dachar
A python package to analyse datasets, apply checks and generate fixes
Web API
Scans / Checks / Fixes
roocs-utils
A base package for common libraries and configurations required by other packages

Applies fixes to datasets before applying subsetting/regridding operations

Analyse datasets and make an inventory of fixes

Credit: Ag Stephens, CEDA

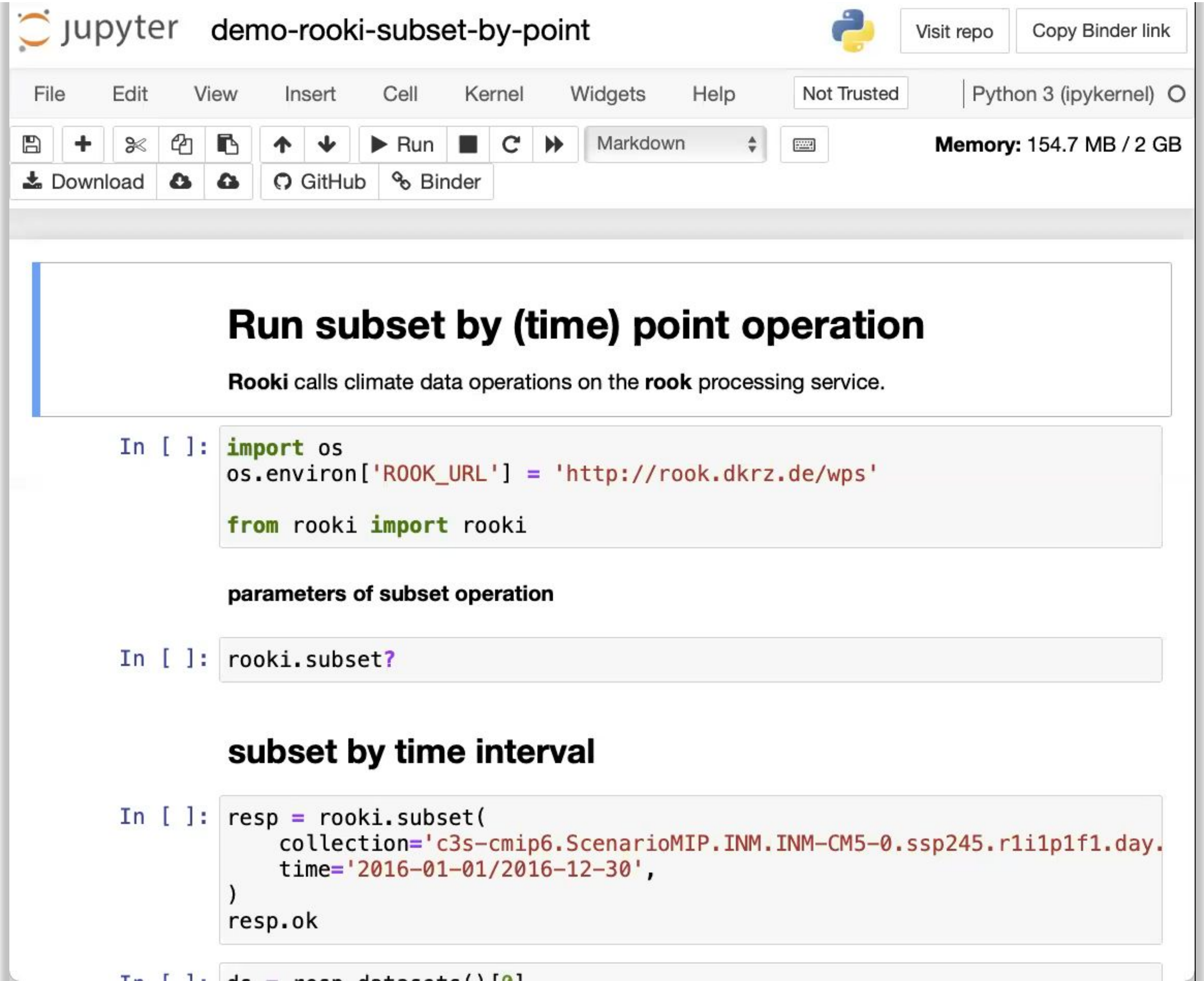
Science and Technology Facilities Council

Natural Environment Research Council

National Centre for Earth Observation

NATURAL ENVIRONMENT RESEARCH COUNCIL

https://nbviewer.org/github/roocs/rooki/blob/master/notebooks/demo/demo-rooki-subset-by-point.ipynb



Run subset by (time) point operation

Rooki calls climate data operations on the **rook** processing service.

```
In [ ]: import os
        os.environ['ROOK_URL'] = 'http://rook.dkrz.de/wps'

        from rooki import rooki
```

parameters of subset operation

```
In [ ]: rooki.subset?
```

subset by time interval

```
In [ ]: resp = rooki.subset(
            collection='c3s-cmip6.ScenarioMIP.INM.INM-CM5-0.ssp245.r1i1p1f1.day.
            time='2016-01-01/2016-12-30',
        )
        resp.ok
```

```
In [ ]: ds = resp.datasets()[0]
```

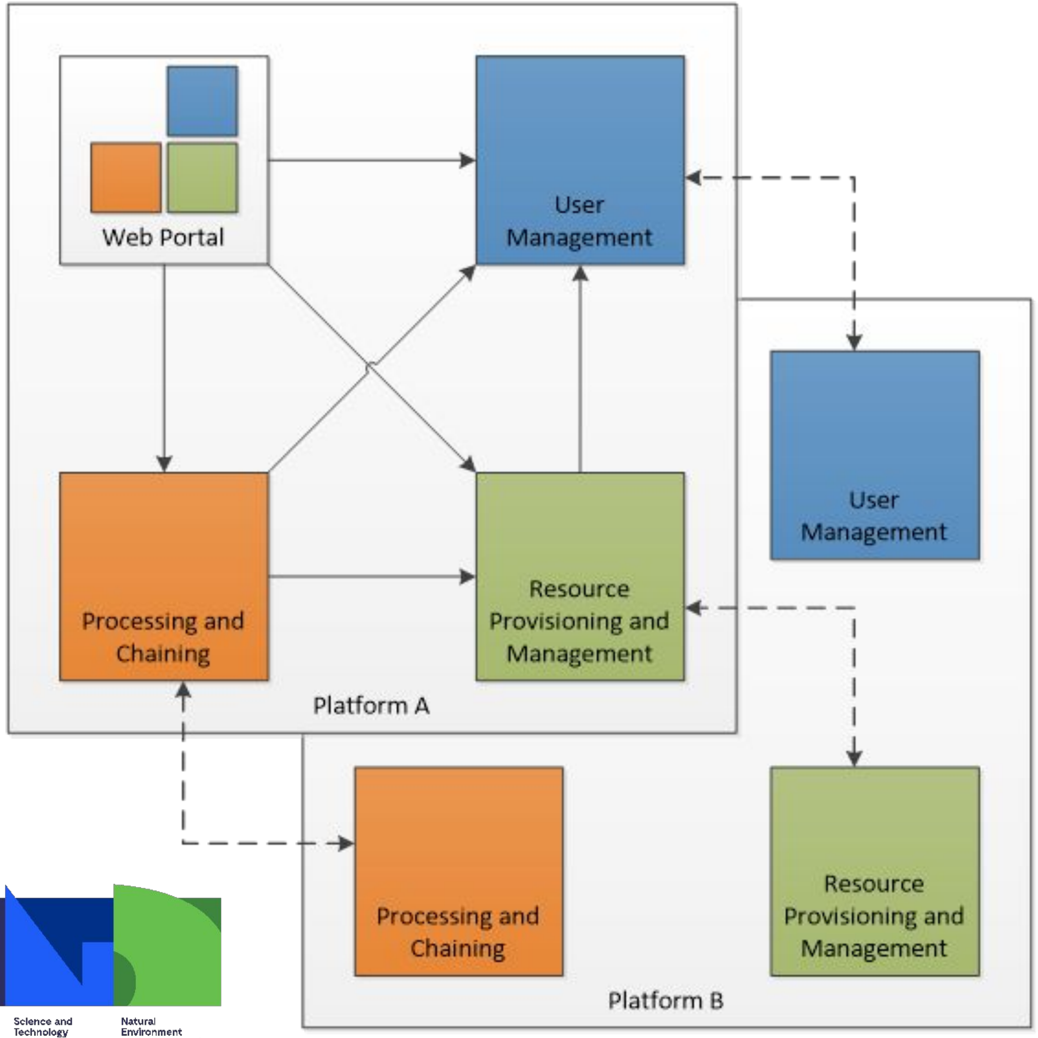
UKRI

Science and Technology Facilities Council

Natural Environment Research Council

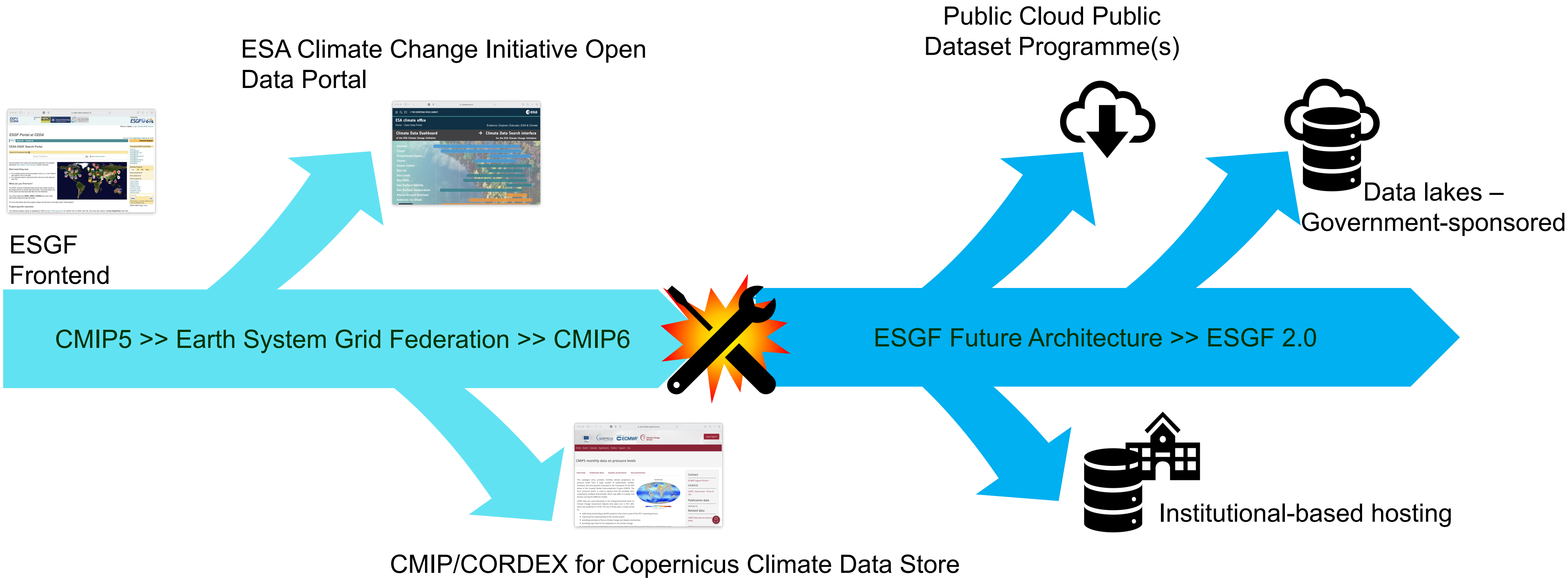
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

ESA Earth Observation Exploitation Platform Common Architecture (EOEPCA)



- Architectural blueprint for the federation of *platforms*
- Interlinked with the work on recent OGC testbeds
- CEDA involved with consultancy role
- Processing and chaining of particular interest
  - ADES and EMS
  - Ability to push customised shrink-wrapped processes to 3rd party WPS instances
- Innovations with ID management: UMA

ESGF – Application and Evolution



ESA Climate Change Initiative Open Data Portal

Public Cloud Public Dataset Programme(s)

ESGF Frontend

Data lakes – Government-sponsored

CMIP5 >> Earth System Grid Federation >> CMIP6

ESGF Future Architecture >> ESGF 2.0

Institutional-based hosting

CMIP/CORDEX for Copernicus Climate Data Store

UKRI Science and Technology Facilities Council Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

ESGF Future Architecture

Platforms and systems administration

Modular, scalable architecture: Containers, Kubernetes

Embrace infrastructure-as-code approach

Search services

Modernise, centralise and simplify

Use community standards: STAC

ID Management and Access Entitlement

Modernise, centralise and simplify

Use industry standards: OpenID Connect / OAuth 2.0

Progress and Achievements

- Container and Container+Kubernetes installs available
- **Deployed on AWS (GFDL) and at CEDA**

- Major community engagement on use of STAC for ESM data
- Prototype developed by CEDA
- Integration tests
- CoG and MetaGrid futures??

- OpenID Connect / OAuth 2.0 done
- New Authorisation system with Open Policy Agent
- Authentication integrated with C4I in test

ESGF Future Architecture

New modes for Data Access + Storage

Augment trad. file serving with object store

New models for aggregation and subsetting, retire OPeNDAP

Compute Services

Important but no consensus for ESGF-wide standard offering yet

Metrics Collection

Leverage advances in industry with standard tooling to exploit - Prometheus and InfluxDb, Grafana

Progress and Achievements

- Factored out TDS
- Test CMIP6 data caches on object store at CEDA and DKRZ

- Ported C3S WPS Data Reduction Services for use in ENES CDI
- Used with Climate4Impact
- Reboot of Compute Working Team

- New Metrics system integrated with CMCC

UKRI

Science and Technology Facilities Council

Natural Environment Research Council

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

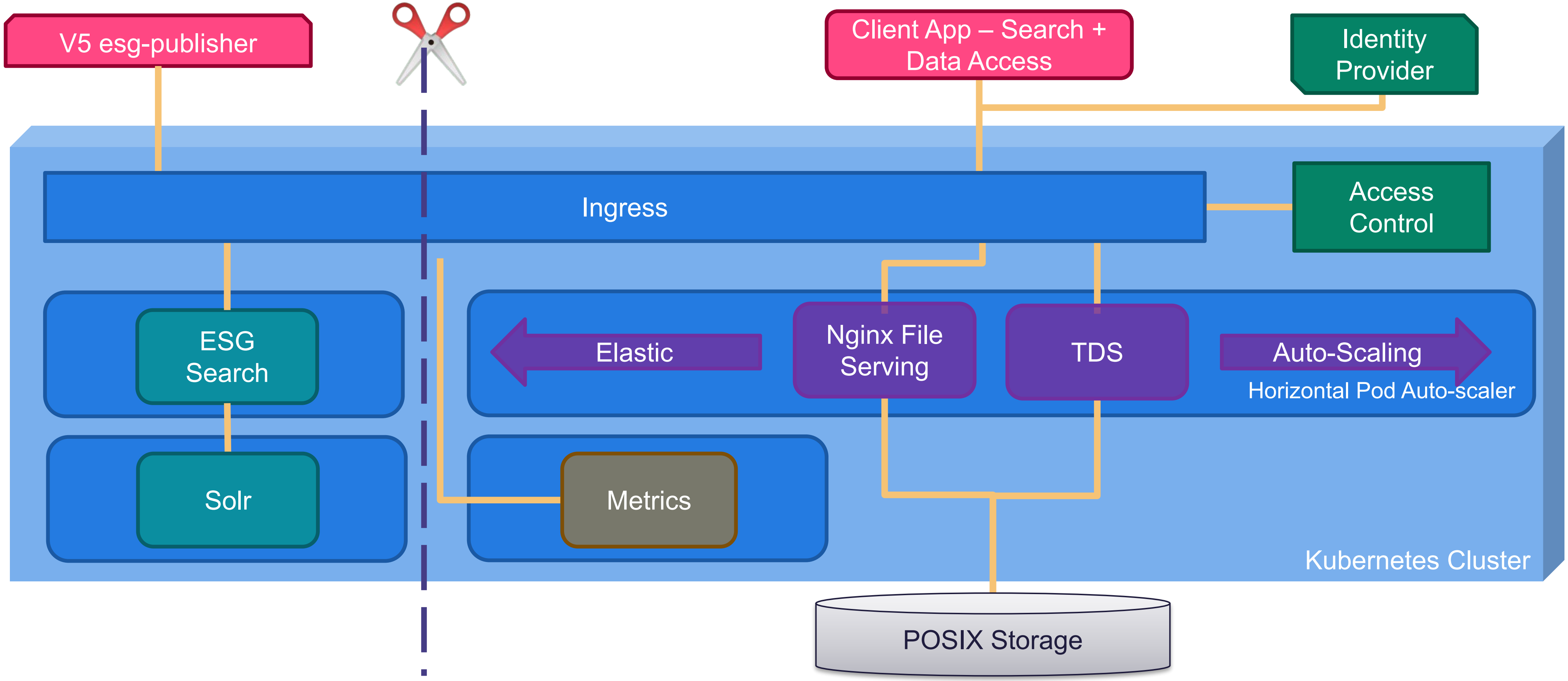
National Centre for Atmospheric Science

NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation

NATURAL ENVIRONMENT RESEARCH COUNCIL

Future Architecture Node – Phase 1



Science and Technology Facilities Council

Natural Environment Research Council

Centre for Environmental Data Analysis

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

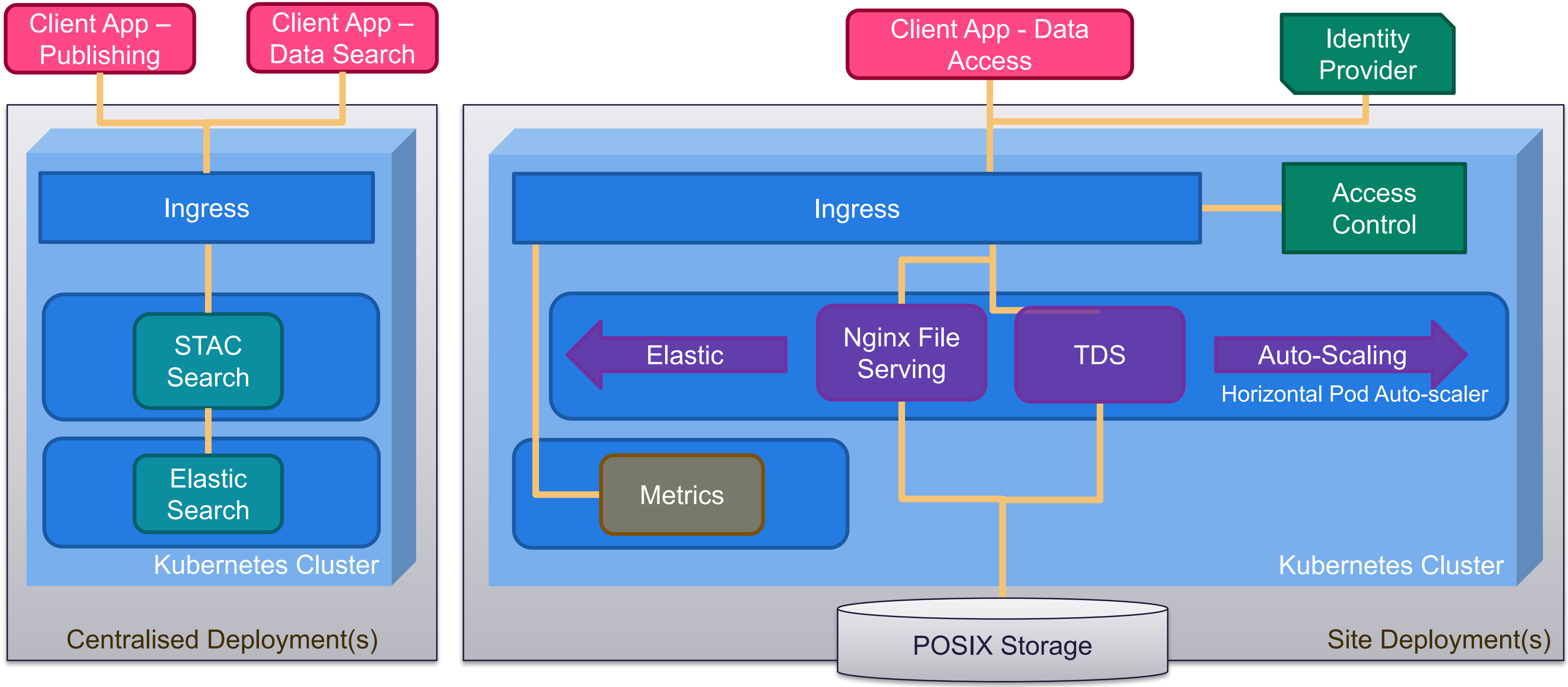
National Centre for Atmospheric Science

NATURAL ENVIRONMENT RESEARCH COUNCIL

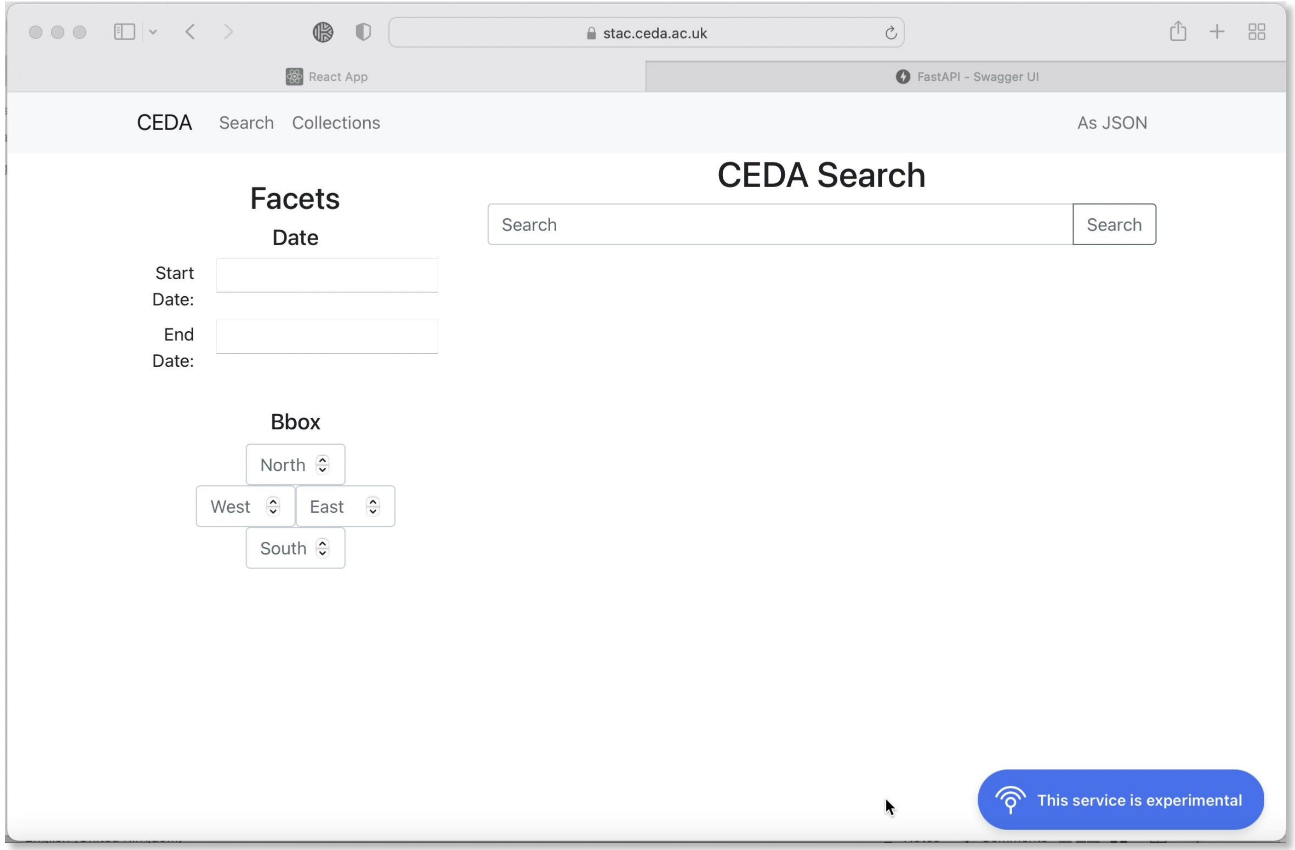
National Centre for Earth Observation

NATURAL ENVIRONMENT RESEARCH COUNCIL

Future Architecture Node – Phase 2



STAC API for ESGF



- Implementation of STAC API
- ElasticSearch backend
- Filter extensions to support faceted search
- Fully featured STAC equivalent API to ESG Search
- Simple frontend created to demonstrate its features

UKRI Science and Technology Facilities Council Natural Environment Research Council

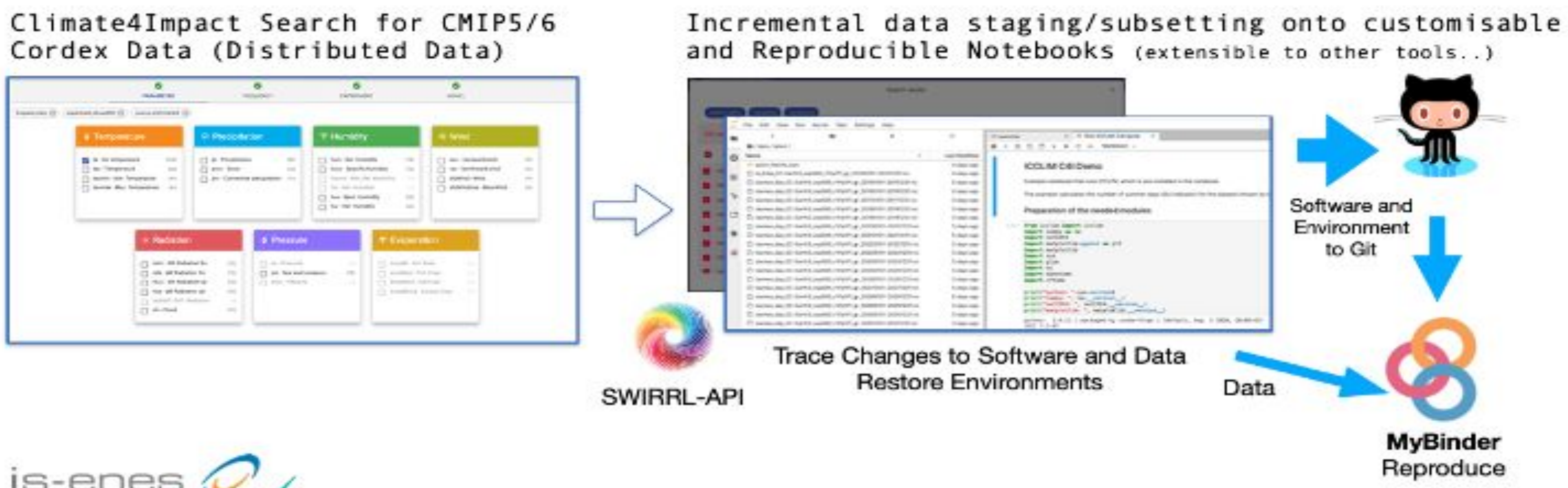
Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

IS-ENES3 - Data Analytics using Notebooks/icclim

Objective: Extend C4I with Data Driven & Reproducible Workspaces



Climate4Impact Search for CMIP5/6 Cordex Data (Distributed Data)

Incremental data staging/subsetting onto customisable and Reproducible Notebooks (extensible to other tools..)

Software and Environment to Git

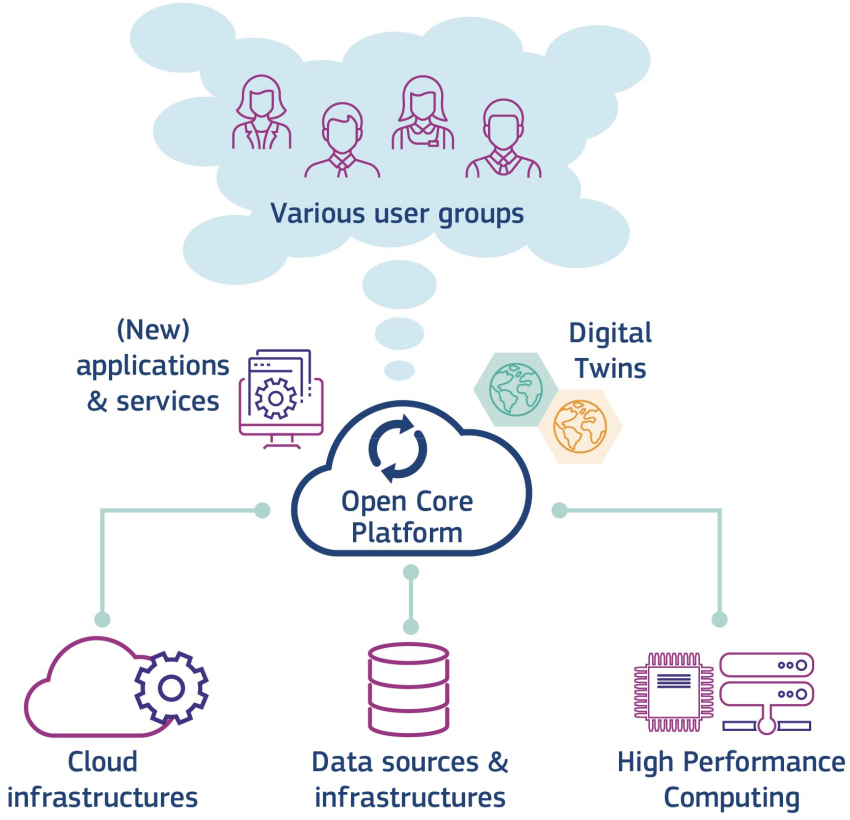
SWIRRL-API

Trace Changes to Software and Data
Restore Environments

Data

MyBinder
Reproduce

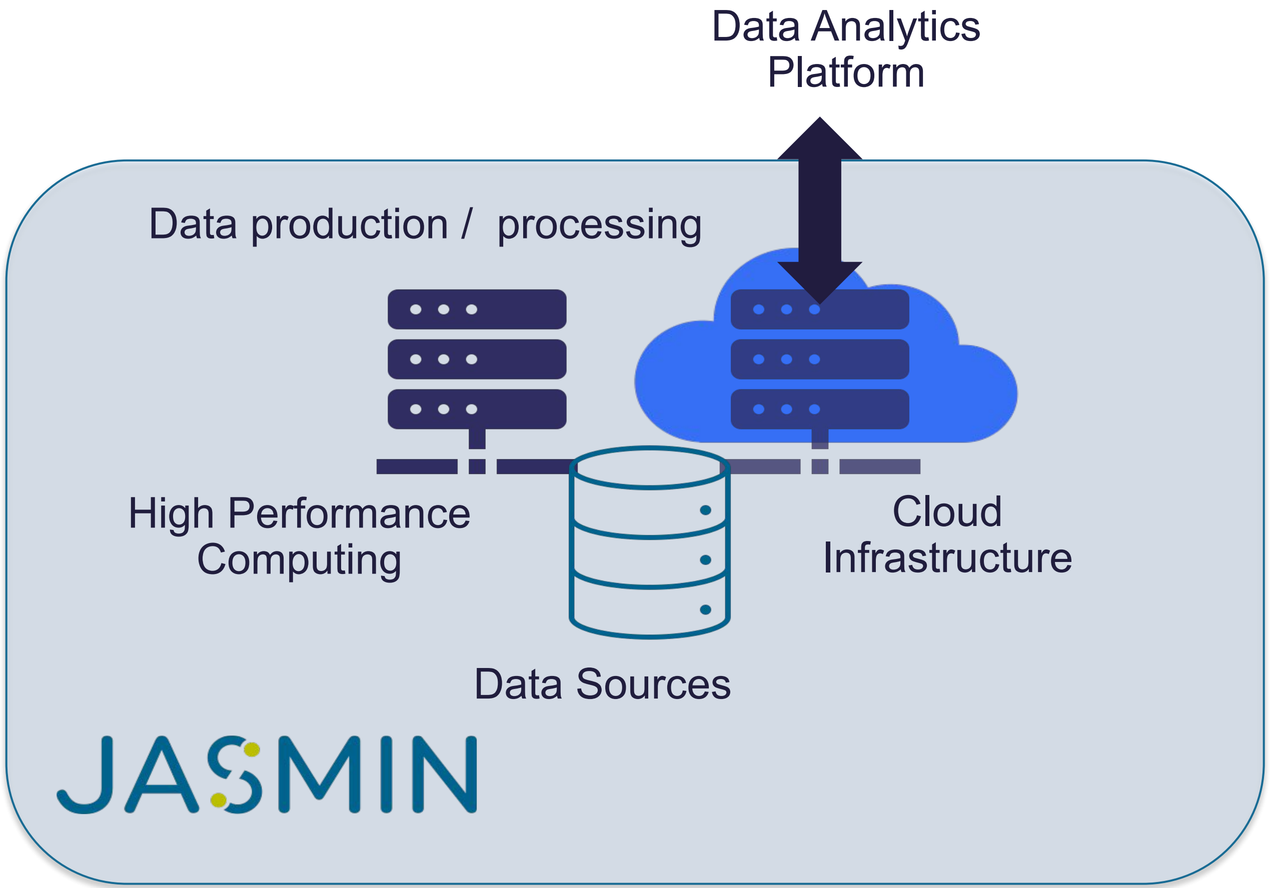
DestinE and Blueprint Architecture


Various user groups
(New) applications & services
Digital Twins
Open Core Platform
Cloud infrastructures
Data sources & infrastructures
High Performance Computing

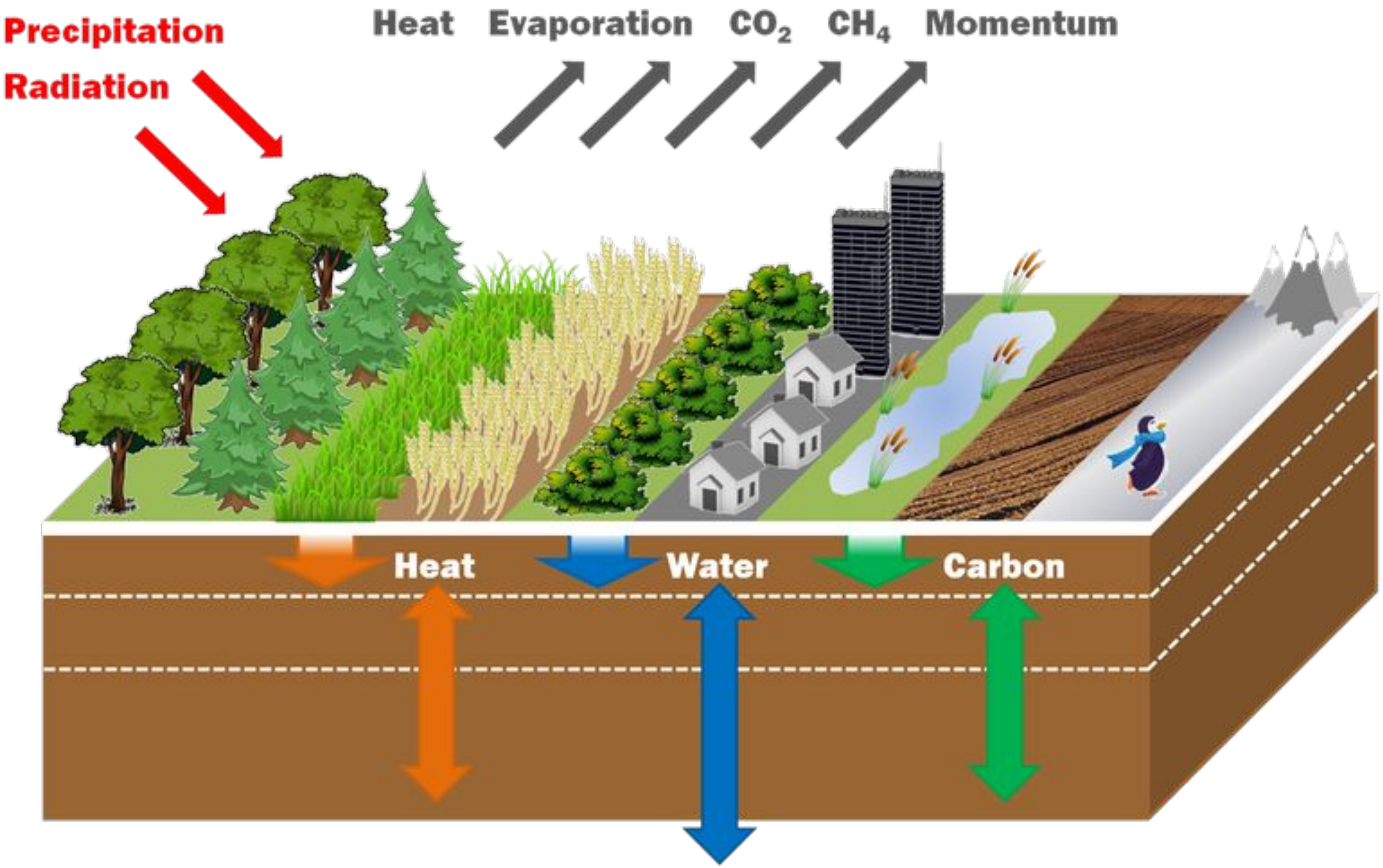
Destination Earth (DestinE) - major EU initiative:

- “to develop a very high precision digital model of the Earth (a ‘digital twin’) to monitor and predict environmental change and human impact to support sustainable development”

JASMIN

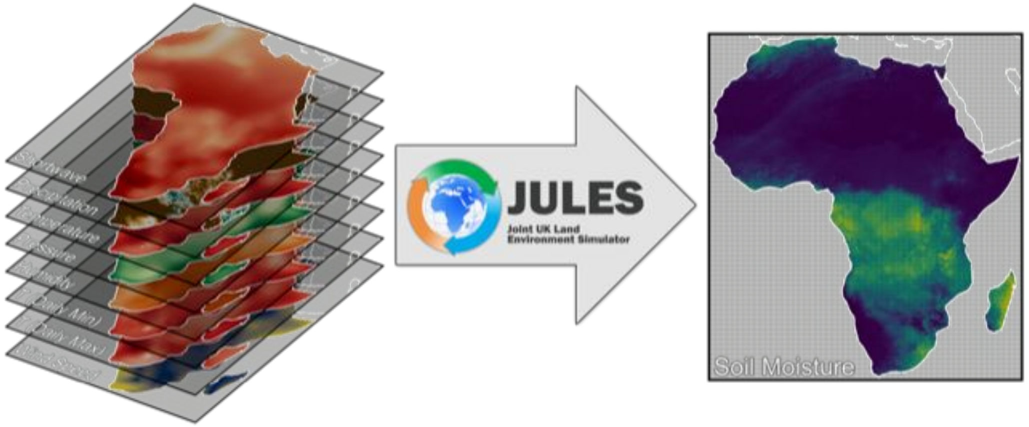

Data Analytics
Platform
Data production / processing
High Performance
Computing
Cloud
Infrastructure
Data Sources
JASMIN

ESA Digital Twin Earth (DestinE) Precursor - land surface modelling and climate



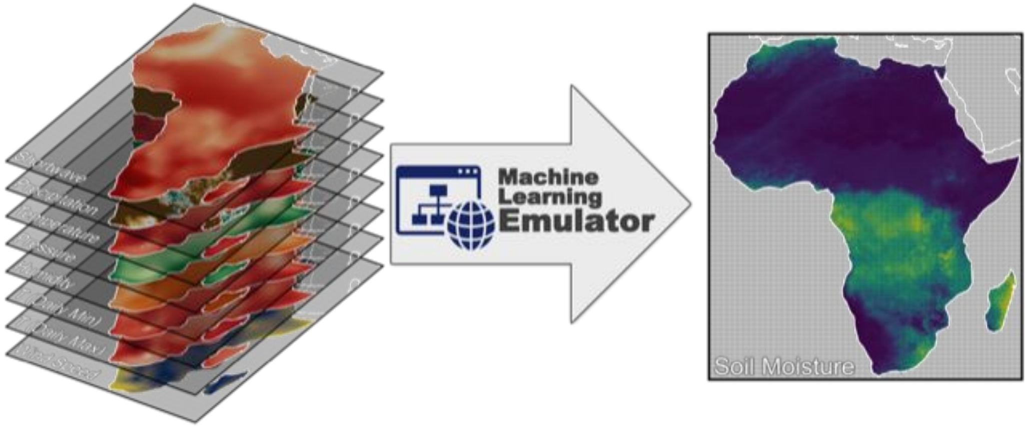
- Using JULES (Joint UK Land Environment Simulator)
  - the land surface component in the Met Office Unified Model
- Improvements with Data Assimilation
  - LaVEnDAR (The Land Variational Ensemble Data Assimilation fRamework)
  - Feed in satellite observations – SIF and SMAP data

What could be the future impact of climate change on the soil moisture?



- JULES driven with climate projections from ISIMIP data (Inter-Sectoral Impact Model Intercomparison Project)

Make a surrogate AI model to JULES

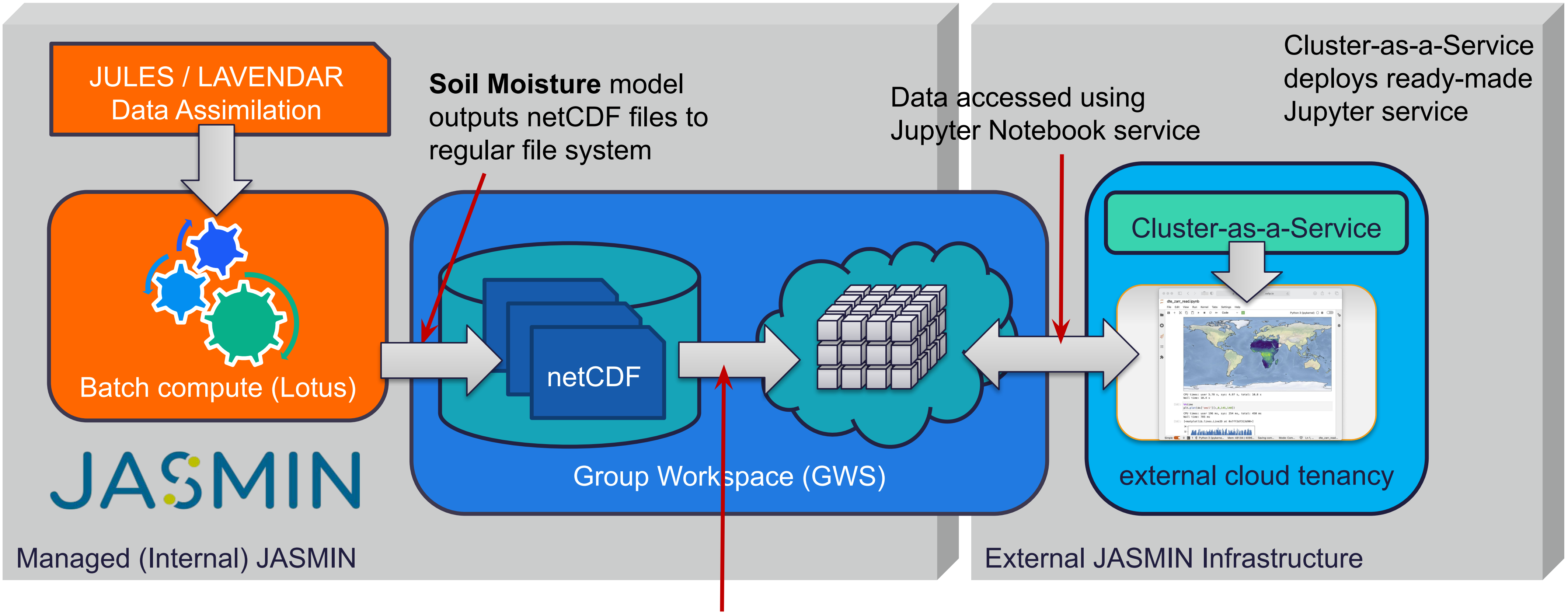


- Experimented with Machine Learning (ML) techniques
- Goal: a general-purpose algorithm -

```
time series of daily weather data □ time
series of soil moisture data
```

- Successfully applied XGBoost (eXtreme Gradient Boosting) algorithm.
- trained on up to 1000 grid cells, representative of the various biomes in continental Africa
- Demonstrated to accurately emulate JULES output at other locations
- The credibility of the model is enhanced by its transparency and explainability

Digital Twin Precursor on JASMIN: HPC for data production, cloud for analysis



Soil Moisture model outputs netCDF files to regular file system

Data accessed using Jupyter Notebook service

Cluster-as-a-Service deploys ready-made Jupyter service

Move data into object store so that it can be accessed by Jupyter Service on JASMIN cloud

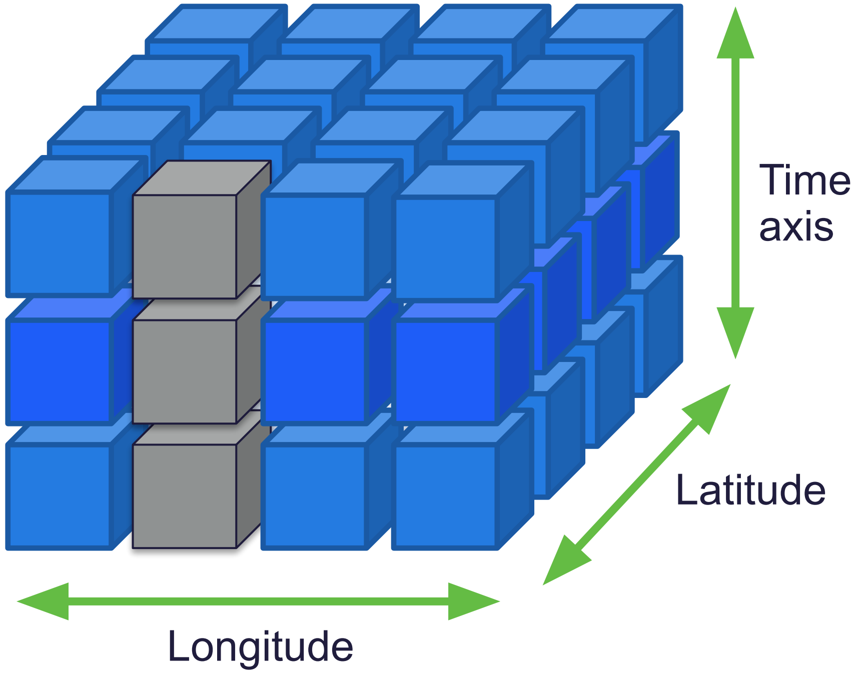
UKRI Science and Technology Facilities Council Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

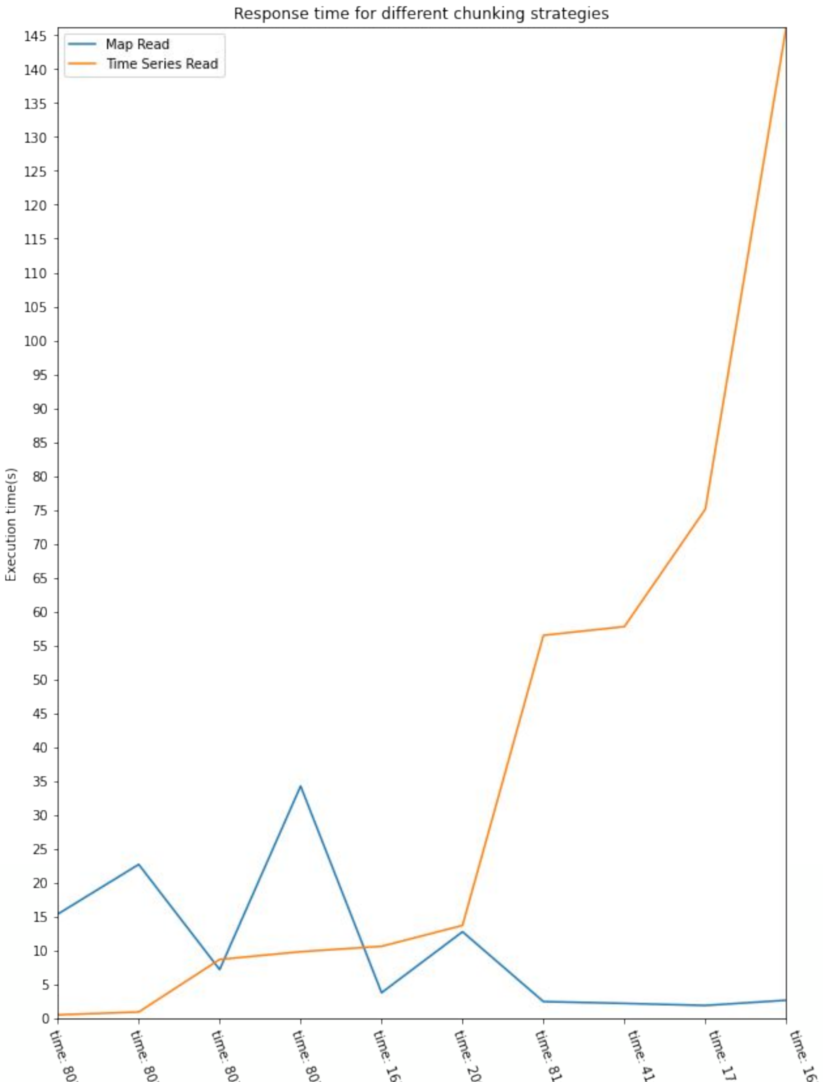
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Arrangement of data and efficient access



- Data output from models as netCDF format
- Data in files arranged in spatial dimensions one per time step
- But predominate access pattern for analysis of climate data in the project is time series query (grey blocks)

Object Store: Different storage strategies showed radically different performance

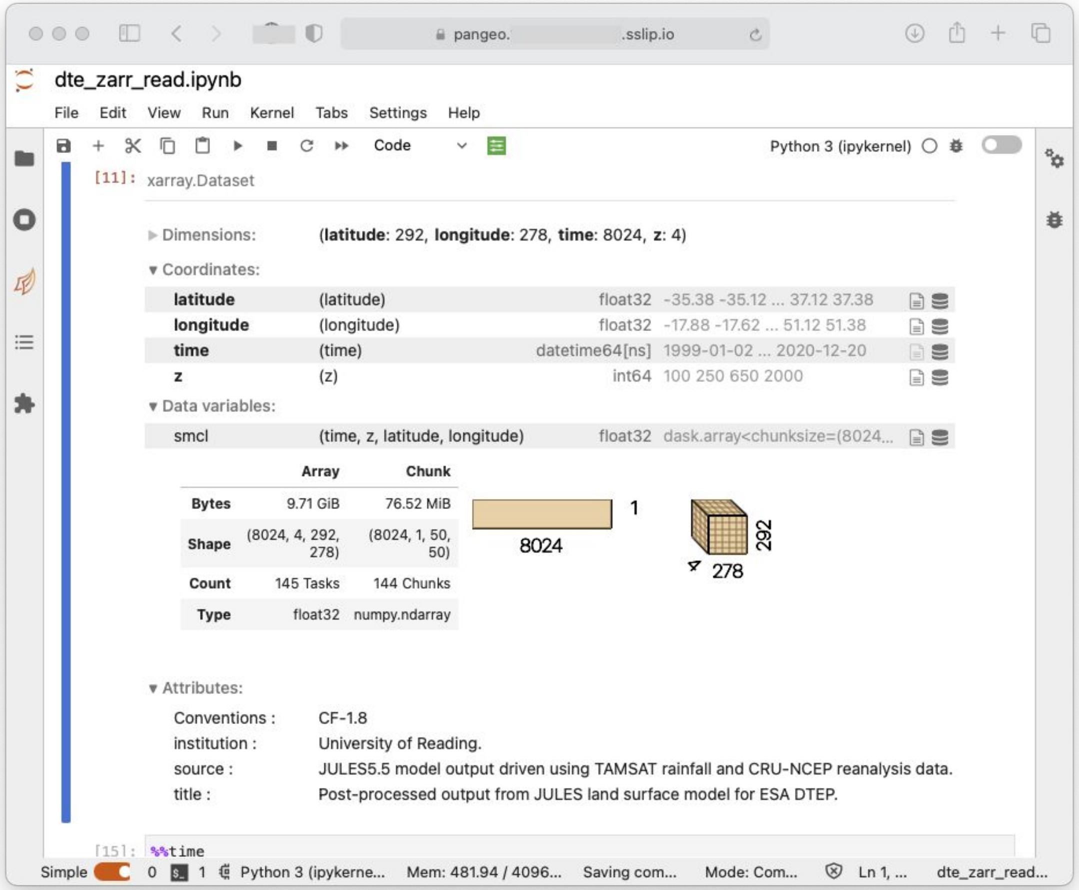


- We experimented with different storage chunking arrangements
- 20-year dataset of soil moisture

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Using Object Store for re-arrangement of data to suite our access patterns



```
[11]: xarray.Dataset

  ▶ Dimensions:       (latitude: 292, longitude: 278, time: 8024, z: 4)
  ▼ Coordinates:
     latitude         (latitude)                          float32  -35.38 -35.12 ... 37.12 37.38
     longitude        (longitude)                         float32  -17.88 -17.62 ... 51.12 51.38
     time             (time)                      datetime64[ns]  1999-01-02 ... 2020-12-20
     z                (z)                                   int64  100 250 650 2000
  ▼ Data variables:
     smcl             (time, z, latitude, longitude)      float32  dask.array<chunksize=(8024...
```

	Array	Chunk
Bytes	9.71 GiB	76.52 MiB
Shape	(8024, 4, 292, 278)	(8024, 1, 50, 50)
Count	145 Tasks	144 Chunks
Type	float32	numpy.ndarray

```
  ▼ Attributes:
     Conventions :    CF-1.8
     institution :    University of Reading.
     source :         JULES5.5 model output driven using TAMSAT rainfall and CRU-NCEP reanalysis data.
     title :          Post-processed output from JULES land surface model for ESA DTEP.

[15]: %%time
```

- Using zarr and xarray Python libraries to store and access the data
- Chunked data into a series of strips along the time axis

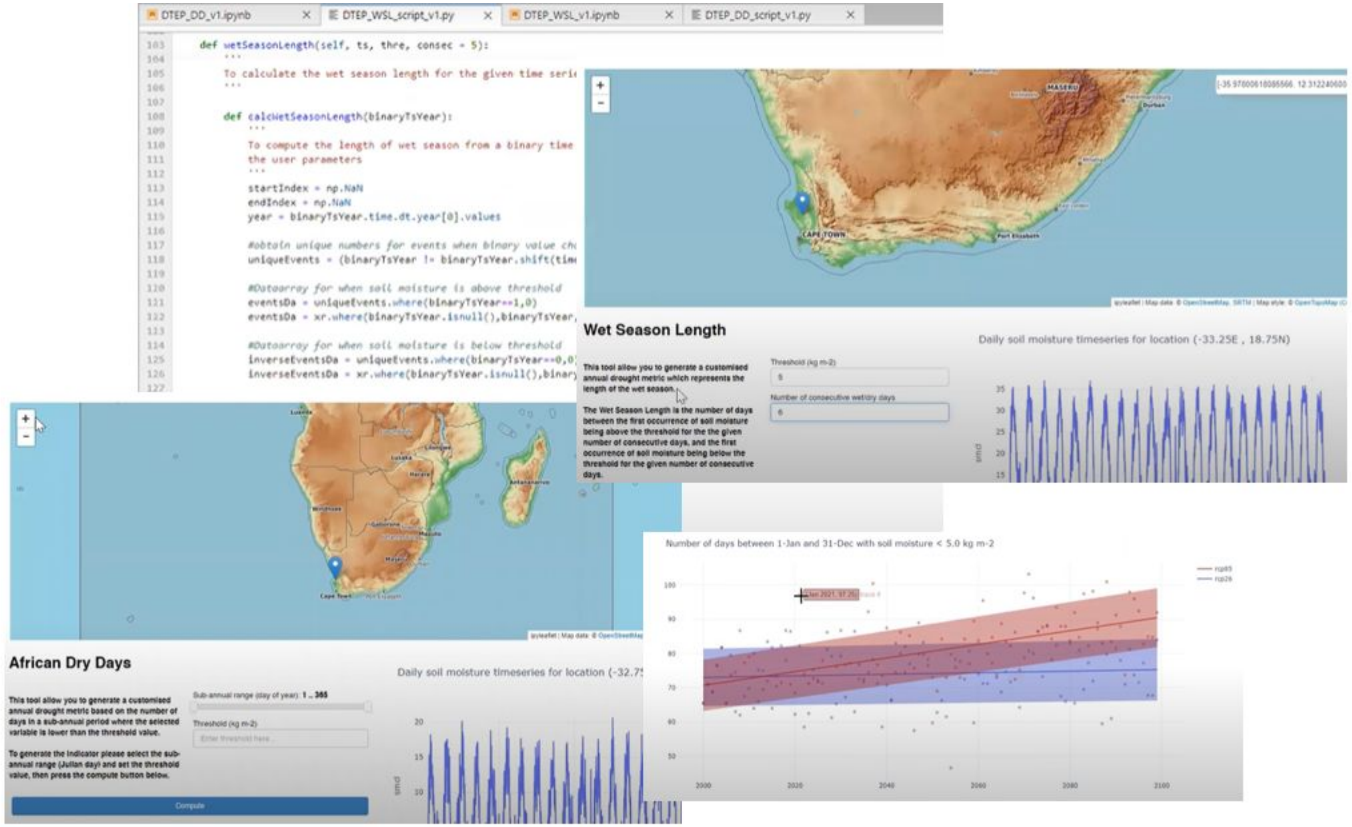
re for Environmental Analysis
ND TECHNOLOGY FACILITIES COUNCIL
ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Facilities Council
Research Council

Rechunking of data made possible interactive maps with long time series



```
    def wetSeasonLength(self, ts, thre, consec = 5):
        '''
        To calculate the wet season length for the given time seri
        '''


        def calcWetSeasonLength(binaryTsYear):
            '''
            To compute the length of wet season from a binary time
            the user parameters
            '''
            startIndex = np.NaN
            endIndex = np.NaN
            year = binaryTsYear.time.dt.year[0].values

            #obtain unique numbers for events when binary value ch
            uniqueEvents = (binaryTsYear != binaryTsYear.shift(tim

            #Dataarray for when soil moisture is above threshold
            eventsDa = uniqueEvents.where(binaryTsYear==1,0)
            eventsDa = xr.where(binaryTsYear.isnull(),binaryTsYear,

            #Dataarray for when soil moisture is below threshold
            inverseEventsDa = uniqueEvents.where(binaryTsYear==0,0
            inverseEventsDa = xr.where(binaryTsYear.isnull(),binary
```

Wet Season Length

This tool allow you to generate a customised annual drought metric which represents the length of the wet season.

The Wet Season Length is the number of days between the first occurrence of soil moisture being above the threshold for the the given number of consecutive days, and the first occurrence of soil moisture being below the threshold for the given number of consecutive days.

Threshold (kg m-2): 5

Number of consecutive wet/dry days: 6

African Dry Days

This tool allow you to generate a customised annual drought metric based on the number of days in a sub-annual period where the selected variable is lower than the threshold value.

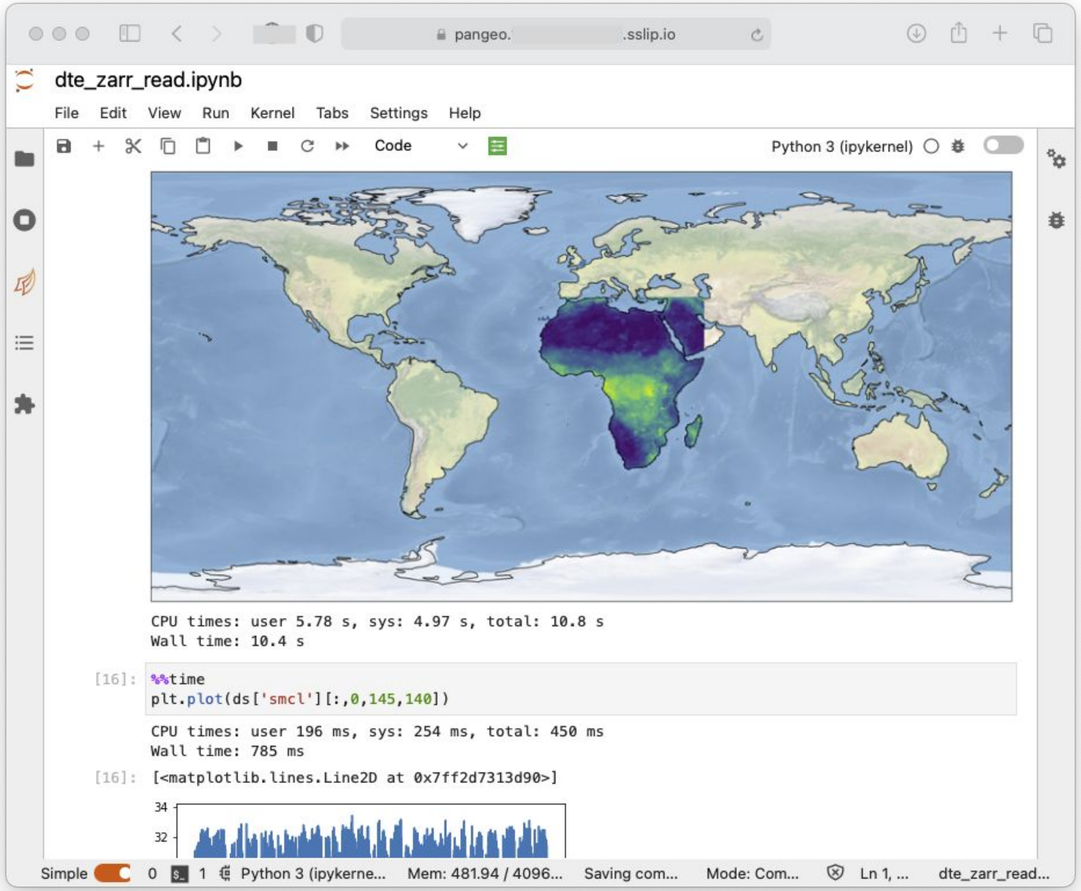
To generate the Indicator please select the sub-annual range (Julian day) and set the threshold value, then press the compute button below.

Sub-annual range (day of year): 1 .. 365

Threshold (kg m-2): Enter threshold here...

+
−
MASERU
Durban
CAPE TOWN
Port Elizabeth
Luanda
Lusaka
Harare
Lilongwe
Antananarivo
Windhoek
Gaborone
Maputo
Cape Town
Port Elizabeth
ipyleaflet | Map data © OpenStreetMap, SRTM | Map style: © OpenTopoMap
Daily soil moisture timeseries for location (-33.25E , 18.75N)
Daily soil moisture timeseries for location (-32.75
smd
35
30
25
20
15
10
Compute
Number of days between 1-Jan and 31-Dec with soil moisture < 5.0 kg m-2
rcp85
rcp26
Jan 2021, 97.25
100
90
80
70
60
50
2000
2020
2040
2060
2080
2100

Take home message: object store for analysis-ready cache specific to project needs


pangeo.[redacted].sslip.io
dte_zarr_read.ipynb
File Edit View Run Kernel Tabs Settings Help
Code
Python 3 (ipykernel)
34
32
Simple 0 1 Python 3 (ipykerne... Mem: 481.94 / 4096... Saving com... Mode: Com... Ln 1, ... dte_zarr_read...

```
CPU times: user 5.78 s, sys: 4.97 s, total: 10.8 s
Wall time: 10.4 s

[16]: %%time
      plt.plot(ds['smcl'][:,0,145,140])

CPU times: user 196 ms, sys: 254 ms, total: 450 ms
Wall time: 785 ms
[16]: [<matplotlib.lines.Line2D at 0x7ff2d7313d90>]
```

- Object store can be efficient for access on cloud
- It is essential to orient data storage to suit predominant access patterns
- Good news – re-writing data into different orientations was fast

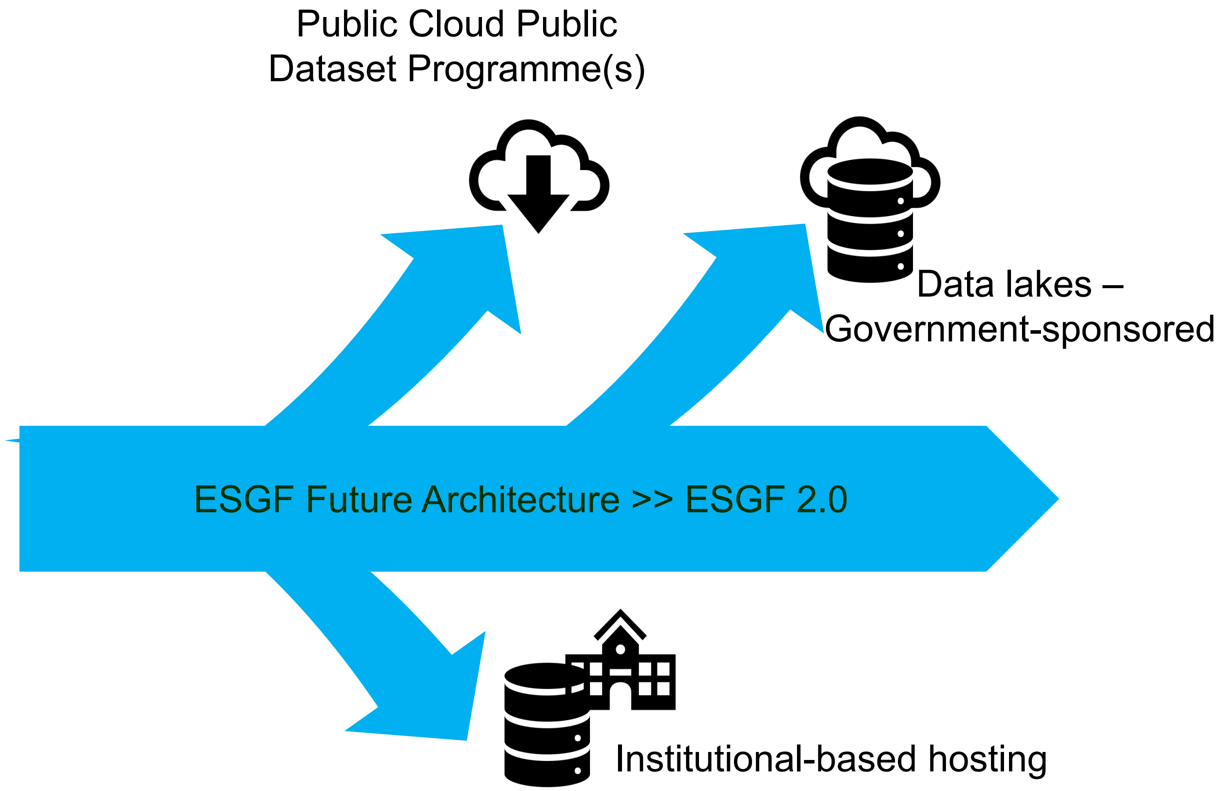
re for Environmental Analysis
ND TECHNOLOGY FACILITIES COUNCIL
ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Facilities Council
Research Council

Futures



Public Cloud Public Dataset Programme(s)

Data lakes – Government-sponsored

ESGF Future Architecture >> ESGF 2.0

Institutional-based hosting

Acknowledgements + Further Info

The Consortium

Coordinated by **CNRS-IPSL**, the IS-ENES3 project gathers **22** partners in **11** countries

cnrs
with
cea
Sorbonne Université
Met Office
National Centre for Atmospheric Science
DKRZ
BSC
CERFACS
cmcc
Koninklijk Nederlands Meteorologisch Instituut
SMHI
UK Research and Innovation
UC
DLR
netherlands eScience
Norwegian Meteorological Institute
METEO FRANCE
MANCHESTER 1824
WAGENINGEN UNIVERSITY & RESEARCH
CHARLES UNIVERSITY
NORCE
LiU LINKÖPING UNIVERSITY

ESGF Future Architecture Report: https://doi.org/10.5281/zenodo.3928222

IS-ENES3 website
https://is.enes.org/

Contact us at
is-enes@ipsl.fr

@ISENES_RI
@cedanews
@PhilipJKershaw

Subscribe to the
IS-ENES3 H2020
Youtube channel !

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

UKRI Science and Technology Facilities Council · Natural Environment Research Council

Centre for Environmental Data Analysis
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL