UNIVERSITY of WASHINGTON
eScience Institute

Moving satellite radar processing and analysis to the Cloud

Pangeo Developers Workshop
August 14, 2018

Scott Henderson
eScience Postdoctoral Fellow
University of Washington
scottyh@uw.edu

Introduction

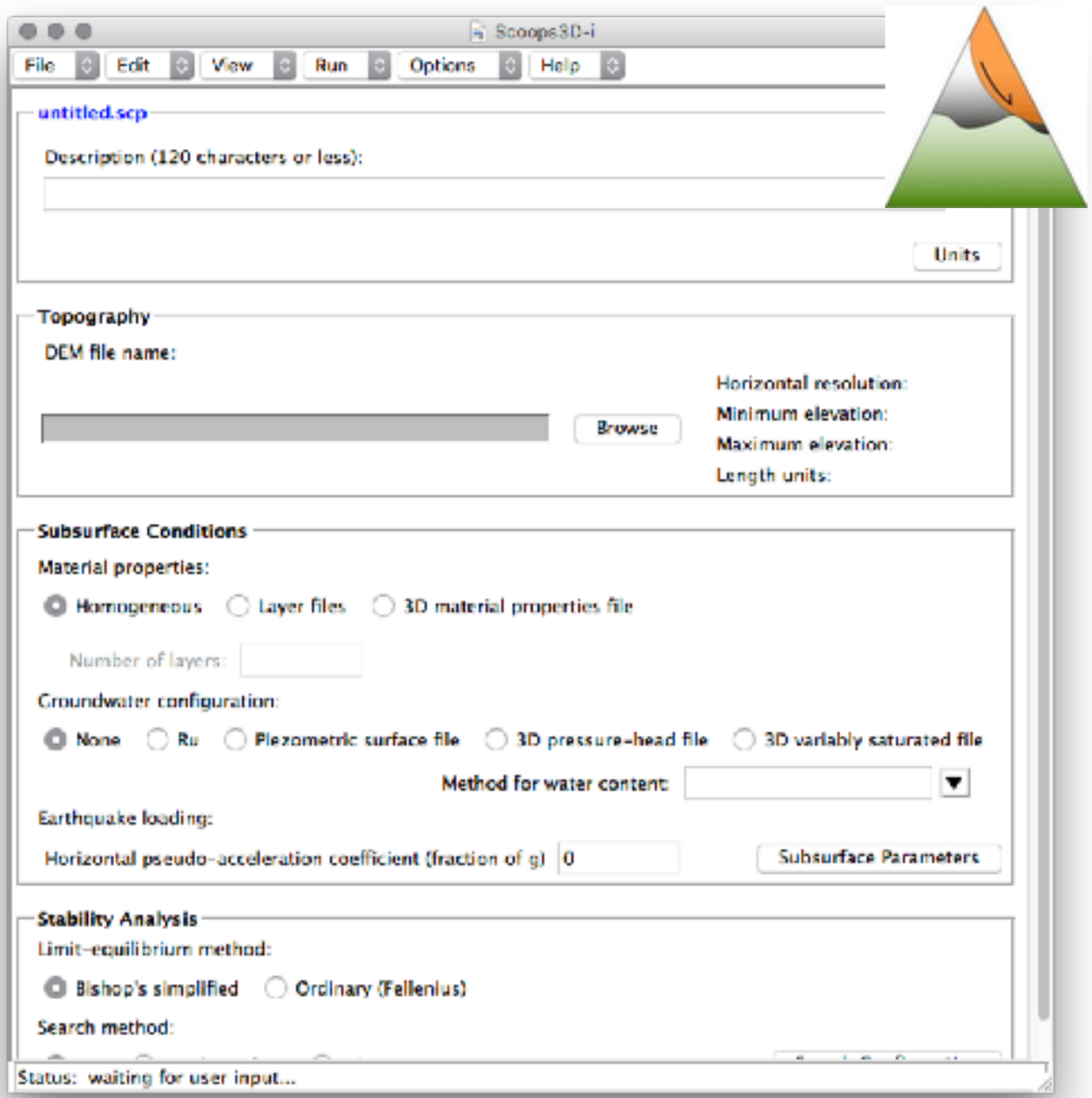
- PhD in Geological Sciences

- First worked with Python in 2008 at USGS, building a graphical interface w/ Tkinter & haven't looked back!

- Most computational training on-the-side - mostly "precious data", more and more "big data"

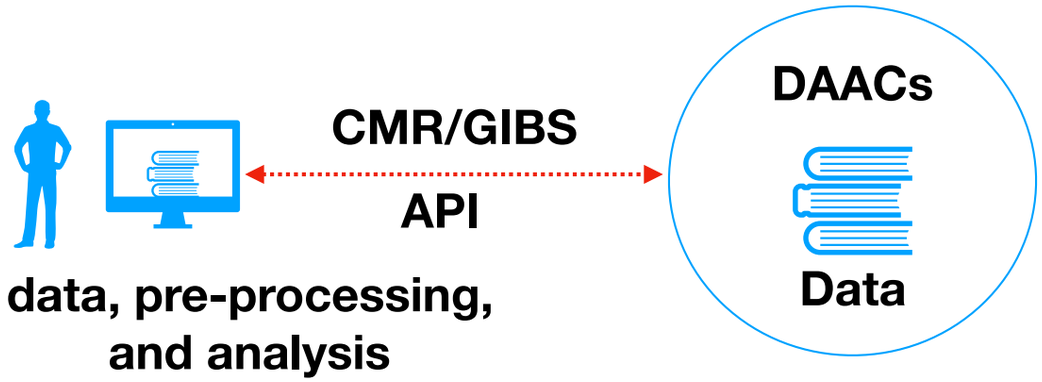
- **Research focus on constraining physical models of solid earth deformation (earthquakes, landslides, volcanic eruptions) with satellite remote sensing**

https://landslides.usgs.gov/research/software.php

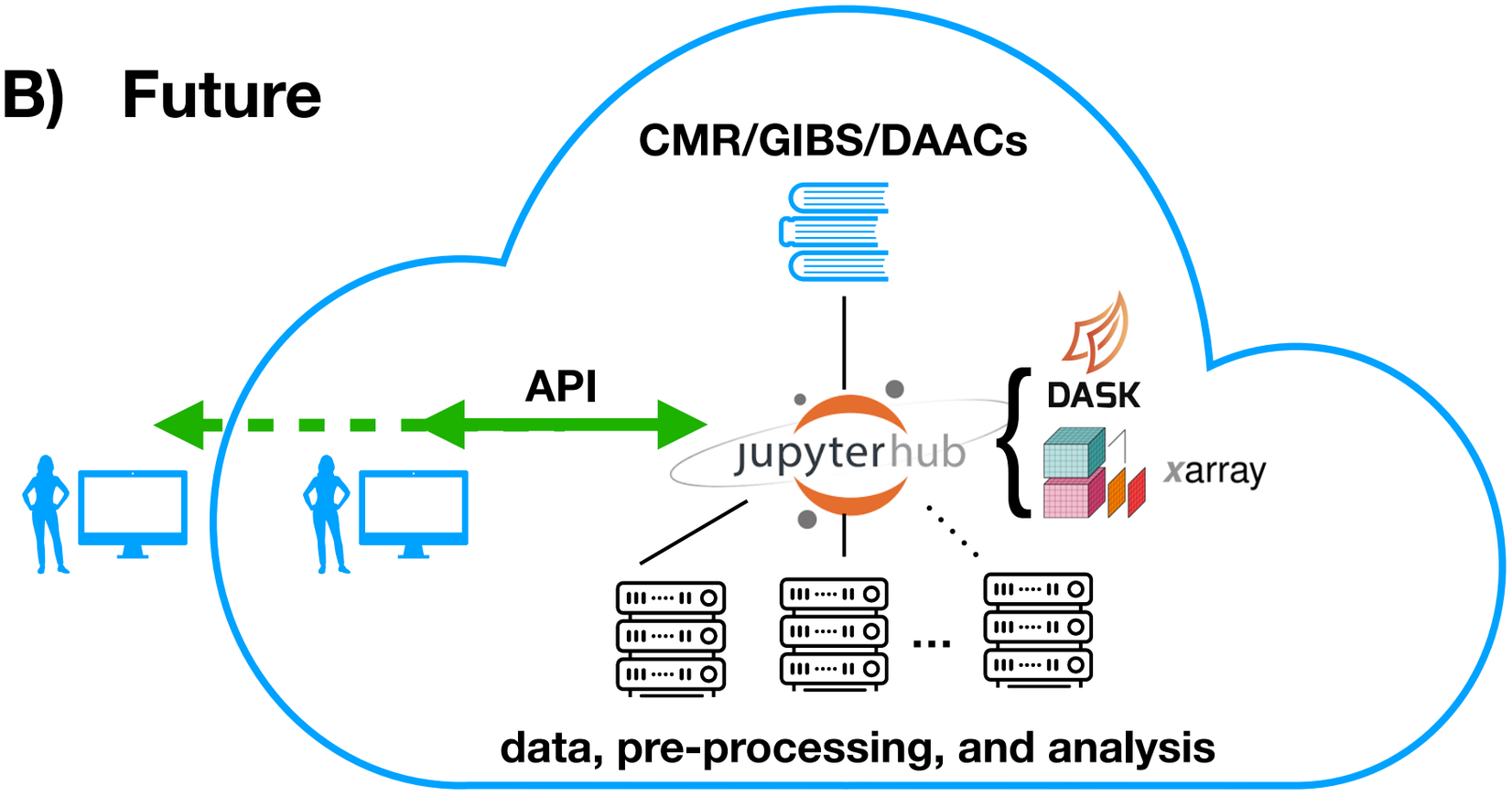


NASA ACCESS 2017 Grant

A) Current



B) Future



PANGEO

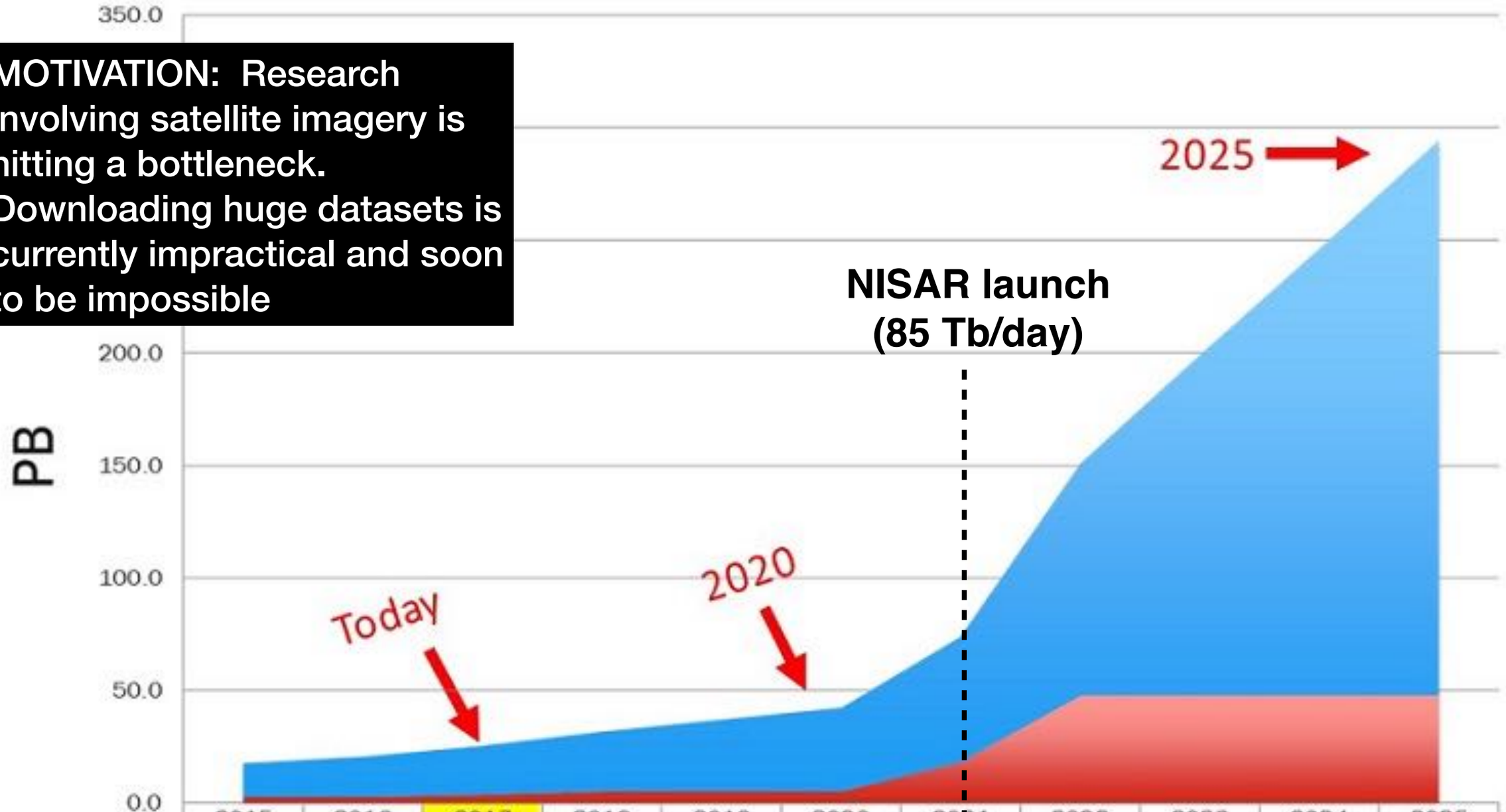
NASA

UW collaborators

Joe Hamman
Anthony Arendt
Rob Fatland
Scott Henderson
Amanda Tan
Don Setiawan

NASA Earth Observing System Data and Information System (EOSDIS) Archive Growth

MOTIVATION: Research involving satellite imagery is hitting a bottleneck. Downloading huge datasets is currently impractical and soon to be impossible



	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
umulative Archive Size (PB)	15.0	17.7	21.6	26.8	32.0	37.2	55.6	103.4	151.1	198.9	246.6
chive Growth Rate (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

Archive Growth Rate (PB) Cumulative Archive Size (PB)

Implications

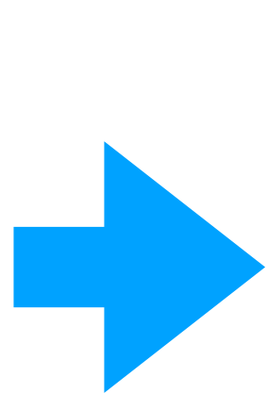
- NASA moving towards commercial cloud storage (Amazon Web Services, AWS) https://earthdata.nasa.gov/about/eosdis-cloud-evolution
  - Major benefit for scientists in virtually unlimited computing resources
  - But, scientific community is not accustomed to Cloud computing infrastructure and costs

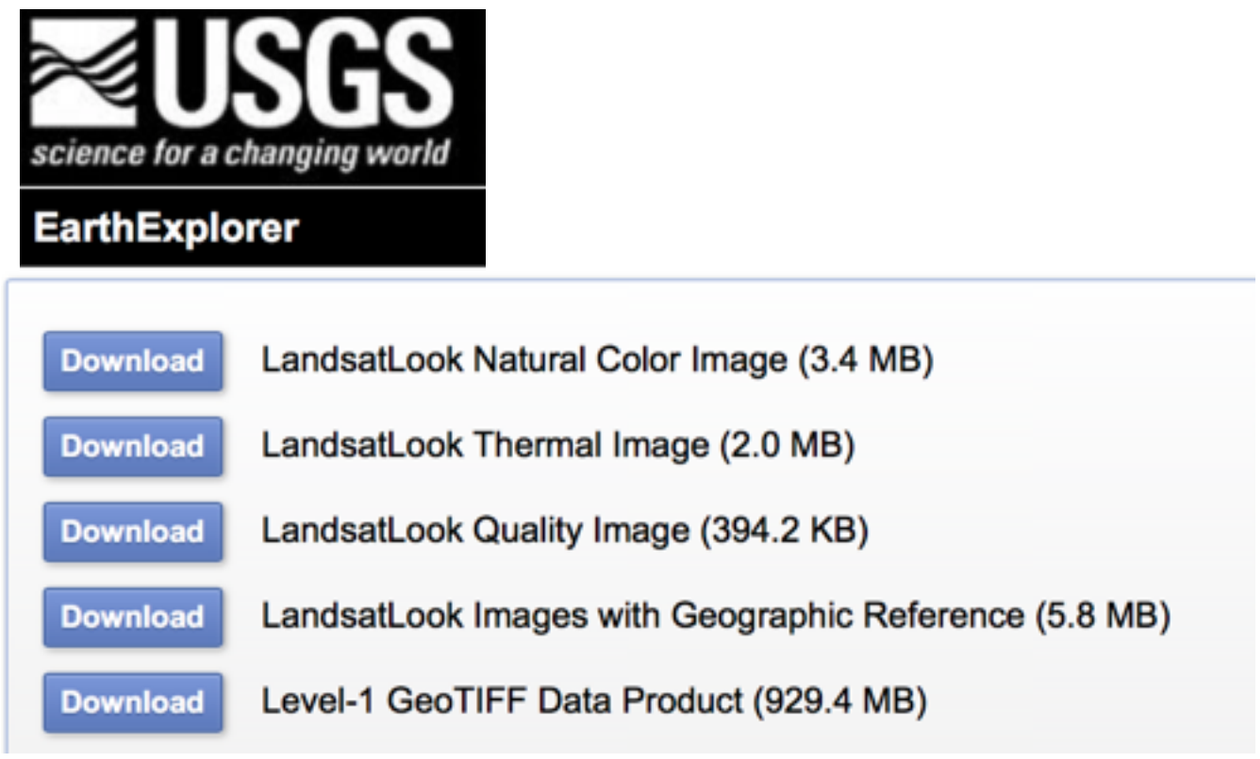
Transition is slow…

EARTHDATA Search

EARTHDATA LOGIN

(or CMR API)





* After 10+ clicks and link redirections, entire image downloaded from LPDAAC

* **But Landsat-8 archive is mirrored on both Google Cloud and AWS as cloud-optimized geotiffs! Wouldn't it be nice to search and retrieve a link to the file you want in the Cloud you are working in? People are starting to achieve this:**

* **https://github.com/radiantearth/stac-spec/blob/master/roadmap.md**

New opportunity to conduct InSAR studies on AWS Cloud

https://aws.amazon.com/earth/

→ ***Beta release: 2018**

My thoughts on “Cloud-optimized”

- Easily discover imagery (both raw data NASA or derived)
- Download image subsets instead of the entire files based on georeferenced windows or polygons
- Run algorithms where imagery is stored, and download only results
- Easy dissemination of results via URLs and maps
- Scale analysis (global, high spatial and temporal resolution)
- Workflows deployable on any cloud-provider (AWS, GC, Azure…)
- Shouldn’t have to worry about security
- Costs should be transparent

- **Allows for HTTP GET range requests to only retrieve portions of full file**
- **Metadata and overviews stored in front for speedy access**
- **Otherwise, it's the same-old geotiff and works in GIS programs!**

***How to create and test your own:**

https://trac.osgeo.org/gdal/wiki/CloudOptimizedGeoTIFF

Cloud-Optimized Geotiffs + SpatialTemporal Asset Catalogs

https://dinosar-stac.netlify.com

ARD

Analysis Ready Data

STAC

SpatioTemporal Asset Catalog

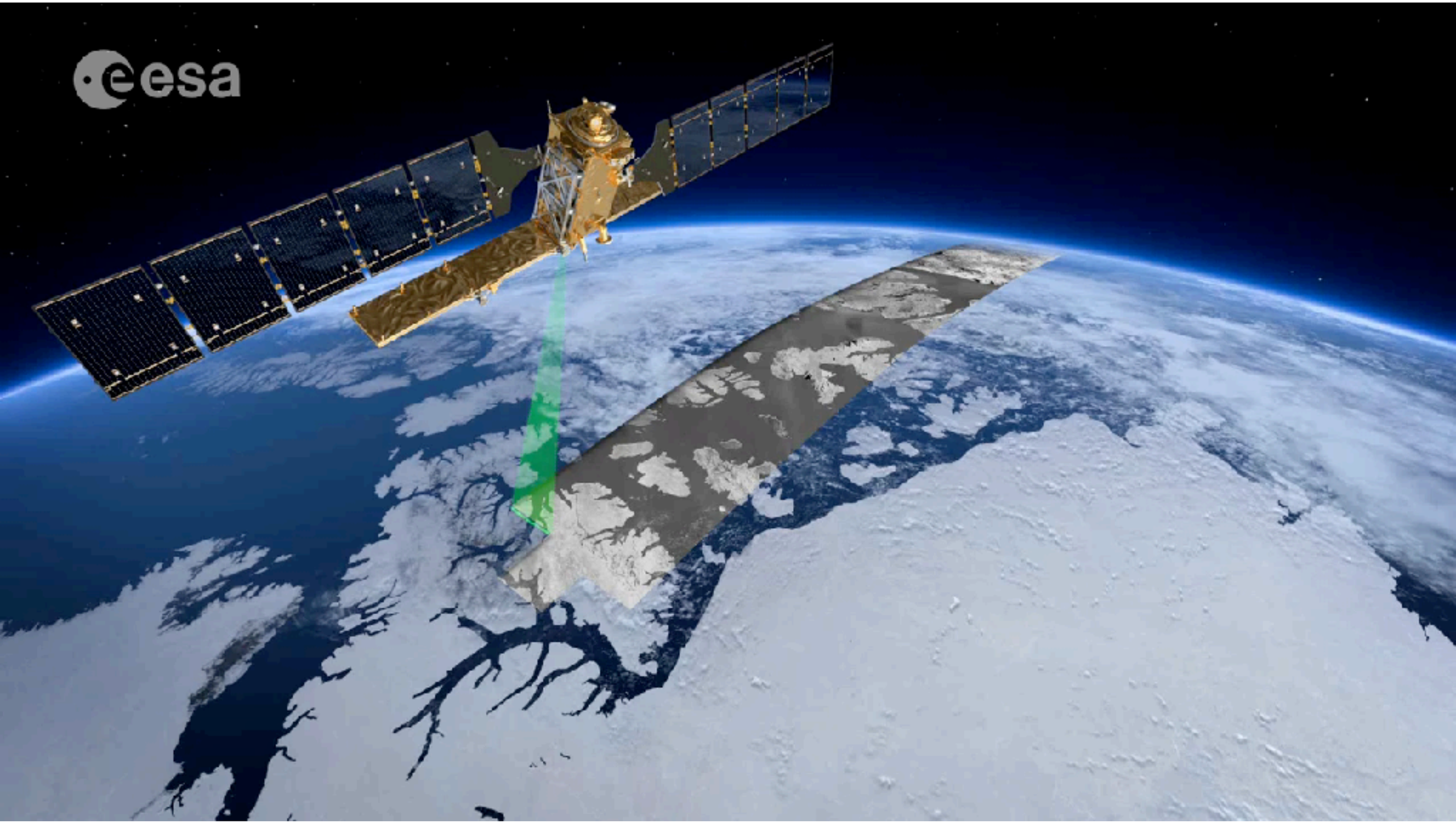
The first Satellite Data Interoperability Workshop is happening next week!

(August 13, 2018)

ACCESS 2017:

Science use cases

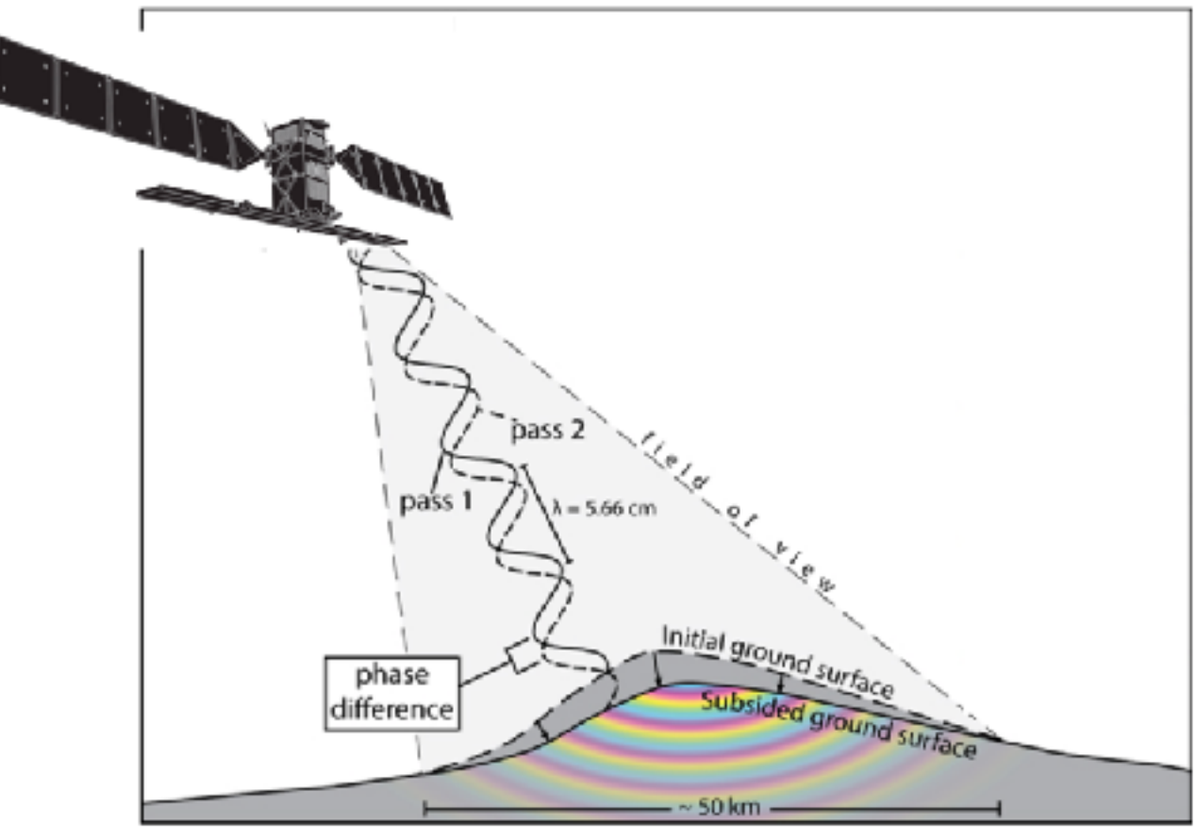
Why is satellite radar so important?

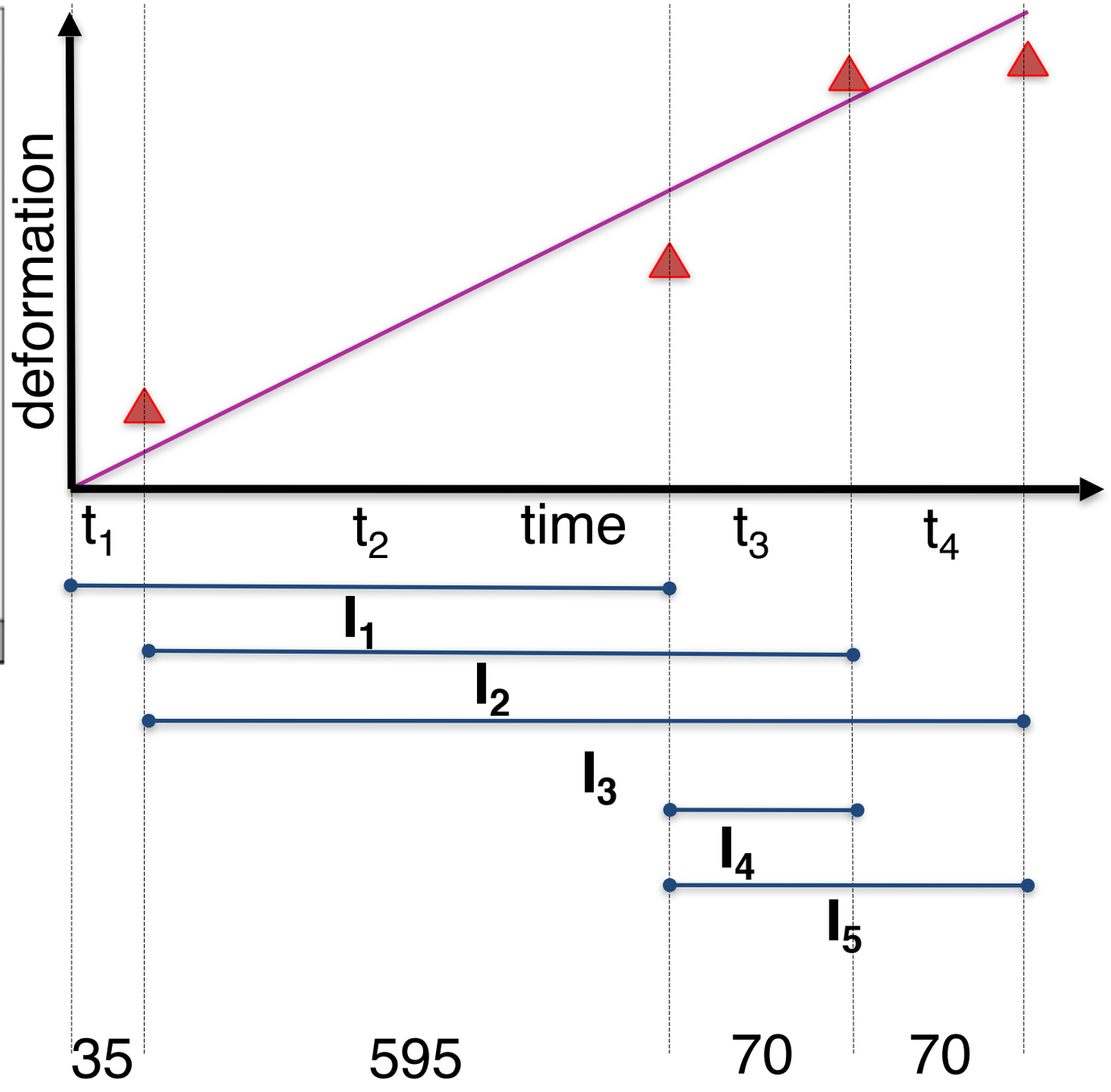


Artist’s rendition of SAR image acquisition (amplitude return shown)

What is “InSAR”?

“Interferometric Synthetic Aperture Radar”

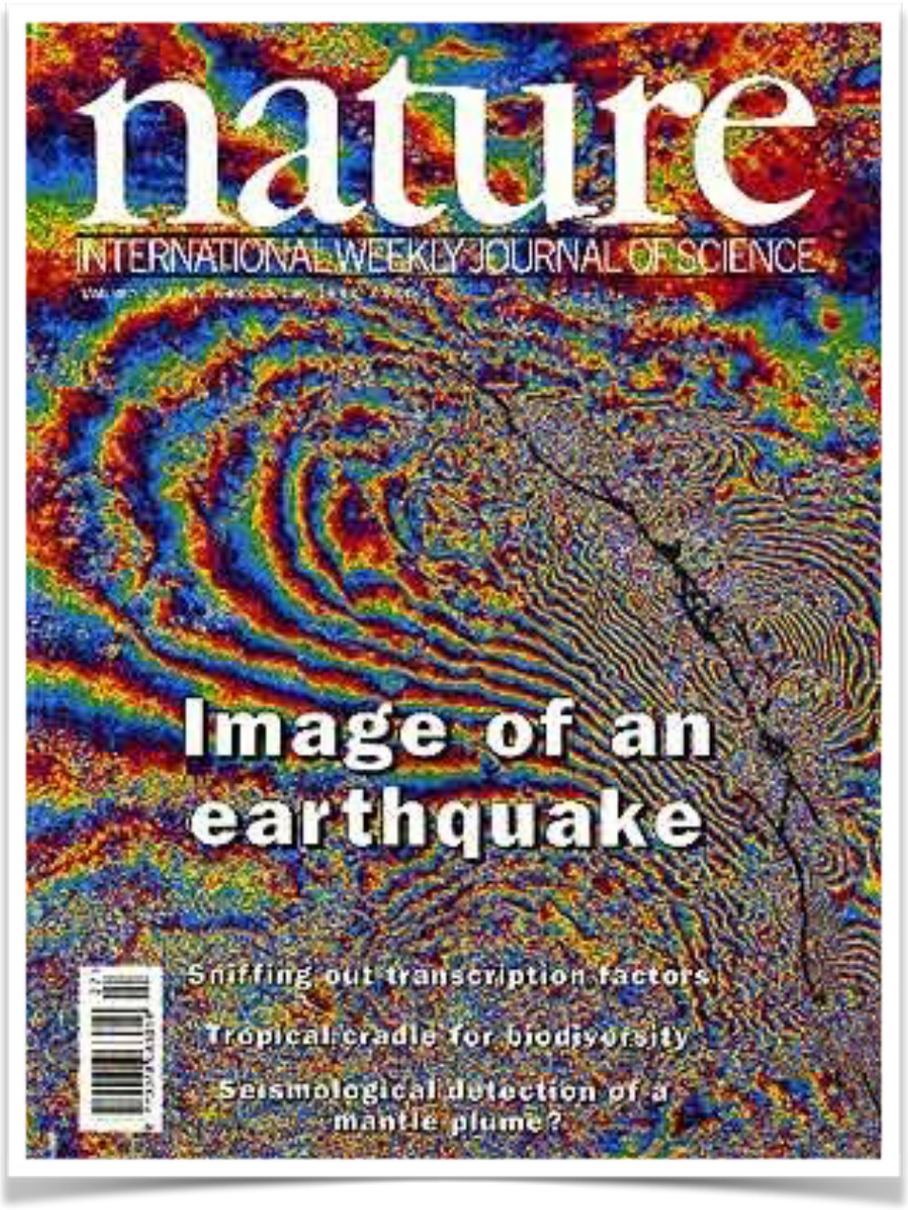




* Phase component from radar return on two dates is differenced to measure **cm-scale** ground displacements

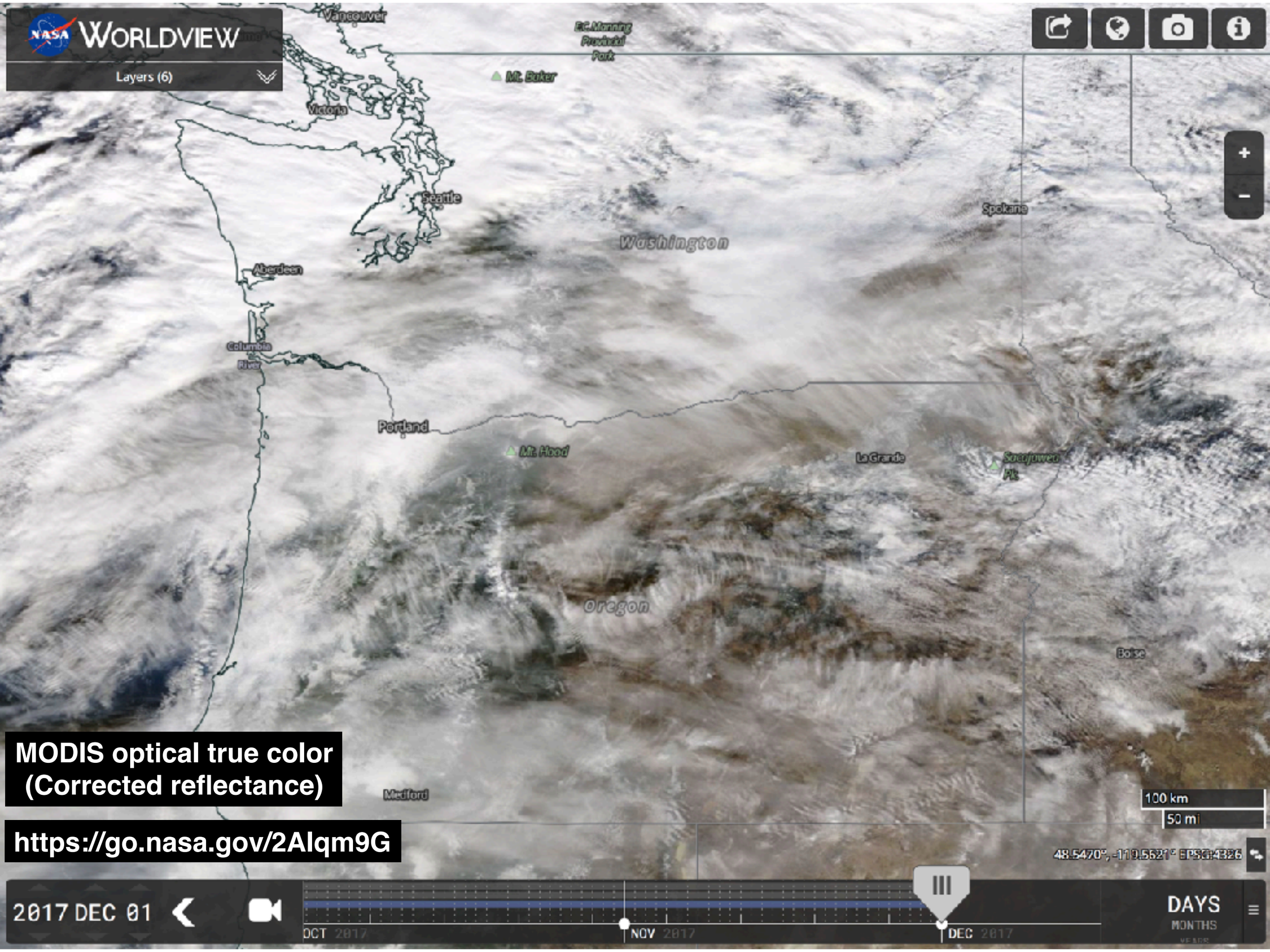
* Displacement images cover **100’s of kilometers with 30-meter pixels**

The displacement field of the Landers earthquake mapped by radar interferometry

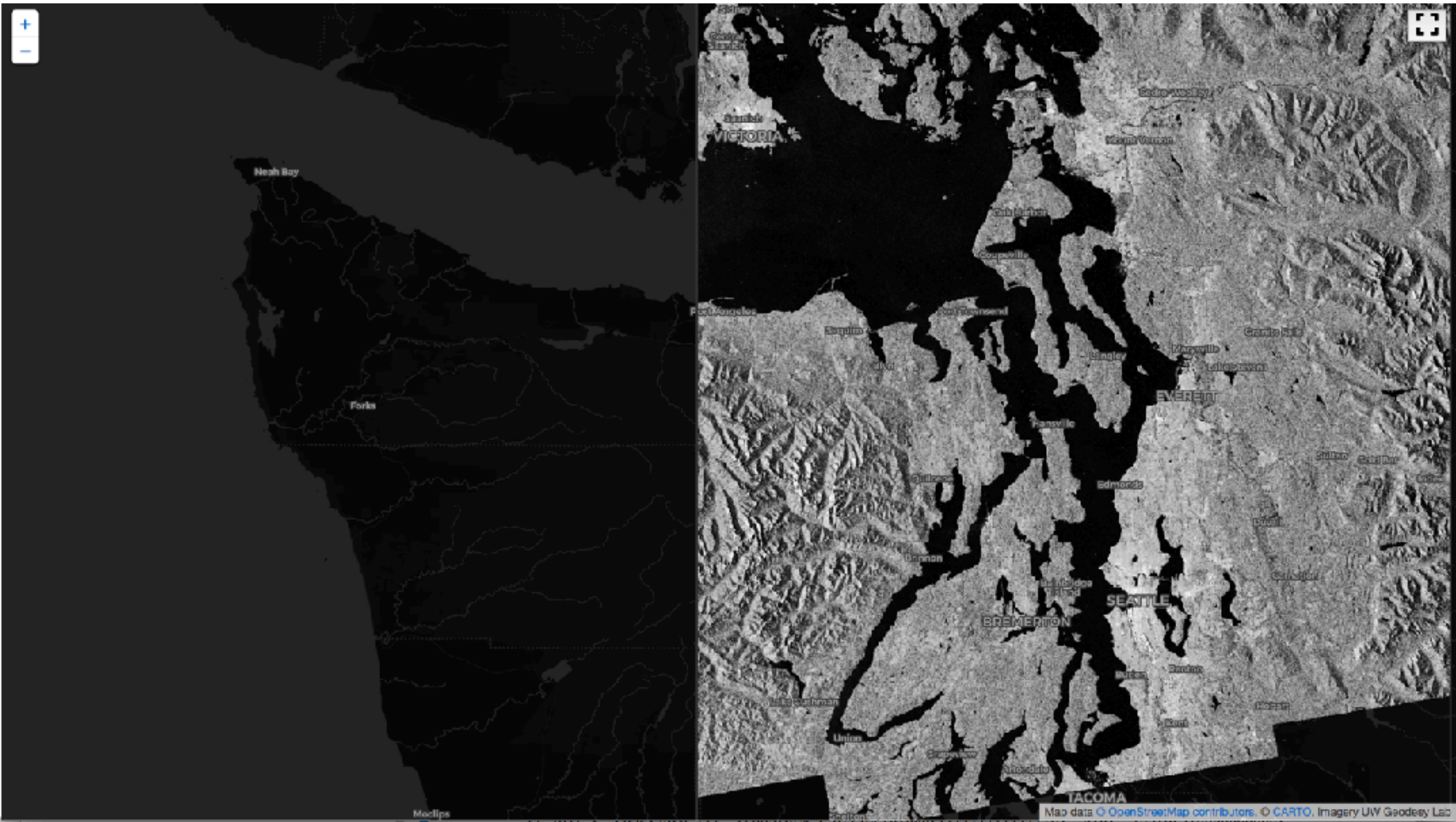


*** July 1993 - the beginning of InSAR used to study Earth Hazards.**

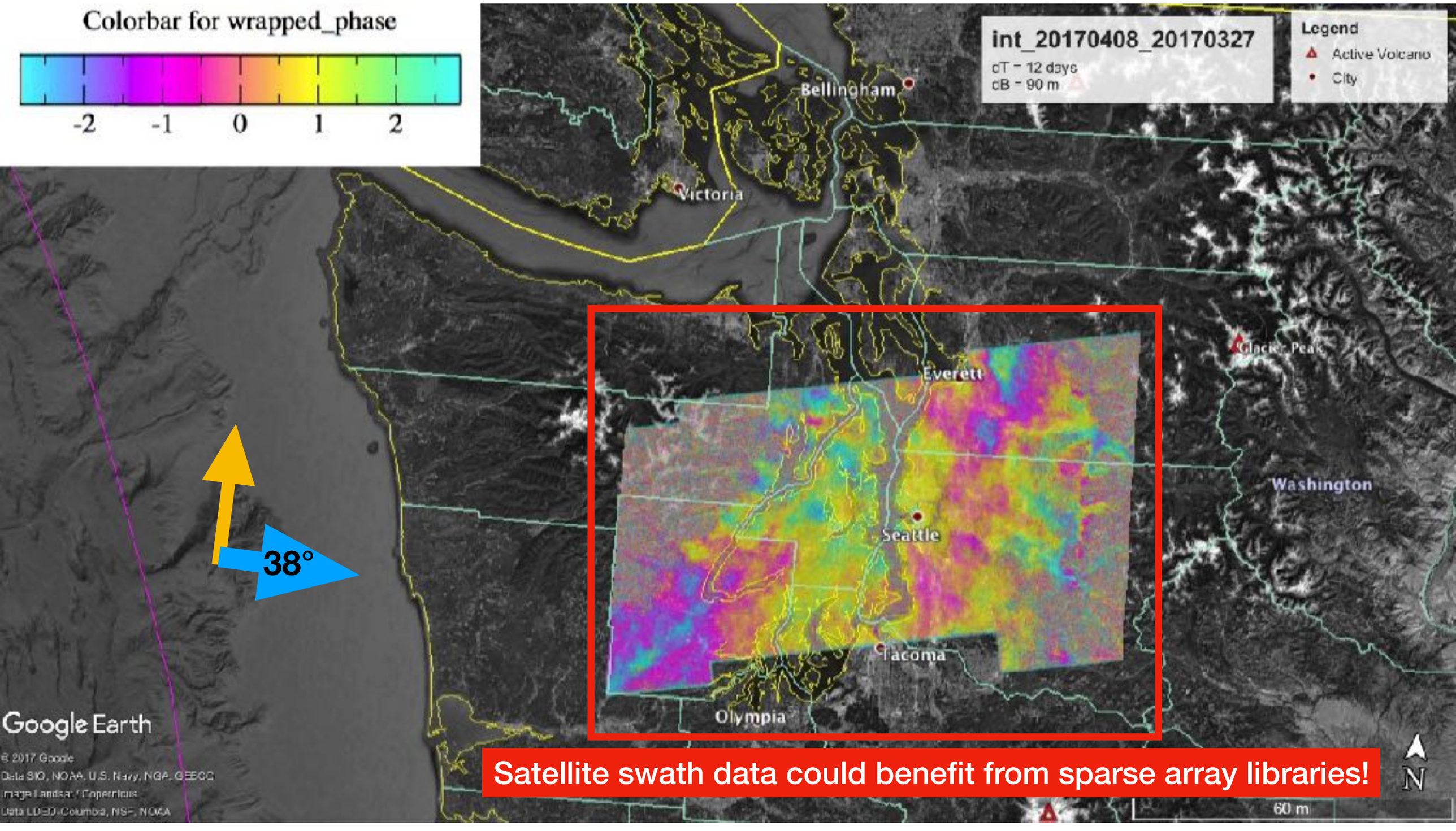
ESA ERS satellite data


WORLDVIEW
Layers (6)
Vancouver
Victoria
Seattle
Aberdeen
Columbia River
Portland
Mt. Baker
E.C. Manning Provincial Park
Washington
Spokane
Mt. Hood
La Grande
Oregon
Boise
Medford
MODIS optical true color
(Corrected reflectance)
https://go.nasa.gov/2AIqm9G
100 km
50 mi
2017 DEC 01
OCT 2017
NOV 2017
DEC 2017
DAYS
MONTHS

SAR solves the cloudy day problem ;)



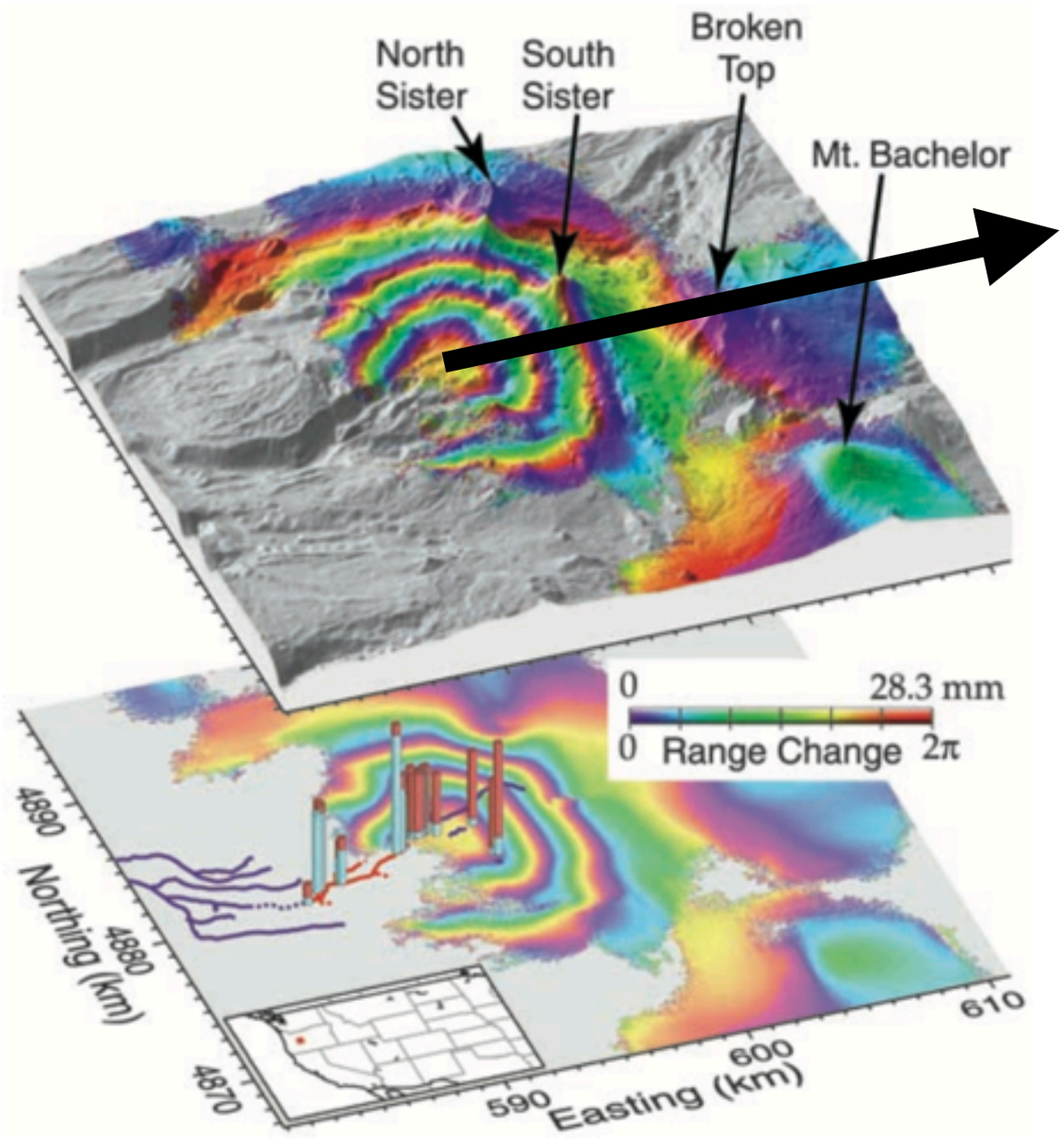
Radar phase gives information about atmospheric water vapor and surface change

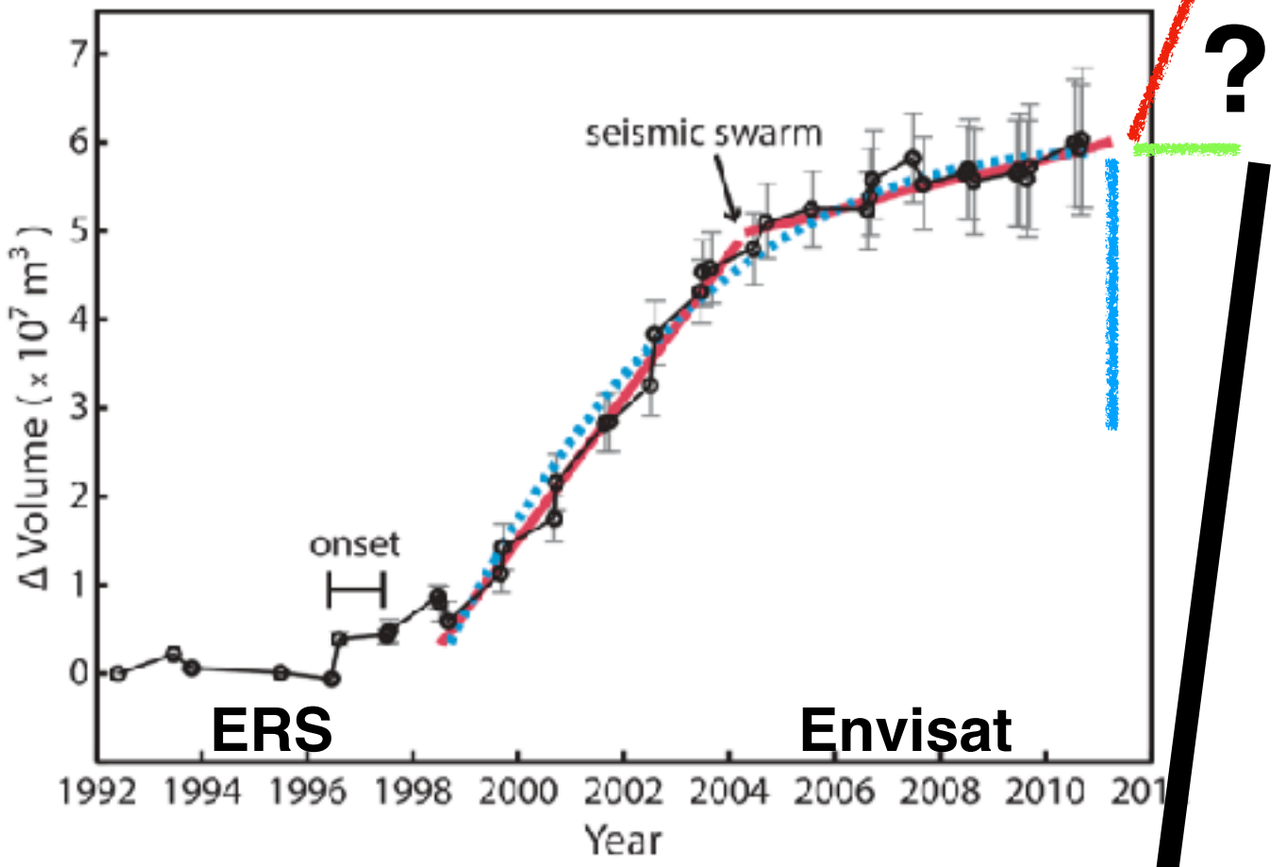


Satellite swath data could benefit from sparse array libraries!

Surveillance of volcanoes: Aseismic inflation

Aseismic deep magmatic intrusion at Three Sisters, Oregon





Riddick & Schmidt 2011, G^3

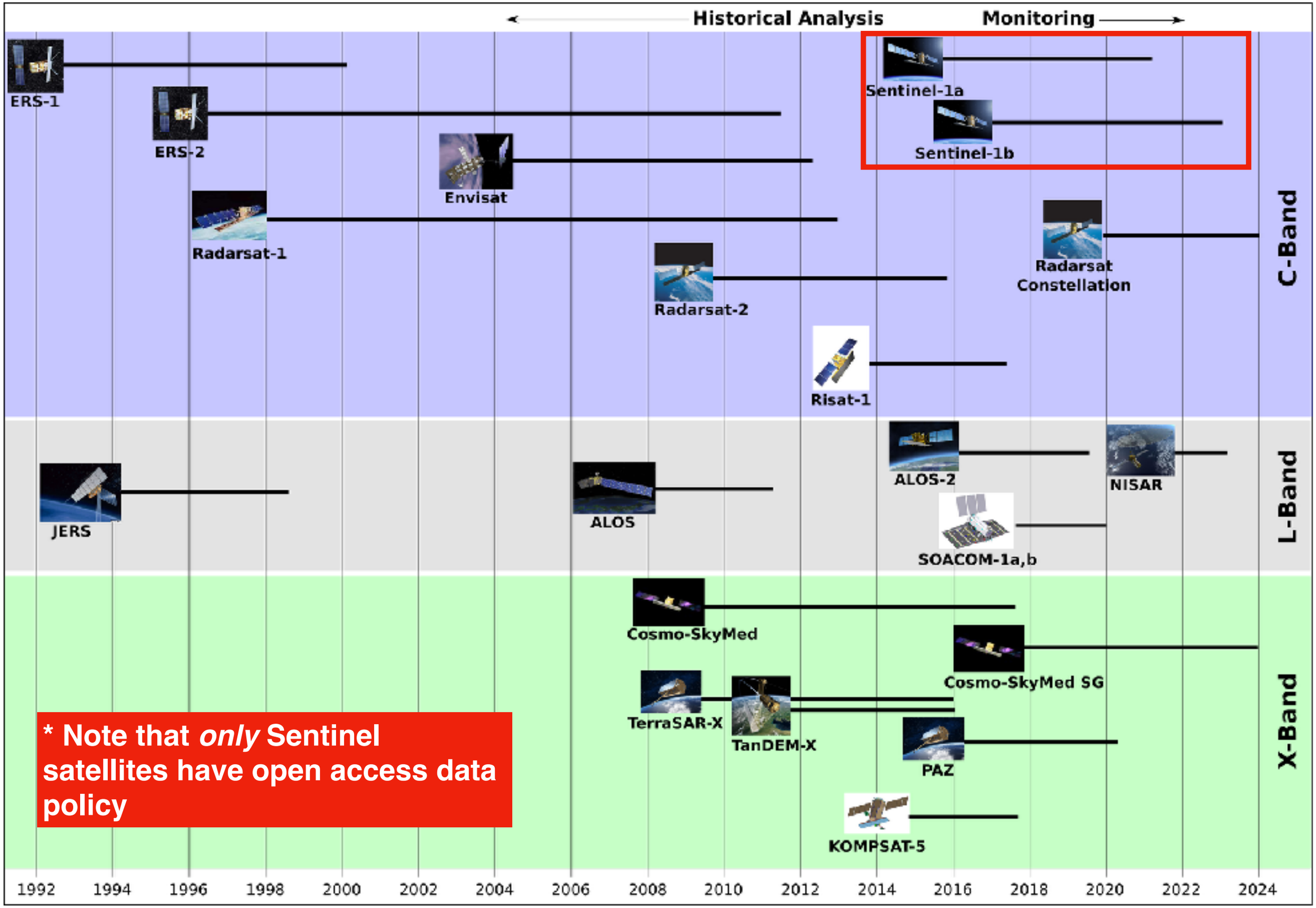
Weekly observations, rather than monthly or yearly, will allow more precise timing of dynamic transitions, and correlation with other measurements (GPS, seismic, gas)

Wicks et al 2002, GRL

https://volcanoes.usgs.gov/volcanoes/three_sisters/three_sisters_geo_hist_129.html

Synthetic Aperture Radar Satellites

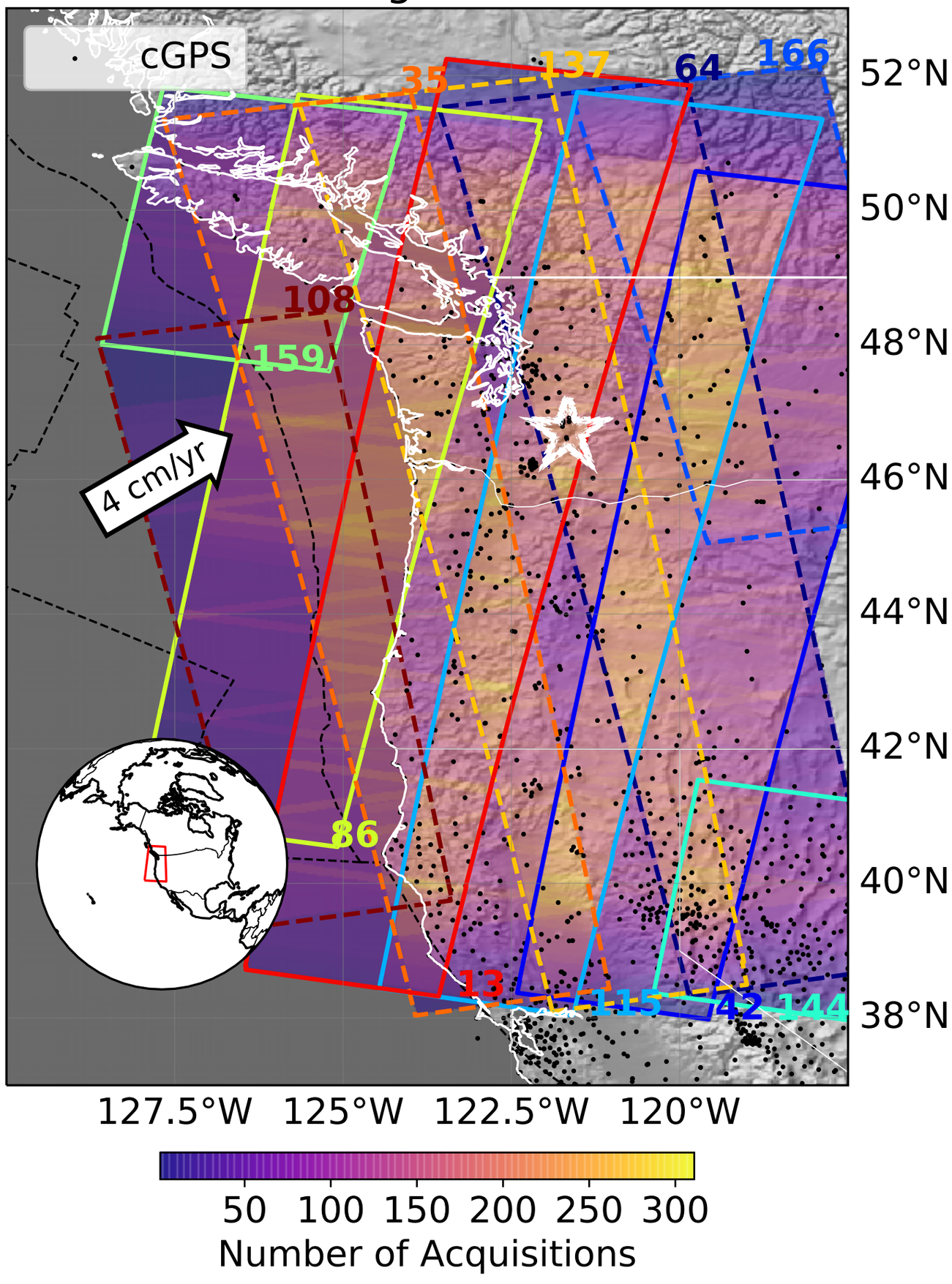
*Today



*** Note that *only* Sentinel satellites have open access data policy**

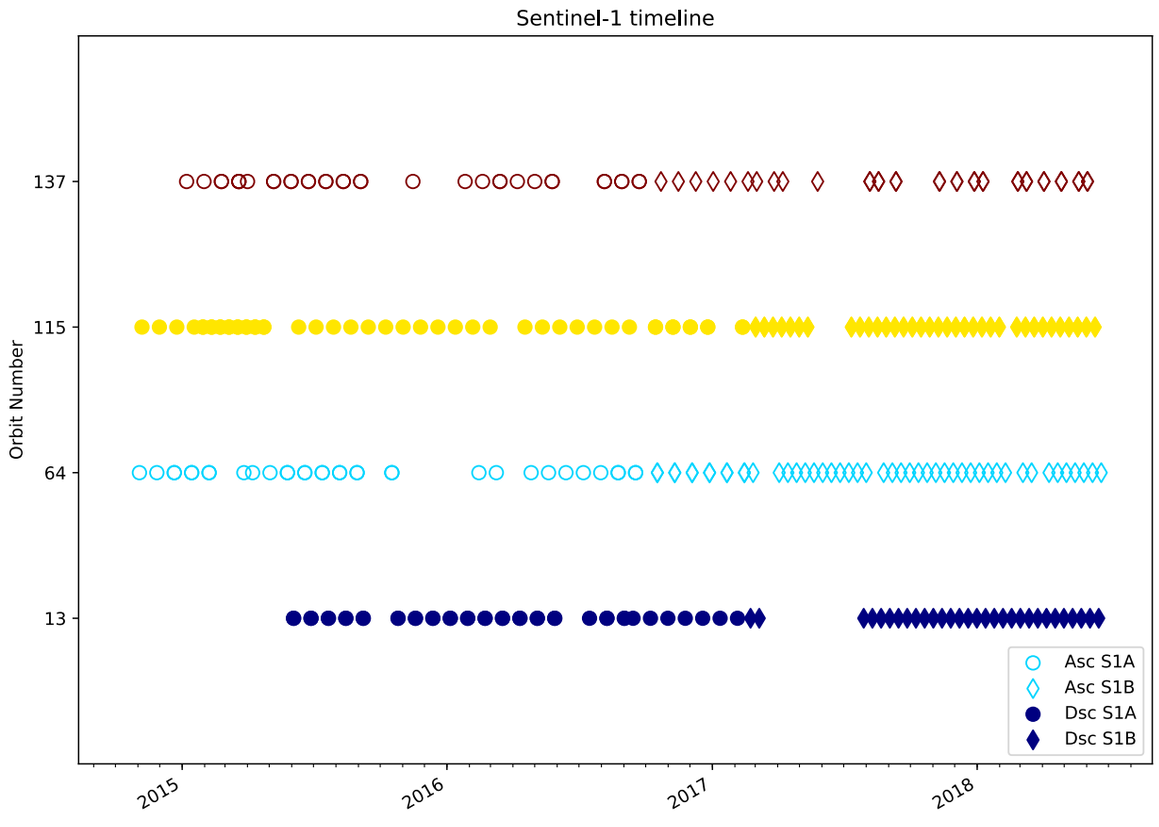
https://www.unavco.org/instrumentation/geophysical/imaging/sar-satellites/sar-satellites.html

Sentinel-1 Coverage 10/2014 - 01/2018



***Minimum requirements for entire region: 25 Tb (30-m resolution)**

*** Currently scaling up processing for entire region, starting with Washington State**



InSAR in the Cloud

https://github.com/scottyhq/dinoSAR

amazon web services

"dinoSARaws"

GOAL:
facilitate batch processing of S1 InSAR

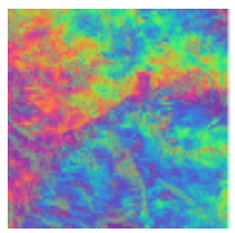
AWS Batch

+

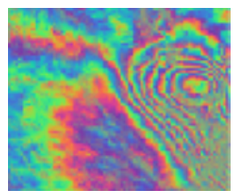
docker

(Ubuntu w/ ISCE 2.1.0)

1)



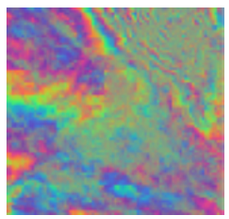
2)



•
•
•

Amazon EFS

n)



***Cloud Optimized Geotiff format (COG)**

Amazon S3

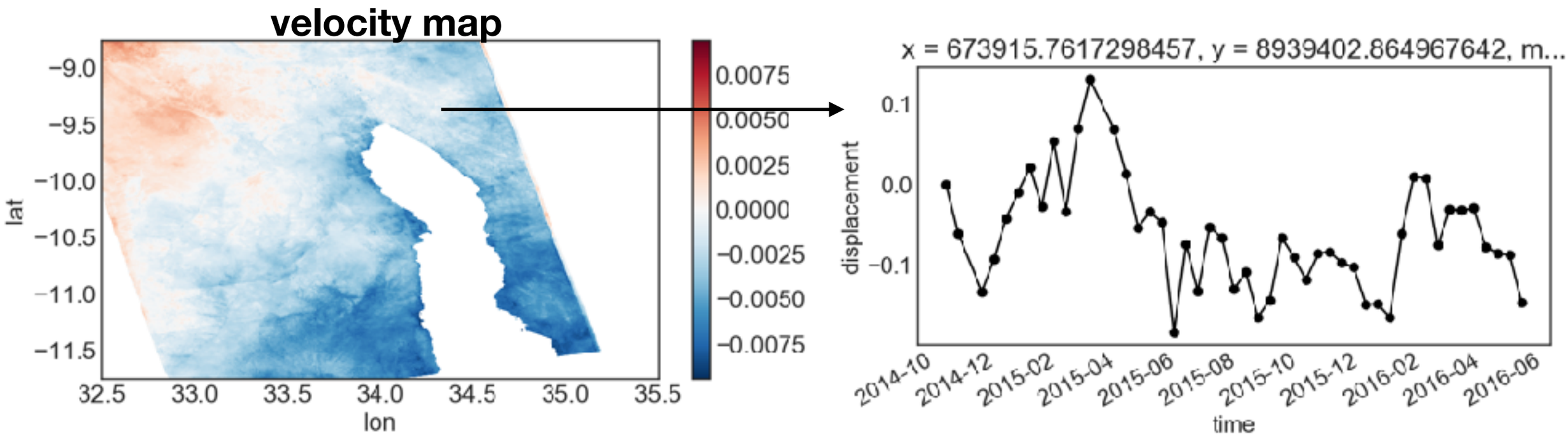
+ STAC JSON metadata

https://dinosar-stac.netlify.com

Initial tests w/ Xarray and Dask

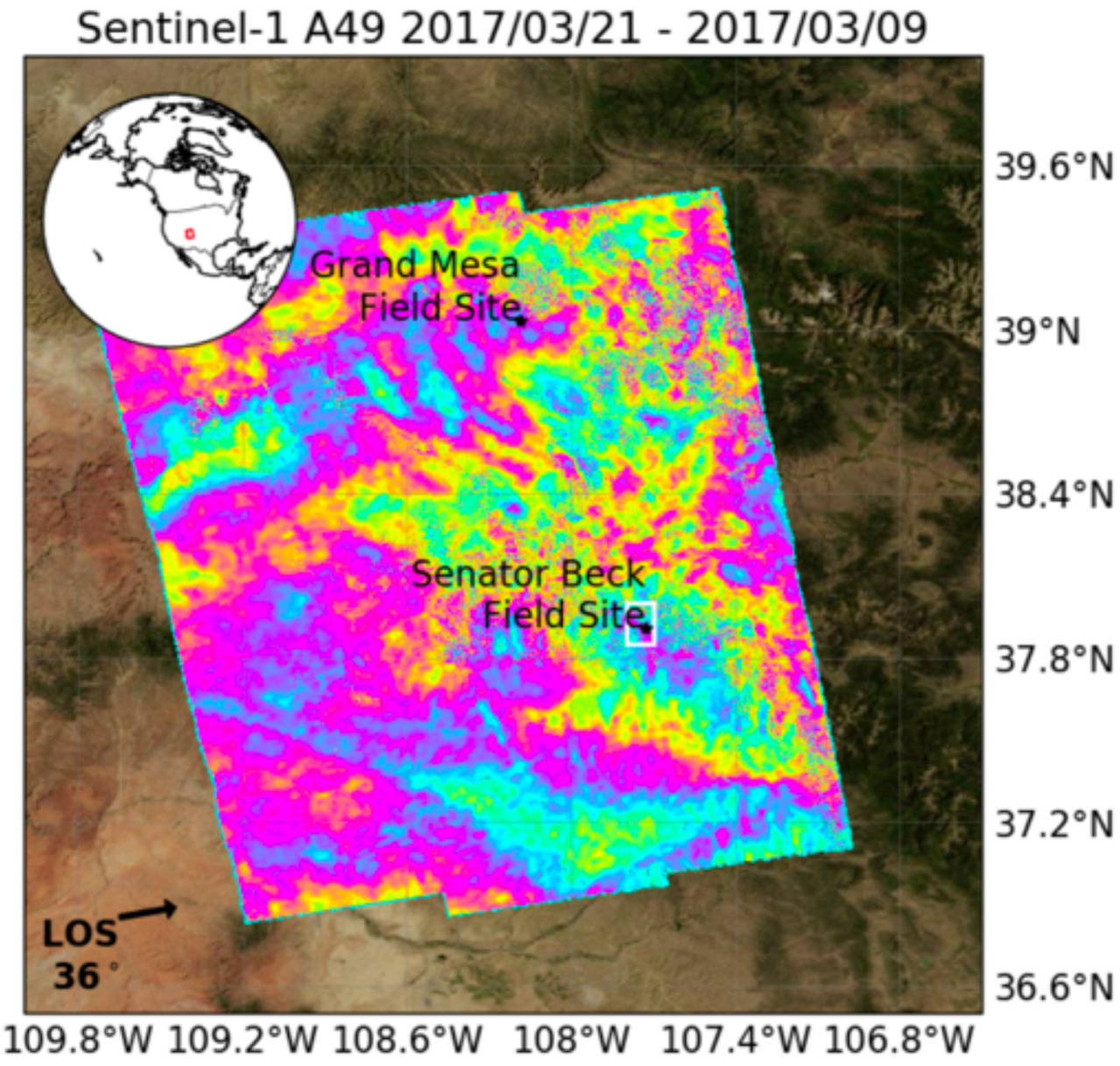
```
<xarray.DataArray 'stack' (pair: 119, lat: 1350, lon: 1350)>
dask.array<stack, shape=(119, 1350, 1350), dtype=float32, chunksize=(1, 1350, 1350)>
Coordinates:
  * lat      (lat) float64 -11.75 -11.75 -11.75 -11.74 -11.74 -11.74 -11.74 ...
  * lon      (lon) float64 32.5 32.5 32.51 32.51 32.51 32.51 32.52 32.52 ...
  * pair     (pair) <U21 'int_20141117_20141012' 'int_20141117_20141024' ...
Attributes:
    description:  SEGMeNT S1 A101 Timeseries
    created:      2017-11-02 00:00:00
    author:       Scott Henderson
    ISCE:         2.1.0
```

*Hard to ingest data to Xarray, but once there, great convenience

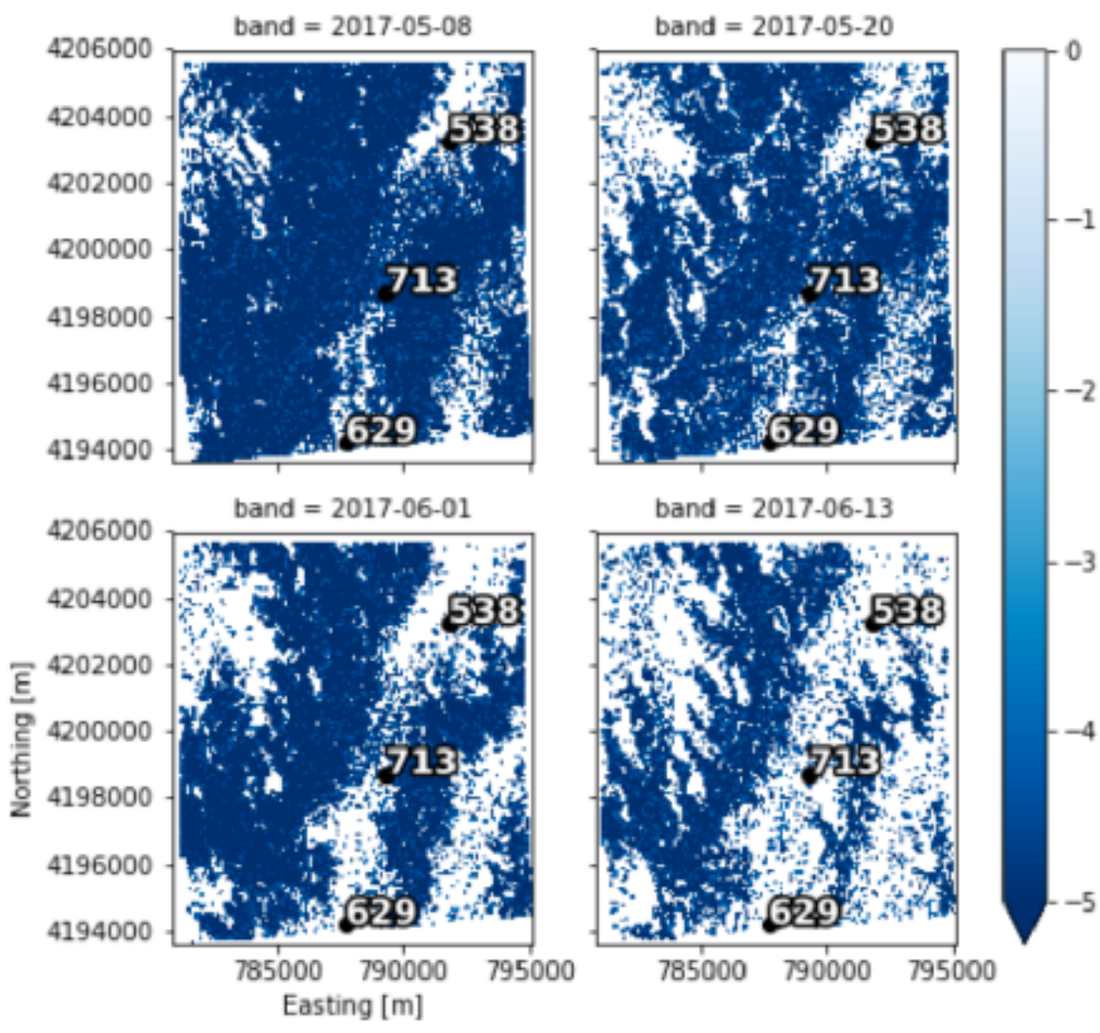


InSAR for Snow

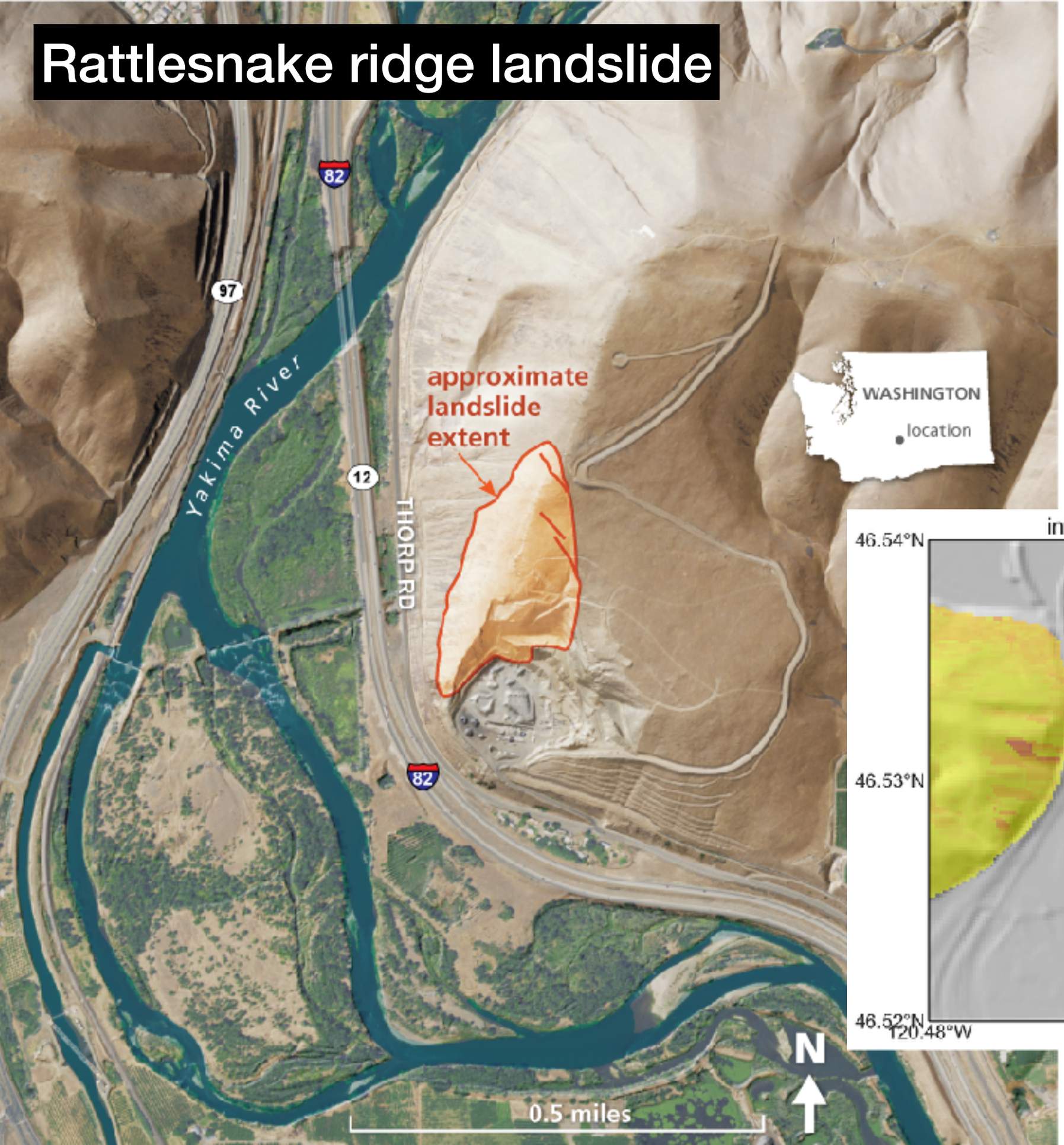
Possible high-resolution maps of SWE?



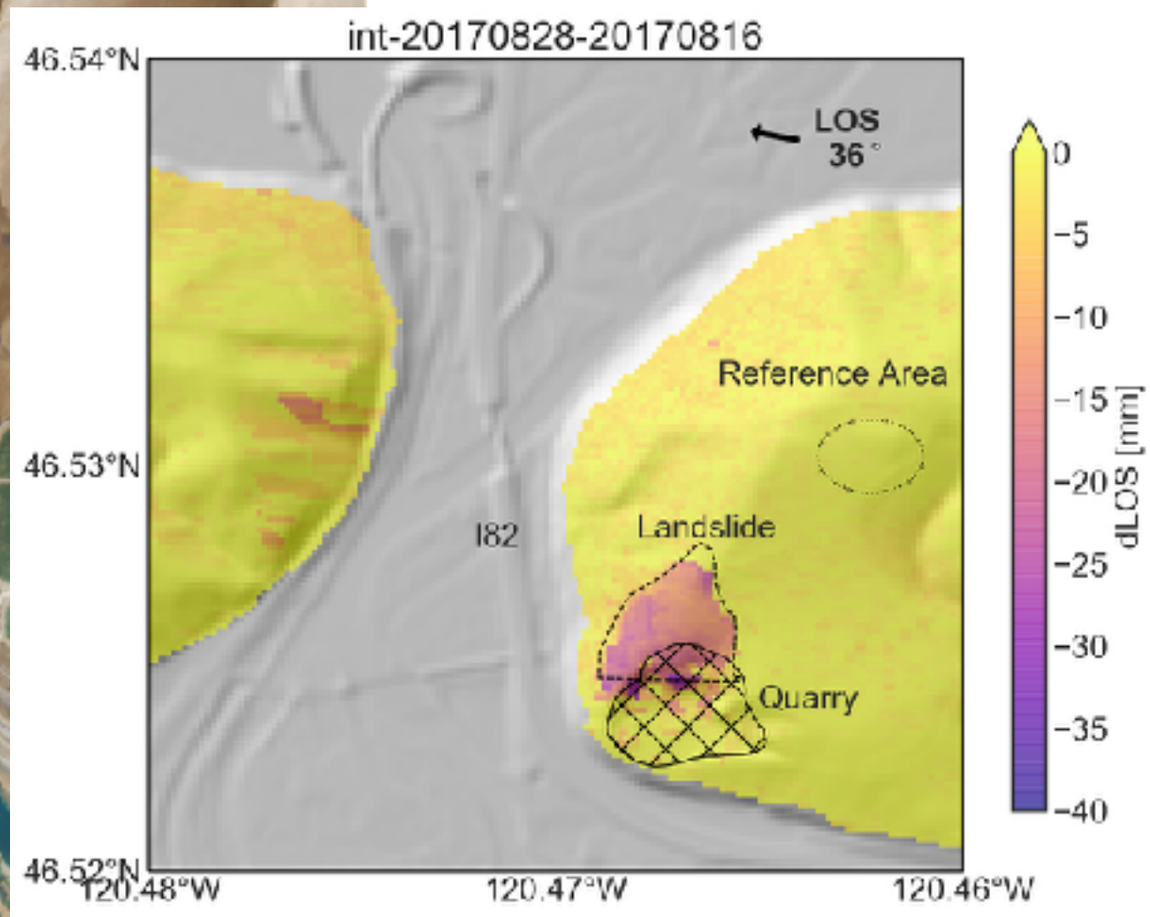
Backscatter maps
snowmelt through seasons



Rattlesnake ridge landslide



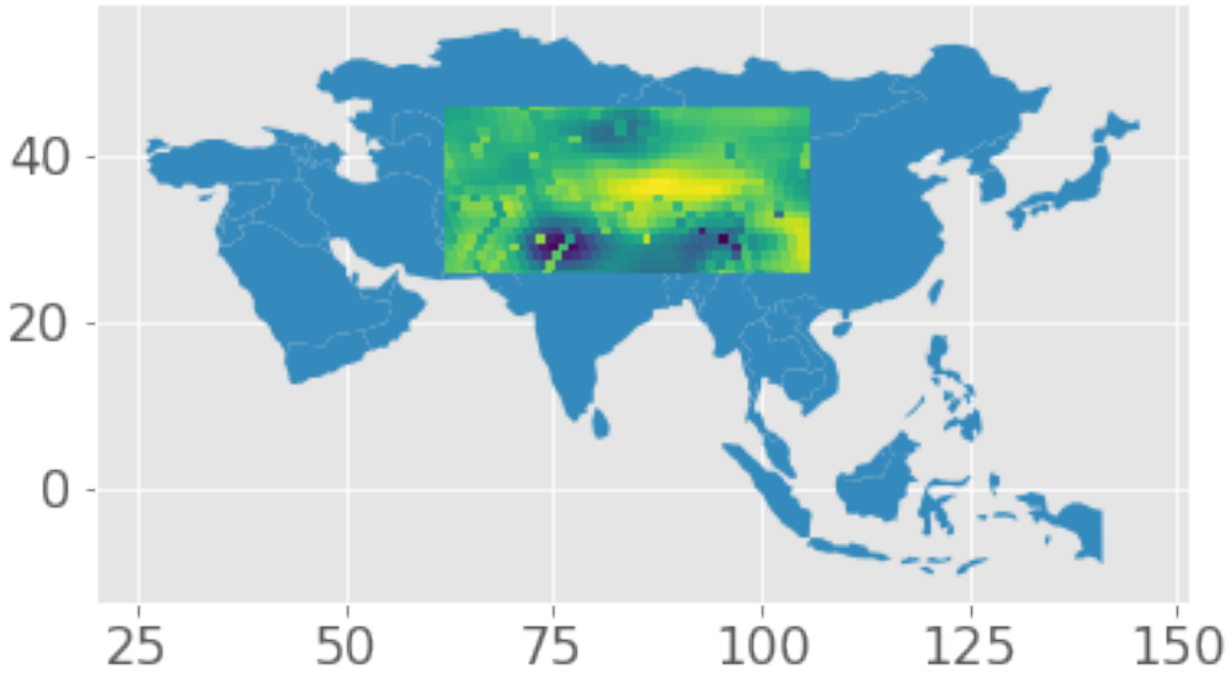
*** Need to process at InSAR at 15m resolution to adequately measure motion**

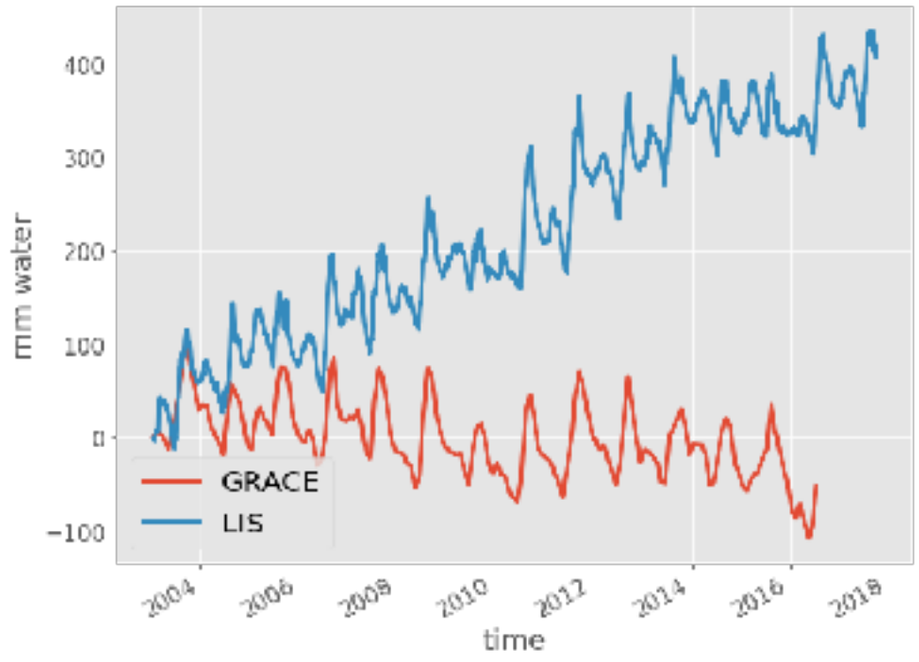


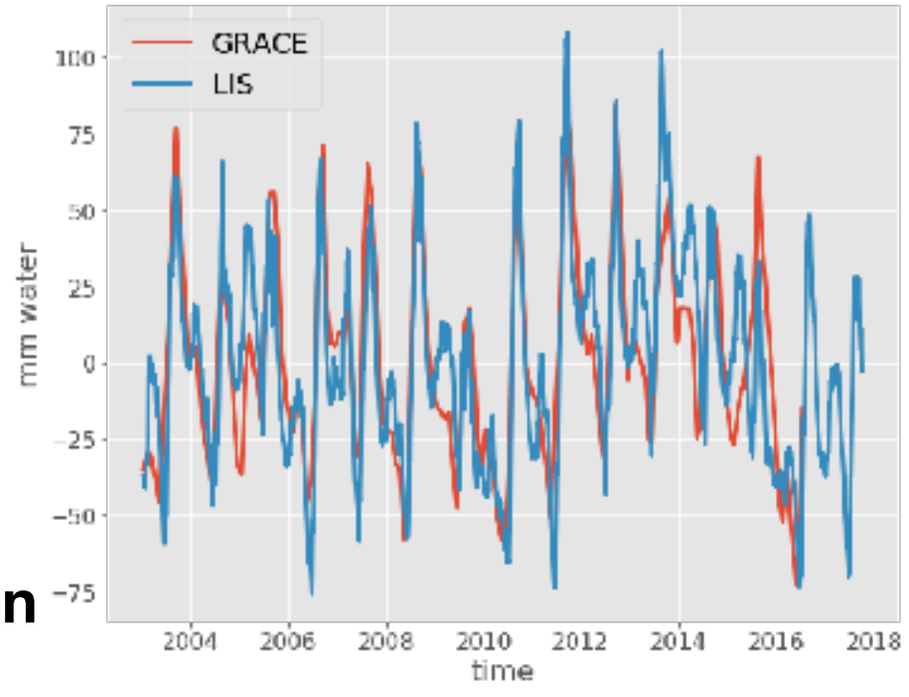
Henderson et. al. - in prep

GRACE in the Cloud

https://github.com/NASA-Planetary-Science/HiMAT/blob/master/Projects/GRACE_MASCON/GRACE_water_balance_validation.ipynb



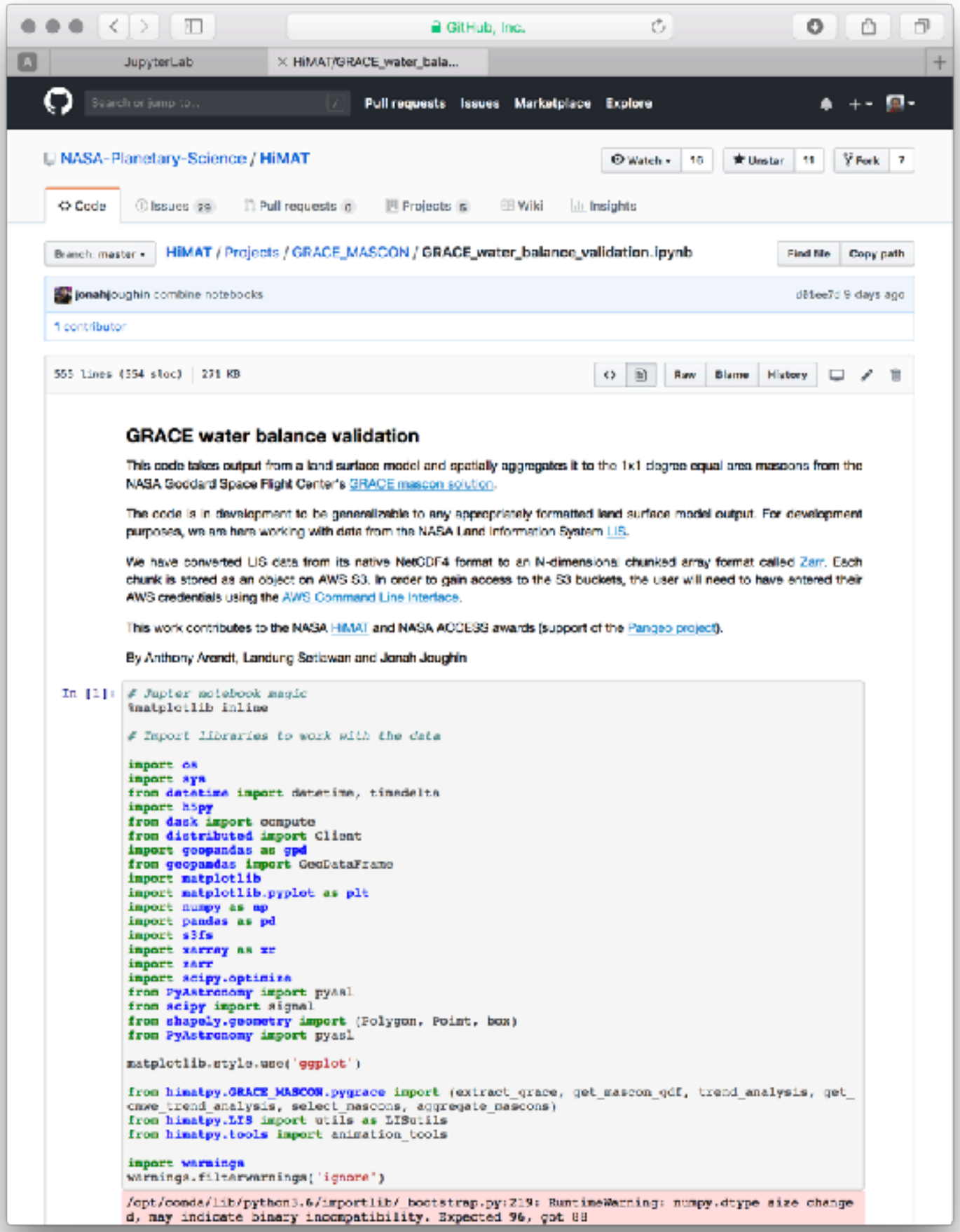




* **Output from a high resolution (1km) land surface model from NASA Land Information System (LIS) is spatially aggregated to 1x1 degree equal area GRACE mascons for all of High Mountain Asia (HiMAT)**

*** Work by Anthony Arendt, Jonah Joughin, Don Setiawan**

https://github.com/communitysnowobs/pangeo



HashiCorp Terraform

- **Single User Pangeo on AWS EKS**
- **Dask Kubernetes**
- **~300 Gb stored in Zarr format on S3**

To run:

>>> terraform apply
>>> kubectl apply -f aws-auth.yml
>>> kubectl create -f kube_specs/aws/

To close:

>>> terraform destroy

~$2 dollars / hour
(depending on EC2 instances)

Conclusions and Questions

- What specifically are the costs for keeping a Pangeo JupyterHub running now and into the future? Can costs be reduced significantly with design tweaks?
    - 10 years from now, if I want to re-run a pangeo.pydata.org notebook to reproduce an InSAR analysis, will I be able to?
- How to get best of both worlds with Cloud-Optimized Geotiffs + Xarray, Zarr, and Dask.
- Research with satellite data often leads to re-used large derived datasets. Also valuable to agencies for hazard monitoring and modeling. We need:
    - 1) Scalable computing
    - 2) Easier dissemination of results

Thanks!