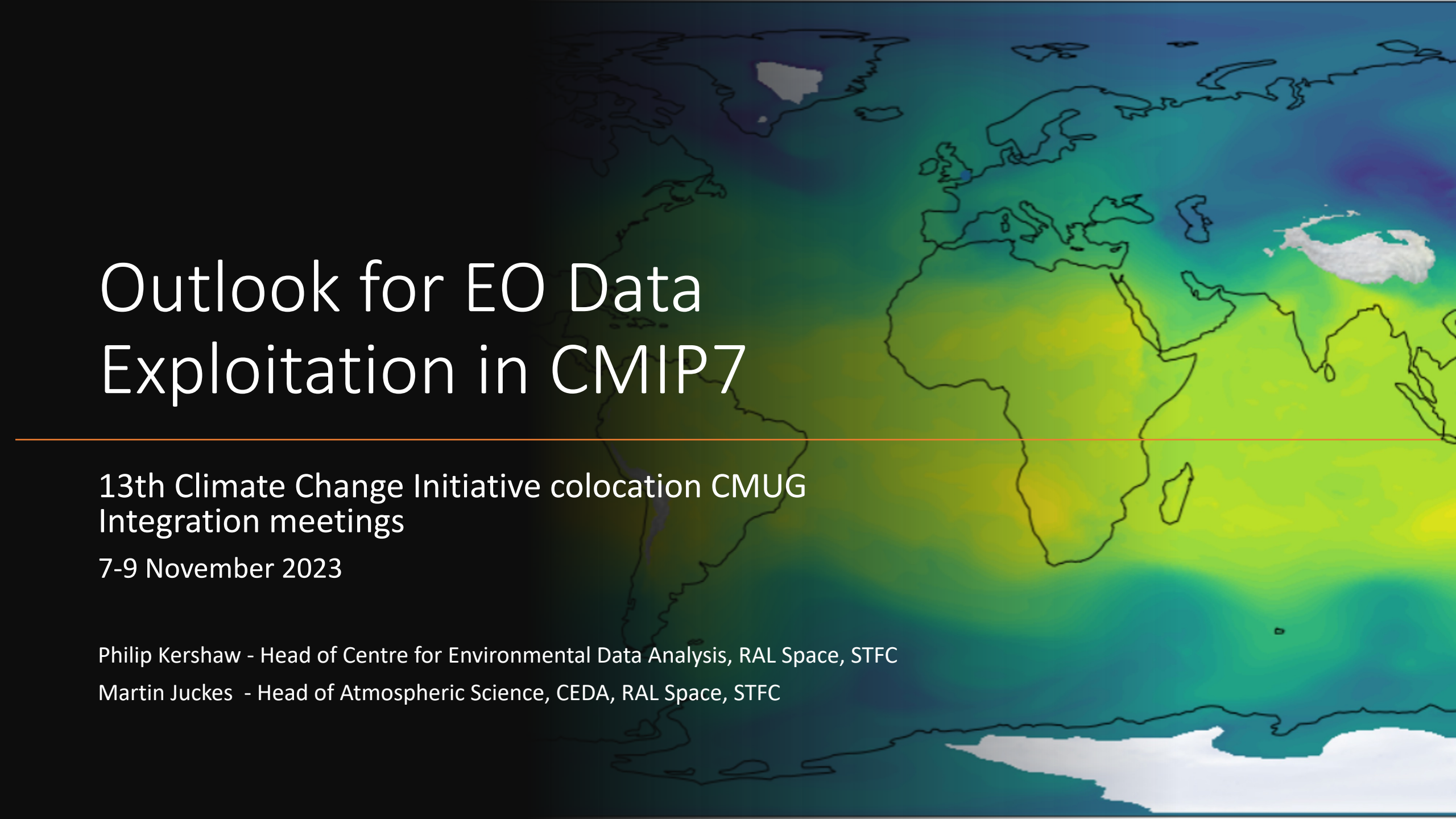


Outlook for EO Data Exploitation in CMIP7

13th Climate Change Initiative colocation CMUG Integration meetings

7-9 November 2023

Philip Kershaw - Head of Centre for Environmental Data Analysis, RAL Space, STFC

Martin Juckes - Head of Atmospheric Science, CEDA, RAL Space, STFC

CEDA, CMIP, WCRP and IPCC

- CEDA's mission: provide data and information services for environmental science on behalf of NERC
- Support climate research -
  - Bringing CMIP and other significant global datasets into our data archive on the JASMIN analysis platform
  - Publishing UK climate simulations and observational datasets to make them accessible to the global scientific and science assessment community
  - Working with peer organisations around the world to create an efficient and scalable global data infrastructure for climate model data (ESGF)
  - Working with the community and peers to establish standards and governance.

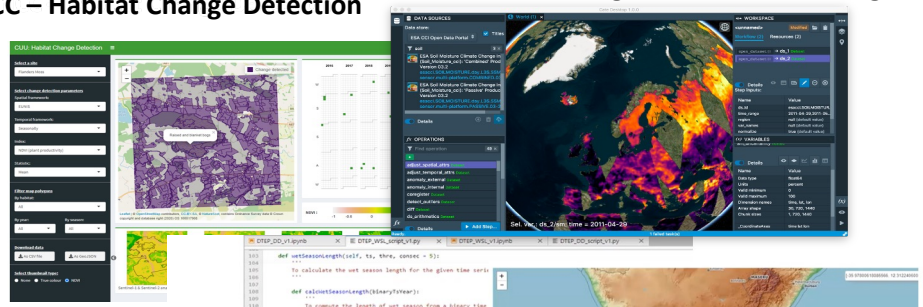
This presentation …

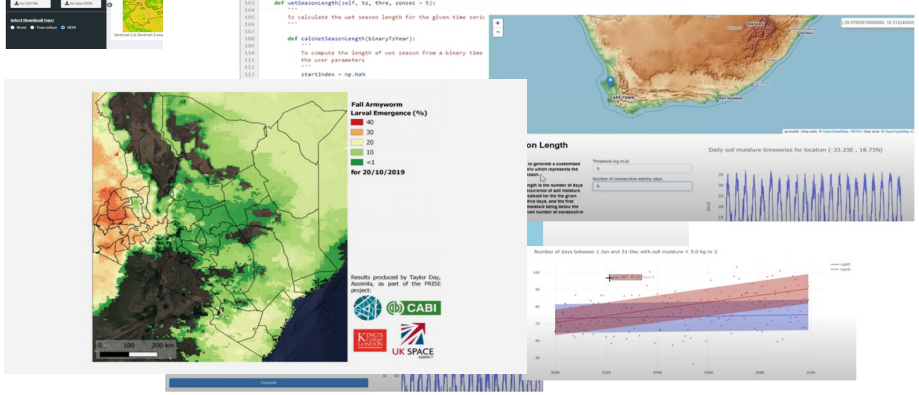
- Data exploitation and the role of data analysis *Platforms*
- Platforms in the wider context of the Earth sciences – their evolution and how they fit into a wider ecosystem of digital infrastructure and services
- Earth System Grid Federation (ESGF) – recent developments
- The EO DataHub –UK initiative to build a new data analysis platform
- Data exploitation in the context of CMIP7, planning, funding, governance
- Thoughts on futures – How does a platform approach best support climate model evaluation taking advantage of EO data

Platform Development Approach

JNCC – Habitat Change Detection

ESA Climate Change Initiative Knowledge Exchange Project





PRISE – Assimila, CABI, KCL, STFC

ESA Digital Twin Earth Climate Explorer - Soil moisture prediction

- Platform: **an underlying software infrastructure** to support the development of applications and web services
- The Platform supports application development through the provision of
  - Hosting – compute and storage (typically cloud)
  - Data
  - Supporting software services and APIs

One example - JASMIN:

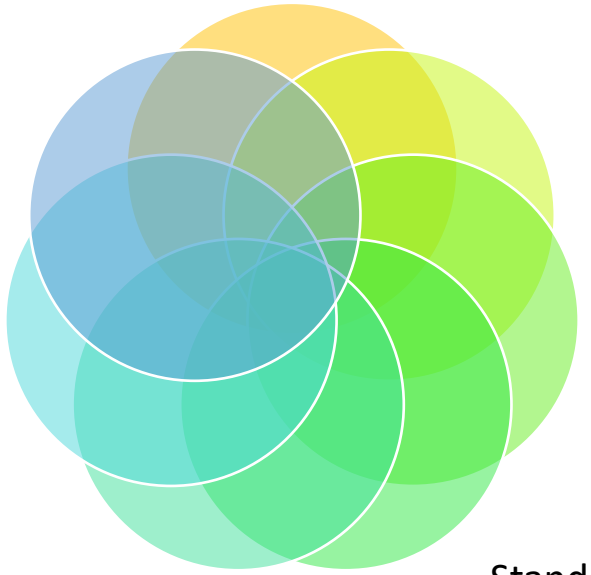
- 10 years' experience with this model serving NERC user community and working with partners in industry and the public sector
- ~40 PB disk and 15k cores

Platforms and Broader International Context

ESA Thematic Exploitation Platforms

Climate Platforms: Pangeo

Copernicus Climate Data Store, DIAS



NERC Digital Solutions Programme, NERC DataLabs

Public cloud offerings: Google Earth Engine, MS Planetary computer, Earth on AWS

Standards: OGC, ESA EO Exploitation Platform Common Architecture

EO DataHub

UKRI 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

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

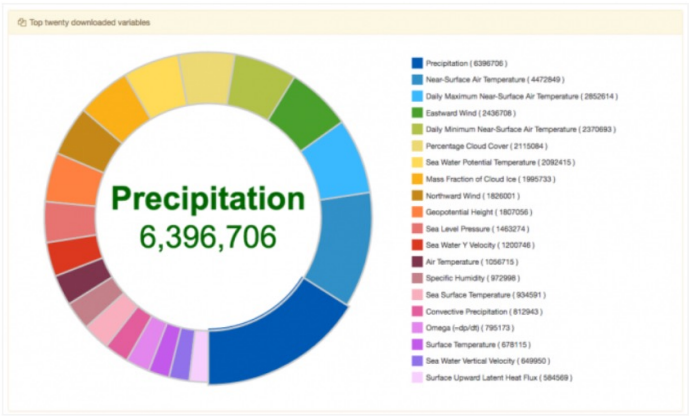
Earth System Grid Federation

37	188	26	15,075,633	1,350,998,839	41.57
Projects	Countries	Data nodes	published datasets	files downloaded	PB downloaded

ESGF Federation

This user interface provides a set of data usage and publication metrics across the Earth System Grid Federation. Statistics refer to the period January 2018 to present.

Data usage



Cross-project and project-specific sections, with a rich set of charts and tables, provide different views about the data downloaded across the ESGF federation.

Data publication

A view of the total amount of data published and available through the ESGF infrastructure gives users an in-depth view about the ESGF data archive.

Ref: http://esgf-ui.cmcc.it/esgf-dashboard-ui/

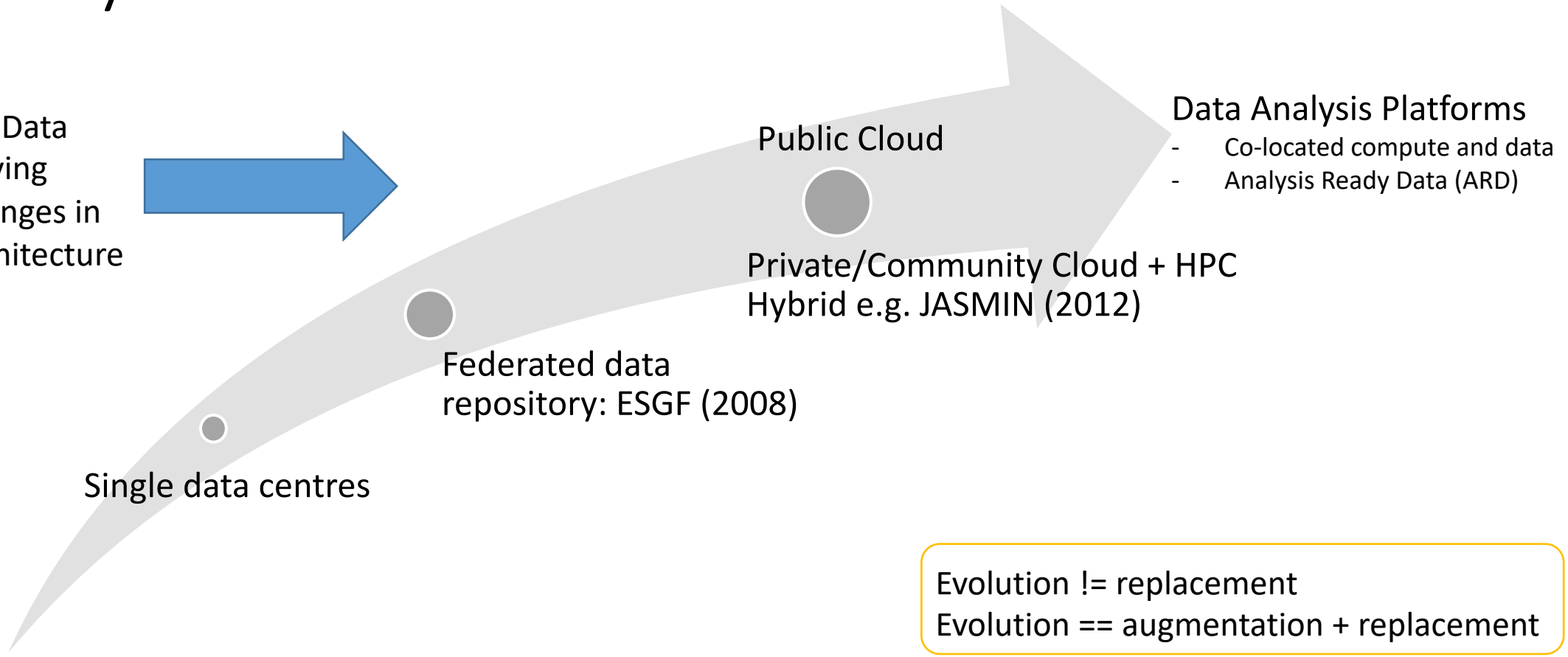
UKRI Science and Technology Facilities Council Natural Environment Research Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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

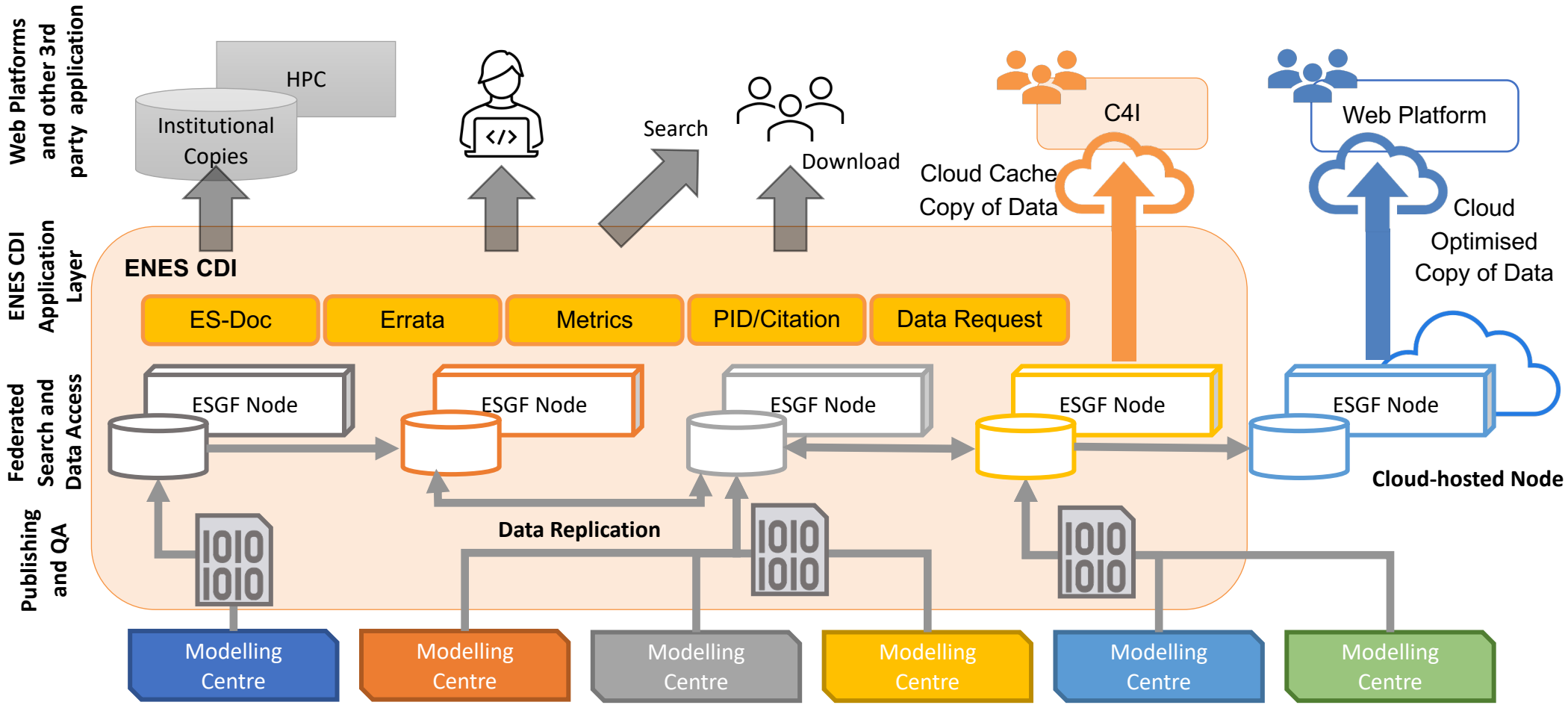
Evolution in architectures for data access and analysis*



Evolution != replacement
Evolution == augmentation + replacement

* Slide from AGU presentation 2020: Cloud futures for CMIP data – evaluating object storage models and re-evaluating federation for data distribution

ESGF and the ENES Research Infrastructure

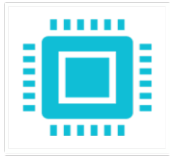


National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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

ESGF Development Plans

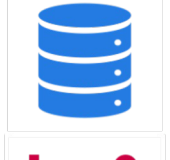


Installation and Systems Administration – Container-based installation developed and being rolled out across the federation

Search Services – new search system under development using Globus (US) and ElasticSearch (Europe). Use common STAC API

Identity and Access Management – uses OpenID Connect / Oauth 2.0 – Globus (US) and EGI Checkin (European) Identity Providers

Compute Services – ENES testbed in operation using web processing service (DKRZ) integrated with web frontend (Climate4Impact)



New modes for data access and storage

Metrics Collection – container-based installation has hooks for linking to CMCC Dashboard

EO DataHub – A new platform for the UK

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Met Office

NPL

CATAPULT
Satellite Applications

UK SPACE AGENCY

- 2-year project, started Feb ‘23 - 10m investment funded from DSIT EO Transitional Package

- Goals:
  - Be a new ‘single point’ EO Data infrastructure, that builds on current UK EO assets and brings together UK EO data offerings from public and commercial centres
  - Enable new EO services and tools to be developed and accessed by the UK EO data community by providing a transformational layer
  - Address key challenges in EO data access and discovery, interoperability, transparency, and trustworthiness
  - Support industry, public sector and academic communities

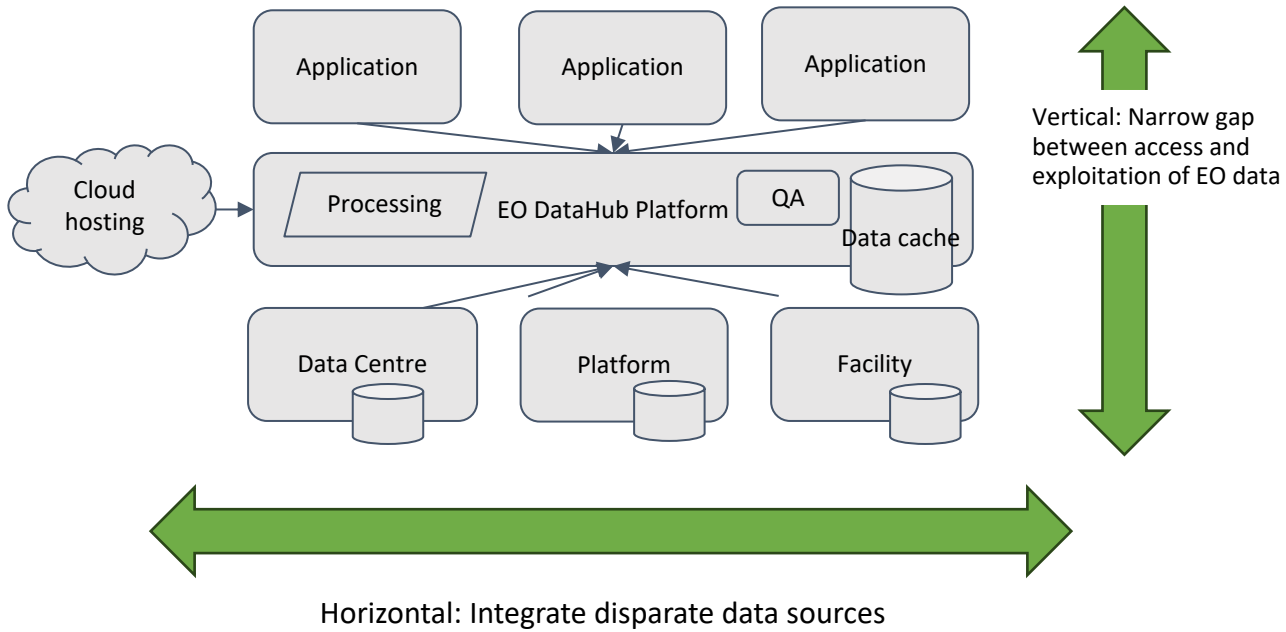
UKRI Science and Technology Facilities Council Natural Environment Research Council

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

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

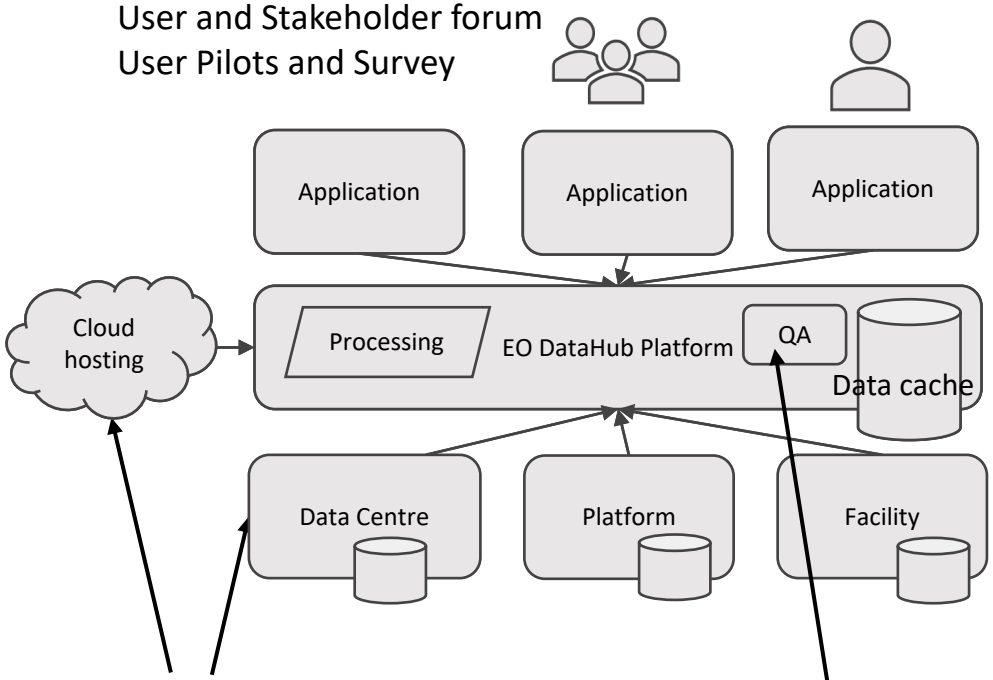
EO DataHub – Architecture Concept



EO DataHub Project Work Areas

User engagement

To inform the overall development esp. Applications
User and Stakeholder forum
User Pilots and Survey



Applications

- ITT will fund development of at least three applications which exploit the Hub Platform
- Themes: Climate, Infrastructure and Utility Systems (Inc. Water & Energy), National Resilience

Hub Platform: software development and operations ITT released

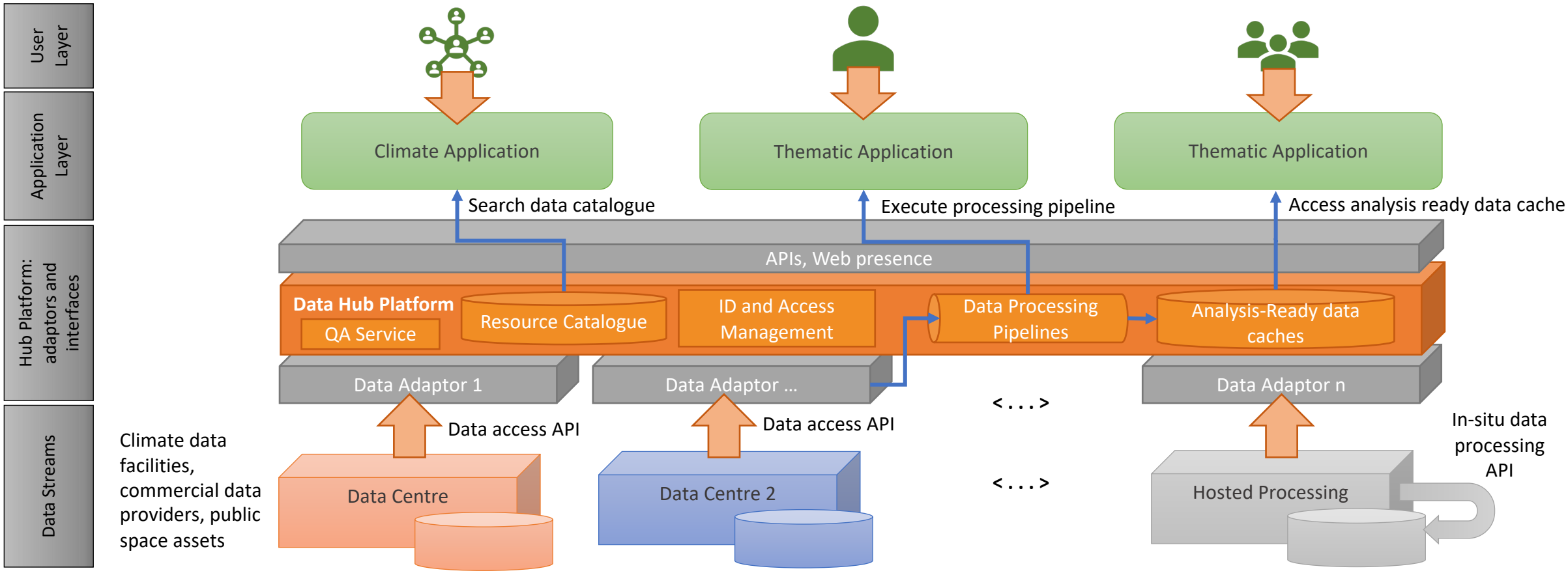
Data Streams: provide data services

- Research community data provider: CEDA – Sentinel, UKCP, CMIP6, CORDEX
- Commercial data provider(s): sourced from ITT to be released – likely hi-res optical and SAR

Hosting: Public cloud + some on-premise for data product generation (***EOCIS Project***)

Quality Assurance: provide quality information to inform users on suitability of data for given purpose

EO DataHub High-Level Architecture



UKRI 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

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

Core Functional Areas

Data Streams + access interfaces

allow integration of different data providers and production and caching of outputs on the Hub

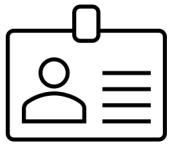
OGC W*S – WMS, DataCubes, Object store – COG, Zarr

Resource Catalogue

support search of data and processing resources

Quality Assurance for data

STAC, OpenSearch, OGC, …



Identity and Access Management, Accounting and Metrics

grant access to secured resources, account and charge for use as required

OpenID Connect, OAuth 2.0, …

Data Analysis, Processing and workflow management

Analyse data, develop and incorporate custom code, execute and integrate into workflows

Jupyter, Dask, CWL, Apache Airflow, …

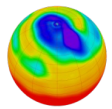
Web presence

Presentation and glue

UKRI 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



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

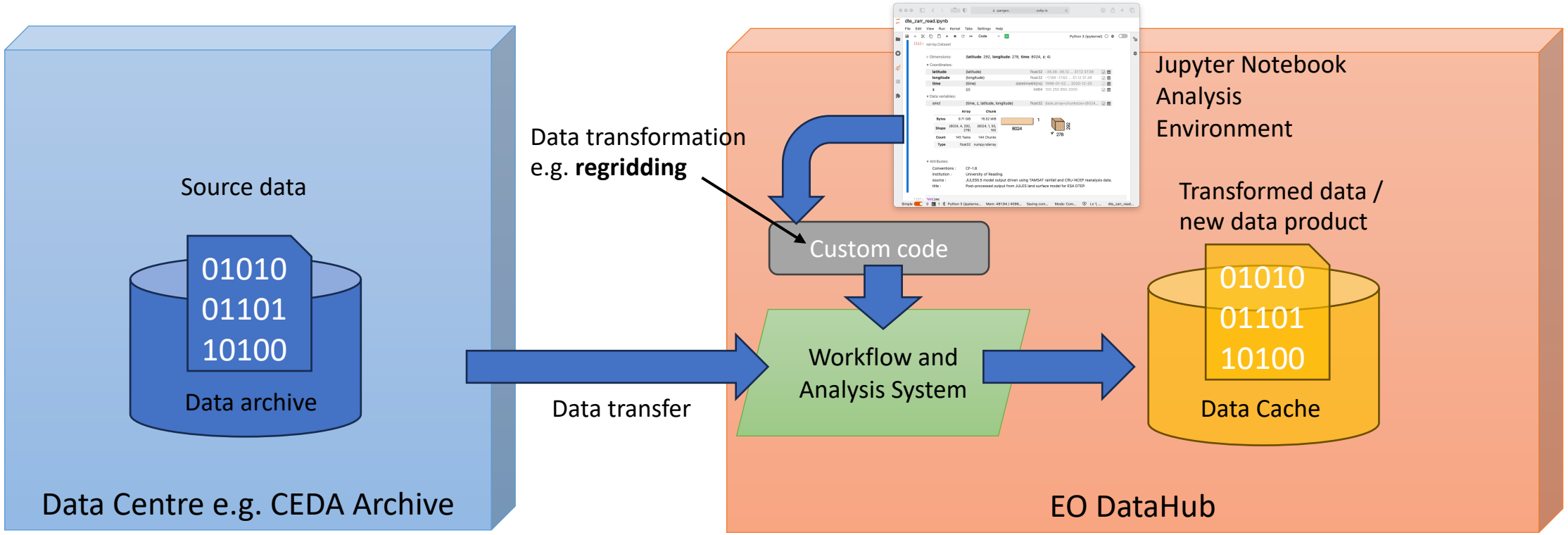
Data Processing to facilitate Data Analysis and Exploitation

- Support for data processing and orchestration of workflows are a core component for an EO Platform

- This is a fundamental capability to support the transformation of source data into forms for further analysis

- Such processing pipelines could be devised to facilitate integration and intercomparison of model outputs and observations

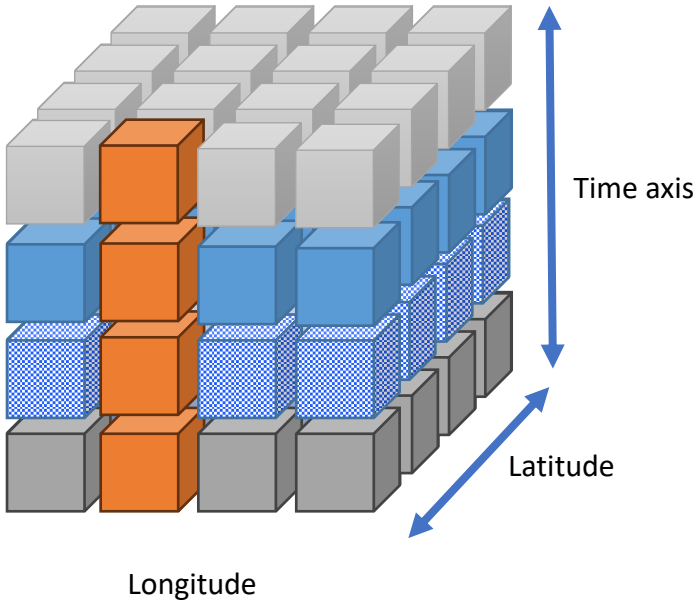
Data Processing to facilitate Data Analysis and Exploitation



Data Formats and Access

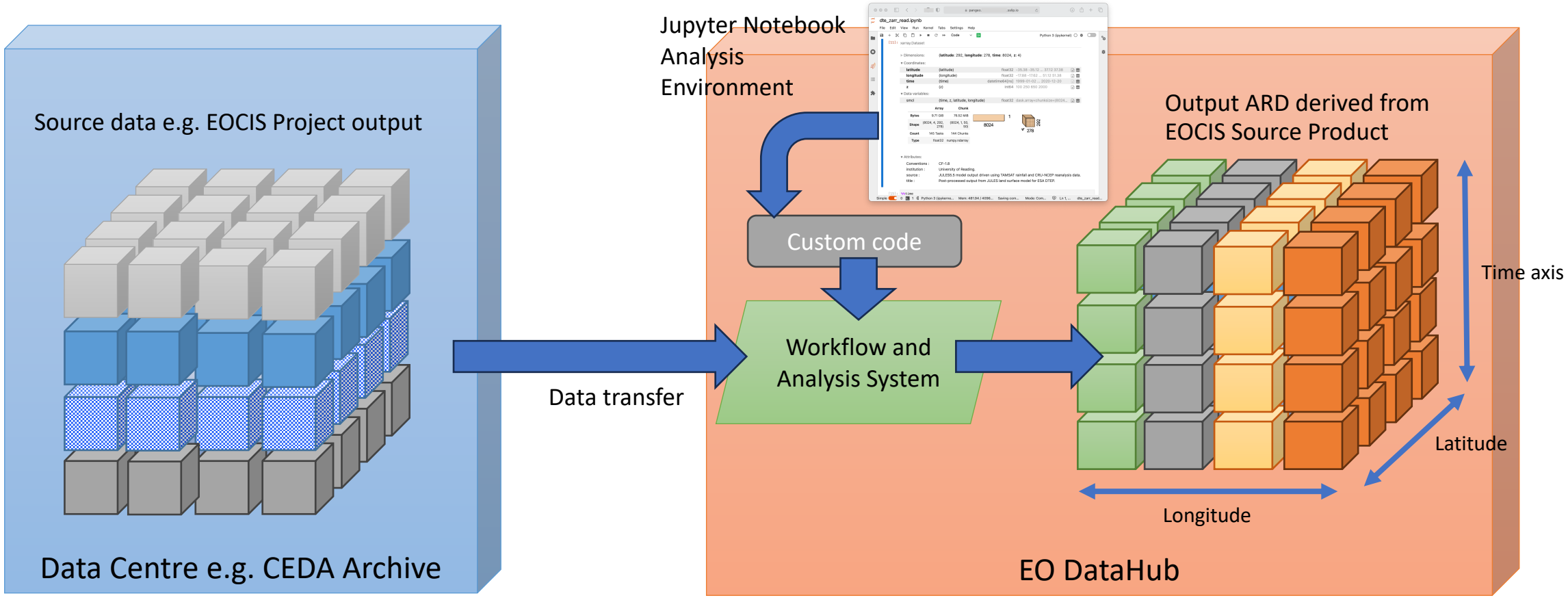
- Analysis-Ready data
- Cloud Optimised GeoTIFF (COG)
  - Popular for EO datasets
  - Enables efficient access of files through the use of HTTP Range GET operations on data stored in object store
- CF-netCDF
  - Predominate standard for climate data
  - Zarr has become popular for cloud storage of netCDF data on object store
  - Kerchunk uses alternative strategy preserving the original netCDF file format but presenting a Zarr compatible interface to consuming client libraries
- Jupyter + xarray + Dask for access and analysis

Data Chunking to optimise read performance



- Example scenario –
- Data processing creates an output file one per time step
  - In the diagram vertically stacked coloured slices as the processing proceeds
- Later, for analysis, a user wants to sample an individual location as a time series (orange vertical)
  - This involves potentially opening hundreds of individual files(!)

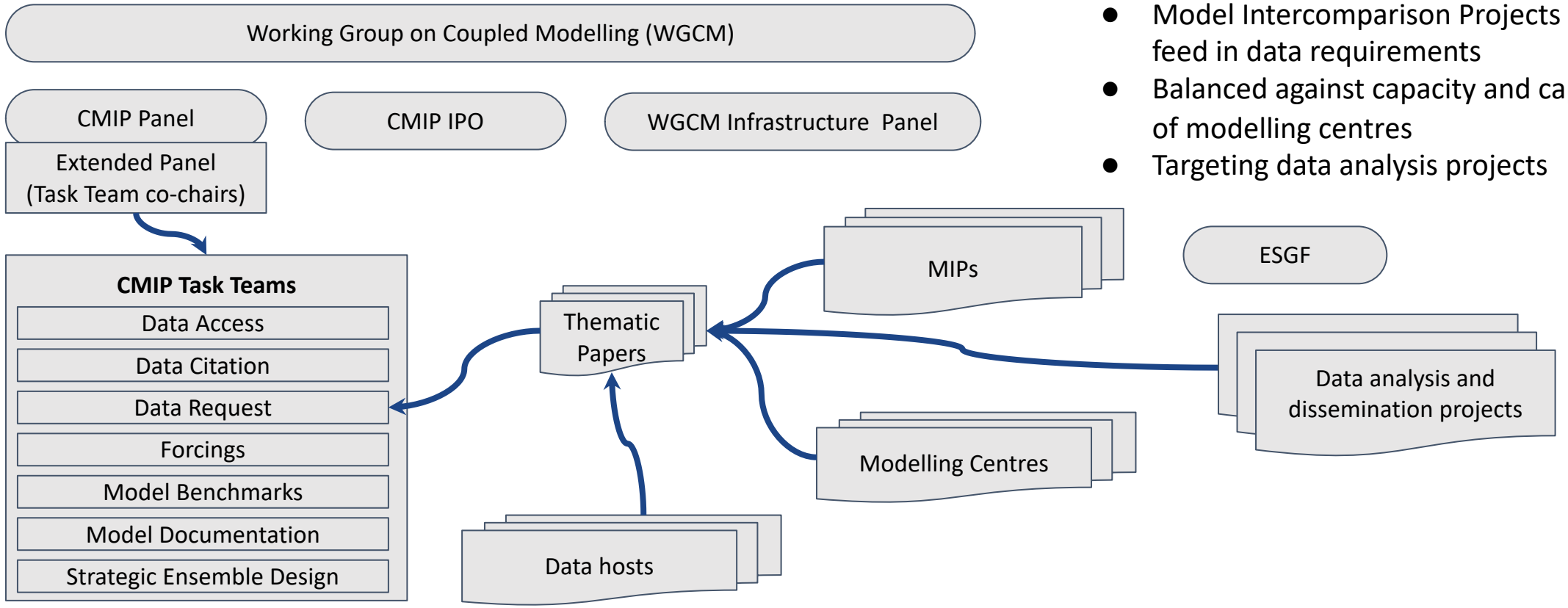
Data Processing Pipeline: rechunking to suit custom data analysis access pattern



EO Data and CMIP7 – Engagement, Processes, and Governance

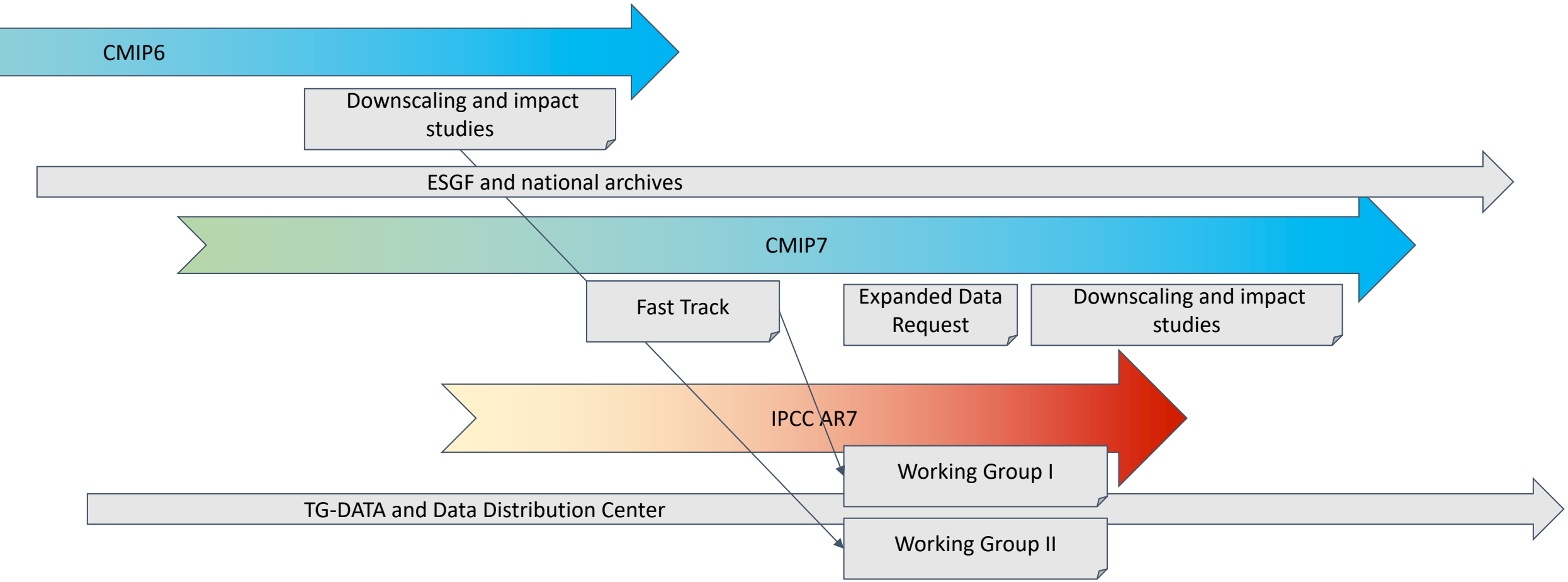
- Effective engagement with the community and participation in the processes and governance for CMIP essential to deliver successful software services
- CMIP has model intercomparison projects that have a targeted focus - this might be an opportunity for specific integration with EO data - in terms of model processes
- How might you do this? - propose a model intercomparison project (MIP)
  - MIPs - need a co-ordinator, needs enough interest from the community

CMIP Data Requirements Workflows

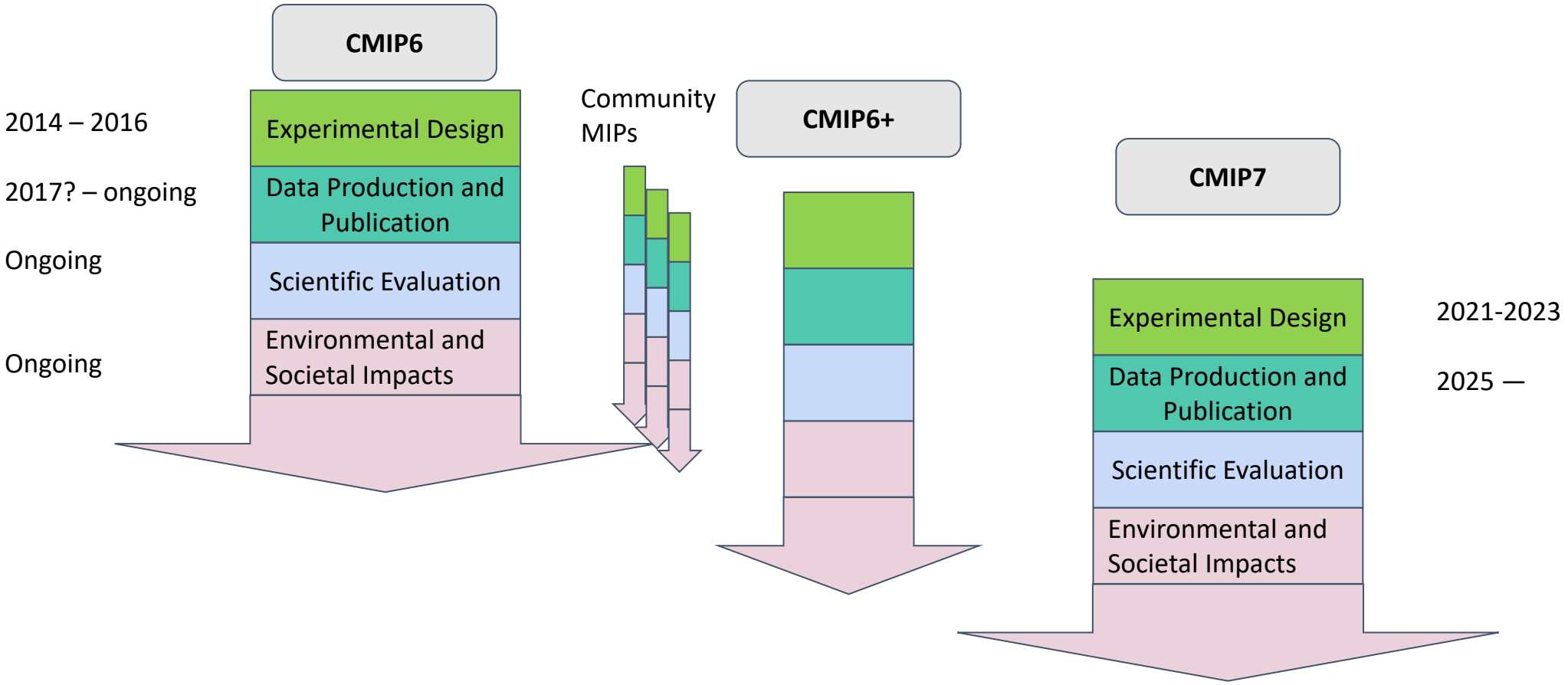


- Model Intercomparison Projects (MIPs) feed in data requirements
- Balanced against capacity and capabilities of modelling centres
- Targeting data analysis projects

Overlapping Phases of CMIP Exploitation



Overlapping Phases of CMIP Exploitation (cont.)



Specific Areas for Potential Engagement

- Supporting evaluation of physical climate simulations for IPCC AR7 WG I – CMIP7 Fast Track
  - Publication of data through Obs4MIPS
  - Developing community tools and platforms
  - Propose a dedicated MIP
- Supporting evaluation of impact modelling in IPCC AR7 WG II – CORDEX downscaling of CMIP6 and distributed community efforts
- Supporting transparency and access to EO Data
- IPCC Task Group on Data (Governance body appointed by IPCC)
- IPCC Data Distribution Centre (Data curation experts funded to work under IPCC governance)

Summary

- New developments in ESGF include a container-based deployment, integration of modern identity and access management and new search services
- ESGF provides the underpinning infrastructure to support a platform-based approach for data analysis
- EO DataHub – new platform which brings together climate model and EO datasets (including new products from EOCIS project)
- New initiatives for interfacing climate models and observations could explore how they might integrate with the CMIP process e.g. initiate a dedicated MIP

Vision for Ecosystem

Leverage platform services to build applications which exploit the model and obs data

Platform provides services for data transformation and exploitation

ESGF underpinning infrastructure + CCI and other EO data sources