National Aeronautics and Space Administration

Optimizing Data Formats for Earth Information System Fire Portal

Goddard SPACE FLIGHT CENTER

Johana Chazaro Cortes, California Baptist University
Marina Dunn, University of California Riverside
Dieu My Nguyen, University of Colorado Boulder

Mentor: Alexey N. Shiklomanov, PhD, GSFC-618

The EIS Portal is a cloud-based project designed to support the understanding of fire activity through an interactive website.

Part 1: Johana Chazaro Cortes - Intro and API work

- Developing APIs that allow for easy access & analysis of data products.

Part 2: Marina Dunn - Web development

- Test possible cyberinfrastructure technologies to sustain a public-facing website.

Part 3: Dieu My - Storage optimization

- Find the most efficient way to store and access multi-dimensional datasets on cloud storage.

Part 1 - Introduction and API Work
By: Johana Chazaro Cortes

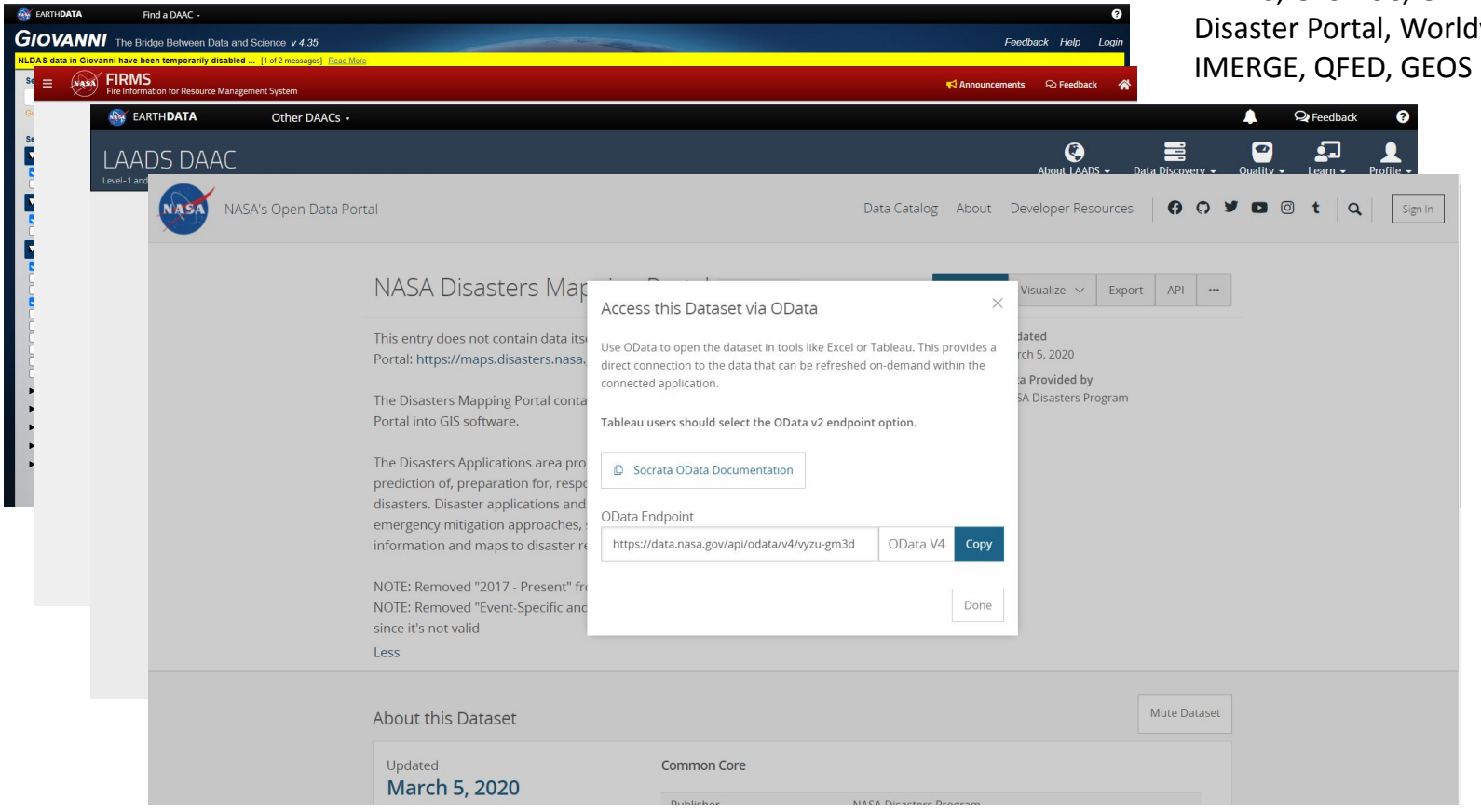
CBU

B.S. Computer Science Spring ‘22

https://github.com/joChazaro

Lifelong dream of seeing buffalo!

NASA fire data exist on various platforms and formats.



Interfaces:

FIRMS, GESDISC, GMAO FP, GFED, NASA Disaster Portal, Worldview, GWIS, IMERGE, QFED, GEOS

As NASA transitions to the Cloud, we are developing ways to make fire data more accessible.

		EIS Fire Portal	EARTHDATA	WORLDVIEW	FIRMS Fire Information	GFED	GMAO	GIOVANNI
Active Fires	MODIS	✓	✓	✓	✓	✓	✗	✗
	VIIRS	✓	✓	✓	✓	✓	✗	✗
	GOES 16/17	✓	✗	✗	✗	✗	✗	✗
	NIFC Fire Perimeters	✓	✗	✗	✗	✗	✗	✗
Fire Emissions	GEOS-FP Output	✓	✗	✗	✗	✗	✓	✗
	AERONET	✓	✗	✗	✗	✗	✓	✗
	TROPOMI	✓	✗	✗	✗	✗	✗	✗
	QFED	✓	✗	✗	✗	✗	✗	✗
	GFED	✓	✗	✗	✗	✓	✗	✗
Fire Weather	IMERG	✓	✓	✓	✗	✗	✗	✓
	GEOS-FP Output	✓	✗	✗	✗	✗	✓	✗
Capabilities	Time-Series Analysis	✓	✗	✗	✗	✓	✓	✓
	Subsetting & Aggregation	✓	✗	✗	✗	✓	✗	✓
	Custom-Derived Variables	✓	✗	✗	✗	✗	✗	✗

The focus on these scripts is to allow access to the analysis ready data in a much easier way through APIs.

```
#SET START DATE
if start_date is None:
    start_date = zarr.time.min().values
else:
    start_date = pd.to_datetime(start_date)
assert start_date >= minDate, f"Start date should be after {minDate}"
#SET END DATE
if end_date is None:
    end_date = zarr.time.max().values
else:
    end_date = pd.to_datetime(end_date)
assert end_date <= maxDate, f"End date should be before {maxDate}"

print("Processing request, please wait.")
polygon = zarr[var].sel(lat=slice(coord_array[1], coord_array[3]),
        lon=slice(coord_array[0], coord_array[2])).sortby("time").sel(time = slice(start_date, end_date))
#plot mean of var over polygon
polygon.mean(["lat", "lon"]).dropna("time").plot()
```

These scripts are influenced by common procedures our scientists have to repeatedly do for fire data analysis

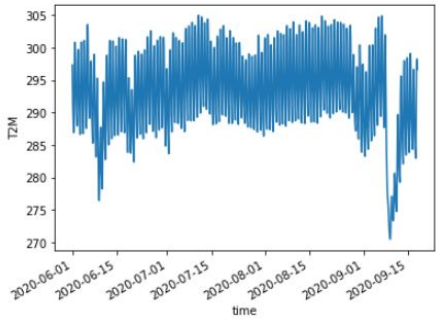
```
print("Processing request, please wait.")
polygon = zarr[var].sel(lat=slice(coord_array[1], coord_array[3]),
        lon=slice(coord_array[0], coord_array[2])).sortby("time").sel(time = slice(start_date, end_date))
#plot mean of var over polygon
polygon.mean(["lat", "lon"]).dropna("time").plot()
```

Sample function call

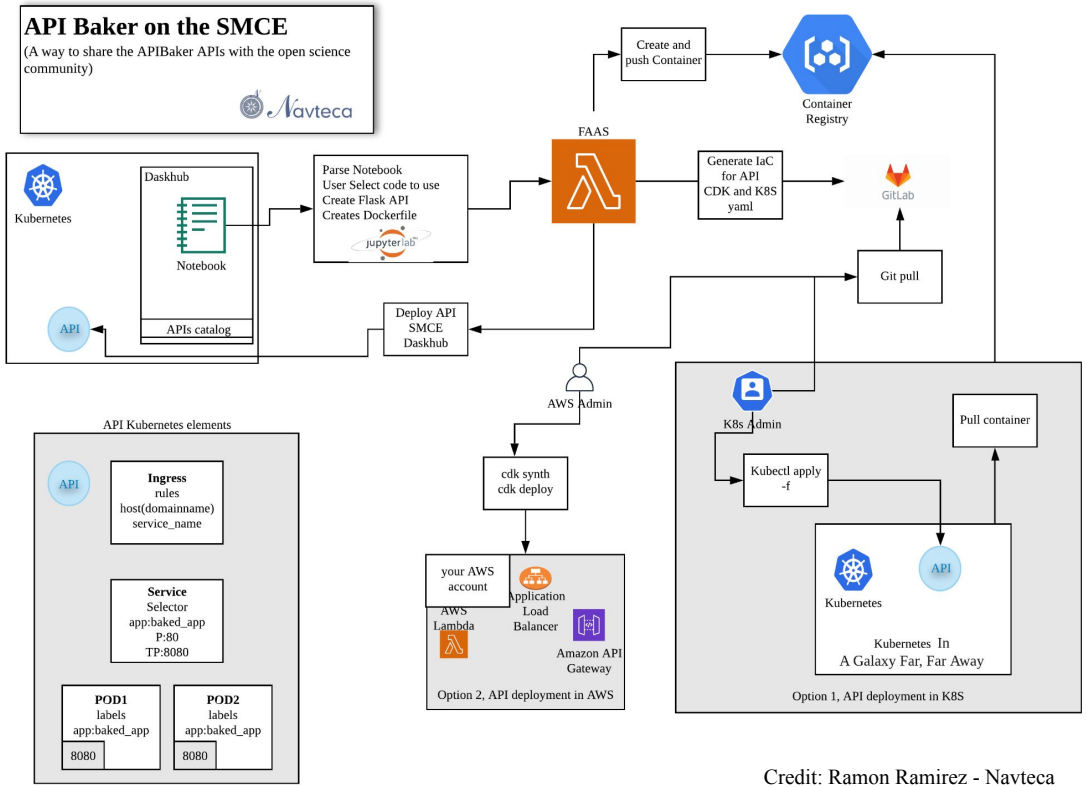
coord_array = [lon1, lat1, lon2, lat2]

```
coord_array = [
    -109.060062,
    36.992426,
    -102.041574,
    41.003444
  ]
polygon_timeseries_function(coord_array, None, "2020-09-18", path = "eis-dh-fire/geos-fp-global/tavg.zarr/", var = "T2M")
```

```
Processing request, please wait.
```



These scripts are designed to work cohesively with API Baker and to used with the future EIS-fire website



Credit: Ramon Ramirez - Navteca

Part 2
Web development
By: Marina Dunn

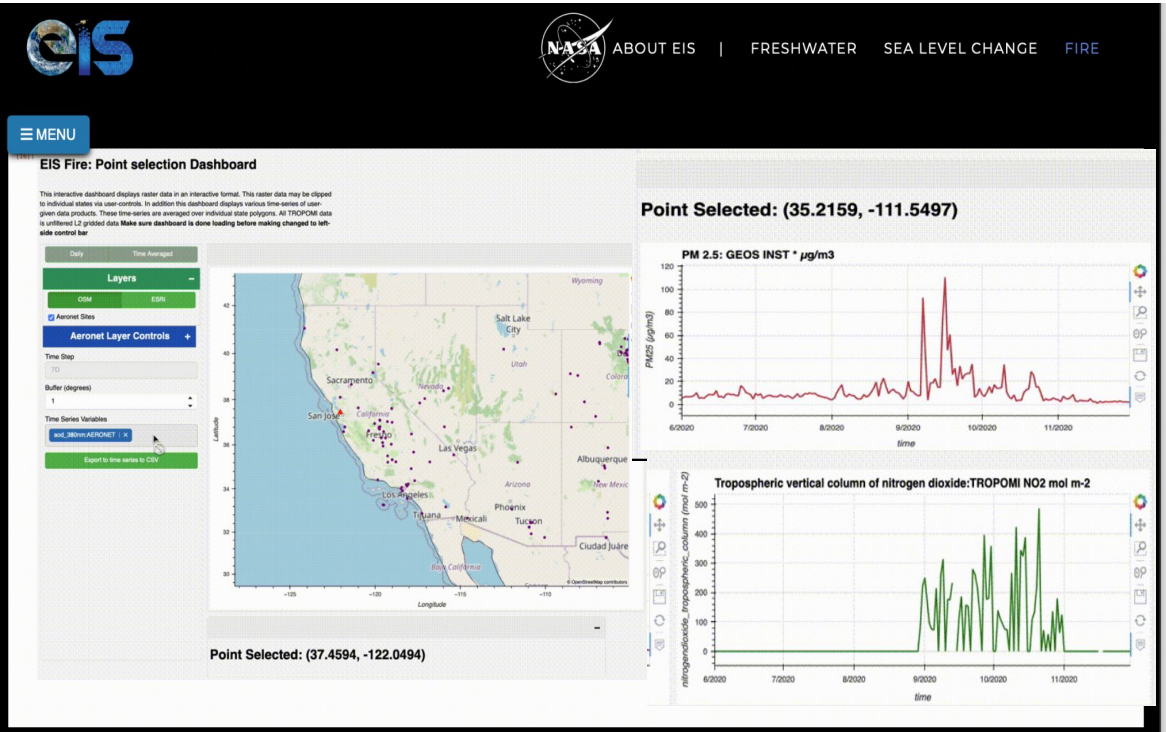
U of AZ
BS. Astronomy

UC Riverside
MS. Engineering - Data Science in progress

marinadunn.github.io

Current EIS site only allows certain users access to dashboards

source: https://eis-fire.mysmce.com



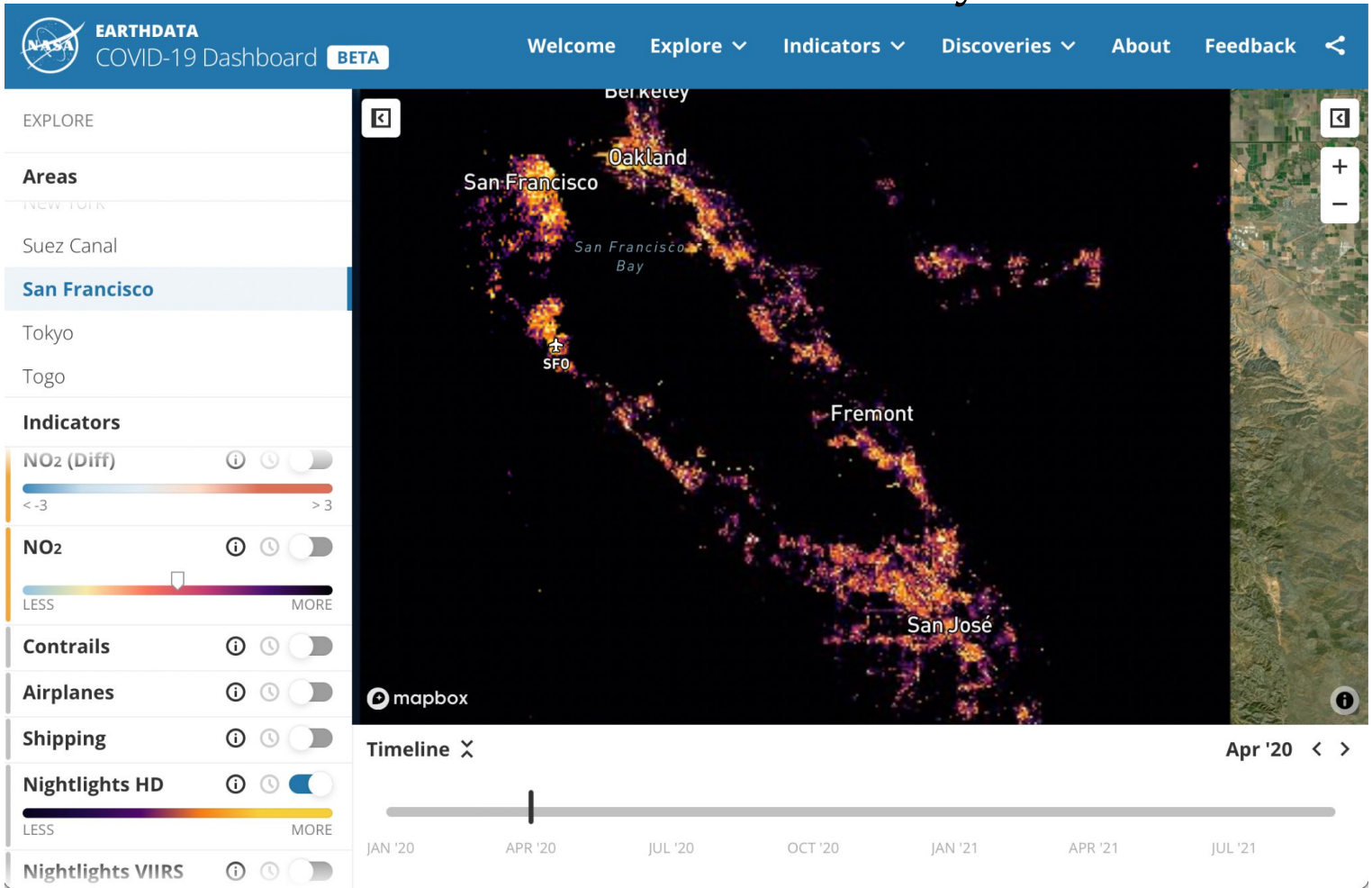
An example Jupyter notebook-based dashboard developed during the pilot study. Here, a user can easily retrieve time series for variety of fire-related data products — including satellite measurements of atmospheric chemistry (e.g., TROPOMI), ground-based measurements of aerosols (e.g., AERONET), model estimates of near-surface PM 2.5 (GEOS), and composite indices of meteorological fire risk — by just clicking a location on a map. Click here to see the source code for this notebook.

Datasets

Add, modify or remove the desired Zarr dataset file. Following the convention: `"DATASET" : "path/to/zarr/file.zarr`

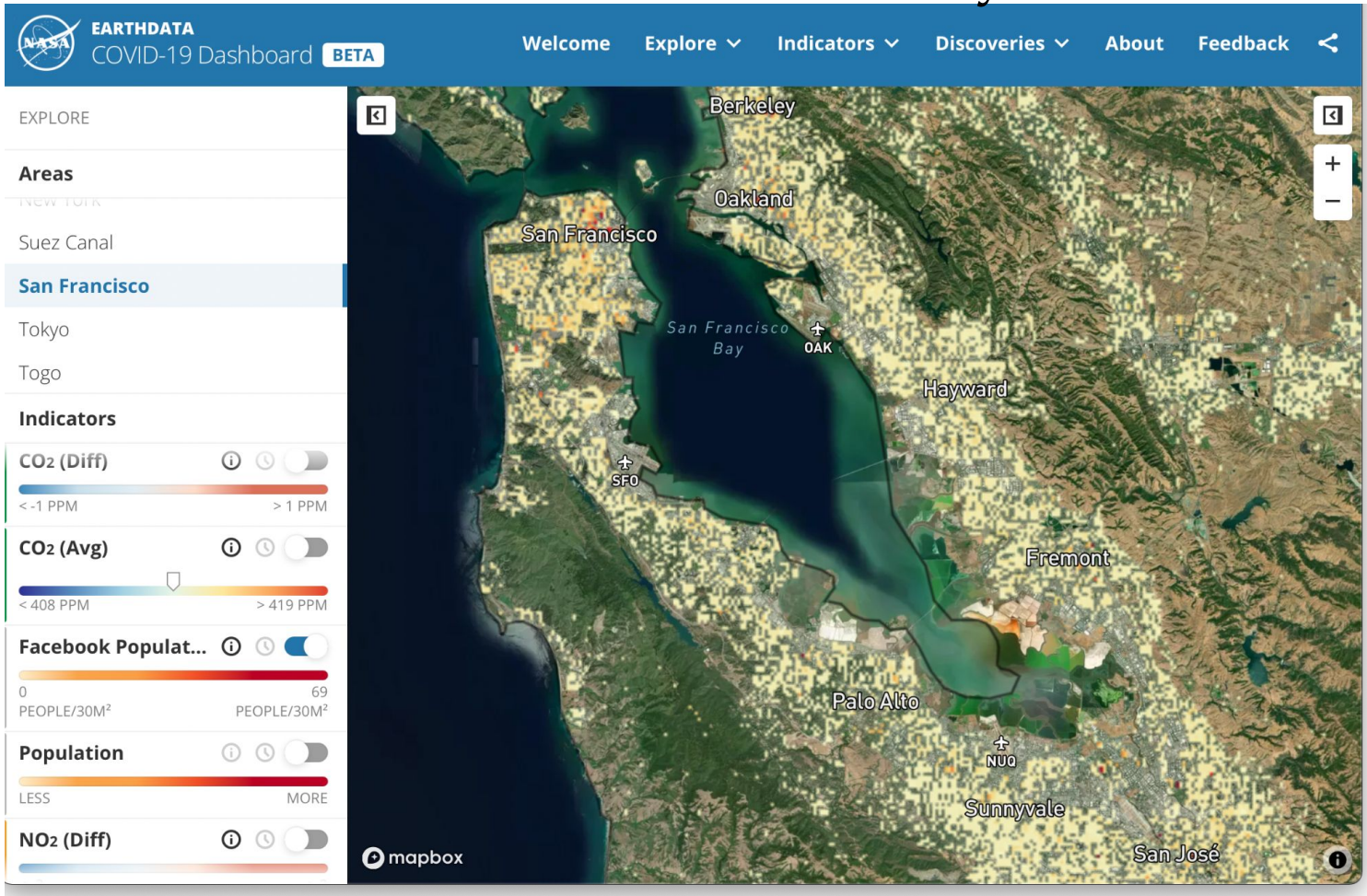
```
raster_filepath = {
    'IMERG FWI': "eis-dh-fire/zarr-rechunked-10/imerg-fwi.zarr",
    'GEOS FP CONUS': "eis-dh-fire/zarr-rechunked-10/geos-fp-zarr/conus.zarr",
    'QFED': 'eis-dh-fire/zarr-rechunked-10/qfed.zarr/',
    'GEOS INST': 'eis-dh-fire/zarr-chunked-30/inst-30.zarr'
}


# TROPOMI DATA
base_path = '/home/jovyan/efs/eis-fire-tropomi'
tropomi_filepath = {
    'TROPOMI AEROSOL-INDEX': "tropomi_aer_ai.zarr",
    'TROPOMI CH4': "tropomi_ch4.zarr",
    'TROPOMI CO': 'tropomi_co.zarr',
    'TROPOMI NO2': 'tropomi_no2.zarr',
    'TROPOMI OZONE': 'tropomi_o3.zarr'
}
for k, v in tropomi_filepath.items():
    tropomi_filepath[k] = os.path.join(base_path, v)


# ---
# Variables from raster datasets to use.
# Follow {'DATASET': ['var1', 'var2'],
#         'DATASET': ['var1', 'var2']
# ---
raster_variables = {
    'IMERG FWI': ['IMERG.FINAL.v6_FWI'],
    'GEOS FP CONUS': ['T2M', 'U2M', 'V2M', 'PRECTOT'],
    'QFED': ['co.biomass'],
    'GEOS INST': ['DUSMASS25', 'SSSMASS25', 'BCSMASS', 'OCSMASS', 'SO4SMASS']
}

# See Cartopy docs for more projection options.
projection = ccrs.PlateCarree()

# List desired defualt variables here
defaultVars = ['IMERG.FINAL.v6 Fire Weather Index:IMERG FWI',
               'Vertically integrated CO column:TROPOMI CO',
               'Tropospheric vertical column of nitrogen dioxide:TROPOMI NO2',
               'CO2 Biomass Emissions:QFED']

# Define title and optional subtitle that will show up in the dashboard
title = '# EIS Fire: NIFC PSA Fires Dashboard'
subtitle = 'This interactive dashboard displays raster data in an interactive format.' + \
' This raster data may be clipped to individual states via user-controls. '+ \
'In addition this dashboard displays various time-series of user-given data products. '+ \
'These time-series are averaged over individual state polygons. All TROPOMI data is unfiltered L2
```

New site will be built on Earthdata cyberinfrastructure



source: https://earthdata.nasa.gov/covid19

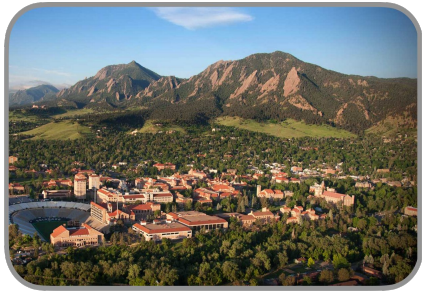
New site will be built on Earthdata cyberinfrastructure



source: https://earthdata.nasa.gov/covid19

Part 3
Storage Optimization
By: Dieu My Nguyen

U of AZ
BS. Neuroscience
BA. Creative Writing



U of CO Boulder
MS. Computer Science
PhD. In progress

Peleg Lab

dieumy.t.nguyen@nasa.gov

dieumynguyen.github.io
peleglab.com

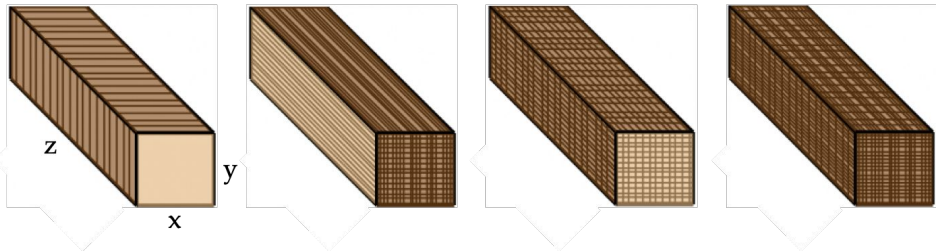
Problem & goal

- How does the chunking scheme affect usage and analysis of multi-dimensional datasets?

- What are the optimal chunking strategies for storing datasets on the cloud?

--- Test performance of different chunking strategies ---

Data cube



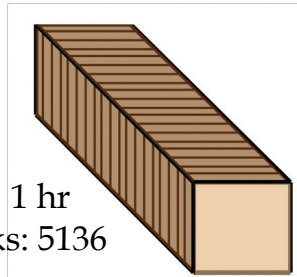
Dataset & chunking strategies

GEOS-FP
dataset in Zarr format

From GEOS - high resolution global atmospheric model

Analyses and forecasts produced in real time

Default chunking scheme

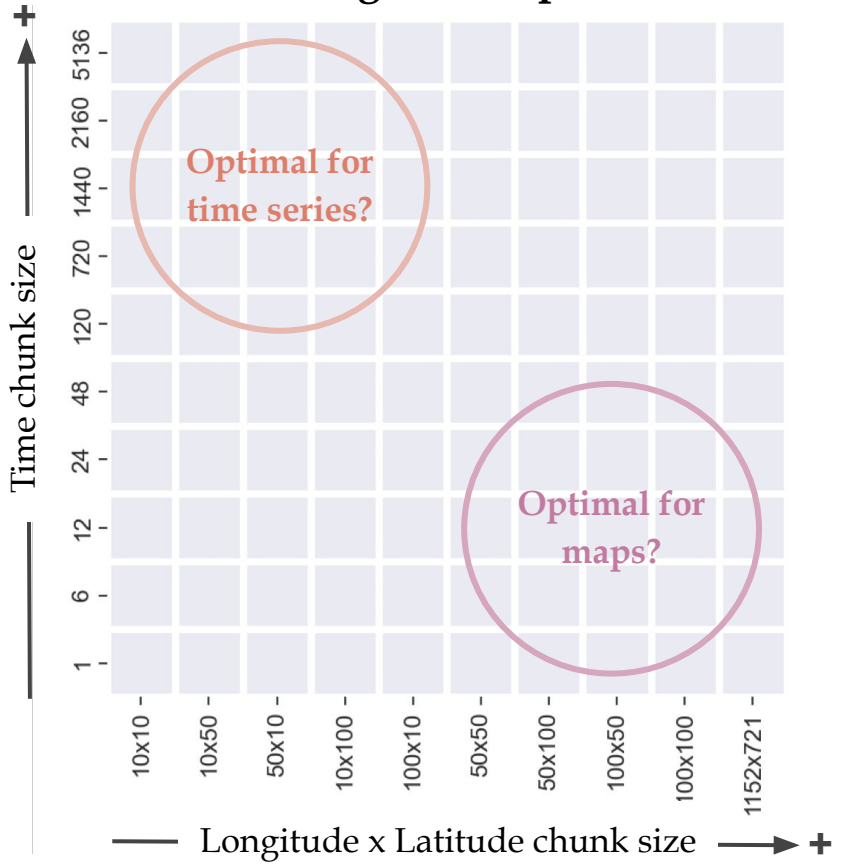


Time
Chunk size: 1 hr
Num chunks: 5136

Latitude
Chunk size: 721 deg
Num chunks: 1

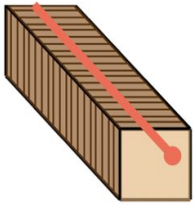
Longitude
Chunk size: 1152 deg
Num chunks: 1

Strategies to explore



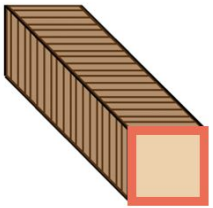
Trade-off between time series and map in CPU time

Task 1. Drawing time series at single location

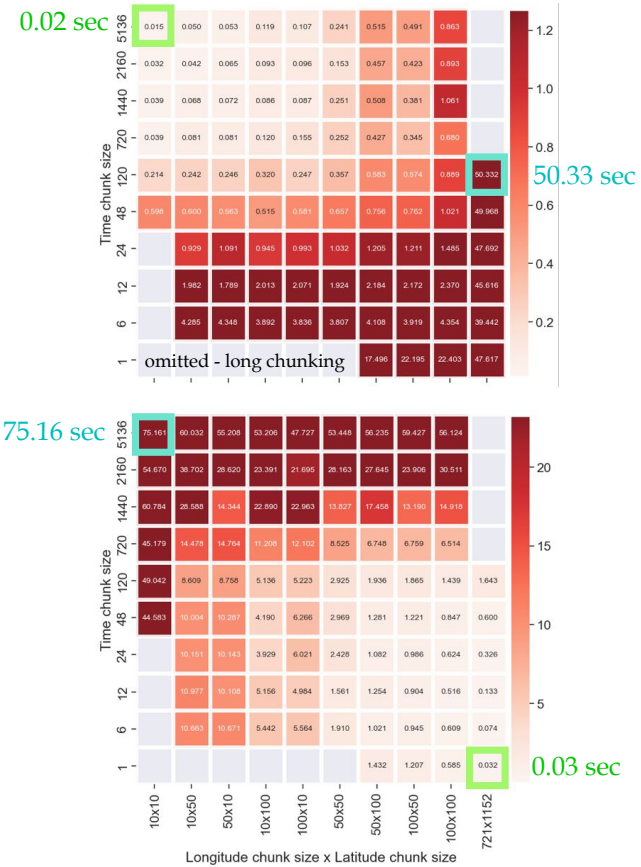


longitude: -122.19
latitude: 47.61

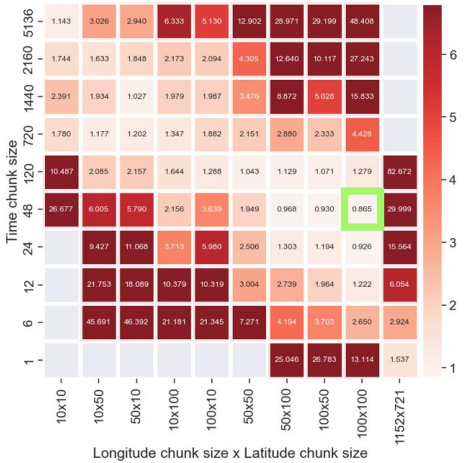
Task 2. Drawing map at 1 time step



time: 2020-06-01 T00:00:00

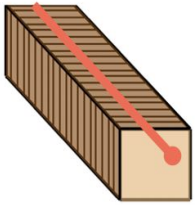


Product of task 1 and task 2



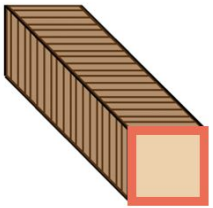
Trade-off between time series and map in peak memory

Task 1. Drawing time series at single location

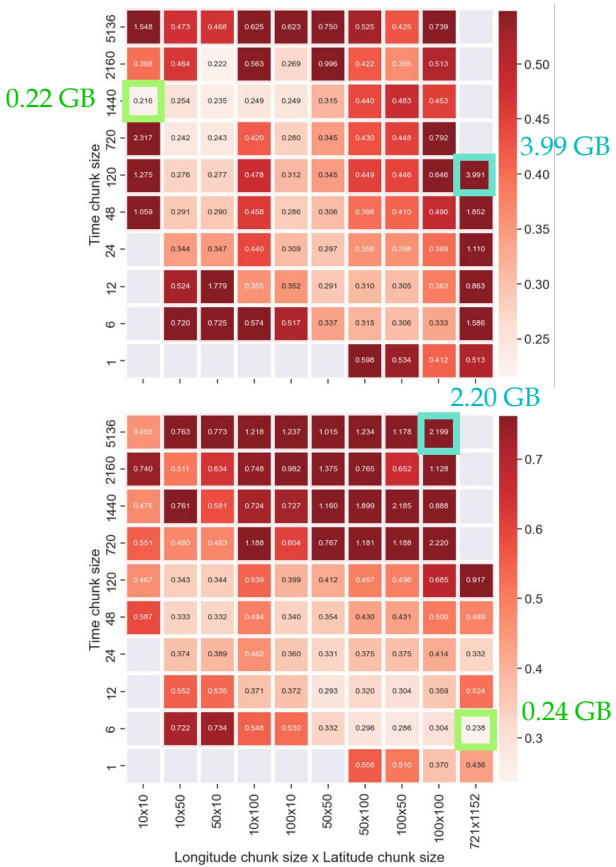


longitude: -122.19
latitude: 47.61

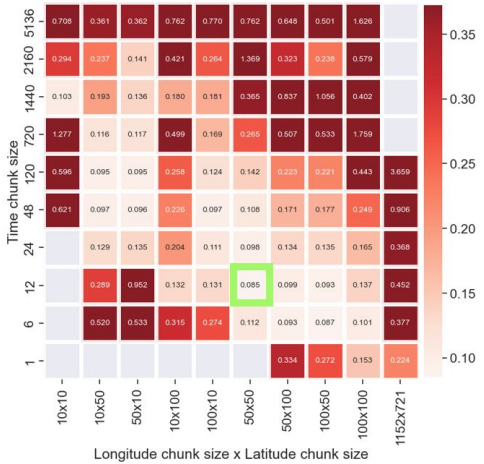
Task 2. Drawing map at 1 time step



time: 2020-06-01
T00:00:00

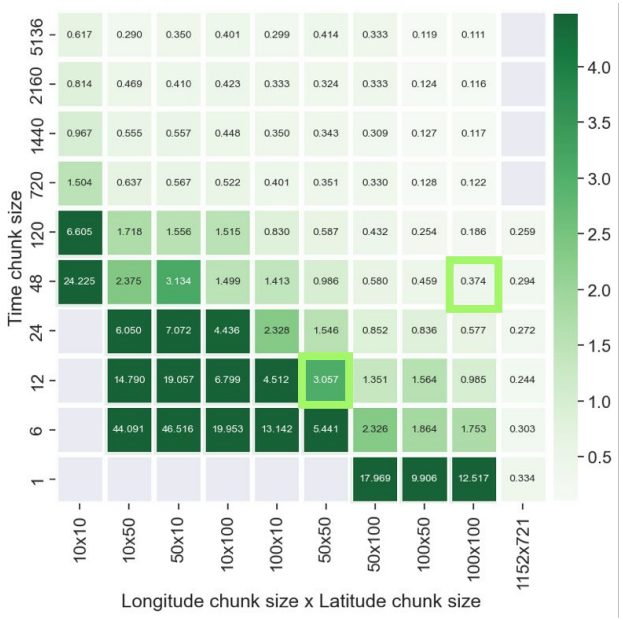


Product of task 1 and task 2

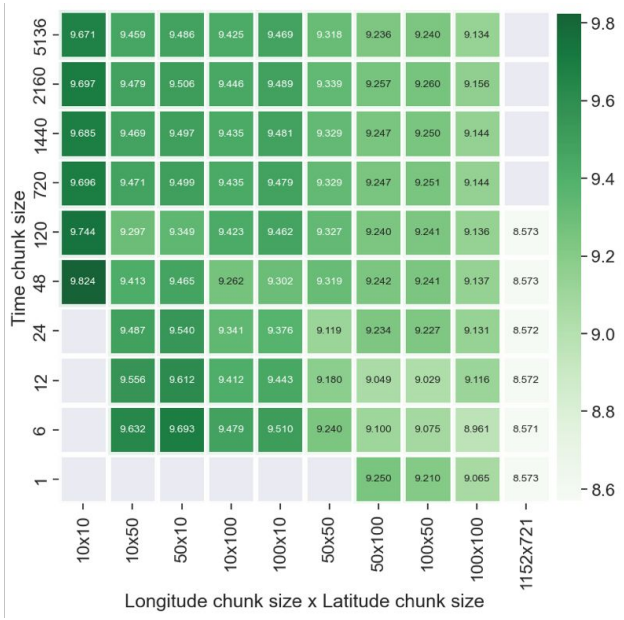


Strategy also affects rechunking time & storage size

Rechunking time (hr)

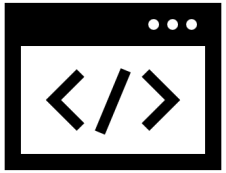


Storage size (GB)



Some compression effect

Future Direction



Lots of work to be done to achieve a publicly available website that optimally interacts with Zarr.

Apply optimized chunking strategies to new data sets using Pangeo Forge.

Automate new processes used by clients with APIs while accessing rechunked datasets.

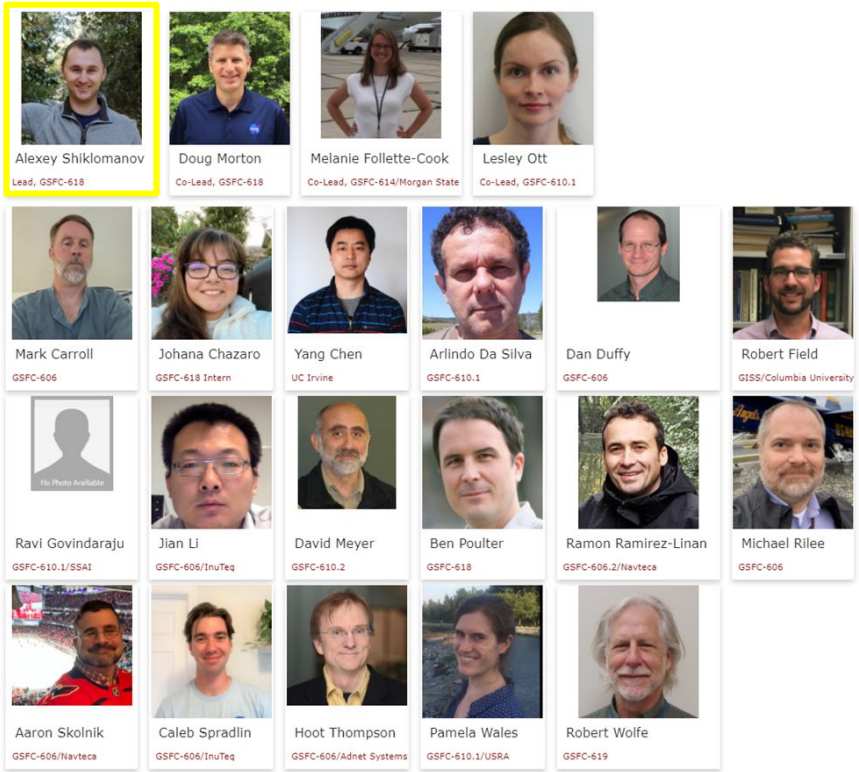
Acknowledgements

Jeremy Raupp (GSFC/NAVTECA)

Sean Harkins, Aimee Barciauskas (DevelopmentSeed)

Kata Martin, Joe Hamman, Jeremy Freeman (CarbonPlan)

EIS-Fire Team



References

CarbonPlan

Pangeo-Forge

NASA-IMPACT

Papers:

Fer, I., Gardella, A. K., Shiklomanov, A. N., Campbell, E. E., Cowdery, E. M., Kauwe, M. G. D., et al. (2021). Beyond Ecosystem Modeling: A Roadmap to Community Cyberinfrastructure for Ecological Data-Model Integration. *Global Change Biology*, *27*(1), 13–26. https://doi.org/10.1111/gcb.15409

Ramachandran, R., Baynes, K., Murphy, K., Jazayeri, A., Schuler, I., & Pilone, D. (2017). CUMULUS: NASA's cloud based distributed active archive center prototype. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (pp. 369–372). https://doi.org/10.1109/IGARSS.2017.8126972

Summary

Present deliverables:

- APIs that allow for easy access & science with open-science principles in mind
- Efficient cloud data storage and chunking strategies for optimized access to NASA data
- Applicable understanding of cyberinfrastructure requirements to develop a sustainable website

Prospective applications:

- Prototype wholistic website that applies previous work as scalable features.

Thank you!