Integration of OSDF with NCAR's data infrastructure: Interim Project Report

Authors: Douglas Schuster, Harsha Hampapura
Oct 1, 2025

Introduction

NSF NCAR's laboratories and programs conduct research across a broad range of Earth System Science topics, generating large volumes of data. Historically, these datasets have been stored in different repositories, with only a subset discoverable through NCAR's data catalog. The current system supports two main groups of users:

a) Remote users who use the legacy "download, clean and analyze" model. This model requires the data to be downloaded to a local storage system and reorganized/cleaned before any analysis can take place; and
b) Users who have access to NCAR's HPC resources. We will refer to them as on-prem users. These users have direct access to a subset of these datasets but do not have a systematic way to search for data short of asking their colleagues for a directory location.

However, this practice has created significant barriers in data search, discovery, and access to NCAR's datasets and does not fully conform to the FAIR principles [Wilkinson et al] that data repositories should satisfy. In particular, the experience of discovering and accessing data can vary widely based on the data repository. In short, we are not realizing the full potential of **NCAR's wealth of datasets**.

An integrated research Data Commons, called Geoscience Data Exchange in NCAR's context, can solve these problems and transform access to NSF NCAR research data. In particular, such a Data Commons will:

● Make data accessible to everyone—from graduate students just starting out to experienced scientists. A consistent user experience across platforms reduces the learning curve and allows researchers to focus more on discovery than on how to find or access data.

● Second, this infrastructure will be deeply integrated. That means bringing together tools for data exploration, analysis, assimilation, and modern workflows like AI and machine learning—all within a shared ecosystem. This integration allows users to move seamlessly from discovery to analysis, and develop and train AI models more efficiently.

● Finally, we see this as a community-first effort. This infrastructure positions NSF NCAR to both support and help lead broader community CI initiatives. It aligns us with the open science movement and strengthens our ability to collaborate across institutions and disciplines.

The aim of building such a data commons aligns perfectly with this NDC (National Discovery Cloud) funded project which aims to `Integrate NCAR's data infrastructure with the Open Science Data Federation (OSDF)'. The OSDF is a `federated platform for delivering datasets from repositories to compute in an effective, scalable manner.' The OSDF infrastructure, which involves a distributed network of data origins and caches, is ideally suited to support performant distributed and democratic access to NCAR's data, while broadening the reach of these datasets and making them interoperable with other geoscience and interdisciplinary datasets.

FY24

Hardware installation and initial workflows

NSF NCAR acquired and deployed hardware to host on-premises OSDF origin and cache servers. We then worked with PNRP and PATh technical staff to install, configure and operate the required OSDF software containers on our servers in order to integrate those systems with the broader OSDF ecosystem. As a result of this work, all data published in NSF NCAR's Geoscience Data Exchange (GDEX, https://gdex.ucar.edu), which includes 17 PB data spread across 1600+ datasets, are now accessible through the OSDF, including popular retrospective atmospheric re-analyses and community Earth System Science model projections. Additionally, researchers that use NSF NCAR's compute services now have performant access to datasets provided on all of OSDF's origin servers through our on-prem OSDF cache server.

In order to support performant data access in reference computational workflows, we developed and archived several intake-ESM catalogs to support analysis-ready and cloud-optimized data access and identified key datasets to develop initial workflows.

FY25

In FY25, we focused on ironing out data access issues that were affecting NCAR's data origins. As part of this effort, the Pelican team began sharing access logs with NCAR, which provide metrics such as successful and failed access attempts for data served from NCAR's OSDF origin. These logs, available here, offer critical insights into system performance and user access patterns. For example, since the OSDF was deployed as a production service in May 2025, users from 5507 Unique IP addresses have transferred 300 TB of data through over 37 million web requests from NCAR's access point on the OSDF infrastructure. In parallel, the NCAR team collaborated with an intern from the SIParCS (Summer Internships in Parallel Computational Science) program to develop an **OSDF cache outage map.** This tool shows, in real time, the status of caches serving NCAR's data—including the nearest functional cache and any failed caches—thereby improving transparency and user experience.

Development of Reference Computational Workflows

In addition to this, we developed reference computational workflows that utilize NCAR datasets through OSDF origins. We published around **16 workflows** spanning more than **eight** unique datasets, including CESM2-LENS, ERA5, JRA-3Q, NA-CORDEX, SAAG dataset, DART reanalysis and CONUS 404. These workflows support a range of scientific analyses such as temperature bias correction, GMST anomaly computation, precipitation analysis, and climatological averaging. We also developed and published multiple workflows that access CMIP6 Zarr datasets hosted on AWS, further showcasing the utility and flexibility of OSDF-enabled data access from multiple data origins. In order to perform cross-platform testing, we deployed these workflows on four distinct computational platforms: Texas Advanced Computing Center's (TACC) Stampede3, NCAR's Casper, Jetstream2, and the Open Science Pool (OSPool). This demonstrates the portability and scalability of the workflows across a mix of HPC and cloud-native environments and reinforces their practical usability in real-world computing scenarios. All workflows have been carefully documented and published publicly on GitHub to support community adoption and reproducibility. The GitHub link for this repository is https://github.com/NCAR/osdf_examples ; the Zenodo DOI can be found here: https://zenodo.org/records/16863133.

A subset of these workflows has been archived in a **Pythia Cookbook,** which was developed during an NCAR-OSDF project supported Aug 2025 Pythia Cookoff session, and is due to be published in the Project Pythia cookbook gallery this Fall. The cookbook can be viewed at https://projectpythia.org/osdf-cookbook/. The first chapter of this cookbook introduces the **Pelican** software and the OSDF infrastructure. Subsequent chapters provide examples of scientific workflows (contributed by different Pathfinder teams) that use this infrastructure to ingest data. The NCAR team contributed to the **'NCAR examples'** chapter, where we explained the role of GDEX and provided two example notebooks that stream data from different datasets archived in the RDA. A stable version of the cookbook has been published on Zenodo with the DOI: https://zenodo.org/records/16802785. We also collaborated with other NDC grant recipients under the umbrella of the Pathfinder project to archive their research workflows in this cookbook. These demonstrate how to successfully ingest data from various OSDF origins, including the AWS open data origin, in scientific workflows using the Pelican Python client (PelicanFS).

During the course of this project, we identified and communicated infrastructure gaps within the NDC ecosystem. We reported the inability to deploy Dask on the OSPool and began a collaboration with a team at the University of Notre Dame that works on a package called floability to address this limitation. The first prototype of this collaboration was successfully executed on TACC. We also documented issues with cache selection and uptime reliability across various OSDF origins that serve NCAR datasets. These findings provide concrete, actionable insights for improving OSDF's data access infrastructure.

Community Engagement

During the course of this project, we maintained regular communication with the Pelican team via Slack and recurring meetings. This ongoing engagement allowed us to provide targeted feedback based on our experience developing and maintaining PelicanFS-based workflows, contributing to infrastructure enhancements and alignment with scientific workflow needs and expectations of the geoscience community.

In addition, we have actively collaborated with all the teams of the Pathfinder project to help remove data access bottlenecks in their computational workflows. Beyond identifying NCAR-hosted datasets useful to collaborators, we played an active leadership role in the success of the Pathfinder project by developing initial workflows that were both relevant to their research and designed to leverage the OSDF data infrastructure.

More broadly, we have engaged extensively with the Earth systems science community through presentations, leadership roles, and active participation in key initiatives. We presented the progress on the Data Commons project at several NSF NCAR ESDS forums. Beyond this, we presented ongoing work on NCAR's collaboration with the OSDF at the High Throughput Computing 2025 conference, reaching a broad audience focused on scalable research computing. We also gave multiple presentations at NDC Pathfinder meetings, ensuring regular engagement with the NDC community and dissemination of progress. In order to reach a broader audience, we organized a session titled **"Advancing Earth System Science by innovating distributed streaming data access and data-proximate computation"** at the Earth System Information Partners July meeting. The session was aimed at two different groups of participants: (1) open data providers who might be interested in learning how to leverage the capabilities of OSDF to provide distributed access to their geoscience datasets, and (2) researchers who are interested in using these datasets. In addition to a vibrant discussion on the use of OSDF infrastructure, the session involved talks by other Pathfinder collaborators. Finally we provided participant support to 11 participants of the Aug 2025 Pythia cookoff, which included 3 representatives from HBCUs.

Future Work

Looking ahead, several areas of improvement and expansion have been identified in order to strengthen NCAR's integration with the OSDF ecosystem and broaden its impact on the geoscience community.

1. **Dataset mutability.** At present, objects are treated as immutable within the OSDF setup. This results in undefined behavior when datasets are updated at their data origins. Developing mechanisms to handle dataset versioning and updates consistently is essential.
2. **Performance variability across origins.** We have observed slow data downloads from certain origins (e.g., AWS). Addressing these bottlenecks will be critical for supporting high-performance scientific workflows at scale.

3. **Expanding dataset coverage.** Many crucial Earth System Science datasets—notably from NOAA and NASA—are not currently available through AWS or connected to OSDF. Incorporating these datasets will significantly improve interoperability with NCAR's holdings and strengthen OSDF's value as a federated infrastructure for Earth system science.
4. **Cache selection and reliability.** There is a pressing need for OSDF's director services to automatically select alternative caches that are spatially proximate to the computational resource when a cache is unavailable or overloaded. We plan to provide feedback to the Pelican team to guide the development of improved cache-selection algorithms.
5. **Workflow development and training.** More workflows need to be developed that span multiple datasets and showcase cross-origin analysis. To accelerate adoption and build community expertise, we plan to organize a workshop or hackathon to train researchers in leveraging the OSDF data access infrastructure for their own science.
6. **Leveraging AI-assisted workflows.** Large Language Models (LLMs) present an opportunity to automatically generate basic data visualization workflows that can later be vetted and refined by domain experts. This approach could accelerate exploratory analysis, lower barriers for new users, and expand the range of reproducible workflows available to the community.

Bibliography

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18