

JupyterHubs and containerization to streamline and accelerate science innovation

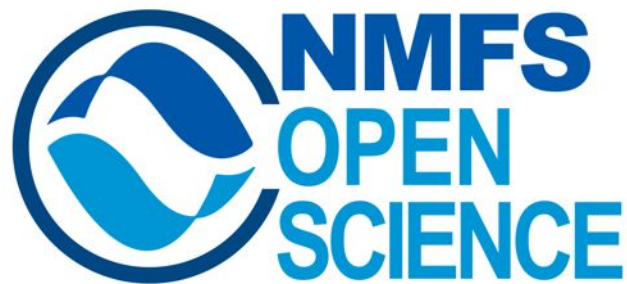
May 17, 2024 EDMW 2024

NOAA Fisheries Open Science

Elizabeth Eli Holmes, Ph.D

NMFS Open Science, Lead

Since 2022, working with NASA DAAC (data centers) colleagues on
open infrastructure for geospatial analysis for cloud-native data



NOAA
FISHERIES

NOAA Fisheries Open Science JupyterHub



Login to the JupyterHub

Sign in with GitHub

Funded by the [NOAA and Microsoft Cooperative Research and Development Agreement \(CRADA\)](#) to help advance NOAA's mission to create a Climate-Ready Nation. This JupyterHub is used for internal NOAA training and workshops in cloud computing and scientific programming to support this mission. These trainings focus on analysis with geospatial earth climate data and output from climate models, processing of acoustics data from passive and active surveys, and bioinformatics toolsets for analysis of fisheries genetic data.

NOAA Fisheries HackHours Home Topics ▾



HackHours 2024

Python - ArcGIS
Python - CMIP6 climate forecasts
Python - Parallel Computing
R - VAST
Python - echopype 2
Python - PACE Ocean Color

Select Language ▾

Powered by [Google Translate](#)

HackHours 2024

Select an event to the left

- [May 17](#). Visualizing the data from the [PACE satellite](#) mission (Python)
- May 10. Accessing Earthdata in Python. Test run of our Eli Holmes (Python)
- May 3. Accessing Earthdata in R. Test run of our R tutor
- [April 26th](#). More acoustics data with [echopype](#). Wu-Jur
- [April 19th](#). Parallel computing with [dask](#) and [coiled.io](#).
- [April 12th](#). Exploring CMIP6 data with pangeo Python t
- [April 5th](#). Using ArcGIS via the [arcgis Python API](#). Tim I
- March 29th. Accessing acoustics data from AWS Open I
- March 22nd. 1pm PT [CoastWatch tutorials](#) in R. Sunny
- March 15th. [CoastWatch tutorials](#). Sunny Hospital and
- March 8th. Using precipitation estimates from IMERG t

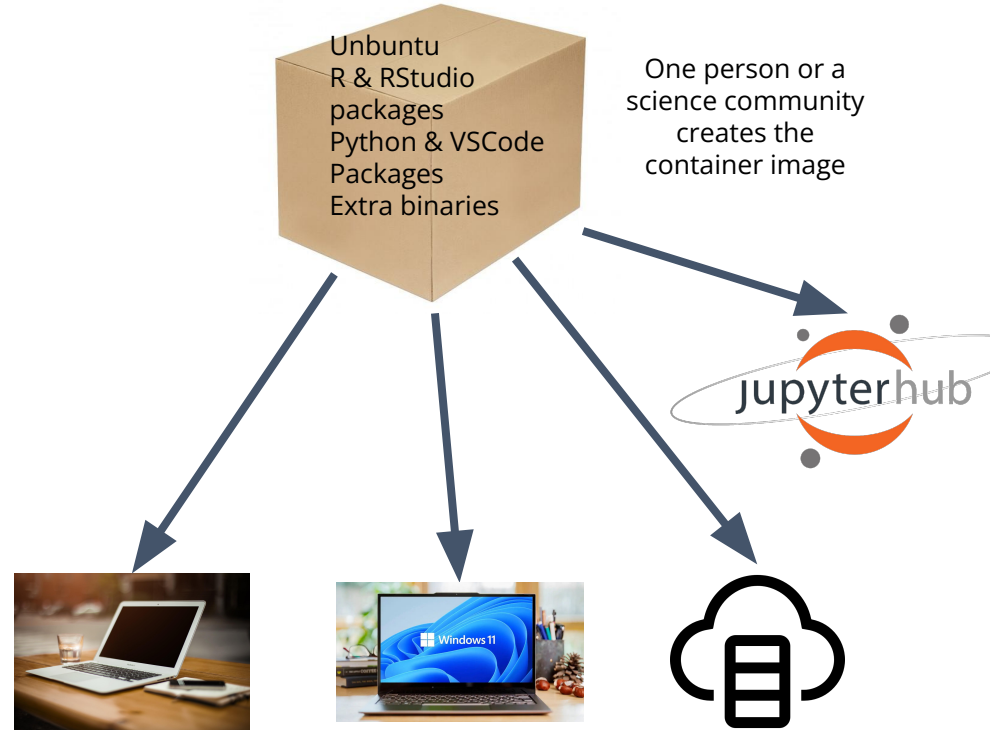


NOAA
FISHERIES

Containerization




Install MacOS
Install R
Install RStudio
Install packages
Install Python
Install VSCode



No one installs anything. You run the container.


JupyterHubs: a multi-user platform for running containers/images


OS
pangeo +
machine-learning
tools + one of our
python modules



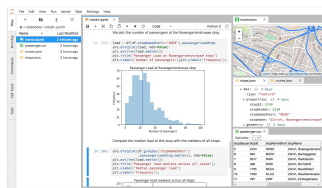
OS
rocker/geospatial
with R 4.2 + suite
of packages

OS
Python + R +
C++
Atlantis model

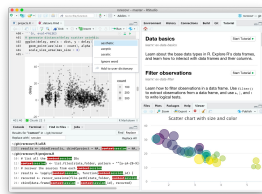

microbiome
bioinformatics


acoustics
pipelines

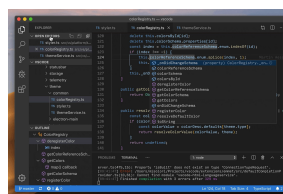
Container is all our
packages and the os
environment set up



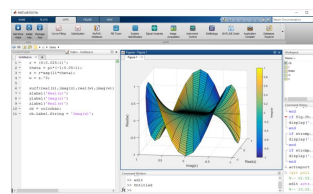
jupyterlab



rstudio



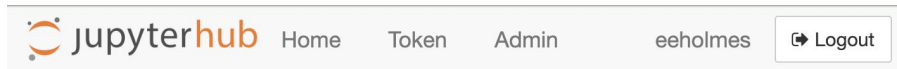
vscode



matlab

Why are JupyterHubs so popular?

User experience is familiar and spin-up easy



Server Options

Select image and RAM

Start a container with at least a chosen share of capacity on a node of this type

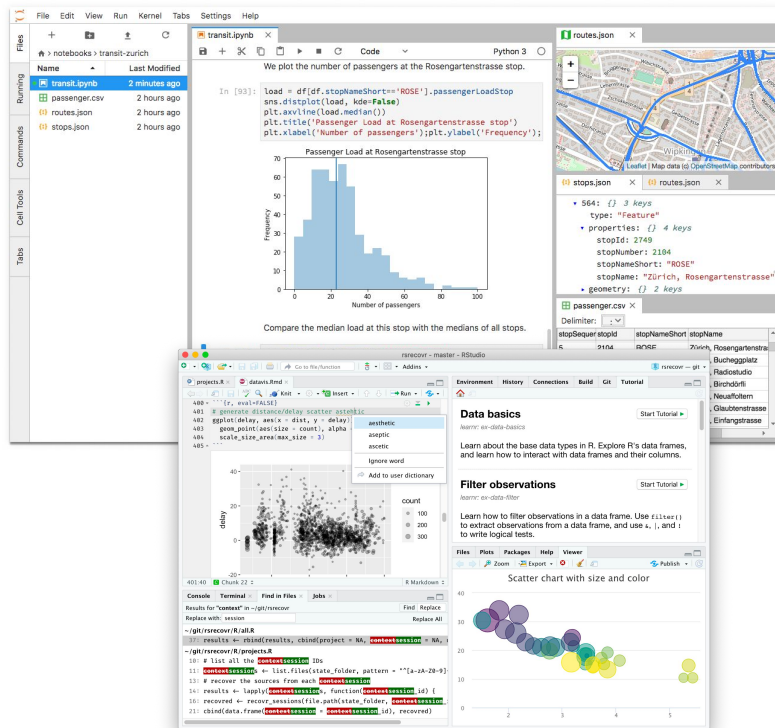
Image

Py-R - Base geospatial image 4.4-3.10

Resource
Allocation

1.9 GB RAM, upto 3.7 CPUs

Start



Share complex compute environments easily

The screenshot shows the JupyterHub web interface. At the top, there's a navigation bar with 'Home', 'Token', 'Admin', and a user profile 'eeholmes' with a 'Logout' button. A dropdown menu is open, listing various pre-configured images. The menu items are: 'Py - Tourmaline Snakemake workflow for QIIME 2 v.2023.5', 'Py - ArcGIS Python 3.9', 'R-Py - NASA TOPS Env Jus - boettiger-lab nasa-tops latest', 'Py - CMIP6-Cookbook', 'Py-R - CoastWatch - nmfs-opensci coastwatch latest', 'Py - Echotype with pangeo nmfs-opensci echotype latest', 'Py - Python w tensorflow - eeholmes iopython 20230714', 'R - R geospatial w sdmTMB - eeholmes iorocker 20230729', 'Py-R - base image 4.4-3.10' (which is highlighted with a blue bar and a checkmark), 'Py - Openscapes Python 39dffde', 'R - Openscapes rocker a7596b5', and 'Py - pangeo base-notebook 2024.01.23'. On the left, there's a sidebar with 'Select image' and 'Start a container' buttons. Below the dropdown, there are labels for 'Image', 'Resource', and 'Allocation'. At the bottom, there's an orange 'Start' button.

jupyterhub Home Token Admin eeholmes Logout

Select image

Start a container

Image

Resource

Allocation

Py - Tourmaline Snakemake workflow for QIIME 2 v.2023.5

Py - ArcGIS Python 3.9

R-Py - NASA TOPS Env Jus - boettiger-lab nasa-tops latest

Py - CMIP6-Cookbook

Py-R - CoastWatch - nmfs-opensci coastwatch latest

Py - Echotype with pangeo nmfs-opensci echotype latest

Py - Python w tensorflow - eeholmes iopython 20230714

R - R geospatial w sdmTMB - eeholmes iorocker 20230729

Py-R - base image 4.4-3.10

✓ Py-R - Base geospatial image 4.4-3.10

Py - Openscapes Python 39dffde

R - Openscapes rocker a7596b5

Py - pangeo base-notebook 2024.01.23

Start

NMFS Open Science Docker Stack

THE DOCKER STACK IS IN ACTIVE DESIGN and DEVELOPMENT

Beta release targeted for June 1, 2024.

These are a collection of container images to provide standardized environments for Python and R computing built off the [Rocker](#), [Pangeo](#) and Jupyter base images. This repo holds the (mostly) stable docker stack for specific pipelines used in Fisheries. Why use a container? The main reason is that geospatial, bioinformatics, and TMB/INLA environments can be hard to get working right. Using a Docker image means you use a stable environment. Watch this video from Yuvi Panda (Jupyter Project) [video](#) and read about the Rocker Project in the R Project Journal [article](#) by Carl Boettiger and Dirk Eddelbuettel.

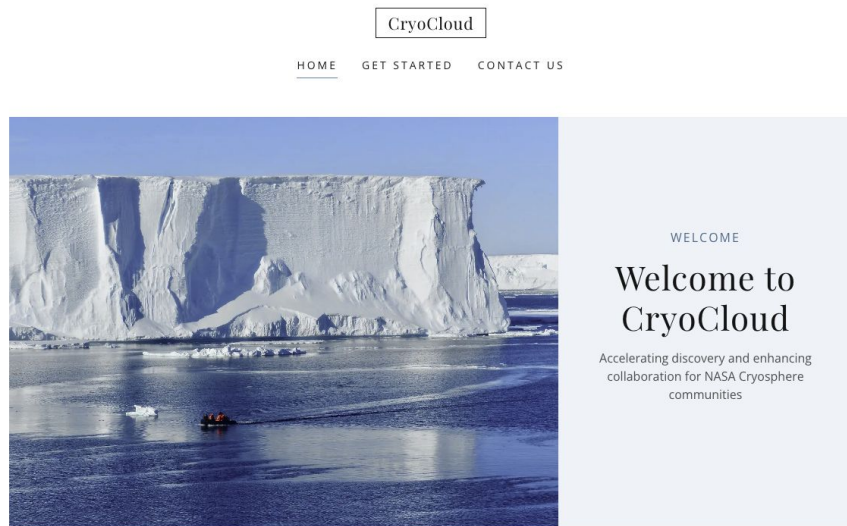
Stable set of images

There are many other images in the `images` folder that are experimental in nature. There are also experimental images in the branches.

It removes the whole software and package installation problem—even for complex environments

Description	Size	Link	Dockerfile
Python based Openscapes		nmfs-opensci-python-base	Dockerfile

JupyterHubs can accelerate collaboration

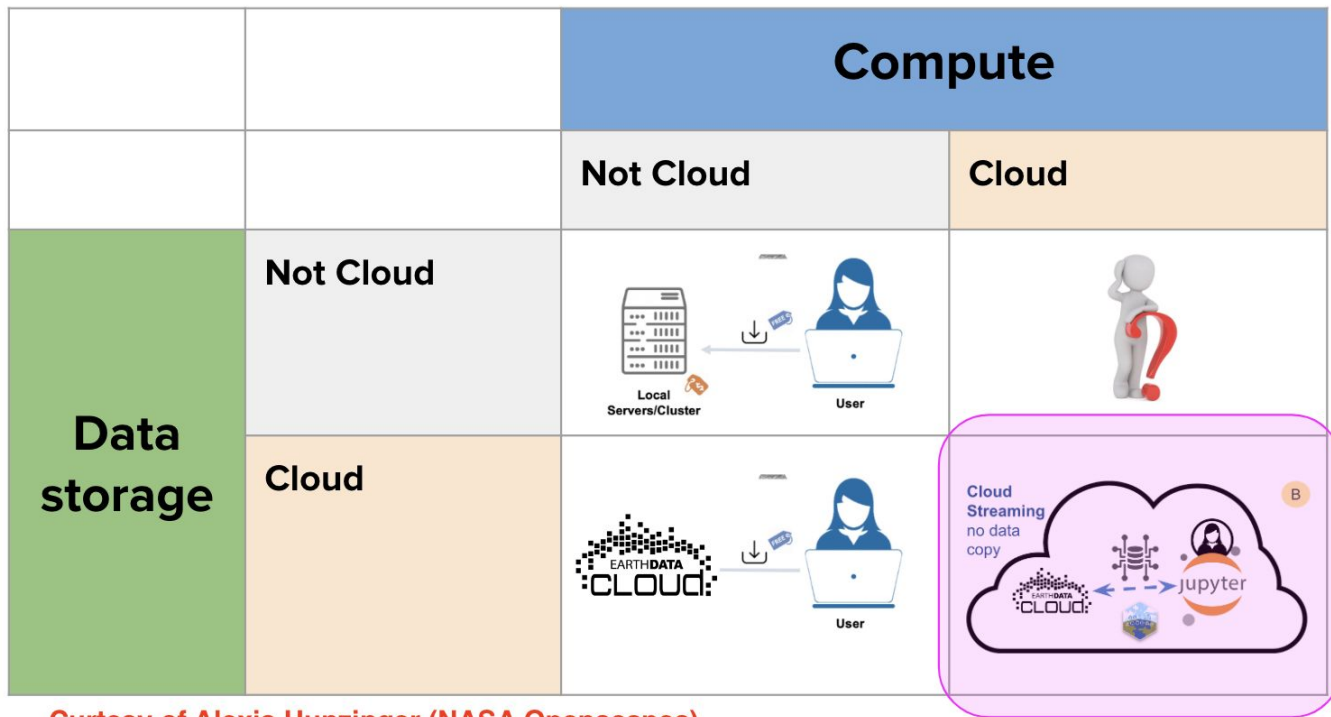


CryoCloud example (Tasha Snow, PI).
ICESat-2 community



James Colliander (presenting for PI
Tasha Snow at JupyterCon 2023):
Accelerating Discovery for NASA
Cryosphere Communities with
JupyterHub

JupyterHubs allow you to get “next” to the data



Data IO is often the bottleneck.

Being next to the data buckets means you can “attach” a huge cloud drive to your virtual computer*.

Cloud-native provides new options for massively parallel computing.

Courtesy of Alexis Hunzinger (NASA Openscapes)

*if all data were cloud-optimized it might not matter, but it's not.

JupyterHubs use Open Infrastructure

Infrastructure (JupyterHub) is cloud-provider agnostic and can be run on 'bare-metal' or even a laptop.

Base "Images" are infrastructure agnostic. Can be dropped into:
Jupyter server,
docker/podman (offline),
Binder, GitHub Codespaces,
GitLab

How does one run a JupyterHub?

Common ways people run JupyterHubs

Self-hosted hubs
on 'bare-metal'
servers or own
cloud account



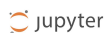
managed hubs



Log in to continue

Welcome to the Visualization
Exploration, and Data Analysis (VEDA)
Project: 2i2C JupyterHub

This is a pilot service running on open source
infrastructure. See the VEDA Pilot documentation for
usage and deployment information.



Institution
managed hubs



Operated by the Division of Computing, Data Science, and Society

Log in to continue

Welcome to the University of
California, Berkeley **DataHub**.



Easy creation: A JupyterHub on Kubernetes on GCP in 5 minutes

Open a Cloud Shell and run these lines of code

```
gcloud container clusters create \
  --machine-type n1-standard-2 \
  --num-nodes 2 \
  --zone us-west1-a \
  --cluster-version latest \
  jhub

kubectl create clusterrolebinding cluster-admin-binding \
  --clusterrole=cluster-admin \
  --user=yourname@gmail.com

gcloud beta container node-pools create user-pool \
  --machine-type n1-standard-2 \
  --num-nodes 0 \
  --enable-autoscaling \
  --min-nodes 0 \
  --max-nodes 3 \
  --node-labels hub.jupyter.org/node-purpose=user \
  --node-taints hub.jupyter.org_dedicated=user:NoSchedule \
  --zone us-west1-a \
  --preemptible \
  --cluster jhub
```

```
curl
https://raw.githubusercontent.com/helm/helm/HEAD/scripts/get-helm-3
```

```
helm repo add jupyterhub https://hub.jupyter.org/helm-chart/
helm repo update
```

```
helm upgrade --cleanup-on-fail \
  --install jhub1 jupyterhub/jupyterhub \
  --namespace jhubk8 \
  --create-namespace \
  --version=3.3.4 \
  --values config.yaml
```

DONE!!!

Cost for running your own JupyterHub. My experience.

In the cloud for 3 people full-time
\$500 / month

But there is the time to set-up. About a week of futzing if you have never done it. 2-4 hrs of futzing otherwise.
AWS, Google, Azure, any other cloud provider

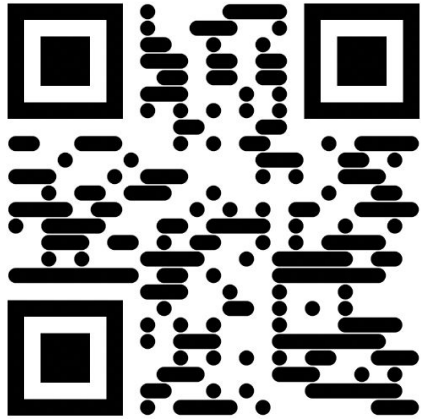
Large community JupyterHub
focused on geospatial analysis

- \$2500/mo admin and support (outside)
- \$1000-1500/mo compute + storage

See a JupyterHub Demo

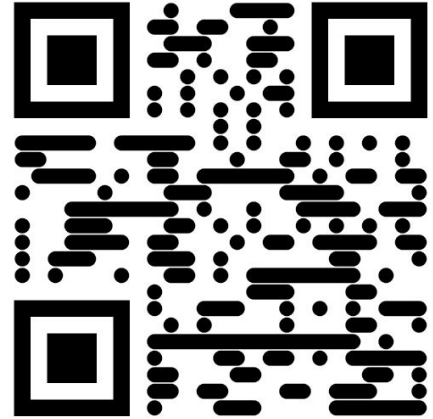
Basic

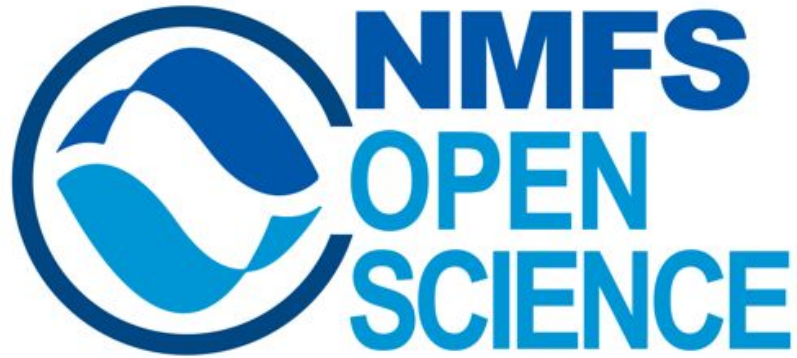
<https://youtu.be/S-OXTg9yadQ>



Bring your own image

<https://youtu.be/WvyepapjTEE>





<https://nmfs-opensci.github.io/>

Look for link to our internal site!