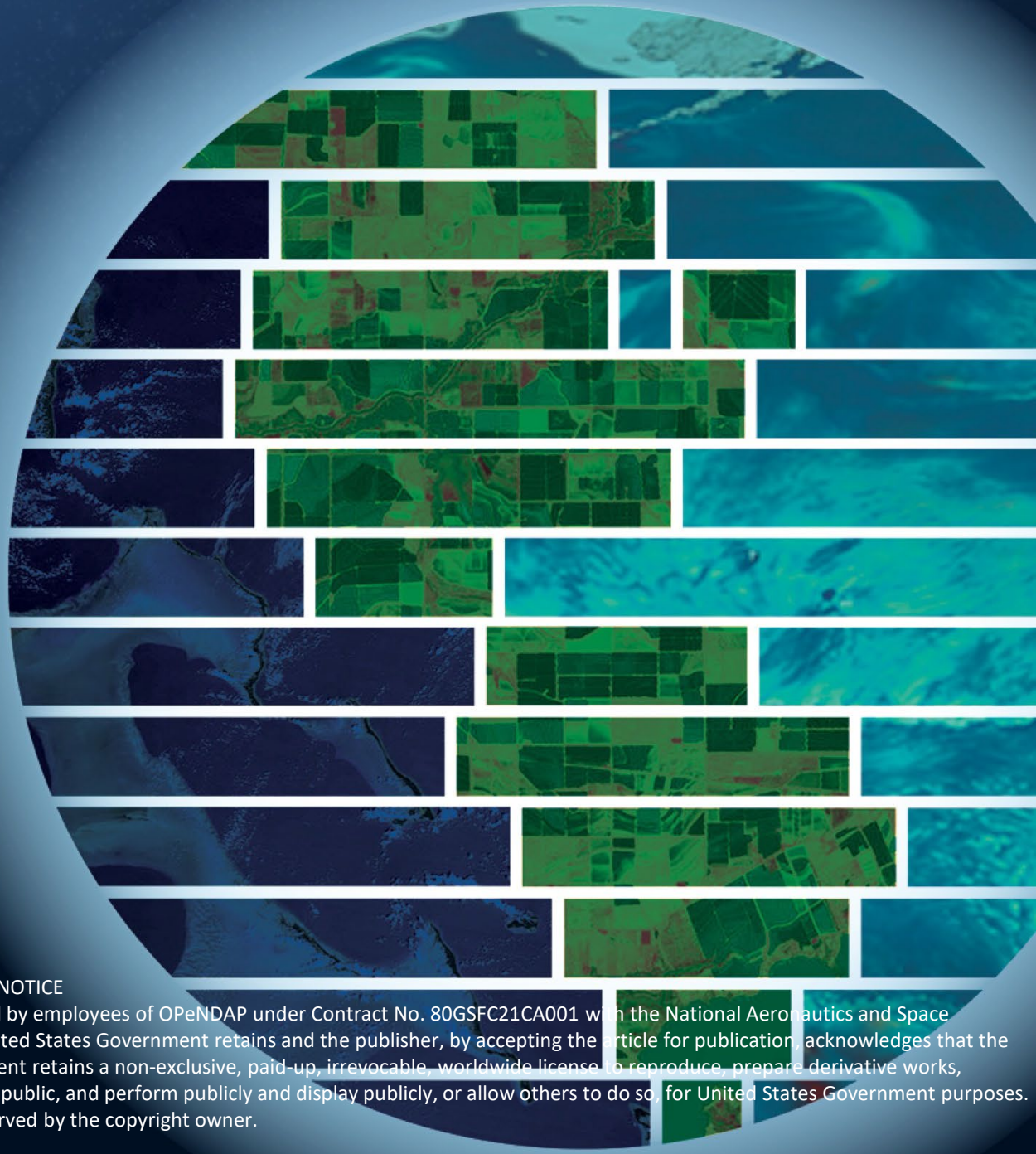
National Aeronautics and Space Administration

EARTHDATA

Cloud-Optimized Data Access



Presentation Date 07/24/2025

James Gallagher
EED3 Contractor, President/CEO OPeNDAP

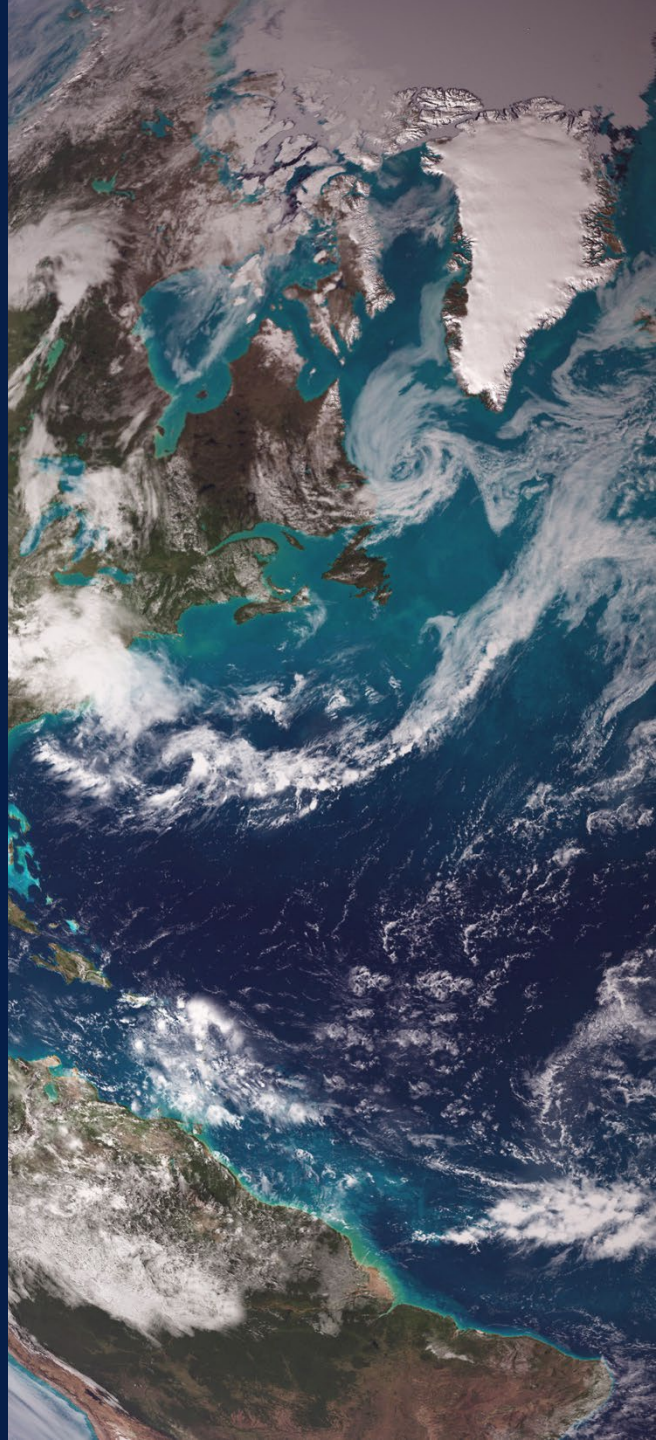
Miguel Jimenez-Urias, PhD.
Scientific Community Director. OPeNDAP

Agenda

Core concepts review

Technologies

Call to action



Cloud-Optimized Data Access

- Cloud storage: scalability, durability, distributed access
- Why optimization matters: performance, cost, usability

Core Concepts – Formats

- Cloud-optimized: GeoTIFF (COG), Zarr, COH5
- Legacy (Pre-cloud): HDF4/5, HDF-EOS, NetCDF3/4
- Legacy formats need ***extra*** metadata for efficient cloud access

COG: Cloud Optimized GeoTIFF
COH5: Cloud Optimized HDF5
HDF: Hierarchical Data Format
EOS: Earth Observing System
NetCDF: Network Common Data Format
Zarr: A shared data storage format

EARTHDATA
Earth Science Data and Information System (ESDIS) Project

Core Concepts – Access Patterns

- HTTP range requests: partial reads while files remain at rest
- Client-side metadata indexing enables efficient subsetting
- Subsetting is a requirement for large, remote files
- REST architecture: metadata documents define access state

Note: In this presentation, “metadata” refers specifically to *chunk manifests*—not to semantic or descriptive metadata. Chunk manifests define the structural layout of data objects and guide software in efficiently interpreting and accessing them.

HTTP: Hyper Text Transfer Protocol
REST: Representational State Transfer

DMR++

- Extension of OPeNDAP DMR/XML with chunk indexing metadata
- Enables serverless subsetting and remote access
- Reduces data transfer via pre-fetched metadata
- Used in NASA data systems for HDF4/5 and EOS data

DMR: Data Metadata Response

XML: Extensible Markup Language

Kerchunk

- From the Pangeo community: JSON metadata catalogs
- Supports NetCDF, HDF5 and others
- Works with Xarray/Zarr backends for seamless reads
- No data rewriting required – ideal for legacy data access

JSON: Javascript Object Notation
Xarray: A python language interface for data access

Icechunk

- Created by Earthmover, Inc. – general-purpose engine
- Cloud-native, transactional Zarr-compatible backend
- Supports ACID transactions, version control, Git-like ops
- Handles HDF5, NetCDF4, GRIB, GeoTIFF – no reformat needed

ACID: Atomicity, Consistency, Isolation, and Durability
GRIB: General Regularly-distributed Information in Binary form

EARTHDATA
Earth Science Data and Information System (ESDIS) Project

Comparative Summary

- Each supports HDF5 in various ways, taken as a group, quite a breadth of formats are covered
- Client- and server-side uses of the metadata are implemented
- Icechunk: Powerful versioning and virtualization features

Operational Considerations

- Storage and compute cost implications
- Ease of integration with existing workflows
- Community and ecosystem support
- Widespread access and metadata reuse best practices

Next: Presentations

- Brief presentations of existing implementations
  - Types of chunk manifests (aka metadata)
  - Client-side use of chunk manifests
- Breakout groups

Breakout Group Topics

1. **What do clients need?** Exploring use cases and the metadata required to support them
2. **Can we agree on a core structure?** Defining shared characteristics for a common data model
3. **How should we package structure?** Designing chunk manifest formats for interoperability

Instructions: Choose one, two, or all three topics to explore. Begin an open conversation with your group and take notes as you go. These notes will help inform future working groups tasked with formally advancing these concepts into community-driven standards.

Turn today's discussion into momentum.

Your ideas from the breakout groups can inform practical, shared solutions.

CONTACT
jgallagher@opendap.org
mjimenez@opendap.org

WEBSITE
www.opendap.org

EARTHDATA

earthdata.nasa.gov

Thank You Participating

This work was supported by NASA/GSFC under Raytheon Company
contract number 80GSFC21CA001