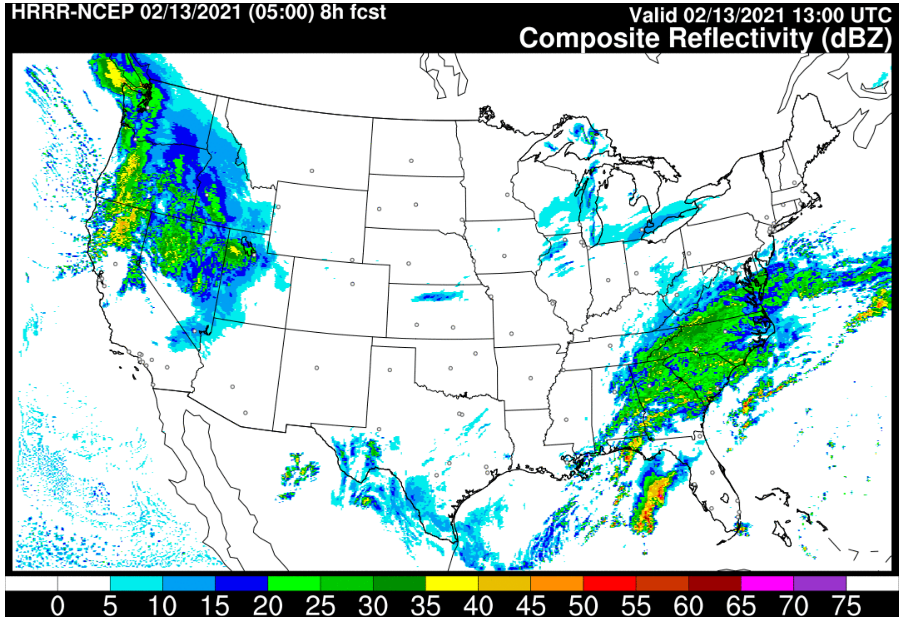
DATA ANALYTICS APPLIED TO SATELLITE-DERIVED PRECIPITATION ESTIMATES AND HIGH-RESOLUTION MODEL OUTPUT

Taylor A. Gowan

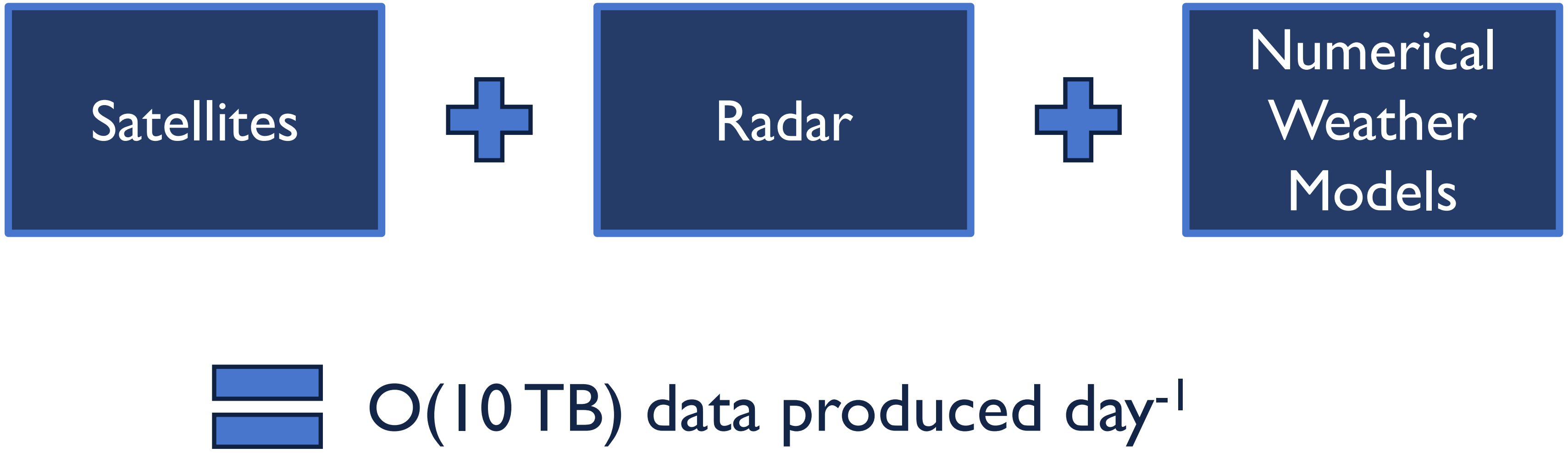
Ph.D. Dissertation Defense

Committee: John Horel (Chair), Erik Crosman, George Huffman, Jim Steenburgh, and Court Strong

ARCHIVAL AND ANALYSIS OF HIGH-RESOLUTION RAPID REFRESH MODEL OUTPUT USING ZARR FILES IN THE CLOUD


HRRR-NCEP 02/13/2021 (05:00) 8h fcst
Valid 02/13/2021 13:00 UTC
Composite Reflectivity (dBZ)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

The Big Data Problem

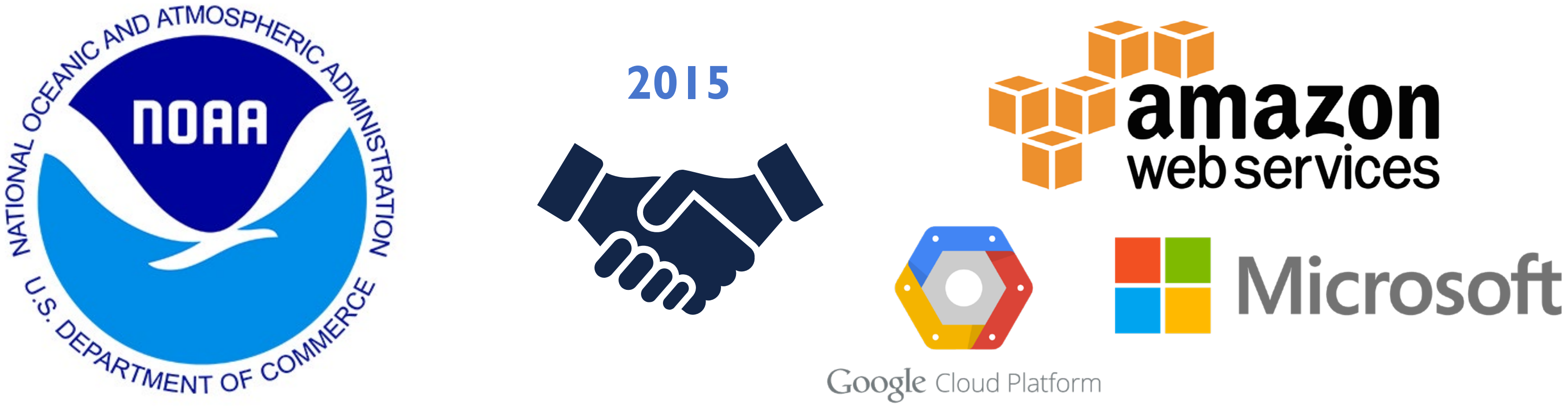


Satellites + Radar + Numerical Weather Models

= O(10 TB) data produced day^{-1}

https://www.noaa.gov/organization/information-technology/big-data-program

The Response:
NOAA Big Data Program


NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE
2015
amazon web services
Google Cloud Platform
Microsoft

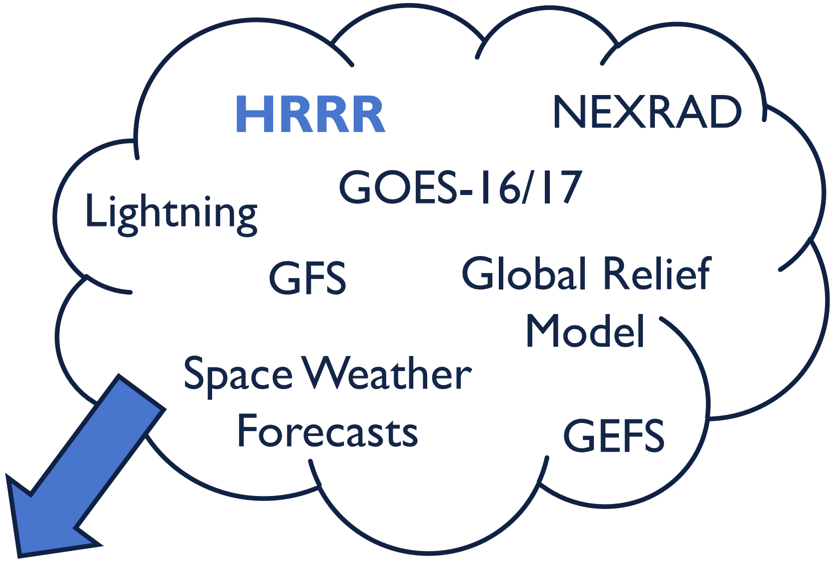
Infrastructure as a Service (IaaS) providers have the capacity to store the increasing volume of data available and provide public access and computer resources to end users

The Response:
NOAA Big Data Program

amazon web services

Microsoft

Google Cloud Platform



Most of these high-volume datasets are stored in hypercube formats e.g., GRIB2, netCDF4

WORLD
METEOROLOGICAL
ORGANIZATION

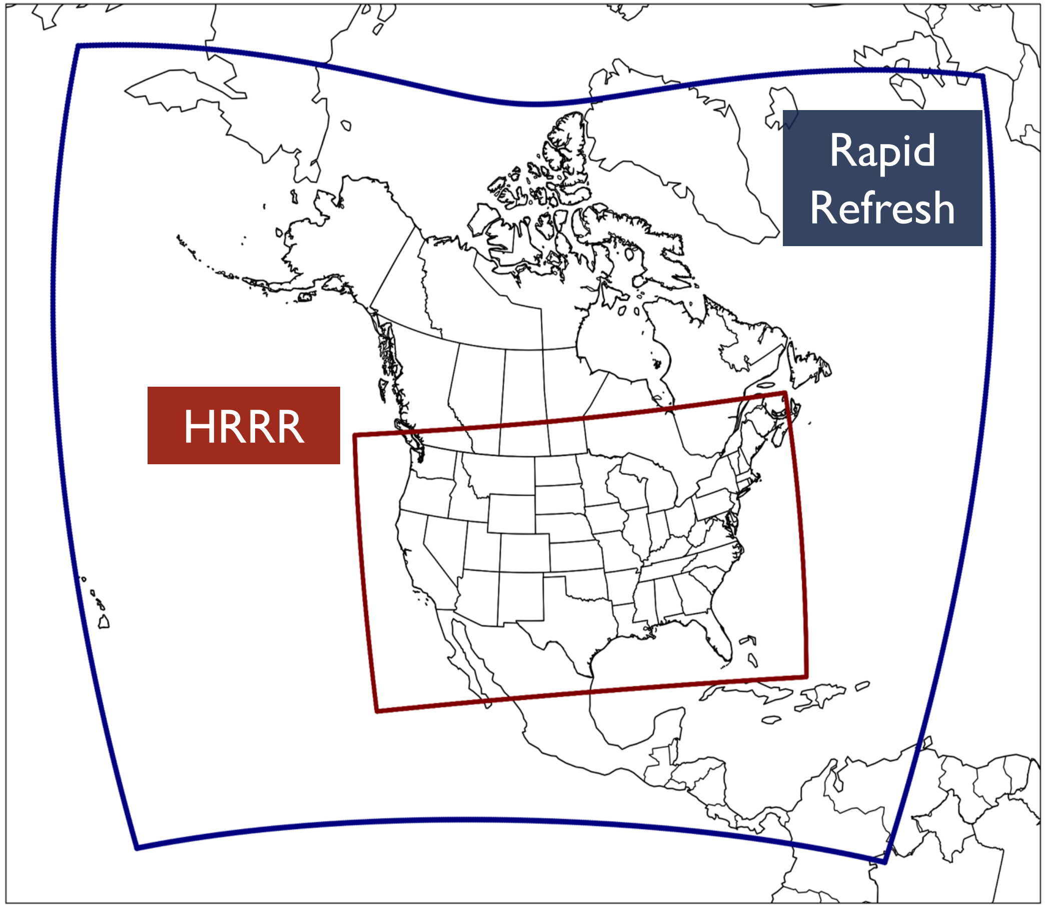
GRIB2 Format: GRIdded Binary Second Edition

- Self-describing file format used to efficiently store and transmit two-dimensional data fields (*Wang 2014*)

- Data is compressed using a similar method to JPEG image compression (*Silver and Zander 2017*)

- GRIB2 files are large when decompressed and are difficult to read efficiently, even when using compatible Python libraries

High-Resolution Rapid Refresh version 4 (HRRR)



- 3-km convection-permitting numerical weather model – 1.9 million grid spaces over CONUS domain
- Run hourly by the National Center for Environmental Prediction (NCEP)
- Output used for forecasting, research, development

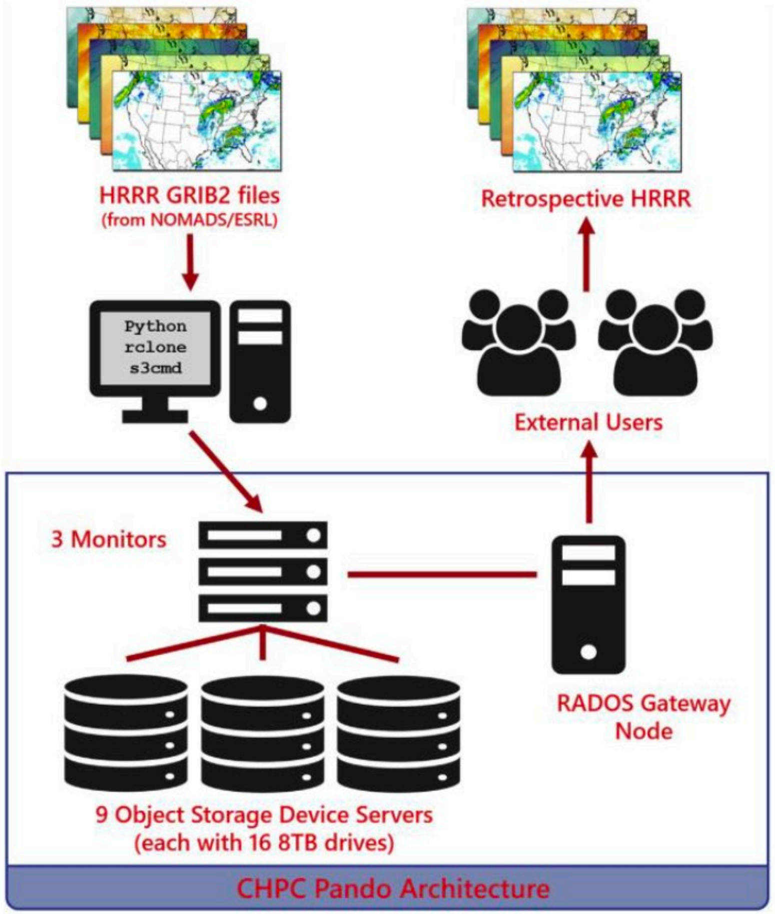
High-resolution model, high-volume output…

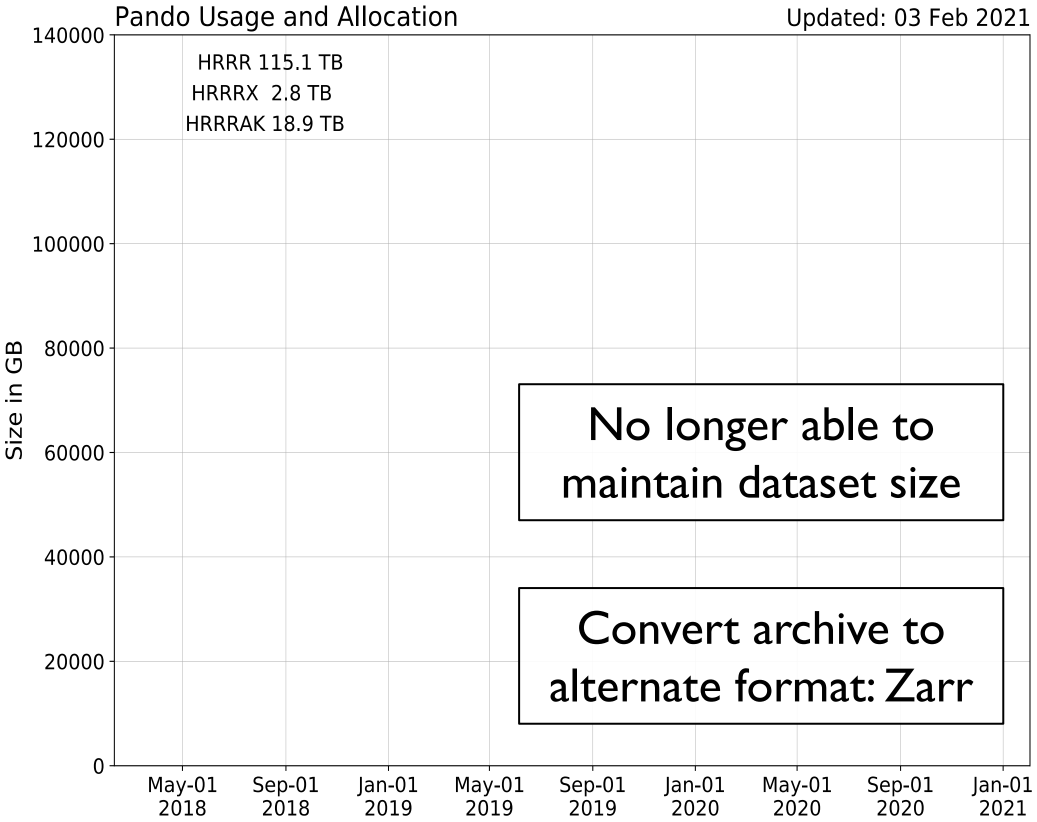
HRRR CONUS		Forecast Length for Initialization Times		Number of Output Files		
v.	First Date	0, 6, 12, 18 UTC	Other Hours	Surface	Pressure	Native
1	9/30/2014	15	15	102	659	778
2	8/23/2016	18	18	135	687	1110
3	7/12/2018	36	18	151	70	
4	12/2/2020	48	18	173	711	1136

576 files day^{-1}

Utah HRRR Archive

2016-2020 Pando was the only HRRR archive





No longer able to maintain dataset size

Convert archive to alternate format: Zarr

Blaylock et al. (2017)

An alternative model output format – Zarr

Developed by Alistair Miles (2016) and supported by the MRC Centre for Genomics and Global Health

https://zarr.readthedocs.io/

- Inspired by HDF5
- Creates N-dimensional arrays with any NumPy dtype
- Arrays can be chunked in any dimension
- Cloud-compatible
- Back-end compatibility with xarray, dask, iris

https://zarr-developers.github.io/

What exactly is a chunk?

An N-dimensional subset of a Zarr array (zarray) whose **shape**, **data type**, and **memory** specifications are based on the user-defined selections for the application

Say we have an array of size (1500,1500)

But we want our chunks to be a more manageable size of (500,500), or 9 equal chunks

Zarr Array
1500 x 1500

https://zarr-developers.github.io/

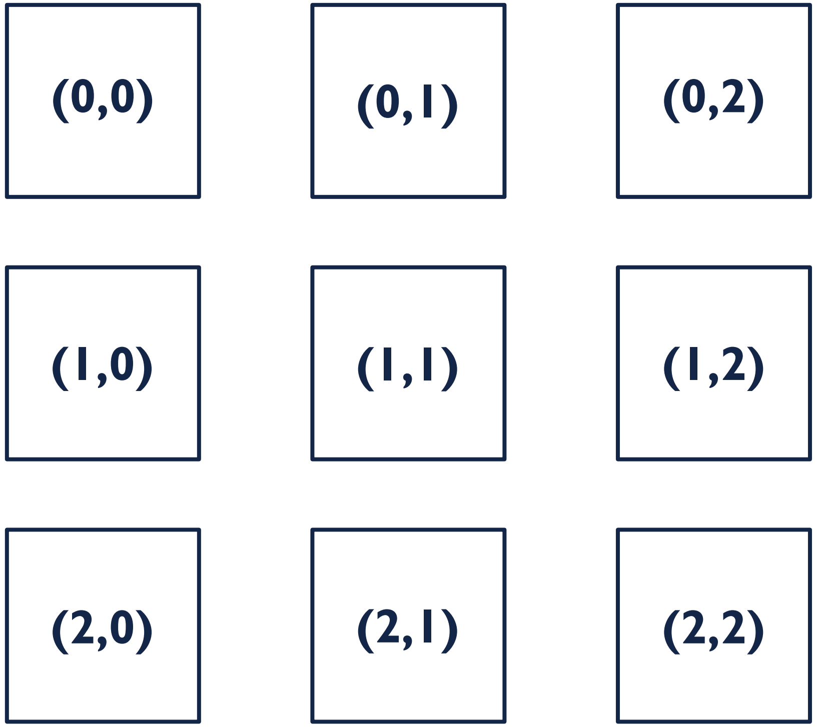
What exactly is a chunk?

A user-defined N-dimensional subset of a Zarr array (zarray) whose shape, data type, and memory specifications are predefined based on the application

Say we have an array of size (1500,1500)

But we want our chunks to be a more manageable size of (500,500), or 9 equal chunks

Each resulting chunk will be labeled with numbers corresponding to its location in the zarr array



(0,0)	(0,1)	(0,2)
(1,0)	(1,1)	(1,2)
(2,0)	(2,1)	(2,2)

https://zarr-developers.github.io/

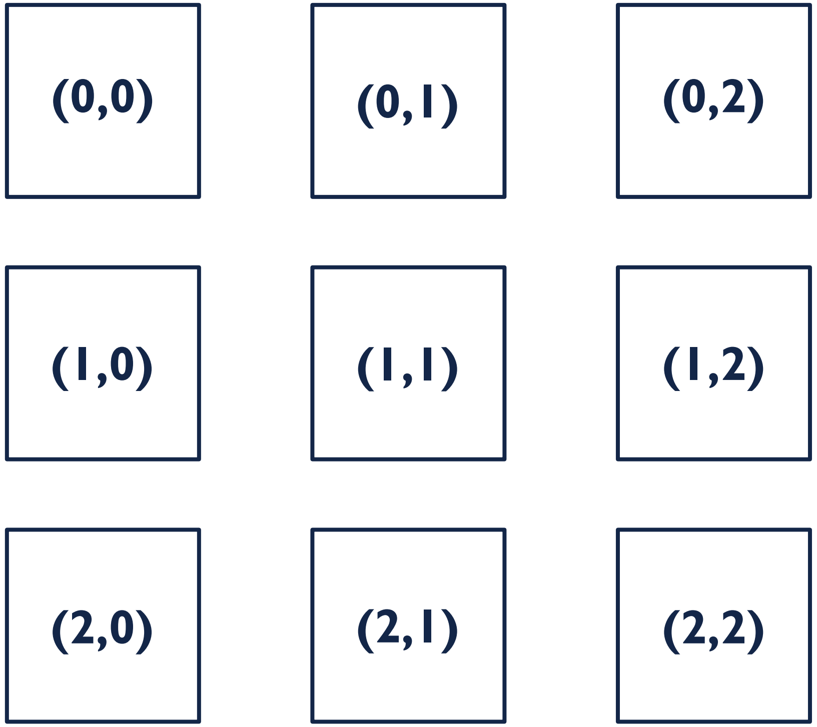
What exactly is a chunk?

A user-defined N-dimensional subset of a Zarr array (zarray) whose shape, data type, and memory specifications are predefined based on the application

Each chunk can be encoded and compressed based on the user specification

Encoding: Byte order, character code, byte length

Compression: LZMA, Blosc, LZ4, Zstandard, Zlib



(0,0)	**(0,1)**	**(0,2)**
(1,0)	**(1,1)**	**(1,2)**
(2,0)	**(2,1)**	**(2,2)**

So, does anyone use it?

CMIP
The World Climate Research Programme's
Coupled Model Intercomparison Project

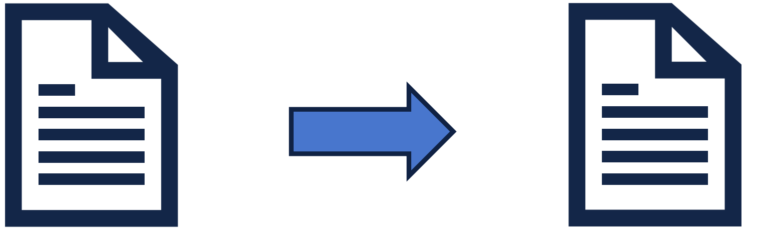
MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

PANGEO

lyft Level 5

Met Office

HRRR-Zarr Dataset



HRRR Surface GRIB2

HRRR Surface Zarr

- ✓ Keep Climate and Forecast naming conventions
- ✓ Store in Amazon Web Service Cloud

python™

NumPy

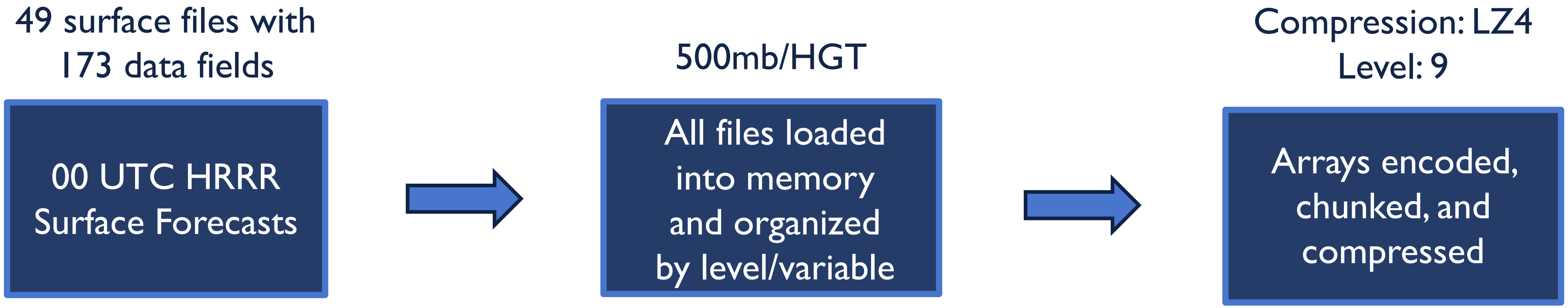
Zarr

Iris

xarray

Boto 3

HRRR-Zarr Workflow



49 surface files with 173 data fields

00 UTC HRRR Surface Forecasts

500mb/HGT

All files loaded into memory and organized by level/variable

Compression: LZ4 Level: 9

Arrays encoded, chunked, and compressed

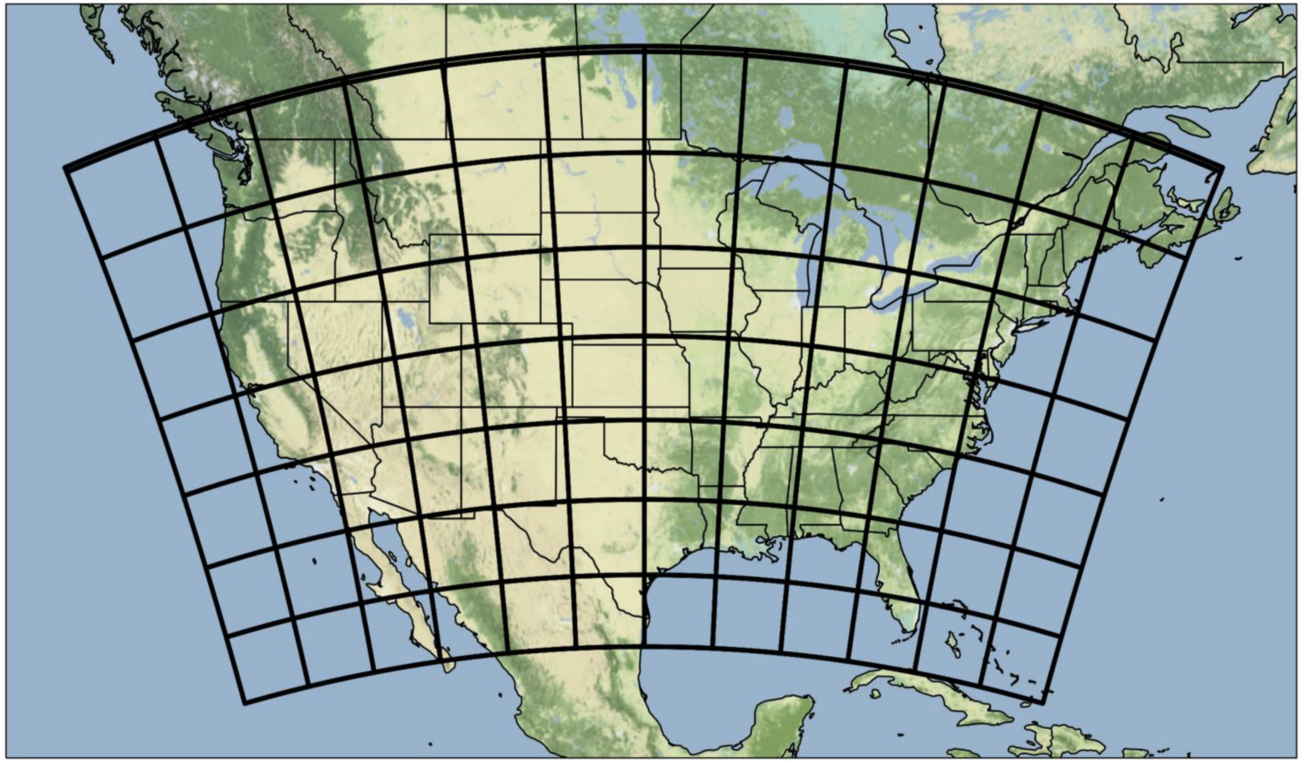
20201015_00z_anl.zarr	**F00**
20201015_00z_fcst.zarr	**F01 – F48**

Consider 20201015_00z_fcst.zarr

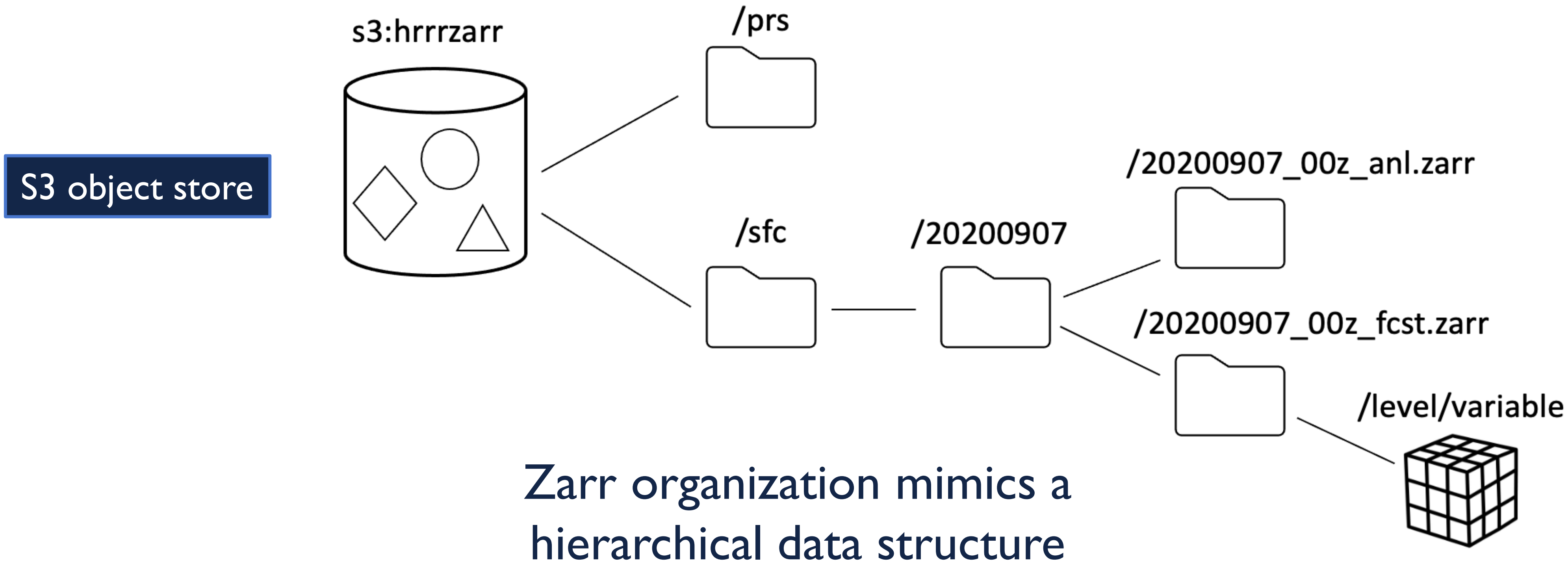
- Each data array (variable) is of size (48,1799,1059)

- We want to subset the array along the time dimension, into chunks of size (48,150,150) for easy time series construction

- Produces 96 chunks across the HRRR domain ~500 kB – 1 MB



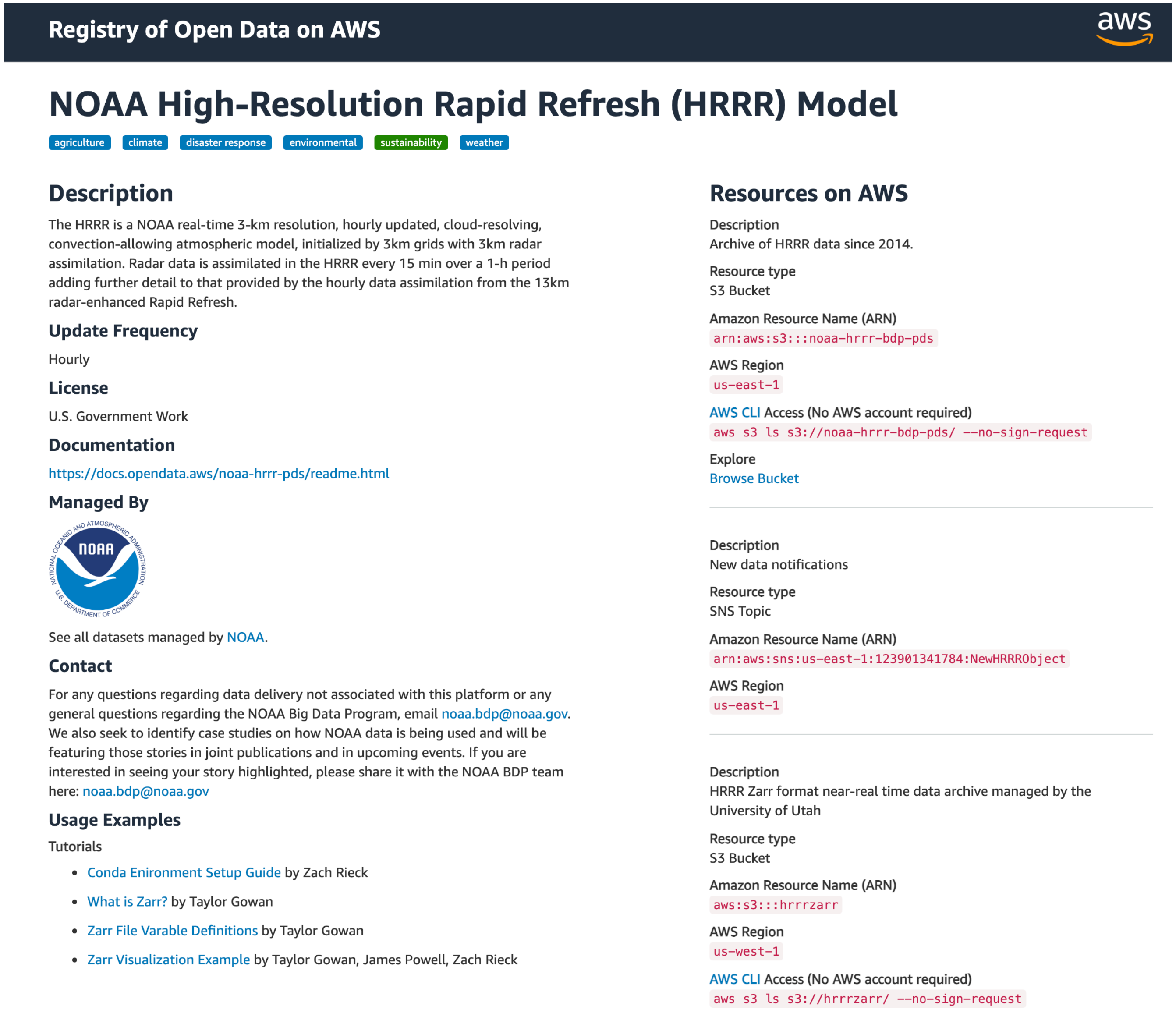
AWS Simple Storage Service (S3)



Zarr organization mimics a hierarchical data structure

Publicly available through the Registry of Open Data on AWS

https://registry.opendata.aws/noaa-hrrr-pds/



Registry of Open Data on AWS

NOAA High-Resolution Rapid Refresh (HRRR) Model

Description

The HRRR is a NOAA real-time 3-km resolution, hourly updated, cloud-resolving, convection-allowing atmospheric model, initialized by 3km grids with 3km radar assimilation. Radar data is assimilated in the HRRR every 15 min over a 1-h period adding further detail to that provided by the hourly data assimilation from the 13km radar-enhanced Rapid Refresh.

Update Frequency

Hourly

License

U.S. Government Work

Documentation

https://docs.opendata.aws/noaa-hrrr-pds/readme.html

Managed By

See all datasets managed by NOAA.

Contact

For any questions regarding data delivery not associated with this platform or any general questions regarding the NOAA Big Data Program, email noaa.bdp@noaa.gov. We also seek to identify case studies on how NOAA data is being used and will be featuring those stories in joint publications and in upcoming events. If you are interested in seeing your story highlighted, please share it with the NOAA BDP team here: noaa.bdp@noaa.gov

Usage Examples

Tutorials

- Conda Enironment Setup Guide by Zach Rieck
- What is Zarr? by Taylor Gowan
- Zarr File Varable Definitions by Taylor Gowan
- Zarr Visualization Example by Taylor Gowan, James Powell, Zach Rieck

Resources on AWS

Description
Archive of HRRR data since 2014.

Resource type
S3 Bucket

Amazon Resource Name (ARN)
`arn:aws:s3:::noaa-hrrr-bdp-pds`

AWS Region
`us-east-1`

AWS CLI Access (No AWS account required)
`aws s3 ls s3://noaa-hrrr-bdp-pds/ --no-sign-request`

Explore
Browse Bucket

Description
New data notifications

Resource type
SNS Topic

Amazon Resource Name (ARN)
`arn:aws:sns:us-east-1:123901341784:NewHRRRObject`

AWS Region
`us-east-1`

Description
HRRR Zarr format near-real time data archive managed by the University of Utah

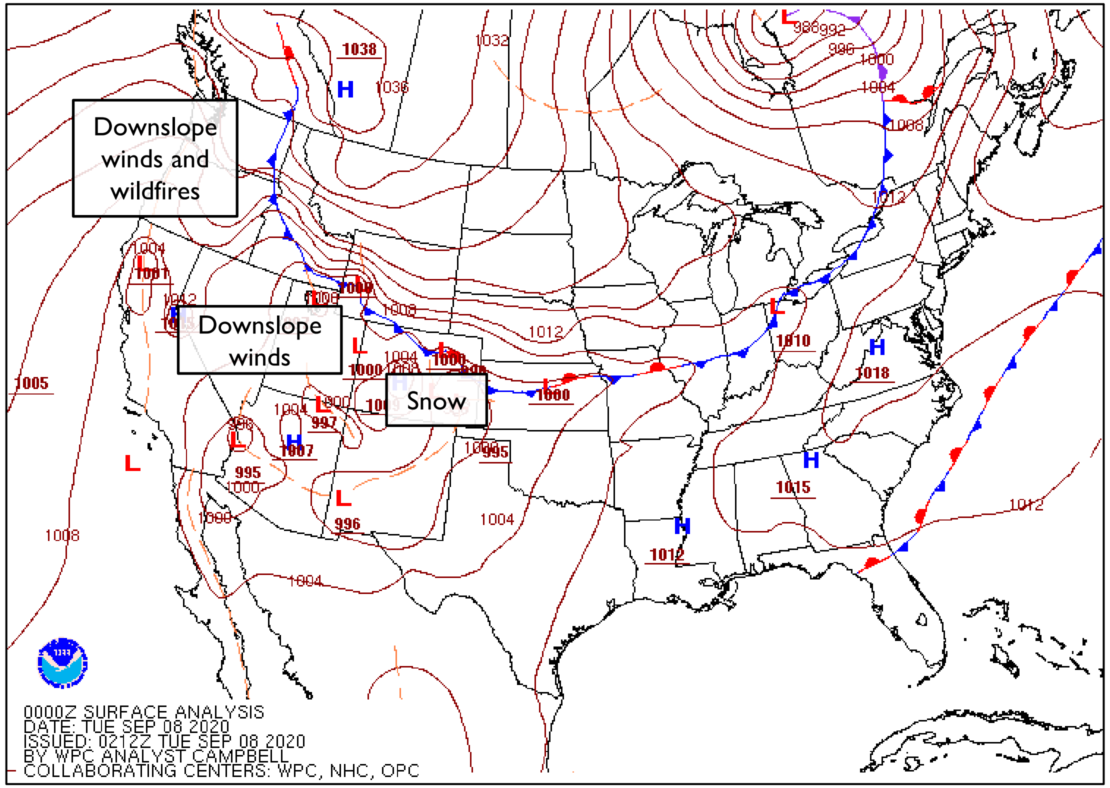
Resource type
S3 Bucket

Amazon Resource Name (ARN)
`aws:s3:::hrrrzarr`

AWS Region
`us-west-1`

AWS CLI Access (No AWS account required)
`aws s3 ls s3://hrrrzarr/ --no-sign-request`

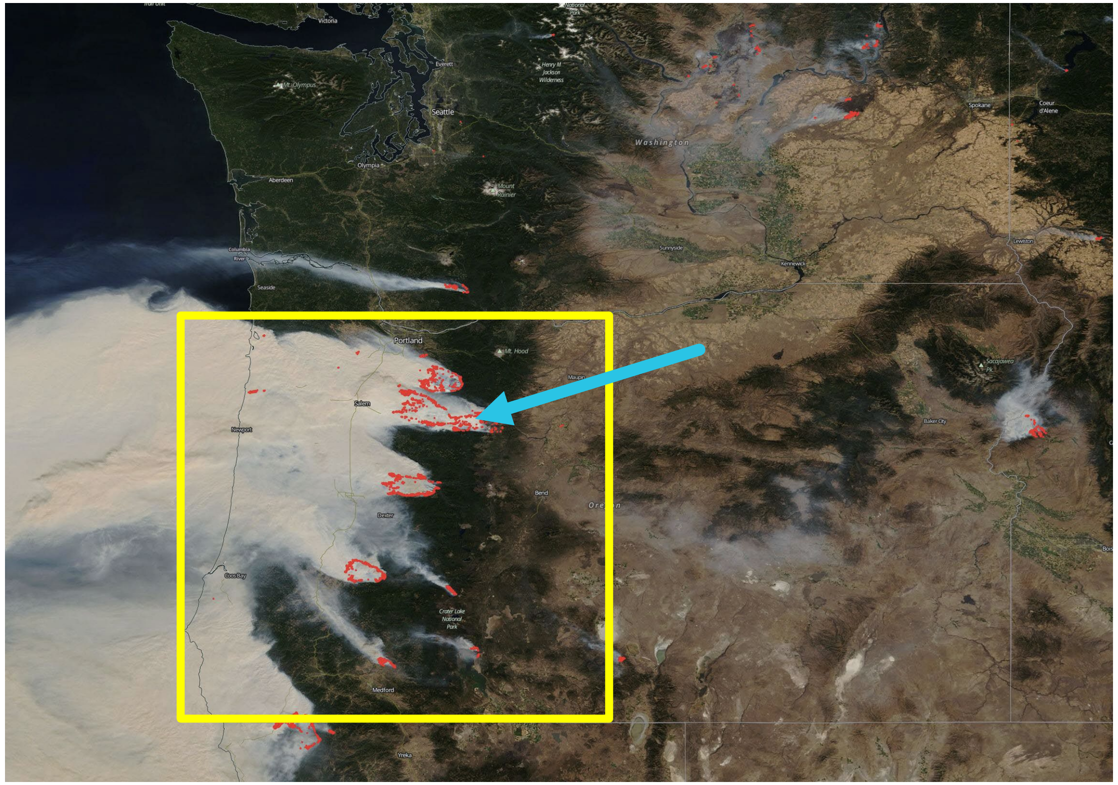
Data Application: Labor Day Weather Event
7-9 September 2020



Downslope winds and wildfires

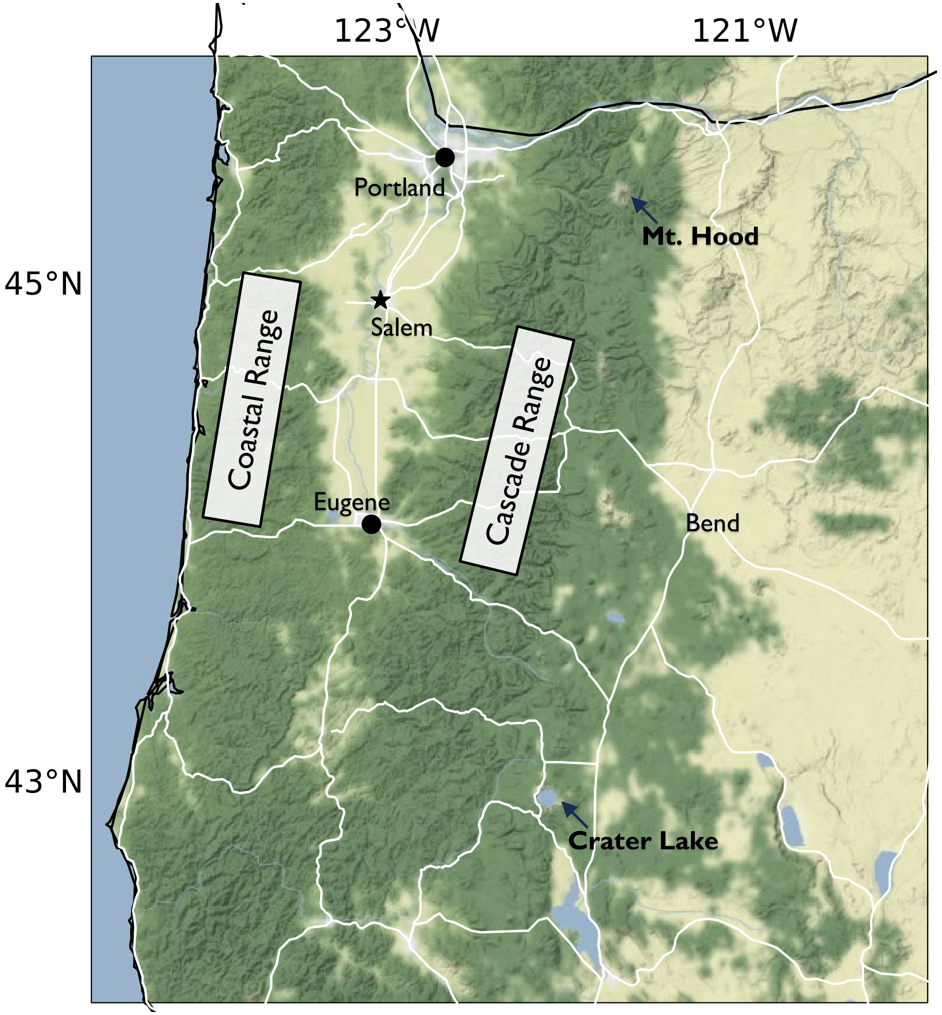
Downslope winds

Snow



https://worldview.earthdata.nasa.gov/

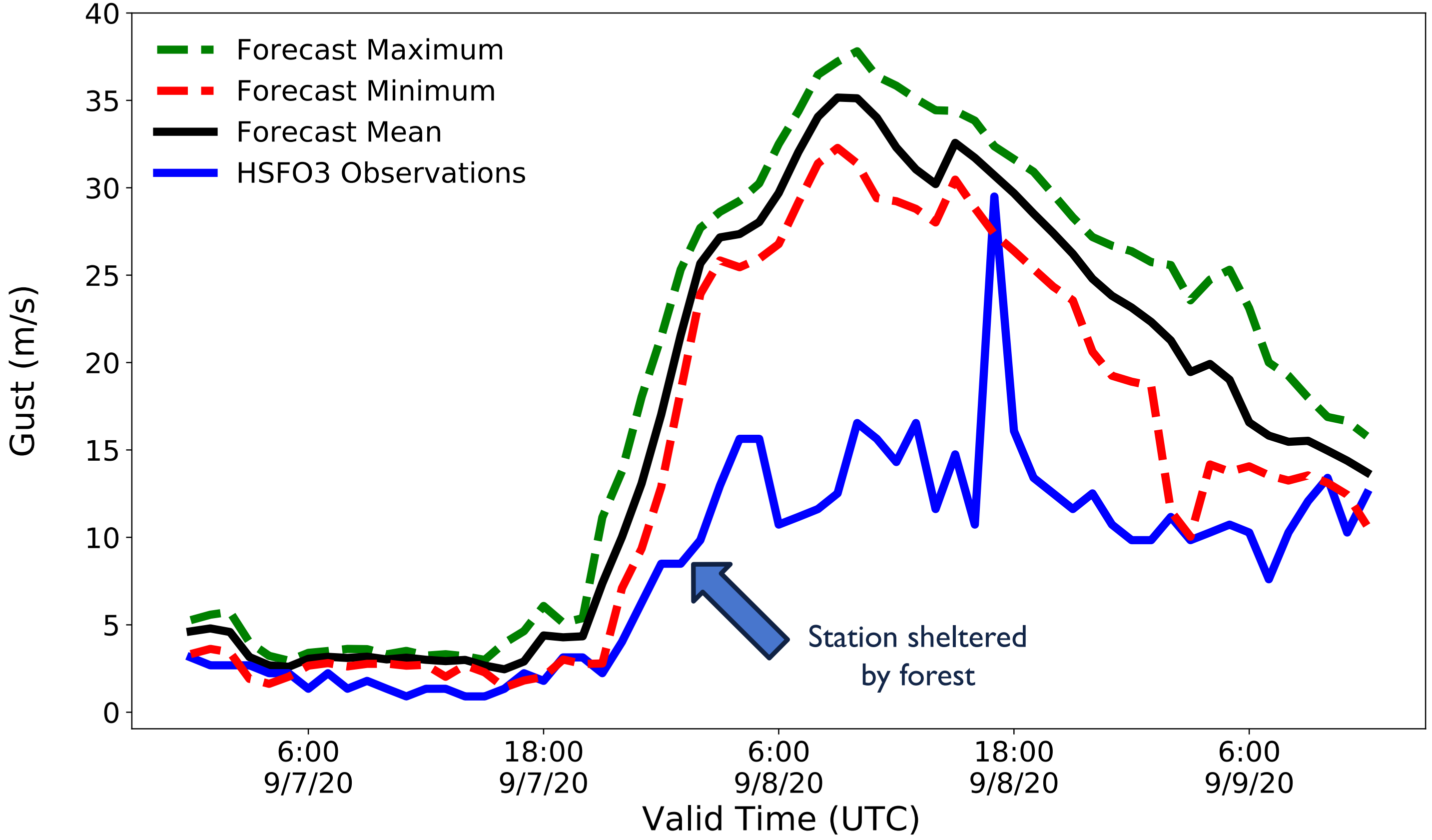
HRRR-Zarr Use Cases



Using HRRR-Zarr data for examples of high-throughput applications
(e.g., operational forecasting, machine learning)

- Time series
  - Forecast spread for a point
- Spatial plots
  - Time-lagged Ensemble (Probabilities)
  - Empirical Cumulative Distributions and exceedance values

Forecast Spread near Beechie Creek Fire



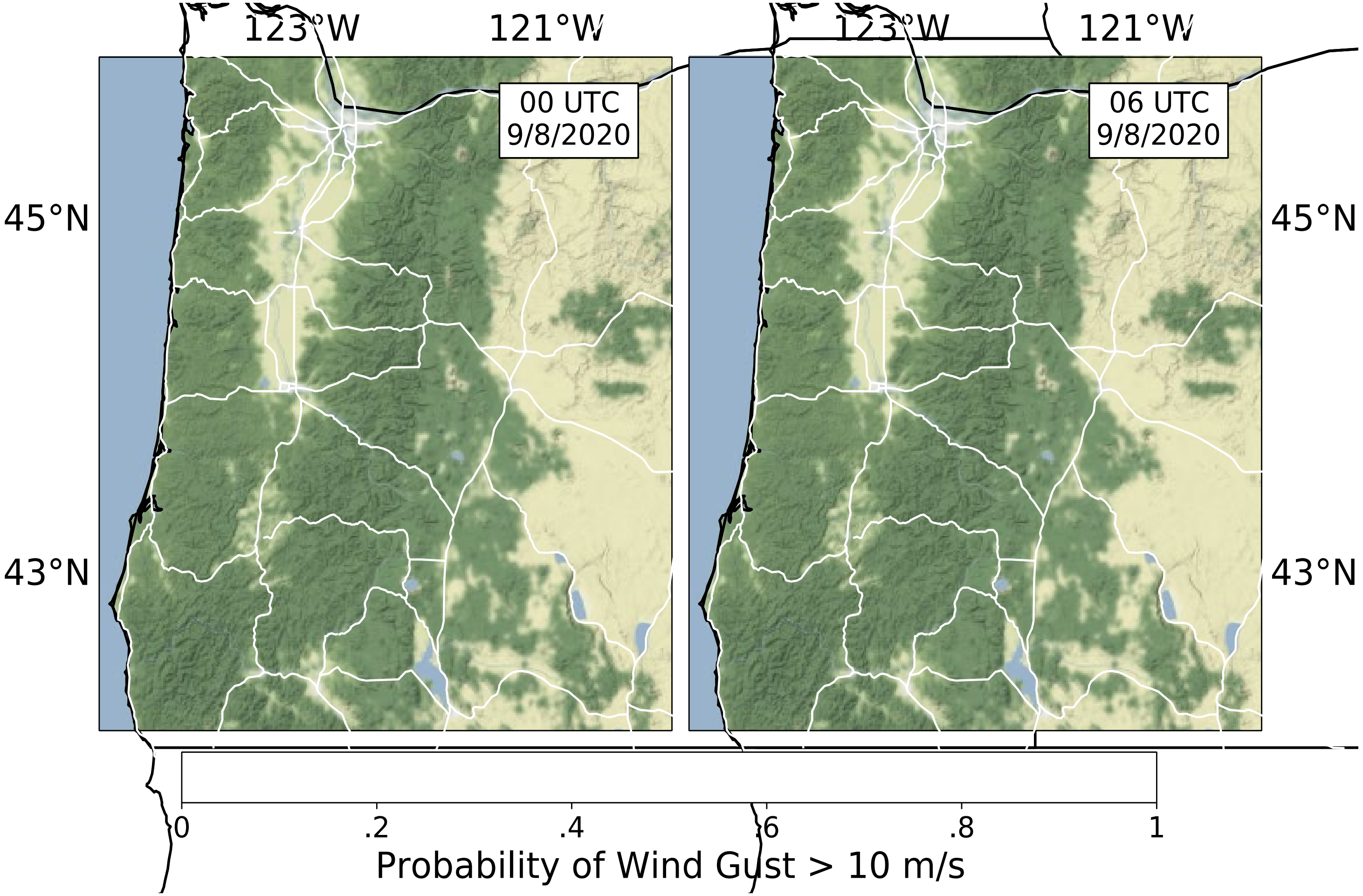
Station sheltered by forest

21 forecasts valid for each time

- 96 Zarr chunks (500 kB)
- 1260 GRIB2 files (150 MB)
- 3.3 seconds on CHPC's Kingspeak34
  No multiprocessing

Probabilistic Guidance

$$\frac{Total\ Forecasts}{Forecasts\ > 10\ m\ s^{-1}}$$

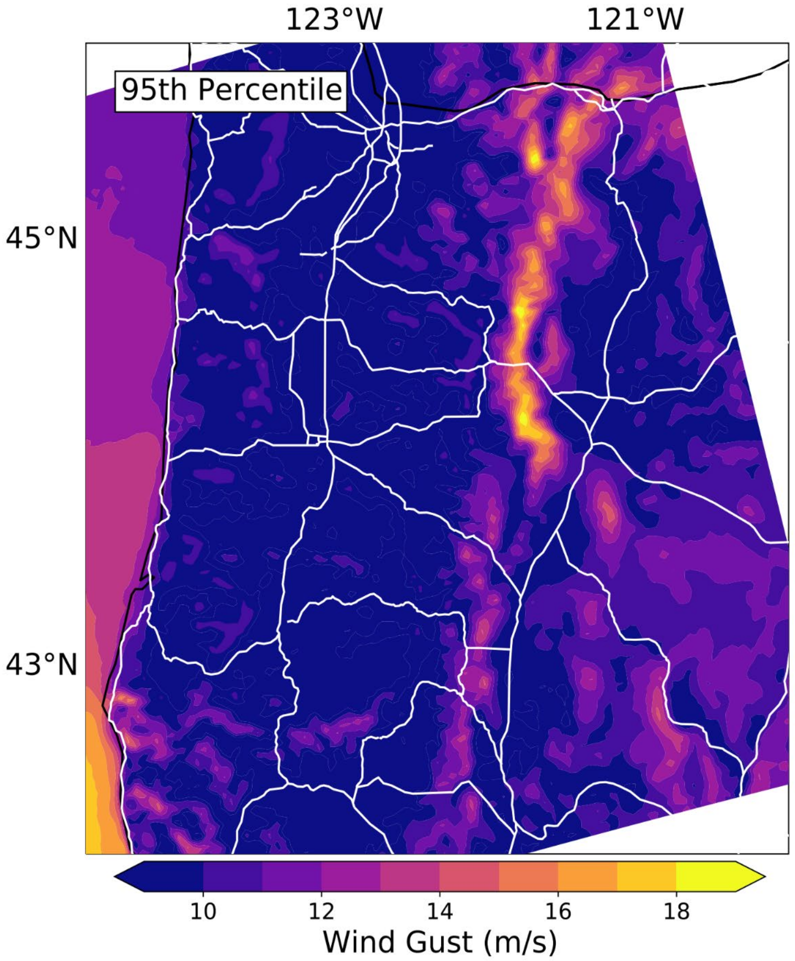


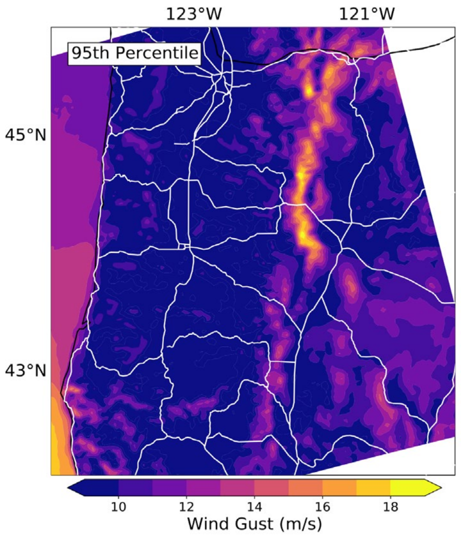
Empirical Cumulative Distributions

Goal is to compare data to a "climatology" distribution to identify outlier or extreme events

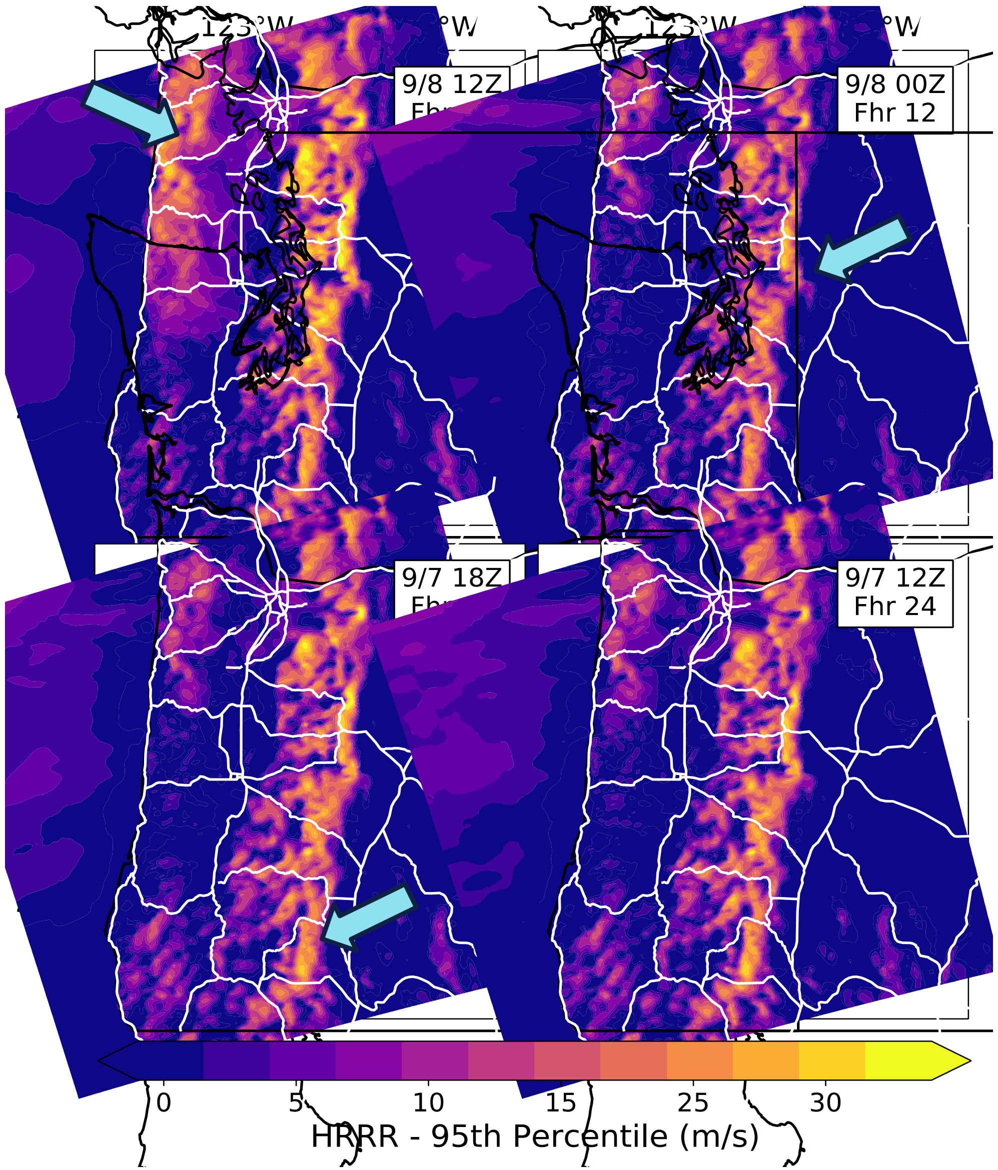
Blaylock et al. (2018) used the Open Science Grid to calculate HRRR distributions

Use all HRRR wind gust analyses (F00) from the month of September 2016-2019 to create a distribution for each grid point in the chunk





123°W
121°W
95th Percentile
45°N
43°N
10
12
14
16
18
Wind Gust (m/s)



9/8 12Z
Fhr
9/8 00Z
Fhr 12
9/7 18Z
Fhr
9/7 12Z
Fhr 24
0
5
10
15
25
30
HRRR - 95th Percentile (m/s)

Problem: Efficiently accessing and processing high-volume model output for machine learning and forecasting applications

Proposed Solution: Convert the High-Resolution Rapid Refresh (HRRR) GRIB2 model output archive to Zarr format

- We converted the HRRR archive to an alternative format, Zarr, for its compatibility with cloud environments and its flexibility
- Each HRRR model run forecasts are condensed into two Zarr files (a forecast and analysis) which contain all data fields and are named using CF conventions
- Each Zarr array is subset into 96 chunks for a more efficient and customizable user experience

Problem: Efficiently accessing and processing high-volume model output for machine learning and forecasting applications

Proposed Solution: Convert the High-Resolution Rapid Refresh (HRRR) GRIB2 model output archive to Zarr format

- **Caveat**: Zarr is optimal for many applications, but GRIB2 files are best for running model simulations and other tasks that require full-domain grids
- The HRRR-Zarr dataset is now stored on AWS, thanks to the Amazon Sustainability Initiative and Open Data Program