While you wait

- Connect to TAMU VPN and Login to Terra

```
ssh <username>@terra.tamu.edu
```

- Go to your scratch dir

```
cd $SCRATCH
```

- Clone the notebook repository from github

```
git clone https://github.com/abishekg7/intro_xarray_dask.git
```

(OR) Copy notebooks from Terra scratch

```
cp   -r    /scratch/training/intro_pangeo/notebooks .
```

High Performance
Research Computing
DIVISION OF RESEARCH

TEXAS A&M UNIVERSITY
Oceanography

Introduction to Xarray and Dask

Spring 2021 HPRC Short Course
Apr 23, 2021

Abishek Gopal
Assistant Research Scientist
iHESP, Texas A&M Oceanography
Texas A&M High Performance Research Computing

Helpful HPRC resources

- Terra quick start guide
  - https://hprc.tamu.edu/wiki/Terra:QuickStart
- Introduction to HPRC – Short course
  - https://hprc.tamu.edu/training/intro_hprc.html
- Submit tickets to help@hprc.tamu.edu

Upcoming relevant HPRC short courses

- Apr 23: Introduction to CUDA
  - **Instructor:** Jian Tao
  - **Time:** Friday, April 23, 1:30PM - 4:00PM
- Apr 30: Introduction to Fortran
  - **Instructor:** Jian Tao
  - **Time:** Friday, Apr 30, 10:00AM - 12:30PM
- Apr 30: Introduction to Containers on Terra
  - **Instructor:** Richard Lawrence
  - **Time:** Friday, October 30, 1:30PM - 4:00PM

https://hprc.tamu.edu/training/index.html

Acknowledgements

- Course materials adapted from detailed Xarray and Dask tutorial notebooks
  - https://github.com/xarray-contrib/xarray-tutorial
  - https://github.com/dask/dask-tutorial

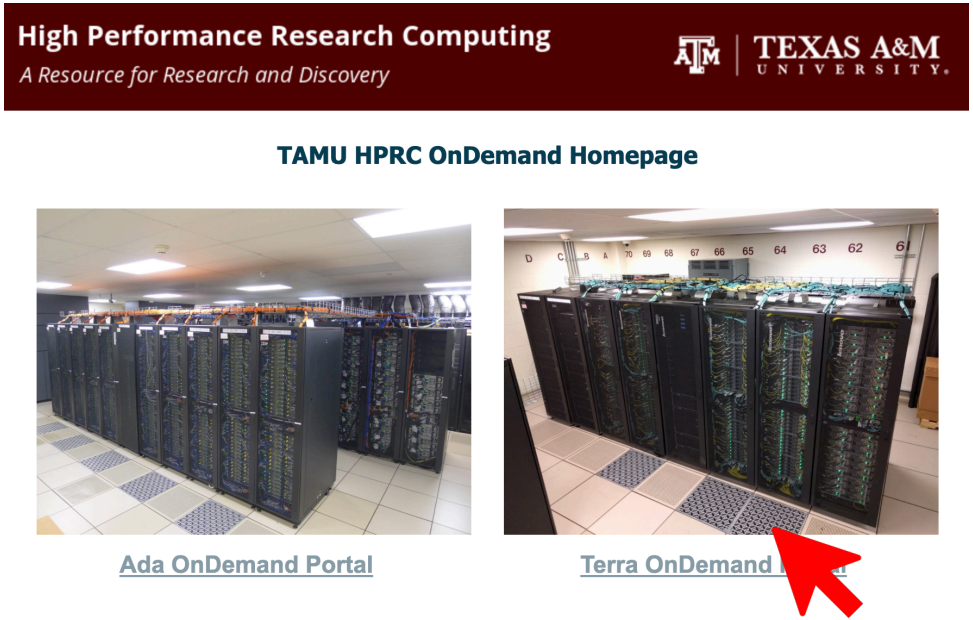
- Lisa Perez and the HPRC team

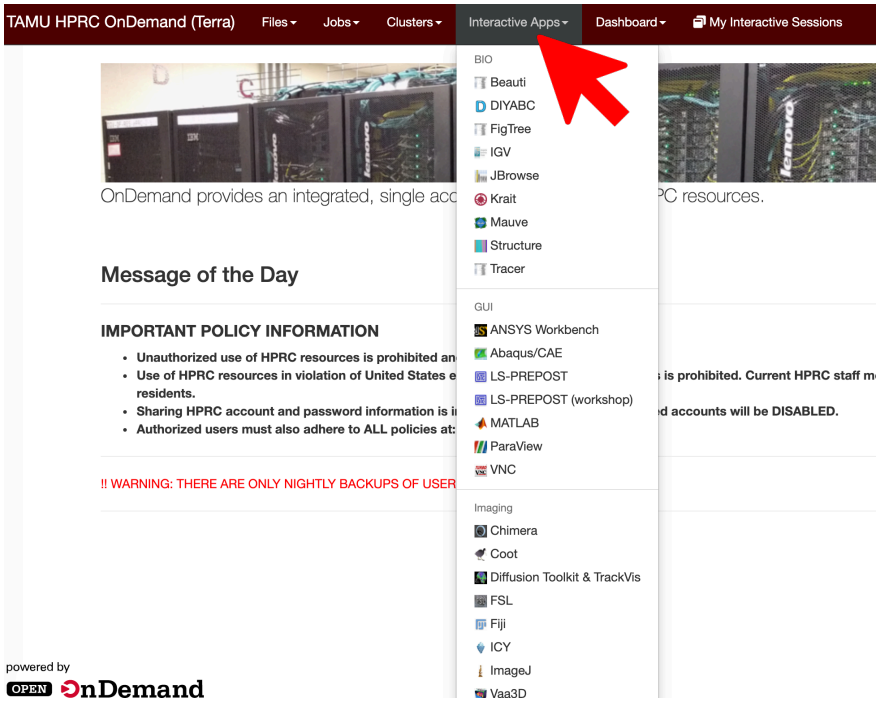
- Dapeng Li and Sanjiv Ramachandran, iHESP

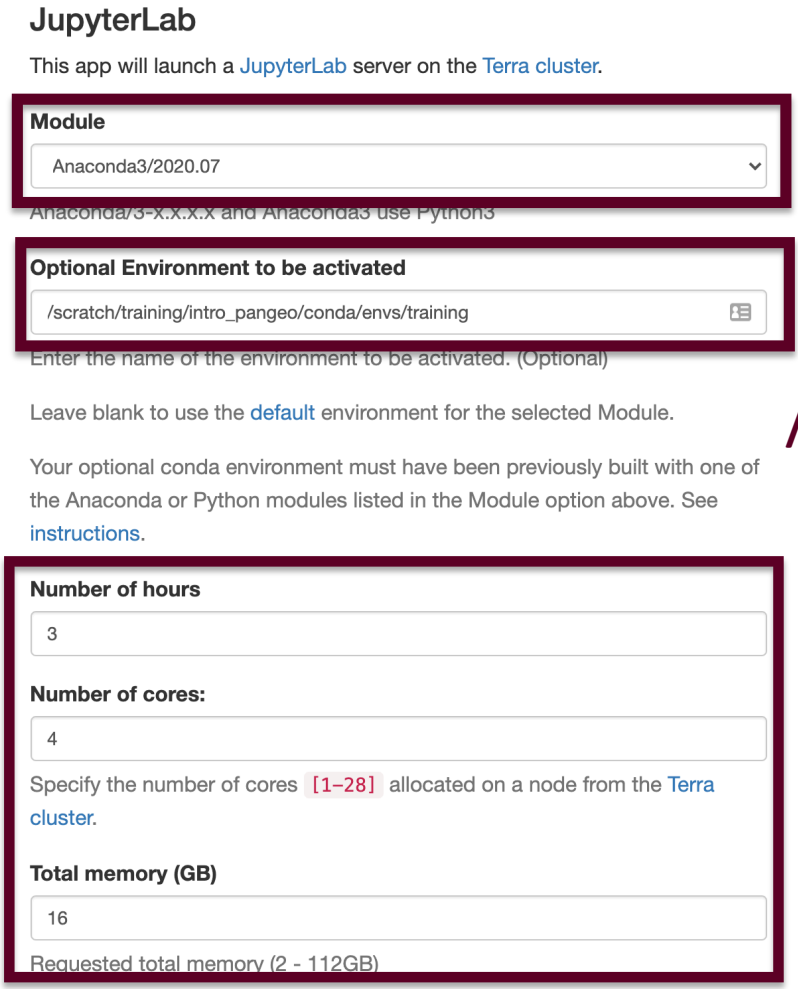
Launching a JupyterLab notebook from Terra portal

1. Go to https://portal.hprc.tamu.edu/

2. Interactive Apps -> JupyterLab







Choose module

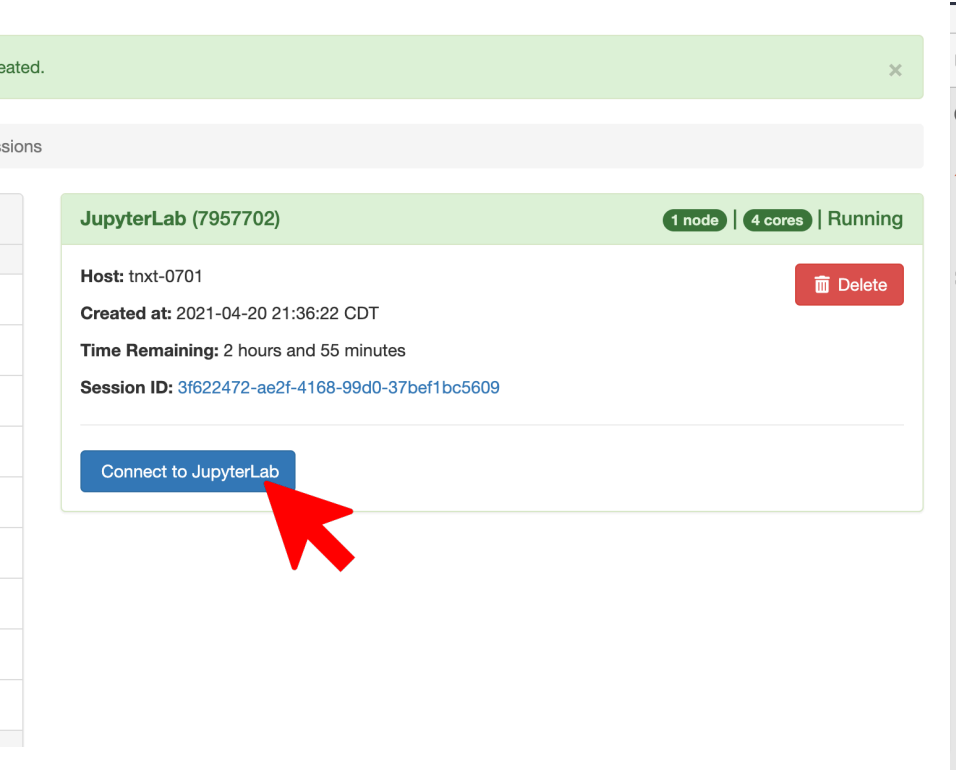
Enter environment path

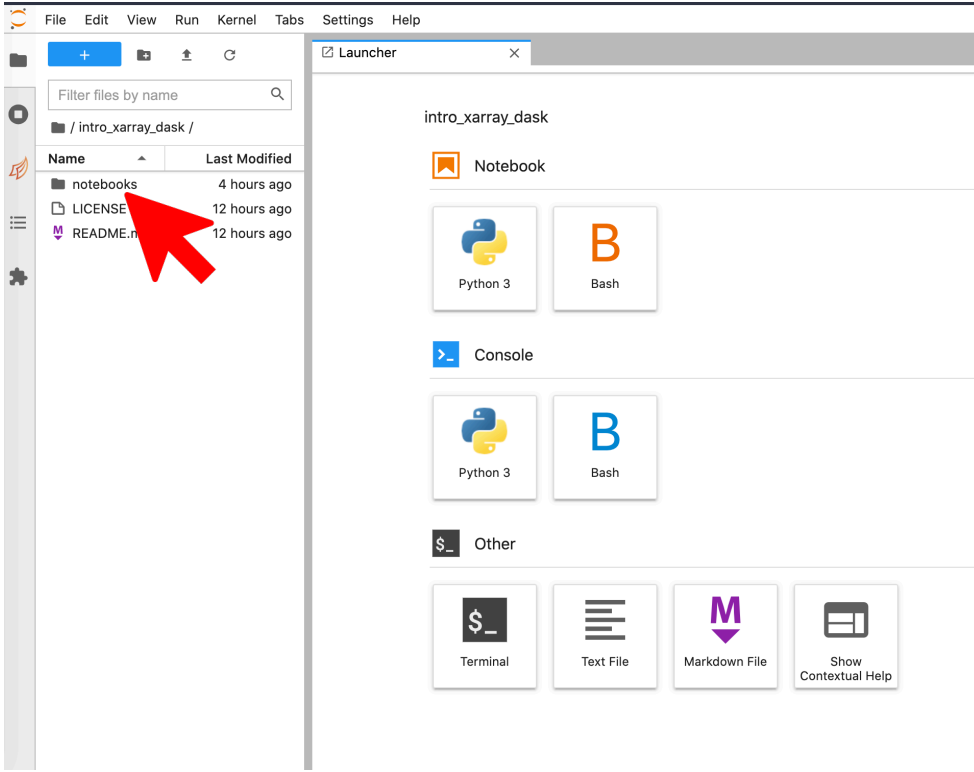
/scratch/training/intro_pangeo/conda/envs/training

Specify job wall time, cores and memory

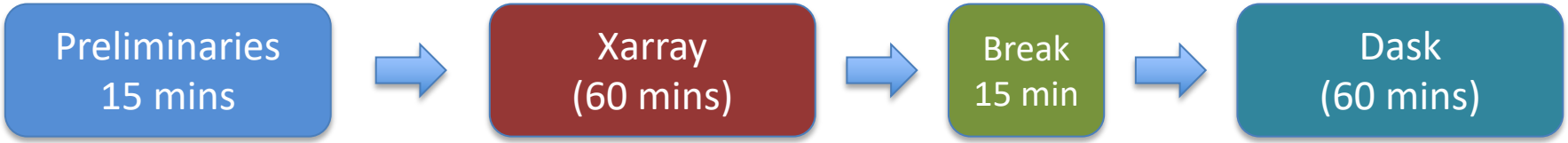
Hit Launch

Connect to JupyterLab session





Course structure



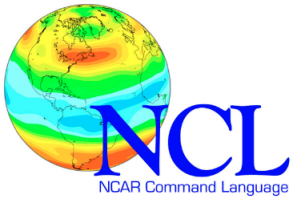
- Intro to the Pangeo stack
- Xarray data structures
- Reading and writing netCDF files
- Plotting with xarray, matplotlib and cartopy
- Dask chunking and lazy loading

- Faster computations with dask (high-level)
- Computations in xarray + xgcm
- Explicit parallelization in dask
- Using the dask dashboard to understand memory usage

Expectations for this course

- Get an overview of latest Python libraries designed to support geoscientific analysis
- Learn about the data structures in xarray, how to load and visualize netCDF files, and some basic operations
- Learn about benefits of lazy loading, and how dask can implicitly parallelize computations in xarray
- Explore other geoscience packages built on top of xarray

Current/last generation of post-processing tools

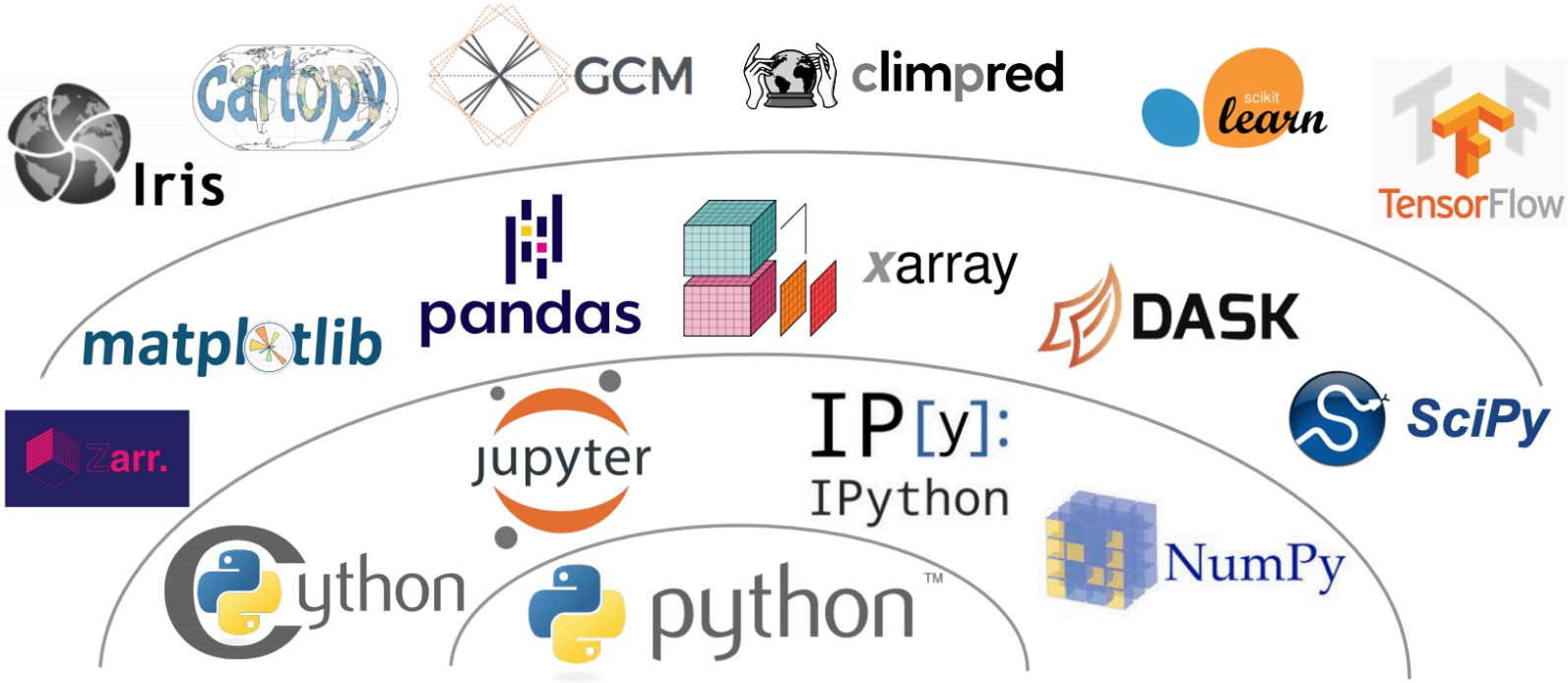


NOAA/ PMEL
FERRET

Max-Planck-Institut
für Meteorologie
CDO

- Mature tools/languages for working with moderate resolution datasets
- Often optimized to do specific tasks really well/fast.
- Not designed with high-resolution datasets in mind.

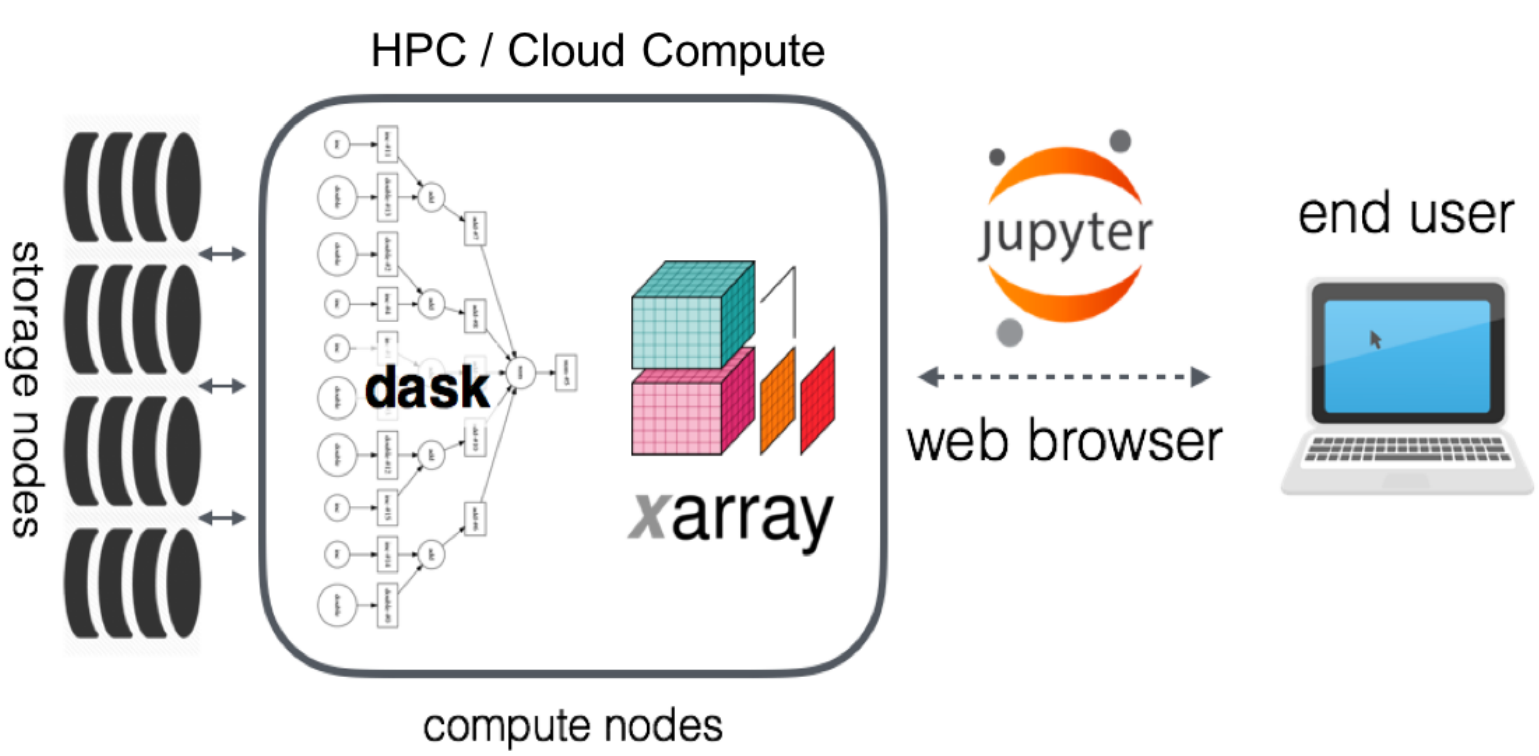
Python geo-scientific software stack



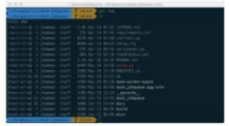
Credit: Ryan Abernathey. Inspired by Jake VanderPlas PyCon 2019

Pangeo

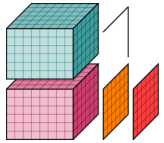
https://pangeo.io/architecture.html



BUILD YOUR OWN PANGEO

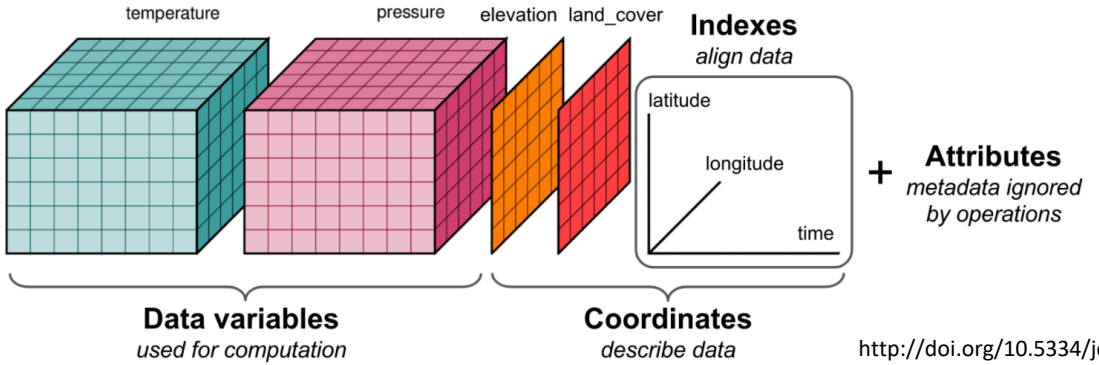
Storage Formats	HDF	OPeNDAP	Cloud Optimized COG/Zarr/Parquet/etc.
ND-Arrays	NumPy	DASK	More coming…
Data Models	xarray	Iris	pandas
Processing Mode	jupyter Interactive	Batch 	Serverless
Compute Platform	HPC 	Cloud  Google Cloud Platform	Local

https://www.ecmwf.int/sites/default/files/elibrary/2018/18737-why-pangeo-what-it-and-why-we-need-it.pdf



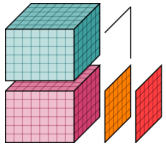
xarray

"pandas for N-dimensional arrays"



http://doi.org/10.5334/jors.148

- Builds on NumPy by applying metadata such as dimensions, coordinates, data variables and attributes to raw NumPy arrays.
- Inherits Pandas functionality
- xarray.Dataset is an in-memory representation of the netCDF file format
- xarray works seamlessly with the dask library to enable parallel computations more easily



xarray

Apply operations over named dimensions

```
x.sum('time')
```

Select values by label or logical conditions, instead of integer location

```
x.loc['2014-01-01']
x.sel(time='2014-01-01')
```

Easily use the split-apply-combine paradigm with groupby

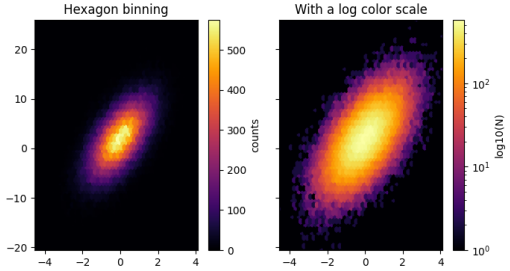
```
x.groupby('season').mean()
```

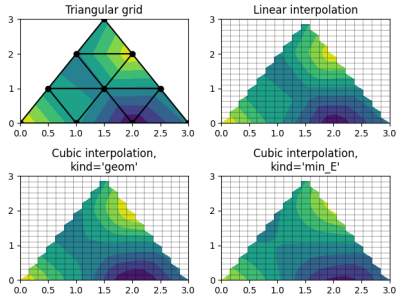
Keep track of arbitrary metadata in the form of a Python dictionary

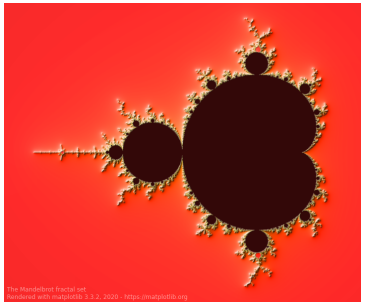
```
x.attrs
```

matplotlib

A comprehensive library for creating static, animated, and interactive visualizations in Python.



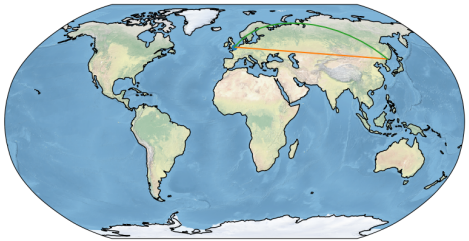


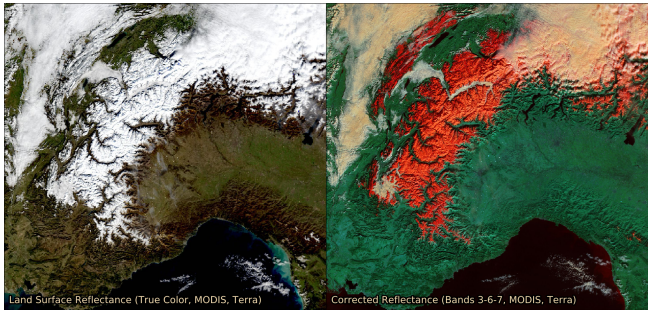


https://matplotlib.org/gallery/

Cartopy

Cartopy adds understanding of map projections to matplotlib plots





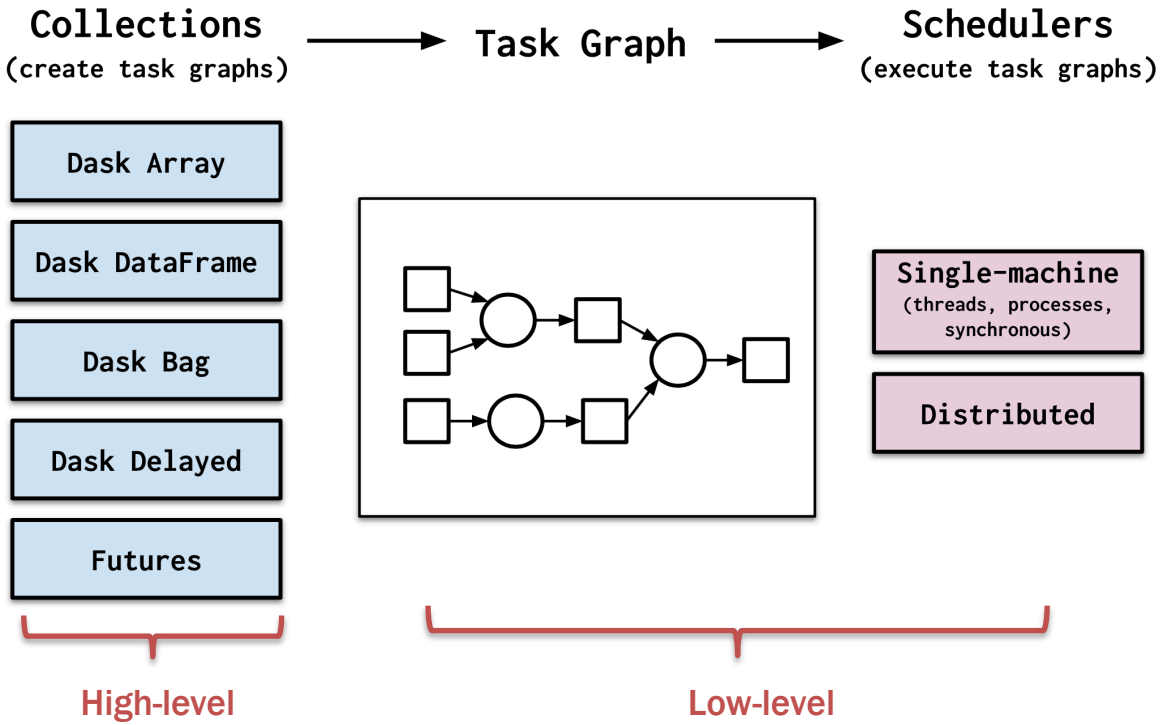
https://scitools.org.uk/cartopy/docs/latest/gallery/index.html

Short break!
(15 minutes)

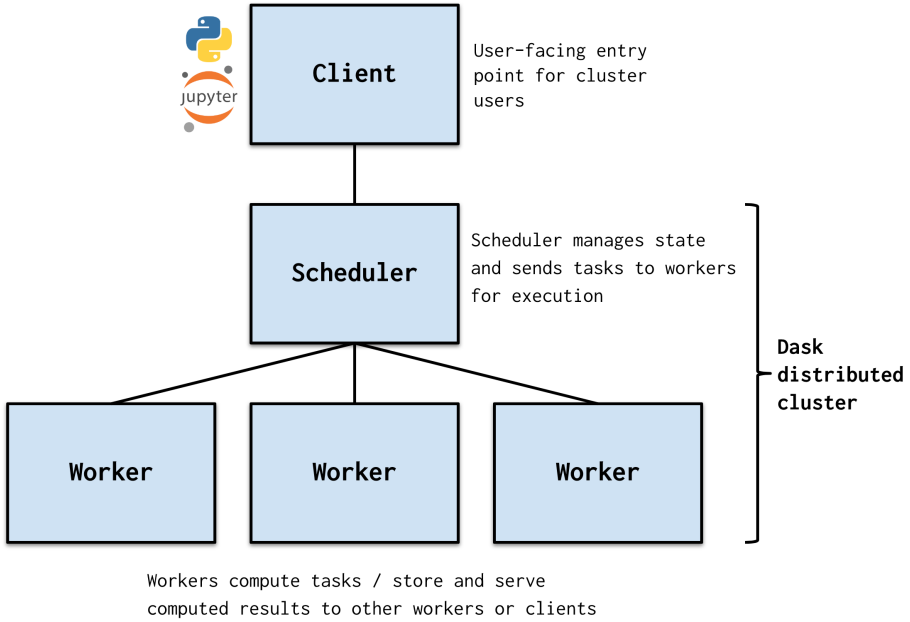
DASK

https://docs.dask.org/en/latest/

Dask provides multi-core and distributed parallel execution on larger-than-memory datasets.



Dask client-scheduler-worker



Instructions on using the dask dashboard on Terra

To start the distributed scheduler, we can simply do

[1]:
```
from dask.distributed import Client
client = Client(n_workers=4,threads_per_worker=4,memory_limit='2GB')
client
```

[1]:

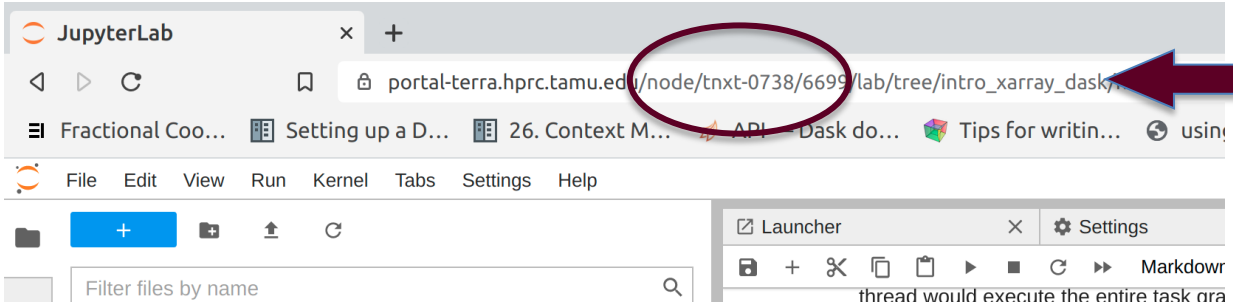
Client

Scheduler: tcp://127.0.0.1:46491
Dashboard: http://127.0.0.1:8787/status

Cluster

Workers: 4
Cores: 16
Memory: 7.45 GiB

Note down dashboard port
Ex. 8787



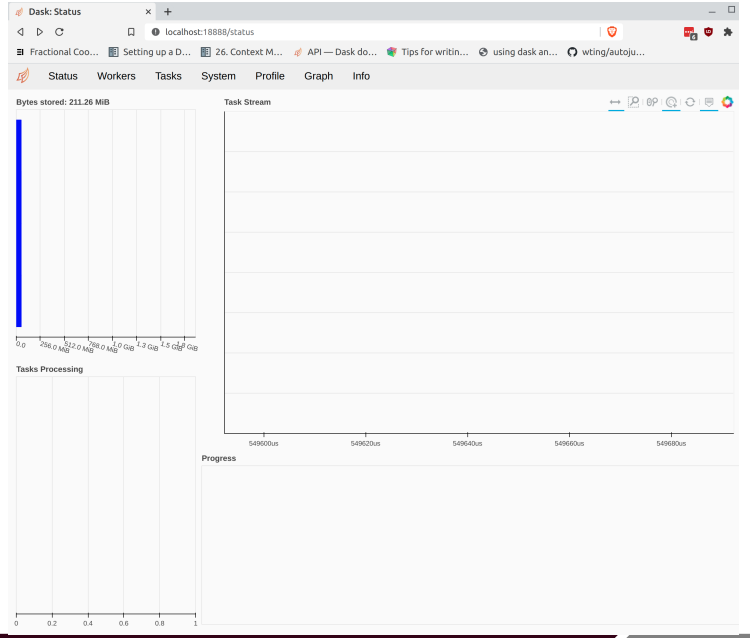
Note down Terra compute node. Ex. tnxt-0738

Instructions on using the dask dashboard on Terra

On your local machine

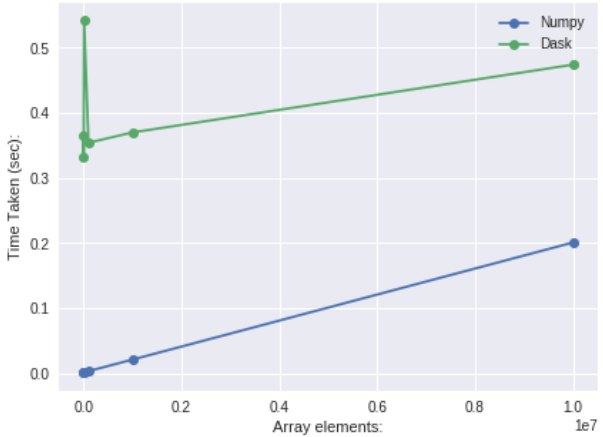
```
ssh -N -L 18888:tnxt-0738:8787 <username>@terra.tamu.edu
```

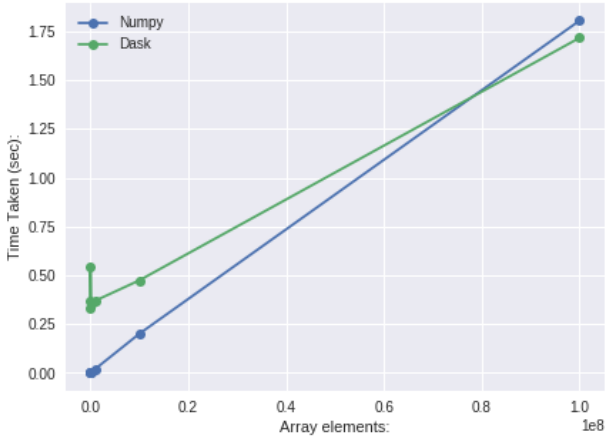
Point your local browser to http://localhost:18888

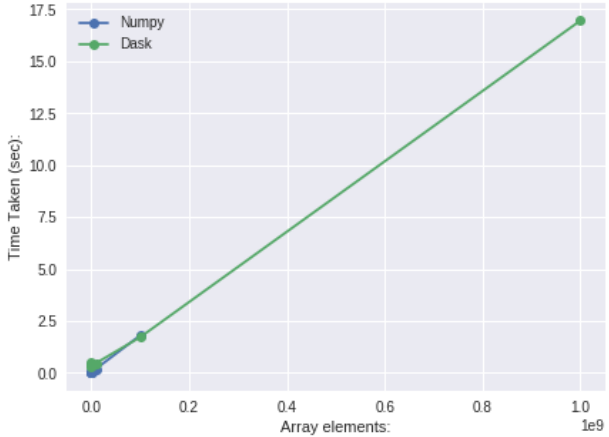


Dask vs NumPy

- NumPy is faster than Dask for a smaller problem size
- For larger problems, Dask achieves better scalability
- Larger datasets require correspondingly large amounts of memory with NumPy, and this is where dask's lazy loading shines







https://towardsdatascience.com/speeding-up-your-algorithms-part-4-dask-7c6ed79994ef

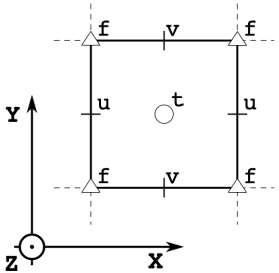
Key Takeaways

- The Pangeo framework rethinks how we analyze large datasets
  - Dask enforced lazy-loading + parallelization
  - In its developmental stages, and will take a few more years to reach the breadth of NCL
  - For newer analysis tools development, consider using Pangeo

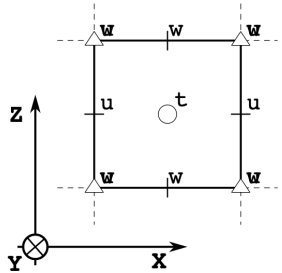
- NCO, CDO, etc are still extremely handy for specific tasks

Some great Python modules to go along with xarray and dask!

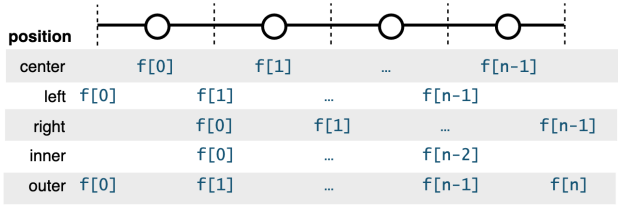
- xarray doesn't implicitly understand GCM grids
- xgcm wraps xarray to add an understanding of grid topology
- Implements spatial derivative operators
- Understands only C-grids for now, but other works are in progress
- New: Grid-aware vertical interpolation



C-grid — horizontal view



C-grid — vertical view



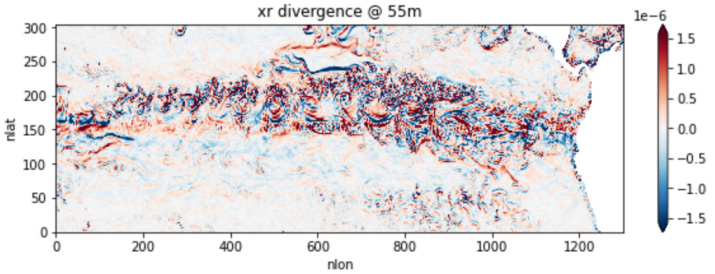
The different possible positions of a variable `f` along an axis.

https://xgcm.readthedocs.io/en/latest/grids.html

pop-tools

Tools to support analysis of POP2-CESM model solutions with xarray

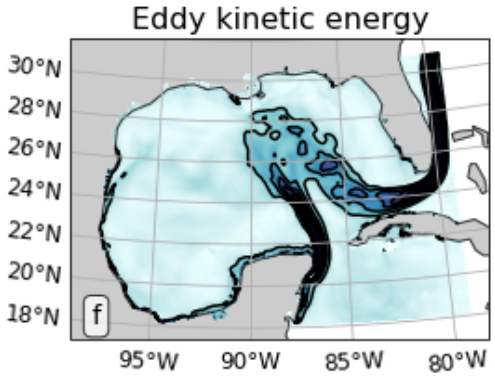
- Wraps xgcm to provide support for POP2 grids.
- Inherits spatial derivative operators from xgcm
- Support for POP2 region masks

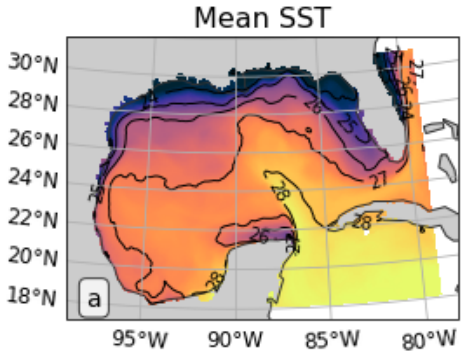


https://pop-tools.readthedocs.io/en/latest/

xroms

- Wraps xgcm to provide ROMS-specific grid manipulations and functions of interest to oceanographers.
- Developed by Kristen Thyng, Rob Hetland, et al. at TAMU
- Wraps **cf-xarray** to generalize coordinate and dimension calling.
- Wraps **xcmocean** to automatically choose colormaps for plotting!

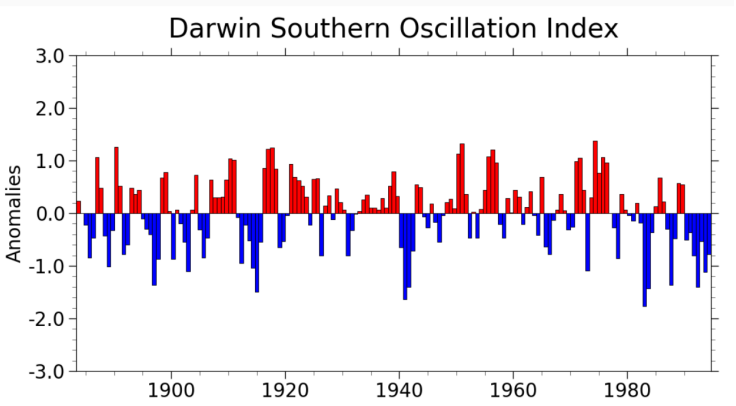


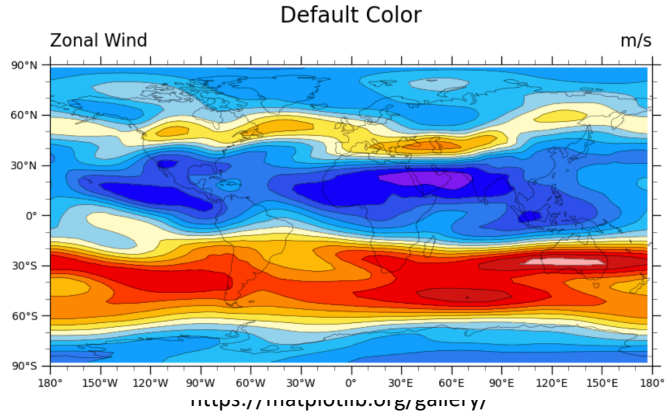


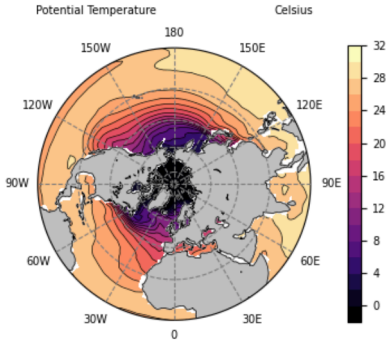
https://github.com/kthyng/xroms

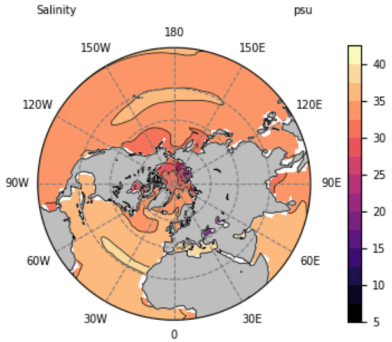
GeoCAT

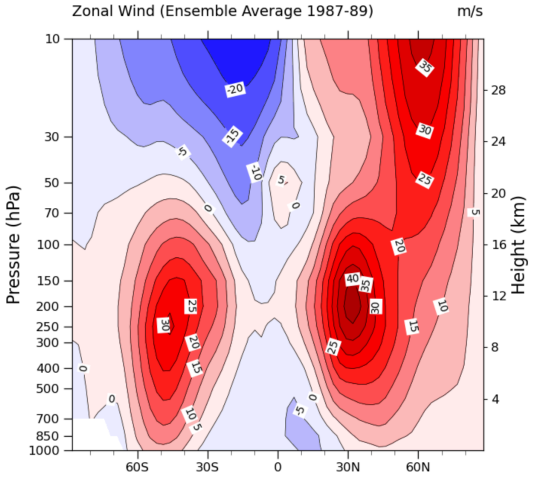
A collection of Python utilities related to NCL







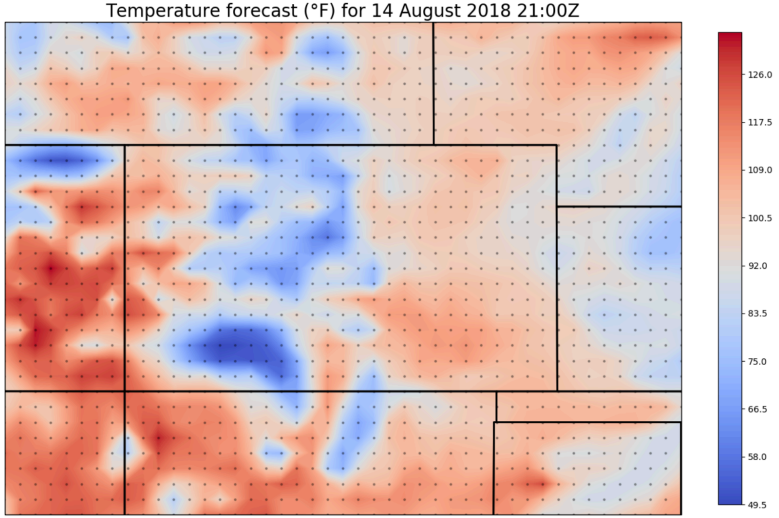


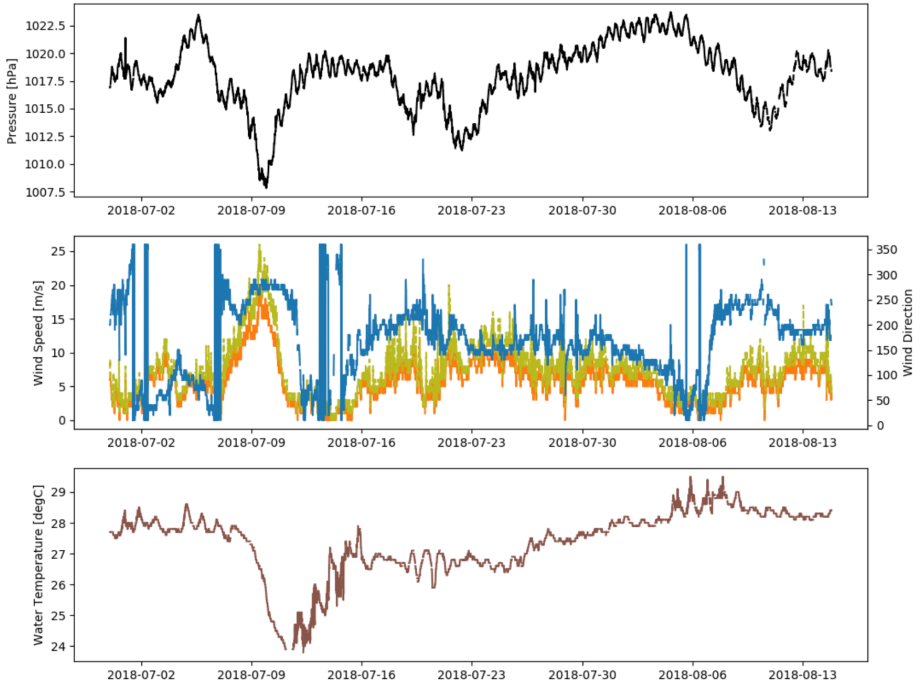


https://geocat-examples.readthedocs.io/en/latest/gallery/index.html

Siphon

A collection of Python utilities for downloading data from remote data services

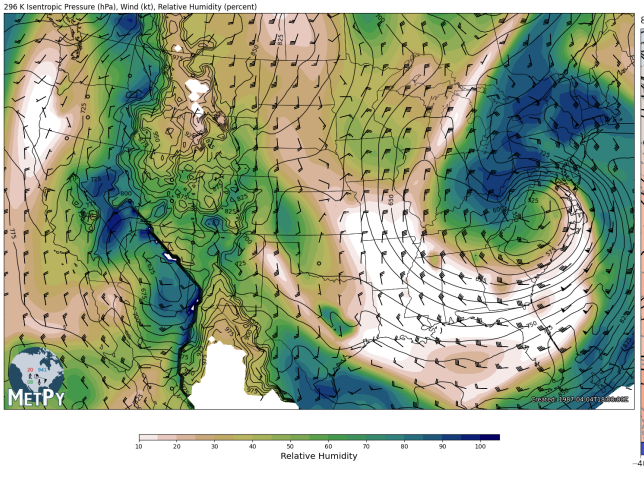


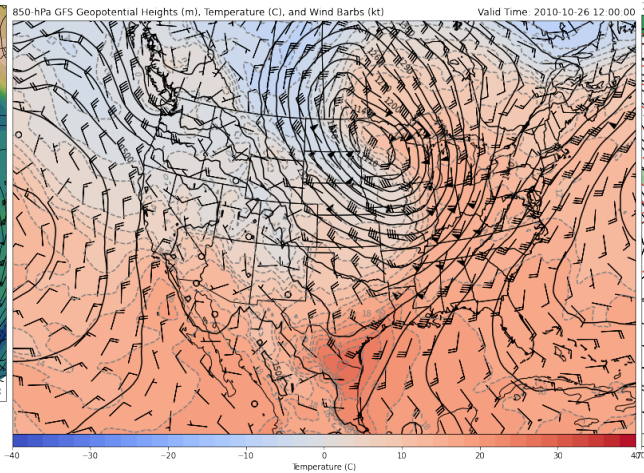


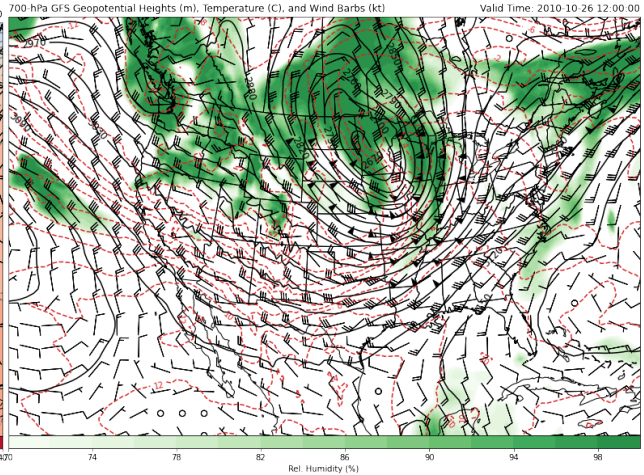
https://matplotlib.org/gallery/

https://scitools.org.uk/cartopy/docs/latest/gallery/index.html

A collection of Python tools for reading, visualizing, and performing calculations with weather data.







https://unidata.github.io/python-training/gallery/gallery-home/

Additional Python resources

- Previously offered HPRC short courses
  - Introduction to Python
    - https://hprc.tamu.edu/training/intro_python.html
  - Introduction to Scientific Python
    - https://hprc.tamu.edu/training/intro_scientific_python.html
  - Introduction to Python for MATLAB users
    - https://hprc.tamu.edu/training/python_matlab.html

- NumPy for MATLAB users (Quick reference)
  - http://mathesaurus.sourceforge.net/matlab-numpy.html

Additional resources

- Official Documentation
  - xarray docs
  - Dask docs

- Ask for help:
  - Use the python-xarray on StackOverflow
  - GitHub Issues for bug reports and feature requests
  - dask tag on Stack Overflow, for usage questions
  - github issues for bug reports and feature requests
  - Pangeo forums http://discourse.pangeo.io/

Questions?