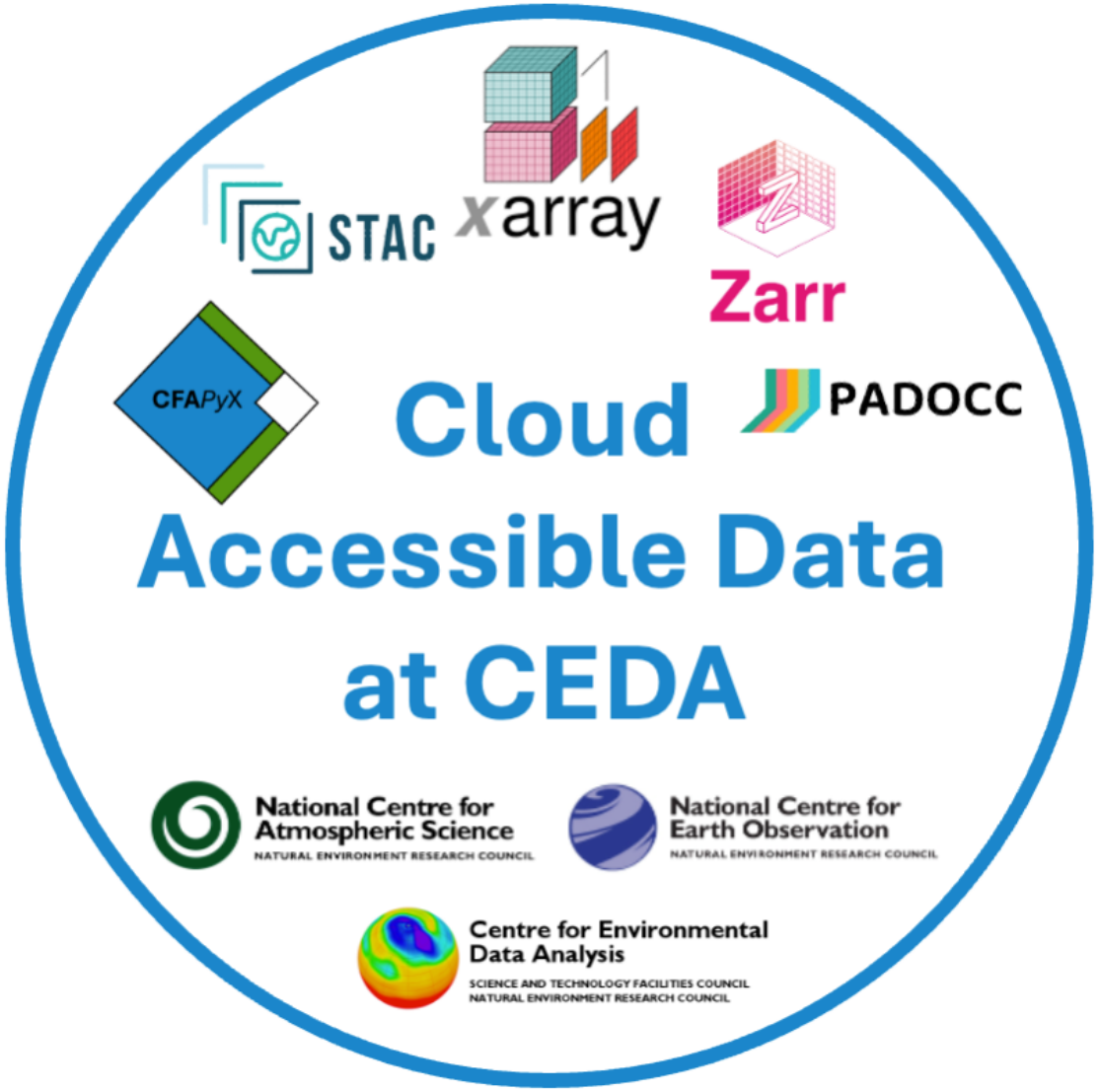
PADOCC

"Pipeline to Aggregate Data for Optimised Cloud Capabilities"

- Scalable pipeline for conversion to Kerchunk/Zarr
- Product Validation
- Attribute/Metadata correction.



CEDA-DataPoint

"Access point to CEDA STAC Collections and Cloud products"

- STAC API uses pystac-client.
- Lazily loaded metadata/references
- Abstracted access to datasets (configuration handled by DataPoint)

Daniel Westwood (daniel.westwood@stfc.ac.uk)

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

PADOCC

```python
infile  = 'padocc/tests/data/myfile.csv'
# Input CSV has Identifier, Path/To/Datasets, {updates}, {removals}

groupID = 'padocc-test-suite'
workdir = '/home/username/padocc-workdir'

mygroup = GroupOperation(
    groupID,
    workdir=workdir,
    label='test_group'
)

mygroup.init_from_file(infile)

mygroup.run('compute', mode='kerchunk')
```

- Supports groups of datasets (N files in each dataset)
- Perform operations on all members of a group (each member has a project code [proj_code])
- Configure parallel deployment to SLURM (batch job manager)
- Python interface or CLI entrypoints available.
- ‘Group’ object – can access any files generated during processing (logs/cache files/scan results)

Daniel Westwood (daniel.westwood@stfc.ac.uk)

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

CEDA-DataPoint

- Python package installable with pip (currently at v0.2.0)

```
pip install ceda-datapoint
```

- DataPointClient configured for read-only access to CEDA STAC Collections.
- `*search*` uses same syntax as pystac for running queries.
- Options to group `*cloud assets*` from a search into a `*cluster*` of datasets.
- Able to open with simple `*open_dataset*` method – returns an Xarray Dataset.

```
from ceda_datapoint import DataPointClient

client = DataPointClient(org='CEDA') # All public connection kwargs are known by default for CEDA
client
```

```
<DataPointClient: CEDA-DP:nyxeq2wt78>
```

With the DataPointClient a user can find the collections offered by the organisation/STAC API and view the search terms for each collection (or just for one collection)

```
client.list_collections()
```

```
cci: cci
cmip6: CMIP6
cordex: CORDEX
eocis-sst-cdrv:
eocis-sst-cdrv:
land_cover: La
sentinel1: Sen
sentinel2_ard:
ukcp: UKCP
```

```
ds = cluster[0].open_dataset()
ds
```

✓ 2.1s

```
<frozen importlib._bootstrap>:241: RuntimeWarning: numpy.ndarray size changed, may indicate binary ir
```

```
search = client
search
```

✓ 0.0s

```
<DataPointSearch:
```

```
cluster = sear
cluster
```

✓ 0.0s

```
<DataPointCluster:
```

xarray.Dataset

▶ Dimensions: (**time**: 251288, axis_nbounds: 2, **lat**: 128, **lon**: 256)

▼ Coordinates:

time	(time)	float64	6.027e+04 6.027e+04 ... 9.168e+04
lat	(lat)	float64	-88.93 -87.54 ... 87.54 88.93
lon	(lon)	float64	0.0 1.406 2.812 ... 357.2 358.6

▼ Data variables:

time_bounds	(time, axis_nbounds)	float64	...
height	()	float64	...
huss	(time, lat, lon)	float32	...

▶ Indexes: (3)

▶ Attributes: (52)

Daniel Westwood (daniel.westwood@stfc.ac.uk)

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

Documentation Links

PADOCC

- Github: https://github.com/cedadev/padocc (*UNDER CONSTRUCTION - watch for v1.3 release*)
- Sphinx Docs: https://cedadev.github.io/padocc/

CEDA-DataPoint

- Github: https://github.com/cedadev/DataPoint (v0.2.0)
- Sphinx Docs: https://cedadev.github.io/DataPoint/

Daniel Westwood (daniel.westwood@stfc.ac.uk)

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