Data Product Development Guide for Data Producers

Version 2.0, May 20, 2024

STATUS OF THIS MEMO

This memo provides information to the National Aeronautics and Space Administration (NASA) Earth Science Data Systems (ESDS) community. This memo describes a "Suggested Practice" and does not define any new ESDIS Standards. Distribution of this memo is unlimited.

CHANGE EXPLANATION

Many changes have been implemented in this revision to Version 1.1. New material, including Figure 3, has been added regarding data products from airborne and field campaign investigations. Section 3.3 on cloud-optimized formats and services is new. Section 8 has been expanded to include Earthdata Pub. Appendices D and E have been revised to include mappings of the attributes to the Unified Metadata Model (UMM) profiles, and expanded to include the Findability, Accessibility, Interoperability, and Reusability (FAIR) sub principles.

COPYRIGHT NOTICE

SUGGESTED CITATION

Ramapriyan H. K., P. J. T. Leonard, E. M. Armstrong, S. J. S. Khalsa, D. K. Smith, L. F. Iredell, D. M. Wright, G. J. Huffman, and T. R. Walker "Data Product Development Guide (DPDG) for Data Producers version 2.0. NASA Earth Science Data and Information System Standards Office, 20 May 2024. https://doi.org/10.5067/DOC/ESCO/RFC-041VERSION2

ABSTRACT

This Data Product Development Guide (DPDG) for Data Producers was prepared for the Earth Observing System Data and Information System (EOSDIS) by the DPDG Working Group, one of the Earth Science Data System Working Groups (ESDSWGs), to aid in the development of NASA Earth Science data products.

The DPDG is intended for those who develop Earth Science data products and are collectively referred to as "data producers." This guide is primarily for producers of Earth Science data products derived from remote sensing, *in situ* and model data that are to be archived at an EOSDIS Distributed Active Archive Center (DAAC). However, producers of other Earth Science data products will also find useful guidance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The National Aeronautics and Space Administration's (NASA's) Earth Observing System Data and Information System (EOSDIS) is a major capability in the Earth Science Data Systems (ESDS) Program [1]. EOSDIS Science Operations (i.e., data production, archive, distribution, and user services), which are managed by the Earth Science Data and Information System (ESDIS) Project [2], are performed within a system of interconnected Science Investigator-led Processing Systems (SIPSs) and discipline-specific data centers called Distributed Active Archive Centers (DAACs).

This Data Product Development Guide (DPDG) for Data Producers was prepared for EOSDIS by the Earth Science Data System Working Groups (ESDSWGs) [3] under the supervision of the ESDIS Project to aid in the development of NASA Earth Science data products. This version is a major update to the DPDG V1.1 [4] and includes material relevant to data products from airborne and field campaign investigations, with new examples and edits to several sentences throughout the document as well as the addition of a new Figure 3. Additional information about airborne and field campaign data delivery and management can be found at [5]. Section 3.3 on cloud-optimized formats and services is new. Section 8 has been expanded to include Earthdata Pub. Appendices D and E have been revised to include mappings of the attributes to the Unified Metadata Model (UMM) profiles, and expanded to include the Findability, Accessibility, Interoperability, and Reusability (FAIR) sub principles.

The DPDG is intended for those who develop Earth Science data products and are collectively referred to as "data producers" (see Appendix B). This guide is primarily for producers of Earth Science data products derived from remote sensing, *in situ*, and model data that are to be archived at an EOSDIS DAAC. However, producers of other Earth Science data products will also find useful guidance.

There is an abundance of documents (e.g., regarding standards, conventions, best practices, and data formats) to direct developers in all aspects of designing and implementing data products. Moreover, some DAACs have developed guides for particular data producers and specific scientific communities [6] [7] [8] [9] [10] [11]. The DPDG aims to compile the most applicable parts of existing guides into one document that logically outlines the typical development process for Earth Science data products. Emphasis has been given to standards and best practices formally endorsed by the ESDIS Standards Coordination Office (ESCO) [12], findings from ESDSWGs, and recommendations from DAACs and experienced data producers. Ultimately, the DPDG provides developers with guidelines for how to make data products that best serve end-user communities—the primary beneficiaries of data product development. The DPDG also guides the developers to ensure that the data products are designed to be Findable, Accessible, Interoperable, and Reusable (FAIR) [13], and adhere to NASA's open science objectives and information policies [14] [15].

The data products are assumed to be archived at a DAAC, and it is vital that data producers work closely with the DAACs to which their products are assigned to obtain details not covered in this document. This document indicates in the respective sections where such close communications between the data producers and their assigned DAACs are needed. Examples of areas requiring such communications are: product design, understanding user requirements, selecting a data format, product naming, metadata requirements, testing and receiving user feedback, optimizing product

formats for use in the cloud, selection of keywords for the products to facilitate user searches, version numbering, obtaining Digital Object Identifiers (DOIs), and product delivery and publication schedules.

The ESDIS Project and the DAACs are actively engaged in migrating data products and services to the cloud, and, for this reason, we include guidance regarding cloud-optimized formats and services (Section 3.3). It is stressed that data producers should work with their assigned DAAC early in the lifecycle of product architecture and implementation when optimizing their data for cloud distribution and computing.

The organization of the rest of this document is illustrated in Figure 1, which shows the various steps in data product development and delivery. The numbers in parentheses in the figure indicate the sections where the individual steps are discussed. Sections 9, 10 and 11, not shown in the figure, cover a bibliography, authors' addresses, and a list of contributing authors and editors, respectively. Finally, five appendices provide a list abbreviations and acronyms, a glossary, details of selected tools useful for testing, as well as important global and variable-level attributes that should be included in the product metadata.

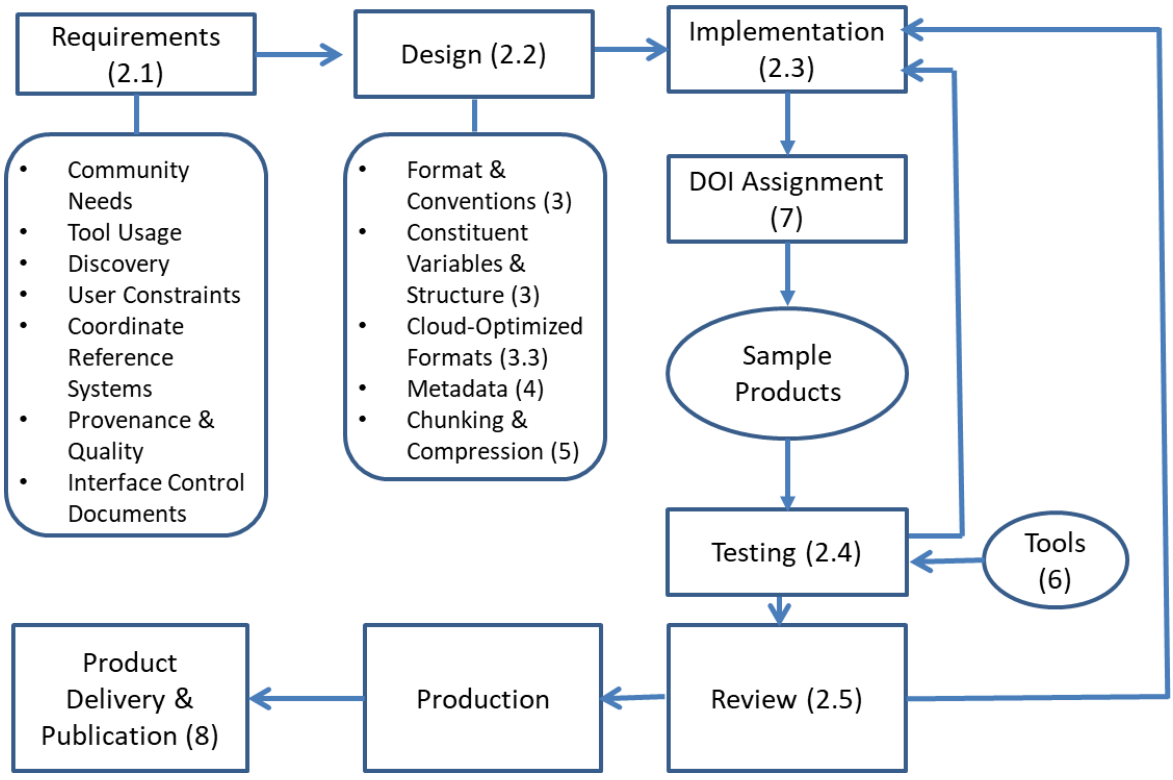

Figure 1. The flow of activities for data product development, production, and delivery (the numbers in parentheses in the figure indicate the sections where the individual steps are discussed).

2 DATA PRODUCT DESIGN PROCESS

For this guide, a data product is defined as a set of data files that can have multiple variables (a.k.a. parameters), which compose a logically meaningful group of related data [16]. This concept is

equivalent to a data collection in the Common Metadata Repository (CMR) [17], and is known colloquially as a dataset (see Appendix B for further explanation of these terms).

Based on the Earth Observing System (EOS) heritage [18], standard Earth Science data products have the following characteristic. They:

- have a peer-reviewed algorithmic basis.
- have wide research and application utility.
- are routinely produced over spatially and temporally extensive subsets of data.
- are available whenever and wherever the appropriate input data are available.

Since the beginning of the EOS Program, an extensive set of data products has been produced and archived that satisfy the above criteria that define standard products. Experience from those products and observations of related issues regarding their interoperability and metadata content have led to several recommendations from the ESDSWGs. Following those recommendations, which have been incorporated into the following sections, will result in a good data product design. It would benefit users of standard data products (as well as users of other data products that do not necessarily meet all the standard product criteria) if the guidance provided below is followed for the product design.

2.1 Requirements: Determining User Community Needs

At the beginning of the design process, the key elements are to 1) identify the expected user communities, and understand their needs regarding data formats, data structures, and metadata, in addition to what is required for data search and discovery; and 2) identify the needs for data tools and services. Developers can acquire this information by surveying the scientific literature, browsing predecessor data, holding and attending data applications workshops, and working with the DAAC that will archive the data product. While it is important to do this early in the product design process, it is equally important to remember that as Earth Science evolves to serve novel uses, user communities can change significantly. This is especially true in the case of long-term projects that might involve several reprocessing cycles, where such changes would need to be accommodated in later versions of the products. It is possible that a given data product would be archived at more than one DAAC (e.g., in the case of airborne campaigns), and the DAACs may have different format and/or metadata requirements. In such cases, some negotiations may be needed between the data producers and the assigned DAACs to understand the reason for the differences and see if the DAACs can accept a common approach.

Once user communities are identified, the key questions to be considered are the following:

- How might these data be used by the identified communities?
- Which tools and services will the community need to use the data?
- Are there common workflows applied to the data (e.g., subset >> quality filter >> re-grid)?
- What are the prevalent data formats, data and metadata standards, and data structures used by the community?
- What common keywords would assist in data discovery?

- What constraints are faced by the user community (e.g., timeliness, network bandwidth, processing capacity, and disk storage)?
- What temporal and spatial resolutions, coordinate systems, and map projections are commonly used in or required by the community?
- What information on data provenance (i.e., data product history and lineage) and quality will the users need for their purposes?
- What other information do the users need to assess the suitability of the data?
- How should the data product be designed to make it useful to unexpected user communities (e.g., make the files self-describing)?
- What associated knowledge should be preserved [19] in addition to the data for the benefit of future users, when data producers are no longer available for consultation?
- How should the associated knowledge be preserved (implementation guidance is available at [20])?
- How does a shift to a cloud environment affect the preceding questions?

When the data collection is sufficiently complex in size, required services, or management, then an Interface Control Document (ICD) [21] may be necessary. An ICD provides the definition, management, and control of interfaces that are crucial to successfully transferring the data and metadata from the producer to the DAAC. The assigned DAAC can assess whether an ICD is needed for any given data collection.

2.2 Design: What Constitutes a Data Product Design

A data product design should address the following:

- Data format and associated conventions (Section 3).
- Identification and structure of constituent variables[1] (Section 3).
- Metadata (Section 4).
- Data chunking, internal compression, and packing (Section 5).

2.3 Implementation: Creating Sample Data Files

The data producer should create sample data files to support the evaluation of the data product design. This needs to start early in the dataset development phase for the testing and evaluation of sample products and to provide timely feedback. Ideally, this process should begin when a single sample file has been developed and well before the full data collection has been completed. The more realistic the sample data, the more helpful it is for evaluating the usefulness and appropriateness of the design. Even a product populated with random science data values can be suitable for checking usability; however, sample data coordinates should be accurately populated, as these are critical to the use of tools on a data product. The software library supporting the selected standard data format and various tools (Section 6) can be used to quickly create sample data files.

[1] In this document we use the term “variable.” Other commonly used synonyms to the term “variable” exist, such as “parameter”. See Appendix B for a full explanation.

2.4 Testing: Evaluating Sample Data Products

Testing of data products should be performed with the tools and services that user communities are expected to employ (see Section 6). Typically, testing will identify structural issues of the data, such as missing or mis-identified variables and attributes, and ordering of dimensions within the variables. The data products should also be tested using compliance checkers (Section 6.2). Data producers should consult with their assigned DAACs regarding the available compliance checkers and how to use these tools. Ideally, an iterative approach should be followed: supply the data product to the assigned DAAC and to selected representatives of the expected user communities, integrate feedback, re-test the product, and solicit additional feedback.

2.5 Review: Independent Evaluation of the Data Product

Soliciting external, independent evaluations can improve the quality and usability of a data product. The following are recommendations for establishing and conducting such reviews throughout the product life cycle.

- Obtain reviews from the distributors of the data product (i.e., the relevant DAAC or DAACs).
- To maintain objectivity, the evaluators should not be directly involved in the development of the data product. Ideally, evaluators should include representatives of the expected user communities and other subject matter experts.
- Perform multiple reviews during the development process to receive guidance well before the release of the product. Any resulting modifications should be documented.
- Feedback should be sought on the format, content, and quality of the data product. Four aspects of quality are defined in [22], namely scientific quality, product quality, stewardship quality, and service quality. At this stage, the scientific and product quality are of primary concern. The format, content, and quality should also help improve the applicability of the data product for specific uses.
- Responses to reviewers' comments should be documented and provided to the reviewers, so that any misunderstanding of the comments can be clarified.
- Reviews should address various aspects of the data product, including capabilities for search, access, exploration, analysis, interoperability, and usage.
- Reviews should verify that the data files and their variables have suitable names and enough supporting information to facilitate their understanding.
- While the data product user guides published by the DAAC hosting the data (which are provided when users retrieve a data product or file) are essential support for the usage of a data product, data producers should strive to make their products as self-describing as possible (i.e., with embedded metadata that describe the format and the meaning of the data).

It is also helpful to have a mechanism for the user community to provide feedback on the usability and quality of the data after the products are released. Such feedback should be gathered by the assigned DAAC and be conveyed as needed to the data producers to help improve the data products.

3 SELECTING A DATA PRODUCT FORMAT

Selecting an encoding (i.e., format) for a data product involves weighing the advantages and disadvantages of the applicable formats. Below are important items for DAACs and data producers to consider together when selecting a format:

- Is the format open (i.e., has an openly published specification)?
- Has a format already been specified in an existing document (e.g., an Interface Control Document)?
- Does the format provide for the widest possible use of the data product, including potentially new applications and research beyond the original intentions?
- Is the format widely used in the target user community for similar data or similar data analysis workflows?
- Was the format used for past long-term observations or models and can it therefore provide for or enable more efficient data processing and interoperability [23] with those observations or models?
- Would it serve the user community better if the data were written in the same legacy format as was used for past or long-term observations for consistency, or in a format compatible with data from other agencies, e.g., the National Oceanic and Atmospheric Administration (NOAA) or the United States Geological Survey (USGS), to increase interoperability?
- Does the format enable efficient data analysis workflows on both global and local scales? Local (as opposed to global scale) applications will often require frequent subsetting, reprojection, or reformatting of the data for combination or intercomparison with *in situ* point observations and physical models.
- Does the format enable efficient data analysis work flows over long time periods as well as relatively near real time?
- Does the format support “self-describing” files (see Appendix B), meaning the files contain sufficient metadata that describe the contents of the file?
- Has the ordering of dimensions been considered for facilitating readability by end users [24] (Rec. 2.10)? The ordering of dimensions can have a significant impact on the ease with which the data can be read.
- Does the format provide for efficient use of storage space (e.g., internal compression of data arrays, see Section 5), keeping file sizes practical for intended users, while minimizing the need to access multiple external files?
- Is the format supported by popular third-party applications (see Section 6) and programming environments, which could expand the user base and promote further development of tools and services?
- Is the version of the format supported by user community tools? The version of the format can be important, because new versions may not be readable by the libraries and tools currently in use by the user community.
- Are resources available to support a long-term commitment to the format? This includes people who can develop and maintain libraries, tools, and documentation for working with the format.
- Have optimizations, novel distribution mechanisms (e.g., streaming), or usage patterns (e.g., data downloading, in-cloud use) (see Section 3.3), etc. been considered? This may depend on the capabilities of the host facilities and the envisioned use cases.

Data producers should consider the interplay of data format standards with other ESCO [12] approved standards for metadata, data search, and access [25]. Producers should consult the web pages associated with these standards to understand the strengths, weaknesses, applicability, and limitations (SWAL) of each data format in those contexts. These web pages also contain information on deprecated standards and practices.

Note that although the selected data format may not satisfy all users of a product, given sufficient user demand for a particular output format, EOSDIS can usually provide reformatting services to cater to a variety of preferences in the user community. The data producers are advised to consult with their assigned DAACs to determine what reformatting services are available or can be implemented to satisfy the expected user community demand.

3.1 Recommended Formats

While several acceptable formats are listed by ESCO [25], the highly preferred format for EOSDIS data products is network Common Data Form Version 4 (netCDF-4) [26], which uses the Hierarchical Data Format Version 5 (HDF5) [27] data storage model. Although files in netCDF-4 can in theory be written via the HDF5 library API, inadvertent use of certain HDF5 features can render files unreadable by the rich ecosystem of netCDF tools. Therefore, we recommend use of the netCDF-4 library API. The ESCO review of the netCDF-4/HDF5 File Format [26] lists the SWAL of the format, which the reader may find useful.

Some of the advantages of using netCDF-4 are:

- Files are “self-describing,” meaning they allow for inclusion of metadata that describe the contents of the file (see Appendix B).
- Supports many data storage structures, including multidimensional arrays and raster images, and naturally accommodates hierarchical groupings of variables.
- Includes access to useful HDF5 features, and can be used in concert with HDF5 tools such as HDFView [28].
- Supports internal data compression.
- Is supported by several important programming languages and computing platforms used in Earth Science.
- Provides efficient input/output on high-performance computing systems.
- Readily allows for conversion to a cloud-optimized format.

Also, a well-established standard called the Climate and Forecast (CF) Metadata Conventions (hereafter, CF Conventions–see Appendices B, D and E) [29] specifies a set of metadata that provide a definitive description of what the data in each variable represent, and the spatial and temporal properties of the data. The CF Conventions were developed for netCDF; thus, they are sometimes referred to together as “netCDF/CF.”

If starting a new project with a user community that does not have a preferred format, then netCDF-4 (or a cloud-optimized version of netCDF-4) should be used. Data producers using legacy formats should work towards migrating to a more contemporary format.

3.1.1 NetCDF-4

A netCDF-4 file can include global attributes, dimensions, groups[2], group attributes, variables, and variable-level attributes. The global attributes provide general information regarding the file (e.g., author information, data product version, date-time range, product DOI). Dimensions can represent: 1) spatio-temporal quantities (e.g., latitude, longitude, time); 2) other physical quantities (e.g., atmospheric pressure, wavelength); and 3) instrumental quantities (e.g., along track, cross track, waveband). A netCDF variable is an object that usually contains an array of numerical data. The structure of a variable is specified by its dimensions. The dimensions included at a given level in the hierarchy can be applied to variables at or below that level. Variable-level attributes provide specific information for each variable (e.g., coordinates, units, valid range). Groups can be created to contain variables with some commonality (e.g., ancillary data, geolocation data, science data). Group attributes apply to everything in a group. Global attributes are attached to the root group.

Note that “dimensions” and “coordinates” are two terms in netCDF/CF that should not be confused with each other. For example, in a Level 2 (L2) swath file, the dimensions can be “along_track” and “cross_track,” while the corresponding coordinates can be “latitude” and “longitude” (illustrated in Figure 2). The coordinates for each variable are specified via the CF `coordinates`[3] attribute. Another example is *in situ* data collected during an airborne campaign (illustrated in Figure 3), where the dimensions are “observation number” and “trajectory number”, and the coordinates are “time”, “longitude”, “latitude” and “altitude”.

[2] Recently, the CF Conventions have been updated to include rules for files with group hierarchies [117].
[3] Words or phrases in this document that are colored `purple` indicate officially recognized CF attribute names or best practice names.

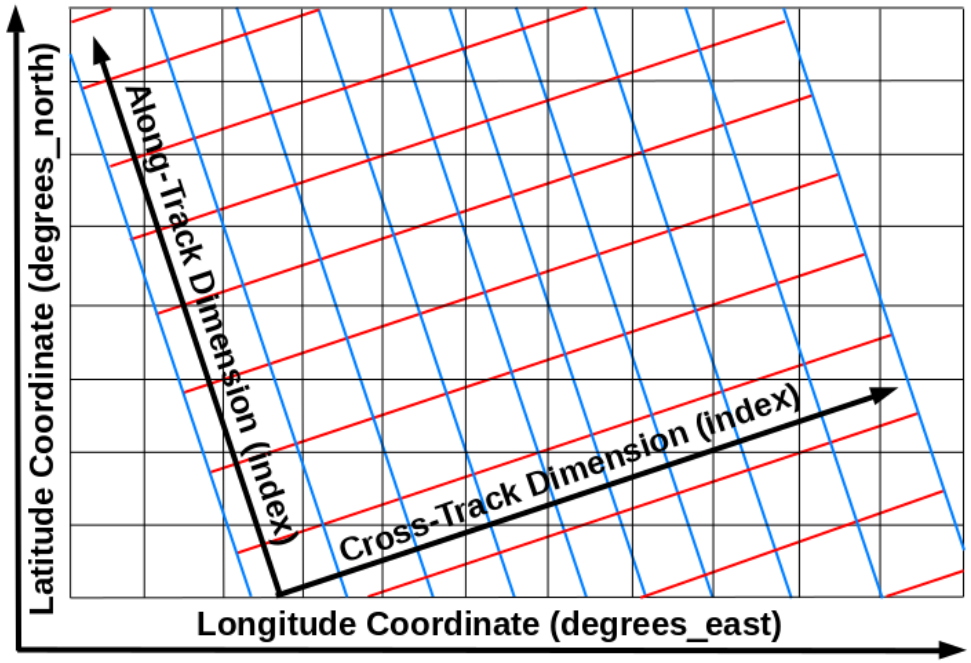

Figure 2. The horizontal dimensions (along_track and cross_track) and coordinates (latitude and longitude) for a swath product. In other products the horizontal dimensions and coordinates can be one and the same (e.g., latitude and longitude).

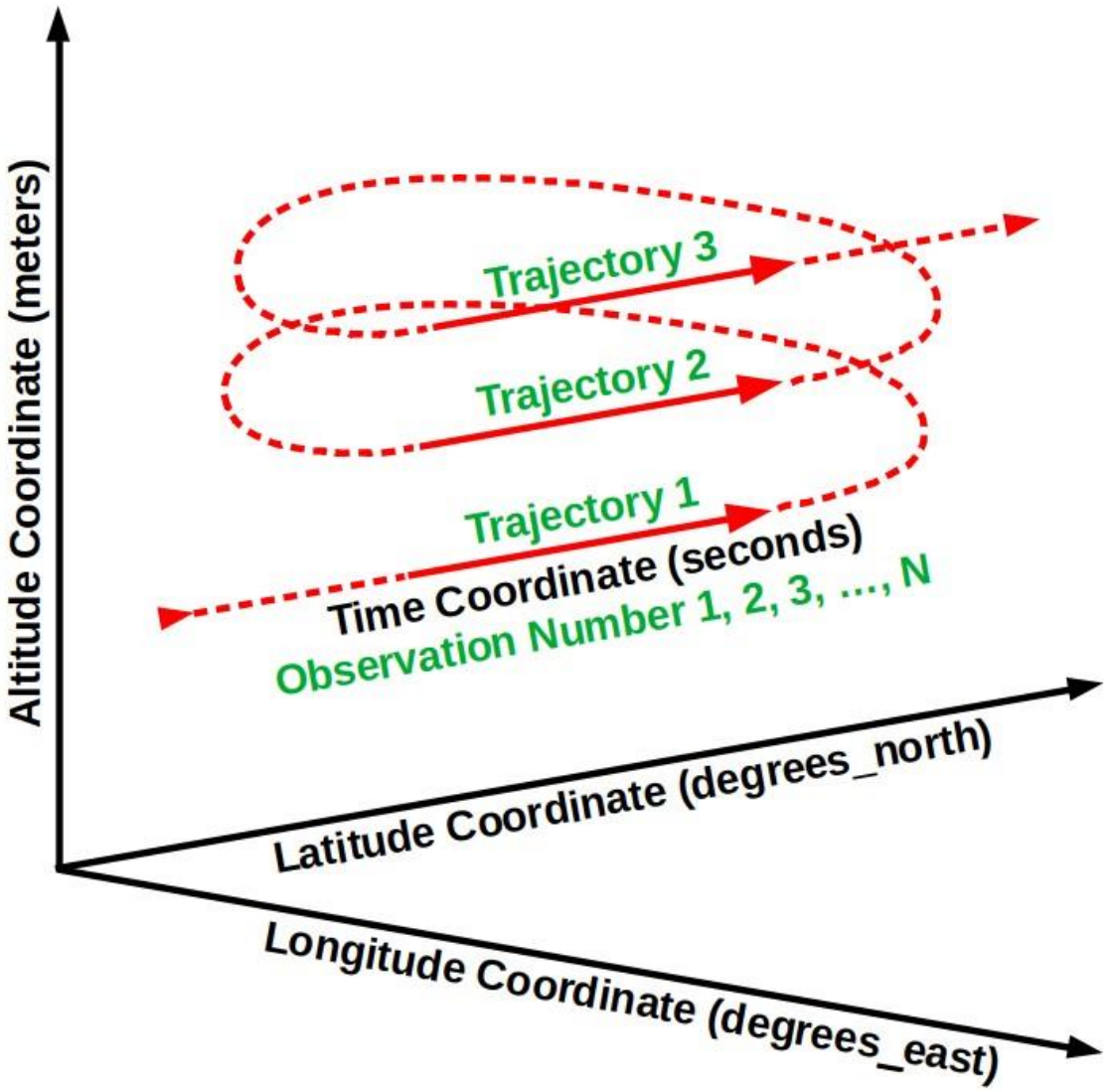

Figure 3. An example illustrating the dimensions and coordinates in a netCDF file for in situ data collected during an airborne campaign. Data acquisition occurs along the solid portion of the red flight lines, treated as three separate trajectories. The dimensions are Observation Number (the along-trajectory observation counter) and the Trajectory Number (the trajectory counter). The coordinates are Time, Longitude, Latitude and Altitude.

Data structures are containers for geolocation and science data. Guidance regarding swath structures in netCDF formats is provided in *Encoding of Swath Data in the CF Convention* [30]. The ESDSWG Dataset Interoperability Working Group (DIWG) has provided guidance regarding grid structures in netCDF-4 in [24] (Rec. 2.8-2.12) and [31] (Rec. 3.6). NOAA has provided a set of netCDF format templates for various types of data products [32] although these should be considered as informative, not normative. Data producers can obtain guidance and samples from their DAAC. Earthdata Search [33] can also be used to acquire a variety of data in different formats and structures.

3.1.2 GeoTIFF

The GeoTIFF (Georeferenced Tagged Image File Format, *.tif) format is a georeferenced raster image that uses the public domain Tagged Image File Format (TIFF) [34], and is used extensively in the Geographic Information System (GIS) [35] and Open Geospatial Consortium (OGC) communities [36]. Although the types of metadata that can be added to GeoTIFF files are much more limited than with netCDF-4 and HDF5, the OGC GeoTIFF Standards Working Group is planning to work on reference system metadata in the near term. Both data producers and users find this file format easy to visualize and analyze, and so it has many uses in Earth Science. OGC GeoTIFF Standard, Version 1.1 is an EOSDIS recommended format [37]. Recently, a cloud-optimized profile for GeoTIFF (called COG) has been developed to make retrieval of GeoTIFF data from Web Object Storage (see Appendix B) more efficient [38] [39]. Also, OGC has published a standard for COG [40] [41]. See [37] for a discussion of SWAL of GeoTIFF. DIWG has recommended that data producers include only one variable per GeoTIFF file [23].

3.2 Recognized Formats

In some cases, where the dominant user communities for a given data product have historically used other formats, it may be more appropriate to continue to use those formats instead of the formats recommended above. If such formats are not already on ESCO's list of approved data formats, they can be submitted to ESCO for review and approval following the Request for Comments instructions [42].

3.2.1 Text Formats

NASA DAACs archive numerous datasets that are in "plain text", typically encoded using the American Standard Code for Information Interchange (ASCII). Unicode, which is a superset of ASCII, is used to represent a much wider range of characters, including those used for languages other than English. The list of ESCO's approved standards using ASCII includes: International Consortium for Atmospheric Research on Transport and Transformation (ICARTT), NASA Aerogeophysics ASCII File Format Convention, SeaBASS Data File Format, and YAML Encoding ASCII Format for GRACE/GRACE-FO Mission Data. Recommendations on the use of ASCII formats are presented in the ASCII File Format Guidelines for Earth Science Data [43].

It should be noted that the comma-separated value (CSV) format is also a plain text format, as are Unidata's Common Data Language (CDL), JavaScript Object Notation (JSON), and markup languages such as HTML, XML and KML. The main advantages of encoding data in ASCII are that no special software is required to read a file and the contents are human readable. The main disadvantage is that file size, if not compressed, will be much larger than if the equivalent data were stored in a well-structured binary format such as netCDF-4, HDF5 or GeoTIFF. Another disadvantage of ASCII is that print-read consistency can be lost. Different programs reading a file could convert numerical values expressed in ASCII to slightly different floating-point numbers. This could complicate certain aspects of software engineering such as unit tests.

3.2.2 ICARTT

The ICARTT format [44] arose from a consensus established across the atmospheric chemistry community for visualization, exchange, and storage of aircraft instrument observations. The format is text-based and composed of a metadata section (e.g., data source, uncertainties, contact information, and brief overview of measurement technique), and a data section. Although it was primarily designed for airborne data, the format is also used for non-airborne field campaigns.

The simplicity of the ICARTT format allows files to be created and read with a single subprogram for multiple types of collection instruments and can assure interoperability between diverse user communities. Since typical ICARTT files are relatively small, the inefficiency of ASCII for storage is not a serious concern. See [44] for a discussion of SWAL of the ICARTT format.

3.2.3 Vector Data and Shapefiles

The OGC GeoPackage is a platform-independent and standards-based data format for geographic information systems implemented as an SQLite database container (*.gpkg) [45]. It can store vector features, tile matrix sets of imagery and raster maps at various scales, and extensions in a single file.

OGC has standardized the Keyhole Markup Language (KML, *.kml) format that was created by Keyhole, Inc. and is based on the eXtensible Markup Language (XML) [46]. The format delivers browse-level data (e.g., images) and small amounts of vector data (e.g., sensor paths, region boundaries, point locations), but it is voluminous for storing large data arrays. KML supports only geographic projection (i.e., evenly spaced longitude and latitude values), which can limit its usability. The format combines cartography with data geometry in a single file, which allows users flexibility to encode data and metadata in several different ways. However, this is a disadvantage to tool development and limits the ability of KML to serve as a long-term data format for archive. OpenGIS KML is an approved standard for use in EOSDIS. As noted in the recommendation, KML is primarily suited as a publishing format for the delivery of end-user visualization experiences. There are significant limitations to KML as a format for the delivery of data as an interchange format [47].

A Shapefile is a vector format for storing geometric location and attribute information of geographic features, and requires a minimum of three files to operate: the main file that stores the feature geometry (*.shp), the index file that stores the index of the feature geometry (*.shx), and the dBASE table that stores the attribute information of features (*.dbf) [48] [49]. Geographic features can be represented by points, lines, or polygons (areas). Geometries also support third and fourth dimensions as Z and M coordinates, for elevation and measure, respectively. Each of the component files is limited to 2 gigabytes. Shapefiles have several limitations that impact storage of scientific data. “For example, they cannot store null values, they round up numbers, they have poor support for Unicode character strings, they do not allow field names longer than 10 characters, and they cannot store both a date and time in a field” [50]. Additional limitations are listed in the cited article.

GeoJSON [51] is a format for encoding a variety of geographic features like Point, LineString, and Polygon. It is based on JSON and uses several types of JSON objects to represent these features, their properties, and their spatial extents.

3.2.4 HDF5

HDF5 is a widely-used data format designed to store and organize large amounts of data. NetCDF-4 (Section 3.1.1) and HDF-EOS5 (Section 3.2.4) are both built on HDF5. NetCDF-4 is the recommended format for new Earth Science data products as this format is generally more easily utilized by existing tools and services. However, as detailed in the Section 3.3.3, there are emerging strategies for enhancing HDF5 for improved S3 read access that represent important usage and performance considerations for Earth Science data distributed via the NASA Earthdata Cloud.

3.2.5 HDF-EOS5

HDF-EOS5 was a specially developed data format for the Earth Observing System based on HDF5, which has been widely used for NASA Earth Science data products and includes data structures specifically designed for Earth Science data.

HDF-EOS5 employs the HDF-EOS data model [52] [53], which remains valuable for developing Earth Science data products. The Science Data Production (SDP) Toolkit [52] and HDF-EOS5 library provide the API for creating HDF-EOS5 files that are compliant with the EOS data model.

In choosing between HDF-EOS5 and netCDF-4 with CF conventions, netCDF-4/CF is recommended over HDF-EOS5 due to the much larger set of tools supporting the format.

3.2.6 Legacy Formats

Legacy formats (e.g., netCDF-3, HDF4, HDF-EOS2, and ASCII) are those used in early EOS missions, though some missions continue to produce data products in these formats. Development of new data products or new versions of old products from early missions may continue to use the legacy format, but product developers are strongly encouraged to transition data to the netCDF-4 format for improved interoperability with data from recent missions. Legacy formats are recommended for use only in cases where the user community provides strong evidence that research will be hampered if the data formats are changed.

3.2.7 Other Formats

Some data products are provided by data producers in formats that are not endorsed by ESCO. These can include ASCII files with no header, simple binary files that are not self-describing, comma-separated value (CSV) files, proprietary instrument files, etc. Producers of such data are not necessarily NASA-funded, such as some participants in field campaigns; thus, they are not under any obligation to conform to NASA's format requirements or could lack adequate resources to do so.

There are other formats that are currently evolving in the community, stemming from developments in cloud computing, big data, Analysis-Ready Data (ARD) [54], that are discussed in Section 3.3.

3.3 Cloud-Optimized Formats and Services

Following the ESDS Program's strategic vision to develop and operate multiple components of NASA's EOSDIS in a commercial cloud environment, the ESDIS Project implemented the Earthdata Cloud architecture that went operational in July 2019, using Amazon Web Services (AWS) [55]. Key

EOSDIS services, such as CMR and Earthdata Search, were deployed within it. Additionally, the DAACs are moving the data archives they manage into the cloud.

The AWS Simple Storage Service (S3) offers scalable solutions to data storage and on-demand/scalable cloud computing, but also presents new challenges for designing data access, data containers, and tracking data provenance. AWS S3 is a popular example of object-based cloud storage, but the general characteristics noted in this document are applicable for object-based cloud storage from other providers as well. Cloud (object) storage is typically accessed through HTTP "range-get" requests in contrast to traditional disk reads, and so packaging the data into optimal independent "chunks" (see Section 5) is important for optimizing access and computation. Furthermore, the object store architecture allows the data to be distributed across multiple physical devices, in contrast to local contiguous storage for traditional data archives, with the data content organization often described in byte location metadata (either internally or in external "sidecar" files). Thus, many cloud storage "formats" are better characterized as (data) content organization schemes (see Appendix B), defined as any means for enhancing the addressing and access of elements contained in a digital object in the cloud.

Cloud "optimized" data containers or content organization schemes that are being developed to meet the emerging cloud compute needs and requirements, include Cloud-Optimized GeoTIFF (COG), Zarr, for-the-cloud versions of HDF5 and netCDF-4 (including NCZarr), and cloud-optimized point-cloud data formats (see also [56] for additional background). COG, Zarr, HDF5 and netCDF-4 (see Sections 3.3.1, 3.3.2, and 3.3.3, respectively) continue to remain preferred formats for raster data while lidar and point-based irregular data are more appropriate for point cloud formats (see Section 3.3.4). These cloud storage optimizations, although described in well-defined specifications, are still advancing and growing in maturity with regard to their use and adaptations in cloud-based workflows, third party software support, and web services (e.g., OPeNDAP, THREDDS, OGC WCPS). However, none of these formats require in-cloud processing for scientific analysis and can work with local computer operating systems and libraries without issue once the data have been downloaded. Analysis-Ready, Cloud-Optimized (ARCO) data, where the cloud data has been prepared with complete self-describing metadata following a standard or best practice, including the necessary quality and provenance information, and well-defined spatial and temporal coordinate systems and variables, offer a significant advantage for reproducible science, computation optimization, and cost reduction.

Data producers should carefully optimize their data products for partial data reads (via HTTP or direct S3 access) to make them as cloud friendly as possible. This requires organizing the data into appropriate producer-defined chunk sizes to facilitate access. The best guidance thus far is that S3 reads are optimized in the 8-16-megabyte (MB) range [57] presenting a reasonable range of chunk sizes. The Pangeo Project [58] reported chunk sizes ranging from 10-200 MB when reading Zarr data stored in the cloud using Dask [59] and the desired chunking often depends on the likely access pattern (e.g., chunking in small Regions of Interest (ROIs) for long time series data requests vs. chunking in larger ROI slices for large spatial requests over a smaller temporal range). However, on the other end of the spectrum, chunks that are too small, on the order of a few megabytes, typically impede read performance in the cloud. Data producers are advised to consult with their assigned DAAC regarding the specific approaches to their products including the chunking implementation.

3.3.1 Cloud-Optimized GeoTIFF

The COG data format builds on the established GeoTIFF format by adding features needed to optimize data use in a cloud-based environment [39] [40]. The primary addition is that internal tiling (i.e., chunking) for each layer is enabled. The tiling features enable data reads to access only the information of interest without reading the whole file. Since COG is compatible with the legacy GeoTIFF format it can be accessed using existing software (e.g., GIS software).

3.3.2 Zarr

Zarr is an emerging open-source format that stresses efficient storage of multidimensional array data in the cloud and fast parallel input/output (I/O) computations [60] [61]. Its data model supports compressed and chunked N-dimensional data arrays, inspired in part by the HDF5 and netCDF-4 data models. Its consolidated metadata can include a subset of the CF metadata conventions that are familiar to existing users of netCDF-4 and HDF5 files, and allow for many useful time-series and transformation operations through third party libraries such as xarray [62]. Zarr stores chunks of data as separate objects in cloud storage with an external consolidated JSON metadata file containing all the locations to these data chunks. A Zarr software reader (e.g., using xarray in python) only needs a single read of the contents of the consolidated metadata file (i.e., the sidecar file) to determine exactly where in the Zarr data store to locate data of interest, substantially reducing file I/O overhead and improving efficiency for parallel CPU access.

3.3.3 NetCDF-4 and HDF5 in the Cloud

Much of NASA Earth Science data has been historically stored in netCDF-4 and HDF5 data files. Besides maintaining continuity to legacy data products, there are other important data lifecycle reasons to continue to use these formats, including data packaging, data integrity and self-describing characteristics. The challenge is how to best optimize the individual files for cloud storage and access. Here, data chunking plays a leading role in this optimization with the general guidelines on this subject found in the introduction to Section 3.3. It has been demonstrated that it is possible to translate the annotated Dataset Metadata Response (DMR++) [63] sidecar files that are generated for many of NASA's HDF5 files that have been migrated to the Earthdata Cloud into a JSON file with the key/value pairs that the Zarr library needs [64], making the HDF5 directly readable as Zarr stores.

Further cloud optimization of HDF5 files, specifically, requires enhancing the internal metadata HDF structure via the "Paged Aggregation" feature at the time of file creation (or modification via *h5repack*), so that the internal file metadata (i.e., not the global metadata) and data are organized into a single or a few pages of specified size (usually on the order of Mebibytes) to improve read file I/O. The exact size is important for parallel I/O operations in the cloud, and other HDF libraries that can cache the pages, further improving performance.

NCZarr is an extension and mapping of the netCDF-enhanced data model to a variant of the Zarr storage model (see Section 3.3.2).

Additional discussion of cloud optimization of netCDF-4 and HDF5 files via data transformation services is provided in Section 3.3.5.

3.3.4 Point Cloud Formats

A point cloud is commonly defined as a 3D representation of the external surfaces of objects within some field of view, with each point having a set of X, Y and Z coordinates. Point cloud data have traditionally been associated with lidar scanners such as on aircraft; in addition, *in situ* sensors such as those mounted on ocean gliders and airborne platforms can also be considered as point cloud data sources. The key characteristic is that these instruments produce a large number of observations that are irregularly distributed and thus are "clouds" of points.

There are many emerging formats in this evolving genre [65]. Some noteworthy formats include Cloud-Optimized Point Cloud (COPC), Entwine Point Tiles (EPT) and Parquet. COPC builds on existing point cloud formats popular in the lidar community known as LAS/LAZ (LASer file format/LAS compressed file format) and specifications from EPT. EPT itself is an open-source content organization scheme and library for point cloud data that is completely lossless and uses octree-based storage format and contains metadata in JSON. Parquet is a column-based data storage format that is suitable for tabular style data (including point cloud and *in situ* data). Its design lends itself to efficient queries and data access in the cloud. GeoParquet is an extension that adds interoperable geospatial types such as Point, Line and Polygon to Parquet [66].

3.3.5 Additional Data Transformation and Access Services

For data analysis in the cloud, it is often preferred to optimize data for parallel I/O and multi-dimensional analysis. This is where Zarr excels and a number of transformation services from netCDF-4 and HDF5 files to Zarr have emerged to support this need. Traditional file-level data access and subsetting via the OPeNDAP web service has also evolved to meet the needs of cloud storage. Many of these tools enable Zarr-like parallel and chunked access capabilities to be applied onto traditional netCDF-4 and HDF5 files in AWS S3. While these services are not critical for producing data products, it is important for data producers to be aware of their use by Earth Science data consumers.

3.3.5.1 Harmony Services

The name "Harmony" refers to a set of evolving open source, enterprise-level transformation services for data residing in the NASA Earthdata Cloud [67]. These services are accessed via a well-defined and open API, and include services for data conversion and subsetting.

Harmony-netcdf-to-zarr [68] is a service to transform netCDF-4 files to Zarr cloud storage on the fly. It aggregates individual input files into a single Zarr output that can be read using xarray calls in Python. As additional files become available, this service must be rerun to account for the new data.

Subsetting requests for trajectory (1D) and along track/across track data in netCDF and HDF files are executed using the harmony L2-subsetter service while geographically gridded Level 3 or 4 data use the Harmony OPeNDAP SubSetter service (HOSS). The Harmony-Geospatial Data Abstraction Library (GDAL)-adapter service supports reprojection.

3.3.5.2 Kerchunk

Kerchunk is a Python library to generate a “virtual” Zarr store from individual netCDF-4 and HDF5 files by creating an external metadata JSON sidecar file that contains all the locations to the individual input data chunks [69]. The “virtual” Zarr store can be read using xarray and the original netCDF-4 and HDF5 files remain unmodified in content and location. As Kerchunk leverages the fsspec library for storage backend access, it enables end users to more efficiently access parallel chunks from cloud-based S3, as well as other remote access such as data over Secure Shell (SSH) or Server Message Block (SMB).

3.3.5.3 OPeNDAP in the Cloud

The OPeNDAP Hyrax server is optimized to address the contents of netCDF-4 and HDF5 files stored in the cloud using information in the annotated DMR++ [63] sidecar file. The DMR++ file for a specific data file encodes chunk locations and byte offsets so access and parallel reads to specific parts of the file is optimized.

4 METADATA

4.1 Overview

Metadata are information about data. Metadata could be included in a data file and/or could be external to the data file. In the latter case there should be a clear connection between the metadata and the data file. As with the other aspects of data product development, it is helpful to consider the purpose of metadata in the context of how users will interact with the data and how metadata are associated with (i.e., structurally linked to) the data.

Metadata are essential for data management: they describe where and how data are produced, stored, and retrieved. Metadata are also essential for data search/discovery and interpretation, including facilitating the users’ understanding of data quality. A data producer has a responsibility to provide adequate metadata describing the data product at both the product level and the file level. The DAAC that archives the data product is responsible for maintaining the product-level metadata (known as collection metadata in the CMR [17]), which is a high-performance, high-quality, continuously-evolving metadata system that catalogs all data and service metadata records for EOSDIS. These metadata records are registered, modified, discovered, and accessed through programmatic interfaces leveraging standard protocols and APIs.

The ESDIS Project employs tools to interact with CMR based on the Unified Metadata Model (UMM). Profiles have been defined within the UMM based on their function or content description, such as Collection, Service, Variable, or Tool as shown in Table 1.

Table 1. UMM-Model descriptions (see Appendix B for definition of "granule")
https://www.earthdata.nasa.gov/unified-metadata-model-umm

EOSDIS Concepts	UMM Profiles	Profile Short Name
Collections	Collection	UMM-C
Granules	Granule	UMM-G
Services	Service	UMM-S
Variables	Variable	UMM-Var
Visualizations	Visualization	UMM-Vis
Tools	Tools	UMM-T
Elements Common to multiple UMM Component Models	Common	UMM-Common

Data product- and file-level metadata are stipulated by the policies of the DAAC that will host the data, and subject to the somewhat minimal requirements that the CMR imposes. However, EOSDIS data are expected to exceed these in order to render the data more searchable and usable. Metadata that are rich in content enable the creation of the UMM-Var, UMM-Service, or UMM-Tool records listed in Table 1. These tools and services could include subsetting of the data by variable, time, or location, and/or might enable the creation of a time series, for example. Other metadata provide information about the provenance of the data, product quality, the names of scientists or teams that created the product and the funding sources that supported the creation of the products. Data producers should work with the DAACs to determine the best approach for their products.

Metadata can be created for and associated with a data product through several methods that are different from the metadata used by the UMM/CMR system. Software libraries, such as netCDF and HDF, make populating file-level metadata straightforward. Metadata can be assigned to any object within the file (i.e., variable-level attributes), or to the file as a whole (i.e., global attributes). As such, most of the recommendations that follow apply to netCDF/HDF files, but a producer of products in other formats should aspire to conform to the degree that those formats permit. A further discussion of file-level metadata is provided in Appendices D and E.

Global attributes are meant to apply to all information in the file, and can vary from file to file within a data product. However, to maximize the self-describing nature of a file, a data producer can also include product-level metadata (i.e., information that is identical for all files) within each file. File-level metadata should be embedded in the file itself if using self-describing formats like netCDF. The DAACs may require that the metadata be provided both embedded in files and as a separate metadata file. The assigned DAAC will ensure that the physically separate metadata are properly associated with the data file to which they refer. Data product files may not contain all the available data product metadata (e.g., may not contain everything in related texts such as the Algorithm Theoretical Basis Document (ATBD)), but they must contain enough metadata to enable data search and discovery and scientific analysis using tools capable of recognizing metadata standardized for interoperability based on recommendations and standards in this document and provided by the DAACs.

4.1.1 Data Product Search and Discovery

Data product users likely experience the impact of metadata for the first time during the search and discovery process—when they are searching for data products that meet their needs. As mentioned above, the metadata that support this process are typically ingested into the CMR by the DAACs. These metadata are used either by the CMR search engine or by other search engines that harvest metadata (e.g., data.gov or Google).

Key success criteria for metadata during the discovery process include:

- Intelligible, descriptive data product names (Section 4.2).
- Precise temporal and spatial coverage (Sections 4.4 and 4.5).
- Accurate and complete list of applicable Global Change Master Directory (GCMD) Science Keywords [70].
- Concise but readable (including machine readable) description of the data product.

Given the expectation to find a specific product among thousands of data products in the archives (over 52,750 as of February 14, 2024 in the EOSDIS catalog, the CMR), it is crucial to use GCMD keywords [70], especially for the platform (e.g., Nimbus-7 or DC-8), instrument (e.g., TOMS or HIRAD), and science (e.g., OZONE) keywords. For airborne and field campaign data it is important to include the campaign or project short name or acronym (e.g., Delta-X, AboVE, or EXPORTS). Reference [71] provides links through which various categories of keywords can be downloaded in a variety of formats such as comma-separated values (CSV) or can be directly viewed with the GCMD Keyword Viewer [72]. When the keywords in the current list of GCMD are not directly applicable to a data product, the data producers are advised to follow the proper GCMD list update procedure [73] in consultation with their assigned DAACs. Data producers should work closely with their assigned DAACs in obtaining and selecting keywords for their products.

4.1.2 File Search and Retrieval

Once a user has identified a data product to pursue, the user typically needs only some files, and not all of the files, for the data product. When metadata are standardized, data search engines, such as Earthdata Search for CMR [33], support the specification of spatial and temporal criteria (i.e., search filters). Therefore, it is best to precisely specify the spatial extent of a given file to limit “false positives” in search results. For example, a four-point polygon provides a more precise specification of spatial extent than a bounding box (Figure 4). Data producers should consult with their assigned DAACs regarding methods the DAACs are using to specify bounding regions before deciding whether a different approach is needed for their products.

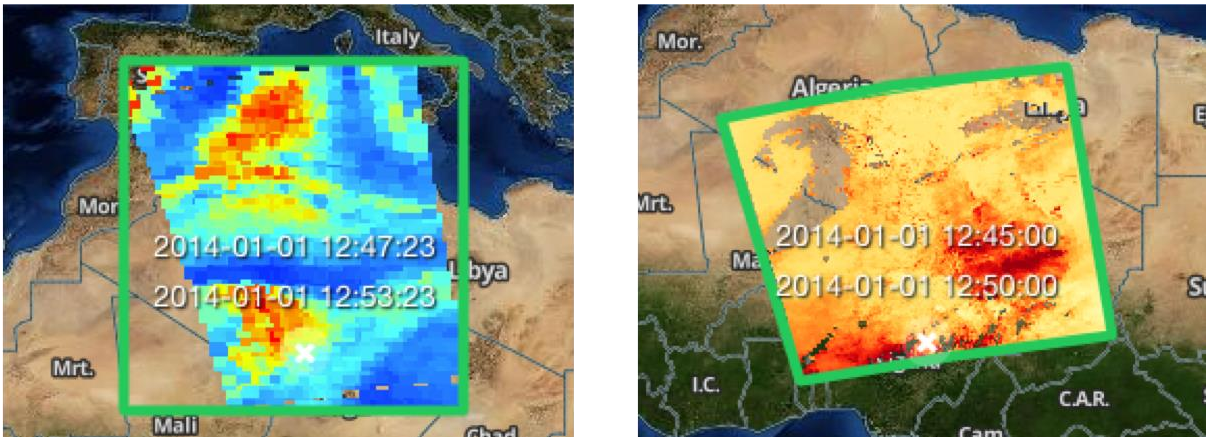

*Figure 4. Screenshots from Earthdata Search [33] for scenes from two Level 2 satellite data products. **Left**: The green box represents the spatial extent of the file, as specified in the metadata, which results in large areas of "no data" in the southwest and northeast corners, where a user selection would generate a false positive result. **Right**: True data coverage specified with a four-point polygon, filled with a browse image, showing negligible areas where a user selection could produce a false positive.*

4.1.3 Data Usage

High-quality metadata are essential for data readability, not only for human users, but also for software, such as decision support systems or models. The aim should be for data producers to generate files that are usable without further adjustments by a data product user—a benchmark known as ARD (e.g., [54]), whose metadata typically includes:

- Coordinate information: spatial and temporal coordinates in standard form.
- Interpretation aids: units [31] (Rec. 3.2 and 3.3), fill (missing) value identification [31] (Rec. 3.7), offset and scale factors [24] (Rec. 2.2, 2.5 and 2.6), data uncertainty.
- Human-readable and scientifically meaningful/standard variable names (see Section 4.1.4).

In addition, data producers should be mindful of other types of potentially useful documentation to facilitate the understanding of data, including but not limited to provenance information (see Section 4.6.1), particularly identifiers of the data inputs and algorithm version, and pointers to documentation such as ATBDs, and data quality assessment(s). (Some of this information is sometimes found in data producers' README files as well). For field campaigns, this may include instrument placement and operation documents, such as images of instrument placement on an aircraft. Such documentation should be produced as early as possible in the dataset production lifecycle.

Since the primary purpose of CMR is to support search and discovery, at present, not all the data usage information is ingested into CMR. However, as automated data system capabilities evolve to provide higher levels of service for new data products, we expect CMR to include and handle more data-usage information.

4.2 Naming Data Products, Files, and Variables

4.2.1 Data Products

The name of a data product is critical to its discovery in Earthdata Search and tools developed by the DAACs. The DAACs have internal rules and guidelines for naming data products, so selection of long and short names should be a joint effort between the data producer and the assigned DAAC. DAAC involvement in naming also helps in making the data products discoverable and unique within EOSDIS. Data products must be assigned both a long name, which is meant to be human readable (and comprehensible), and a short name, which aids precise searching by keyword. The data product long name and short name are considered global attributes in that they are not associated with any particular variable but are related to the data product as a whole.

4.2.1.1 Long Name

The data product's long name (called `LongName`[4] in CMR) is a name that is scientifically descriptive of the product. It should be as brief (but as complete) as possible, as it expands on the corresponding short name, which can sometimes be indecipherable. The attribute `LongName` is synonymous with the CF `title` attribute used in netCDF documentation (see Appendix D.1).

Data producers should seek a long name that will be understandable to the target user community and is also unique within EOSDIS. A reasonable data product name may already be in use (e.g., beginning with "MODIS/Aqua" or "MODIS/Terra"), and care should be taken to avoid naming conflicts by consultation with the assigned DAAC and other relevant data producers. Interoperability should also be considered when choosing names (see Section 4.2.2).

We provide the following recommendations regarding the formulation of data product long names:

- The data source, usually the acronym/abbreviation for the project responsible for producing the data, but can also include the instrument (e.g., HIRDLS, the High Resolution Dynamics Limb Sounder; AVIRIS, the Airborne Visible / Infrared Imaging Spectrometer), platform (satellite, aircraft, ship, etc.) (e.g., TRMM, the Tropical Rainfall Measuring Mission; P-3, a NASA aircraft), or program (e.g., MEaSUREs, **M**aking **Ea**rth **S**ystem Data Records for **U**se in **R**esearch **E**nvironment**s**); include both the instrument and platform names to eliminate ambiguity (e.g., MODIS/Aqua; HIRAD/ER-2). For campaigns, it may be important to include the campaign name along with instrument and platform (e.g., OLYMPEX/HIWRAP/ER-2).
- Science content (e.g., Aerosol, Precipitation Rate, Total Column Water Vapor).
- The general spatial coverage type of the data (e.g., gridded, swath, orbit, point).
- The temporal coverage per file (e.g., instantaneous, 5-minute, orbital, daily, monthly).
- Processing level (e.g., L2 or Level 2) [74].
- Spatial resolution (e.g., 3 km; if there are multiple resolutions in the product, then the highest resolution is typically stated).

[4] Not to be confused with the CF `long_name` attribute for individual variables.

- Version number (optional; details should be resolved with the assigned DAAC).

Examples of data products that follow this naming convention are provided in Table 2. Note that not all the contents suggested in the above bullets are included in names shown in each of the examples.

4.2.1.2 Short Name

EOSDIS developed the standard Earth Science Data Type (ESDT) [75] naming conventions to provide an efficient way to archive data products by name and for convenience in working with toolkits. The short name (called `ShortName` in CMR) is an abbreviated version of the product name. In short names, alphanumeric and underscore ("_") are the only acceptable characters. The restriction on the usage of spaces and special characters is to ensure compatibility with Earthdata Search and other search systems. The short name is included in the metadata as the global attribute `ShortName` and included in the data product's documentation. Data producers should contact the DAAC responsible for archiving the data product to check if there are additional restrictions on short names, such as consistency across systems.

Table 2. Examples of data products named using some of the recommended naming conventions. Deviations from this format are sometimes necessary owing to the characteristics of specific data products.

Short Name	Long Name	Comments
OCO2_L2_Met	OCO-2 Level 2 meteorological parameters interpolated from global assimilation model for each sounding	Includes: data source (OCO-2), scientific content (meteorological parameters) spatial coverage (global), processing level (Level 2). Missing: spatial resolution, temporal coverage, and version number
MYD09GQ	MODIS/Aqua Near Real Time (NRT) Surface Reflectance Daily L2G Global 250m SIN Grid	Includes: data source (MODIS/Aqua), scientific content (surface reflectance), spatial coverage (global), temporal coverage (NRT), processing level (L2G), spatial resolution (250m). Missing: version number
MOD05_L2	MODIS/Terra Total Precipitable Water Vapor 5-Min L2 Swath 1 km and 5 km – NRT	Includes: data source (MODIS/Terra), scientific content (Total Precipitable Water Vapor), spatial coverage (swath), Temporal coverage, (5 min; NRT) processing level (L2), Spatial resolution (1km and 5km) Missing: version number

Short Name	Long Name	Comments
GPM_PRL1KU	GPM DPR Ku-band Received Power L1B 1.5 hours 5 km	Includes: data source (GPM DPR Ku-band), scientific content (Received Power), temporal coverage (1.5 hours), processing level (L1B), spatial resolution (5 km), Missing: spatial coverage, version number
GLDAS_NOAH10_M	GLDAS Noah Land Surface Model L4 Monthly 1.0 x 1.0 degree	Includes: data source (GLDAS), scientific content (Noah Land Surface Model), temporal coverage (Monthly), processing level (L4), spatial resolution (1.0 x 1.0 degree) Missing: spatial coverage, version number
SWDB_L3M10	SeaWiFS Deep Blue Aerosol Optical Depth and Angstrom Exponent Monthly Level 3 Data Gridded at 1.0 Degrees	Includes: data source (SeaWiFS), scientific content (Deep Blue Aerosol Optical Depth and Angstrom Exponent), spatial coverage, temporal coverage (Monthly), processing level (Level 3), spatial resolution (Gridded at 1.0 Degrees) Missing: spatial coverage, version number
AIRX2RET	AIRS/Aqua L2 Standard Physical Retrieval (AIRS+AMSU) V006 (AIRX2RET) at GES DISC	Includes: data source (AIRS/Aqua; AIRS+AMSU), scientific content (Standard Physical Retrieval), processing level (L2), version number (V006) Missing: spatial coverage, temporal coverage, spatial resolution
M2I3NPASM	MERRA-2 inst3_3d_asm_Np: 3d,3-Hourly,Instantaneous,Pressure-Level,Assimilation,Assimilated Meteorological Fields 0.625 x 0.5 degree V5.12.4	Includes: data source (MERRA2), scientific content (Pressure-Level, Assimilation, Assimilated Meteorological Fields), spatial coverage (3d), temporal coverage (3-Hourly, Instantaneous), spatial resolution (0.625 x 0.5 degree), version number (V5.12.4) Missing: processing level
ATLAS_VEG_PLOTS_1541	Arctic Vegetation Plots ATLAS Project North Slope and Seward Peninsula, AK, 1998-2000	Includes: data source (ATLAS), location (north slope and peninsula, AK), and temporal coverage (1998 - 2000) Missing: processing level, spatial and temporal resolution, version number

Short Name	Long Name	Comments
CARVE_L1_FTS_SPECTRA_1426	CARVE: L1 Spectral Radiance from Airborne FTS Alaska, 2012-2015	Includes: campaign (CARVE), instrument (FTS), data product level (L1), region of study (Alaska), and temporal coverage (2012-2015) Missing: spatial and temporal resolution, version number
DISCOVERAQ_Texas_AircraftRemoteSensing_B200_GCAS_Data	DISCOVER-AQ Texas Deployment B-200 Aircraft Remotely Sensed GCAS Data	Includes: data source (campaign - DISCOVER-AQ, deployment - Texas, platform - B-200, and instrument - GCAS), region of study (Texas) Missing: spatial and temporal coverage and resolution, version number, processing level
AirMOSS_L2_Precipitation_1417	AirMOSS: L2 Hourly Precipitation at AirMOSS Sites, 2011-2015	Includes: campaign (AirMOSS), data level (L2), temporal resolution (hourly) and temporal coverage (2011-2015), variable (precipitation) Missing: data source (rain gauge), spatial coverage and resolution, region, version number
DeltaX_Sonar_Bathymetry_2085	Delta-X: Sonar Bathymetry Survey of Channels, MRD, Louisiana, 2021	Includes: campaign (Delta-X), data source (Sonar), region (Mississippi River Delta), variable (bathymetry), temporal coverage 2021 Missing: version number, processing level, temporal and spatial resolution

4.2.2 Files

There is no universal file-naming convention for NASA Earth Science data products, apart from the DIWG recommendations regarding the components of file names provided in [31] (Rec . 3.8-3.11). However, file names should be unique and understandable to both humans and machines as well as contain information that is descriptive of the contents.

The date-time information in the file names should adhere to the following guidelines (detailed in [31], Rec. 3.11):

- Adopt the ISO 8601 standard [76] [77] for date-time information representation.
- The start time should appear before the end time in the file name.
- Date-time fields representing the temporal extent of a file's data should appear before any other date-time field in the file name.
- All date-time fields in the file name should have the same format.

4.2.3 Variables

A fundamental consideration for variable names is that care should be exercised in the use of special characters or spaces in the name to ensure that they are readable and interpretable by commonly used software. Also, to promote usability and human readability the names should be meaningful. This may include community best-practice names including modified standard names such as the CF Standard Names [78] . An example of a community approach to construction of variable names, used for ICARTT files (see Section 3.2.1), is contained in the Atmospheric Composition Variable Standard Name Convention document [79]].

To optimize discovery and usability, variable names should comply, with community best-practice names and endorsed standard names via the variable-level CF `long_name` and `standard_name` attributes, respectively (see Appendix E).

4.3 Versions

Data products are uniquely identified by the combination of `ShortName` and `VersionID` within the CMR [17]. Data producers can also specify a `product_version` as an Attribute Convention for Data Discovery (ACDD) global attribute [80] . In most cases, the `product_version` and `VersionID` are identical although there may be some exceptions to this (e.g., when selected files associated with a limited reprocessing of data have a different value for the `product_version`). However, if reprocessed data files have significant differences in terms of science content, then these files should be organized into a separate data product with a different `VersionID`. Guidance for setting version numbers should be sought from the assigned DAAC.

The software version used to generate a data product is specified via the CMR attribute `PGEVersion`. In most cases, the `product_version` and the `PGEVersion` differ.

4.4 Representing Coordinates

Earth Science data files should be produced with complete information for all geospatial coordinates to help enable software application capabilities for data visualization, mapping, reprojection, and transformation. Encoding geolocation based on the CF Conventions maximizes the ability to use the data in multiple tools and services. Please note that coordinates should not be confused with dimensions - sometimes these two things are one and the same, but this is not always the case, as explained in Section 3.1.1 (see Figures 2 and 3).

4.4.1 Spatial Coordinates

Variables representing latitude and longitude must always explicitly include the CF `units` attribute, because there are no default values for the units of latitude and longitude. The recommended unit of latitude is degrees_north (valid range: -90 to 90 degrees) and unit of longitude is degrees_east (valid range: -180 to 180 degrees).

Consider the spatial accuracy required to represent the position of data in the file when choosing the latitude and longitude variable datatypes. Use double precision for latitude and longitude datatypes if meter-scale geolocation of the data is required.

To support the widest range of software tools while avoiding storage of redundant geospatial coordinate data, practice the following guidelines:

- Specify coordinate boundaries by adding the CF `bounds` attribute [24] (Rec. 2.3).
    - For example, a producer can annotate the “latitude" coordinate with the CF `bounds` attribute with value “latitude_bnds.” The “latitude_bnds” variable would be a multi-dimensional array of the intervals for each value of variable “latitude.”
- Include horizontal attributes and, as necessary, vertical attributes.
    - For example, a producer can include the CF attribute `units`: degrees_north for the latitude coordinate variable; degrees_east for the longitude coordinate variable; and "m" for a height coordinate variable.
- Store all coordinate data for a single file in coordinate variables only. No coordinate data, or any part thereof, should be stored in attributes, or as variable and/or group names [31] (Rec. 3.5).
- Files are required to contain the most applicable type of geospatial coordinates for the data. The decision whether to provide any additional types of geospatial coordinates is left to the data producer.
- A grid mapping variable named “crs” (Coordinate Reference System) can provide the projection and datum information via its attached attributes. The projection information is specified by attaching to the "crs" variable the CF `grid_mapping_name` attribute and other CF attributes. The `crs_wkt` attribute can be attached to the "crs" variable to provide the datum information via OGC Well-Known Text (WKT). Other variables in the file can make use of this projection and datum information by attaching the CF `grid_mapping` attribute with value "crs". See [31] (Rec. 3.6) for three realistic implementations of the "crs" variable.
- Specify both the horizontal (geodetic) datum, which is typically WGS84, and the vertical datum, which may be either an ellipsoid such as WGS84, or a local datum which would yield orthometric heights.
- The latitude and longitude coordinate variables can have the CF `axis` attribute attached with values Y and X, respectively. Grid products that do not explicitly include variables named latitude and longitude can also include the CF `axis` attribute for the horizontal coordinate variables.

Note that different data user communities prefer different ordering of the latitude and longitude in the data files. Data producers should target the dominant user community for their products to decide upon the order, indicate clearly what the order is, and use self-describing file formats [23].

4.4.2 Temporal Coordinate

The CF Conventions represent time as an integer or float, with the `units` attribute set to the time unit since an epochal date-time, represented as YYYY-MM-DDThh:mm:ss (e.g., "seconds since 1993-01-01T00:00:00Z"). Use Coordinated Universal Time (UTC) instead of local time unless there is a strong justification for not doing so.

For gridded observations (e.g., Level 2G or Level 3), data can be aggregated using the time coordinate axis to record the different time steps. For example, this technique can be used to aggregate sub-daily observations (e.g., hourly, 4-hourly) into a single daily file.

When files contain only a single time slice, a time axis coordinate vector variable of size 1 should be included, so that the time information is easy to locate and the degenerate (i.e., one value of time for the entire file) time axis can improve performance when aggregating additional files over time (see [24], Rec. 2.9; [31], Rec. 3.4). The time coordinate variable can have the CF `axis` attribute attached with value T.

Just as for the latitude and longitude coordinate variables (Section 4.4.1), temporal boundaries for each value of the time coordinate variable can be specified by including a time bounds variable (e.g., time_bnds), which is named via the value of the CF `bounds` attribute attached to the time variable.

4.4.3 Vertical

Some data have a vertical dimension, and, therefore, a variable should be included to describe the vertical axis. The most commonly used values to describe vertical coordinates are layer, level, pressure, height, and depth. It is important to identify the vertical coordinates using the most common standard terminology, and to include the following information:

- `long_name` This CF attribute can be something as simple as "vertical level" and can also be used to clarify the CF `units` attribute. The valid values for the CF `units` attribute are provided by the Unidata units library (UDUNITS) package [82] and include units of pressure, length, temperature, and density. In general, the CF `units` attribute should not be included for variables that do not have physical units [31] (Rec. 3.3). Also, we recommend adding the CF `standard_name` attribute to describe the coordinate variable.
- `positive` This CF attribute refers to the direction of the increasing coordinate values, with a valid value of either `up` or `down`. If the vertical coordinate follows units of pressure, then this attribute is not required. Variables representing dimensional height or depth axes must always explicitly include the CF `units` and `positive` attributes because there is no default value.
- `axis` Setting the value of this CF attribute to Z indicates that a coordinate variable is associated with the vertical axis.

4.5 Data Quality

It is essential that users of scientific data products have access to complete (to the degree to which all knowledge is available at the time) and properly articulated (i.e., correctly described for the user to logically understand, discern, and make well-informed decisions) information regarding the data

quality, including known issues and limitations. This information will help to inform users about the potential applications of the data products and prevent data misuse. Therefore, data products should include metadata pointing to public documentation of the processes used for assessing data quality. Also, data producers should supply the documentation to the DAACs for archiving and distribution. Data producers can work with the DAACs and review boards to provide data quality information through an existing community-standardized format for describing quality (e.g., the data quality metadata model found in GHRSST [11]). For example, data quality information can be provided as a part of a data product user guide. Data quality information can also be included in file-level metadata and/or as data quality layers.

The recommended contents for capturing and ensuring data quality are provided below. More detailed explanations on these, along with examples, are found in the Data Management Plan Template for Data Producers [83]. In the discussion below, we use the term documentation to refer to somewhat extensive information that is typically stored separately from data files, with the metadata in the files including pointers (URLs) to such information.

4.5.1 Quality Information in Data Product Documentation

This subsection provides guidance regarding the information that should be covered in documents considered too extensive to be contained within the data files themselves.

1. Document the process used, including data flows and organizations involved in assuring data quality. Provide the references to the ICDs between organizations that have been or will be developed. If the ICD does not exist or is a work in progress, include the names and email addresses of the lead authors responsible for drafting the ICD. See [84] for an example of an ICD. In the cases (e.g., airborne investigations and field campaigns) where formal ICDs are not produced, provide references to Data Management Plans where data quality procedures are described.
2. Provide documentation of the calibration/validation (Cal/Val, see Appendix B) approach used, including sources of data, duration of the process, the targeted uncertainty budget that was used to assess performance, and the conditions under which Cal/Val are conducted. As Cal/Val data sources change or are reprocessed, ensure that the information is kept up to date in a publicly accessible location with reference to the relevant geospatial and temporal coverage information that is directly applicable to those Cal/Val data products.
3. Provide a description of how quality flags or indicators (see Appendix B) are used in the product and explain their meanings. The following are general considerations regarding quality flags and indicators:
    a. Define and create indicators to represent the quality of a data product from different aspects (e.g., data dropout rate of a sea surface temperature data product).
    b. Ensure that quality flags are related to a quantifiable metric that directly relates to the usefulness, validity, and suitability of the data.
    c. Identify quantifiable data quality criteria, such as confidence levels and the values of quality flags, which can be used as criteria for refining search queries.

d. Provide ancillary quality and uncertainty flags to facilitate detection of areas that are likely to contain spurious data (e.g., ice in unexpected places).

e. Provide pixel-level (or measurement-level) uncertainty information where possible and meaningful. Provide the confidence level (e.g., 95%) to indicate the statistical significance.

f. Provide data quality variables and metadata along with detailed documentation on how the metadata are derived and suggestions on how to interpret them or use them in different applications.

g. Provide definition and description of each data quality indicator, including the algorithms and data products used to derive the quality information and description of how each quality indicator can be used.

h. Provide examples of idealized quality flag/indicator combinations that would likely yield optimal quality filtering (i.e., minimized bias, uncertainty, and spurious observations) for science in a particular domain of research.

4. A quality summary should also be documented and disseminated whenever a new dataset or a new version of a dataset is published. The quality summary should at least be a high-level overview of strengths and limitations of the dataset and should be directly traceable and reproducible by the variables within the dataset, such as by referencing the quality flags and indicators used to derive the summary. For example, the quality summary may describe the overall percentage of data that are either missing from the dataset (due to pre-processing Quality Assessment/Quality Control (QA/QC)) or that may be optionally discarded (at the discretion of the data user) due to quality conditions that are expressed by the quality flags and indicators.

5. Provide documentation of the methods used for estimating uncertainty and how uncertainty estimates are included in the data product. Provide documentation of known issues and caveats for users and consider leveraging DAAC resources for more expedient updating and publication of this information (e.g., forums, web announcements). Also include citations and references to the data used in the validation process.

4.5.2 Quality Information in Product Metadata

This subsection provides guidance regarding the quality information that should be provided via the metadata within the data files themselves, if possible.

1. Include the uncertainties in the delivered data, with the level of detail dependent on the size of the uncertainty information. For example, provide uncertainty expressed per data (pixel) value, per file, or per data product.

2. Provide pointers (URLs or citations) to the ancillary data products that are used for quality assessments, Cal/Val, validation of uncertainty budget, as well as quantification and characterization of uncertainty.

3. Implement quality flags and measurement state information in CF-compliant attributes [29]: `flag_values`, `flag_masks`, and `flag_meanings` [29] (Section 3.5). The choice of `flag_values` vs. `flag_mask` depends on the use case. The `flag_values` and `flag_masks` may or may not be used together. An example of a complex case in which both

can be used is illustrated in [29], Section 1.7, example 3.5. In all cases, `flag_meanings` is used in conjunction with either `flag_masks` or `flag_values`.

Consider compliance with metadata standards related to data quality – International Organization for Standardization (ISO) 19157 [85], CF Conventions [29] including those for flags and indicators, ACDD [80], and ISO 8601 [76] [77]. Plan on using an automated compliance checker (Section 6.2).

4.6 Global Attributes

Global attributes (i.e., those that apply to an entire file rather than to a particular group or variable in a file) improve data discoverability, documentation, and usability. Descriptions of the recommended global attributes, according to CF, Attribute Conventions for Data Discovery (ACDD) [80], and other conventions, are found in Appendix D.

4.6.1 Provenance

Data provenance captures the origins, lineage, custody, and ownership of data. Including provenance in the file-level and product-level metadata helps ensure transparency and reproducibility. Provenance is also essential for debugging while data products are being developed, verified, validated, and evaluated for quality. File-level provenance can be combined with the data-product-level provenance to help the user ascertain the overall data product provenance. When describing provenance, include information about the context of the data production run (e.g., list of input data and ancillary files, production time) and the environment used to create the data product (e.g., software version, processing system, processing organization). While capturing provenance metadata can be challenging, it is generally good to capture sufficient lineage information for an independent third party to be able to reproduce the generation of a science data product.

The recommended provenance metadata for the processing environment are shown in Appendix D.6.

4.7 Variable-Level Attributes

Variable-level attributes (i.e., those that apply to a specific variable) improve interoperability, documentation, and usability. The recommended variable-level attributes given by the CF and the ACDD conventions are found in Appendix E. Several of these variable-level attributes have been discussed above, such as `long_name`, `standard_name`, `units`, `flag_values`, `flag_masks`, and `flag_meanings`).

5 DATA COMPRESSION, CHUNKING, AND PACKING

Data compression and chunking are two storage features provided by the HDF5 library and are available through the netCDF-4 API. HDF5's internal compression can reduce the space taken up by variables, especially those with many fill values or value repetition. The saved space can pay significant dividends in both storage space and transmission speed over the network. HDF5 includes a compression method referred to as "deflation", based on the common compressor gzip (itself based on the Lempel-Ziv algorithm). Deflation levels run from 1 to 9, with storage efficiency and time to compress increasing with each level. A level of 5 is often a good compromise. NetCDF-

4/HDF5 variables are individually compressed within the file, which means that applications only need to uncompress those variables of interest, and not the whole file, as would be necessary for external compression methods such as gzip or bzip. NetCDF-4/HDF5 variables can also be "chunked," which means that each variable is stored as a sequence of separate chunks in the file. If compression is used, each chunk is compressed separately. This allows read programs to decompress only the chunks required for a read request, and not the entire variable, resulting in even greater I/O efficiencies. Chunking also can allow a calling program to retrieve segments of data more efficiently when those data are stored in Object Storage (see Appendix B, Glossary). Note that the DIWG recommends using only the DEFLATE compression filter on netCDF-4 and netCDF-4-Compatible HDF5 Data [86]. Also, applying the netCDF-4/HDF5 "shuffle" filter before deflation can significantly improve the data compression ratio for multidimensional netCDF-4/HDF5 variables.

Chunking is appropriate for variables with a large number of values, particularly for multidimensional variables. It is helpful to consider the most likely pattern of data usage. However, where this is unknown or widely varied, "balanced chunking" is recommended, i.e., balanced access speeds for time-series and geographic cross-sections, the two most-common end-member geometries of data access. For example, Unidata has an algorithm for balanced chunking [87]. The DIWG recommends using balanced chunking for multidimensional variables contained in grid structures [24] (Rec. 2.11). Further recommendations regarding chunk sizes for use in the cloud are covered in Section 3.3 above.

In addition, the following command-line utilities can be used to chunk and compress files after the files have been written:

- h5repack (part of the HDF5 library [27]).
- nccopy (part of the netCDF library [88]).
- ncks (part of the NCO package [89]).

These utilities are also useful in experimenting with different compression levels and chunking schemes.

Another way to reduce data size is to apply a scale and offset factor to the data values, allowing the values to be saved as a smaller data type, such as a 2-byte integer. This technique, known as "packing," is appropriate for data with a limited dynamic range. The attributes needed to reconstruct the original values are `scale_factor` and `add_offset`.

The equation to reconstruct the original values is:

final_data_value = (`scale_factor` * packed_data_value) + `add_offset`

The values for `scale_factor` and `add_offset` may be selected by the data producer, or automatically computed to minimize storage size by a packing utility such as **ncpdq** (part of the NCO package [90]). The DIWG recommends using packing only when the packed data are stored as integers [24] (Rec. 2.6).

An additional benefit of packing is that it can sometimes make the data more compressible via deflation. That is, packing followed by the netCDF-4/HDF5 "shuffle" filter followed by deflation can result in very significant data compression.

6 TOOLS FOR DATA PRODUCT TESTING

The steps indicated below should be followed to test compliance and usability of a new data product and to respond to any issues found during testing:

1. Inspect the contents of the data file (e.g., via ncdump, h5dump, Panoply, HDFView, or a similar tool) to check for correctness of the metadata, and that the data/metadata structures agree with the product user guide (Section 6.1).
2. Use automated compliance checkers to test the product against metadata conventions (Section 6.2).
3. Inspect and edit the metadata using tools described in Section 6.3, if problems are found.
4. Modify the data production code as required, once all the necessary changes are known.
5. Test the product with the tools that will likely be used on the product (Section 6.4).
6. Validate that the packaging decisions (see Section 5) result in the desired size/performance trade-off.

6.1 Data Inspection

Dumping the data and inspecting them is a useful first check or troubleshooting technique. It can reveal obvious problems with standards compliance, geolocation (see Appendix C.1), and consistency with the data product user guide (if available). Useful tools for data inspection of netCDF-4 and HDF5 files are summarized in Table 3 (but also see [91]).

Table 3. Useful tools for inspecting netCDF-4 and HDF5 data. The "Type" column indicates the interfaces supported by the tools (command-line interface (CLI) or graphical user interface (GUI)).

Tool	Type	Access	Capabilities
HDFView	GUI	Website [28]	Read netCDF-4 and HDF5 files; views any data object; select “Table” from menu bar and then “export to text” or “export to binary.”
Panoply	GUI	Website [92]	Read netCDF-4 and HDF5 files; create images and maps from the data contained in a data product.
h5diff	CLI	HDF5 library [27]	Compare a pair of netCDF-4 or HDF5 files. The differences are reported as ASCII text.
h5dump, h5ls	CLI	HDF5 library [27]	Dump netCDF-4 and HDF5 file content to ASCII format.
IDV	CLI	IDV [93]	Integrated Data Viewer. 3D geoscience visualization and analysis tool that gives users the ability to view and analyze geoscience data in an integrated fashion.
ncdump	CLI	NetCDF-4 C library [88]	Dump netCDF-4 file content to ASCII format.
ncompare	CLI	NASA github [94]	Compare a pair of netCDF-4 files. Runs directly in Python and provides an aligned and colorized difference report for quick assessments of groups, variable names, types, shapes, and attributes. Can generate .txt, .csv or .xlsx report files.
ncks	CLI	NCO Toolkit [95]	Read and dump netCDF-4 and HDF5 files.

Tool	Type	Access	Capabilities
NCL	CLI	NCAR Command Language [96]	Interpreted language designed for scientific data analysis and visualization[5]

6.2 Compliance Checkers

Compliance checkers should be used while data products are being developed to ensure that the metadata fields are all populated and are meaningful. The following are recommended compliance checkers:

- Metadata Compliance Checker (MCC) is a web-based tool and service designed by the Physical Oceanography DAAC (PO.DAAC) for netCDF and HDF formats [97].
- CFChecker developed by Decker [98].
- Dismember developed by NCO [99].
- Integrated Ocean Observing System (IOOS) Compliance Checker [100].
- National Centre for Atmospheric Science (NCAS) CF Compliance Checker [101].
- Center for Environmental Data Analysis (CEDA) [102].

6.3 Internal Metadata Editors

Data editors can be useful in tweaking metadata internal to the data files when problems surface during testing. Once the metadata (or data) have been corrected, of course, the data processing code typically needs to be modified for the actual production runs. Several useful tools available for editing data in netCDF-4 and HDF5 formats are summarized in Table 4 (but also see [91]).

Table 4. Useful tools for editing netCDF-4 and HDF5 metadata and data. The "Type" column indicates the interfaces supported by the tools (command-line interface (CLI) or graphical user interface (GUI)).

Tool	Type	Access	Capabilities
HDFView	GUI	Website [28]	Create, edit, and delete content of netCDF-4 and HDF5 files.
h5diff	CLI	HDF5 library [27]	Compare a pair of netCDF-4 or HDF5 files. The differences are reported as ASCII text.
ncatted	CLI	NCO Toolkit [95]	Edit netCDF-4 global, group and variable-level attributes.
ncks	CLI	NCO Toolkit [95]	For netCDF-4: subset, chunk, compress, convert between versions, copy variables from one file to another, merge files, print.
ncompare	CLI	NASA github [94]	Compare a pair of netCDF-4 files. Runs directly in Python and provides an aligned and colorized difference report for quick assessments of groups, variable names, types, shapes, and attributes. Can generate .txt, .csv or .xlsx report files.
ncrename	CLI	NCO Toolkit [95]	Rename groups, dimensions, variables, and attributes of netCDF-4 files.
ncdump	CLI	NetCDF-4 C library [88]	Print the internal metadata of netCDF-4 files.
ncgen	CLI	NetCDF-4 C library [88]	Convert ASCII files to netCDF-4 format.

[5] NCL was put into maintenance mode in 2019. See https://geocat.ucar.edu/blog/2020/11/11/November-2020-update.

6.4 Other Community-Used Tools

Two generalized GUI tools in particular work with a large variety of netCDF and HDF products: Panoply [92] and HDFView [28]. Thus, these are recommended for at least minimal testing of data products before their release to users. Appendix C provides illustrations of these tools. In addition, it is helpful to test with tools that are in wide use by the target community for a data product, such as GIS tools for land processes products. Some of these tools are shown in Table 5. This table provides representative examples of data analysis tools, and is intended to spark a discussion of what tools the target community(communities) is(are) using.

Table 5. Other community-used tools

Tool	Source	URL
ArcGIS	Esri	https://www.arcgis.com/
AppEEARS	NASA Land Processes DAAC	https://lpdaac.usgs.gov/tools/appeears/
ENVI	NV5 Geospatial Software	https://www.nv5geospatialsoftware.com
ERDAS IMAGINE	Hexagon	https://hexagon.com/products/erdas-imagine
GMT	University of Hawai'i at Mānoa	https://www.generic-mapping-tools.org/
Google Earth Engine	Google	https://earthengine.google.com/
Grid Analysis and Display System (GrADS)	George Mason University	http://cola.gmu.edu/grads/
GRASS	OSGeo Project	https://grass.osgeo.org/
HDFLook	HDF-EOS Tools and Information Center	https://hdfeos.org/software/HDFLook.php
HEG	NASA Land Processes DAAC	https://lpdaac.usgs.gov/tools/heg/
IDL	NV5 Geospatial Software	https://www.nv5geospatialsoftware.com/Products/IDL
IDRISI	Clark Labs	https://clarklabs.org/terrset/idrisi-gis/
Multi-Mission Algorithm and Analysis Platform (MAAP)	NASA and ESA	https://www.earthdata.nasa.gov/esds/maap

Tool	Source	URL
MATLAB	Mathworks	https://www.mathworks.com/products/matlab.html
Octave	GNU Octave	https://octave.org/
Python	Python Software Foundation	https://www.python.org/
Quantum GIS (QGIS)	OSGeo Project	https://qgis.org/en/site/
R	The R Project for Statistical Computing	https://www.r-project.org/
SeaDAS	NASA Ocean Biology DAAC	https://seadas.gsfc.nasa.gov/
Sentinel Application Platform (SNAP)	European Space Agency	https://step.esa.int/main/download/snap-download/

7 DATA PRODUCT DIGITAL OBJECT IDENTIFIERS

A Digital Object Identifier (DOI) is a unique alphanumeric character string (i.e., handle) that can be used to identify a data product. A DOI is permanent, such that when it is registered, it can be used to locate the object to which it refers permanently. Since their introduction in 2000, DOIs have been routinely assigned to journal articles and cited by the scientific community. Use of DOIs for data products, however, is more recent but equally important for universal referencing and discoverability of data, as well as for proper attribution and citation.

The DOI handle is composed of a prefix that includes characters to identify the registrant and a suffix that includes the identification number of the registered object. In addition to the DOI handle, a web address is assigned by the DOI registration service provider. For a data product, its DOI typically leads to a web landing page (for guidelines, see [103] [104]) that provides information about the data product and services for users to obtain the data. One of the key benefits of assigning a DOI to a data product is that even if the web address changes, the DOI remains valid. This means that a DAAC can change the web address of a data product without affecting the validity of references made in published literature. In addition, the data publisher could change, but the DOI is unaffected. For a detailed description of DOIs, see the DOI Handbook [105]. The ESDIS Project has established procedures for managing DOIs for EOSDIS data [106]. The format of the DOIs managed by the ESDIS Project is *[prefix]/[suffix].* Here the prefix always starts with 10 followed by a number, as in 10.5067, and is assigned to the agency (ESDIS) for its repositories whereas the suffix uniquely identifies the object. The suffix can be semantic (containing meaningful information about the digital object) or opaque (any combination of alphanumeric characters, usually generated randomly and not having any semantic content). Examples of the URLs for two products, showing sematic and opaque DOIs respectively, are https://doi.org/10.5067/ECOSTRESS/ECO2LSTE.001 and https://doi.org/10.5067/D7GK8F5J8M8R.

Data producers should work with the assigned DAAC to obtain DOIs for their data products. The requests for DOI registration are made to the ESDIS Project by a DAAC.

The ESDIS Project uses a two-step process for registering DOIs for most DAACs. First, DOIs are reserved, so that data producers can start using them in the metadata while generating the products. Information about the DOI should be included in the data product metadata. In particular, the DOI resolving authority (i.e., https://doi.org) and the DOI identifier must be included as global attributes (see [107]). When a data product is ready to be delivered to the DAAC for public release, the DOI is registered. Until the DOI is registered, it can be modified or deleted (withdrawn). However, once registered, the DOI becomes permanent.

8 PRODUCT DELIVERY AND PUBLICATION

The responsibilities for generating data products and making them available to the user community are shared between data producers and the DAACs. Earthdata Pub [108] is a set of tools and guidance to help data producers publish Earth Science data products with an assigned NASA DAAC. In the case of an unknown DAAC assignment, except in the case of ROSES-funded projects, Earthdata Pub can be used to submit information about a potential data product for consideration and possible assignment by NASA. ROSES-funded projects should work with their Program Scientists to have a DAAC assigned as early as possible after their funding has been approved. Earthdata Pub creates a consistent data publication experience across the DAACs and provides the primary point of interaction for data producers and the DAAC staff. Using Earthdata Pub data producers can: 1) request to publish data at a DAAC, 2) submit information and files required to publish data in one place, 3) track the publication status of a request in real-time, and 4) communicate directly with DAAC staff support. Earthdata Pub includes resources to help data producers at each step along the way. Simple forms and workflows provide a guided data publication process.

Even though the details of the processes leading to data delivery and publication vary depending on the type of data as well as the DAAC that the data producer is working with, the primary phases of the data publication processes, from the perspective of the DAAC, are generally the same: 1) obtain the data, documentation and metadata, and related information from data producers; 2) work with data producers to generate CMR compliant metadata and additional documentation (e.g., user guide) describing the data[6]; 3) generate or adapt appropriate data software readers as needed; and 4) release the data and documentation for access by the user community. Each of the 12 EOSDIS DAACs have historically established publication workflows that account for the heterogeneous suite of missions, instruments, data producers, data formats, and data services managed by EOSDIS. Earthdata Pub allows each DAAC to maintain unique workflow steps while combining common steps. The specifics of data delivery and publication, such as schedules, interfaces, workflow, and procedures for submitting data product updates, are best established by communications between the data producers and their assigned DAACs through the Earthdata Pub. Data producers and DAACs will agree to a Level of Service [109] for the data. Levels of Service are the services applied to data during archiving and preservation to optimize the data usability and access.

To allow ample time to discuss data format and packaging, data producers should contact the assigned DAAC through Earthdata Pub as soon as sample data are ready. As of May 2024, the DAACs are actively on-boarding to Earthdata Pub. The data producer should check the list of DAACs in

[6] Data documentation should include at least a user guide. If the product is a geophysical retrieval, then providing an Algorithm Theoretical Basis Document is recommended as well.

Earthdata Pub or contact the assigned DAAC to verify the DAAC's preferred process. The general process for adding new data products to EOSDIS as well as the requirements and responsibilities of data producers and DAACs are shown in [110]. To get started publishing data, visit Earthdata Pub: https://pub.earthdata.nasa.gov.

9 REFERENCES

[1] ESDIS, "Earth Science Data Systems (ESDS) Program," 6 November 2023. [Online]. Available: https://www.earthdata.nasa.gov/esds. [Accessed 9 January 2024].

[2] ESDIS, "Earth Science Data and Information System (ESDIS) Project," 16 January 2020. [Online]. Available: https://earthdata.nasa.gov/esdis. [Accessed 9 January 2024].

[3] ESDIS, "Earth Science Data System Working Groups," 1 March 2021. [Online]. Available: https://www.earthdata.nasa.gov/engage/esdswg. [Accessed 16 May 2024].

[4] H. K. Ramapriyan and P. J. T. Leonard, "Data Product Development Guide (DPDG) for Data Producers version1.1. NASA Earth Science Data and Information System Standards Office," 21 October 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/data-product-development-guide-for-data-producers. [Accessed 9 January 2024].

[5] ESDS Program, "Airborne and Field Resources for Investigation Scientists and Data Managers," 7 December 2022. [Online]. Available: https://www.earthdata.nasa.gov/esds/impact/admg/evs. [Accessed 7 January 2024].

[6] GES DISC, "GES DISC Data and Metadata Recommendations to Data Providers," 31 March 2022. [Online]. Available: https://docserver.gesdisc.eosdis.nasa.gov/public/project/DataPub/GES_DISC_metadata_and_data_formats.pdf. [Accessed 9 January 2024].

[7] P. Meyappan, P. S. Roy, A. Soliman, T. Li, P. Mondal, S. Wang and A. K. Jain, "Documentation for the India Village-Level Geospatial Socio-Economic Data Set: 1991, 2001," NASA Socioeconomic Data and Applications Center (SEDAC), 12 March 2018. [Online]. Available: https://doi.org/10.7927/H43776SR. [Accessed 9 January 2024].

[8] PO DAAC, "PO.DAAC data management best practices," [Online]. Available: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices. [Accessed 9 January 2024].

[9] NSIDC DAAC, "Submit NASA Data to NSIDC DAAC," 2024. [Online]. Available: https://nsidc.org/data/submit-data/submit-nasa-data-nsidc-daac/assigned-data. [Accessed 9 Janaury 2024].

[10] ORNL DAAC, "Submit Data," [Online]. Available: https://daac.ornl.gov/submit/. [Accessed 9 January 2024].

[11] GHRSST Science Team (2010), "The Recommended GHRSST Data Specification (GDS) 2.0, Document Revision 5," 9 October 2012. [Online]. Available: https://doi.org/10.5281/zenodo.4700466. [Accessed 9 January 2024].

[12] ESDIS, "ESDIS Standards Coordination Office (ESCO)," 18 May 2022. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco. [Accessed 9 January 2024].

[13] M. D. Wilkinson, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data,* vol. 3, 15 March 2016.

[14] NASA, "Science Information Policy," 25 April 2023. [Online]. Available: https://science.nasa.gov/researchers/science-data/science-information-policy. [Accessed 9 January 2024].

[15] ESDS Program, "Open Data, Services, and Software Policies," NASA, 25 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/engage/open-data-services-and-software. [Accessed 9 January 2024].

[16] ESDIS, "EOSDIS Glossary," 28 January 2020. [Online]. Available: https://www.earthdata.nasa.gov/learn/glossary. [Accessed 9 January 2024].

[17] ESDIS Project, "Common Metadata Repository," 12 May 2021. [Online]. Available: https://earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr. [Accessed 9 January 2024].

[18] G. Asrar and H. K. Ramapriyan, "Data and Information System for Mission to Planet Earth," *Remote Sensing Reviews,* vol. 13, pp. 1-25. https://doi.org/10.1080/02757259509532294, 1995.

[19] H. K. Ramapriyan, J. F. Moses and D. Smith, "NASA Earth Science Data Preservation Content Specification, Revision C.," 3 May 2022. [Online]. Available: https://earthdata.nasa.gov/esdis/eso/standards-and-references/preservation-content-spec. [Accessed 9 January 2024].

[20] H. K. Ramapriyan, J. F. Moses and D. Smith, "Preservation Content Implementation Guidance, Version 1.0," 25 January 2022. [Online]. Available: https://doi.org/10.5067/DOC/ESO/RFC-042. [Accessed 9 January 2024].

[21] NASA, "Interface Management," [Online]. Available: https://www.nasa.gov/seh/6-3-interface-management. [Accessed 7 January 2024].

[22] H. Ramapriyan, G. Peng, D. Moroni and C.-L. Shie, "Ensuring and Improving Information Quality for Earth Science Data and Products," *D-Lib Magazine,* vol. 23, no. July/August 2017 Number7/8 https://doi.org/10.1045/july2017-ramapriyan, 2017.

[23] DIWG, "Dataset Interoperability Recommendations for Earth Science," Dataset Interoperability Working Group, 17 November 2022. [Online]. Available: https://wiki.earthdata.nasa.gov/display/ESDSWG/Dataset+Interoperability+Recommendations+for+Earth+Science. [Accessed 9 January 2024].

[24] DIWG, "Dataset Interoperability Recommendations for Earth Science, ESDS-RFC-028v1.3," ESDIS Project, 19 June 2020. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/dataset-interoperability-recommendations-for-earth-science. [Accessed 9 January 2024].

[25] ESCO, "Standards and Practices," ESDIS Project, 21 November 2023. [Online]. Available: ttps://www.earthdata.nasa.gov/esdis/esco/standards-and-practices. [Accessed 9 January 2024].

[26] ESCO, "netCDF-4/HDF5 File Format," ESDIS Project, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/netcdf-4hdf5-file-format. [Accessed 9 January 2024].

[27] ESCO, "HDF5 Data Model, File Format and Library – HDF5 1.6," ESDIS Project, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/hdf5. [Accessed 9 January 2024].

[28] The HDF Group, "HDF View," [Online]. Available: https://www.hdfgroup.org/downloads/hdfview/. [Accessed 9 January 2024].

[29] ESDIS, "Climate and Forecast (CF) Metadata Conventions," ESDIS Standards Coordination Office, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/climate-and-forecast-cf-metadata-conventions. [Accessed 9 January 2024].

[30] A. Jelenak, "Encoding of Swath Data in the Climate and Forecast Convention," 19 June 2018. [Online]. Available: https://github.com/Unidata/EC-netCDF-CF/blob/master/swath/swath.adoc. [Accessed 9 January 2024].

[31] DIWG, "Dataset Interoperability Recommendations for Earth Science: Part 2, ESDS-RFC-036v1.2," June 2020. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/dataset-interoperability-recommendations-for-earth-science. [Accessed 9 January 2024].

[32] NCEI, "NCEI NetCDF Templates v2.0," 7 December 2015. [Online]. Available: https://www.nodc.noaa.gov/data/formats/netcdf/v2.0/. [Accessed 9 January 2024].

[33] ESDIS Project, "Earthdata Search," 18 September 2023. [Online]. Available: https://www.earthdata.nasa.gov/learn/earthdata-search. [Accessed 9 January 2024].

[34] Adobe, "TIFF," [Online]. Available: https://www.adobe.com/creativecloud/file-types/image/raster/tiff-file.html. [Accessed 9 January 2024].

[35] wikipedia, "Geographic information system," wikipedia, 8 January 2024. [Online]. Available: https://en.wikipedia.org/wiki/Geographic_information_system. [Accessed 9 January 2024].

[36] OGC, "Open Geospatial Consortium," [Online]. Available: https://www.ogc.org/. [Accessed 9 January 2024].

[37] ESCO, "GeoTIFF File Format, ESDS-RFC-040v1.1," 1 March 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/geotiff. [Accessed 16 May 2024].

[38] COG, "Cloud Optimized GeoTIFF: An imagery format for cloud-native geospatial processing," [Online]. Available: https://www.cogeo.org/. [Accessed 9 January 2024].

[39] ESCO, "Cloud Optimized GeoTIFF," April 2024. [Online]. Available: https://doi.org/10.5067/DOC/ESCO/ESDS-RFC-049v1. [Accessed April 2024].

[40] OGC, "OGC Cloud-Optimized GeoTIFF Standard," 14 July 2023. [Online]. Available: http://www.opengis.net/doc/is/COG/1.0. [Accessed 18 January 2024].

[41] C. Plain, "New Standard Announced for Using GeoTIFF Imagery in the Cloud," 3 January 2024. [Online]. Available: https://www.earthdata.nasa.gov/learn/articles/new-cloud-optimized-geotiff-standard?. [Accessed 12 February 2024].

[42] ESCO, "Instructions to RFC Authors," ESDIS Project, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-references/instructions-to-rfc-authors. [Accessed 9 January 2024].

[43] ESCO, "ASCII File Format Guidelines for Earth Science Data, ESDS-RFC-027v1.1," May 2016. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/ascii-file-format-guidelines-for-earth-science-data. [Accessed 9 January 2024].

[44] E. Northup, G. Chen, K. Aikin and C. Webster, "ICARTT File Format Standards V2.0," January 2017. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/icartt-file-format. [Accessed 9 January 2024].

[45] OGC, "OGC GeoPackage Encoding Standard Version 1.3.1," OGC, 16 November 2021. [Online]. Available: https://www.geopackage.org/spec131/. [Accessed 9 January 2024].

[46] W3C, "Extensible Markup Language (XML)," 11 October 2016. [Online]. Available: https://www.w3.org/XML/. [Accessed 9 January 2024].

[47] ESCO, "OGC KML," ESDIS Standards Coordination Office, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/ogc-kml. [Accessed 9 January 2024].

[48] Esri, "Shapefiles," Esri, [Online]. Available: https://doc.arcgis.com/en/arcgis-online/reference/shapefiles.htm. [Accessed 9 January 2024].

[49] Wikipedia, "Shapefile," Wikipedia, 27 August 2023. [Online]. Available: https://en.wikipedia.org/wiki/Shapefile. [Accessed 9 January 2024].

[50] Esri, "Geoprocessing considerations for shapefile output," 24 April 2009. [Online]. Available: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Geoprocessing%20considerations%20for%20shapefile%20output. [Accessed 9 January 2024].

[51] Internet Engineering Task Force (IETF), "The GeoJSON Format," August 2016. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc7946. [Accessed 7 January 2024].

[52] ESCO, "HDF-EOS5 Data Model, File Format and Library," ESDIS Standards Office, 20 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/hdf-eos5. [Accessed 9 January 2024].

[53] A. Taaheri and K. Rodrigues, "HDF-EOS5 Data Model, File Format and Library," May 2016. [Online]. Available: https://cdn.earthdata.nasa.gov/conduit/upload/4880/ESDS-RFC-008-v1.1.pdf. [Accessed 9 January 2024].

[54] CEOS, "CEOS Analysis Ready Data," Committee on Earth Observation Satellites, 18 October 2021. [Online]. Available: http://ceos.org/ard/. [Accessed 9 Januarhy 2024].

[55] ESDS Program, "Earthdata Cloud Evolution," 10 August 2023. [Online]. Available: https://www.earthdata.nasa.gov/eosdis/cloud-evolution. [Accessed 8 January 2024].

[56] Cloud-Native Geospatial Foundation, "Cloud-Optimized Geospatial Formats Guide," 2023. [Online]. Available: https://guide.cloudnativegeo.org. [Accessed 8 January 2024].

[57] AWS, "Best Practices Design Patterns: Optimizing Amazon S3 Performance," June 2019. [Online]. Available: https://d1.awsstatic.com/whitepapers/AmazonS3BestPractices.pdf. [Accessed 9 January 2024].

[58] PANGEO TEam, "PANGEO - A community platform for Big Data geoscience," 2023. [Online]. Available: https://pangeo.io/. [Accessed 9 January 2024].

[59] R. Signell, A. Jelenak and J. Readey, "Cloud-Performant NetCDF4/HDF5 Reading with the Zarr Library," 26 February 2020. [Online]. Available: https://medium.com/pangeo/cloud-performant-reading-of-netcdf4-hdf5-data-using-the-zarr-library-1a95c5c92314. [Accessed 9 January 2024].

[60] Zarr Developers, "Zarr-Python Version 2.14.2," 2022. [Online]. Available: https://zarr.readthedocs.io/en/stable/index.html. [Accessed 9 January 2024].

[61] D. J. Newman, "Zarr storage specification version 2: Cloud-optimized persistence using Zarr. NASA Earth Science Data and Information System Standards Coordination Office.," April 2024. [Online]. Available: https://doi/org/10.5067/DOC/ESCO/ESDS-RFC-048v1. [Accessed 13 May 2024].

[62] xarray Developers, "Xarray documentation," 8 December 2023. [Online]. Available: https://docs.xarray.dev/en/stable/. [Accessed 8 January 2024].

[63] OPeNDAP, "DMR++: How to build & deploy dmr++ files for Hyrax," 17 October 2023. [Online]. Available: https://docs.opendap.org/index.php?title=DMR%2B%2B. [Accessed 9 January 2024].

[64] P. Quinn, "Cloud Optimized Formats: NetCDF-as-Zarr Optimizations and Next Steps," Element 84, 29 March 2022. [Online]. Available: https://www.element84.com/blog/cloud-optimized-formats-netcdf-as-zarr-optimizations-and-next-steps. [Accessed 12 January 2024].

[65] S. J. S. Khalsa, E. M. Armstrong, J. Hewson, J. F. Koch, S. Leslie, S. W. Olding and A. Doyle, "A Review of Options for Storage and Access of Point Cloud Data in the Cloud," February 2022.

[Online]. Available: https://www.earthdata.nasa.gov/s3fs-public/2022-06/ESCO-PUB-003.pdf. [Accessed 12 January 2024].

[66] OGC, "GeoParquet," [Online]. Available: https://geoparquet.org/. [Accessed 8 January 2024].

[67] ESDIS, "Earthdata Harmony Documentation," [Online]. Available: https://harmony.earthdata.nasa.gov/docs. [Accessed 8 January 2024].

[68] NASA, "Service for transforming NetCDF4 files into Zarr files within Harmony," [Online]. Available: https://github.com/nasa/harmony-netcdf-to-zarr. [Accessed 12 January 2024].

[69] M. Durant, "kerchunk," 2021. [Online]. Available: https://fsspec.github.io/kerchunk/. [Accessed 8 January 2024].

[70] T. Stevens, "GCMD Keyword Access," 17 March 2021. [Online]. Available: https://wiki.earthdata.nasa.gov/display/CMR/GCMD+Keyword+Access. [Accessed 12 Janaury 2024].

[71] ESDIS, "GCMD Keywords by Category," 12 January 2024. [Online]. Available: https://gcmd.earthdata.nasa.gov/static/kms/. [Accessed 12 January 2024].

[72] ESDIS, "GCMD Keyword Viewer," 2024. [Online]. Available: https://gcmd.earthdata.nasa.gov/KeywordViewer/scheme/all?gtm_scheme=all. [Accessed 8 January 2024].

[73] ESDIS, "NASA's GCMD releases the Keyword Governance and Community Guide Document, Version 1.0," 11 August 2016. [Online]. Available: https://www.earthdata.nasa.gov/news/nasa-s-gcmd-releases-the-keyword-governance-and-community-guide-document-version-1-0. [Accessed 12 January 2024].

[74] ESDS Program, "Data Processing Levels," 13 July 2021. [Online]. Available: https://www.earthdata.nasa.gov/engage/open-data-services-and-software/data-information-policy/data-levels. [Accessed 12 January 2024].

[75] ESDIS, "EOSDIS Glossary - "E"," 12 January 2024. [Online]. Available: https://www.earthdata.nasa.gov/learn/glossary#ed-glossary-e. [Accessed 12 January 2024].

[76] ISO, "ISO 8601-1:2019 Date and time -- Representations for information interchange -- Part 1: Basic rules," February 2019. [Online]. Available: https://www.iso.org/standard/70907.html. [Accessed 12 January 2024].

[77] ISO, "ISO 8601-2:2019 Date and time -- Representations for information interchange -- Part 2: Extensions," February 2019. [Online]. Available: https://www.iso.org/standard/70908.html. [Accessed 12 January 2024].

[78] CF Conventions, "CF Standard Name Table," CF Conventions, 19 January 2024. [Online]. Available: https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html. [Accessed 16 May 2024].

[79] ESCO, "Atmospheric Composition Variable Standard Name Convention," 2023. [Online]. Available: https://doi.org/10.5067/DOC/ESCO/ESDS-RFC-043v1. [Accessed 16 May 2024].

[80] ESIP Documentation Cluster, "Attribute Conventions for Data Discovery," 5 September 2023. [Online]. Available: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery. [Accessed 12 Janaury 2024].

[81] OGC, "Geographic information — Well-known text representation of coordinate reference systems," OGC, 16 August 2023. [Online]. Available: https://docs.ogc.org/is/18-010r11/18-010r11.pdf. [Accessed 26 April 2024].

[82] Unidata, "UDUNITS 2.2.28 Documentation," [Online]. Available: https://docs.unidata.ucar.edu/udunits/current/. [Accessed 12 January 2024].

[83] Earth Science Data Systems (ESDS) Program, HQ SMD, "Data Management Plan (DMP) Template for Data Producers, Version 1.1," 23 June 2020. [Online]. Available: https://wiki.earthdata.nasa.gov/download/attachments/118138197/ESDIS05161_DMP_for_DPs_template.pdf?api=v2. [Accessed 12 January 2024].

[84] ESDIS Project, "ICD Between the ICESat-2 Science Investigator-led Processing System (SIPS) and the National Snow and Ice Data Center (NSIDC) Distributed Active Archive Center (DAAC) - 423-ICD-007, Revision A," NASA GSFC, Greenbelt, MD, 2016.

[85] ISO, "ISO 19157-1:2023 Geographic information — Data quality — Part 1: General requirements," April 2023. [Online]. Available: https://www.iso.org/standard/78900.html. [Accessed 12 January 2024].

[86] DIWG, "Use Only Officially Supported Compression Filters on NetCDF-4 and NetCDF-4-Compatible HDF5 Data," 10 Janaury 2024. [Online]. Available: https://wiki.earthdata.nasa.gov/display/ESDSWG/Use+Only+Officially+Supported+Compression+Filters+on+NetCDF-4+and+NetCDF-4-Compatible+HDF5+Data. [Accessed 12 January 2024].

[87] Developers@Unidata, "Chunking Data: Choosing Shapes," 28 March 2013. [Online]. Available: https://www.unidata.ucar.edu/blogs/developer/en/entry/chunking_data_choosing_shapes. [Accessed 12 January 2024].

[88] UCAR, "Network Common Data Form (NetCDF)," 14 March 2023. [Online]. Available: https://www.unidata.ucar.edu/software/netcdf. [Accessed 12 January 2024].

[89] NCO, "NCO 5.1.6-alpha03 User Guide - 3.32 Chunking," 7 November 2023. [Online]. Available: http://nco.sourceforge.net/nco.html#Chunking. [Accessed 12 January 2024].

[90] NCO, "NCO Users Guide, Edition 5.1.6 - Alpha03," 7 November 2023. [Online]. Available: http://nco.sourceforge.net/nco.html#ncpdq-netCDF-Permute-Dimensions-Quickly. [Accessed 12 January 2024].

[91] ESDIS, "Data Tools," 26 May 2023. [Online]. Available: https://www.earthdata.nasa.gov/learn/use-data/tools. [Accessed 12 January 2024].

[92] NASA, "Panoply netCDF, HDF and GRIB Data Viewer," 1 Janaury 2024. [Online]. Available: https://www.giss.nasa.gov/tools/panoply/. [Accessed 12 Janaury 2024].

[93] UCAR, "Integrated Data Viewer," 28 August 2023. [Online]. Available: https://www.unidata.ucar.edu/software/idv/. [Accessed 16 January 2024].

[94] NASA, "NASA ncompare," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10636759. [Accessed 12 February 2024].

[95] NCO, "Bienvenue sur le netCDF Operator (NCO) site," 8 November 2023. [Online]. Available: http://nco.sourceforge.net/. [Accessed 16 January 2024].

[96] NCAR, "The NCAR Command Language (Version 6.6.2) [Software]," Boulder, Colorado: UCAR/NCAR/CISL/VETS, November 2020. [Online]. Available: http://dx.doi.org/10.5065/D6WD3XH5. [Accessed 16 January 2024].

[97] PO DAAC, "Metadata Compliance Checker," [Online]. Available: https://mcc.podaac.earthdatacloud.nasa.gov. [Accessed 24 April 2023].

[98] M. Decker, "CFchecker," 11 November 2018. [Online]. Available: https://jugit.fz-juelich.de/IEK8-Modellgruppe/cfchecker. [Accessed 16 January 2024].

[99] NCO, "NCO User Guide Version 5.1.6-Alpha03 - 3.15.3 Dismembering Files," 7 November 2023. [Online]. Available: http://nco.sourceforge.net/nco.html#ncdismember. [Accessed 16 January 2024].

[100] IOOS, "IOOS Compliance Checker," 17 May 2023. [Online]. Available: https://compliance.ioos.us/index.html. [Accessed 16 January 2024].

[101] National Centre for Atmospheric Science, "CF Compliance Checker," [Online]. Available: https://cfchecker.ncas.ac.uk/. [Accessed 8 January 2024].

[102] Python Software Foundation, "The NetCDF Climate Forecast Conventions compliance checker," 2024. [Online]. Available: https://pypi.org/project/cfchecker/. [Accessed 8 January 2024].

[103] ESDIS Project, "DOI Landing Page," 13 October 2016. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page. [Accessed 16 January 2024].

[104] ESDIS Project, "DOI Documents," 19 September 2023. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Documents. [Accessed 16 January 2024].

[105] International DOI Foundation, "DOI Handbook," April 2023. [Online]. Available: http://www.doi.org/hb.html. [Accessed 16 January 2024].

[106] ESDIS Project, "Digital Object Identifiers for ESDIS," 28 April 2023. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS. [Accessed 16 January 2024].

[107] ESDIS Project, "DOI Background Information," 28 September 2023. [Online]. Available: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Background+Information. [Accessed 16 Janaury 2024].

[108] Earthdata Pub Team, "NASA Earthdata Pub," [Online]. Available: https://pub.earthdata.nasa.gov/. [Accessed 16 Janaury 2024].

[109] ESDS Program, "Earth Science Data Systems Level of Service Model," 25 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/engage/new-missions/level-of-service. [Accessed 16 Janaury 2024].

[110] ESDS Program, "Adding New Data to EOSDIS," 25 May 2021. [Online]. Available: https://www.earthdata.nasa.gov/engage/new-missions. [Accessed 16 January 2024].

[111] PO.DAAC, "PO.DAAC Data Management Best Practices - Metadata Conventions," [Online]. Available: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#Metadata%20Conventions. [Accessed 16 January 2024].

[112] C. Davidson and R. Wolfe, "VIIRS Science Software Delivery Guide," 2021. [Online]. Available: https://doi.org/10.5067/FA8689BC-E374-11ED-B5EA-0242AC120001. [Accessed 16 January 2024].

[113] GOFAIR, "FAIR Principles," [Online]. Available: https://www.go-fair.org/fair-principles/. [Accessed 19 Jaunary 2024].

[114] Internet Assigned Numbers Authority, "Uniform Resource Identifier (URI) Schemes," 12 January 2024. [Online]. Available: https://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml. [Accessed 16 January 2024].

[115] "Reverse DNS Look-up," [Online]. Available: https://remote.12dt.com/. [Accessed 16 January 2024].

[116] NOAA EDM, "ISO 19115 and 19115-2 CodeList Dictionaries," ESIP, 3 October 2018. [Online]. Available: http://wiki.esipfed.org/index.php/ISO_19115_and_19115-2_CodeList_Dictionaries. [Accessed 16 January 2024].

[117] B. Eaton and et al., "NetCDF Climate and Forecast (CF) Metadata Conventions (section 2.7)," 5 December 2023. [Online]. Available: http://cfconventions.org/cf-conventions/cf-conventions.html#groups. [Accessed 16 January 2024].

10 AUTHORS' ADDRESSES

Hampapuram K. Ramapriyan
Email: Hampapuram.Ramapriyan@ssaihq.com

Peter J. T. Leonard
Email: peter.j.leonard@nasa.gov

Edward M. Armstrong
Email: Edward.M.Armstrong@jpl.nasa.gov

Siri Jodha Singh Khalsa
Email: khalsa@colorado.edu

Deborah K. Smith
Email: Deborah.Smith@uah.edu

Lena F. Iredell

Email: lena.f.iredell@nasa.gov

Daine M. Wright
Email: wrightdm@ornl.gov

George J. Huffman
Email: george.j.huffman@nasa.gov

Tammy. R. Walker
Email: beatytw@ornl.gov

11 CONTRIBUTORS AND EDITORS

Editors' names are shown in **bold**.

11.1 Contributors and Editors for Versions 1 and 1.1

(Affiliations indicated are as of October 2021, the publication date of Version 1.1. Editors' names are shown in **bold**)

Edward M. Armstrong, JPL/CalTech

Walter E. Baskin, SSAI & NASA/LaRC

Jeanne Beatty, ADNET & NASA/GSFC

Robert R. Downs, SEDAC

George Huffman, NASA/GSFC

Aleksandar Jelenak, The HDF Group

Siri Jodha Khalsa, NSIDC DAAC

Amanda Leon, NSIDC DAAC

Peter J.T. Leonard, ADNET & NASA/GSFC

Wen-Hao Li, JPL/CalTech

Christopher Lynnes, NASA/GSFC

David Moroni, JPL/CalTech

John Moses, NASA/GSFC

Dana Ostrenga, ADNET & NASA/GSFC

Hampapuram K. Ramapriyan, SSAI & NASA/GSFC

Justin Rice, NASA/GSFC

Elliot Sherman, SSAI & NASA/GSFC

Tammy Walker, ORNL DAAC

Lalit Wanchoo, ADNET & NASA/GSFC

Yaxing Wei, ORNL DAAC

Jessica N. Welch, ORNL DAAC

Robert Wolfe, NASA/GSFC

11.2 Contributors and Editors for Version 2

(Affiliations indicated are as of February 2024. Editors' names are shown in **bold**)

Edward M. Armstrong, NASA/JPL/CalTech PO.DAAC

Will Ellett, GHRC DAAC

George Huffman, NASA/GSFC

Lena Iredell, ADNET & NASA/GSFC GESDISC

Siri Jodha Khalsa, NSIDC DAAC

Peter J.T. Leonard, ADNET & NASA/GSFC

Juanisa McCoy, Raytheon Technologies

Hampapuram K. Ramapriyan, SSAI & NASA/GSFC ESDIS

Deborah Smith, UAH & NASA/IMPACT

Tammy Walker, ORNL DAAC

Daine Wright, ORNL DAAC

Taylor Wright, GHRC DAAC

Specific Contributions to V2.0:

Hampapuram K. Ramapriyan, SSAI & NASA/GSFC - Co-led the DPDG WG and the editing team. Ensured all comments from reviewers were accounted for and facilitated the team's work. Prepared final document and ensured all references were valid and accessible.

Peter J.T. Leonard, ADNET & NASA/GSFC – Co-led the DPDG WG and the editing team. Contributed to editing the entire document. Performed detailed "quality assurance" review of the final draft of the document.

Edward M. Armstrong, JPL/CalTech PO.DAAC – Wrote and edited Section 3.3 on cloud-optimized data formats and contributed to editing the entire document.

Siri Jodha Khalsa, NSIDC DAAC - Reviewed and contributed text to the metadata sections and contributed to editing the entire document. Provided many useful references to standards and policy documents.

Deborah Smith, UAH & NASA/IMPACT – Reviewed the entire document and modified or added text to ensure that the document addresses concerns of data producers from airborne and field campaign observations. Recommended the idea of a Resource Center for Data Producers.

Lena Iredell, ADNET & NASA/GSFC GES DISC – Reviewed and contributed text to the metadata sections, and showed the relationships between the Unified Metadata Model (UMM) and the attributes called out in Appendices D and E.

Daine Wright, ORNL DAAC – Contributed text in Section 8 pertaining to data publication using EarthDataPub and to editing the entire document.

George Huffman, NASA/GSFC – Actively participated in the editing team and critically evaluated the document from a data producer's point of view to ensure that the language is understandable to novice as well as experienced readers.

Tammy Walker, ORNL DAAC – Reviewed the document from the point of view of a data publisher and contributed to editing the entire document.

Juanisa McCoy, Raytheon Technologies - Contributed text in Section 8 pertaining to data publication using EarthDataPub.

Will Ellett, GHRC DAAC - Contributed text in Section 8 pertaining to data publication using EarthDataPub.

Taylor Wright, GHRC DAAC - Contributed text in Section 8 pertaining to data publication using EarthDataPub.

APPENDIX A. ABBREVIATIONS AND ACRONYMS

ACDD	Attribute Convention for Data Discovery
AIRS	Atmospheric Infrared Sounder (on Aqua)
AMSU	Advanced Microwave Sounding Unit (on Aqua)
API	Application Programming Interface
ARCO	Analysis-Ready, Cloud-Optimized
ARD	Analysis Ready Data
ASCII	American Standard Code for Information Interchange
ATBD	Algorithm Theoretical Basis Document
AWS	Amazon Web Services
Cal/Val	Calibration/Validation
CDL	Common Data Language
CEDA	Center for Environmental Data Analysis
CF	Climate and Forecast Metadata Conventions
CLI	Command Line Interface
CMR	Common Metadata Repository
COG	Cloud Optimized GeoTIFF
COPC	Cloud Optimized Point Cloud
CPU	Central Processing Unit
CRS	Coordinate Reference System
CSV	Comma-Separated Values
DAAC	Distributed Active Archive Center
DIWG	Dataset Interoperability Working Group
DMR++	Dataset Metadata Response
DOI	Digital Object Identifier
DPDG	Data Product Development Guide
DPR	Dual-Frequency Precipitation Radar (on GPM)
EOS	Earth Observing System
EOSDIS	Earth Observing System Data and Information System
EPT	Entwine Point Tiles
ESCO	ESDIS Standards Coordination Office

ESDIS Project	Earth Science Data and Information System Project
ESDS	Earth Science Data System (Program)
ESDSWG	Earth Science Data System Working Group
ESDT	Earth Science Data Type
FAIR	Findable, Accessible, Interoperable, Reusable
GCMD	Global Change Master Directory
GDAL	Geospatial Data Abstraction Library
GeoTIFF	Georeferenced Tagged Image File Format
GES DISC	NASA's Goddard Earth Sciences Data and Information Services Center
GHRSST	Group for High Resolution Sea Surface Temperature
GIS	Geographic Information System
GLDAS	Global Land Data Assimilation System
GPM	Global Precipitation Measurement (Mission)
GSFC	NASA Goddard Space Flight Center
GUI	Graphical User Interface
HDF5	Hierarchical Data Format, Version 5
HDF-EOS5	Hierarchical Data Format - Earth Observing System, Version 5 (based on HDF5)
HIRDLS	High Resolution Dynamics Limb Sounder
HTML	Hypertext Mark-up Language
HTTP	Hypertext Transfer Protocol
ICARTT	International Consortium for Atmospheric Research on Transport and Transformation
ICD	Interface Control Document
IDL	Interactive Data Language
I/O	Input/Output
IOOS	Integrated Ocean Observing System
ISO	International Organization for Standardization
JPL	NASA Jet Propulsion Laboratory
JSON	JavaScript Object Notation
KML	Keyhole Mark-up Language
L1, L2, L3	Level 1, Level 2, Level 3 (data product)

LAADS	Level-1 and Atmosphere Archive and Distribution System
LaRC	NASA Langley Research Center
LAS/LAZ	LASer file format/LAS compressed file format
MCC	Metadata Compliance Checker
MEaSUREs	**M**aking **Ea**rth **S**ystem Data Records for **U**se in **R**esearch **E**nvironment**s**
MERRA	Modern Era-Retrospective Analysis for Research and Applications
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCAS	National Centre for Atmospheric Science
NCO	NetCDF Operator
NetCDF-4	Network Common Data Form, Version 4
NOAA	National Oceanic and Atmospheric Administration
NRT	Near Real Time
NSIDC	National Snow and Ice Data Center
NUG	NetCDF Users Guide
OCO	Orbiting Carbon Observatory
OGC	Open Geospatial Consortium
OPeNDAP	Open-source Project for a Network Data Access Protocol
ORNL DAAC	Oak Ridge National Laboratory DAAC (NASA)
PGE	Product Generation Executable
PO.DAAC	Physical Oceanography DAAC (NASA JPL)
QA	Quality Assurance
QC	Quality Control
RFC	Request for Comments
ROI	Region of Interest
ROSES	Research Opportunities in Space and Earth Science
SDP	Science Data Production (Toolkit)
SEDAC	Socioeconomic Data and Applications Center (NASA)
SIPS	Science Investigator-led Processing System
SMB	Server Message Block
SSH	Secure Shell

SWAL	Strengths, weaknesses, applicability, and limitations
S3	Simple Storage Service
THREDDS	Thematic Real-time Environmental Distributed Data Services
TRMM	Tropical Rainfall Measuring Mission
UDUNITS	A Unidata package that contains an extensive unit database
UMM	Unified Metadata Model
URL	Uniform Resource Locator
USGS	United States Geological Survey
UTC	Coordinated Universal Time
uuid	Universal Unique Identifier
WCPS	Web Coverage Processing Service
WGS84	World Geodetic System 1984
W3C	World Wide Web Consortium
WKT	Well-Known Text markup language
XML	eXtensible Markup Language
YAML	Yet Another Markup Language

APPENDIX B. GLOSSARY

Attribute - Element constituting metadata.

Calibration/Validation (Cal/Val) - Calibration is a demonstration that an instrument or device produces accurate results; validation is a program that provides assurance that a specific process, equipment, or system consistently produces results that meet predetermined acceptance criteria.

Climate and Forecast Metadata Conventions - A set of metadata conventions that were invented for climate and weather forecast data but have since been applied to describe other kinds of Earth Science data, with the intention of promoting the processing and sharing of data.

Content Organization Scheme - A means for enhancing the addressing and access of elements contained in a digital object in the cloud.

Data Collection - A major release of a data product, or of a set of closely related data products, which can be followed by minor releases within the same collection.

Data Format - A standard way that information is encoded for storage in a computer file (see https://www.earthdata.nasa.gov/technology/data-format). An example is HDF5.

Data Distributor - An entity responsible for archiving and distributing data products (e.g., a DAAC).

Data Processing Level – The level of processing that results in data products ranging from raw instrument data to refined analyses that use inputs from various sources [74].

Data Producer – A person or group that directly collects/creates data to be submitted to a NASA DAAC for archiving and public distribution.

Data Product - A set of data files that can contain multiple parameters and that compose a logically meaningful group of related data.

Data Structure - In Earth Science data products, a multi-dimensional container for geolocation and science data tailored for a specific type of instrument acquisition mode or spatial arrangement of the data (e.g., swath, grid, zonal mean, trajectory).

Dataset - A broadly used term that can be used to describe any set of data. The official term “HDF5 dataset” describes a data array in an HDF5 file (equivalent to a NetCDF-4 variable in a NetCDF-4 file or an HDF-EOS field in an HDF-EOS file). An entire data collection is sometimes referred to as a dataset.

Discovery – Successful identification and location of data products of interest.

File Format - A term that is used for both "Data Format" and "Product Format" whose meaning should be understood by its context.

Global Attribute – An attribute that applies to either the entire file or the entire collection of files. Some important examples are provided in Appendix D.

Granule - The smallest aggregation of independently managed (i.e., described, inventoried, retrievable) data at a DAAC. Some web applications and services provided by DAACs allow for the subsetting of granules. One granule usually comprises one file, more rarely multiple

files. The latter is not optimal as it complicates data management by both the archive and users, and utilization by tools and services.

Metadata - Information about data.

NetCDF-4/HDF5 - NetCDF format that uses the HDF5 data storage model.

Product Format - The specific implementation of global attributes, dimensions, groups, variables, and variable-level attributes in a data product, which is specified via the Product Format Specification (PFS) file.

Quality Flag - One or more unique variables within a data file that show what data quality assessments have been performed as well as diagnostics on various aspects of quality. A quality flag can be a byte value with each bit representing a pre-defined quality verification criterion provided as a Boolean expression. For example, see https://oceancolor.gsfc.nasa.gov/resources/atbd/ocl2flags/.

Quality Indicator - One or more unique variables within a data file whose numerical value shows the overall quality of a geophysical measurement. The numerical value should be on a predefined numerical scale. For example, uncertainty per pixel (or measurement), percent cloud cover (in a scene).

Search – An activity attempting to identify and locate data products of interest given user-defined criteria.

Self-Describing File – A file that contains metadata thoroughly describing the characteristics and content of the file.

UMM Profile – One of seven UMM metadata profiles: Collection (UMM-C), Granule (UMM-G), Service (UMM-S), Variable (UMM-Var), Visualization (UMM-Vis), Tools (UMM-T), and Common (UMM-Common).

Variable (a.k.a. parameter) - A named set of data that contains the recorded values of a measurement. In this context, the variable is described by its name and characteristics (UMM-Var).

Web Object Storage - Object storage accessible through https.

APPENDIX C. PRODUCT TESTING WITH DATA TOOLS

C.1 Panoply

Panoply [92], a tool developed and maintained by the NASA Goddard Institute for Space Studies, is particularly useful in verifying file-level metadata, as well as the structure and proper geolocation of data. If Panoply is able to interpret the geolocation information contained in a file it can create color contour plots by latitude-longitude, latitude-vertical, longitude-vertical, time-latitude or time-vertical arrays slices from 2D or larger multidimensional variables. It can also create color contour plots of "generic" 2D arrays from 2D or larger multidimensional variables as well as line plots of data from 1D or larger multidimensional variables.

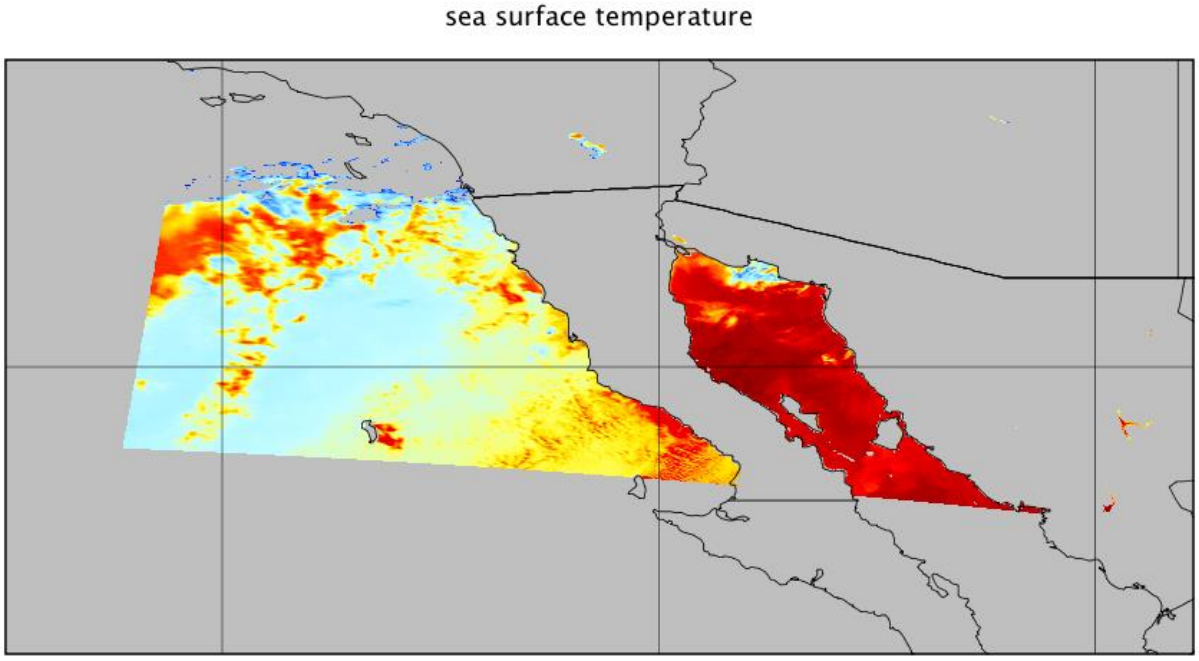

Figure 5. Example of a plot generated with the Panoply software environment demonstrating the display of a georeferenced two-dimensional dataset.

If the geolocation has not been set properly, it will display a two-dimensional image and generate a simple coordinate plot. Similarly, trajectory data will have type “GeoTraj” when properly geolocated; otherwise, it will display a one-dimensional image. Panoply also displays the variable-level and global metadata in the right-hand panel, making it convenient to confirm the map units and coordinate attributes.

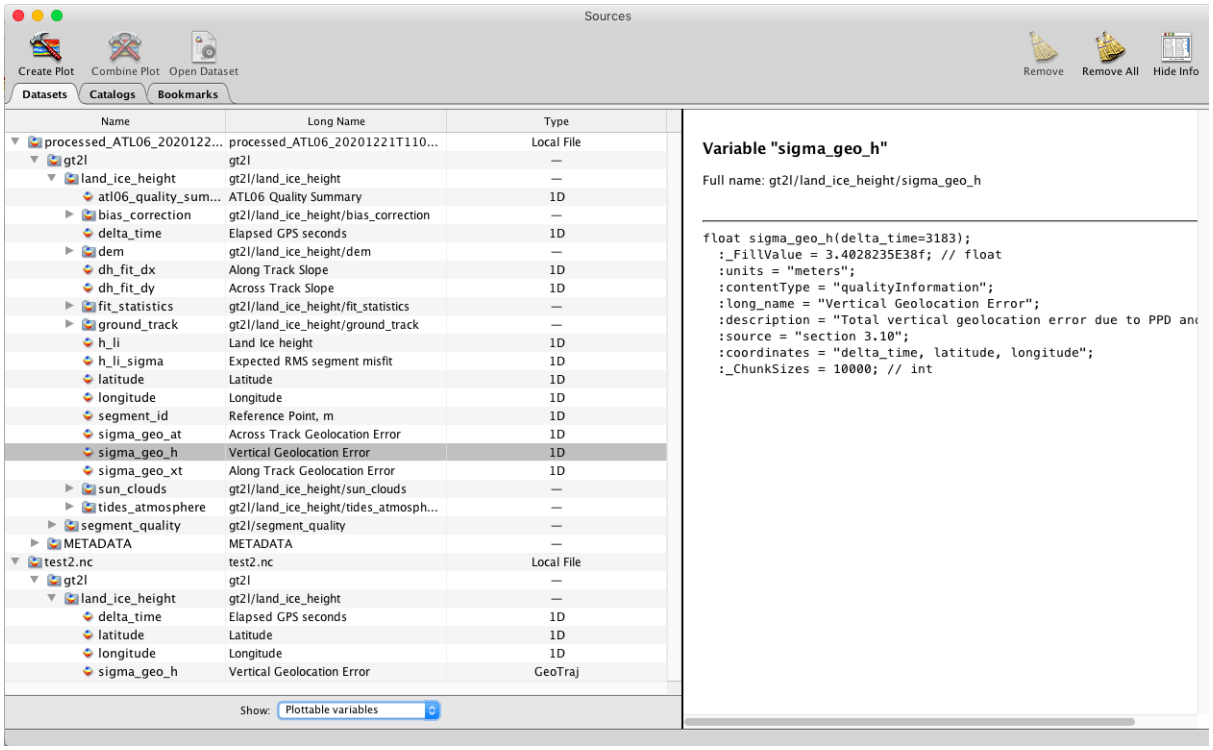

Figure 6. A screenshot of the Panoply software environment demonstrating a deviation from the CF Conventions.

Figure 6 illustrates that the variable “sigma_geo_h” for the data file “processed_ATL06_ … .nc” includes commas in the attribute `coordinates`, which violates the CF Conventions and causes Panoply to recognize the data as 1D. In the data file “test2.nc” there are no commas included in `coordinates`, Panoply recognizes the data as a GeoTraj, and the data can be plotted.

C.2 HDFView

HDFView from The HDF Group [28] is a tool for browsing and editing files in HDF and netCDF formats (e.g., Figure 7). Using HDFView, a user can view and modify the content of a file, view a file hierarchy in a tree structure, create new files, add or delete groups and data, and modify attributes. Unlike Panoply, the visualizations do not include a coastline overlay. However, the HDF-EOS plugin for HDFView can supply the latitude and longitude for each cell location in a data array.

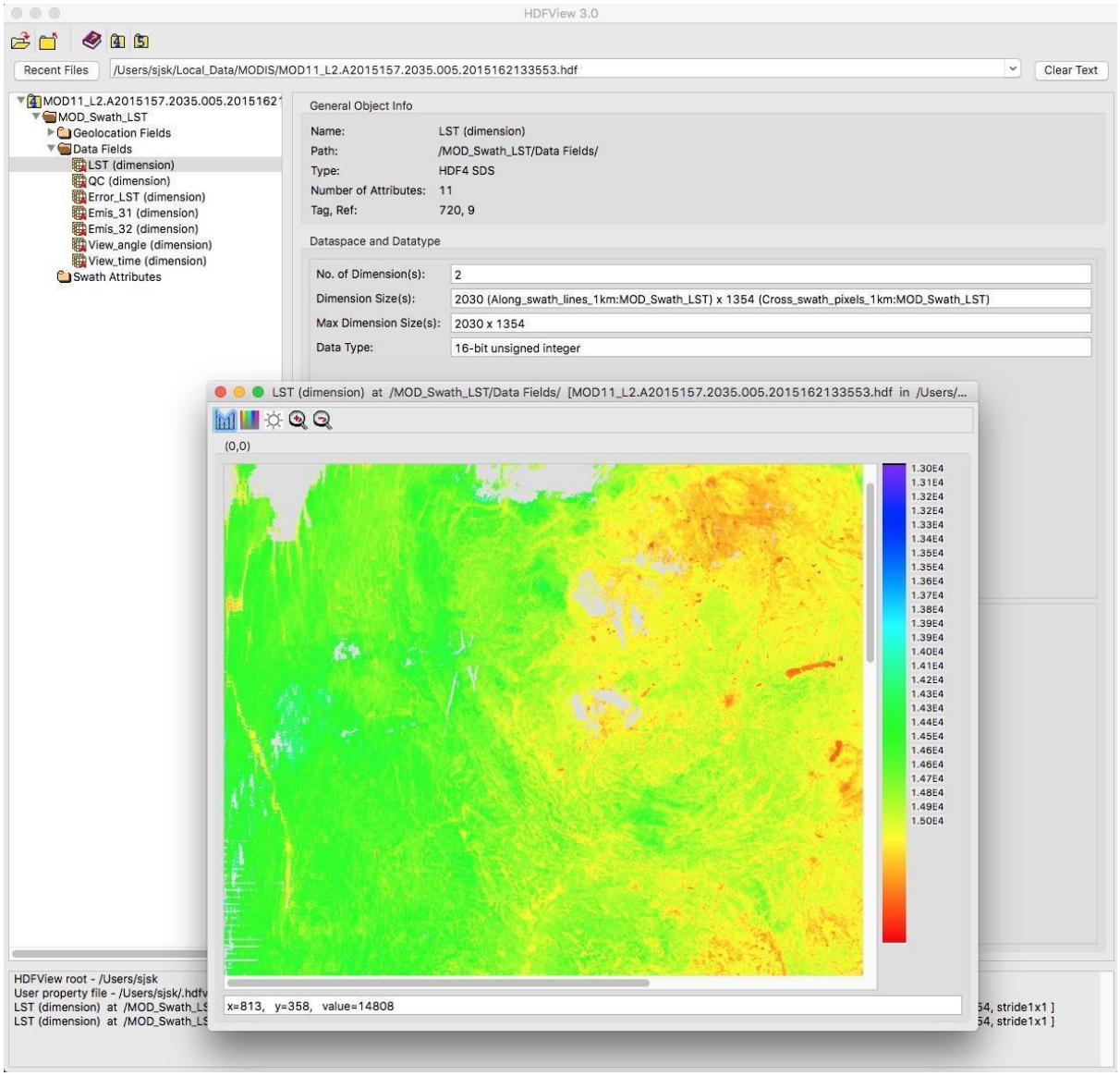

Figure 7. A screenshot of the HDFView software environment demonstrating how to view the dimension "LST" of a data swath.

APPENDIX D. IMPORTANT GLOBAL ATTRIBUTES

In this appendix, we provide a list of recommended global-level metadata attributes, following CF (Version 1.10) [29], ACDD (Version 1.3) [80], and other conventions. These are derived primarily from the CF and ACDD recommendations by PO.DAAC [111]. The attributes where the indicated source is the Goddard Earth Sciences Data and Information Services Center (GES DISC) are from reference [6]. The attributes where the source is shown as Level-1 and Atmosphere Archive and Distribution System (LAADS) DAAC are from reference [112]. Where the CF/ACDD and GES DISC versions have conceptually similar attributes, the GES DISC names for the attributes are shown in parentheses in the Attribute Name column. It is to be noted that while there is considerable commonality in the attribute sets called for by the different DAACs, there are also some differences.

These recommended attributes have been characterized with a justification for use, labeled with how they support *Findability, Accessibility, Interoperability, and Reusability (FAIR)* [13] [113], and mapped to UMM. The attributes are grouped based on the justification (i.e., what purpose they serve from the point of view of a data user) and shown in Sections D.1 through D.6. Section D.6 shows the attributes needed for providing information on provenance, and is divided into three subsections: General, Attribution, and Lineage. Where an attribute is considered to be important for more than one purpose, the Justification column shows both the primary justification (corresponding to the section in which it appears) and the other purpose(s). The column titled UMM-(X) lists the UMM profiles discussed in Section 4.1 that use the specific attributes shown in the corresponding rows. The column titled FAIR lists the mappings of FAIR principles to the attributes. The detailed descriptions of FAIR principles can be found in [113]. Note that all attributes whose sources are CF or ACDD map to FAIR principles I1, I2, and R1.3. Therefore, these are not repeated in the FAIR column.

Data producers are reminded to work with the DAACs to determine which attributes are required, where they are stored in the data products, and if additional attributes not in this list are necessary.

D.1 Interpretability

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
title (LongName/title)	A short phrase or sentence describing the data. In many search systems, this attribute will be displayed in the results list from a search and should be human-readable and of a reasonable length.	CF 1.7, ACDD 1.3, NUG 1.7	Interpretability	UMM-C, UMM-Var	F2
summary (ProjectAbstract)	A paragraph describing the data, analogous to an abstract for a manuscript.	ACDD 1.3	Interpretability	UMM-C	F2
keywords_vocabulary	If using a controlled vocabulary for the words/phrases in keywords, this is the unique name or identifier of the vocabulary from which keywords are taken. If more than one keyword vocabulary is used, each can be presented with a prefix (e.g., "CF:NetCDF COARDS Climate and Forecast Standard Names") and a following comma, so that keywords may optionally be prefixed with the controlled vocabulary key.	ACDD 1.3	Interpretability	UMM-C, UMM-Var	
Conventions (Conventions)	A comma-separated list of the conventions that the data follow. For files that follow this version of ACDD, include the string "ACDD-1.3."	CF 1.7, ACDD 1.3, NUG 1.7	Interpretability	UMM-G, UMM-S, UMM-T, UMM-Var	
platform_vocabulary	Controlled vocabulary for the names used in platform.	ACDD 1.3	Interpretability	UMM-C, UMM-Var	F2
instrument_vocabulary	Controlled vocabulary for the names used in instrument.	ACDD 1.3	Interpretability	UMM-C, UMM-Var	F2
standard_name_vocabulary	The name and version of controlled vocabulary from which variable standard names are taken. Values for standard_name must come from the CF Standard Names vocabulary for the data to comply. Example: "CF Standard Name Table v27." Multiple distinct vocabularies can be separated by commas.	ACDD 1.3	Interpretability	UMM-G, UMM-Var	F2

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
Format	Format of the data, e.g., HDF-EOS5 or netCDF-4.	GES DISC	Interpretability	UMM-C, UMM-G, UMM-S, UMM-T	I1, R1.3
MapProjection	Applies to gridded data. Useful for application data tools, such as ArcGIS.	GES DISC	Interpretability	UMM-S, UMM-T	I1, R1.2, R1.3
DataSetLanguage	Language used within the dataset, LanguageCode=ISO 639 default=English.	GES DISC	Interpretability	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	I1, R1.3

D.2 Discovery

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
platform (source)	Name of the platform that supported the sensor data used to create the data. Platforms can be of any type, including satellite, ship, station, aircraft, or other. Indicate controlled vocabulary used in platform_vocabulary.	ACDD 1.3	Discovery	UMM-C, UMM-Var	F2, R1.2
instrument (source)	Name of the contributing instrument or sensor used to create the data. Indicate controlled vocabulary used in instrument_vocabulary.	ACDD 1.3	Discovery	UMM-C, UMM-Var	F2, R1.2
processing_level (ProcessingLevel)	A textual description of the processing (or quality control) level of the data.	ACDD 1.3	Discovery	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
keywords (ProductParameters/ Keyword)	A comma-separated list of keywords and/or phrases. Keywords can be common words or phrases, terms from a controlled vocabulary (e.g., Global Change Master Directory [70]), or Uniform Resource Identifiers [114] for terms from a controlled vocabulary. See also keywords_vocabulary.	ACDD 1.3	Discovery	UMM-C, UMM-Var	F2
ShortName	An abbreviated name of the product that should contain the version and be descriptive of the data product. The ShortName may only contain alphanumeric numeric characters plus "_", and be under 30 characters	GES DISC	Discovery	UMM-C, UMM-Var	F2, R1.3
GranuleID	For most Level 1 to Level 4 products, this is the file name. It is recommended to include descriptive information, version, and date in the file name.	GES DISC	Discovery	UMM-C, UMM-G, UMM-S, UMM-T	F2, R1.3
OrbitNumber	(For swath data – Level 1 & 2) Sequential number assigned to satellite orbits (integer).	GES DISC	Discovery, Geolocation	UMM-C, UMM-G, UMM-S, UMM-T,	F2, R1.3
EquatorCrossingLongitude	(For swath data – Level 1 & 2) Longitude at which the satellite crosses the equator (decimal degrees).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T	F2, R1.3
EquatorCrossingDate	(For swath data – Level 1 & 2) Date on which the satellite crosses the equator (YYYY-MM-DD).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T	F2, R1.3
EquatorCrossingTime	(For swath data – Level 1 & 2) Time at which the satellite crosses the equator (UTC hh:mm:ss).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T	F2, R1.3
StartLatitude	(For swath data – Level 1 & 2) Nadir latitude at start of swath (+ or – 90 degrees).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1.3
StartDirection	(For swath data – Level 1 & 2) Orbit direction at StartLatitude (A = ascending; D = descending).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S,	F2, R1.3

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
				UMM-T, UMM-Var	
EndLatitude	(For swath data – Level 1 & 2) Nadir latitude at end of swath (+ or – 90 degrees).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1.3
EndDirection	(For swath data – Level 1 & 2) Orbit direction at EndLatitude (A = ascending; D = descending).	GES DISC	Discovery, Geolocation	UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1.3
NumberofOrbits	(For swath data – Level 1 & 2) Number of orbits for the swath (needed if a file has more than 1 orbit).	GES DISC	Discovery, Geolocation	UMM-C, UMM-G	F2, R1.3
StartOrbit	(For swath data – Level 1 & 2) The start orbit number (needed if a file contains multiple orbits).	GES DISC	Discovery, Geolocation	UMM-C, UMM-G	F2, R1.3
StopOrbit	(For swath data – Level 1 & 2) The stop orbit number (needed if a file contains multiple orbits).	GES DISC	Discovery, Geolocation	UMM-C, UMM-G	F2, R1.3
metadata_link	A URL that gives the location of complete metadata. A persistent URL is recommended. The data-collection DOI is preferred for this link.	ACDD 1.3	Discovery, Provenance	UMM-C	F2, F3, A1, I3, R1
product_version (VersionID)	Version identifier of the data as assigned by the data creator. For example, a new algorithm or methodology could result in a new version of a product.	ACDD 1.3	Discovery, Provenance	UMM-C, UMM-G, UMM-Var	F2

D.3 Geolocation

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
geospatial_bounds (SpatialCoverage, ObservationArea)	Describes the data's 2D or 3D geospatial extent in OGC's Well-Known Text (WKT) Geometry format (see the OGC Simple Feature Access specification [81]). The meaning and order of values for each point's coordinates depends on the CRS. The ACDD default is "2D geometry" in the "EPSG:4326" CRS. The default may be overridden with geospatial_bounds_crs and geospatial_bounds_vertical_crs. EPSG:4326 coordinate values are "latitude (decimal degrees_north)" and "longitude (decimal degrees_east)," in that order. Longitude values in the default case are limited to the (-180, 180) range. Example: "POLYGON (40.26 -111.29, 41.26 -111.29, 41.26 -110.29, 40.26 -110.29, 40.26 -111.29)."	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_bounds_crs (SpatialCoverage, ObservationArea)	The CRS of the point coordinates in geospatial_bounds. This CRS may be 2D or 3D, but together with geospatial_bounds_vertical_crs (if supplied), must match the dimensionality, order, and meaning of point coordinate values in geospatial_bounds. If geospatial_bounds_vertical_crs is also present, then only specify a 2D CRS. EPSG CRSs are strongly recommended. If this attribute is not specified, the CRS is assumed to be "EPSG:4326." Examples: "EPSG:4979" (the 3D WGS84 CRS), "EPSG:4047."	ACDD 1.3	Geolocation	UMM-C, UMM-Var	F2, R1
geospatial_bounds_vertical_crs (SpatialCoverage, ObservationArea)	The vertical CRS for the Z axis of the point coordinates in geospatial_bounds. This attribute cannot be used if the CRS in geospatial_bounds_crs is 3D. To use this attribute, geospatial_bounds_crs must exist and specify a 2D CRS. EPSG CRSs are strongly recommended. There is no default for this attribute when not specified. Examples: "EPSG:5829" (instantaneous height above sea level), "EPSG:5831" (instantaneous depth below sea level), "EPSG:5703" (NAVD88 height).	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1, R1.3

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
geospatial_lat_min (SouthernmostLatitude)	Describes a simple lower latitude limit that may be part of a 2D or 3D bounding region. geospatial_lat_min specifies the southernmost latitude covered by the data.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1, R1.3
geospatial_lat_max (NorthernmostLatitude)	Describes a simple upper latitude limit that may be part of a 2D or 3D bounding region. geospatial_lat_max specifies the northernmost latitude covered by the data.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_lat_units (SpatialCoverage)	Units for the latitude axis described in geospatial_lat_min and geospatial_lat_max. These are presumed to be “degrees_north,” but other options from the UDUNITS package [82] can be specified instead.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_lat_resolution (LatitudeResolution)	Information about the targeted spacing of points in latitude. Describing resolution as a number value followed by the units is recommended. For L1 and L2 swath data, this is an approximation of the pixel resolution. Examples: “100 meters,” “0.1 degree.”	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_lon_min (WesternmostLongitude)	Describes a simple longitude limit and can be part of a 2D or 3D bounding region. geospatial_lon_min specifies the westernmost longitude covered by the data. See also geospatial_lon_max.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
geospatial_lon_max (EasternmostLongitude)	Describes a simple longitude limit; may be part of a 2D or 3D bounding region. geospatial_lon_max specifies the easternmost longitude covered by the data. Cases where geospatial_lon_min is greater than geospatial_lon_max indicate the bounding box extends from geospatial_lon_max, through the longitude range discontinuity meridian (either the antimeridian for -180:180 values, or Prime Meridian for 0:360 values), to geospatial_lon_min. For example, “geospatial_lon_min=170” and “geospatial_lon_max=-175” incorporates 15 degrees of longitude (ranges 170 to 180 and -180 to -175).	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_lon_units (SpatialCoverage)	Units for the longitude axis described in geospatial_lon_min and geospatial_lon_max. These are presumed to be "degrees_east," but other options from the UDUNITS package [82]can be specified instead.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_lon_resolution (LongitudeResolution)	Information about the targeted spacing of points in longitude. Describing resolution as a number value followed by units is recommended. For L1 and L2 swath data, this is an approximation of the pixel resolution. Examples: “100 meters,” “0.1 degree.”	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_vertical_min (SpatialCoverage)	Describes the numerically smaller vertical limit and can be part of a 2D or 3D bounding region. See also geospatial_vertical_positive and geospatial_vertical_units.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_vertical_max (SpatialCoverage)	Describes the numerically larger vertical limit and can be part of a 2D or 3D bounding region. See also geospatial_vertical_positive and geospatial_vertical_units.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
geospatial_vertical_resolution (DataResolution)	Information about the targeted vertical spacing of points. Example: “25 meters.”	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_vertical_units (SpatialCoverage)	Units for the vertical axis described in geospatial_vertical_min and geospatial_vertical_max. The default is “EPSG:4979” (height above the ellipsoid, in meters), but other vertical coordinate reference systems may be specified. Note that the common oceanographic practice of using pressure for a vertical coordinate, while not strictly a depth, can be specified using the unit bar. Examples: “EPSG:5829” (instantaneous height above sea level), “EPSG:5831” (instantaneous depth below sea level).	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
geospatial_vertical_positive (SpatialCoverage)	Values include either “up” or “down.” If “up,” vertical values are interpreted as “altitude,” with negative values corresponding to below the reference datum (e.g., under water). If “down,” vertical values are interpreted as “depth,” positive values correspond to below the reference datum. Note that if geospatial_vertical_positive is “down” (depth orientation), geospatial_vertical_min specifies the data’s vertical location furthest from the Earth’s center, and geospatial_vertical_max specifies the location closest to the Earth’s center.	ACDD 1.3	Geolocation	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1

D.4 Temporal Extent

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
time_coverage_start (RangeBeginningDate, and RangeBeginningTime)	Describes the time of the first data point in the data. Use the ISO 8601:2004 date format, preferably the extended format as recommended in ACDD Section 2.6 [80].	ACDD 1.3	Temporal Location	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
time_coverage_end (RangeEndingDate, and RangeEndingTime)	Describes the time of the last data point in the data. Use ISO 8601:2004 date format, preferably the extended format as recommended in ACDD Section 2.6 [80].	ACDD 1.3	Temporal Location	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
time_coverage_duration (TemporalRange)	Describes the duration of the data. Use ISO 8601:2004 duration format, preferably the extended format as recommended in ACDD Section 2.6 [80].	ACDD 1.3	Temporal Location	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1
time_coverage_resolution (TemporalRange)	Describes the targeted time period between each value in the data. Use ISO 8601:2004 duration format, preferably the extended format as recommended in ACDD Section 2.6 [80].	ACDD 1.3	Temporal Location	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1

D.5 Usability

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
cdm_data_type	The data type, as derived from Unidata's Common Data Model (CDM) Scientific Data types and understood by Thematic Real-time Environmental Distributed Data Services (THREDDS). This is a THREDDS dataType, and is different from the CF featureType, which indicates a Discrete Sampling Geometry file.	ACDD 1.3	Usability	UMM-S, UMM-T	A1.1
comment	Miscellaneous information about the data not captured elsewhere.	CF 1.7, ACDD 1.3,	Usability	UMM-C, UMM-S, UMM-T, UMM-Var	F2, R1
acknowledgement	A place to acknowledge various types of support for the project that produced the data.	ACDD 1.3	Usability	UMM-C	F2, R1
license	Provide the Uniform Resource Locator (URL) to a standard or specific license, enter "Freely Distributed" or "None," or describe any restrictions to data access and distribution in free text.	ACDD 1.3	Usability	UMM-C UMM-S, UMM-T, UMM-Var	R1.1
DataSetQuality	Overall assessment of quality of data, including relevant articles. Short summary is preferred.	GES DISC	Usability	UMM-C, UMM-S, UMM-T, UMM-Var	F2, R1
DataProgress	Status of dataset.	GES DISC	Usability	UMM-C, UMM-S, UMM-T	F2, R1
SpatialCompletenessDefinition	Definition of a measure of data quality: e.g., the ratio of grid elements containing valid values to total number of grid elements.	GES DISC	Usability	UMM-C, UMM-S, UMM-T	F2, R1, R1.3
SpatialCompletenessRatio	The data quality information: value for the above-defined measure.	GES DISC	Usability	UMM-C, UMM-S, UMM-T	F2, R1, R1.3

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
FOVResolution	(For swath data – Level 1 & 2) Field-of-view resolution of sensor used to acquire the swath data (if multiple sensors, list FOVResolution of each sensor).	GES DISC	Usability	UMM-C, UMM-G, UMM-S, UMM-T	F2, R1, R1.3

D.6 Provenance

D.6.1 General

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
uuid	A Universal Unique Identifier (uuid) is a 128-bit number used to uniquely identify some object or entity on the web. Depending on the specific mechanisms used, a uuid is either guaranteed to be different or at least extremely likely to be different from any other uuid generated until 3400 A.D. Applied to identify the file.	NASA ESDIS	Provenance (General)	UMM-C	F1, A1
date_created	The date on which the current version of the data was created. Modification of values implies a new version; hence, this would be assigned the date of the most recent values modification. Metadata changes are not considered when assigning this attribute. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [80].	ACDD 1.3	Provenance (General)	UMM-C	F2, R1
ProductionDateTime	Date and time the file was produced. This attribute is similar to date_created. The attribute “date_created” is the preferred attribute.	GES DISC	Provenance (General)		F2, R1
date_modified	The date on which the data was last modified. Note that this applies just to the data, not the metadata. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [80].	ACDD 1.3	Provenance (General)	UMM-C	F2, R1
date_issued	The date on which the data (including all modifications) was formally issued (i.e., made available to a wider audience). Note that these apply just to the data, not the metadata. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [80].	ACDD 1.3	Provenance (General)	UMM-C	F2, R1
date_metadata_modified	The date on which the metadata was last modified. The ISO 8601:2004 extended date format is recommended, as described in ACDD Section 2.6 [80].	ACDD 1.3	Provenance (General)	UMM-C	F2, R1
ValidationData	Description of or reference on how the data were validated.	GES DISC	Provenance (General)	UMM-C	R1

D.6.2 Attribution

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
id (IdentifierProductDOI)	An identifier for the data product, provided by and unique to its naming authority. The combination of naming_authority and ID should be globally unique, but it can be globally unique per se. It can be web addresses, DOIs, meaningful text strings, a local key, or any other unique string of characters. It should not include whitespace characters.	ACDD 1.3	Discovery Provenance (Attribution)	UMM-C	F1, A1
naming_authority (IdentifierProductDOIAuthority)	The organization that provides the initial identifier for the data. The naming authority should be uniquely specified by this attribute. It is recommended to use Reverse Domain Name System (e.g., [115]) for the naming authority. URIs are also acceptable (e.g., “edu.ucar.unidata”).	ACDD 1.3	Provenance (Attribution)	UMM-C	F1, A1
creator_name (ContactPersonName)	The name of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
creator_email (ContactPersonEmail)	The email address of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
creator_url (RelatedURL)	The URL of the person (or other creator type specified by creator_type) principally responsible for creating the data.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
creator_type (ContactPersonRole)	Specifies the type of creator with one of the following: person, group, institution, or position. If this attribute is not specified, the creator is assumed to be a person.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
creator_institution (ContactPersonAddress)	The creator’s institution. The value should be specified even if it matches the value of publisher_institution or if creator_type is an institution.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
institution	The name of the institution principally responsible for originating the data.	CF 1.7, ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
project	The name of the project principally responsible for originating the data. Examples: “PATMOS-X,” “Extended Continental Shelf Project.”	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
program	The overarching program of which the data belongs. A program consists of a set (or portfolio) of related and possibly interdependent projects that meet an	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
	overarching objective. Examples: “GHRSST,” “NOAA CDR,” “NASA EOS,” “JPSS,” “GOES-R.”				
contributor_name	The name of any individuals, projects, or institutions that contributed to the creation of the data. Can be presented as free text, or in a structured format compatible with conversion to NcML (e.g., insensitive to changes in whitespace, including end-of-line characters).	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
contributor_role	The role of any individuals, projects, or institutions that contributed to the creation of the data. Can be presented as free text, or in a structured format compatible with conversion to NcML (e.g., insensitive to changes in whitespace, including end-of-line characters). Multiple roles should be presented in the same order and number as the names in contributor_names.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
publisher_name	The name of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
publisher_email	The email address of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
publisher_url (RelatedURL)	The URL of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
publisher_type	Specifies type of publisher with one of the following: person, group, institution, or position. If this attribute is not specified, the publisher is assumed to be a person.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1
publisher_institution (ProcessingCenter)	The publisher’s institution. If publisher_type is an institution, this attribute should have the same value as publisher_name.	ACDD 1.3	Provenance (Attribution)	UMM-C	F2, R1

D.6.3 Lineage

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
InputDataProductVersion	Input data version.	GES DISC	Provenance (Lineage)	UMM-C	F2, R1, R1.2
InputDataProducts	Input data to the product of interest.	GES DISC	Provenance (Lineage)	UMM-C, UMM-Var	F2, R1, R1.2
history (history, ProductGenerationAlgorithm, and ProductGenerationAlgorithmVersion)	Provides an audit trail for modifications to the original data. This attribute is also in the NUG: "This is a character array with a line for each invocation of a program that has modified the dataset. Well-behaved, generic netCDF applications should append a line containing: date, time of day, username, program name, and command arguments." To include a more complete description, append a reference to an ISO Lineage entity (see NOAA EDM ISO Lineage guidance [116]).	CF 1.7, ACDD 1.3	Provenance (Lineage)		F2, R1, R1.2
source (source, InputOriginalFile)	The method of production of the original data. If it was model generated, this attribute should provide the model and its version. If it is observational, this attribute should characterize it. Examples: "temperature from CTD #1234," "world model v.0.1."	CF 1.7, ACDD 1.3	Provenance (Lineage)		F2, R1, R1.2
references	Published or web references that describe the data or methods used to produce the data. Recommend are URIs (such as a URL or DOI) for manuscripts or other references.	CF 1.7, ACDD 1.3	Provenance (Lineage)		F2, R1, R1.2
PGEVersion	Version of the PGE that the product was generated from.	LAADS DAAC	Provenance (Lineage)		F2, R1, R1.2
PGE_Name	Value like "PGExxx" where xxx is a number assigned to the PGE.	LAADS DAAC	Provenance (Lineage)		F2, R1, R1.2

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
ProcessingEnvironment	Information on the machine where the code was executed.	LAADS DAAC	Provenance (Lineage)		F2, R1, R1.2
PGE_StartTime	Value reflects the time at which the processing was started. Should be a value like YYYY-MM-DD HH:MM:SS.	LAADS DAAC	Provenance (Lineage)		F2, R1, R1.2
PGE_EndTime	Value reflects the time at which the processing was completed. Should be a value like YYYY-MM-DD HH:MM:SS.	LAADS DAAC	Provenance (Lineage)		F2, R1, R1.2
publisher_url (RelatedURL)	The URL of the person (or other entity specified by publisher_type) responsible for publishing the data to users with its current metadata and format.	ACDD 1.3	Provenance (Attribution)		F2, R1
publisher_type	Specifies type of publisher with one of the following: person, group, institution, or position. If this attribute is not specified, the publisher is assumed to be a person.	ACDD 1.3	Provenance (Attribution)		F2, R1
publisher_institution (ProcessingCenter)	The publisher's institution. If publisher_type is an institution, this attribute should have the same value as publisher_name.	ACDD 1.3	Provenance (Attribution)		F2, R1

APPENDIX E. IMPORTANT VARIABLE-LEVEL ATTRIBUTES

In this appendix, we provide a list of recommended attributes constituting variable level metadata according to CF (Version 1.10) [29] and ACDD (Version 1.3) [80] conventions. These are derived primarily from the CF and ACDD recommendations by PO.DAAC [111]. The attributes listed by GES DISC, from reference [6], are shown in parentheses in the Attribute Name column. It is to be noted that while there is considerable commonality in the attribute sets called for by the different DAACs, there are also some differences.

As in Appendix D, these recommended attributes have been characterized with a justification for use, labeled with how they support *Findability, Accessibility, Interoperability, and Reusability (FAIR)* [13] [113], and mapped to UMM. The attributes are grouped based on the justification (i.e., what purpose they serve from the point of view of a data user). The column titled UMM-(X) lists the UMM profiles discussed in Section 4.1 that use the specific attributes shown in the corresponding rows. The column titled FAIR lists the mappings of FAIR principles to the attributes. The detailed descriptions of FAIR principles can be found in [113]. Note that all attributes whose sources are CF or ACDD map to FAIR principles I1, I2, and R1.3. Therefore, these are not repeated in the FAIR column.

Additionally, the order of variable dimensions has standardized recommendations. If any or all of the dimensions of a variable have the interpretations (as given by their units or axis attribute) of time (T), height or depth (Z), latitude (Y), or longitude (X) then those dimensions should appear in the relative order T, then Z, then Y, then X in the CDL definition corresponding to the file.

There are also recommendations for the construction of variable names (http://cfconventions.org/Data/cf-standard-names/docs/guidelines.html). And a standard name look-up list and tool to assist with ensuring the variables are consistent with recommendations. http://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html

Data producers are reminded to work with the DAACs to determine which attributes are required, where they are stored in the data products, if additional attributes not in this list are necessary, and variables names and structure.

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
long_name	A descriptive (i.e., human-readable) name for the variable. Avoid including acronyms, abbreviations, and units.	CF 1.7, ACDD 1.3	Interpretability	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2
standard_name	Reserved for names that are part of the CF. If no CF standard name is appropriate for the variable, this attribute should be excluded. However, the CF is continually evolving: to get a standard name added to the CF, propose one by emailing to cf-metadata@cgd.ucar.edu.	CF 1.7	Interpretability	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2
coverage_content_type	An ISO 19115-1 code to indicate the source of the data, MD_CoverageContentTypeCode [116]. Examples: “image,” “thematicClassification,” “physicalMeasurement,” “auxiliaryInformation,” “qualityInformation,” “referenceInformation,” “modelResult,” “coordinate.”	ACDD 1.3	Interpretability	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	F2, R1.2
units	This attribute is required for all variables that represent dimensional quantities (see [31], Rec. 3.3). We recommend adhering to the CF on the use of the units attribute with the following clarifications: a unitless (i.e., dimensionless, in the physical sense) variable is indicated by excluding this attribute, unless appropriate physical units do exist, then use of dimensionless units identifiers is common practice in the target user community. Values of units should be supported by the UDUNITS-2 library [82]. A variable used in any context other than data storage must not define units.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	
_FillValue	Include only if the variable has missing values. The value should be the same as that of the variable. Using NaN (Not a Number) to specify this attribute is discouraged (see [31], Rec. 3.7). A data producer can create a separate array to document the reasons for missing values by using flag_values and flag_meanings.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	
coordinates	This is necessary if the variable has coordinates that are not the same as the dimension names of the data. In this case, the coordinates attribute identifies the auxiliary variables that contain geospatial or temporal coordinates. For example,	CF 1.7	Data Services		

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
	variables on a trajectory often use the data acquisition time as a single dimension but have auxiliary coordinates that record the latitude and longitude for each datum in the main variable. If a variable “foo” has a coordinate of time but co-exists with variables named “lat” and “lon” with the corresponding geolocation information, then the coordinates variable would read “lat lon” to indicate that the variables hold the appropriate geospatial location information. Without this attribute, tools will not be able to locate the corresponding latitude and longitude for data in a swath, trajectory, or non-rectilinear grid. Note that the auxiliary coordinate variables must still follow the conventions for units that are noted in Section 4.4 and Section 4.5 above. Also, note that the only delimiter between words inside the coordinates attribute should be a space.				
bounds	An attribute for coordinate variables to describe the vertices of the cell boundaries and thus the intervals between cells	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
scale_factor	This is often used to represent floating point numbers as short integers, thus resulting in more compact data (i.e., packed data). To convert the short integer value, it is multiplied by scale_factor and then add_offset is added. These attributes should be floating point numbers, not strings, to work properly. If the scale_factor is “1.0” and add_offset is “0.0,” these attributes are omitted.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
add_offset (addscale_offset)	This is often used to represent floating point numbers as short integers, thus resulting in more compact data (i.e., packed data). To convert the short integer value, it is multiplied by scale_factor and then add_offset is added. These attributes should be floating point numbers, not strings, to work properly. If the scale_factor is “1.0” and add_offset is “0.0,” these attributes should be omitted.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
positive	This is the direction of increasing vertical coordinate values, with a valid value of either up or down.	GES DISC	Data Services	UMM-C, UMM-S, UMM-T, UMM-Var	R1.2

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
valid_min	A scalar specifying the minimum valid value for the variable.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
valid_max	A scalar specifying the maximum valid value for the variable.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
valid_range	A vector of two numbers specifying the minimum and maximum valid values for the variable, equivalent to specifying values for both valid_min and valid_max attributes. The attribute valid_range should not be defined if either valid_min or valid_max are defined.	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
grid_mapping	Describes the horizontal coordinate system. This attribute should indicate a variable that contains the parameters corresponding to the coordinate system. There are typically several parameters associated with each coordinate system. The CF defines a separate attribute for each of the parameters. Examples: “semi_major_axis,” “inverse_flattening,” “false_easting.”	CF 1.7	Data Services	UMM-S, UMM-T, UMM-Var	R1.2
flag_values	An enumerated list of status flags indicating unique conditions whose meaning is described by the commensurate list of descriptive phrases in attribute flag_meanings. The status flags are scaler of the same type as the described variable.	CF 1.7	Quality Filtering	UMM-S, UMM-T, UMM-Var	F2, R1.2
flag_masks	A number of independent Boolean (i.e., binary) conditions using bit field notation and setting unique bits whose values are associated with a list of descriptive phrases in attribute flag_meanings. This attribute is the same type as the variable and contains a list of values matching unique bit fields.	CF 1.7	Quality Filtering	UMM-S, UMM-T, UMM-Var	F2, R1.2
flag_meanings	A list of strings that define the physical meaning of each flag_masks bit field or flag_values scalar field. The strings are often phrasing with words concatenated with underscores, and strings are separated by a single space. The CF allows a single variable to contain both flag_values and flag_masks. In such cases, the interpretation of the flags is slightly tricky. flag_masks is used to “group” a set of flag_values into a nested conditional. See [29], Section 3.5 on how to	CF 1.7	Quality Filtering	UMM-S, UMM-T, UMM-Var	F2, R1.2

Attribute Name	Definitions	Source	Justification	UMM-(X)	FAIR
	interpret flag_meanings. It is recommended that Boolean (i.e., flag_masks) and enumerated flags (i.e., flag_values) be kept in separate variables.				
Comment (comments)	Provides the data producer an opportunity to further describe the variable and inform the user of its contents via a text statement.	CF 1.7	Usability	UMM-C, UMM-G, UMM-S, UMM-T, UMM-Var	R1.2