NCAR Datasets Published in the Cloud

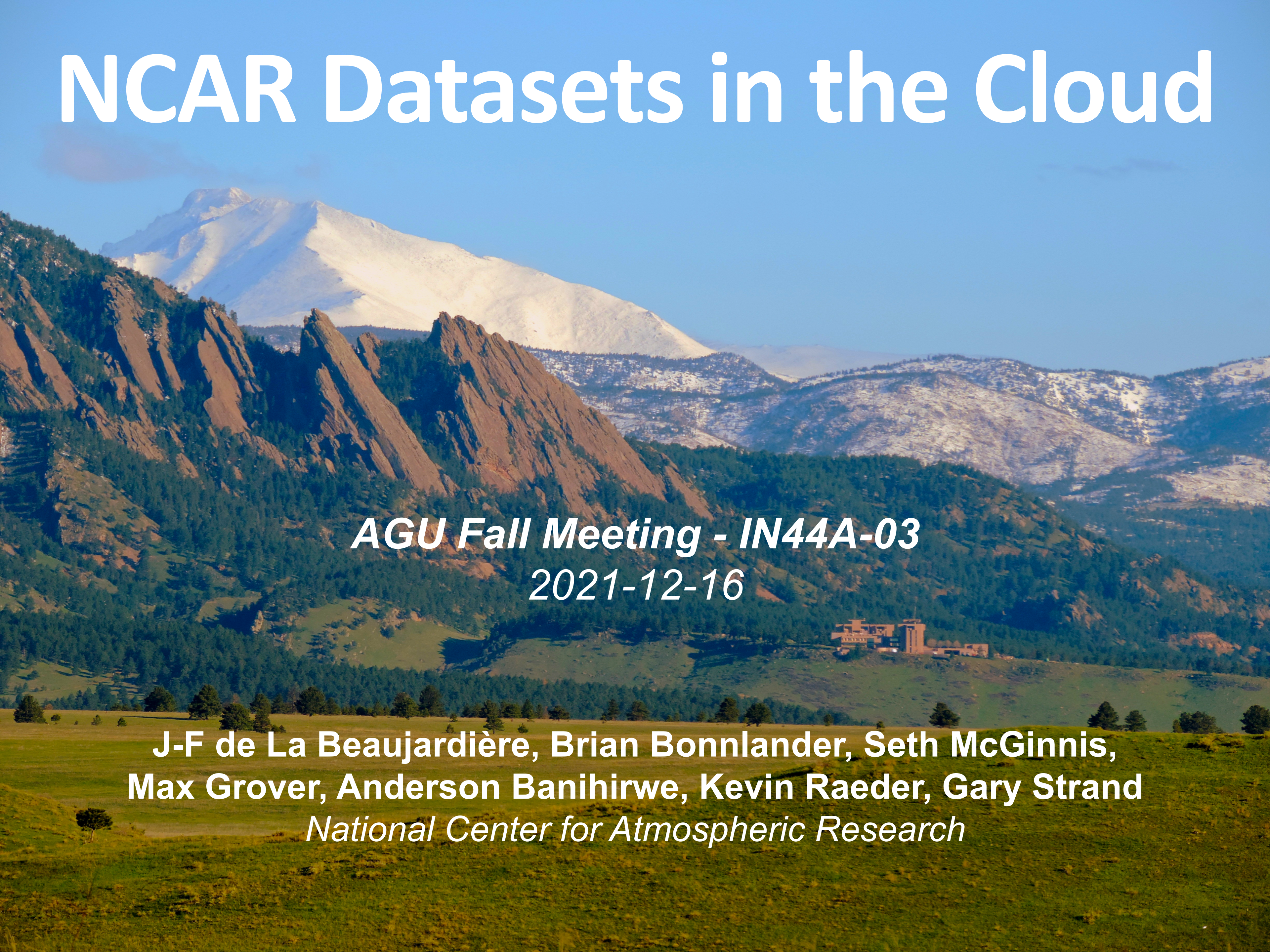
Jeff de La Beaujardière[1], Brian Bonnlander[1], Seth McGinnis[1], Maxwell Grover[1], Anderson Banihirwe[1], Kevin Raeder[1], and Gary Strand[1]

[1]National Center for Atmospheric Research

November 22, 2022

Abstract

The US National Center for Atmospheric Research (NCAR) has published several large datasets in the Amazon Web Services (AWS) cloud, thanks to support from the NCAR "Science at Scale" project, the AWS Open Data Sponsorship program, and the Amazon Sustainability Data Initiative. In each case we selected a subset comprising the most useful variables from the original data, and converted that subset from NetCDF to Zarr before publication. The Zarr format supports the same data model as netCDF and is well suited to object storage and distributed computing in the cloud using the Pangeo libraries in Python. Each dataset has an accompanying Intake-ESM catalog to facilitate data discovery and reading via Xarray, and each also has a sample Jupyter Notebook to illustrate how to access and analyze the data. Egress for these data are free, but users are encouraged to bring their compute to the data. The datasets currently published are: Community Earth System Model Large Ensemble (CESM LENS): https://doi.org/10.26024/wt24-5j82 North American Coordinated Regional Downscaling Experiment (NA-CORDEX): https://doi.org/10.26024/9xkm-fp8 CESM version 2 Large Ensemble (CESM2-LE): https://doi.org/10.26024/y48t-q717 Data Assimilation Research Testbed (DART) Reanalysis: https://doi.org/10.26024/sprq-2d04 This paper will provide information about the datasets and summarize lessons learned from the data conversion and publication.



NCAR Datasets in the Cloud

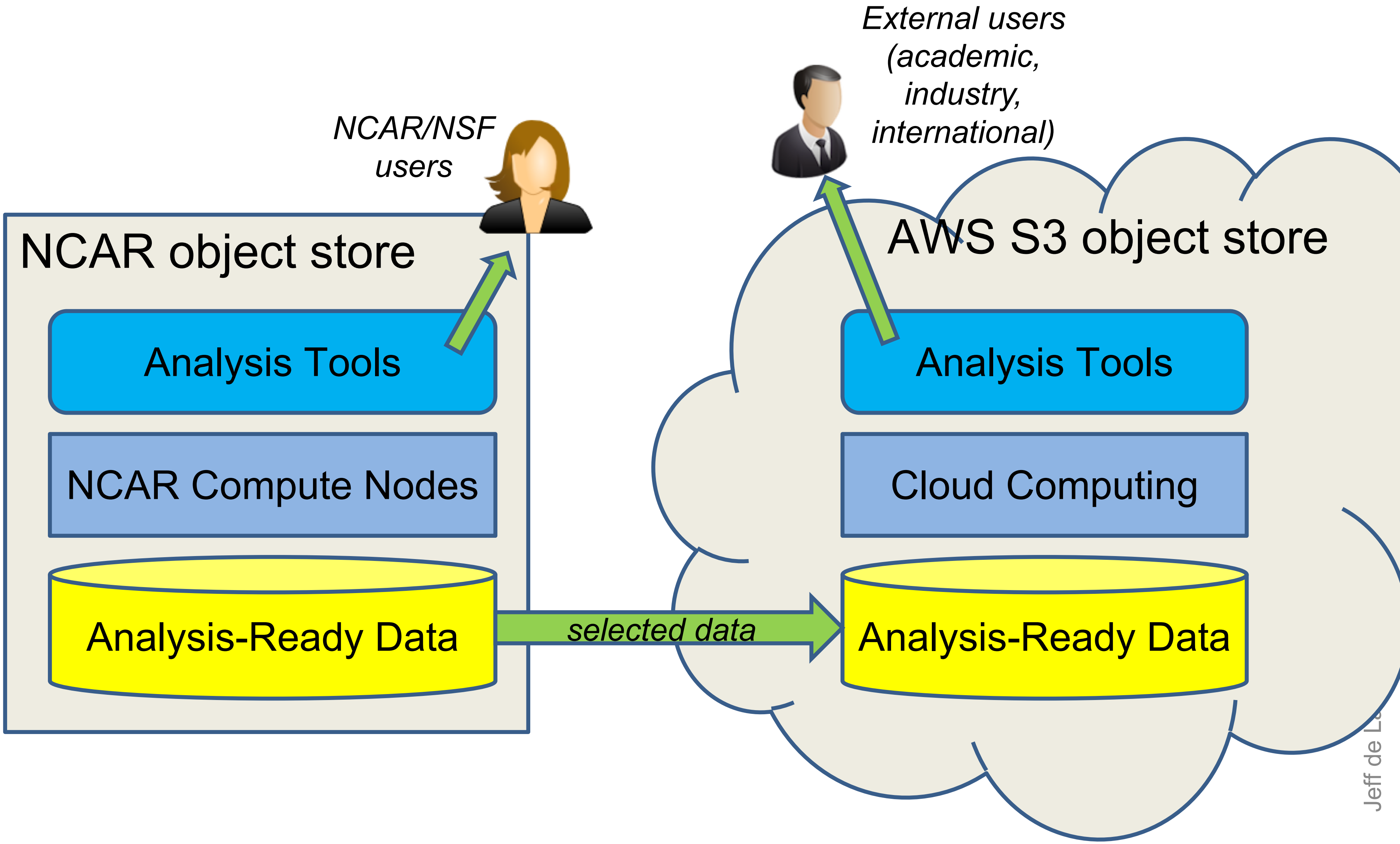
AGU Fall Meeting - IN44A-03

2021-12-16

J-F de La Beaujardière, Brian Bonnlander, Seth McGinnis, Max Grover, Anderson Banihirwe, Kevin Raeder, Gary Strand

National Center for Atmospheric Research

NCAR "Science at Scale" Project: High-Level Concept



NCAR/NSF users

External users (academic, industry, international)

NCAR object store

Analysis Tools

NCAR Compute Nodes

Analysis-Ready Data

selected data

AWS S3 object store

Analysis Tools

Cloud Computing

Analysis-Ready Data

What We Did

- **Converted to Zarr** major subsets of 4 large datasets
  - Binary format optimized for fast parallel reads using Python Xarray
  - Multiple netCDF files combined into single Zarr stores
  - Data are sliced into separate small (100 MB) objects along multiple dimensions: space, time, ensemble member
- **Published data on AWS S3** (us-west-2)
  - Free storage & egress thanks to AWS Open Data Sponsorship Program and Amazon Sustainability Data Initiative
  - Data also on NCAR on-prem Object Store
- **Created Intake-ESM Catalogs** as machine-readable inventory for each dataset
  - Simplifies ingest be Python xarray
  - Mapped unique variable names to CF Standard Names
  - Added info about spatial & temporal coverage
- **Wrote Jupyter Notebooks** showing sample analyses

Published Datasets

- **CESM LENS** - Community Earth System Model Large Ensemble
  - https://doi.org/10.26024/wt24-5j82

- **CESM2 LE** - CESM version 2 Large Ensemble
  - https://doi.org/10.26024/y48t-q717

- **NA-CORDEX** - N. Am. Coordinated Regional Downscaling Experiment
  - https://doi.org/10.26024/9xkm-fp8

- **DART CAM6 Reanalysis** - Data Assimilation Research Testbed Community Atmosphere Model v6 (CAM6) Reanalysis
  - https://doi.org/10.26024/sprq-2d04

Jeff de La Beaujardiere <jeffdlb@ucar.edu>

Data Characteristics

Dataset	Size	Years	Coverage	Ensemble
CESM LE	83 TB	1920-2100	Global 1°	40 members
CESM2 LE	267 TB	1850-2100	Global 1°	100 members
NA-CORDEX	13 TB	1950-2100	N. Am. ¼ or ½°	50 experiments
DART CAM6	2.5 TB	2011-2019	Global 1°	80 members

AWS Registry of Open Data listing:
https://registry.opendata.aws/?search=managedBy:national%20center%20for%20atmospheric%20research

Future Plans

- Add additional variables on request
  - Notably: Plant Functional Type data from DART CAM6
- Notebooks for dataset inter-comparison
- Simple on-demand visualizations using serverless computing

For more info, questions, usage problems, requests for additional data: contact cisl-aws-lens@ucar.edu

J-F de La Beaujardière, PhD
Director, NCAR/CISL Information Systems Division
jeffdlb@ucar.edu
https://orcid.org/0000-0002-1001-9210

Jeff de La Beaujardiere <jeffdlb@ucar.edu>