



OCEAN
OBSERVATORIES
INITIATIVE

Analysis Ready, Cloud Optimized Cabled Data with Zarr


Joe Duprey, Wendi Ruef, Mariela White,
Mike Vardaro

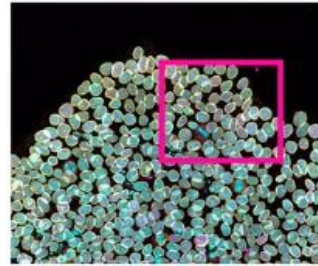
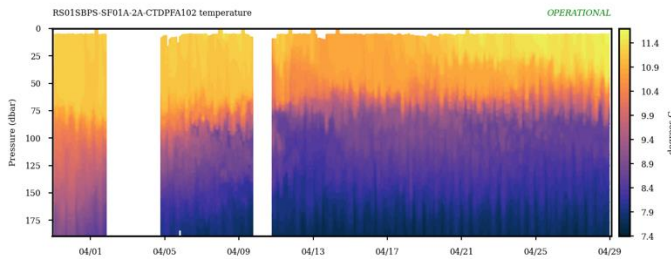
OOIFB-DSC Fall Meeting
October 13, 2024



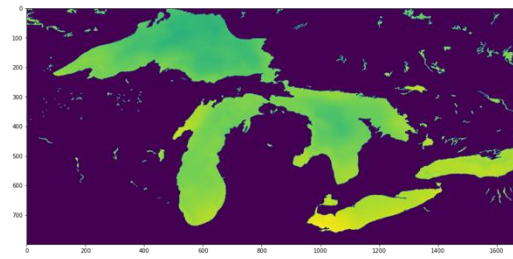
Why we love working with Zarr:

1) Open-source tool with an active community

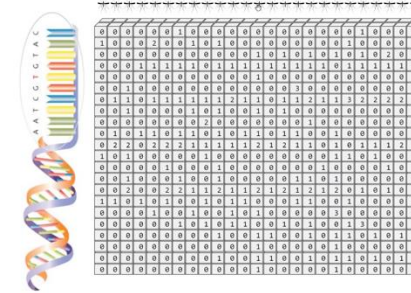
- Actively maintained
- Easy to find guidance for common problems
- User community with shared challenges: big data, consolidating archives
- Available data  accessible data



Moore et al 2023



NASA Openscapes



Miles 2018

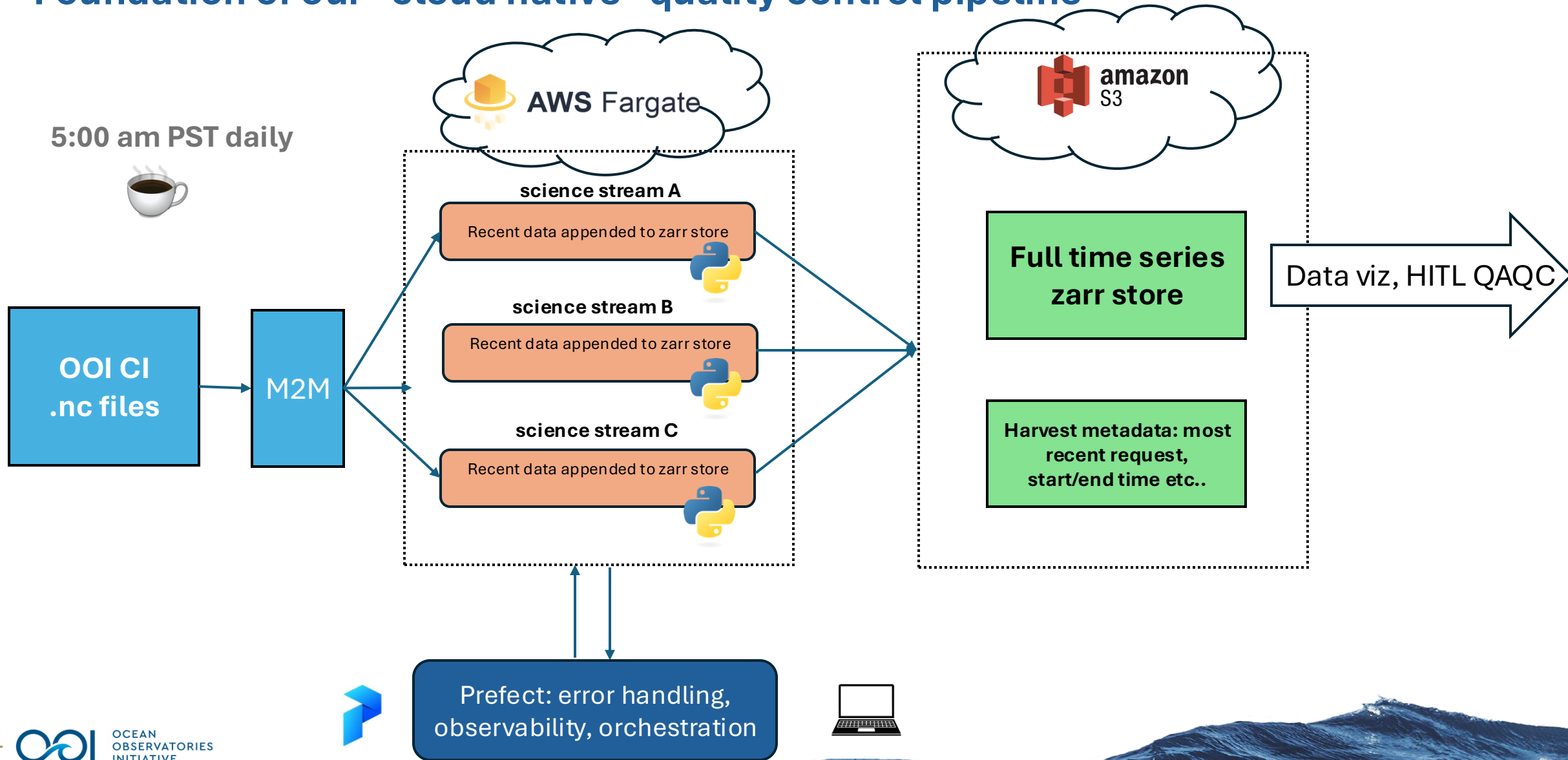
2) Cloud optimized format

- We can run many automated analyses in parallel every morning
- Data team can spend less time tinkering with compute / infrastructure
- More time conducting quality control
- More time improving visualization
- More time expanding QA/QC methodology

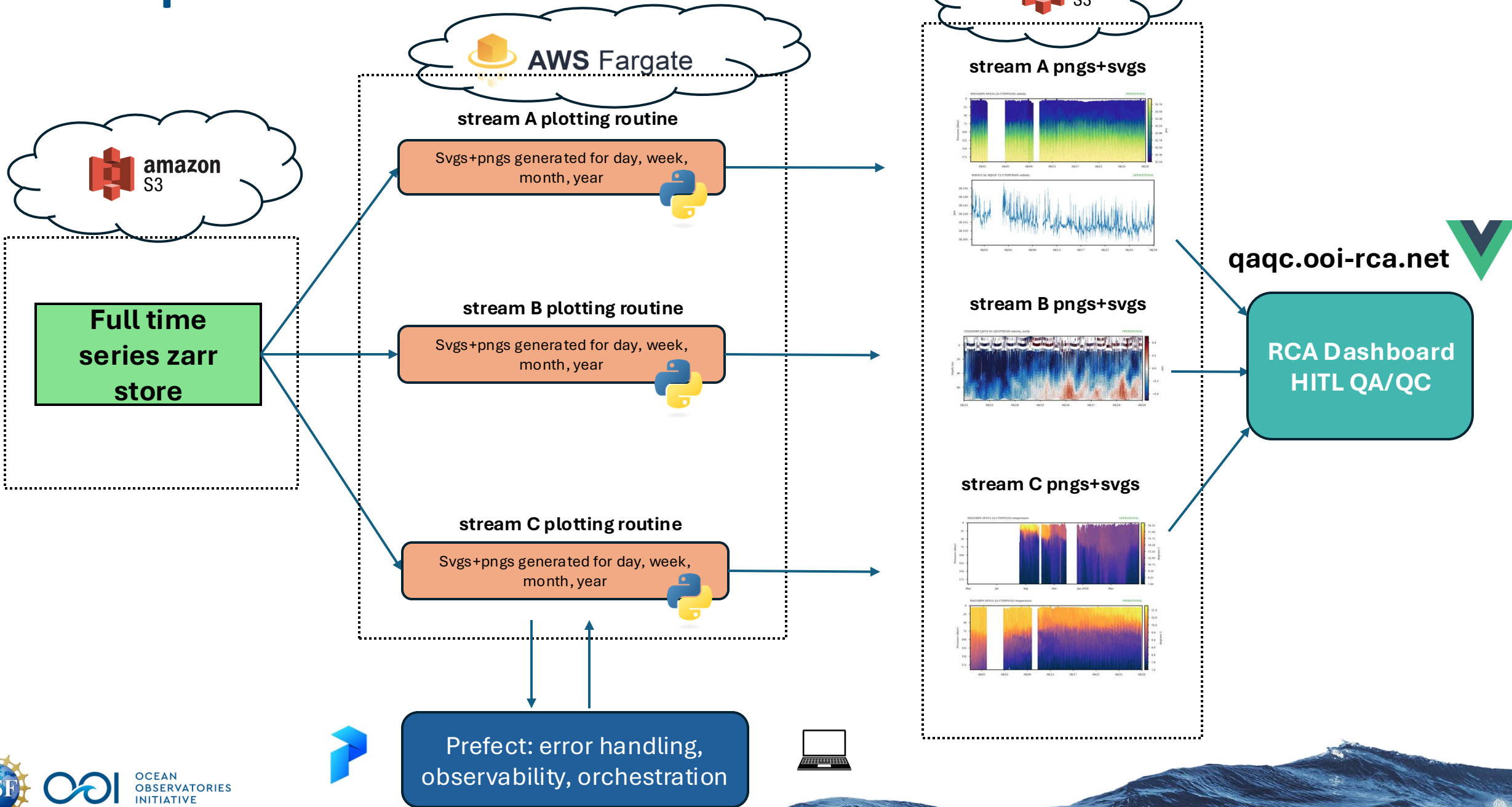


Cloud optimized format...

- Foundation of our “cloud native” quality control pipeline

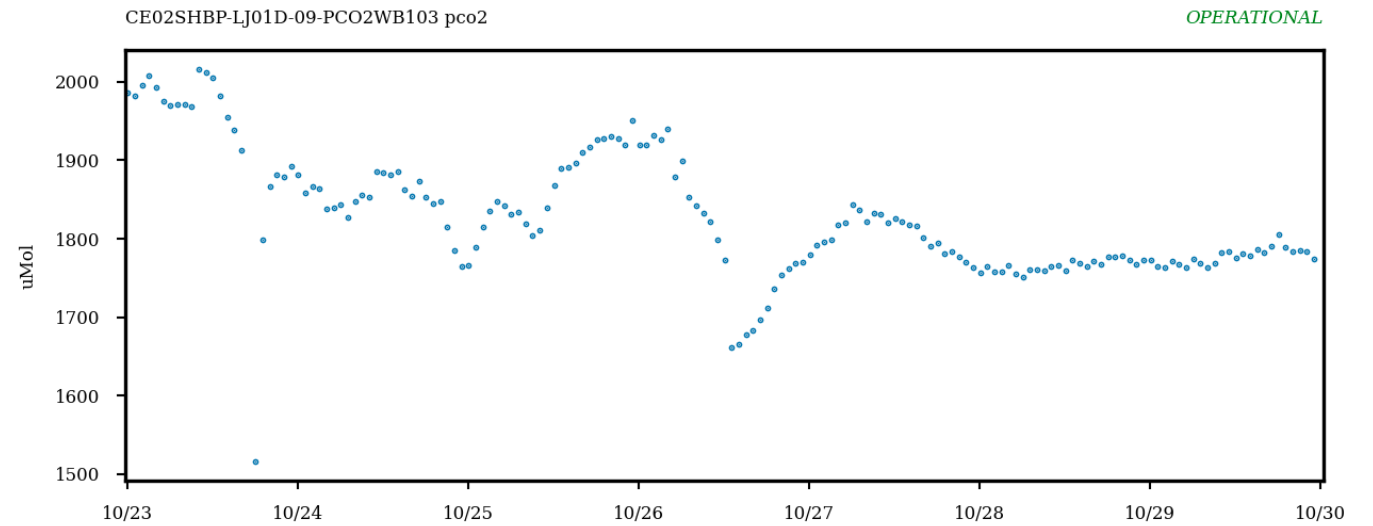
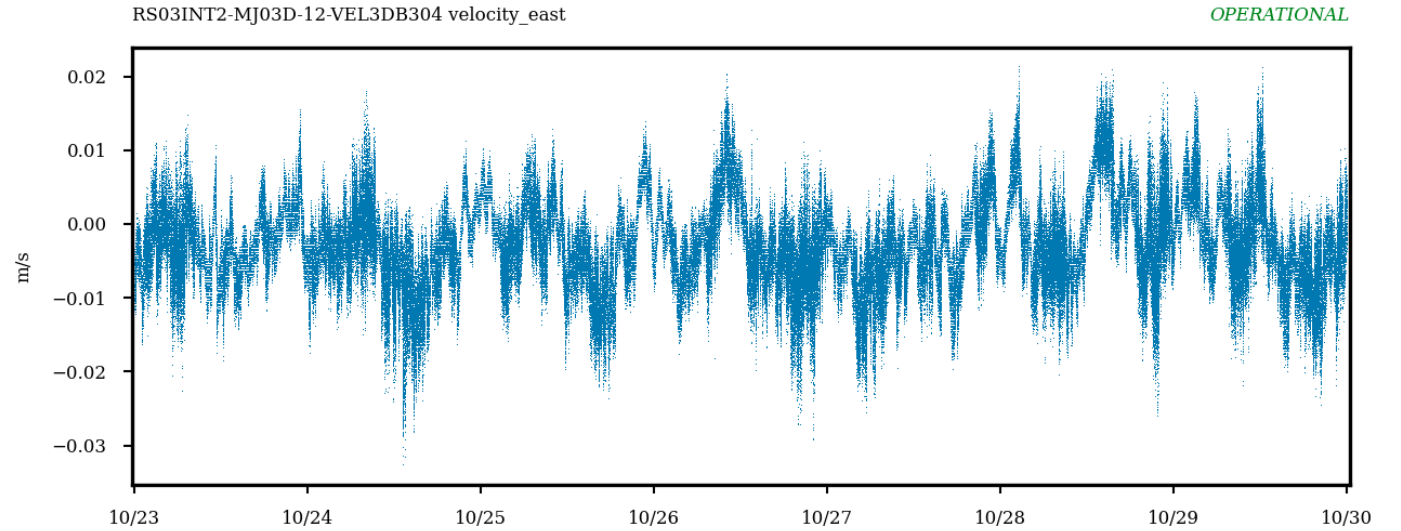


Cloud optimized format...



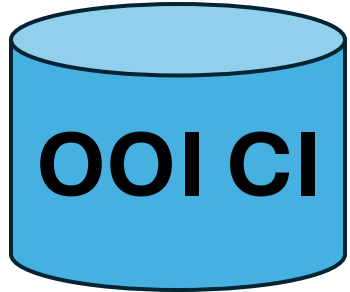
3) Zarr is scalable

- Same pipelines work for low density and high density data
- Simple to append new data to existing store



Zarr is scalable...

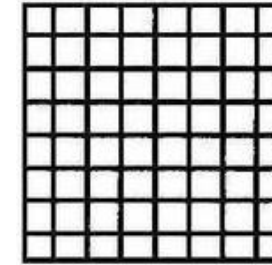
- Our zarr harvest is similar to gold copy pipeline



AWS Fargate

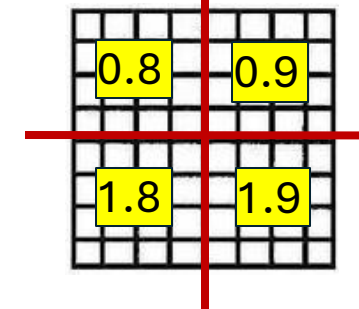
1) Dimensions are validated

2024-10-13



2) Data are chunked

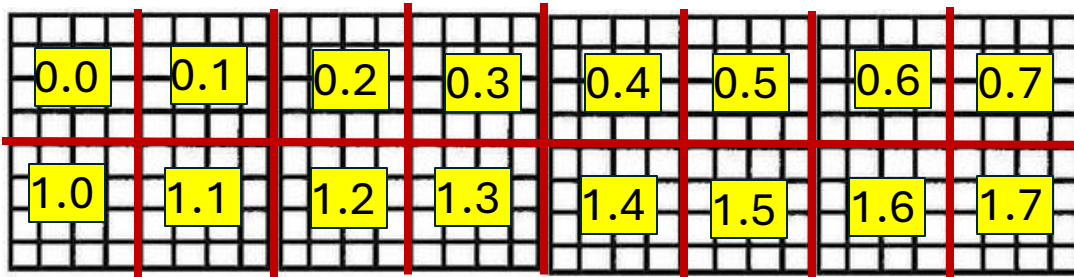
2024-10-13



S3://ooi-rca-data

2014-07-11

2024-10-12



4) Zarr is intuitive

- Consolidated, human-readable metadata
- Can quickly load all metadata into memory without loading data itself:

```
"stream": "adcp_velocity_beam",
"subsite": "CE02SHBP",
"summary": "Dataset Generated by Stream Engine from Ocean Observatories Initiative",
"time_coverage_end": "2024-10-29T11:02:18.433804288",
"time_coverage_start": "2014-09-25T18:10:07.729999872",
"title": "Data produced by Stream Engine version 1.20.13 for CE02SHBP-LJ01D-05-ADCPTB104-streamed-velocity_beam"
},
".zgroup": {
  "zarr_format": 2
},
"bin/.zarray": {
  "chunks": [
    22
  ],
  "compressor": {
    "blocksize": 0,
    "clevel": 5,
    "cname": "lz4",
    "id": "blosc",
    "shuffle": 1
  }
},
```



Zarr is intuitive...

- Lots of array files converted to a single store for an entire time series with an intuitive interface



Parent Directory

deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample.ncml	2024-10-24 18:41	3.1K
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180101T000000.072852-20180115T235959.441816.nc	2024-10-24 18:34	457M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180116T000000.439950-20180131T235959.361761.nc	2024-10-24 18:35	467M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180201T000000.362396-20180215T235959.798609.nc	2024-10-24 18:35	457M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180216T000000.798616-20180303T235959.828334.nc	2024-10-24 18:36	484M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180304T000000.827821-20180320T235959.719274.nc	2024-10-24 18:37	461M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180321T000000.719388-20180405T235959.781239.nc	2024-10-24 18:37	486M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180406T000000.780520-20180421T235959.647306.nc	2024-10-24 18:38	487M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180422T000000.648147-20180507T235959.265704.nc	2024-10-24 18:38	487M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180508T000000.265712-20180523T235959.802046.nc	2024-10-24 18:39	487M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180524T000000.801639-20180608T235959.405142.nc	2024-10-24 18:39	487M
deployment0004_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180609T000000.405150-20180624T010603.227388.nc	2024-10-24 18:40	458M
deployment0004_RS01SBPS-PC01A-4A-D0STAD103-streamed-do_stable_sample.ncml	2024-10-24 18:41	1.4K
deployment0004_RS01SBPS-PC01A-4A-D0STAD103-streamed-do_stable_sample_20171231T235959.074510-20180403T000000.588909.nc	2024-10-24 18:41	585M
deployment0004_RS01SBPS-PC01A-4A-D0STAD103-streamed-do_stable_sample_20180402T235959.589109-20180624T005819.185111.nc	2024-10-24 18:41	537M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample.ncml	2024-10-24 18:41	3.1K
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180624T161351.293750-20180711T235959.995274.nc	2024-10-24 18:29	470M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180712T000000.995787-20180726T235959.652776.nc	2024-10-24 18:29	458M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180727T000000.652249-20180811T235959.366336.nc	2024-10-24 18:30	483M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180812T000000.366434-20180911T235959.816132.nc	2024-10-24 18:30	466M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20180912T000000.816123-20181003T235344.191960.nc	2024-10-24 18:31	470M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181004T000000.189829-20181020T235959.994478.nc	2024-10-24 18:31	473M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181021T000000.994156-20181104T235959.444537.nc	2024-10-24 18:32	458M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181105T000000.444427-20181119T235959.430798.nc	2024-10-24 18:32	458M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181120T000000.430998-20181205T235959.812736.nc	2024-10-24 18:33	487M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181206T000000.812834-20181220T235959.736370.nc	2024-10-24 18:33	459M
deployment0005_RS01SBPS-PC01A-4A-CTDPFA103-streamed-ctdpf_optode_sample_20181221T000000.736781-20181231T235959.646708.nc	2024-10-24 18:34	336M

Zarr is intuitive...

- Code that works with 10 seconds of data can be easily adapted to 10 years

```
[1]: import s3fs
import xarray as xr

[2]: RCA_s3_bucket = "ooi-data/"
fs = s3fs.S3FileSystem(anon=True)

[3]: starttime = '2017-06-16T00:00:00'
endtime = '2024-08-25T00:00:00'

[4]: stream_name = "RS01SBPS-PC01A-05-ADCPTD102-streamed-adcp_velocity_beam"

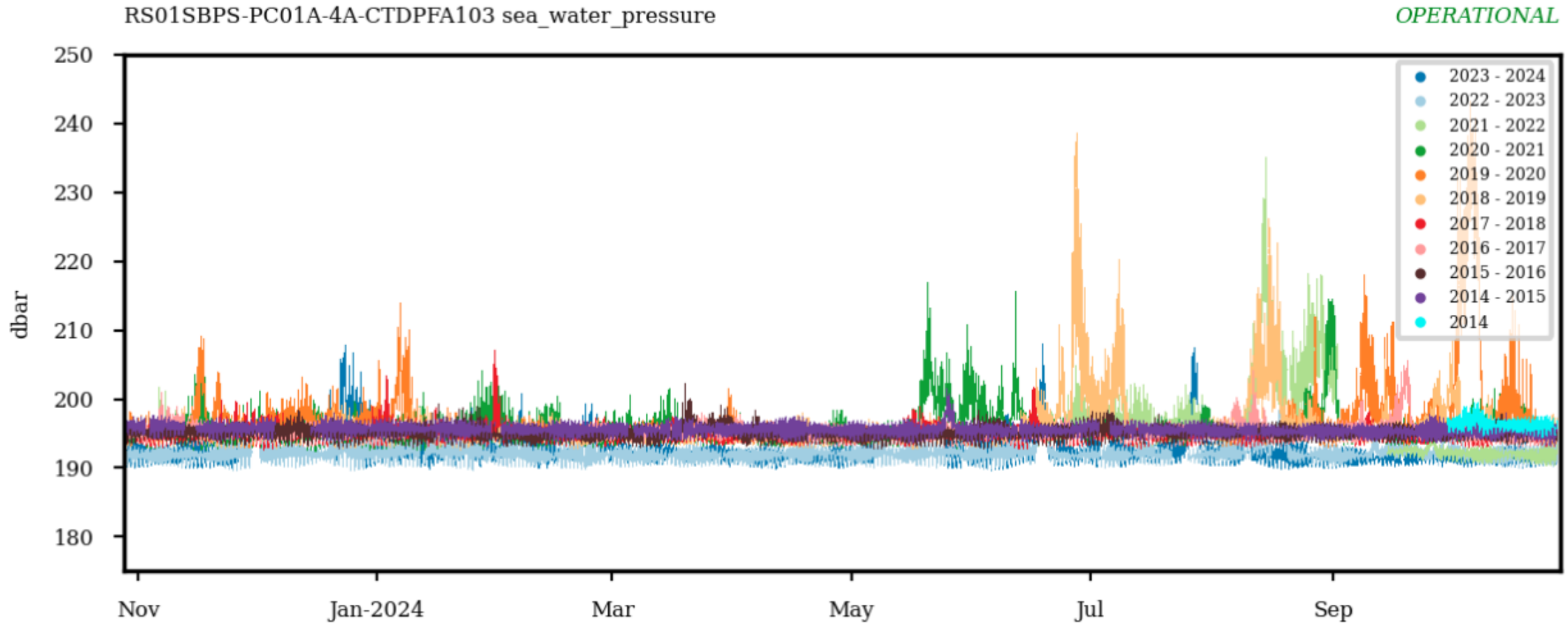
[5]: def load_data(stream_name):
    zarr_dir = RCA_s3_bucket + stream_name
    zarr_store = fs.get_mapper(zarr_dir)
    ds = xr.open_zarr(zarr_store, consolidated=True)
    return ds

[6]: ds = load_data(stream_name)

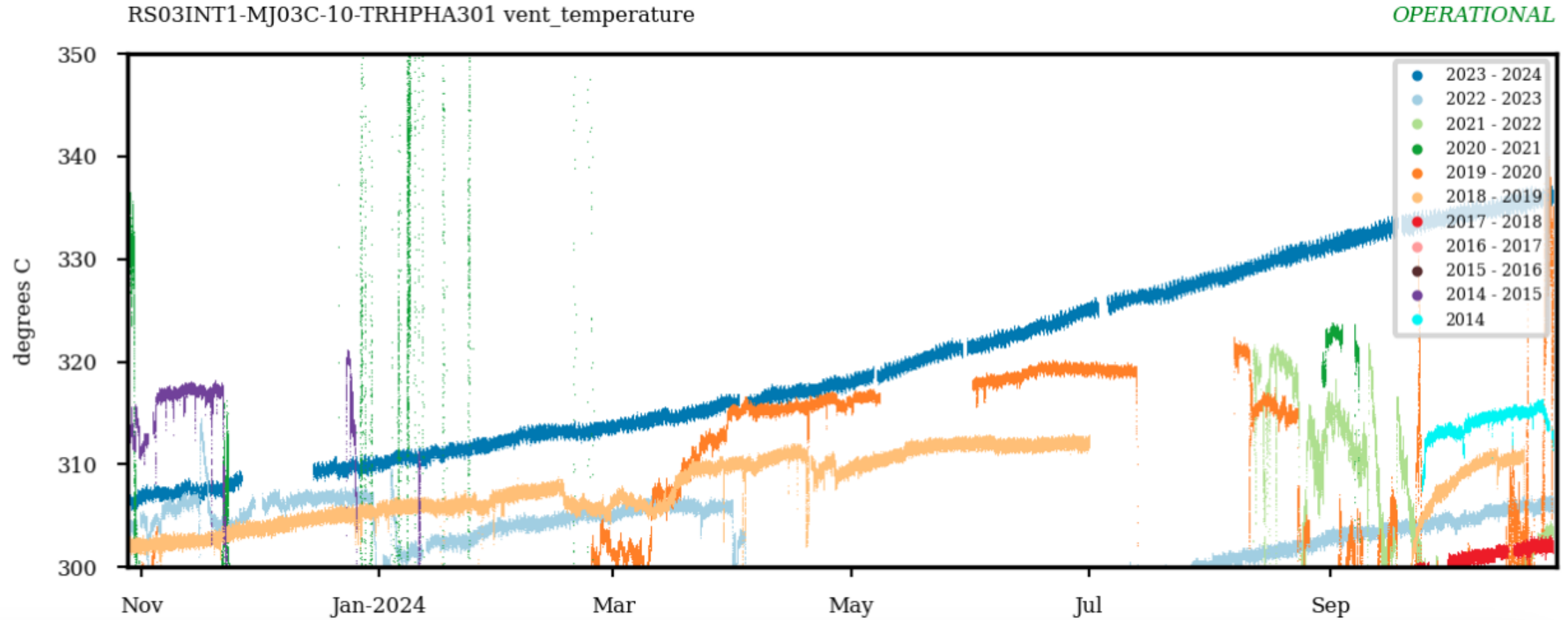
[7]: ds
xarray.Dataset
  Dimensions:                (bin: 40, time: 382404438)
  Coordinates:
    bin                       (bin)          int32  0 1 2 3 4 5 6 ... 34 35 36 37 38 39
    time                       (time)        datetime64[ns]  2014-10-02T14:45:31.139999744 .....
  Data variables:
    (47)
  Indexes: (2)
  Attributes: (62)
```

Zarr is intuitive...

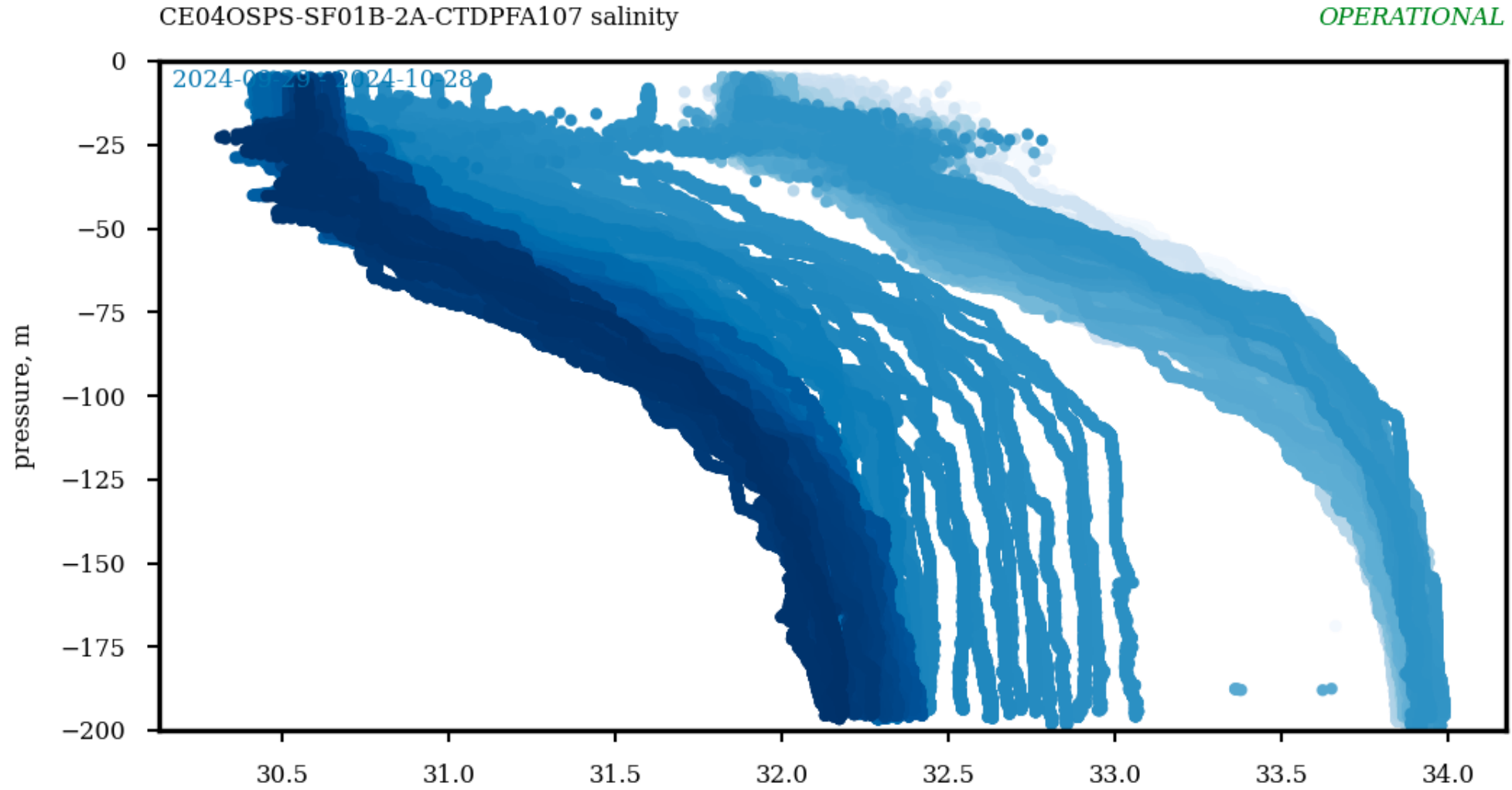
- Some example views from our dashboard:



Zarr is intuitive...



Zarr is intuitive...



5) Zarr plays nice with our scientific python stack

storage



Zarr

analysis



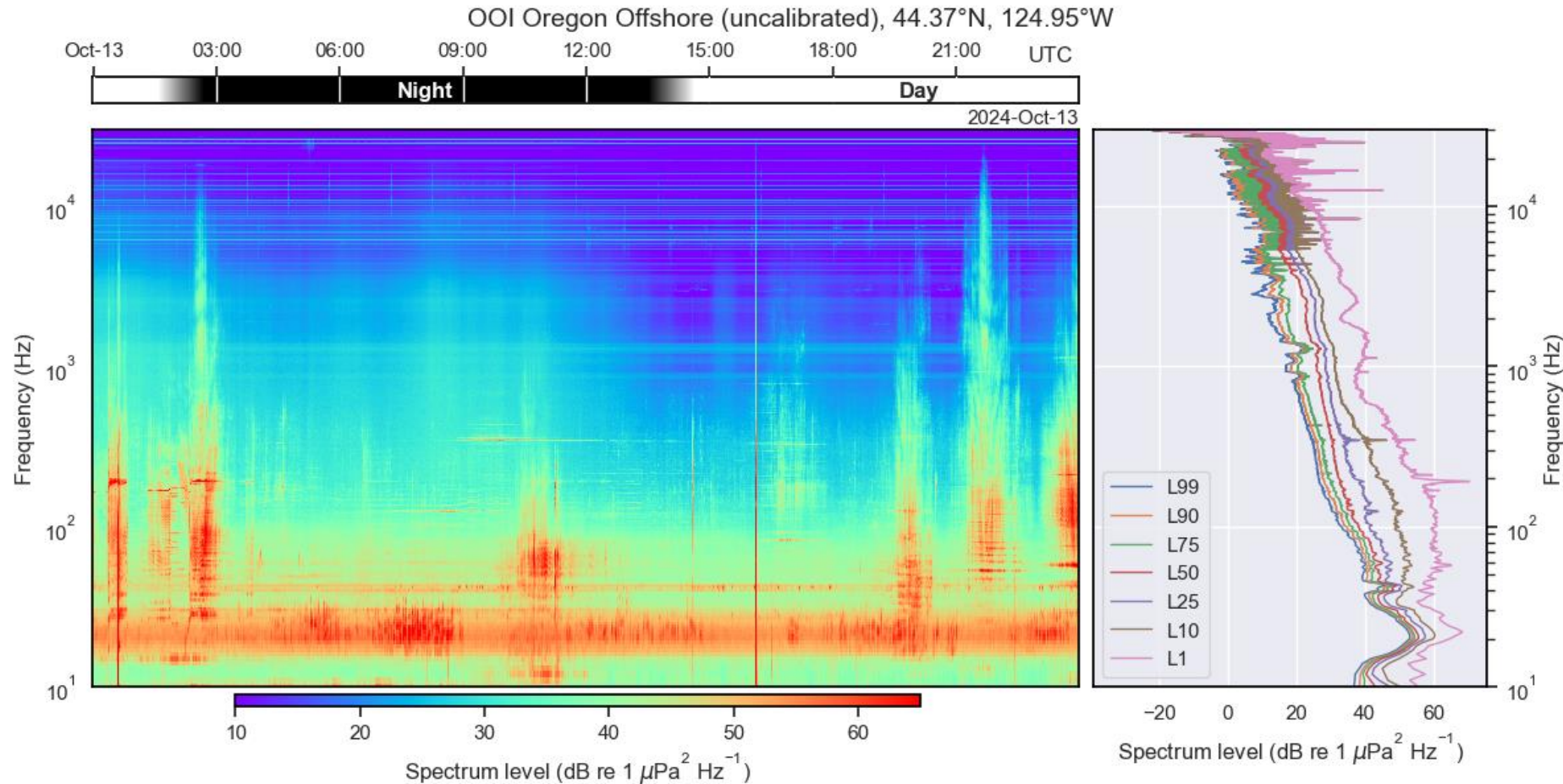
interface

matplotlib



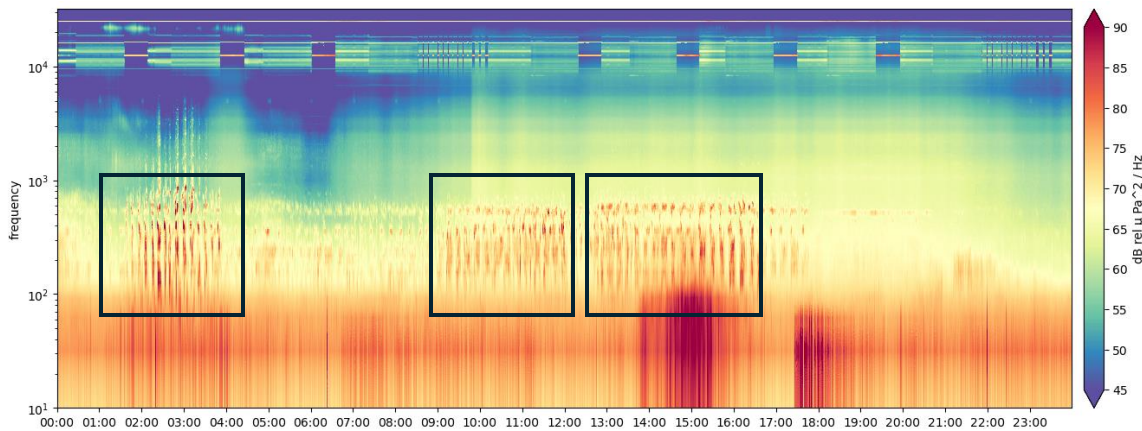
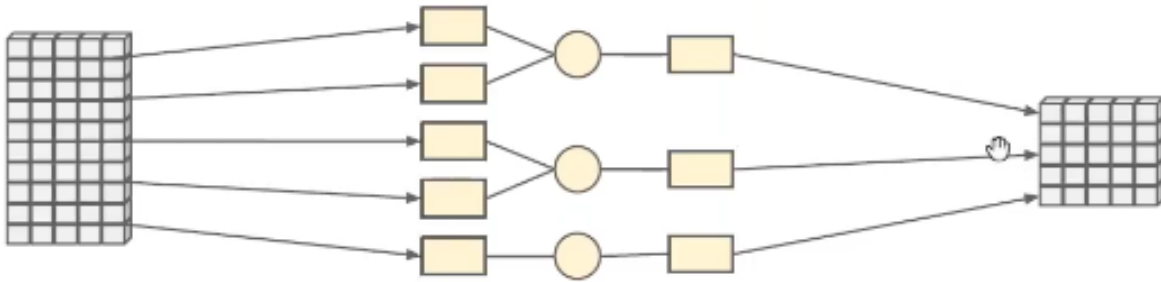
6) Zarr works well across domains and data types

- User communities are already utilizing zarr
- Consolidated spectrogram storage, Complete soundscape



7) Zarr is designed for parallelism

- Lots of room for optimization
- Automated nearest neighbor comparisons
- Machine learning ready – anomaly and event detection





OCEAN
OBSERVATORIES
INITIATIVE

Questions?

