

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

Presenters: Valerio Pascucci Amy Gooch, Aashish Panta,
Xuan Huang, Alper Sahistan, Giorgio Scorzelli ¹

Other contributors: Michela Taufer, Jack Marquez, Heberth Martinez ,
Paula Olaya, Gabriel Laboy, Jay Ashworth ²

¹ University of Utah, ² University of Tennessee Knoxville

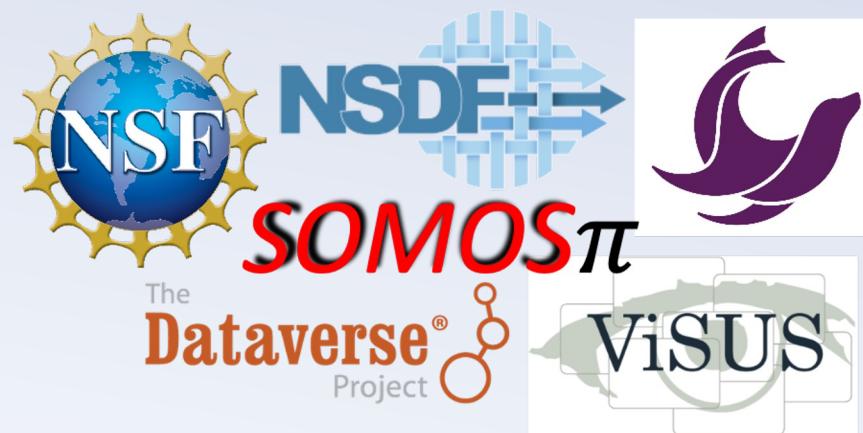


Acknowledgments

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

- NSF through awards 2138811, 2103845, 2334945, 2138296, and 2331152
- [ViSOAR](#)
- [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”

[Create codespace](#)

✓ Image found.
⠦ Building container...



National Science Data Fabric



www.sci.utah.edu



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



Powered by ViSUS



3

Session III



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



Create a new codespace

Repository
To be cloned into your codespace
nsdf-fabric/Tutori... ▾

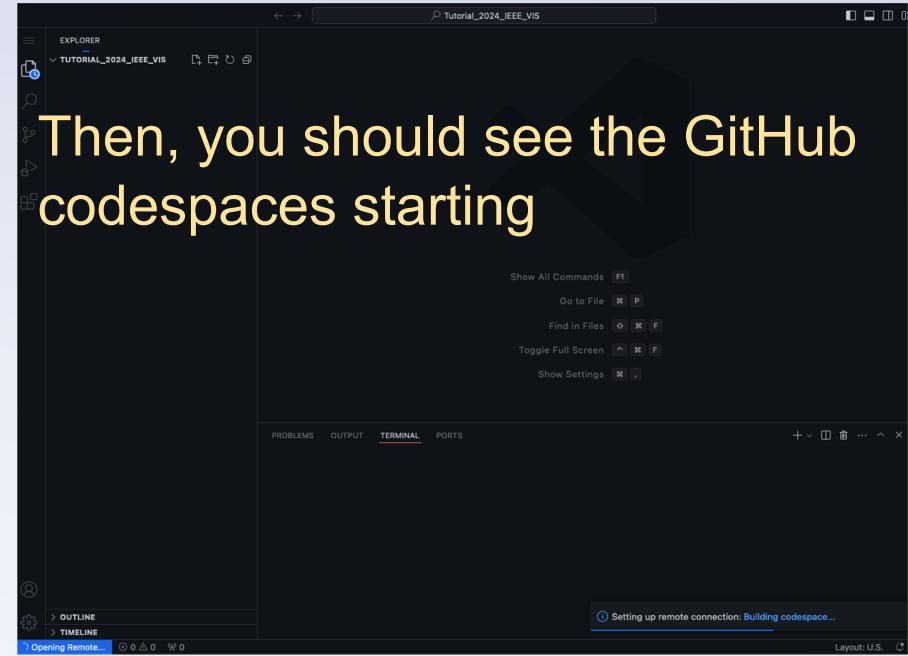
Branch
This branch will be checked out on creation
main ▾

Dev container configuration
Your codespace will use this configuration
NSDF Tutorial - Session III ▾

Region
Your codespace will run in the selected region
US West ▾

Machine type
Resources for your codespace
2-core ▾

Click this button -> Create codespace



→ Let's start:

[NSDF Tutorial 2024](#)



Loading GitHub Codespace can take from 1 to 5 minutes



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



Tutorial Session III Goals



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**



National Science Data Fabric





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?

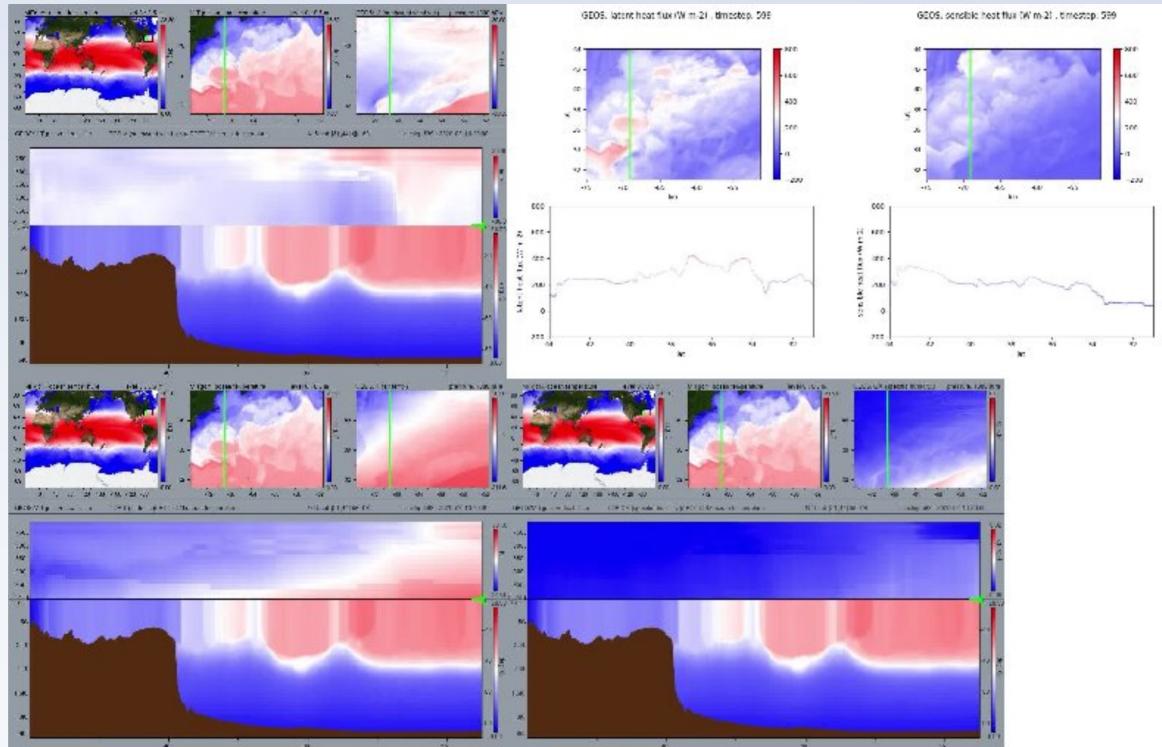
As a student?



2026 IEEE
SciVis
CONTEST



NASA Scientists and Experts in Vis and Climate



Images/Video Copyright NASA and Nina McCurdy, used with permission.



National Science Data Fabric



www.sci.utah.edu



T
THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



JOHNS HOPKINS
UNIVERSITY



8

Difficulty accessing
the full data

Limitations in
computational
power

Need real-time
processing
capabilities



Publicly available NASA Datasets

Datasets :

NASA 1.8 PB DYAMOND dataset¹:
a global atmospheric model and a global ocean model

LLC4320 2.0 PB Ocean dataset²

1) https://gmao.gsfc.nasa.gov/global_mesoscale/dyamond_phasesII/data_access/

2) <https://www.ecco-group.org/data.html>

How big is a petabyte?



$1,000^5 = 1,000 \times 1,000 \times 1,000 \times 1,000 = 1,000,000,000,000,000$ 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.



National Science Data Fabric



www.sci.utah.edu



Powered by ViSUS



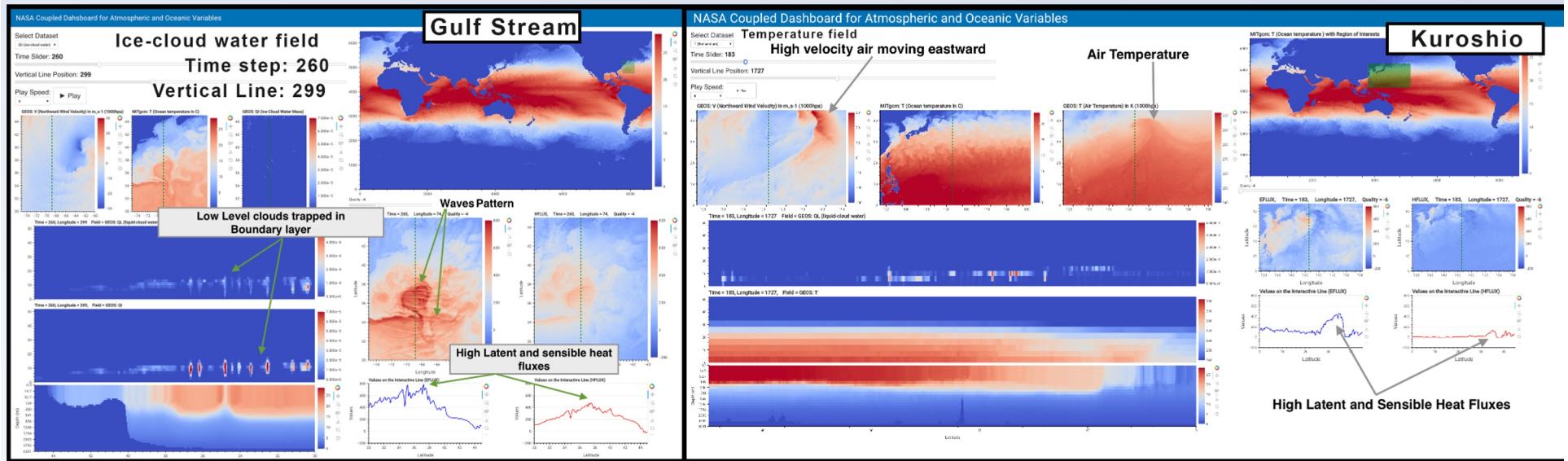
JOHNS HOPKINS
UNIVERSITY

SDSC



10

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". Best Paper, LDAV 2024

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

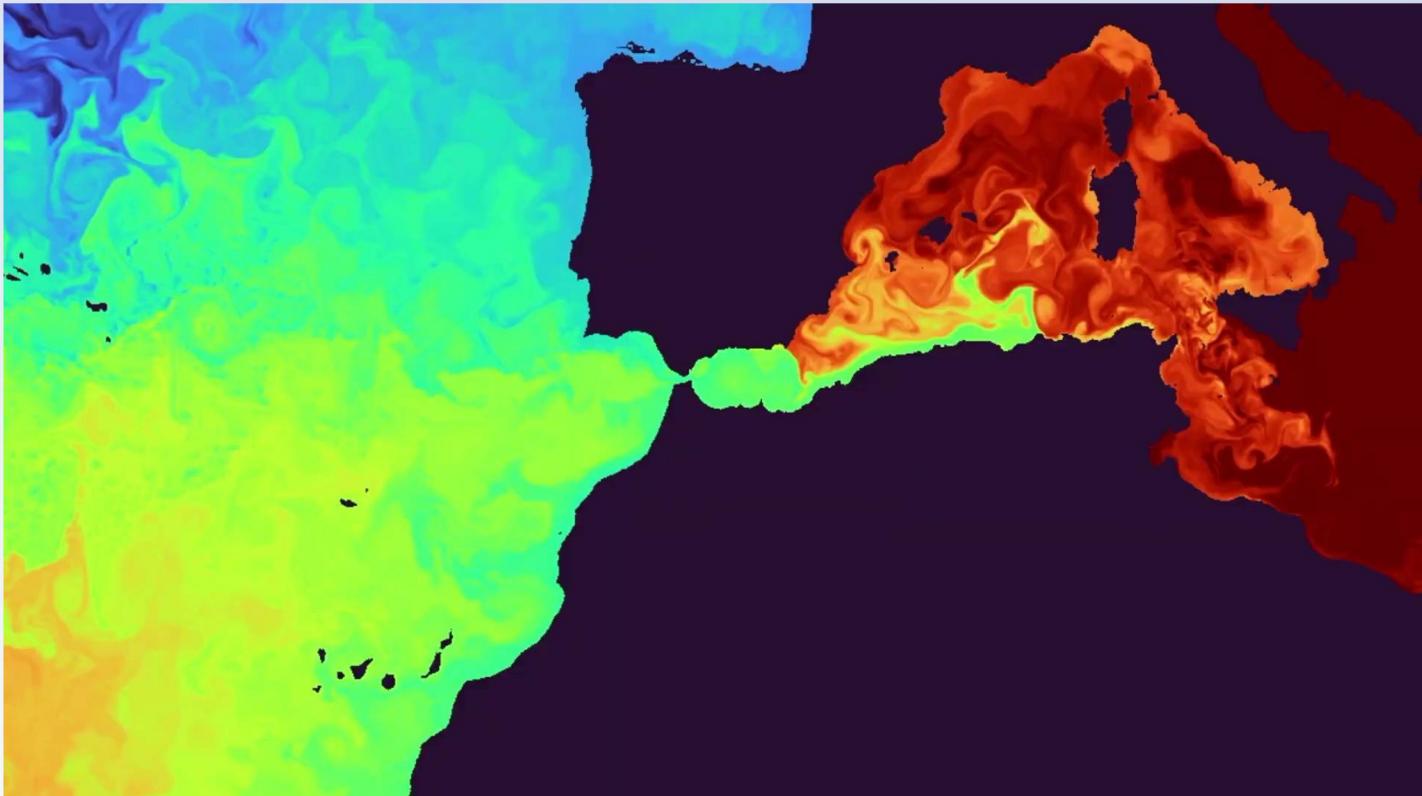


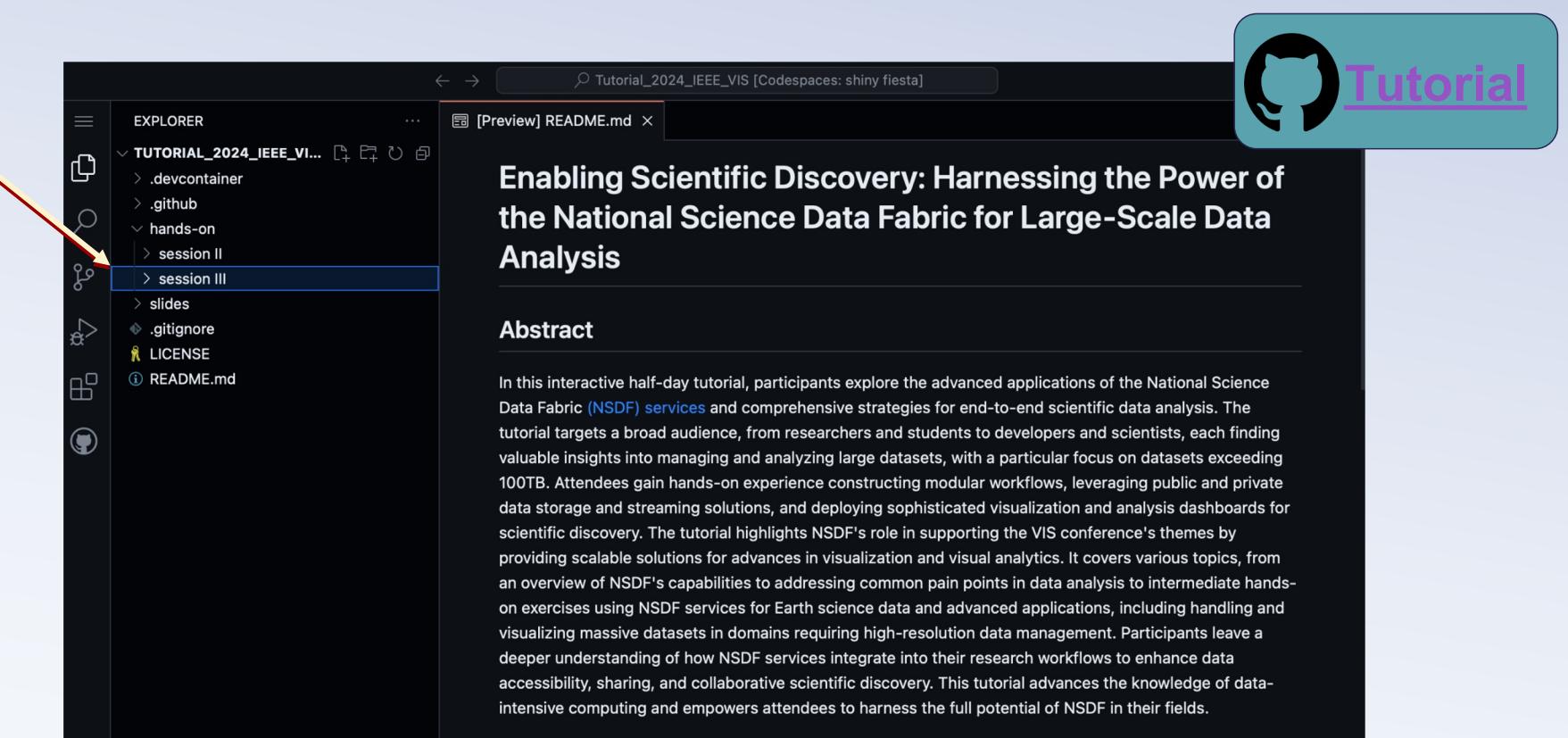
Access
petascale
NASA
data

With OpenViSUS
we can treat it as
a NumPy array.



[Head back to
Codespaces](#)



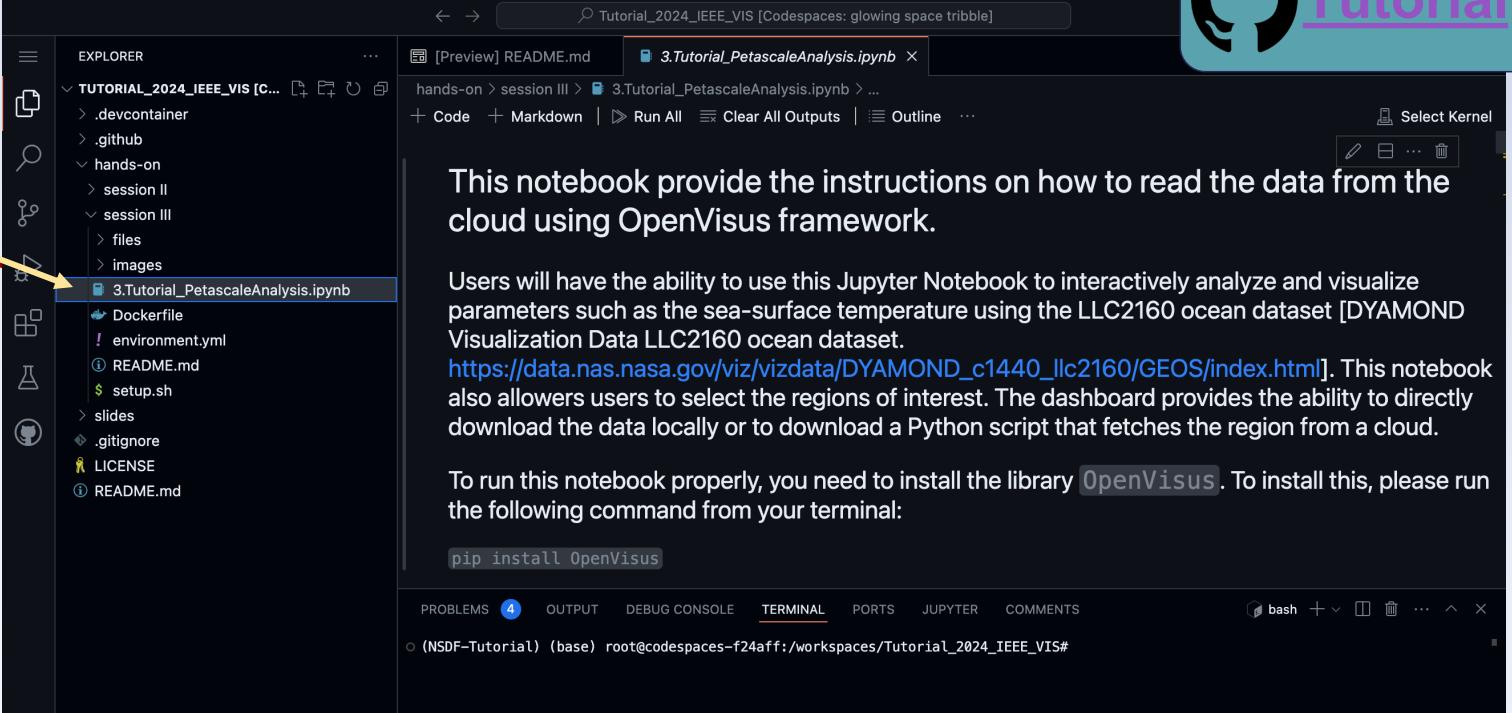


(1) When the codespace is fully loaded, you should see this screen

Select the
file

.ipynb

using the
side bar

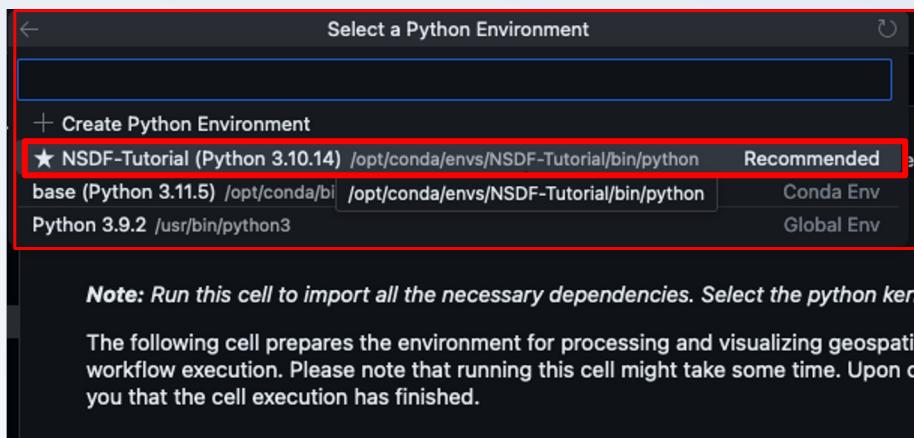
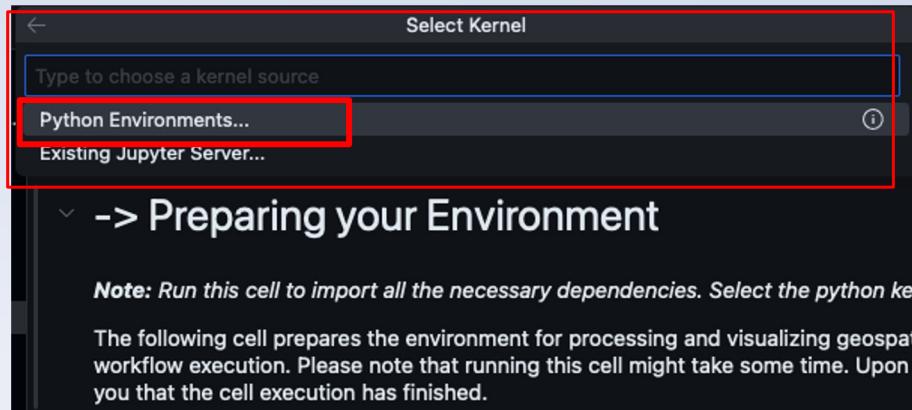


The screenshot shows a Jupyter Notebook interface within a GitHub Codespace. The title bar reads "Tutorial_2024_IEEE_VIS [Codespaces: glowing space tribble]". The left sidebar, titled "EXPLORER", lists files and directories under "TUTORIAL_2024_IEEE_VIS [C...]" including ".devcontainer", ".github", "hands-on", "session II", "session III", "files", "images", "3.Tutorial_PetascaleAnalysis.ipynb" (which is selected), "Dockerfile", "environment.yml", "README.md", "setup.sh", "slides", ".gitignore", "LICENSE", and "README.md". The main notebook area displays the content of "3.Tutorial_PetascaleAnalysis.ipynb". The text in the notebook states: "This notebook provide the instructions on how to read the data from the cloud using OpenVisus framework. Users will have the ability to use this Jupyter Notebook to interactively analyze and visualize parameters such as the sea-surface temperature using the LLC2160 ocean dataset [DYAMOND Visualization Data LLC2160 ocean dataset]. https://data.nas.nasa.gov/viz/vizdata/DYAMOND_c1440_llc2160/GEOS/index.html. This notebook also allows users to select the regions of interest. The dashboard provides the ability to directly download the data locally or to download a Python script that fetches the region from a cloud. To run this notebook properly, you need to install the library OpenVisus. To install this, please run the following command from your terminal: pip install OpenVisus". The bottom of the screen shows a terminal window with the command "pip install OpenVisus" entered. The GitHub logo and the word "Tutorial" are visible in the top right corner.

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

...

(5) Select the
★NSDF-Tutorial
option →



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

```
# You have successfully prepared your environment.
print("You have successfully prepared your environment.")
[1]: ✓ 1.2s
... You have successfully prepared your environment.
```

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



Step 1: Importing the libraries

```
import numpy as np
import os
os.environ["VISUS_CACHE"]=".visus_cache_can_be_erased"
from OpenVisus import *
import matplotlib.pyplot as plt
import time
start_time = time.time()
```

[1]



National Science Data Fabric



Sci
www.sci.utah.edu



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



SDSC

IBM



18

Hands on..



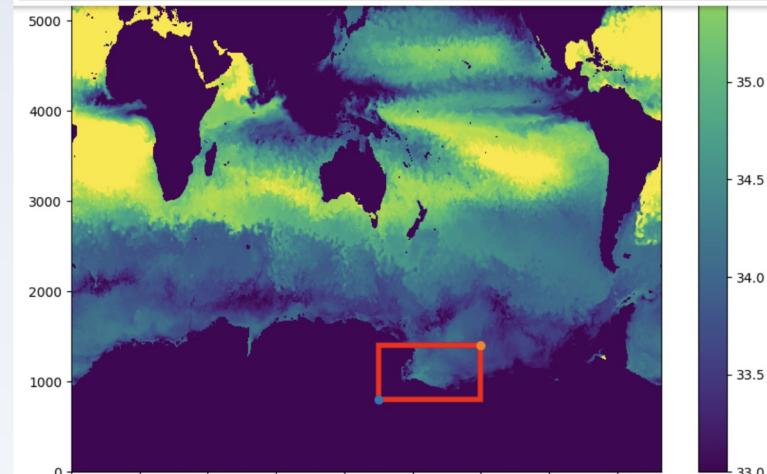
When finished with

3.Tutorial_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?



National Science Data Fabric



www.sci.utah.edu



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



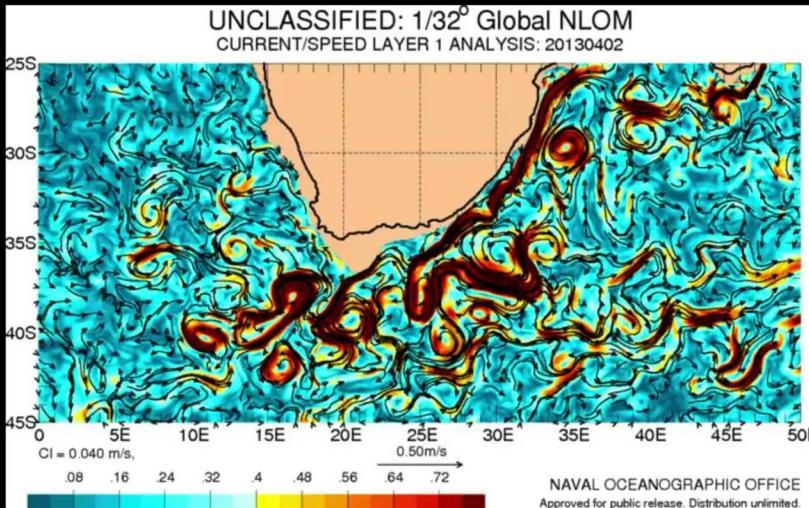
SDSC

IBM



21

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:

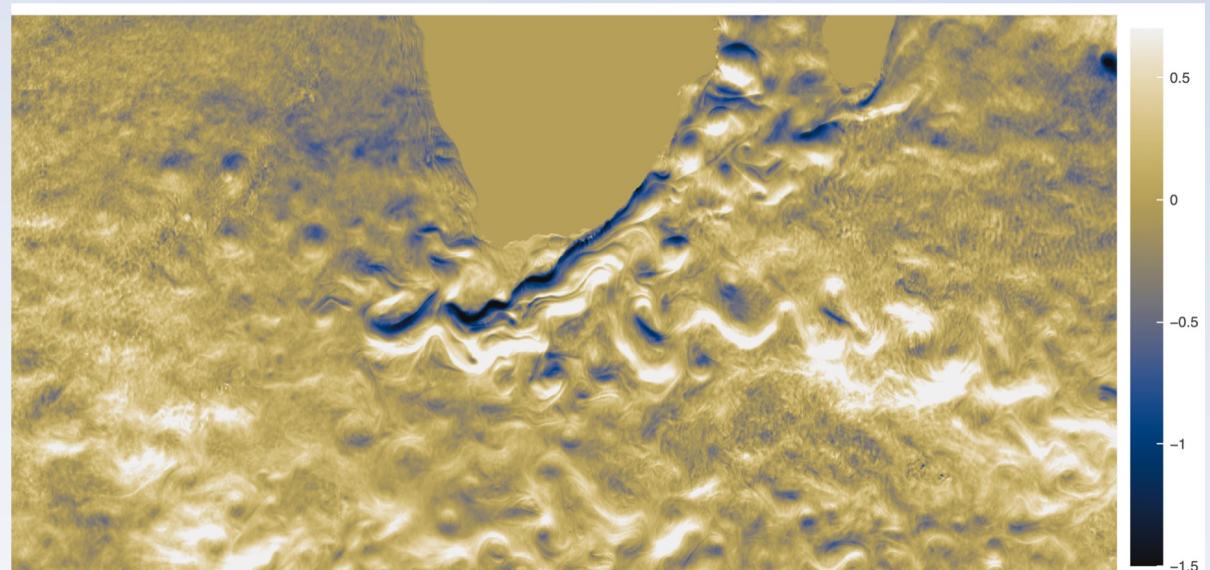


First:

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

Next:

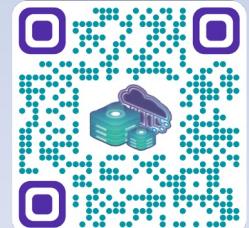
Change the region of interest in Step 8 to the southern tip of Africa



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

Now you have all these datasets

How do you
organize?
share?
view?
archive?



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

Create your own login to upload your own data

or use:

Username
demo@visus.net

Password
demoVisus2022

The screenshot shows the VisSTORE Data Portal interface. On the left, a sidebar lists 'My Datasets' with categories like Agriculture, CEDMAV, CHESS, CHESS_Test, DigitalRocks, and LLNL, each containing various sub-datasets. On the right, a larger panel displays details for the dataset 'Back_60_041420_test'. It includes a preview thumbnail, file type (TIFF), name, and download options. Below this, it shows conversion status ('Converted') and upload history ('Uploaded').



National Science Data Fabric



THE UNIVERSITY OF

TENNESSEE

KNOXVILLE



Powered by VISUS



25

Session IV: Other Datasets

Democratizing Access and Use of Large-scale Data



www.sci.utah.edu



26

Check Out Other NSDF-Dashboards



[QR1: NASA Ocean
Dataset Use Case](#)



[QR3: Material
Science Use Case](#)



[QR2: CHESS Use Case](#)



[QR4: Bellows Use Case](#)



National Science Data Fabric



www.sci.utah.edu



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



Powered by ViSUS



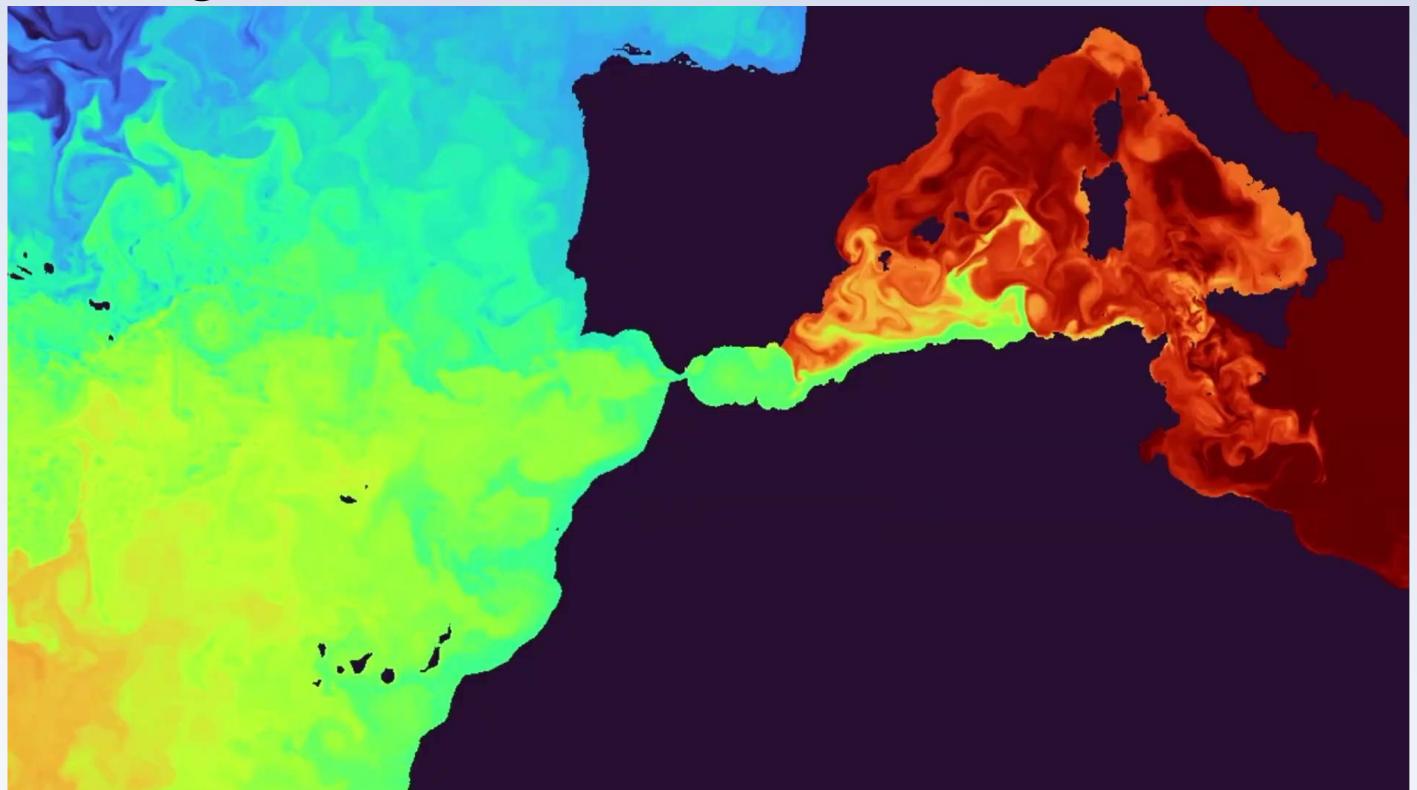
27

Power of Visualization: Water Circulation in the Mediterranean Region

[Dashboard 1](#)



[Run it yourself](#)



National Science Data Fabric

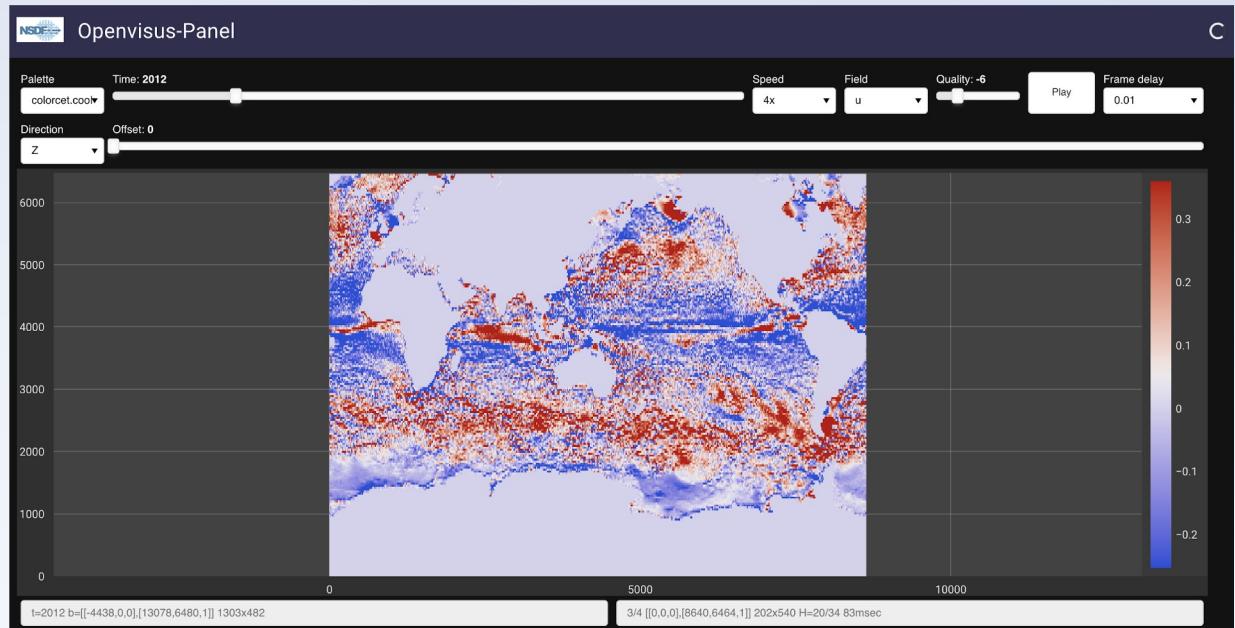


28

Check Out Other NSDF-Dashboards (I)



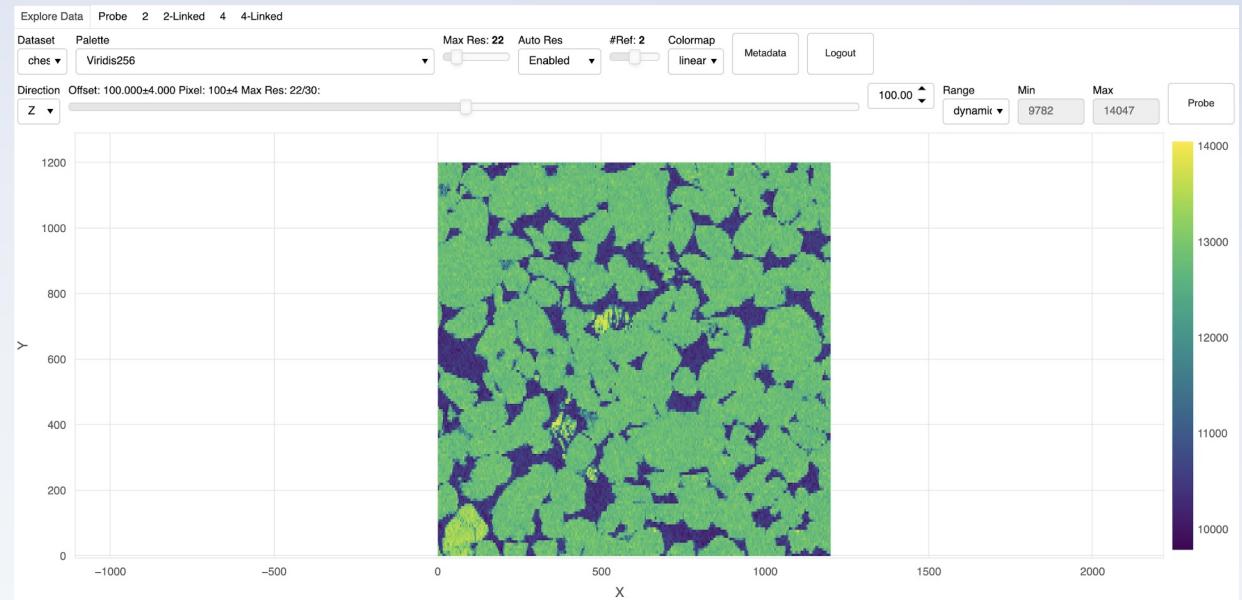
NASA Ocean
Dataset Use Case



Check Out Other NSDF-Dashboards(II)



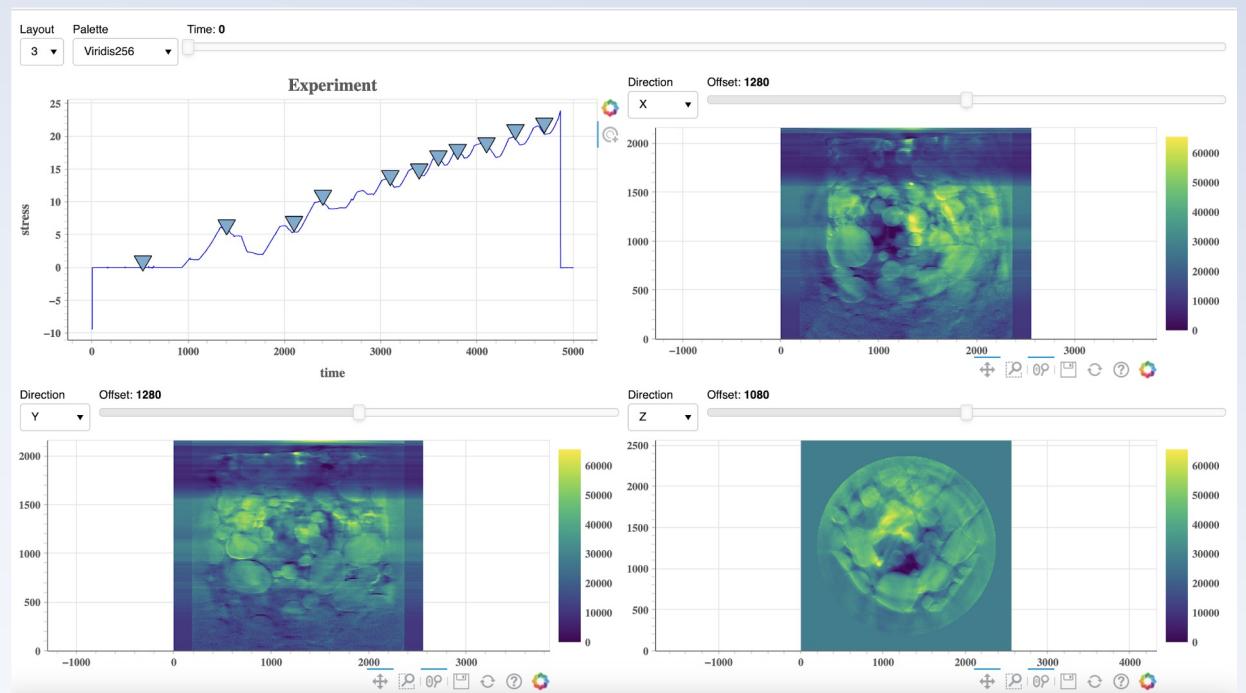
CHESS
Use Case



Check Out Other NSDF-Dashboards (III)



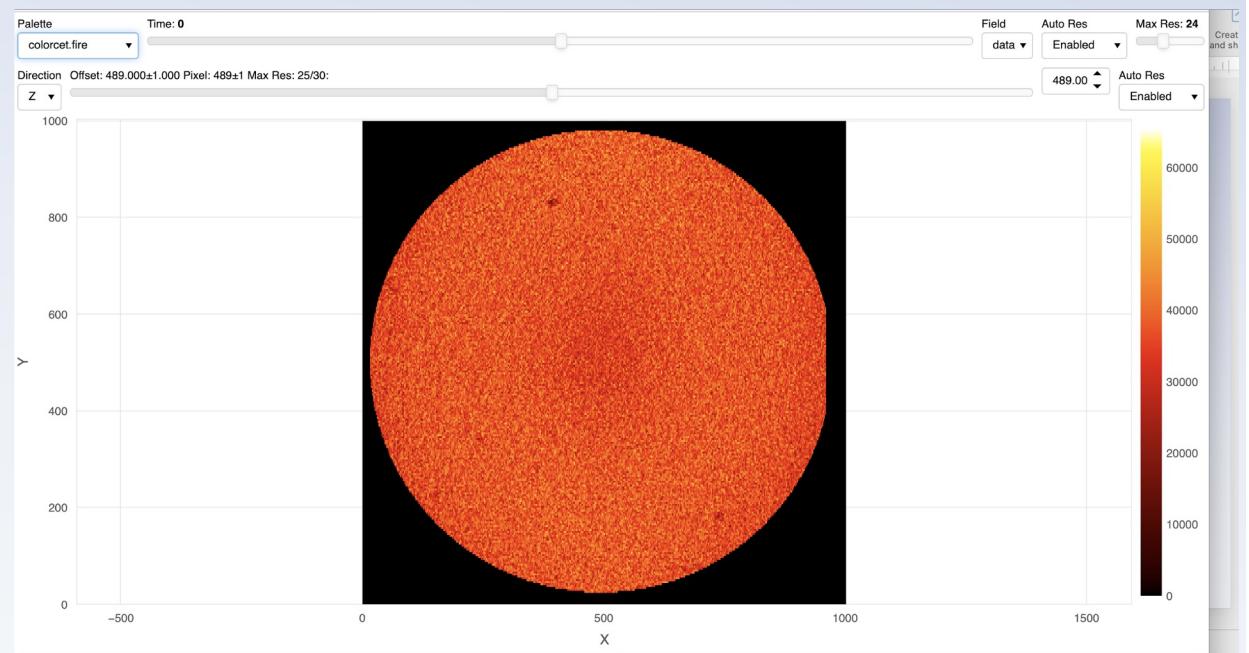
Materials Science
Use Case



Check Out Other NSDF-Dashboards (I)



Bellows
Dataset Use Case



Part IV: Discussion with Tutorial Attendees

Democratizing Access and Use of Large-scale Data

Tutorial Links



[NSDF-Tutorial](#)



[GEOtiled](#)



[VisStore](#)



[SOMOSPIE](#)



[OpenVisus](#)



[ViSOAR](#)



National Science Data Fabric



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE



JOHNS HOPKINS
UNIVERSITY



34

Survey

Share your thoughts with us! (3 mins)



<https://shorturl.at/gPNP0>