

National Aeronautics and Space Administration

Data Publication and Discovery

Shubha Ranjan
Lead, Data Analytics;
Data Publication and Discovery

shubha.ranjan@nasa.gov

May 22, 2024



NASA High-End Computing

Here's what we can do for you

- Assistance with data publication and discovery projects.
  - Provide guidance in solving big data problems, using HECC resources.
  - Framework for sharing datasets stored at NAS with the public.
  - Automatic publication of public datasets as part of data pipelines.

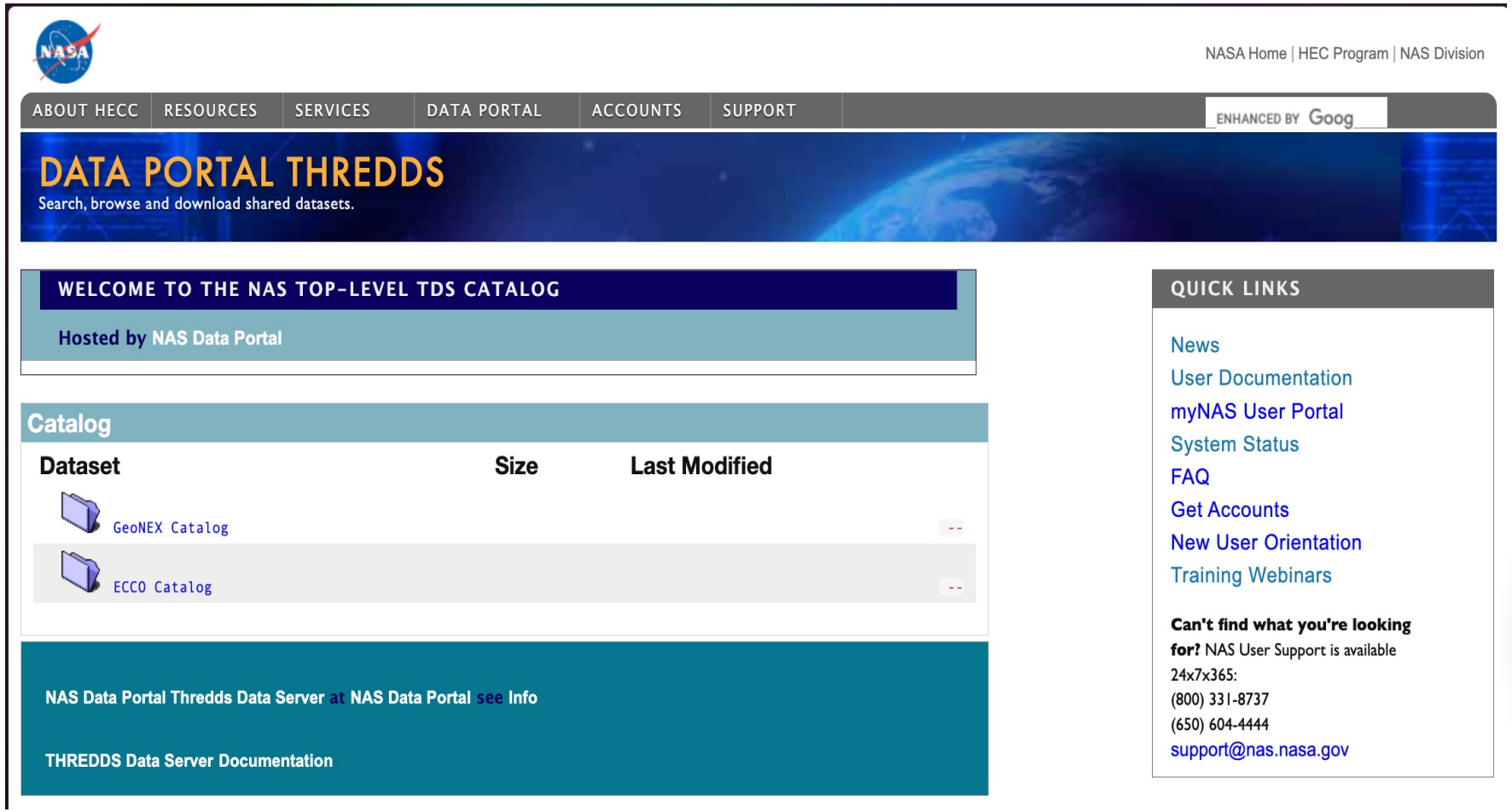


Success stories

- Added nine new datasets and two custom search and query portals over last two years.
- Enhanced availability, reliability, and security.
- Launched the NAS THREDDS Data Server to allow users to download subsets of scientific data that are available on the Data Portal:
  - Service developed by Unidata (UCAR) to compile various data formats into a Common Data Model and serve them through a catalog service.
  - Leverages existing NAS Data Portal assets.
  - Re-exporters—used to make datasets available outside of the NAS secure enclave—were extended to provide THREDDS catalogues.
  - Enables external users to browse available datasets through a web interface and download files using a variety of remote access protocols.
  - Serves gridded data in GeoTIFF or netCDF format, allowing weather and climate services (WCS) clients to specify a subset of a gridded dataset, gather metadata information, and/or visualize the extracted data through common scientific visualizing software packages.
  - Enables existing HECC users to request access to their shared datasets through the THREDDS server.
- The server is available at: **https://data.nas.nasa.gov/thredds**



Success stories



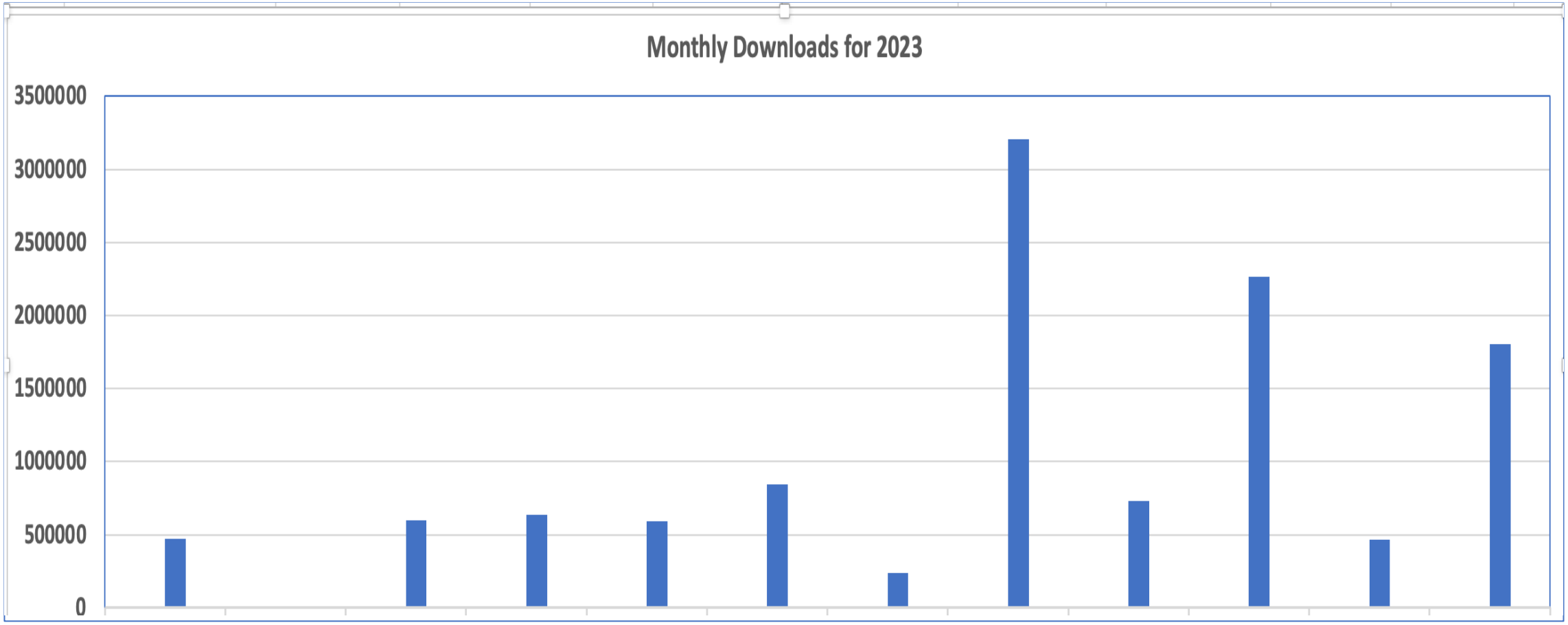


You can also get in-depth support

- Get dedicated filesystems for data storage and/or sharing.
- Web and data portal for sharing data stored on HECC resources with colleagues outside HECC and/or the public.
- Capability to share as Public/Private datasets with internal NASA and/or external users via the NAS Data Portal.
- Private datasets can be used with external users via IdMax external accounts feature.



Metrics for NAS Data Portal downloads in 2023: 23 Million (~2 Million/Month) - 52 terabytes



Looking forward

- We are expecting a steady growth in shared data/publications.
- A planned expansion of the THREDDS Server will enable more scientific datasets.
- The Data Portal hardware infrastructure is being completely revamped to bring in new enhanced storage with up to 200 petabyte (PB) capacity.
  - The current footprint of ~10 PB is expected to grow with the need for more data and compute for increasingly complex data analysis practices agency-wide.
  - Actual storage will be added incrementally as needed.





National Aeronautics and Space Administration

Data Analytics

Shubha Ranjan
Lead, Data Analytics;
Data Publication, and Discovery
shubha.ranjan@nasa.gov

May 22, 2024



NASA High-End Computing

Here's what we can do for you

- Assistance with using machine learning (ML) technology.
  - Provide guidance with machine learning tools and techniques.
  - Provide support for running ML tools on HECC resources.
  - Help with TensorFlow, PyTorch, Jupyter Notebooks, TensorBoard.
  - Provide support for large training sets: Distributed TF, Horovod, Ray
  - Work with NASA teams to help move to AI/ML technologies and/or port your work to scale on HPC cluster GPU nodes.
  - AI/ML services also available on AWS.



You can also get in-depth support

- Work with your pilot projects to develop a framework addressing ML needs that can be applicable for multiple projects.
- Assist with porting and /or scaling code and software packages on HEC systems.
- Provide more focused support to infuse AI/ML and deep learning technology into projects.
- Assist with exploratory data analysis (EDA) for new and proposed projects.

Note: Additional funding may be required depending on the level of effort.



Success stories

- Trained recent Foundation Models for hurricane prediction by SMD.
  - Scaled up to 128 Cabeus A100 nodes for scaling large training runs distributed using Ray.
- Participated in internal SBU determination and benchmarking of new systems, for representation of AI/ML workloads in these metrics.
- Successfully completed AI expansion project.
  - Completed one year of pilot projects with five teams that ported their code over to HECC and reported success; use cases represented Aeronautics Research and Science Mission Directorates (Earth Science, Heliophysics).
- Data Analytics joined Anaconda licensed repositories as required by the agency to ensure security of our systems.



Looking forward

Support for large language models (LLMs)

- Recent increase in use of LLMs at NASA.
- Many users request downloading models from Hugging Face.
- HECC has identified multiple copies of downloaded models.
- We're creating an in-house repo to cache the downloaded base models to be used by multiple users.
  - This would standardize the process and streamline the common model usage.
  - We are also charting a workflow for the best way to curate these models.



We welcome your input – requirements, suggestions, and requests.

National Aeronautics and Space Administration

Ideas?
Questions?
Suggestions?

We want to hear from you!
support@nas.nasa.gov
1 (800)-331-8737

