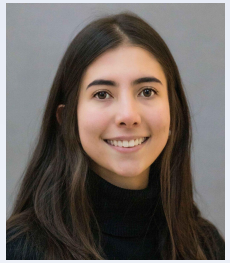
Enabling Scientific Discovery: Harnessing the Power of the National Science Data Fabric for Large-Scale Data Analysis (Session III & IV)

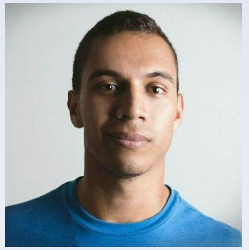
Presenters: Aashish Panta[2], Amy Gooch[2], Jack Marquez[1], Valerio Pascucci[2], and Michela Taufer[1]
Assistants: Heberth Martinez[1]
Other contributors: Nina McCurdy[3], David Ellsworth[3], Patrice Klein[4], Hector Torres[4], Paula Olaya[1], Gabriel Laboy[1], Jay Ashworth[1], Lauren Whitnah[1], and Giorgio Scorzelli[2]

[1]University of Tennessee Knoxville, [2]University of Utah, [3]NASA Ames Research Center, [4]NASA Jet Propulsion Lab










Acknowledgments

The authors of this tutorial would like to express their gratitude to:

- NSF through awards 2138811, 2103845, 2334945, 2138296, and 2331152
- Dataverse
- Seal Storage
- Rodrigo Vargas, Vargas Lab, University of Delaware
- Werner Sun, CHESS, Cornell University
- DOE SBIR Phase II award DE-SC0017152

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Prerequisites

Step 0: Access to GitHub

To run this tutorial, you need to have a GitHub account.

- ➔ You can create one following the instructions here:

  https://docs.github.com/en/get-started/start-your-journey/creating-an-account-on-github#

- ➔ Now you can login into GitHub

  https://github.com/login

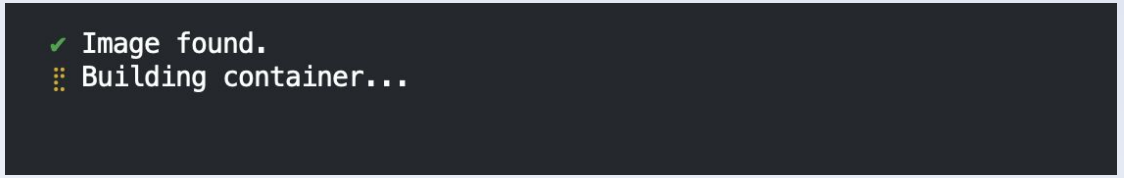
Step 1: Create Codespaces

Use your GitHub account to run this tutorial with GitHub codespaces

- ➔ Access this link:

  NSDF Tutorial 2024

- ➔ Click on green button "Create codespace"



Session III

Tutorial

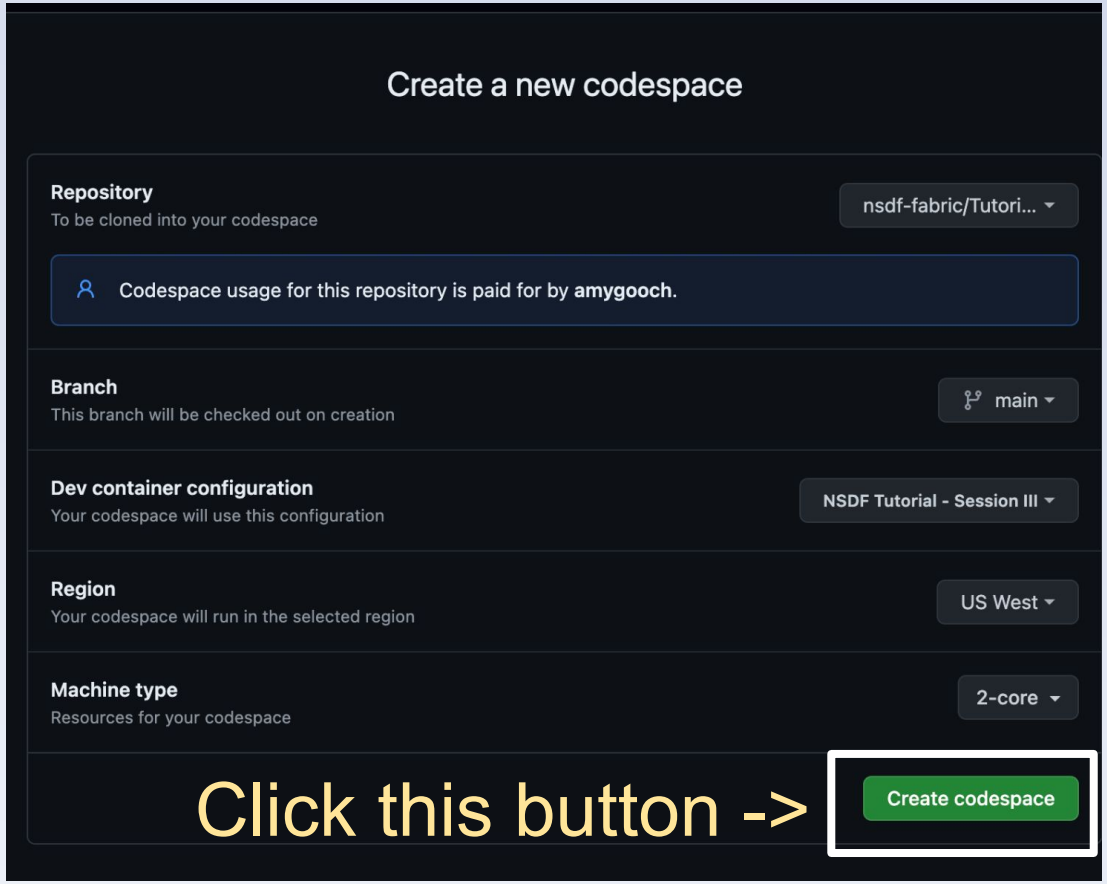
Step 1: Create Codespaces

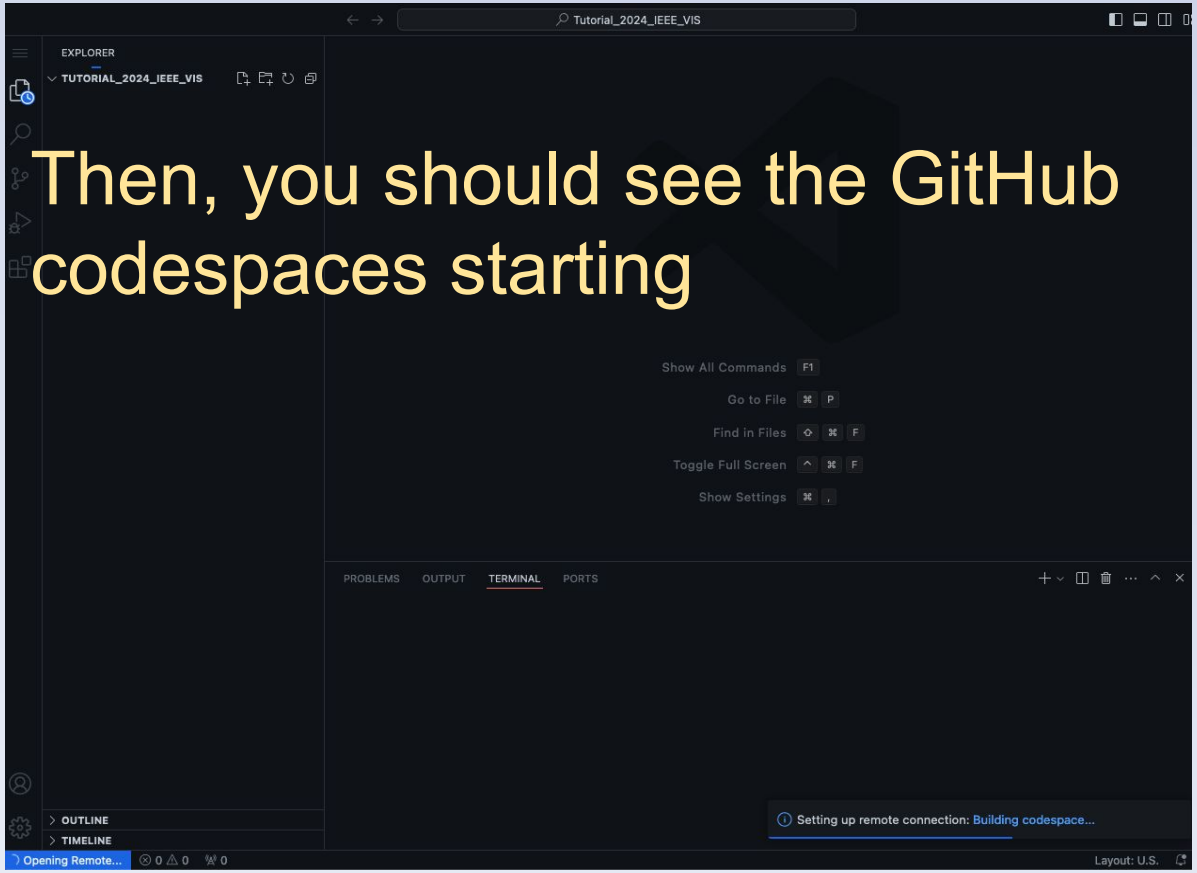
Use your GitHub account to run this tutorial with GitHub codespaces

- ➔ Access this link: NSDF Tutorial 2024
- ➔ Click on green button "Create codespace"

Creating the GitHub Codespace

Tutorial





➔ Let's start:

[NSDF Tutorial 2024]

Loading GitHub Codespace can take from 1 to 5 minutes

Tutorial Session III Goals

Tutorial

This tutorial demonstrates end-to-end analysis of scientific data through NSDF services

Tutorial Goals

- **Access publicly available petascale datasets**
- **Treat the data as a NumPy array and perform statistical analysis**
- Store your data for large-scale **data access, visualization, analysis, and sharing**

Session III: Petascale Data

NASA makes big datasets available

What does it mean to get to that data?

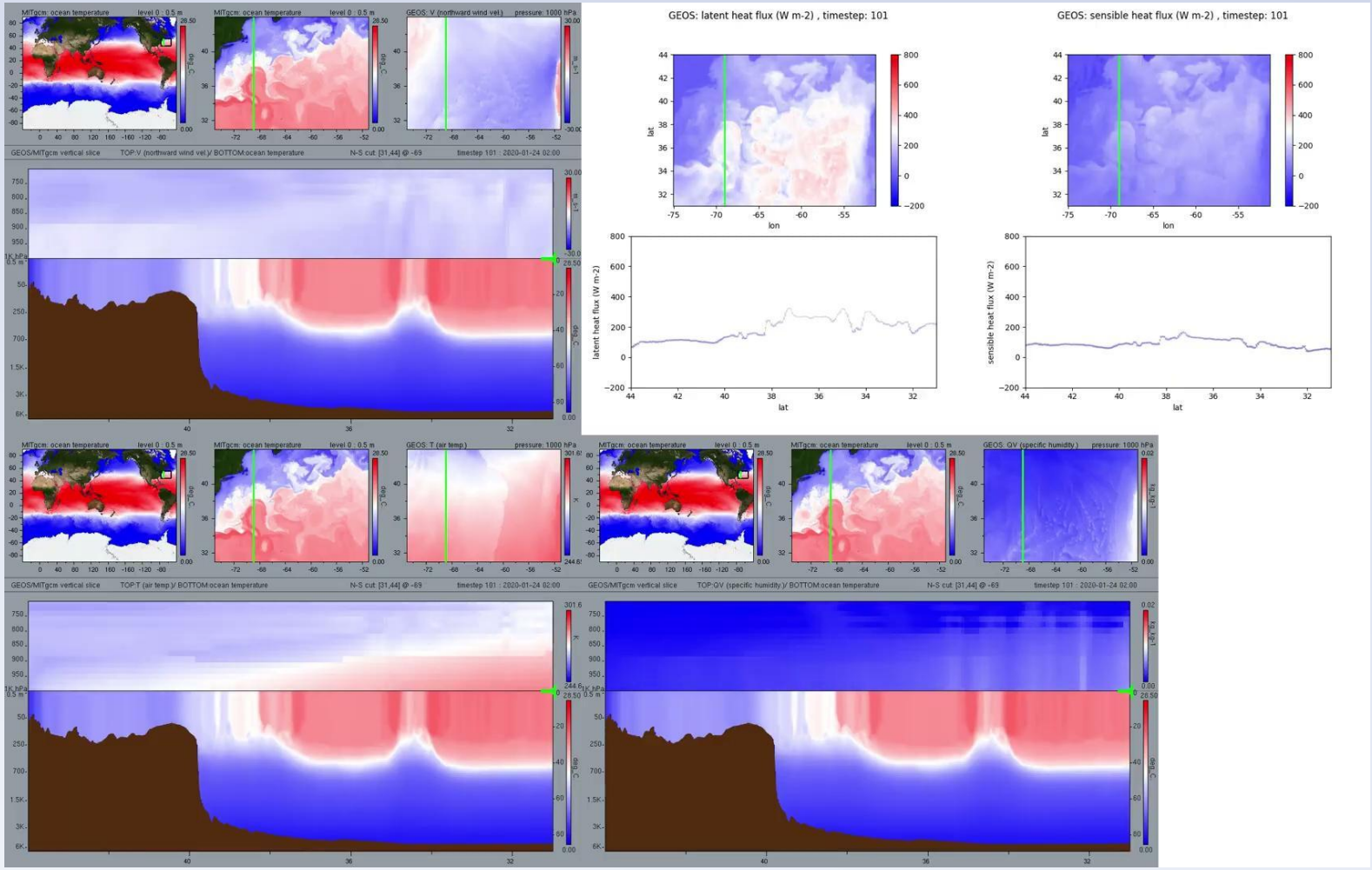
What do you have to do to see it

As a domain expert, such as a climate scientist, geographer?

As a visualization expert?

As a student?

NASA Scientists and Experts in Vis and Climate



- Difficulty accessing the full data
- Limitations in computational power
- Need real-time processing capabilities

Publicly available NASA Datasets

Datasets :

NASA 1.8 PB DYAMOND dataset[1]:
a global atmospheric model and a global ocean model

LLC4320 2.8 PB Ocean dataset[2]

1) https://gmao.gsfc.nasa.gov/global_mesoscale/dyamond_phaseII/data_access/
2) https://www.ecco-group.org/data.html

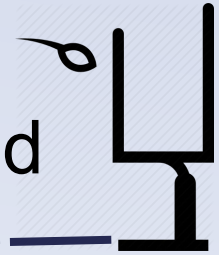
How big is a petabyte?

1,000 ^ 5 = 1,000 x 1,000 x 1,000 x 1,000 x 1,000 = 1,000,000,000,000,000 bytes

11,000 4k Movies

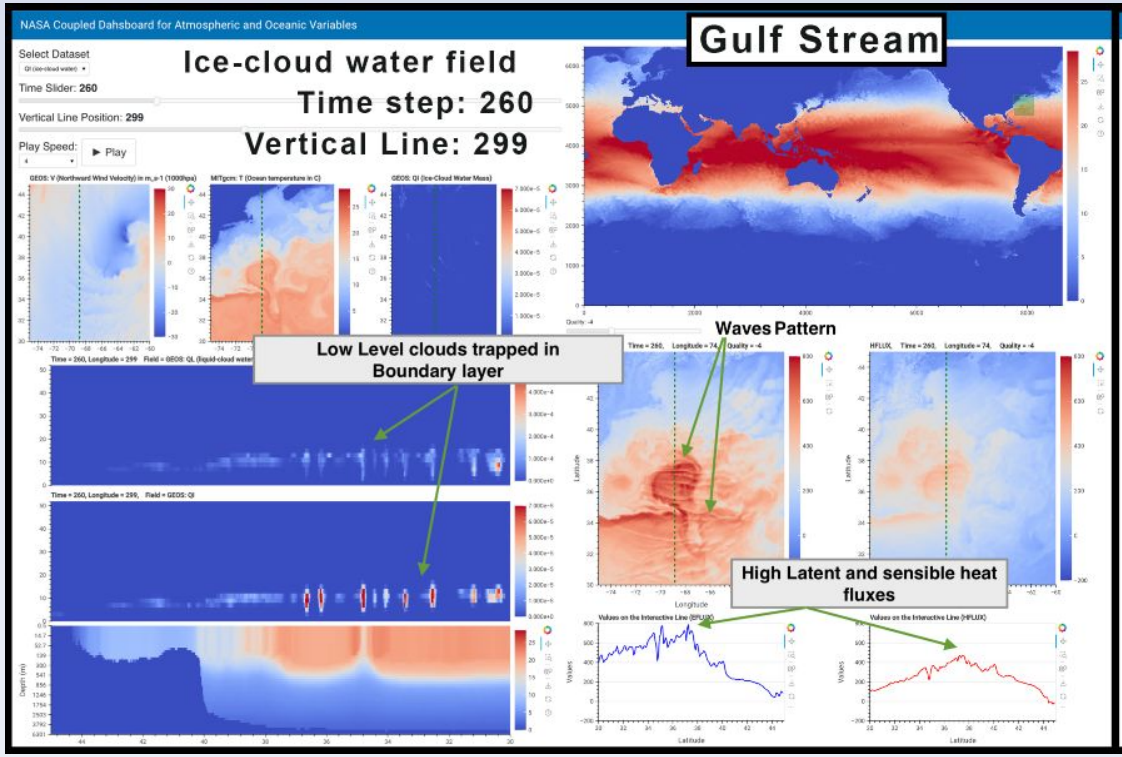
over 2.5 years of non-stop binge watching to get thru a petabyte of 4k movies

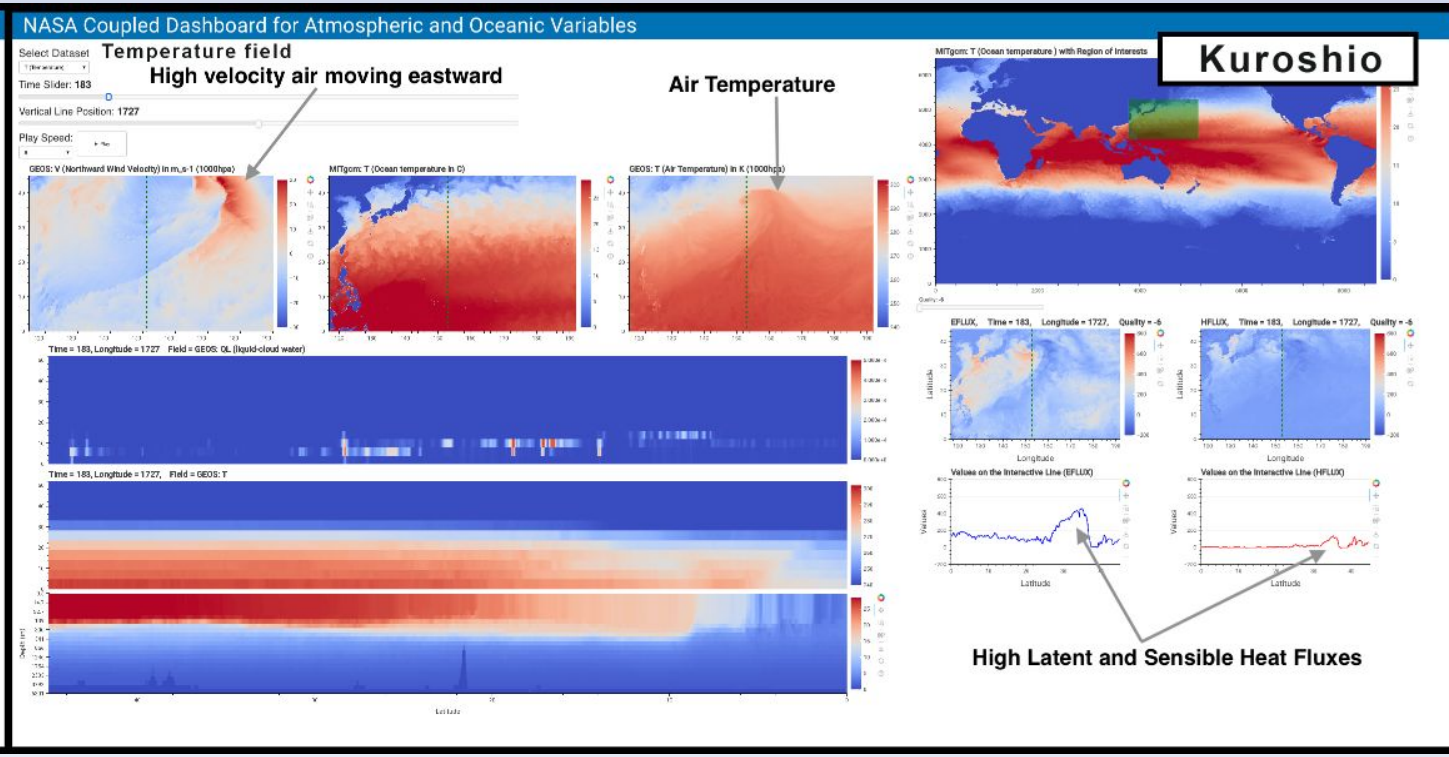
92 football fields of 1GB flash drives put end to end



1 PB is equivalent to taking over 4,000 digital photos per day, over your entire life.

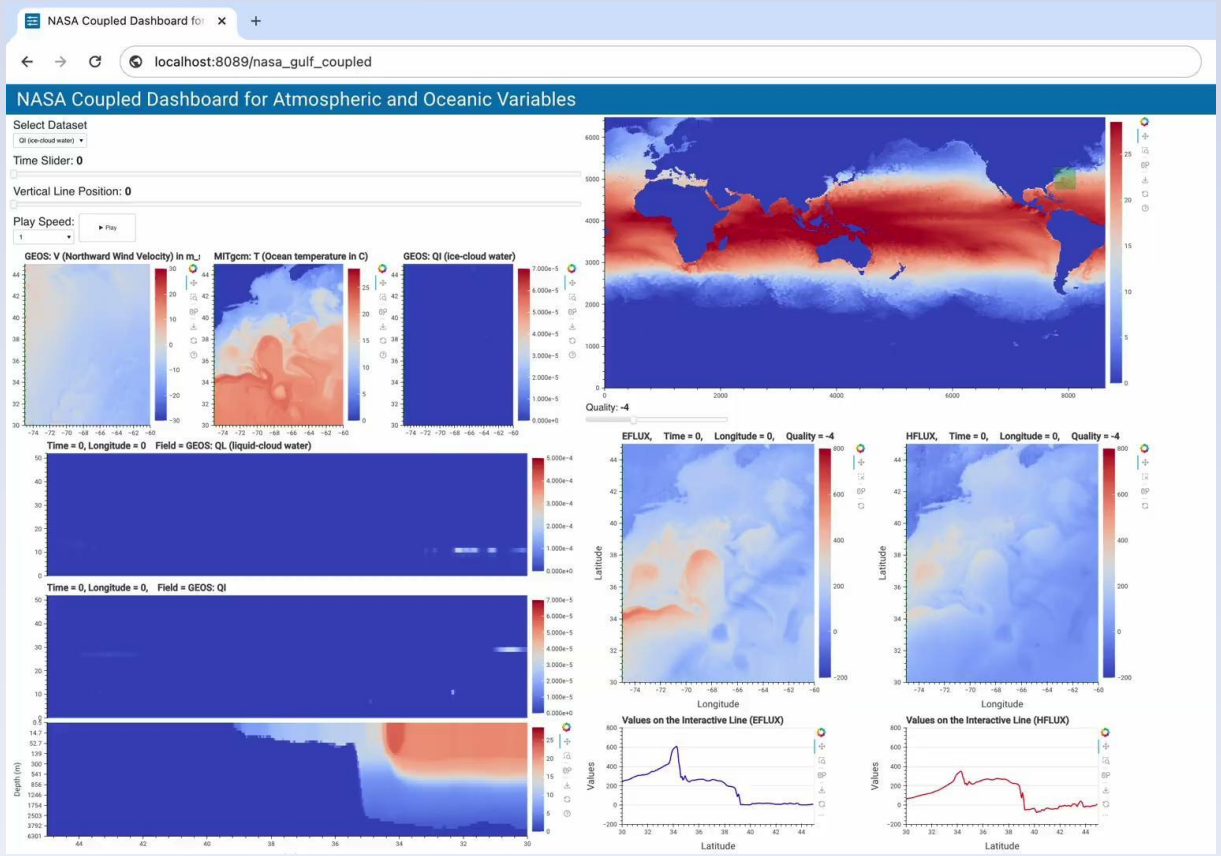
Cloud served optimized data with local caching





Aashish Panta,Xuan Huang, Nina McCurdy, David Ellsworth, Amy A. Gooch, Giorgio Scorzelli, Hector Torres, Patrice Klein, Gustavo A. Ovando-Montejo, Valerio Pascucci.
“Web-based Visualization and Analytics of Petascale data: Equity as a Tide that Lifts All Boats”. LDAV 2024

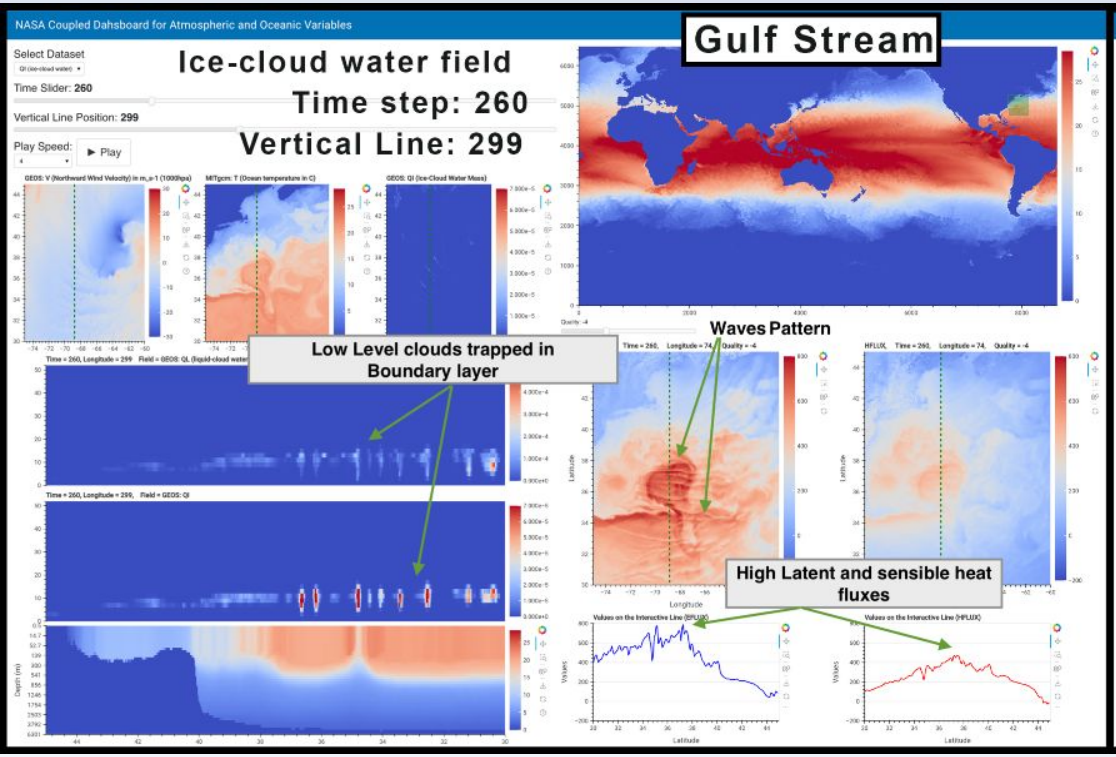
Cloud served optimized data with local caching

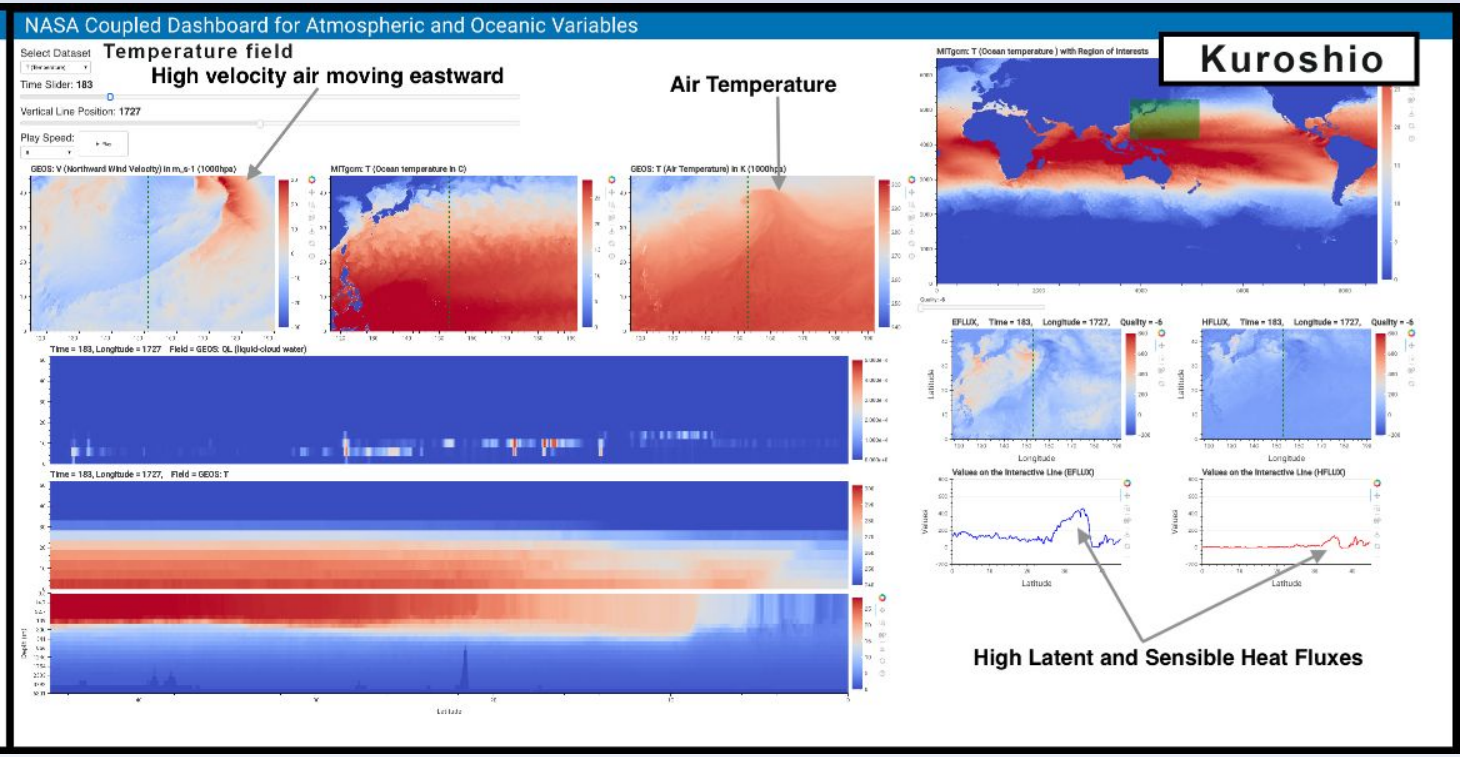


Aashish Panta,Xuan Huang, Nina McCurdy, David Ellsworth, Amy A. Gooch, Giorgio Scorzelli, Hector Torres, Patrice Klein, Gustavo A. Ovando-Montejo, Valerio Pascucci.
“Web-based Visualization and Analytics of Petascale data: Equity as a Tide that Lifts All Boats”. LDAV 2024

Cloud served optimized data with local caching

Github link to access instructions to launch a new dashboard:
https://github.com/sci-visus/Openvisus-NASA-Dashboard





Science of Climate change

The Mediterranean Sea is warming and evaporating more and more with marine heat waves increasing in intensity, duration and frequency

Tracking the changes of the ocean on either side of the Gibraltar Strait is crucial in order to understand the impact of climate change.

Even today, the Mediterranean is considerably saltier than the North Atlantic.

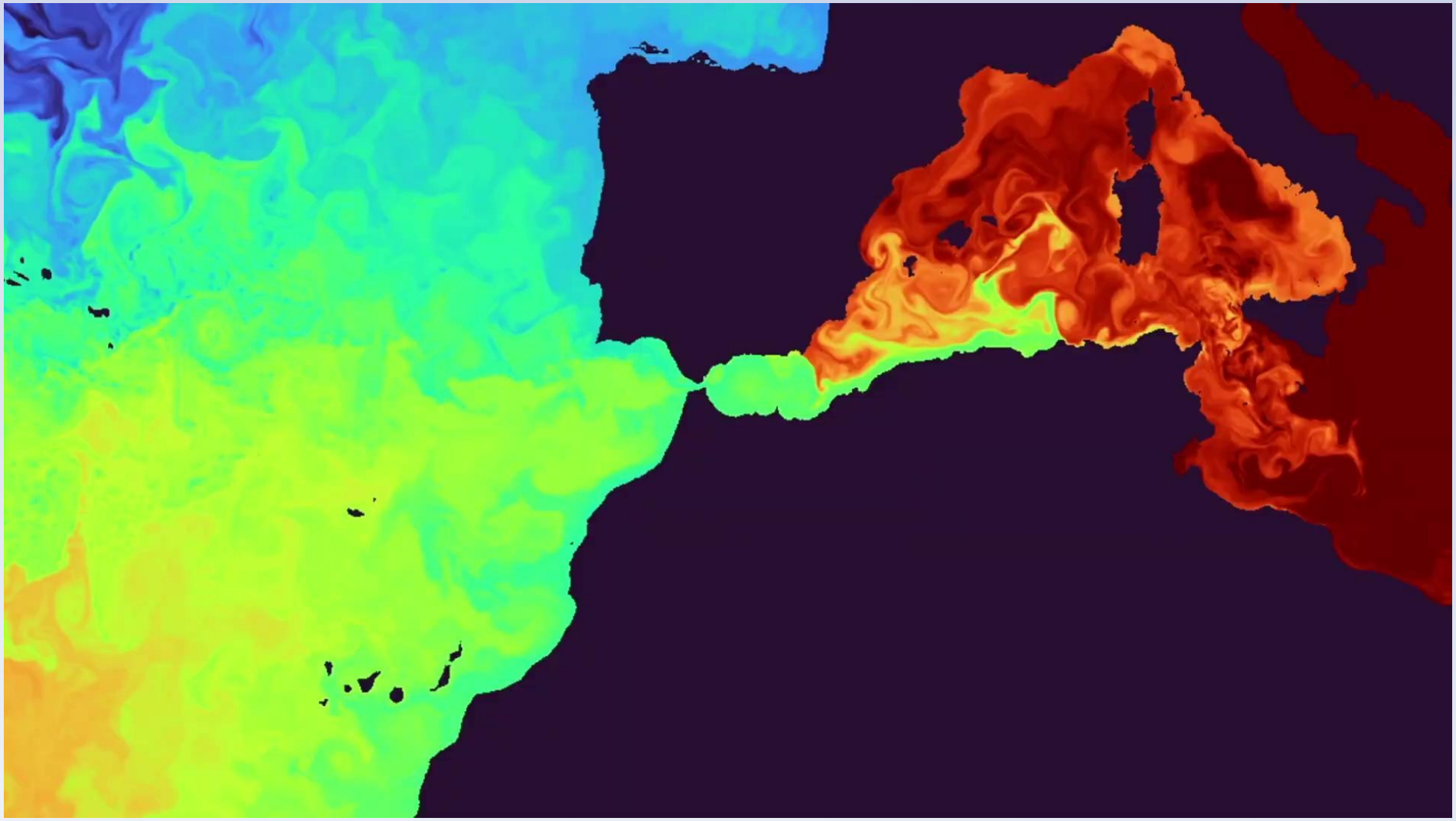
Now you can Visualize the salient changes in the Strait of Gibraltar

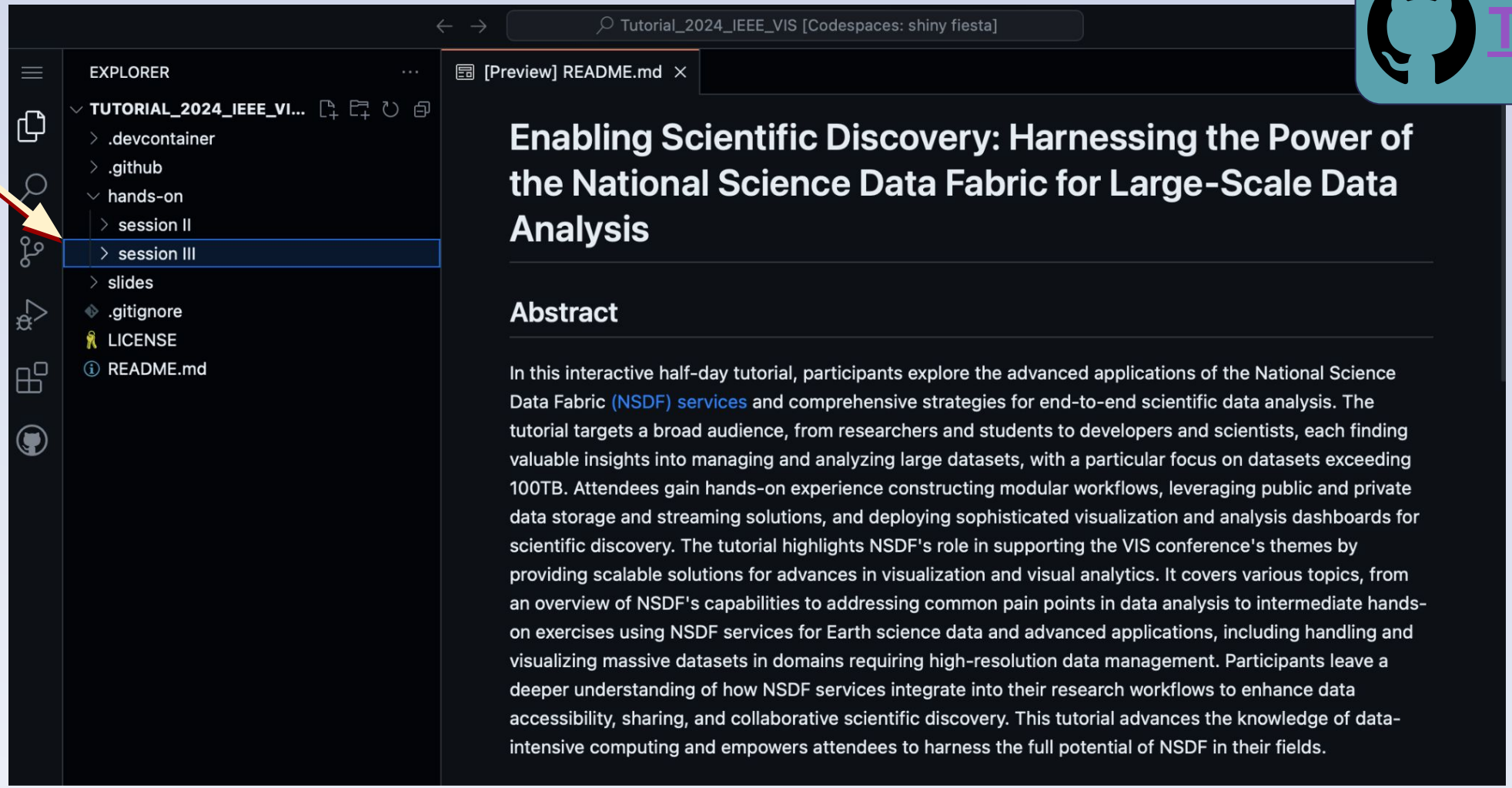
Access petascale NASA data

With OpenViSUS we can treat it as a NumPy array.

Tutorial

Head back to Codespaces





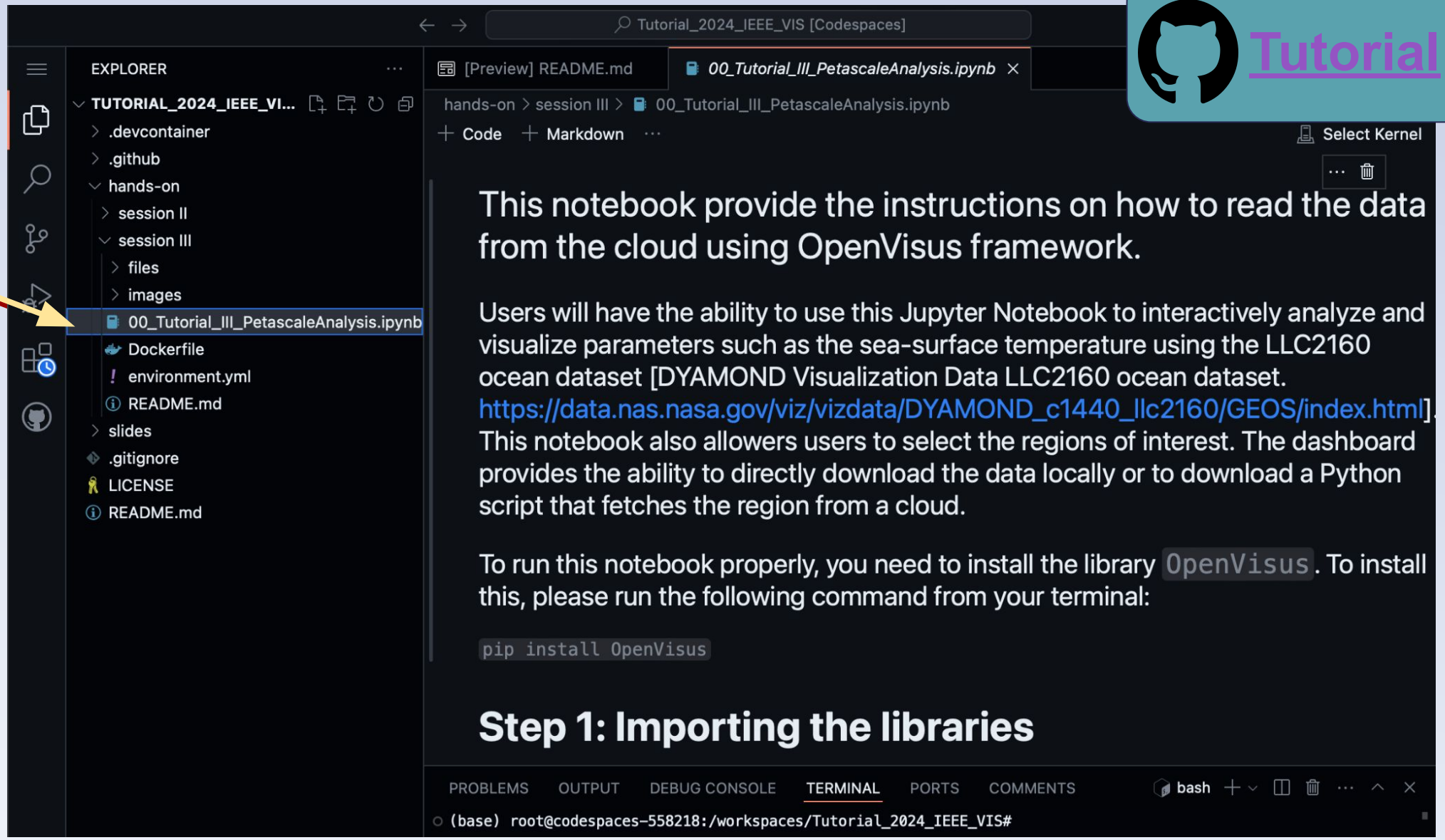
Tutorial

(1) When the codespace is fully loaded, you should see this screen

Select the file

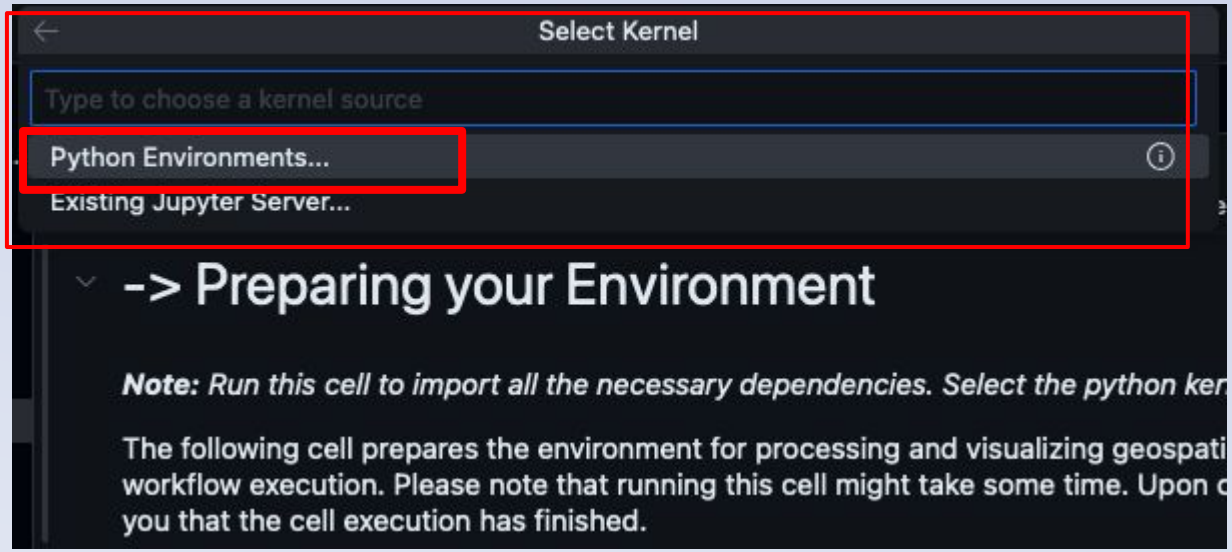
.ipynb

using the side bar

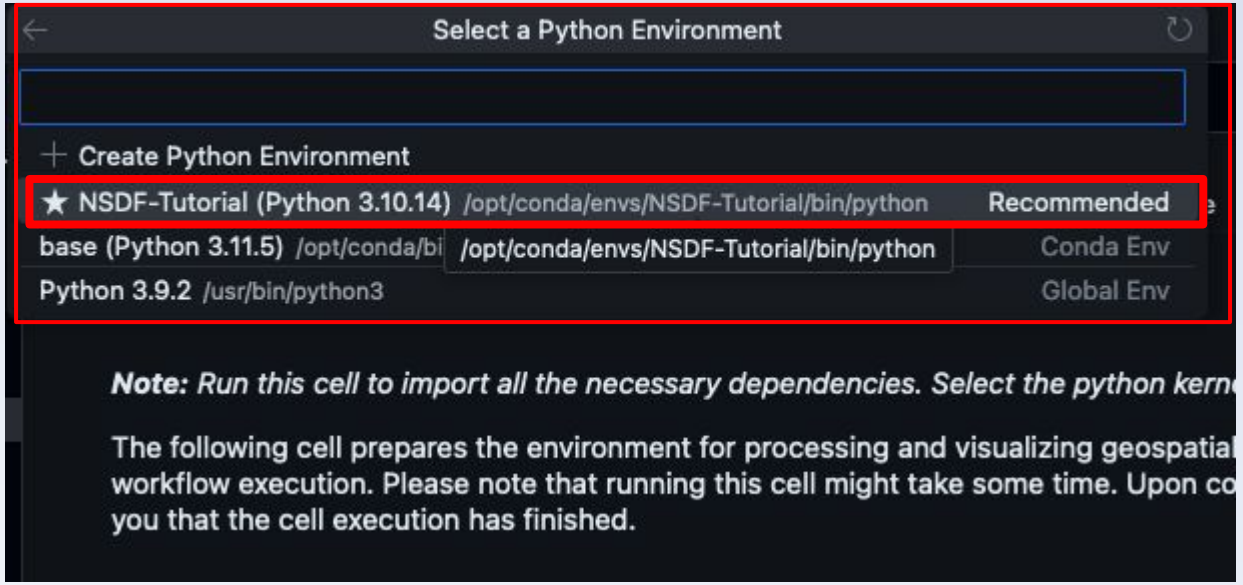


Tutorial

(4) A list to Select Kernel should pop up. Select *Python Environments…*

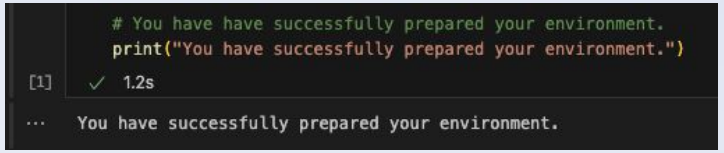


(5) Select the ★*NSDF-Tutorial* option



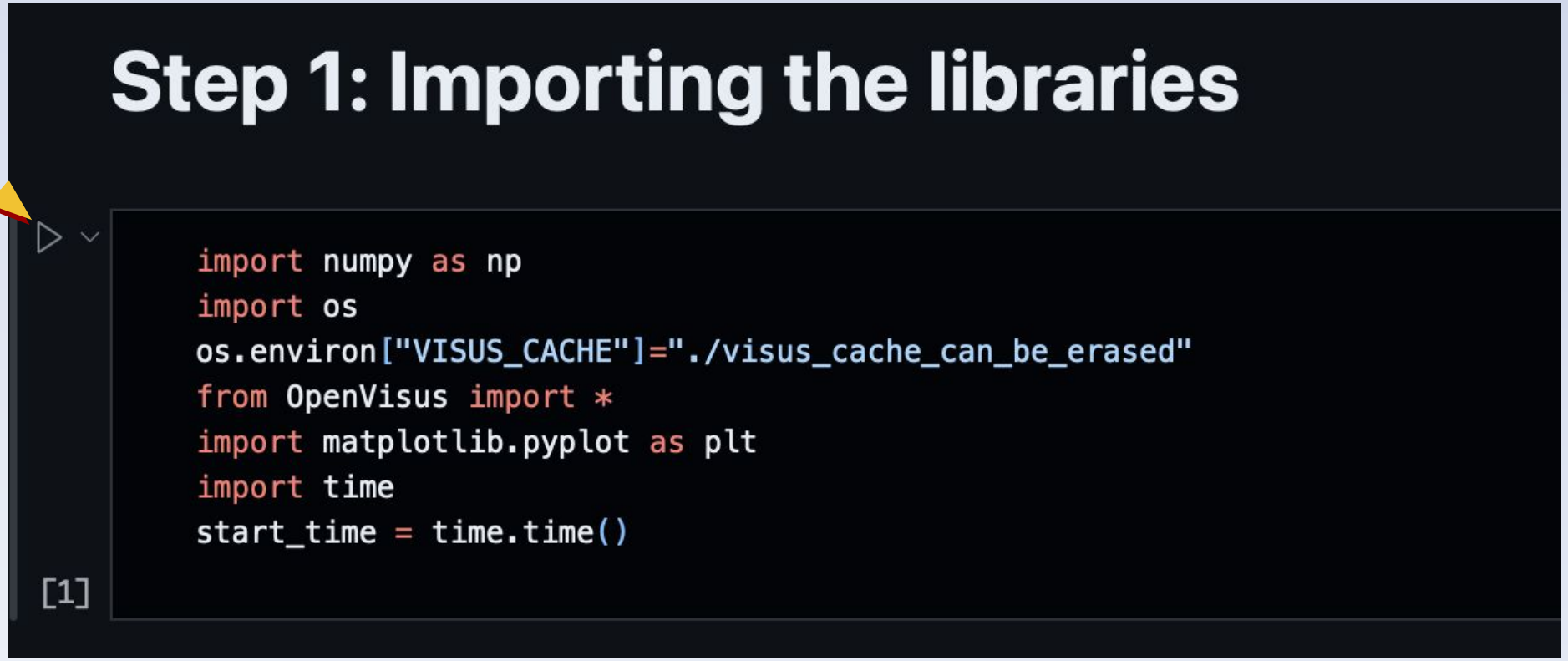
Tutorial

(6) Finally, after running the *Preparing your Environment* cell, you should see a message saying it was successfully prepared



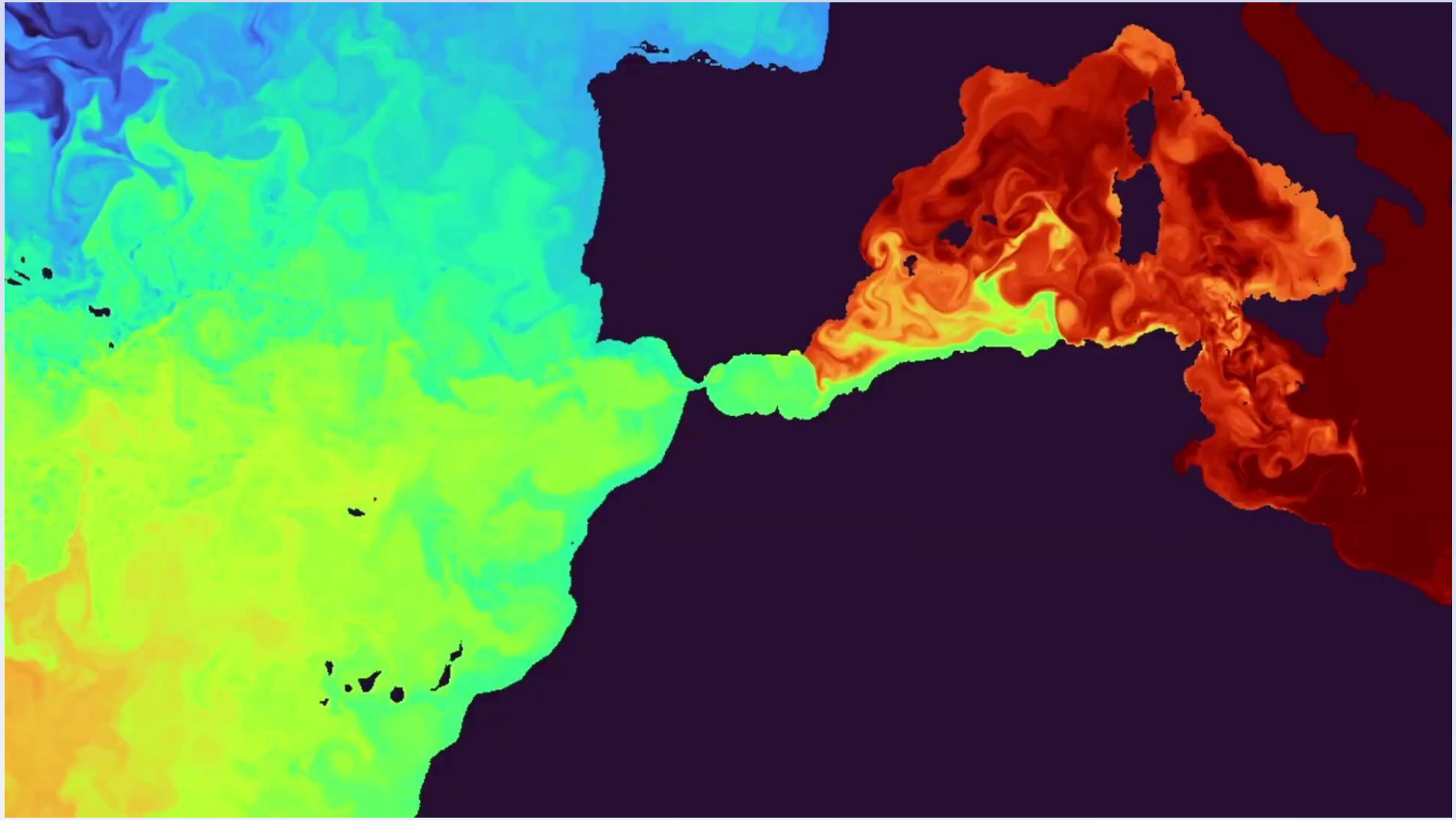
Click on the triangle to play through the section of Python code

Tutorial



[Dashboard 1]

[Run it yourself]



Try it yourself:

Using the NASA LLC2160:

https://github.com/nsdf-fabric/Tutorial_2024_IEEE_VIS/blob/main/notebooks/OpenVisus_NASA_Gustavo2.ipynb

Create a Dashboard:

https://github.com/sci-visus/Openvisus-NASA-Dashboard

Hands on..

When finished with

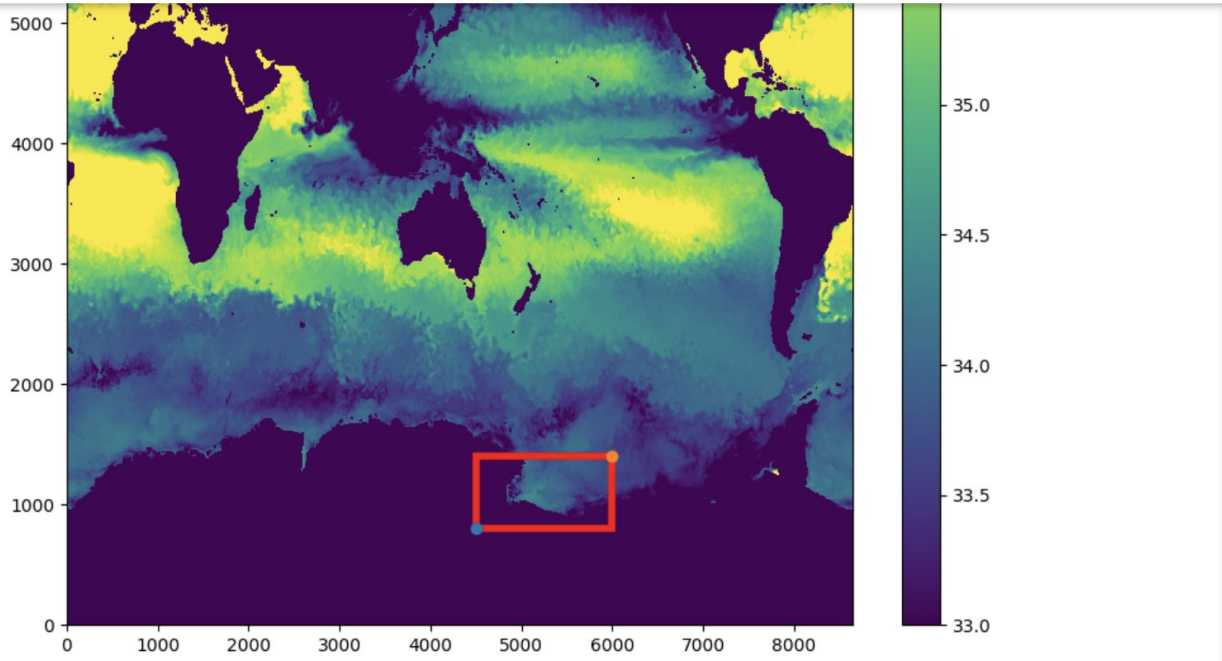
00_Tutorial_III_PetascaleAnalysis.ipynb

You can go to next set of investigations on your own…

(next slide)

Be Curious! Change the Jupyter Notebook:

```
#Step 8a) Test Draw a new region to look at using the unfiltered data:

fig,axes=plt.subplots(1,1,figsize=(10,8))
axes.set_xlim(0,8640)
axes.set_ylim(0,6480)

#Set up a plot of the full data
axp = axes.imshow(data,extent=[0,8640,0,6480], aspect='auto',origin='lower',vmin=33,vmax=36,cmap='viridis')
plt.colorbar(axp,location='right')

#Plot corners of subregion using range of x values (x1,x2) and range of y values (y1,y2)
x1, x2 = [4500,6000]
y1, y2 = [800, 1400]
plt.plot(x1, y1, x2, y2, marker = 'o')

#or you can plot them as a red rectangle
#xy = (top, left)
regionRect =plt.Rectangle(xy=(4500,800), width=1500, height=600, color='r', fill=False, linewidth=4 )
# add the patch to the Axes
axes.add_patch(regionRect)

plt.show()
print('Step 8a done.')
```




```
Step 8a done.
```

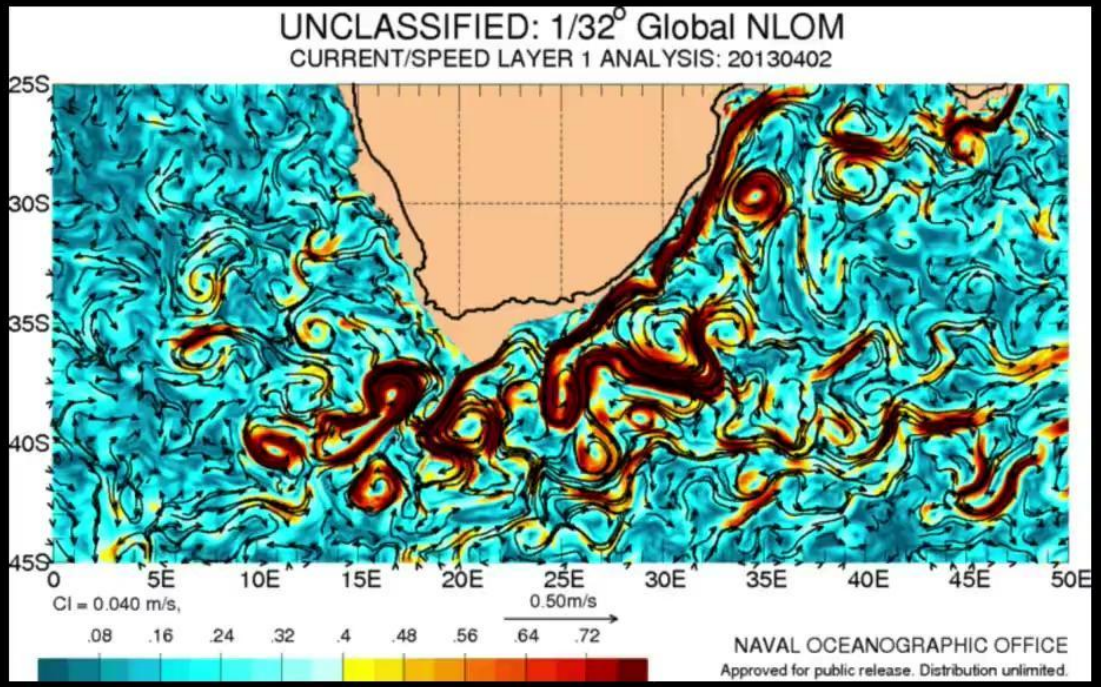
Be Curious! Change the Jupyter Notebook:

Can you visualize the min and max salinity in the Straits of the Gibraltar at a particular time?

HINT: Maybe look at two regions, one on Atlantic side, one in the Mediterranean

What is the min/max over a section of time?

More Domain Science!



Creative Commons Image from http://www7320.nrlssc.navy.mil/global_nlom32/agu.html US Navy NLOM

The Agulhas current is one of the most powerful ocean currents in the world, transporting about 73 billion liters of water per second. A warm water current runs down the east coast of southern Africa, before retroflecting back eastwards into the Indian Ocean.

Dashboard for Figure 7

IEEE Vis Submission: "Web-based Visualization and Analytics of Petascale data: \newline Equity as a Tide to Lift All the Boats"

All Hands Advanced Tutorial Option:

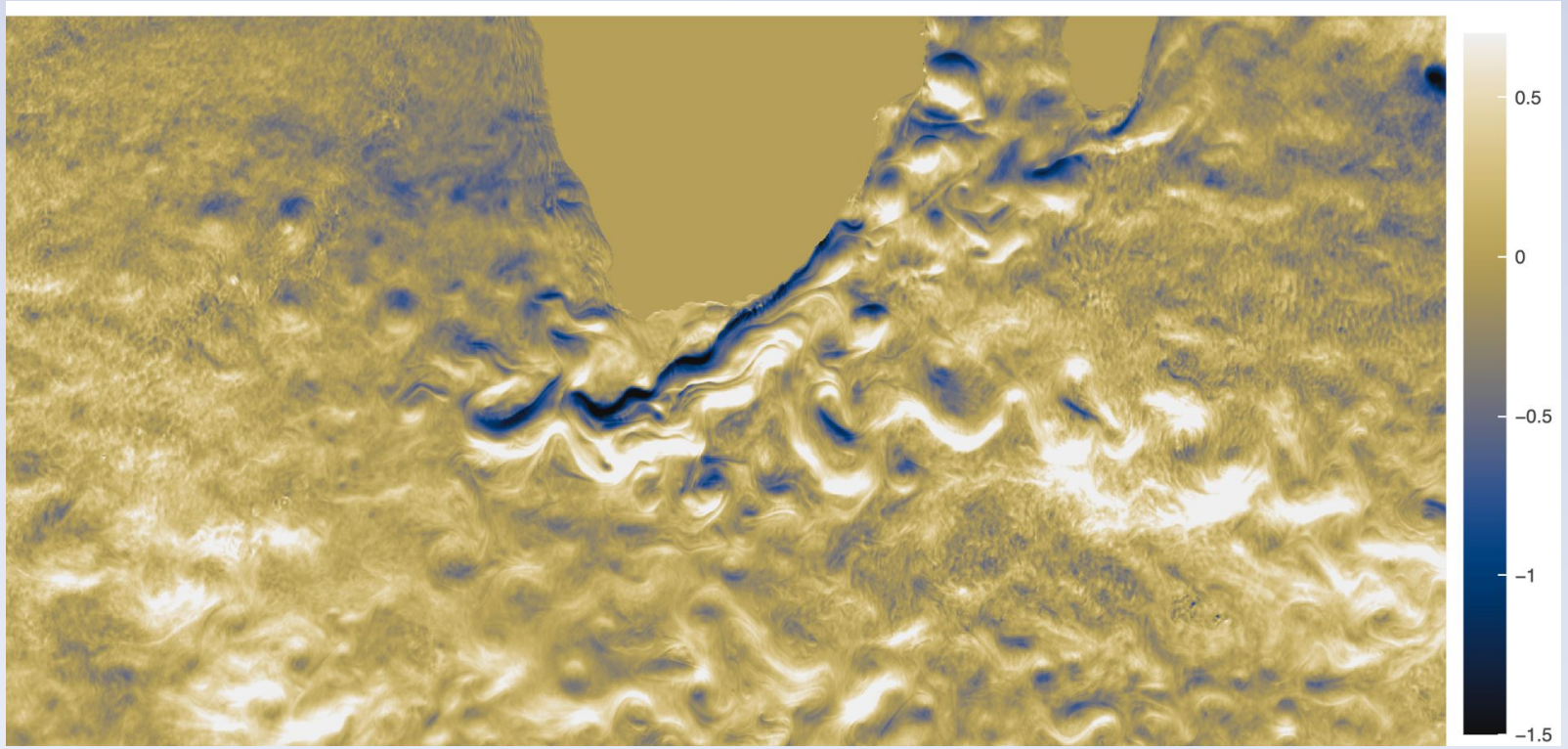
Tutorial

First:

Change the dataset you load in Step 2 to load *eastwest_ocean_velocity_u*

Next:

Change the region of interest in Step 8 to the southern tip of Africa



What is the maximum E-W current of the Aghulas Rings (currents) off the southern tip of South Africa?

Now you have all these datasets

How do you

- organize?
- share?
- view?
- archive?

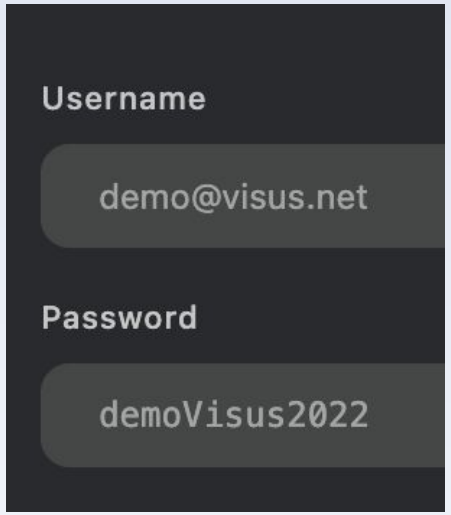
VisSTORE

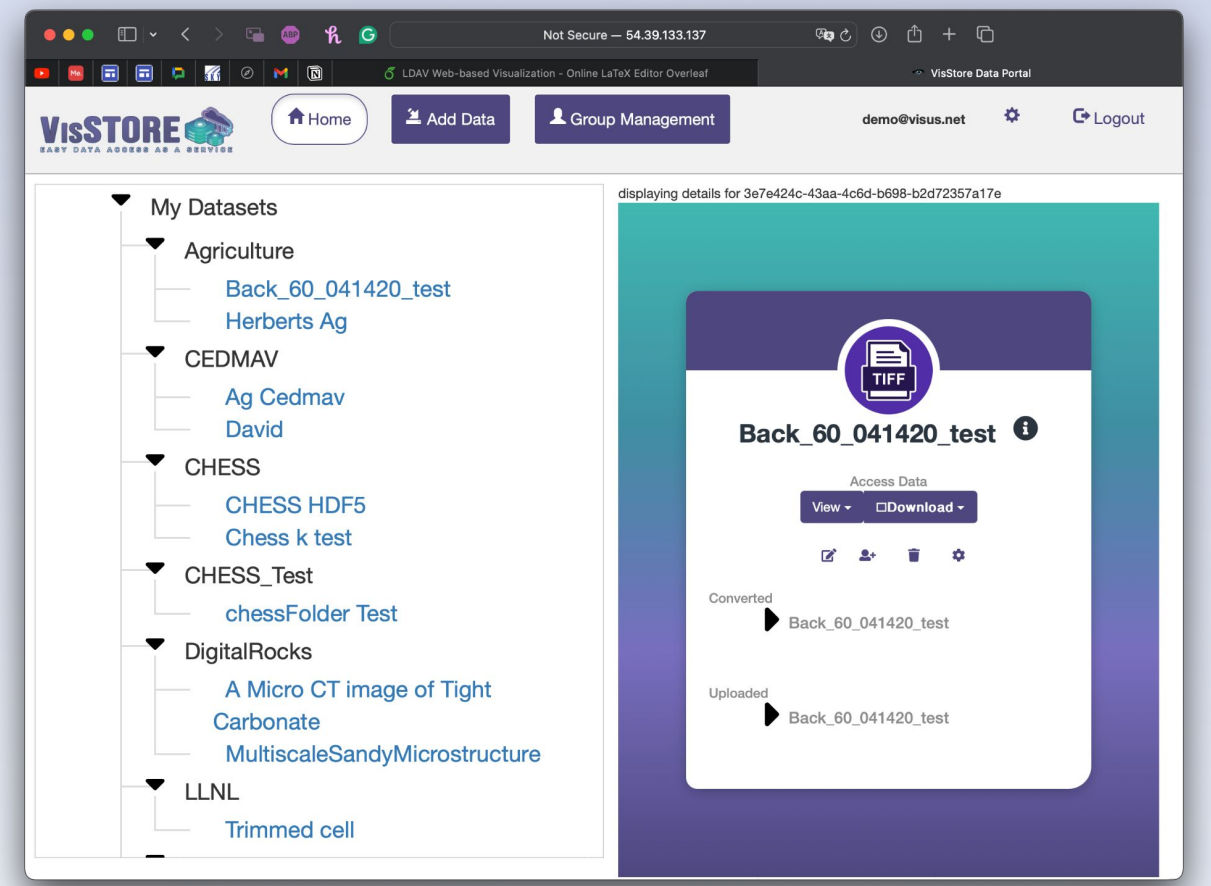
EASY DATA ACCESS AS A SERVICE

Demo: set url in browser to: http://54.39.133.137/

Create your own login to upload your own data

or use:





Session III: Other Datasets

Democratizing Access and Use of Large-scale Data

Check Out Other NSDF-Dashboards

QR1: NASA Ocean Dataset Use Case

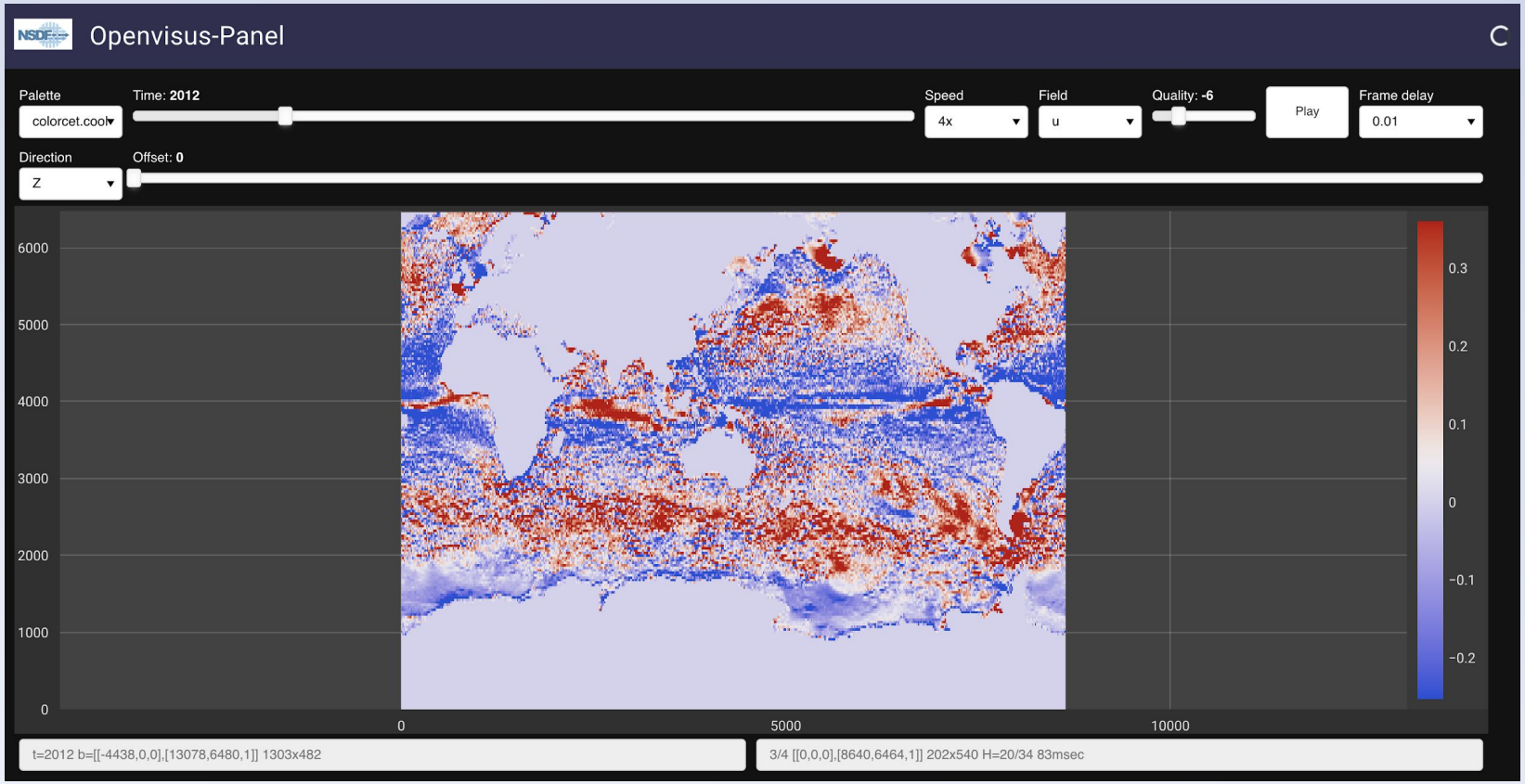
QR2: CHESS Use Case

QR3: Material Science Use Case

QR4: Bellows Use Case

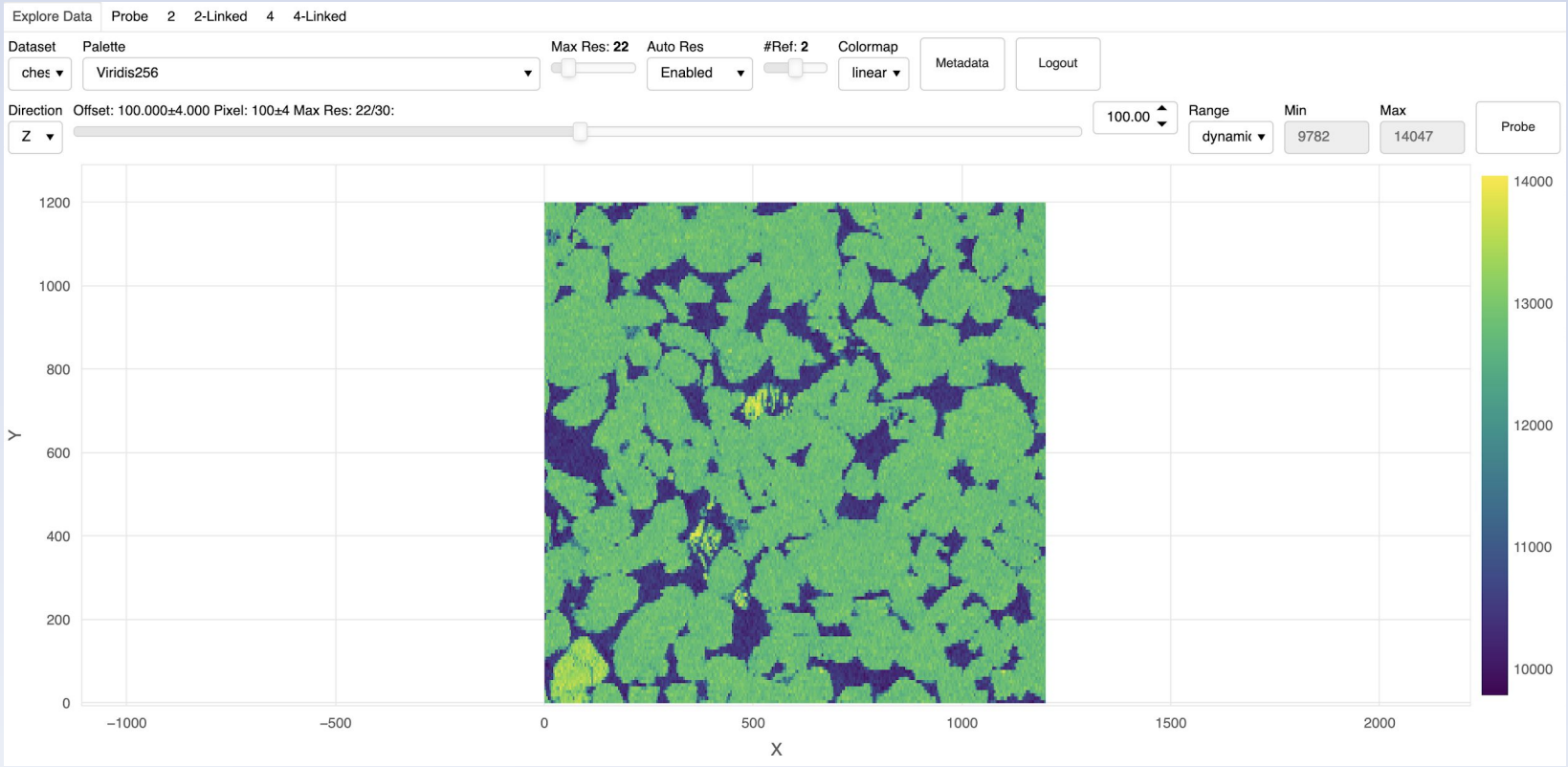
Check Out Other NSDF-Dashboards (I)

[NASA Ocean Dataset Use Case]



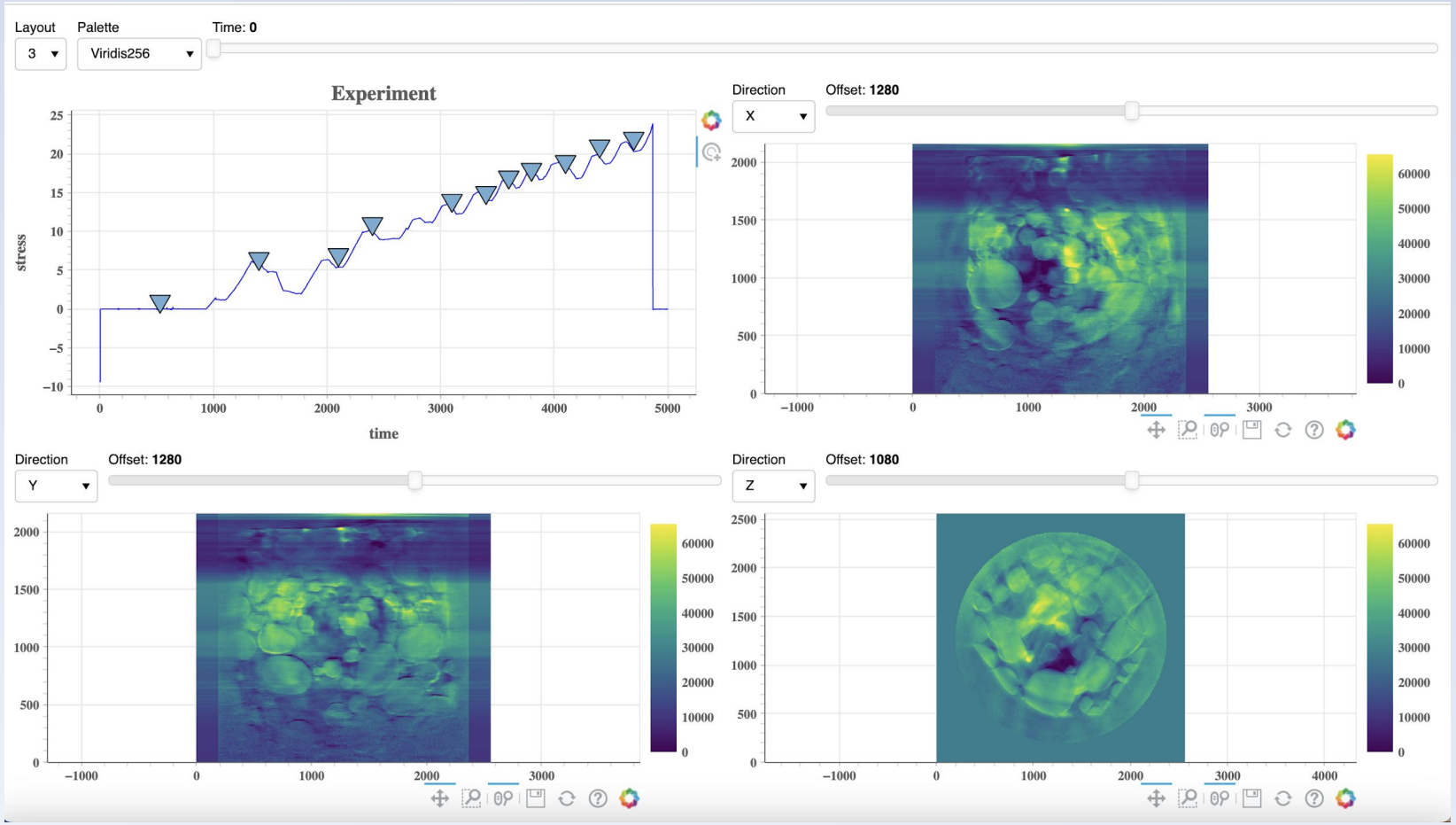
Check Out Other NSDF-Dashboards(II)

CHESS
Use Case



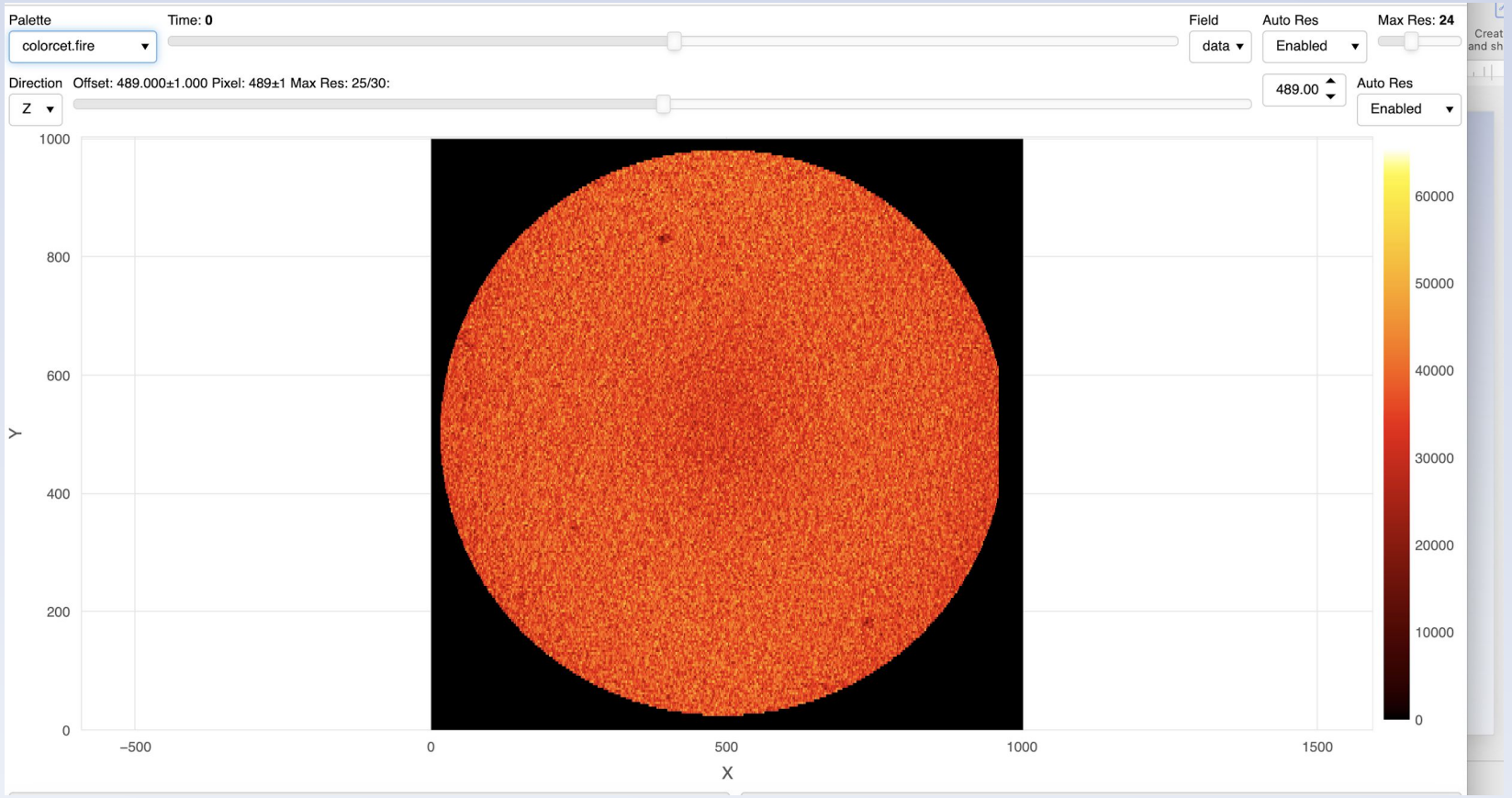
Check Out Other NSDF-Dashboards (III)

[Materials Science Use Case]



Check Out Other NSDF-Dashboards (I)

[Bellows Dataset Use Case]



Part IV: Discussion with Tutorial Attendees

Democratizing Access and Use of Large-scale Data

From Theory to Practice

Construct a modular workflow that combines your application components with NSDF services

Is your application modular? Can you leverage APIs?
Can your application take advantage of the NSDF services?

Upload, download, and stream data to and from **public and private storage** solutions

How large is your data? How do you access, share, and store your data?
Can your data take advantage of private and public storage?

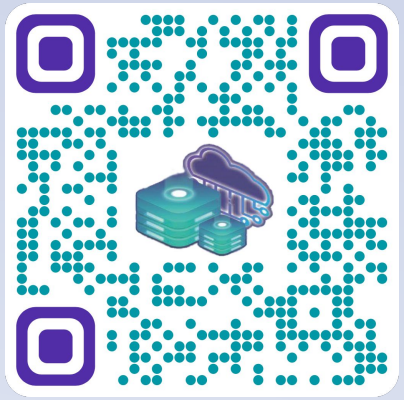
Deploy the NSDF dashboard for large-scale **data access, visualization,** and **analysis**

What type of analysis do you perform on your data?
Can your research take advantage of an interactive dashboard?

Tutorial Links

NSDF-Tutorial

GEOtiled



VisStore

SOMOSPIE

OpenVisus

ViSOAR

Survey

Share your thoughts with us! (3 mins)

https://shorturl.at/jYBxS