

Eniola Awowale¹, Thomas Nicholas², Owen Littlejohns³, Stephan Hoyer⁴, Matt Savoie⁵, Justus Magin⁶, Alfonso Ladino⁷

¹ NASA Goddard Earth Sciences Data and Information Services Center (NASA GES DISC), ²[C]Worthy, ³NASA Earth Observing System Data and Information System (EOSDIS) Evolution and Development (NASA EED-3), ⁴Google, ⁵National Snow and Ice Data Center, ⁶IFREMER, ⁷University of Illinois Urbana-Champaign

Introduction

NASA's Earth Observing System Data and Information System (EOSDIS) contains thousands of Earth Science datasets. EOSDIS data are stored in formats that are well supported by the Earth Science community, including the Hierarchical Data Format (HDF), with its derivatives such as HDF-5 and the Network Common Data Format (NetCDF-4). HDFs support large multidimensional data, are self describing, and have a directory-like hierarchy within a single file, known as "groups". Observational data and associated metadata within a single file can be distributed amongst multiple internal groups, which can also be nested to multiple levels.

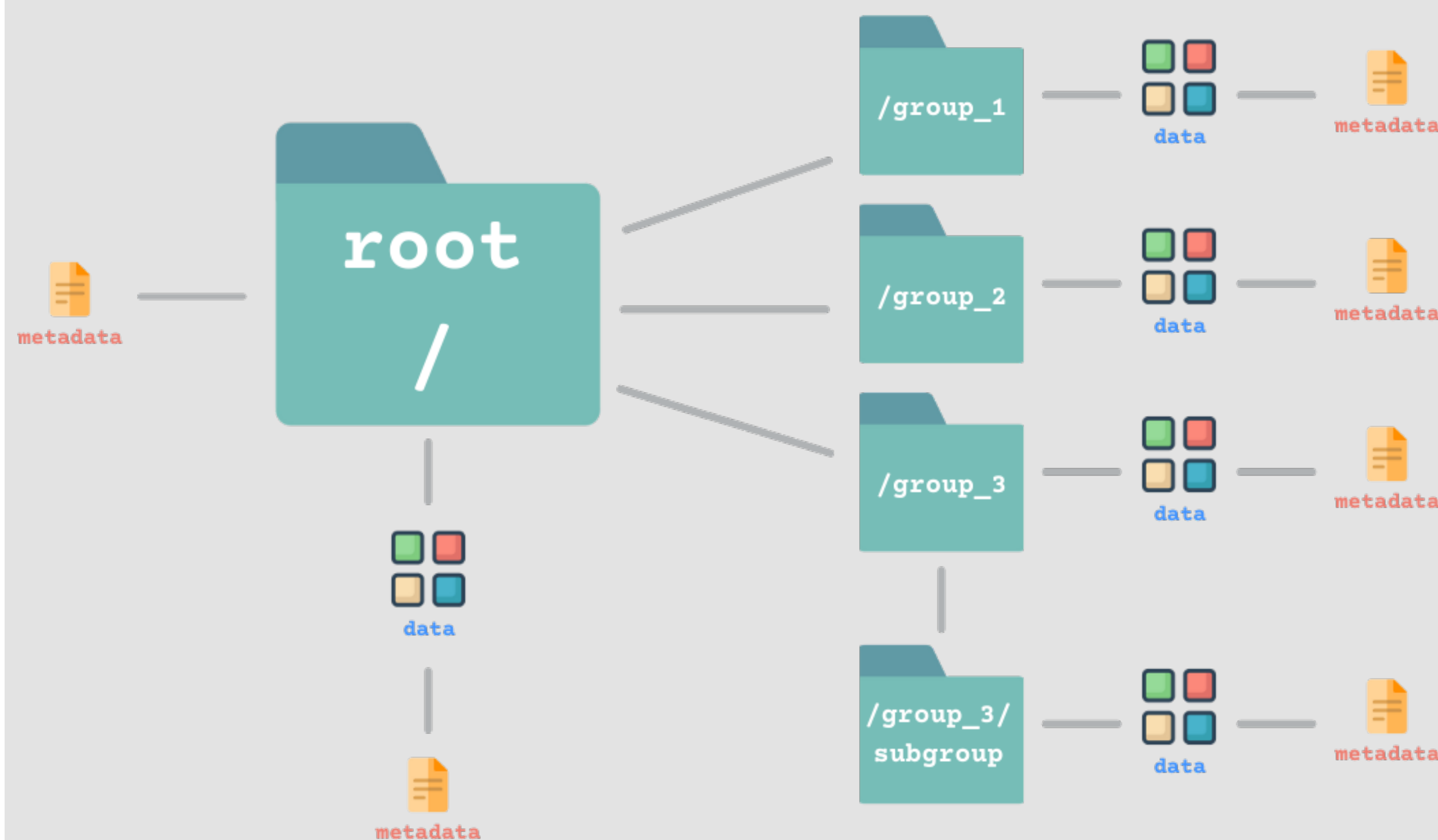


Figure 1. A grouped HDF file

Motivation

Working with datasets that have a group hierarchical structure can be difficult because of the nested structure of groups. Before the 2024.10.0 release, xarray's data model did not accommodate hierarchical HDF files. Users had to traverse the file by opening different HDF groups as separate, unrelated xarray.Dataset objects. This makes accessing Earth Science datasets difficult because related variables can be stored in different hierarchical groups. Additionally, groups and their metadata structures can vary wildly between datasets, making it challenging to create dataset-agnostic tools or services.



Figure 2. The Global Precipitation Measurement Core Observatory satellite in orbit above the Earth

Try it out on GitHub!



Acknowledgements

The authors thank the GES DISC, ESDIS, and xarray scientists and engineers for their support with xarray.DataTree. Special regards to Steven Hessenius, Lucas Sterzinger, Jennifer Adams, Brianna Pagán, Chris Battisto, and Andrey Savtchenko from GES DISC for their help and advice with preparing this poster. Icon images obtained from Wikimedia Commons. Satellite data products are available at disc.gsfc.nasa.gov and search.earthdata.nasa.gov.

Data Model

A DataTree is now xarray's highest-level data structure. It builds on xarray's data model by introducing a hierarchical tree of nodes (each node is itself a DataTree object). DataTree objects have the same properties as xarray.Datasets, including dimensions, coordinates, data variables and attributes. DataTree objects have additional features that help define relationships between groups.

Nodes

Groups in the DataTree model are connected through parent/child node relationships. In Figure 3, groups "/day_1", "/day_2", "/day_3", and "/day_3/misc" are all descendant nodes of the root group ("/"). A group in a HDF file is a node in a DataTree and is also a DataTree object. A root node is a node without a parent node and is the root group in a HDF file (see Figures 1 and 3).

Inheritance

The DataTree model supports inheritance for coordinates, dimensions, and indices that is applied downward to all descendent nodes. When using DataTree with a grouped HDF file, all dimensions and their indices must be aligned with a descendant node's existing dimensions or indices, such that identically named dimensions must have the same length. This inheritance model allows child nodes to use dimensions defined in ancestor nodes and reduces data duplication.

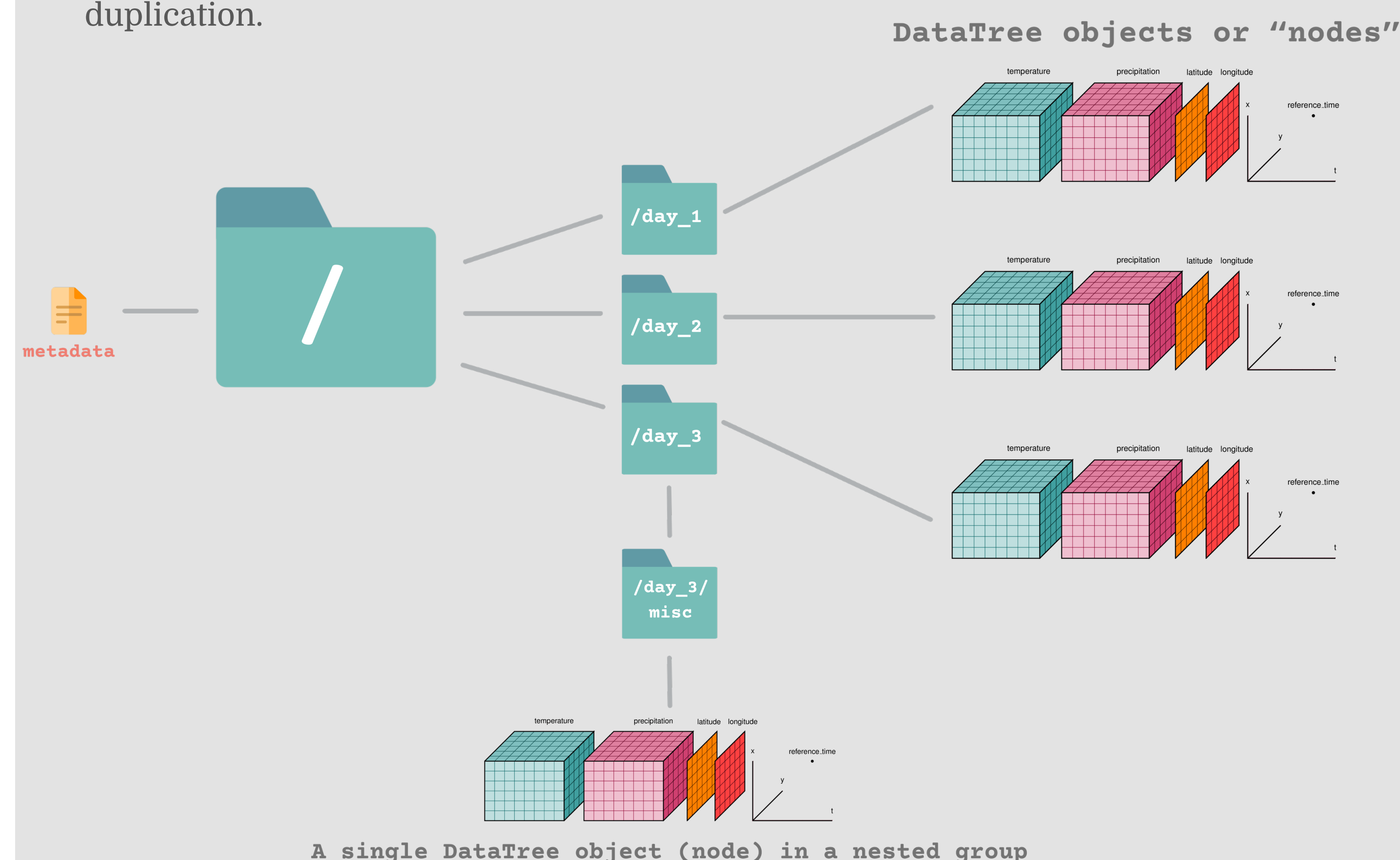


Figure 3. A representation of the DataTree model with timeseries data

Features

- Opens grouped HDF files as a nested tree
- Groups are connected by parent/child relationships
- Access to child and parent groups with file path or attribute syntax
- Computations map over the entire tree, enabling easier data analysis
- Coordinates and dimensions defined in ancestor nodes are automatically applied to all descendent nodes

Usage

We used xarray.DataTree to create a timeseries of the *GPM_3IMERGDL_07* dataset. Xarray's standard plotting functionality for DataArrays was used to plot the daily mean precipitation in Spain during a historic rainfall period in late October 2024.

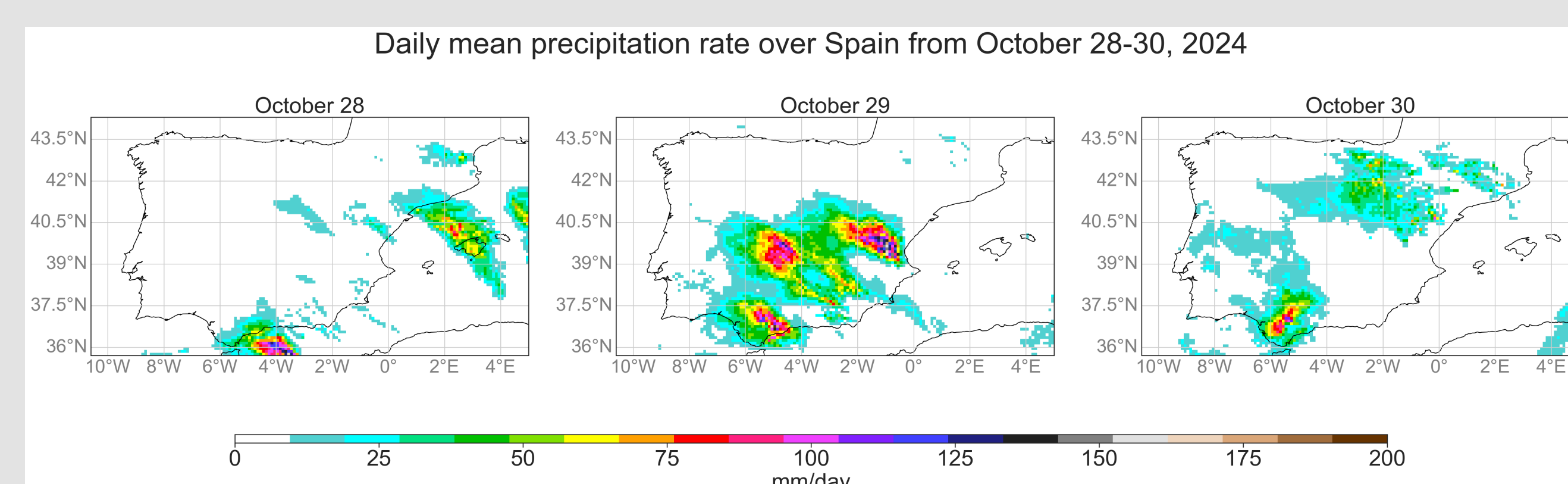


Figure 4. Daily mean precipitation rate (combined microwave-IR) estimate (formerly *precipitationCal*) from *GPM_3IMERGDL_07*

What's next

- Continued development and improvements for xarray.DataTree
- Development of a Jupyter notebook tutorial with ESDIS data utilizing xarray.DataTree
- Adapting NASA ESDIS tools and services to use xarray.DataTree for opening, analyzing and transforming NetCDF-4 and HDF-5 data.