

## Development of Analytics Services

Stephan Kindermann (DKRZ), Alessandro Spinuso (KNMI), Carsten Ehbrecht (DKRZ), Paola Nassisi (CMCC)

# Overview

- ENES CDI Analytics Services
- Analytics service activity (SA, TNA)
- Web processing service
- Future Perspectives

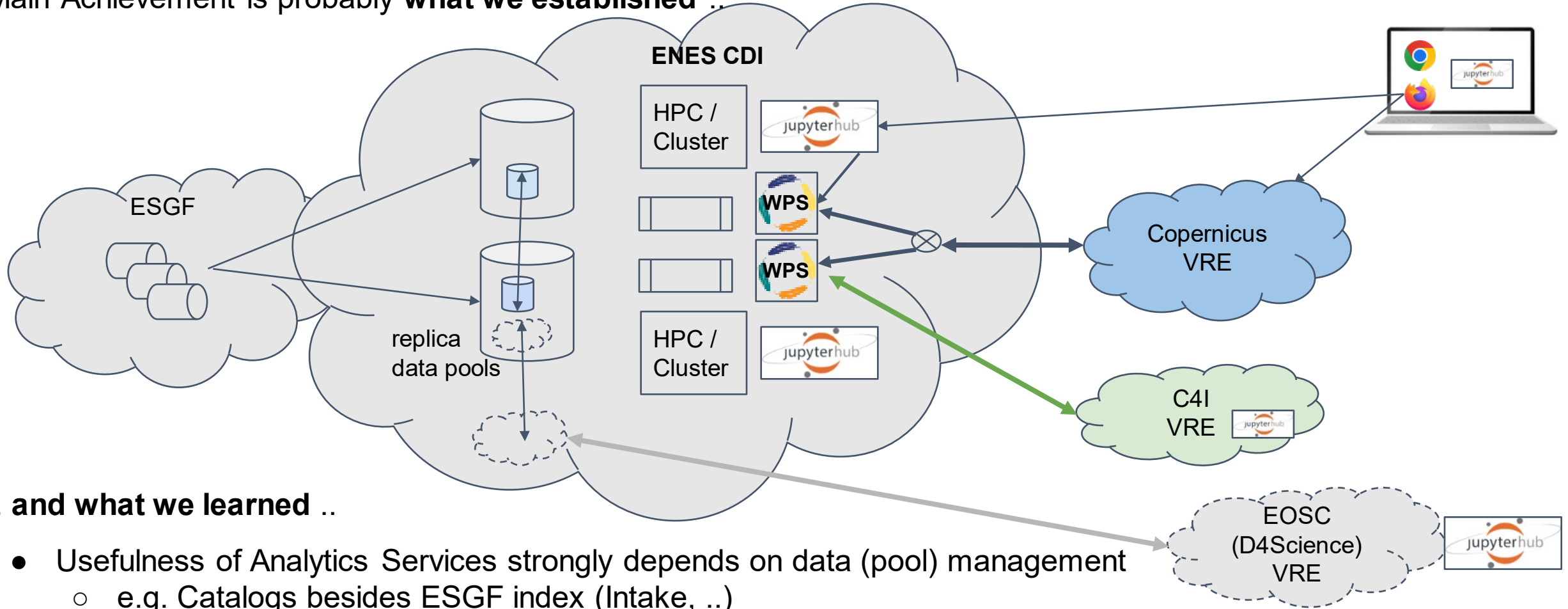
# Processing / Analytics Services

## Two complementary analytics service development efforts:

- A) Harmonize access to institutional processing capabilities with focus on (interactive) Jupyter-Hub based data analytics
  - service provisioning (SA and TNA)
  - IPCC WG support
  - aligned with broader community efforts (PANGEO SW stack, Intake data catalogs etc.)
  - aligned with EOSC efforts (community VRE support, EOSC-pillar example)
  
- A) Standards (OGC) based Processing Web Service provisioning
  - integrating data access and data pre-processing behind one interface
  - used by C4I portal
    - used in production for Copernicus data provisioning
  - based on OGC/GeoPython and bird-house, close collaboration with Canada
  - proposed solution for an ESGF compute service (ESGF CWT)

# Processing / Analytics Services

Main Achievement is probably **what we established ..**



**.. and what we learned ..**

- Usefulness of Analytics Services strongly depends on data (pool) management
  - e.g. Catalogs besides ESGF index (Intake, ..)
- Interfaces to Analytics services (WPS) are important VRE integration points
  - e.g. one interface for data download and (pre-)processing (WPS)
- Cloud (and analysis ready data) based Analytics Services becoming more important

# Interactive Analytics Services: Jupyterhub

## Main Achievements (IS-ENES3)

- Provisioning of Jupyter-hub instances at ENES-CDI ESGF tier1 centers (DKRZ, IPSL and STFC) as well as CMCC with access to replica data pools
  - Intake catalog support to exploit data pools at some sites
- Demonstrated usefulness of consistent community SW stack setup at different sites (at DKRZ and CMCC to support summer school with a failsafe environment)
  - Pangeo related Amazon/Google Cloud CMIP6 community notebooks are supported by ENES CDI analytics service
  - Specific support for ESMValTool based analytics (Jupyterhub kernel)
- Demonstrated integration with external (EOSC based) Jupyterhub based VRE (EOSC-pillar project)
- (Usefulness of data pools with associated interactive analytics services at DKRZ and STFC acknowledged in CMIP6 survey)

# Analytics Services: Service Aspects (SA and TNA)

## Main Achievements:

- Supported 15 group applications during IS-ENES3 (TNA)
- > 4000 registered users for VA compute service
- Service Activities helped to harmonize the Analytics Environments
  - Coordinated Training Material, Coordinated compute environments (xarray,..)
  - Shared approach for local data catalogs: Intake (ESGF catalog not so usefull ..)
- Lessons learned:
  - TNA:
    - relatively low CPU resource requirements
    - relatively small research groups (mostly individuals, exception: Primavera support)
    - some requirements for longer term support e.g. SMEs (Geoskop)
  - VA:
    - short term analytics / on demand requirements → lightweight quick application procedure
    - jupyter hub + data catalog (intake) for local data useful

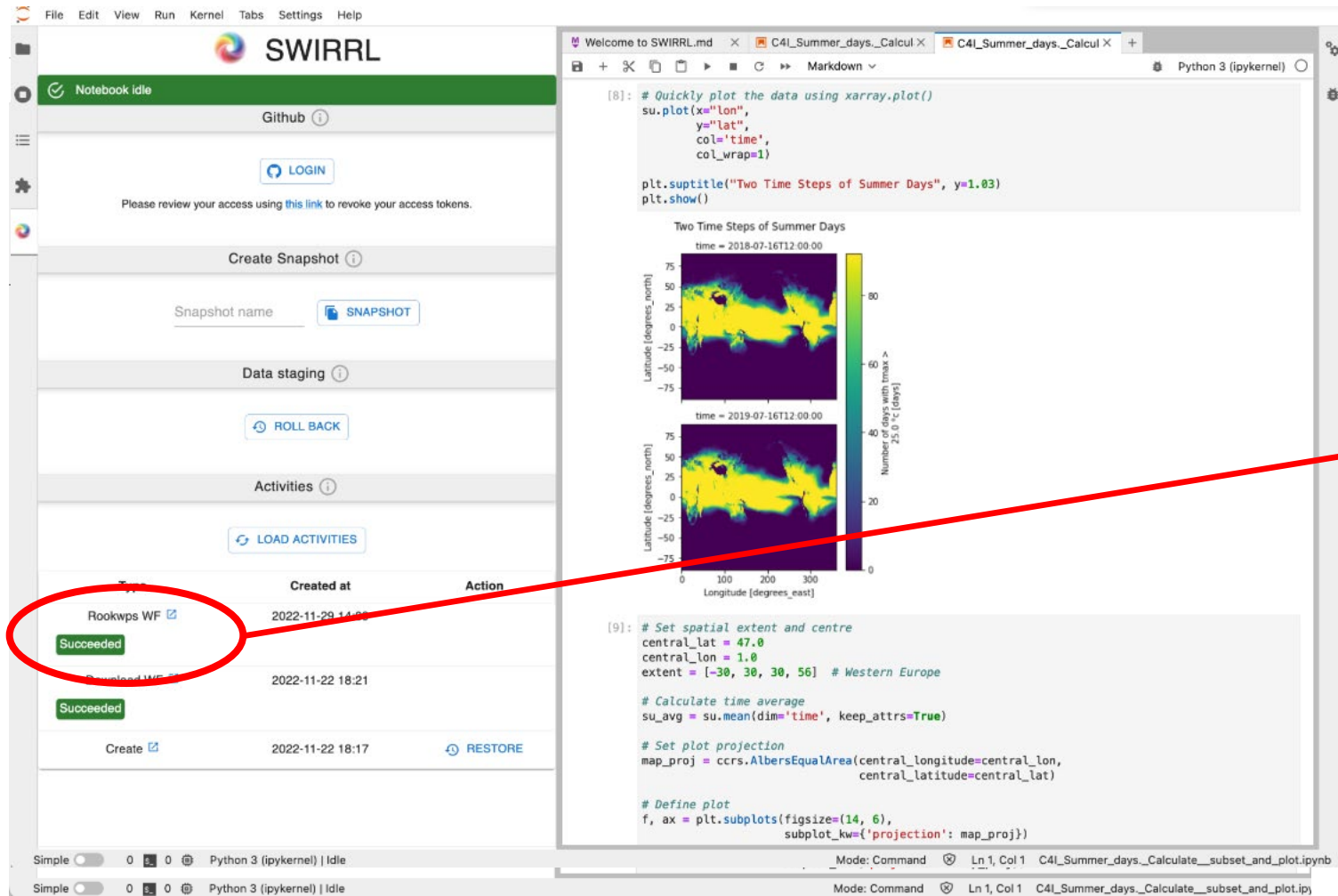
→ requirement to sustain a coordinated “ECAS” service beyond IS-ENES3

# OGC Web processing interface based analytics service (WPS) developments

## Main achievements:

- Demonstrated suitability as a uniform interface solution for data download and data processing in ENES CDI.
- Stable cooperation around different parts of the SW stack and deployment approach (Copernicus, Canada, OGC)
- Demonstrated production deployment readiness of WPS developments in Copernicus service provisioning
  - load balanced deployment at DKRZ and IPSL (and STFC)
  - rooks WPS (subsetting, averaging, ...)
  - relies on distributed ENES CDI data management
- Integrated with C4I
- Provenance reporting for WPS calls (exploited in Copernicus and C4I integration)

# Climate4Impact Workspaces - WPS Integration



The screenshot displays the SWIRRL web interface. On the left, a sidebar contains navigation links: Notebook idle, Github, LOGIN, Create Snapshot, Data staging, and Activities. The main panel shows a table of activities. The first activity, 'Rookwps WF', is circled in red and has a 'Succeeded' status. The second activity, 'Rookwps WF', is also 'Succeeded'. The third activity, 'Create', is in progress. The right panel shows a Jupyter notebook with two plots. The first plot, titled 'Two Time Steps of Summer Days', shows two maps of Europe with a color scale for 'Number of days with Tmax > 25.0 °C (days)'. The second plot, titled 'Set spatial extent and centre', shows a map of Europe with a color scale for 'Number of days with Tmax > 25.0 °C (days)'.

Created at	Action
2022-11-29 14:00	
2022-11-22 18:21	
2022-11-22 18:17	RESTORE

```
{
  "prov:used": [...],
  "provone:hadPart": [
    {
      "prov:wasAssociatedWith": [...],
      "@type": [
        "Resource",
        "prov:Activity",
        "provone:Execution"
      ],
      "rdfs:label": "orchestrate",
      "prov:startedAtTime": "2022-11-29T14:13:02Z",
      "@id": "urn:roocs:orchestrate_16ca3e1f-fee6-4419-a264-65d273a801bf",
      "prov:endedAtTime": "2022-11-29T14:13:45Z"
    },
    {
      "prov:wasAssociatedWith": [
        { ... }
      ],
      "prov:wasActivityOfInfluence": [...],
      "@type": [
        "Resource",
        "prov:Activity",
        "provone:Execution"
      ],
      "rdfs:label": "average_ta_1",
      "roocs:apply_fixes": false,
      "@id": "urn:roocs:average_ta_1_6eabac6b-444e-4e9a-a71b-bcc1b16b9fb1",
      "roocs:dims": "time"
    }
  ],
  "@type": [...],
  "@context": { ... },
  "prov:generated": { ... },
  "swirrl:sessionId": "13abfb95-fd4a-463f-ba61-fc96d40a9b6c",
  "swirrl:message": "Succeeded",
  "prov:wasAssociatedWith": [...],
  "@id": "urn:uuid:49f0d7f5-5cc8-4893-8e17-2a18ed870540",
  "swirrl:jobId": "49f0d7f5-5cc8-4893-8e17-2a18ed870540",
  "prov:endedAtTime": "2022-11-29T13:13:54.220Z",
  "prov:atLocation": "POST /workflow/rookwps/run/",
  "prov:startedAtTime": "2022-11-29T13:09:03.676Z"
}
```

Thanks to the use of PROV by Rooks and SWIRRL, provenance is easily merged, stored and made available to users and machines (interoperable)



# WPS sustainability achievements

## Based on modular Ecosystem

- Client side
  - OGC/OWSLib: WPS client module
  - Birdy: based on OWSLib, used in notebooks
  - Rooki: based on Birdy, used by CDS
- Server side
  - OGC/PyWPS: Python WPS implementation
  - Rook: subsetting service for CMIP and CORDEX
- Libraries:
  - clisops: subset, average, regrid with xarray for CMIP and CORDEX
- Deployment:
  - Ansible playbook
  - Docker

# WPS Sustainability

- ENES CDI WPS approach promoted as an ESGF compute service solution
- Community based development: OGC, Canada (Ouranos), STFC, IPSL, DKRZ, ...
- Demonstrated usefulness of ENES CDI WPS approach in external projects:
  - Copernicus Climate Data Store (data access for CMIP, CORDEX)
  - H2020 Climate Intelligence Project (CLINT): WPS based AI infilling service

# CLINT- WPS based AI Climate Service

Fill gaps in climate dataset using a web processing service

 **ClintAI** Please complete the form below and submit a job.

Fills the gaps in your uploaded climate dataset (HadCRUT).

[ClintAI Logo](#) [Clint AI](#) [Clint Project](#) [HadCRUT on Wikipedia](#) [HadCRUT4](#) [HadCRUT5](#) [Near Surface Air Temperature](#)

Add your HadCRUT file here

URL

Enter a URL pointing to a HadCRUT NetCDF file. Example: [https://www.metoffice.gov.uk/hadobs/hadcrut5/data/current/non-infilled/HadCRUT.5.0.1.0.anomalies.ensemble\\_mean.nc](https://www.metoffice.gov.uk/hadobs/hadcrut5/data/current/non-infilled/HadCRUT.5.0.1.0.anomalies.ensemble_mean.nc)

HadCRUT variant

HadCRUT5

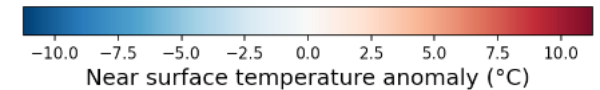
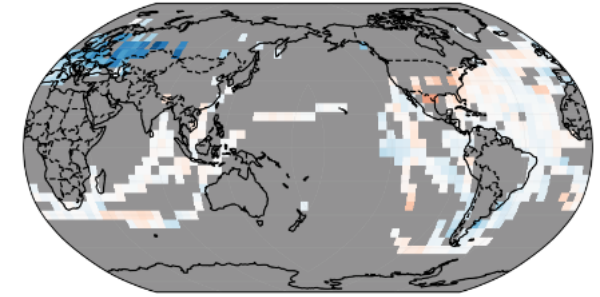
Choose HadCRUT variant of your dataset.

Submit

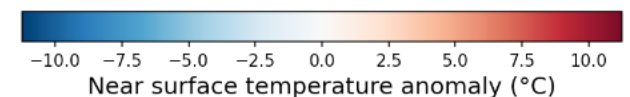
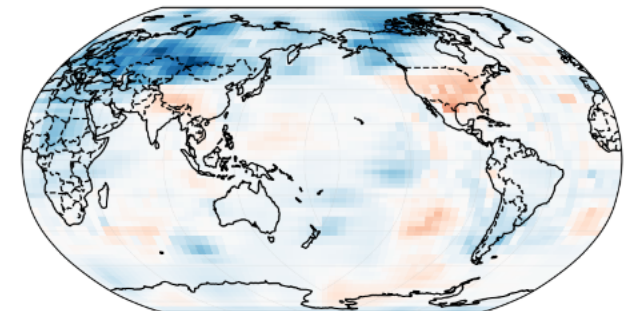
Job Succeeded: [Show Output](#) [Details](#)

```
1 0:00:02 0%: PyWPS Process clintai accepted
2 0:00:04 20%: Infilling ...
3 0:00:06 20%: Infilling ...
4 0:00:08 20%: Infilling ...
5 0:00:10 20%: Infilling ...
6 0:00:15 20%: Infilling ...
7 0:00:20 20%: Infilling ...
8 0:00:20 100%: PyWPS Process ClintAI finished
```

**Before**



**After**



# Future Perspectives:

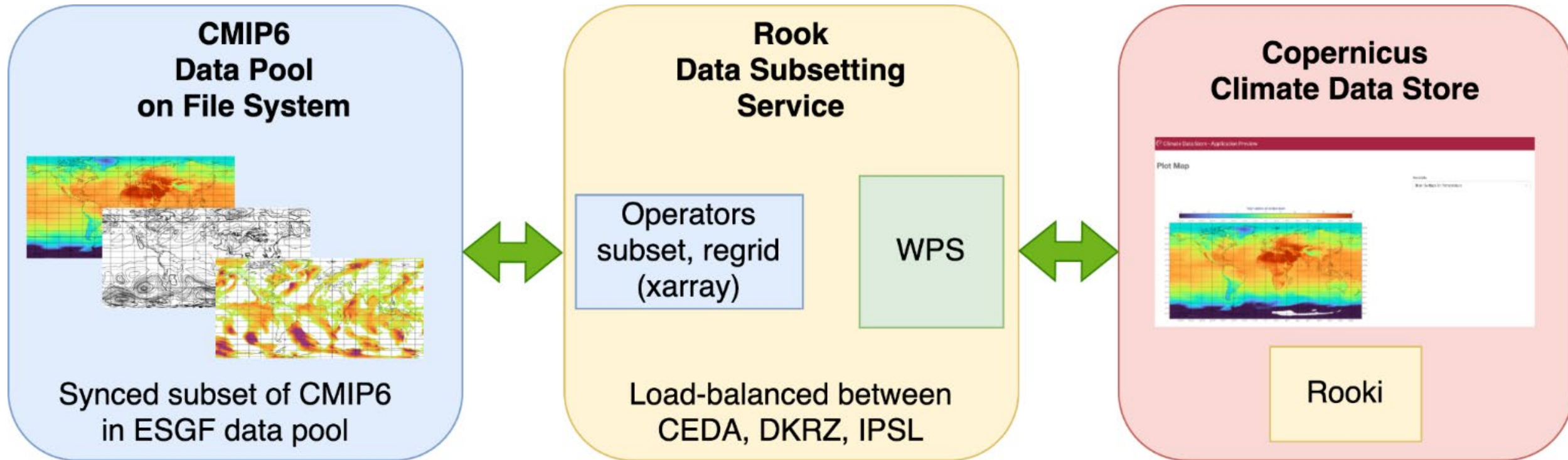
- Good Integration of analytics services with data catalogs is missing
  - changing data pool content not reflected in cross site catalogs (e.g. ESGF)
  - automatic institutional catalog generation vs. manual ESGF publication
  - also hinders cross-institutional data analytics workflows

→ future collaboration on Intake/STAC exploitation “beyond” ESGF
- Exploitation of cloud storage in analytics services:
  - currently local, institutional first steps
  - ENES-CDI collaboration important as well as coordination with European infras (e.g. EOSC, EGI, ..) and broader community efforts (Pangeo)

Add on slides ...

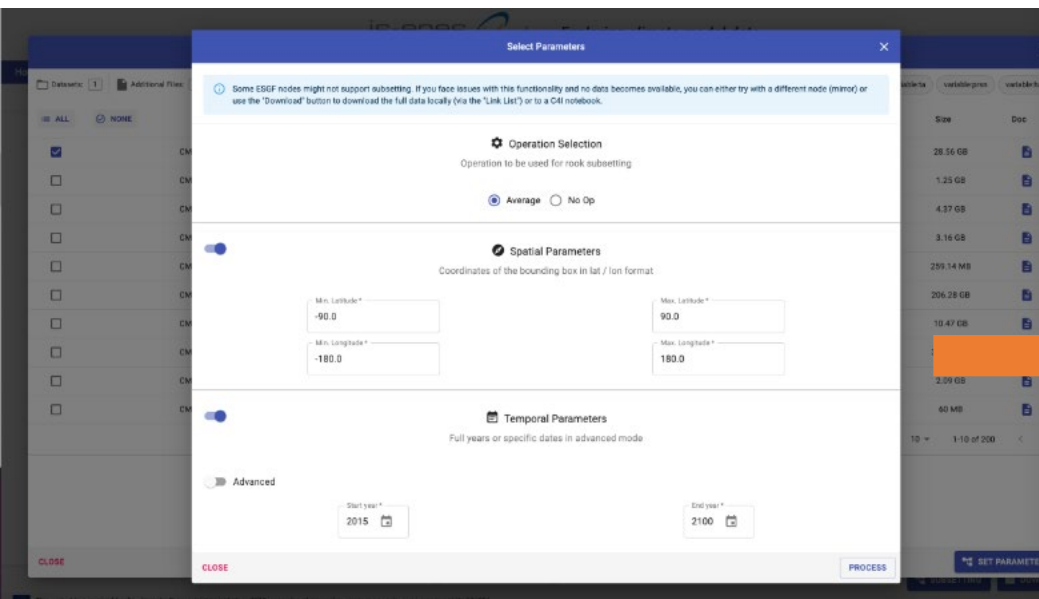
# Copernicus Climate Data Store

Rook WPS used by CDS to access CMIP and CORDEX

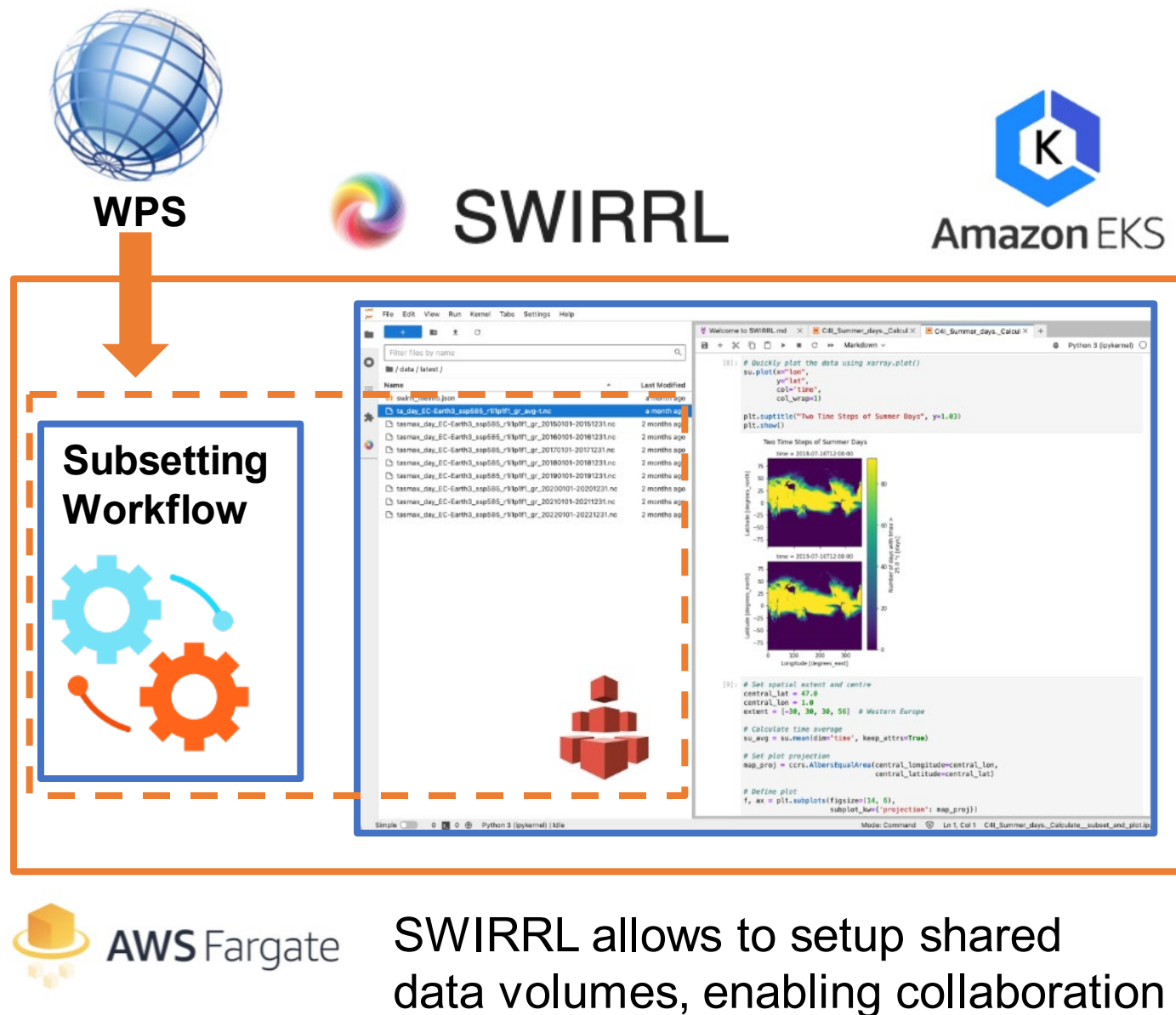


# Climate4Impact Workspaces - WPS Integration

Data Selection and Subsetting configuration enabled for ESGF/WPS nodes



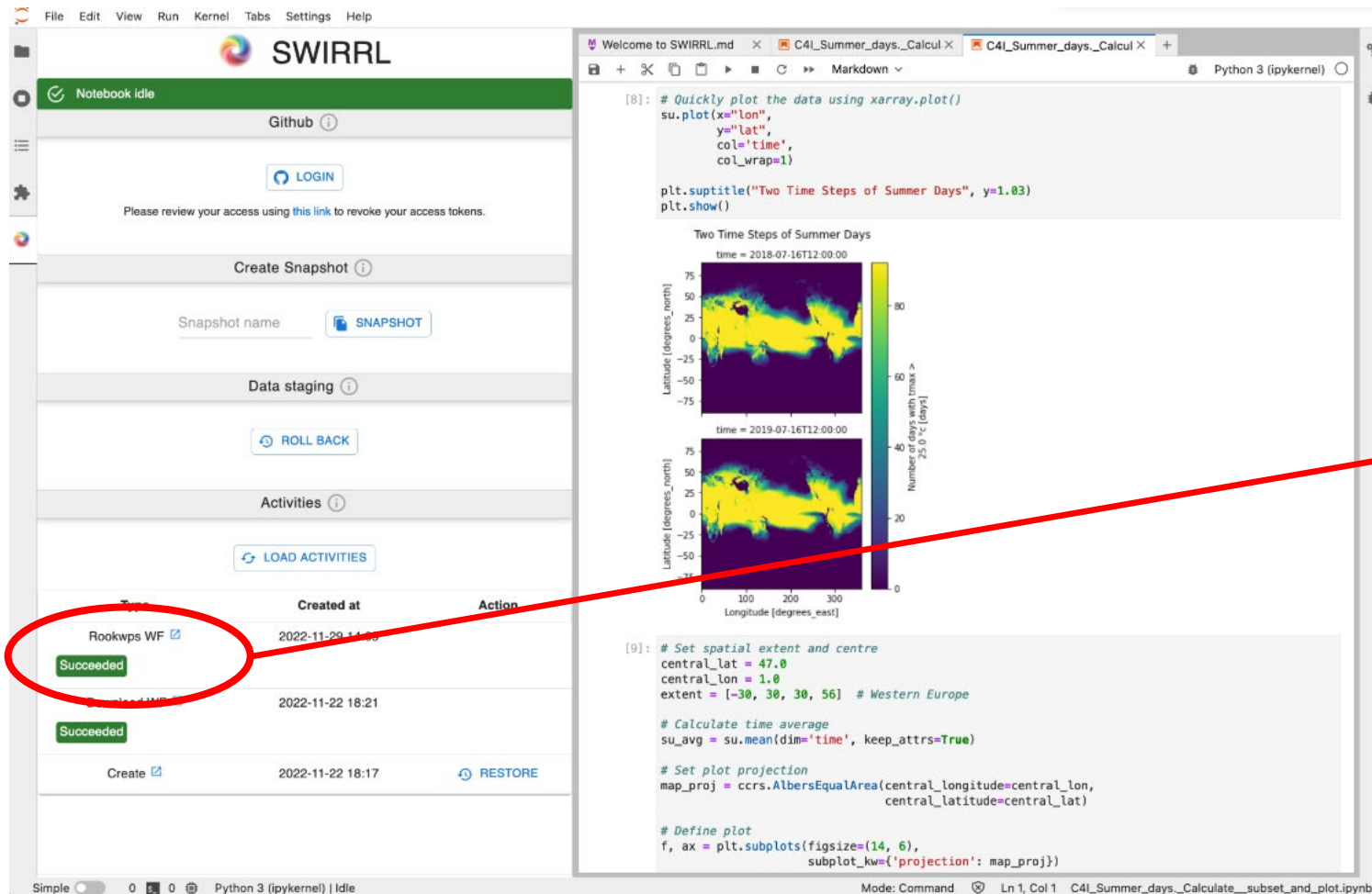
subsets are staged to the user workspace



SWIRRL allows to setup shared data volumes, enabling collaboration



# Climate4Impact Workspaces - WPS Integration



The screenshot shows the SWIRRL interface with a notebook titled 'C4I\_Summer\_days\_Calcul'. The notebook contains two heatmaps showing the number of days with a temperature above 20°C for the years 2018 and 2019. The first heatmap is for 2018-07-16T12:00:00 and the second is for 2019-07-16T12:00:00. The y-axis is Latitude (degrees\_north) and the x-axis is Longitude (degrees\_east). A red circle highlights the 'Rookwps WF' activity in the 'Activities' table, which shows a 'Succeeded' status. The table also includes columns for 'Created at' and 'Action'.

Activity	Created at	Action
Rookwps WF	2022-11-29 14:00	Succeeded
Calculate WF	2022-11-22 18:21	Succeeded
Create	2022-11-22 18:17	RESTORE



```
{
  "prov:used": [...],
  "provone:hadPart": [
    {
      "prov:wasAssociatedWith": [...],
      "@type": [
        "Resource",
        "prov:Activity",
        "provone:Execution"
      ],
      "rdfs:label": "orchestrate",
      "prov:startedAtTime": "2022-11-29T14:13:02Z",
      "@id": "urn:roocs:orchestrate_16ca3e1f-fee6-4419-a264-65d273a801bf",
      "prov:endedAtTime": "2022-11-29T14:13:45Z"
    },
    {
      "prov:wasAssociatedWith": [
        { ... }
      ],
      "prov:wasActivityOfInfluence": [...],
      "@type": [
        "Resource",
        "prov:Activity",
        "provone:Execution"
      ],
      "rdfs:label": "average_ta_1",
      "roocs:apply_fixes": false,
      "@id": "urn:roocs:average_ta_1_6eabac6b-444e-4e9a-a71b-bcc1b16b9fb1",
      "roocs:dims": "time"
    }
  ],
  "@type": [...],
  "@context": { ... },
  "prov:generated": { ... },
  "swirrl:sessionId": "13abfb95-fd4a-463f-ba61-fc96d40a9b6c",
  "swirrl:message": "Succeeded",
  "prov:wasAssociatedWith": [...],
  "@id": "urn:uuid:49f0d7f5-5cc8-4893-8e17-2a18ed870540",
  "swirrl:jobId": "49f0d7f5-5cc8-4893-8e17-2a18ed870540",
  "prov:endedAtTime": "2022-11-29T13:13:54.220Z",
  "prov:atLocation": "POST /workflow/rookwps/run/",
  "prov:startedAtTime": "2022-11-29T13:09:03.676Z"
}
```

Thanks to the use of PROV by Rooks and SWIRRL, provenance is easily merged, stored and made available to users and machines (interoperable)



# Climate4Impact Workspaces - on-demand calculations - icclim

Workflow Monitoring

GitHub Authentication

Snapshot Controls

Data Staging Rollback

Activities History and Provenance

Notbook idle

GitHub

LOGIN

Please review your access using [this link](#) to revoke your access tokens.

Create Snapshot

Snapshot name

Data staging

Activities

Type	Created at	Action
Library Update	2021-06-15 16:00	<input type="button" value="RESTORE"/>
<input type="text" value="pip install xarray"/>		
Workflow	2021-06-09 12:51	
Workflow	2021-06-09 12:31	
Workflow	2021-06-09 12:17	
Snapshot	2021-06-09 10:39	<input type="button" value="OPEN"/>

Simple ☐ 1 ☐ 0 Python 3 | idle

Welcome to SWIRRL.md

C4I\_Averaged\_Temperature

C4I\_Summer\_days\_Calculs

Terminal 1

# Contour filled colors

p = su\_avg.plot.contourf(levels=levels,

cnorm='RdBu\_r',

extend='both',

transform=ccrs.PlateCarree

# Plot information

plt.suptitle("Two Time Steps of Europe Summer Days",

# Add the coastlines to axis and set extent

ax.coastlines()

ax.gridlines()

ax.set\_extent(extent)

# Save plot as png

plt.savefig('c4i\_su\_contours\_icclim.png')

Two Time Steps of Europe

height = 2.0, spatial\_ref = 0

IS-ENES Climate Data Infrastructure for Climate 4 Impact > C4I Use Cases as Jupyter Notebooks

C4I Use Cases as Jupyter Notebooks

Project ID: 25761638 [Request Access](#)

13 Commits 1 Branch 0 Tags 1.5 MB Files 1.5 MB Storage

A collection of Jupyter Notebooks implementing some Use Cases.

master notebooks /

Some small fixes. Added deltaT\_deltaP Notebook. Tested also with icclim v5.0.0-b3.

Christian Page authored 2 days ago

7d663d8e

Name	Last commit	Last update
C4I_Averaged_Temperature_An...	Some small fixes. Added deltaT_deltaP Not...	2 days ago
C4I_Summer_days_Calculate_...	Some small fixes. Added deltaT_deltaP Not...	2 days ago
C4I_deltaT_deltaP_Anomaly_20...	Some small fixes. Added deltaT_deltaP Not...	2 days ago
README.md	small readme and notebook edits	4 months ago

<https://gitlab.com/is-enes-cdi-c4i/notebooks>

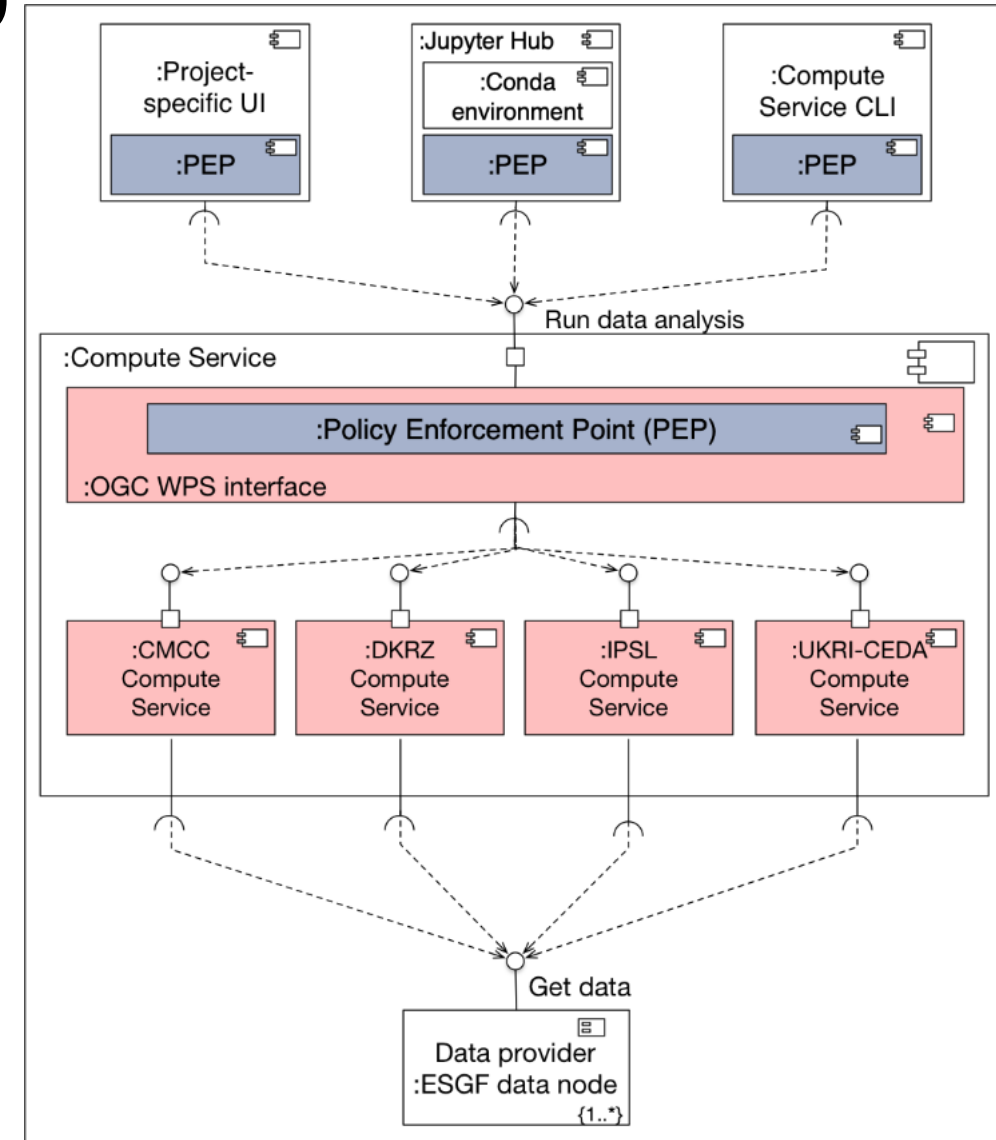
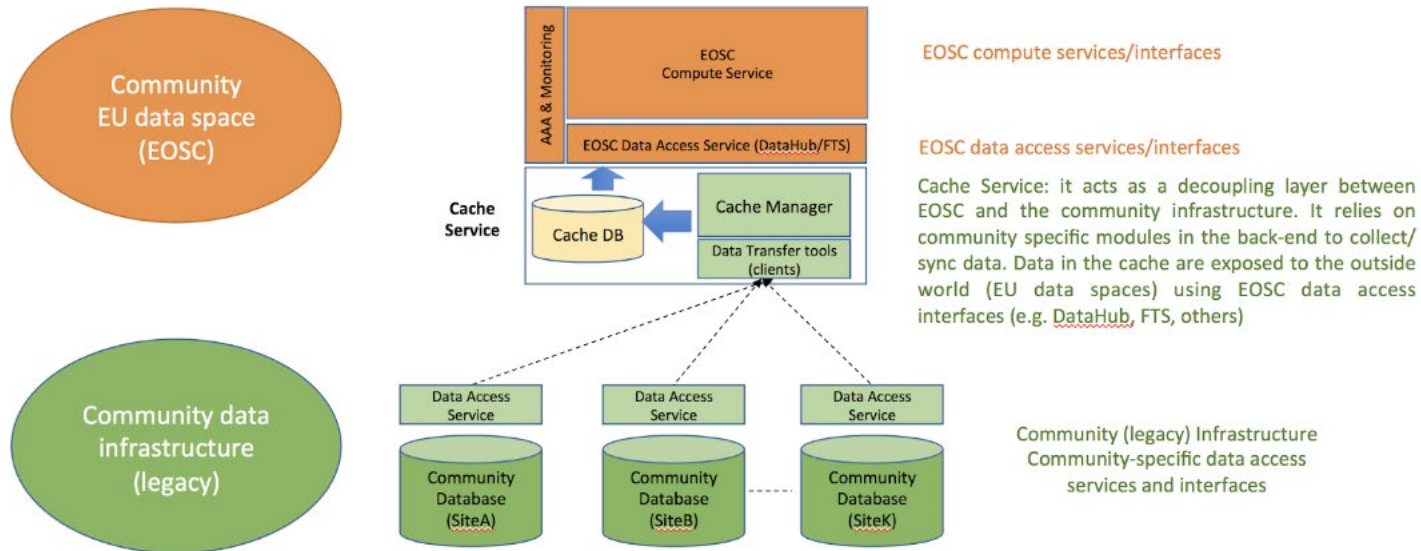
# processing service roadmap

## Milestones and Deliverables

<https://docs.google.com/document/d/1gvcRyeAvO10rk-sfzo0kfdmGW-BfMAud/edit#>

<https://marketplace.eosc-portal.eu/services/enes-data-space>

<https://docs.google.com/document/d/1y1izcPtSnCv8HIJJa9g8BJU8orzK-sn/edit>



## THE CONSORTIUM

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



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



Our website  
<https://is.enes.org/>



Follow us on Twitter !  
**@ISENES\_RI**



Contact us at  
[is-enes@ipsl.fr](mailto:is-enes@ipsl.fr)



Follow our channel  
**IS-ENES3 H2020**