Harnessing R and the Cloud for hydrology

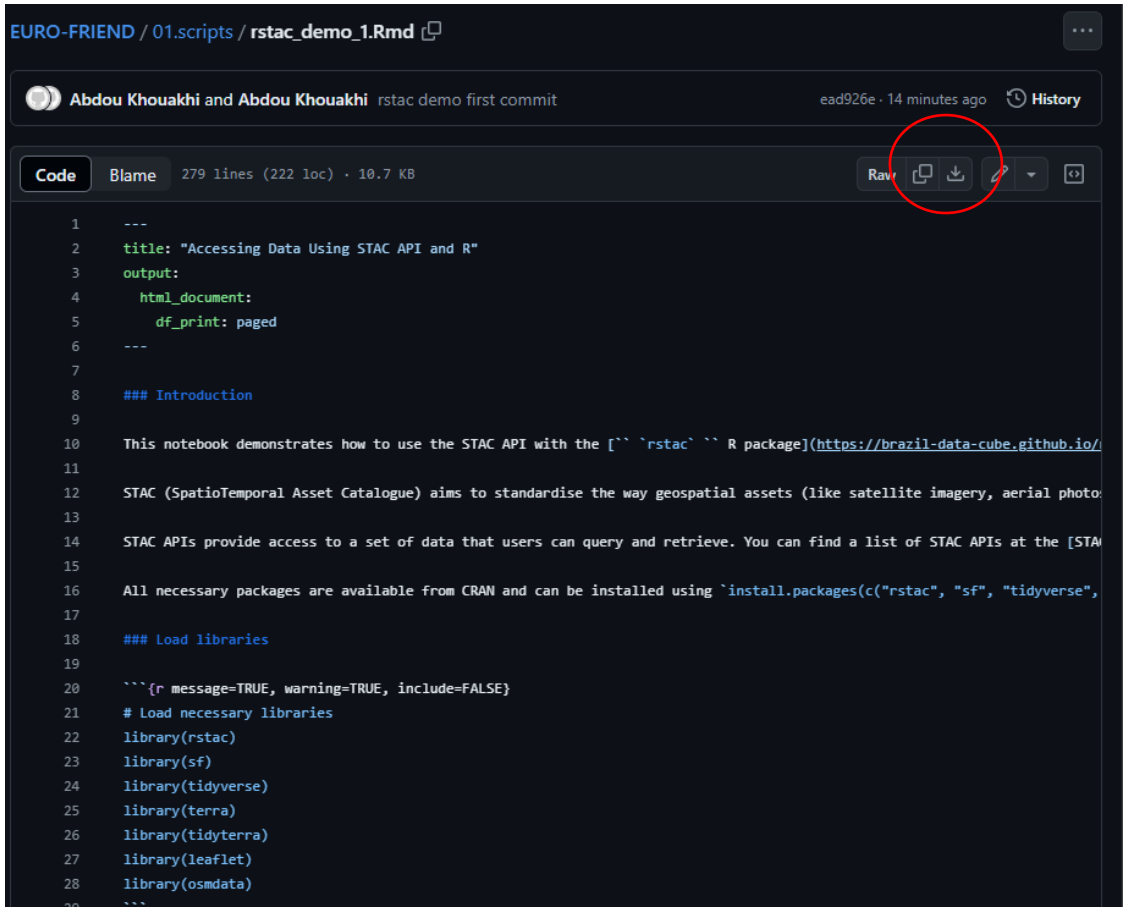
Abdou Khouakhi
a.khouakhi@cranfield.ac.uk

EURO-FRIEND Training

www.cranfield.ac.uk

Pre-demo setup: download and open rstac_demo_1.Rmd

- Go to https://github.com/khouakhi/EURO-FRIEND
- Navigate to the *01.scripts* folder.
- Click on *rstac_demo_1.Rmd*.
- Download the raw file by clicking on the “Download Raw File" button.
- Open the downloaded *rstac_demo_1.Rmd*



Aim

To introduce the concept of Cloud-Native Geospatial (CNG) and demonstrate how to use the SpatioTemporal Asset Catalogue (STAC) API to retrieve data, using the Microsoft Planetary Computer STAC API as an example.

Learning objective:

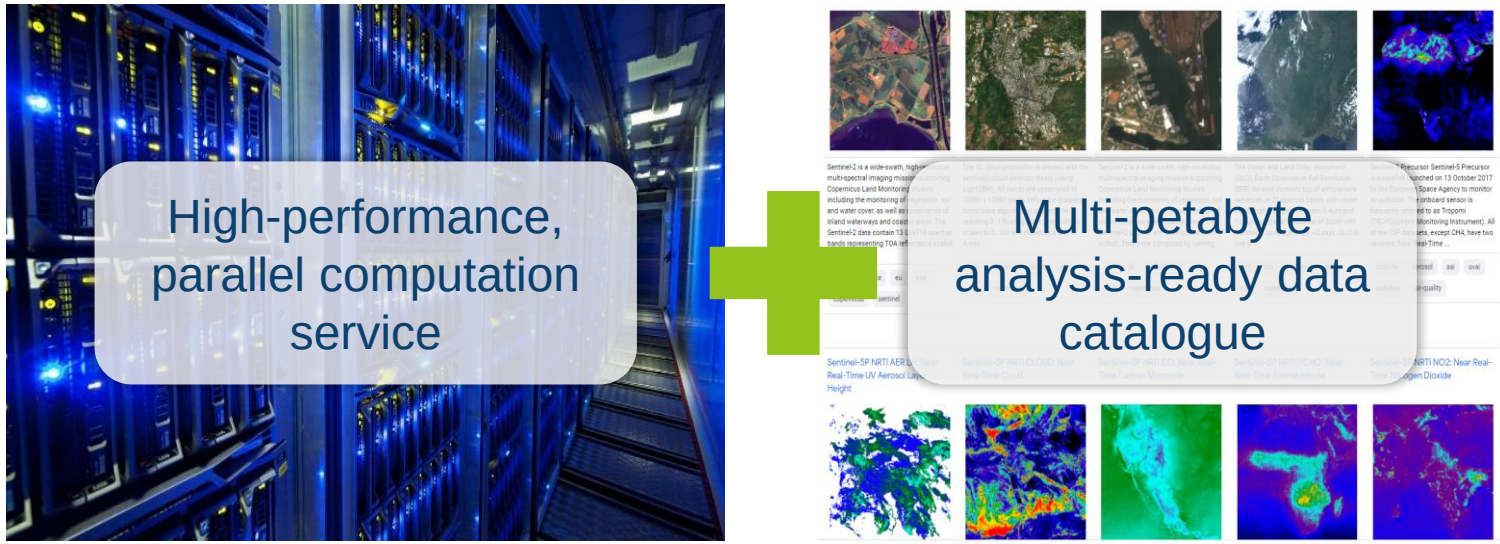
1. Describe the concept of cloud-native geospatial technologies
2. Demonstrate how to interact with STAC APIs using R
3. Retrieve spatial data for your area of interest from the MPC

Cloud-native geospatial:

- Leverages cloud technologies for geospatial data management and analysis.
- Focuses on emerging standards (COG, STAC, Zarr, COPC…) that prioritise cloud-based accessibility.
- Uses FAIR principles (Findable, Accessible, Interoperable, Reusable), promoting efficient data use.
- Enables scalable, real-time processing of remote sensing and geospatial data.
- Key platforms: Google Earth Engine (GEE), Microsoft Planetary Computer (MPC), Earth on AWS.

Key Platforms in Geospatial Cloud Computing- GEE

A cloud based platform to access, analyse and visualise geospatial data.

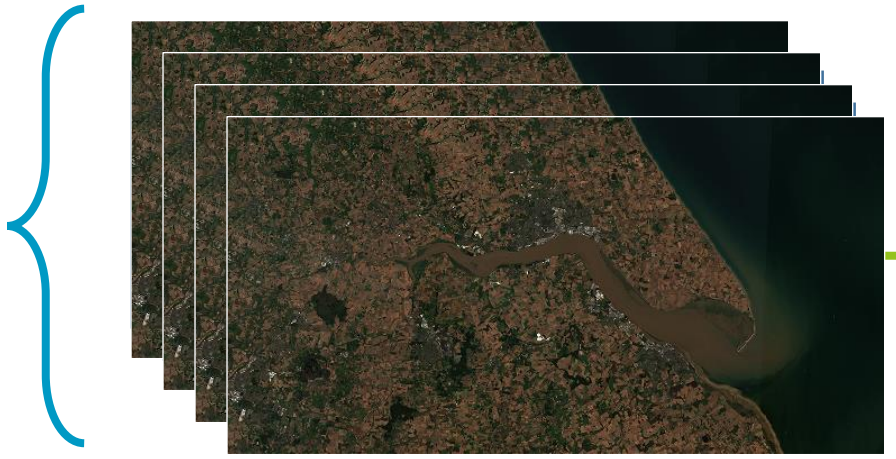


- Easy to access and analyse data;
- Less time to download data;
- Less time in managing computation infrastructure;
- More time on data analysis;
- Reproducibility.

Earth Engine is free for research, education, and non-profit use.

Data structure in GEE

- **Collections:** a stack or time series of images/ features;
- All the images produced by a single sensor are grouped together and presented as a “collection”.



Collection: e.g. Sentinel 2 TOA collection

Filter, Sort, Join, Map, Reduce

Image

- Stack of Georeferenced bands.
- Each band has its own:
  - Mask, Projection, Resolution.
- A list of properties, including:
  - Date, Bounding-box.

How to access the data?

- Data is accessed and controlled through an Internet-accessible application programming interface (API)
- Users access the API either through a client library or through a web-based interactive development environment built on top of that client library.

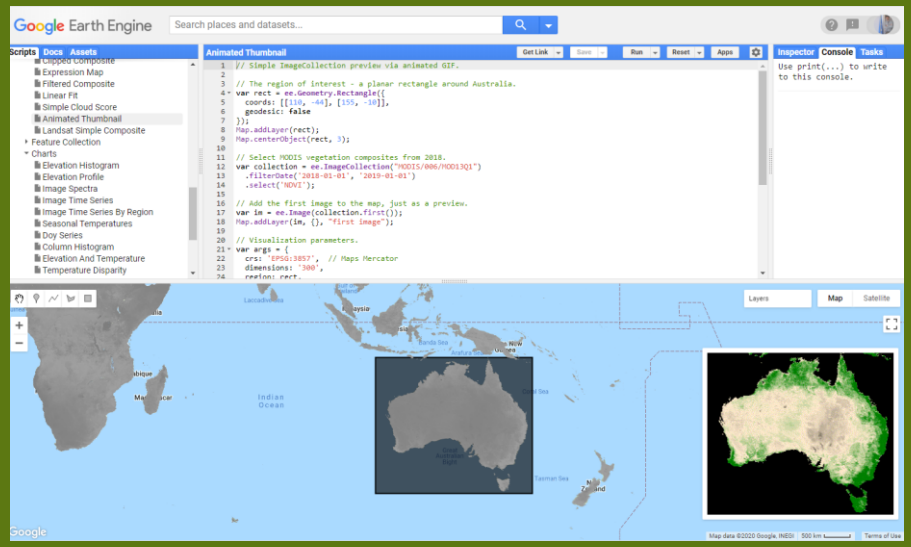
Two APIs:

- Javascript API (JS code Editor)
- Python API (Jupyter Notebook)

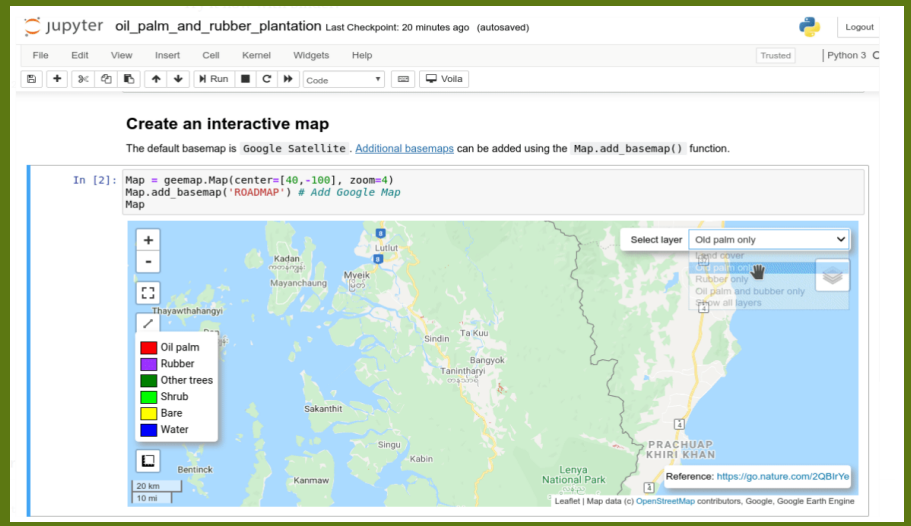
- R (rgee package)
- Explorer (explore visualise)
- QGIS (QGIS EE plugin)

See example of rgee for hydrology
https://github.com/hydrosoc/rhydro_vEGU21

Web-based JavaScript Code Editor



Interactive Jupyter notebook



Microsoft Planetary Computer

- Similar to GEE, MPC is a cloud computing platform that provides access to vast geospatial datasets and computational resources
- Launched more recently in 2021, the platform offers a range of features and benefits for users.
- The platform includes a Data Catalogue, APIs, and Applications
- MPC provides a STAC API, which makes it easier to manage and work with geospatial data.
- The platform supports scalable computing with Dask Gateway, making it possible to analyse large datasets quickly and efficiently.

Data Catalog

The Planetary Computer includes petabytes of environmental monitoring data, in consistent, analysis-ready formats, accessible through our APIs as well as directly available via Azure Storage.

API

The Planetary Computer API makes it easy for users to find exactly the data they need, simplifying search and discovery across our Data Catalog.

Applications

Partners all over the world are building on top of the Planetary Computer platform, providing the actionable information that is critical to sustainability practitioners.

Microsoft planetary computer data catalog

Data Catalog

The Planetary Computer Data Catalog includes petabytes of environmental monitoring data, in consistent, analysis-ready formats. All of the datasets below can be accessed via Azure Blob Storage, and can be used by developers whether you're working within or outside of our Planetary Computer Hub.

Filter datasets

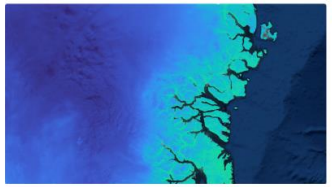
- Featured
- Air Quality
- Biodiversity
- Biomass/Vegetation
- Climate/Weather
- DEMs
- Demographics
- Fire
- Imagery
- Infrastructure
- Land use/Land cover
- SAR
- Snow
- Soils

Featured

Landsat Collection

The Landsat program provides provides a comprehensive, continuous archive of multispectral imagery of the Earth's surface from 1972 to present.

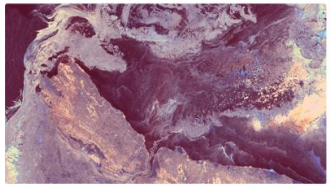
Landsat | USGS | NASA | Satellite | Global | Imagery | ···



MODIS Version 6.1 Products

The MODIS instrument operates on both the Terra and Aqua spacecraft, covering the entire surface of the Earth within one or two days. The derived data products describe atmosphere, cryosphere, land, and ocean features utilized in studies across various disciplines.

MODIS | NASA | USGS | Satellite | Global | Imagery | ···



Sentinel-1 Synthetic Aperture Radar (SAR)

Sentinel-1 comprises a constellation of two polar-orbiting satellites, operating day and night performing C-band synthetic aperture radar imaging.

ESA | Copernicus | Sentinel | C-Band | SAR

Sentinel-2 Level-2A

The Sentinel-2 program provides global imagery in thirteen spectral bands at 10m-60m resolution and a revisit time of approximately five days. This dataset contains the global Sentinel-

https://planetarycomputer.microsoft.com/catalog

Key Cloud-Native Geospatial Formats

STAC (SpatioTemporal Asset Catalog)

- An open specification for geospatial asset metadata.

COG (Cloud Optimized GeoTIFF)

- A format for GeoTIFF files that is optimised for cloud-based workflows.
- Efficient storage, quick access, and compatibility with cloud services.

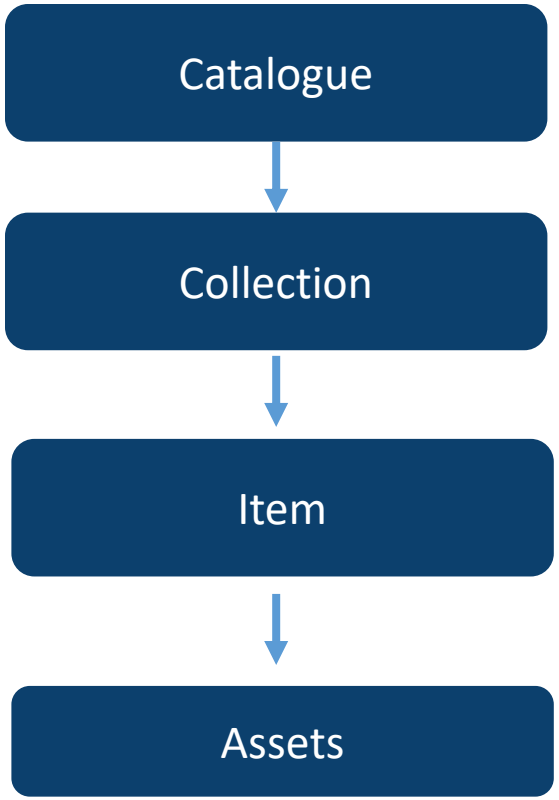
Zarr

- A format for chunked, compressed, N-dimensional arrays.

SpatioTemporal Asset Catalogue (STAC)

STAC (SpatioTemporal Asset Catalog) is a specification for organising and describing geospatial information, making it easier to search, share, and access spatial-temporal data on the web.

STAC Entities Structure



- The top-level entry point for a STAC dataset - organises and provides metadata about the collections and items it contains (e.g. MPC catalogue)

- A grouping of related items, usually representing a dataset or a series of datasets.
- Provides metadata that applies to all items in the collection, such as the spatial and temporal extent (e.g., Landsat 8 Collection)

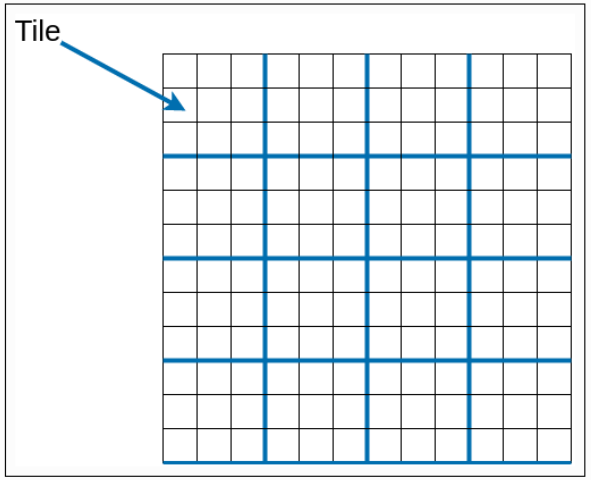
- The core object of a STAC catalogue, representing an individual geospatial asset.
- Contains metadata specific to that item, such as acquisition date, geometry, and links to the actual data assets (e.g., Single Landsat 8 Image)

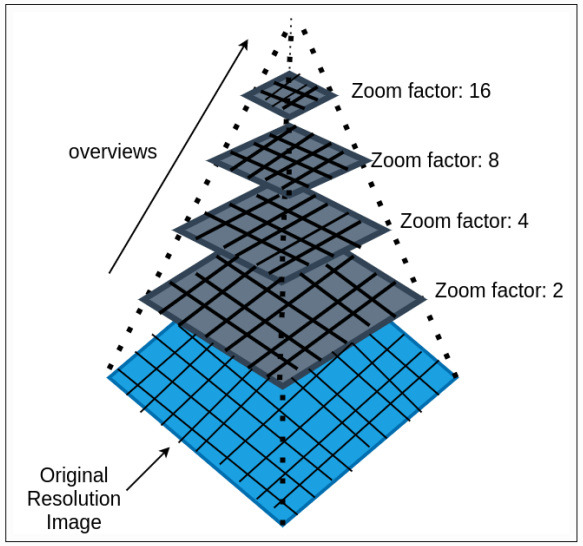
- The actual data files or resources that belong to an item.
- Provides direct links to the data and describes the format and type of each asset.

Cloud Optimised GeoTIFFs (COGs)

COGs are regular GeoTIFF files with some additional features that make them ideal for cloud computing and other web-based services.

- Raw pixel data is stored as an index of 'tiles' within a GeoTIFF file, making it easier to read specific portions of the file.
- Pixel data is organised into multiple 'overviews' that represent different zoom levels, making it quick to select a point of interest by rendering a lower quality.
- COGs use HTTP Get range requests to allow users to request portions of the image file, reducing the amount of data transferred over the network.





Hands-on session

Prerequisites:

- Internet access
- R and RStudio

R notebook

- Use rstac to interact with the MPC data
- Demonstrate how to retrieve data over a river basin of interest
- Examples include getting DEM data and downscaled CMIP6 data

Useful documents

- MPC documentations and tutorials
- MPC Data catalogue
- STAC Browser
- STAC index
- rstac package vignette