Committee on Earth Observation Satellites

Jupyter Notebooks for Capacity development

Esther Conway (NCEO/CEDA), Ag Stephens (NCEO/UKSA), Richard Smith (NCEO/UKSA), Kenton Ross (NASA), Lauren Childs-Gleason (NASA), Yousuke Ikehata (JAXA), Makoto Natsuisaka (JAXA), Matt Paget (CSIRO), Brian Killough (SEO), Rizvi Syed (SEO), Damiano Guerruci (ESA), Giuseppe Troina (ESA) and Julia Wagemann (EUMESAT)

ESA Living Planet Thursday 26th May 2022

Jupyter Notebook Collaboration ?

What is a Jupyter Notebook?

A Jupyter Notebooks allow you to combine

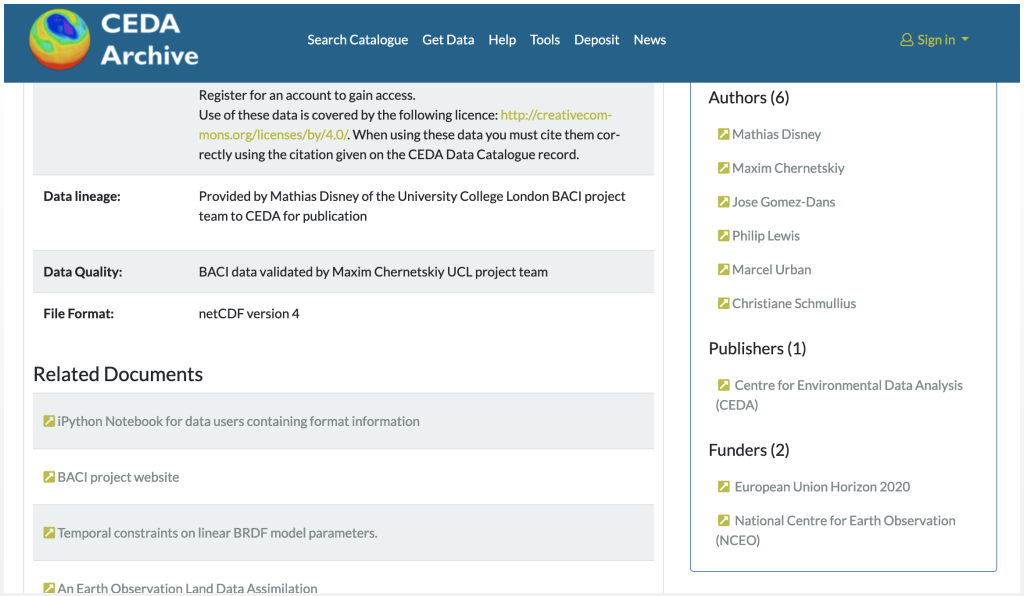
- **rich documentation**
- **live and adaptable code**
- **data visualizations**

As a tool to share your data analysis with others, collaborate, teach, and promote reproducible science.

The Jupyter notebook began as an (IPython Notebook), this use of Python makes it ideal starting point for many people using EO data for the first time. In addition it currently supports around 40 programming languages, including Python, R and Julia (Ju-pyt-R).

Where can they be deployed and why is this important for the CEOS Community ?

- Standalone computer notebooks can be supplied alongside the relevant data
- In classroom/institutional setting using Jupyter Hub adopted by University Data Science Departments
- Data Analysis platforms (JASMIN, Earth Analytics Interoperability Lab)
- As a centralized web services using a Data Cubes etc (SEO Google Earth Engine, ESA PDGS data cube, Digital Earth Africa)

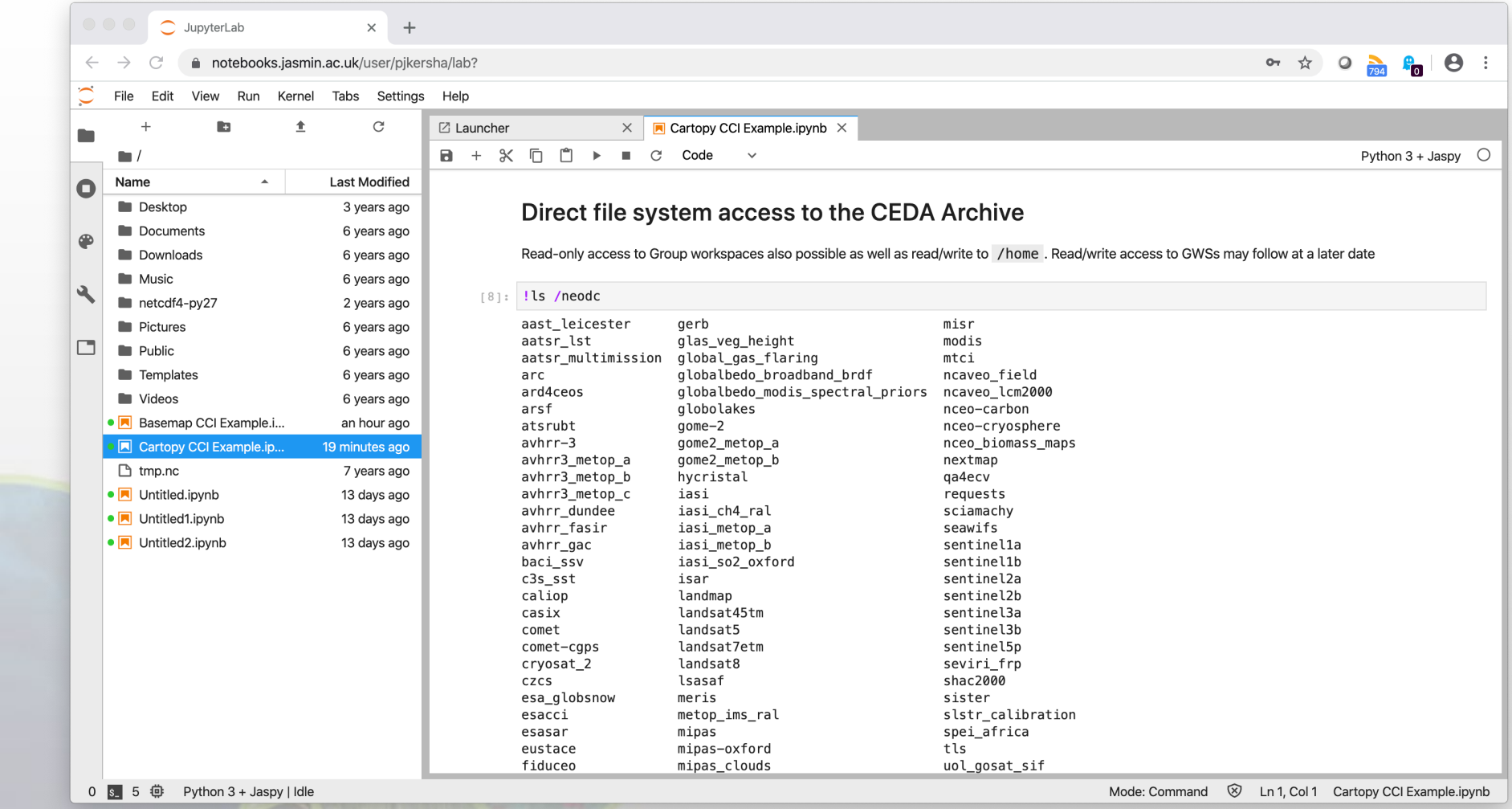


University of Reading

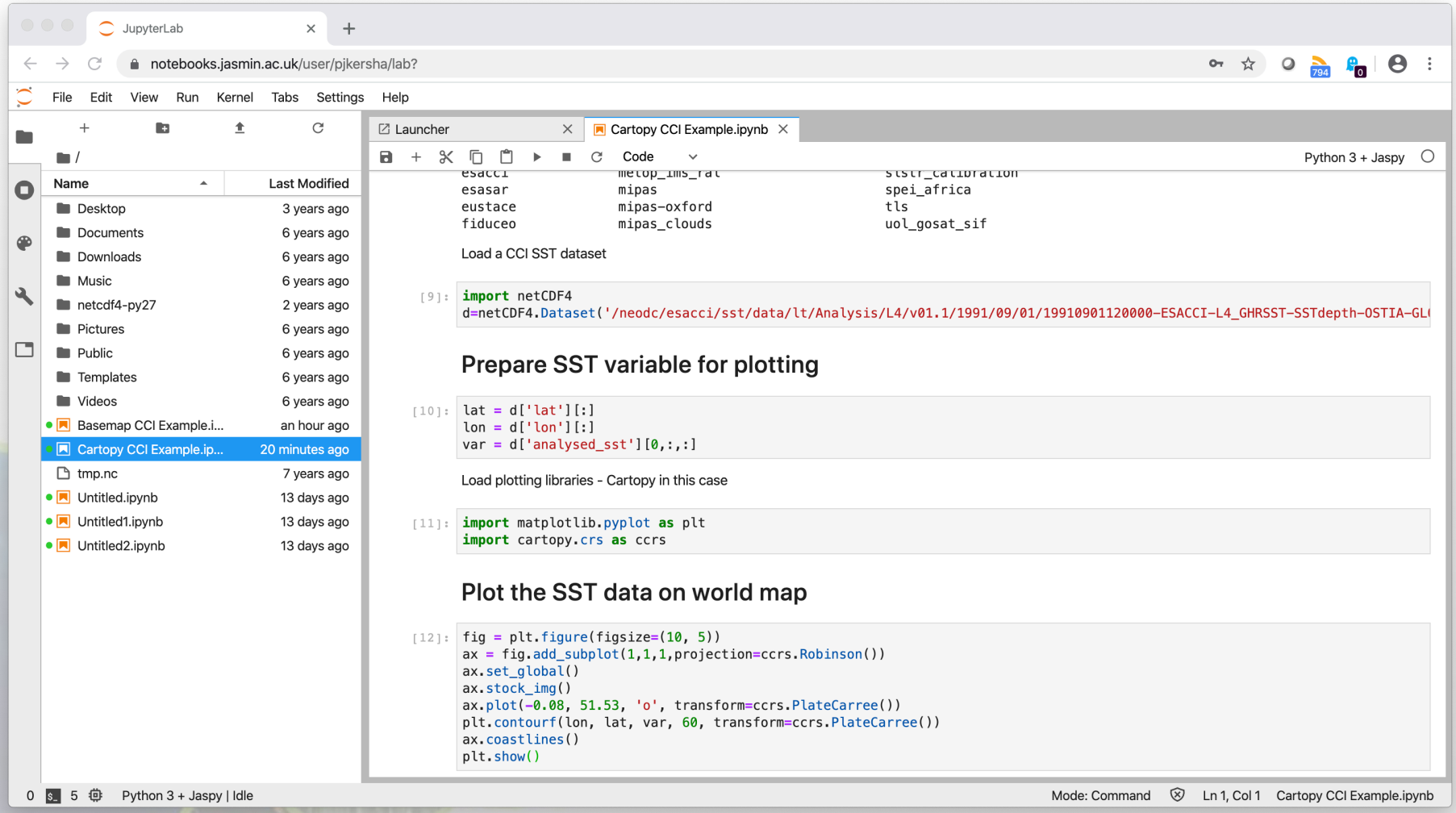
Google Earth Engine

CEOS
Working Group on Information Systems and Services

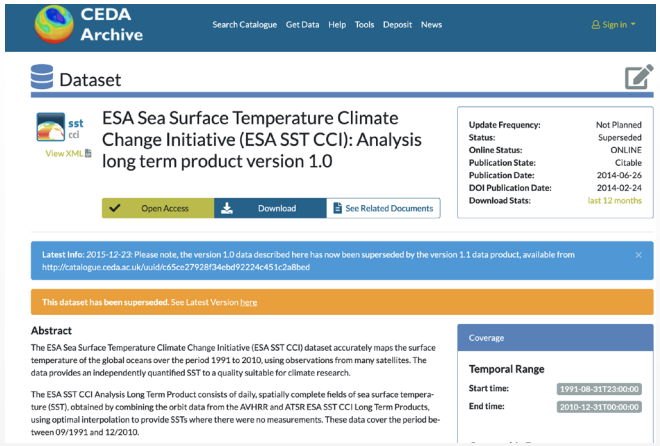
Access to Big Data
Volume, Velocity, Veracity, Variety and Value



Lowers barriers and support the development of important data science skills



Integrate with Support Materials



https://catalogue.ceda.ac.uk/uuid/916986a220e6bad55411d9407ade347c

matplotlib

https://matplotlib.org/stable/index.html

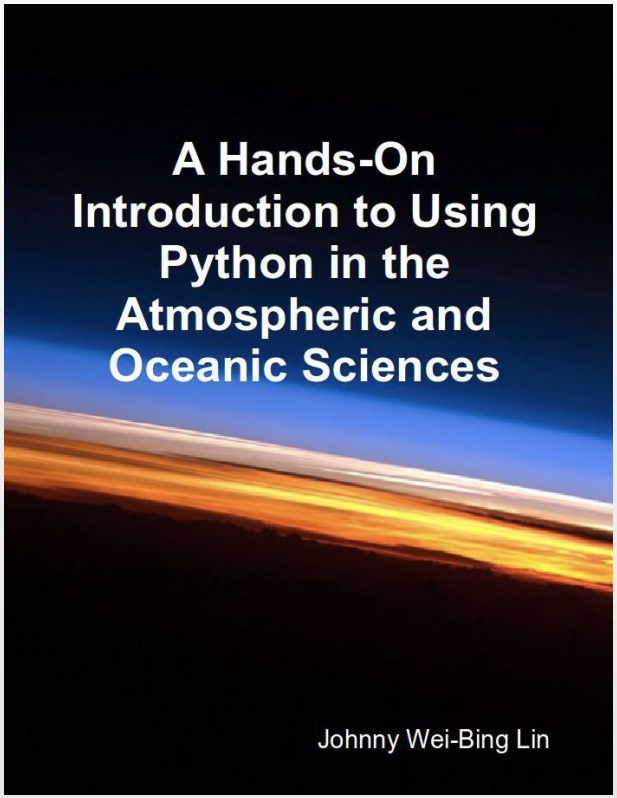
Cartopy

https://scitools.org.uk/cartopy/docs/latest/

https://rasterio.readthedocs.io/en/latest/

software carpentry

https://software-carpentry.org/



http://www.johnny-lin.com/pyintro

Users are able to obtain meaningful results quickly

JupyterLab

notebooks.jasmin.ac.uk/user/pjkersha/lab?

File Edit View Run Kernel Tabs Settings Help

/

Name	Last Modified
Desktop	3 years ago
Documents	6 years ago
Downloads	6 years ago
Music	6 years ago
netcdf4-py27	2 years ago
Pictures	6 years ago
Public	6 years ago
Templates	6 years ago
Videos	6 years ago
Basemap CCI Example.i...	an hour ago
Cartopy CCI Example.ip...	20 minutes ago
tmp.nc	7 years ago
Untitled.ipynb	13 days ago
Untitled1.ipynb	13 days ago
Untitled2.ipynb	13 days ago

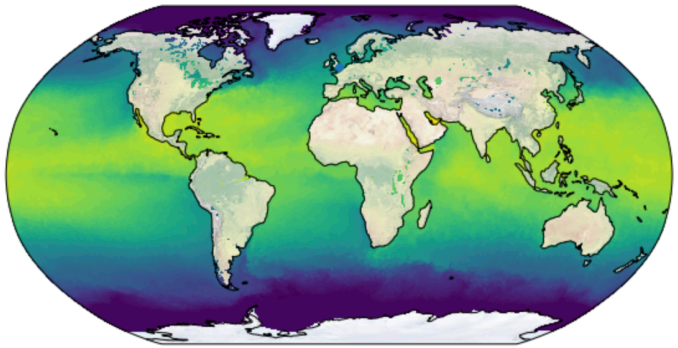
Launcher | Cartopy CCI Example.ipynb

Code — Python 3 + Jaspy

[11]:
```
import matplotlib.pyplot as plt
import cartopy.crs as ccrs
```

Plot the SST data on world map

[12]:
```
fig = plt.figure(figsize=(10, 5))
ax = fig.add_subplot(1,1,1,projection=ccrs.Robinson())
ax.set_global()
ax.stock_img()
ax.plot(-0.08, 51.53, 'o', transform=ccrs.PlateCarree())
plt.contourf(lon, lat, var, 60, transform=ccrs.PlateCarree())
ax.coastlines()
plt.show()
```



[ ]:

0 5 Python 3 + Jaspy | Idle — Mode: Command — Ln 1, Col 1 — Cartopy CCI Example.ipynb

Technical Arc of Data Science Skills

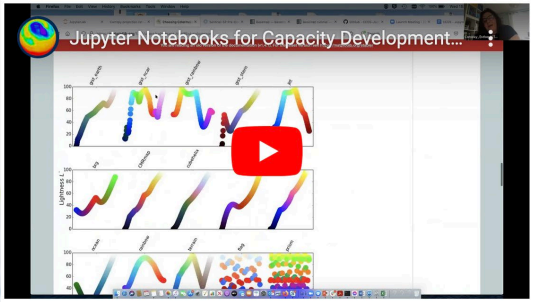
https://ceos.org/meetings/jupyter-notebooks-for-capacity-development-webinar/

CEOS / Meetings / Jupyter Notebooks for Capacity ...

Jupyter Notebooks for Capacity Development Webinar

Event Dates: July 21st - 22nd, 2021

Webinar Recording

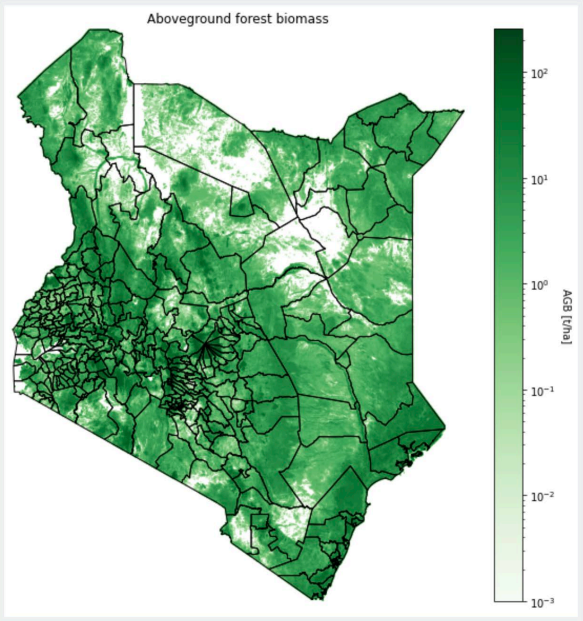


Webinar Feedback Survey

A big thank you to all our webinar participants. To assess the quality and impact of your webinar experience and make future webinars even better, please fill out our quick Webinar Feedback Survey to let us know how we did.

Overview

The CEOS Working Group on Capacity Building and Data Democracy and the Working Group on Information Systems and Services ran a joint webinar on Jupyter Notebooks for Capacity Development. The aim of this webinar was to introduce space agencies and environmental organisations worldwide to Jupyter Notebooks and take a tour of emerging services from CEOS Agencies and their applications. We



This image is above-ground biomass in Kenya in tonnes per hectare. The data is on the CEDA

- Overview of what a Jupyter Notebook is and the benefits of using one
- Demo a simple example of plotting sea surface temperature data, navigating an archive, and adjusting the colour scale
- Describe open-source Jupyter Notebook resources and different platform and environment types
- Discuss how collaborative research and activities such as international Hackathons can also be supported

Jupyter Hub and Notebooks on Data Analysis Platforms

We looked at two examples from the UK's JASMIN Jupyter Notebook service, which can access over 20 petabytes of data on the CEDA archive. We then explored the Sentinel 5p global archive of data and demonstrated how to use a very basic Notebook to use the data and answer valuable questions, e.g. how did pollution levels change in large cities during the Covid-19 pandemic? We also looked at a smaller scale specialist example, regional NCEO biomass maps. This session helped to demonstrate how, in addition to helping users use Jupyter Notebooks to obtain domain-specific information from data, we can also help them learn technical knowledge and skills related to libraries, modules, and shape files.

Open Data Cube and Google Earth Engine

The Open Data Cube (ODC) Google Sandbox is a free and open programming interface that connects users to Google Earth Engine datasets. This open-source tool allows users to run Python application algorithms using Google's Colab Notebook environment. This demonstration showed two examples of Landsat applications focused on scene-based cloud statistics and historic water extent. Basic operation of the tool will support unlimited users for small-scale analyses and training but can also be scaled in size and scope with Google Cloud resources to support enhanced user needs.

ESA PGDS Data Cube and Time Series Data

The ESA PDGS Data Cube is a pixel-based access service that enables human and machine-to-machine interfaces for Heritage Missions (HM), Third-Party Missions (TPM) and Earth Explorer (EE) datasets. The pixel-based access service provides the users with advanced retrieval capabilities, such as time series extraction, data subsetting, mosaicking, band combinations, and index generation (e.g. NDVI, anomalies, and more) directly from the EO-SIP packages with no need for data duplication or data preparation.

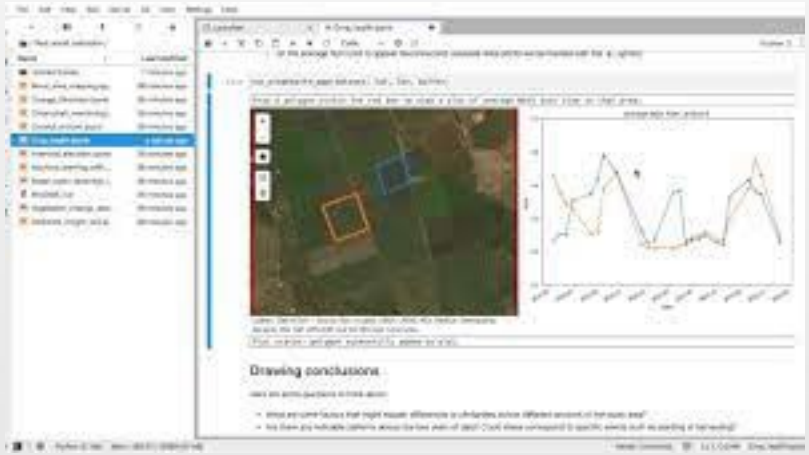
In addition to the web-based Explorer graphic user interface, the ESA PDGS Data Cube service also provides a Jupyter processing environment to allow users to import, write, and execute code that runs close to the data. This demonstration showcased how to retrieve Soil Moisture time-series using the Jupyter environment in order to generate thematic maps (monthly anomalies map) over an area of interest. The benefit of using the pixel-based service with respect to traditional access services in term of resources usage will be also highlighted.

Earth Analytics and Interoperability Lab – Big Data Processing

The CEOS Earth Analytics Interoperability Lab (EAIL) is a platform for CEOS projects to test interoperability in a live Earth Observation (EO) ecosystem. EAIL is hosted on Amazon Web Services and includes facilities for Jupyter Notebooks, scalable compute infrastructure for integrated analysis, and data pipelines that can connect to new and existing CEOS data discovery and access services. This demonstration showed how we use Jupyter Notebooks with the Python Dask Library to efficiently compute and perform large-scale analyses (10s GB) with interactive plotting and scalable compute resources in EAIL.

Expert Panel

- Edward Boamah (Digital Earth Africa)
- Hayley Evers-King (EUMETSAT)
- Sean McCartney (NASA DEVELOP)
- Franz Meyer (University of Alaska Fairbanks)



Why would we create a best practice document

Who will benefit form a Best Practice

- Data Producers
- Authors of Jupyter Notebooks
- Providers of EO data training
- Users of EO data
- EO data archives
- Providers of Data Analysis Infrastructure

JUPYTER NOTEBOOK BEST PRACTICE CONTENT

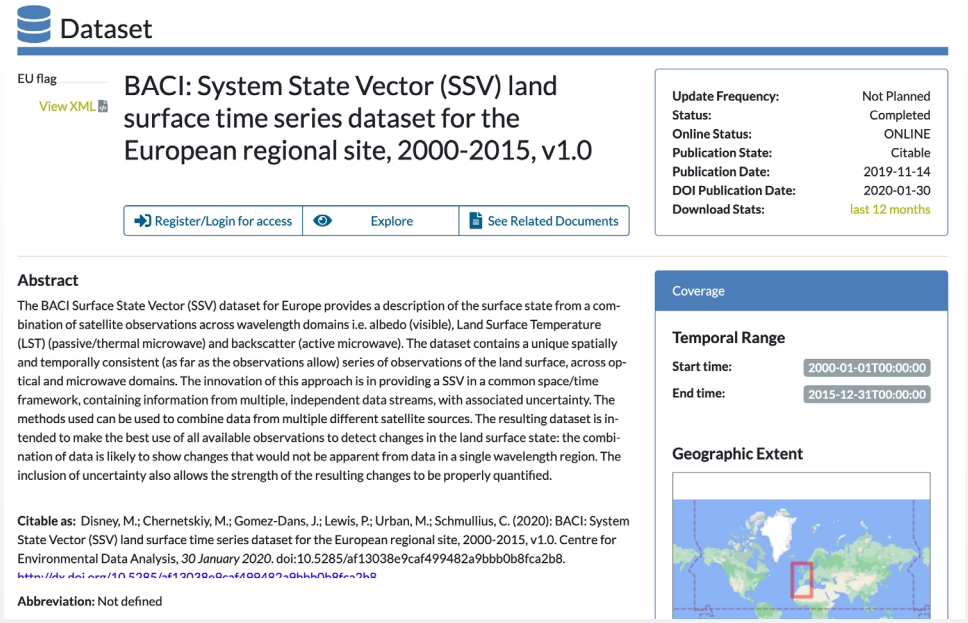
4.1 General Recommendations
4.2 Notebook description and function
4.3 Structure, workflow and documentation
4.4 Technical dependencies and Virtual Environments
4.6 Citation of input data and data access
4.7 Association with archived data
4.8 Incorporation with data cubes
4.9 Version control, preservation and archival
4.10 Open source software licensing
4.12 Publishing software and getting a DOI
4.12 Interoperability and reuse on alternate platforms
4.13 Creating a binder deployment

Developing a Jupyter Notebooks Best Practice

Structure, workflow and documentation

Encourage the development of good workflow and structure within notebooks along with quality documentation. Supporting their reuse and adaptation by new users

- Good software engineering principles
- Use of Functions
- Use of Modules
- Size of Notebook
- Capturing Output



Dataset

EU flag

View XML

BACI: System State Vector (SSV) land surface time series dataset for the European regional site, 2000-2015, v1.0

Register/Login for access | Explore | See Related Documents

Update Frequency:	Not Planned
Status:	Completed
Online Status:	ONLINE
Publication State:	Citable
Publication Date:	2019-11-14
DOI Publication Date:	2020-01-30
Download Stats:	last 12 months

Abstract

The BACI Surface State Vector (SSV) dataset for Europe provides a description of the surface state from a combination of satellite observations across wavelength domains i.e. albedo (visible), Land Surface Temperature (LST) (passive/thermal microwave) and backscatter (active microwave). The dataset contains a unique spatially and temporally consistent (as far as the observations allow) series of observations of the land surface, across optical and microwave domains. The innovation of this approach is in providing a SSV in a common space/time framework, containing information from multiple, independent data streams, with associated uncertainty. The methods used can be used to combine data from multiple different satellite sources. The resulting dataset is intended to make the best use of all available observations to detect changes in the land surface state: the combination of data is likely to show changes that would not be apparent from data in a single wavelength region. The inclusion of uncertainty also allows the strength of the resulting changes to be properly quantified.

Citable as: Disney, M.; Chernetskiy, M.; Gomez-Dans, J.; Lewis, P.; Urban, M.; Schmullius, C. (2020): BACI: System State Vector (SSV) land surface time series dataset for the European regional site, 2000-2015, v1.0. Centre for Environmental Data Analysis, *30 January 2020*. doi:10.5285/af13038e9caf489482a9bbb0b8fca2b8.

Abbreviation: Not defined

Coverage

Temporal Range

Start time: 2000-01-01T00:00:00

End time: 2015-12-31T00:00:00

Geographic Extent

```
1. Opening and reading
1.1 Optical data
1.2 Land Surface Temperature (LST)
1.3 Synthetic Apperture Radar (SAR) backscatter
2. Reprojection
3. Principal Componenet Analysis (PCA)
4. Clustering
4.1 Red and NIR
4.2 Red, NIR and LST
4.3 Red, NIR, LST and micowave
```

Technical Dependencies and Virtual Environments

- Language and versions
- Libraries
- Other Code
- Virtual Machines on Data Analysis Platforms
- Installing libraries from within notebook, recommending conda etc.
- Setting up virtual environments
- Testing

```
import numpy as np
import netCDF4 as nc
import seaborn as sns
from scipy import stats
import pandas as pd
import os
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
warnings.simplefilter('ignore')
```

baci-data-processor.ipynb	restructure	15 months ago
grid_lines.py	restructure	15 months ago

Citation of input data and data access

Currently there is no guidance on how a Jupyter notebook should

- **Cite input data**
- **Explain access requirements i.e. is it open or do need to apply for access**
- **Location and structure of data**

Citable as: Disney, M.; Chernetskiy, M.; Gomez-Dans, J.; Lewis, P.; Urban, M.; Schmullius, C. (2020): BACI: System State Vector (SSV) land surface time series dataset for the European regional site, 2000-2015, v1.0. Centre for Environmental Data Analysis, *30 January 2020*. doi:10.5285/af13038e9caf499482a9bbb0b8fca2b8. http://dx.doi.org/10.5285/af13038e9caf499482a9bbb0b8fca2b8

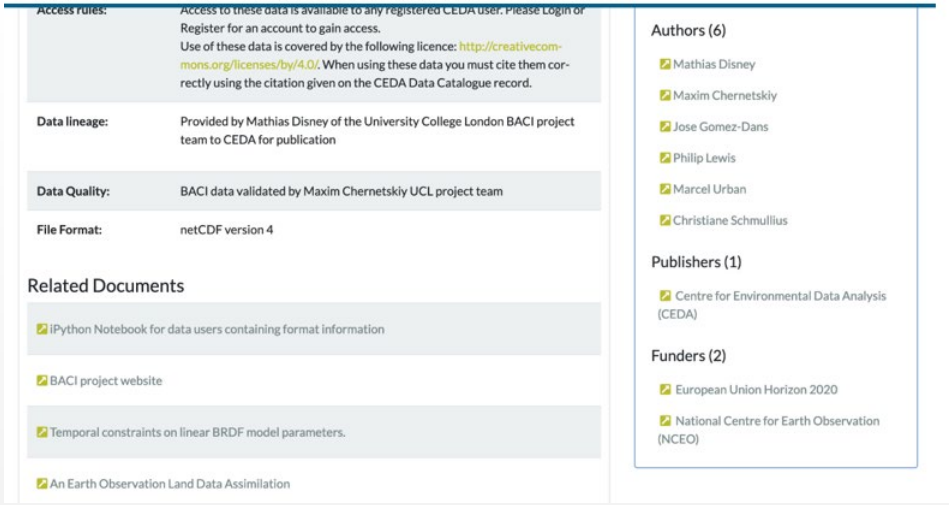
This document explains basics of opening and manipulation of the BACI Surface State Vector (SSV) files in netCDF format. SSV consists of surface reflectance, albedo, land surface temperature (LST) and synthetic apperture radar (SAR) backscatter. All datasets have the same spatial resolution, geographical projection and temporal step.
Presented examples are based on the following python 2.7 libraries:
netCDF4 - working with python netCDF;
GDAL - Geospatial Data Abstraction Library;
seaborn - enhanced data visualization and
scikit-learn - machine learning library.

Datasets can be found in */group_workspaces/jasmin2/baci/sigil/baci_wp2_files/*
All regional sites which are in the geographical Europe are in the same folder */group_workspaces/jasmin2/baci/sigil/baci_wp2_files/13_europe/*
Datasets are in sinusoidal projection and divided by MODIS tiles https://modis-land.gsfc.nasa.gov/MODLAND_grid.html This means that processed regions are larger than BACI regional sites.
Table of Contents:

Association with archived data

There are many issue around how software should be linked to archived data and how permanent. This should be ? Areas for discussion include:

- **Research Objects**
- **AIP's**
- **Part of the catalogue record**
- **Deprecation and update**

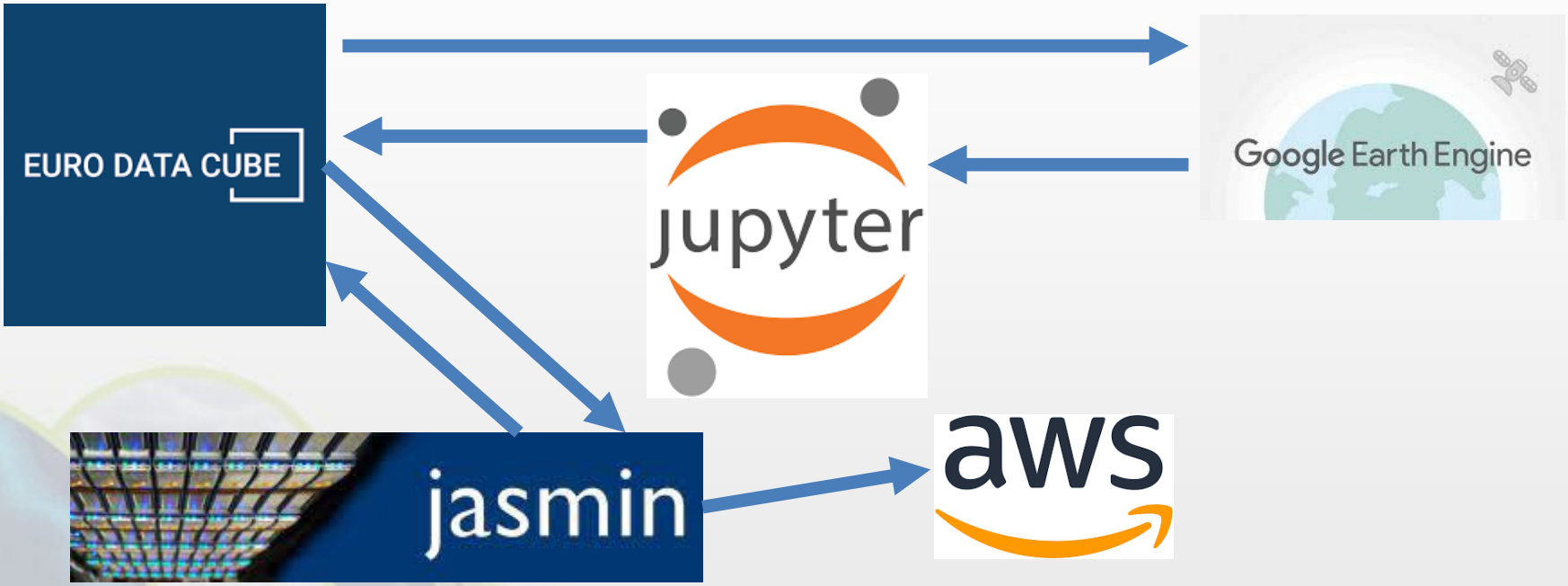



```
"metadata": {
"title": "Discover Sentinel-3 OLCI Level 1 data",
"description": "This notebook introduces you to ...",
"author": "Julia Wagemann",
"image": "./img/img_04.png",
"link": "link_to_jupyterhub/notebook.ipynb",
"github": "link_to_github/notebook.ipynb",
"tags": {
"domain": "atmosphere",
"variable": "toa_reflectance",
"satellite": "Sentinel-3",
"sensor": "OLCI",
"level": "L1"}
```

Incorporation with data cubes

- Establishing Connections with Data Cubes
- Defining what data gets loaded
- Working with xarray
- Porting Notebook from hierarchical file system to a data cube

Interoperability and reuse on alternate platforms



Creating a binder or Colab deployment

What is a Binder?

A Binder (also called a Binder-ready repository) is a code repository that contains at least two things:

Code or content that you'd like people to run. This might be a Jupyter Notebook that explains an idea, or an R script that makes a visualization.

Configuration files for your environment. These files are used by Binder to build the environment needed to run your code. Configuration files may be placed in the root of your repository or in a binder/ folder in the repository's root (i.e. myproject/binder/).

A Binder repository can be built by a BinderHub, which will generate a link that you can share with others, allowing them to interact with the content in your repository.

Jupyter Services and Capabilities Survey

Need to understand the following ?

- **Notebook Services/Platforms current status and future plans**
  - **What data is accessible**
  - **Who is eligible to use**
  - **Number of concurrent users**
  - **Duration of access**
- **Archive OpenDAP access**
- **Communities of interest ?**
- **Core Skills**
- **Input to Best Practice**

Core skills – Navigating the EO Landscape and Data Types

- EO Data Landscape
  - Locating Satellite Data,
  - Discovery Tools
  - Catalogues and Documentation
  - Data Analysis Platforms – including Navigating the Linux shell
  - Data Cubes

- EO Data Types and Content
  - Levels of Satellite Data
  - Atmospheric
  - Land
  - Ocean

Basic Skills

- Data Formats and Structure
  - Text/CSV/Excel
  - NetCDF
  - HDF
  - GeoTiff
- Data Visualization/Manipulation
  - Software and Libraries (Matplotlib, NumPy, Rasterio, GDAL, Basemap, Cartopy)
  - Working with arrays
  - Plotting Data, Colour Scales, Log Scales
  - Projections/Reprojections/Regridding
  - Extracting Statistics
  - Using Shapefiles
  - Writing a NetCDF file
  - Timeseries
  - Combining different satellite data

Notebook Platform Training

- Jupyter Notebook Services
- JASMIN JupyterHub Exemplars
- Binder OpenDAP and Google Colabs
- Using Data Cube 1 - Timeseries (ESA PDGS)
- Using data Data Cube 2 – Google Earth
- Others (Jupyter Notebooks Day)
- Hackathons and Team Challenges
- Working with Github
- Publishing your notebook (Zenodo, Versioning, Licensing & DOI's)

Contact Us !

esther.conway@stfc.ac.uk