11A.3 INVESTIGATION OF USING HDF5 ARCHIVAL INFORMATION PACKAGES (AIP) TO STORE NASA ECS DATA

MuQun Yang*[1], Ruth Duerr[2], Choonghwan Lee [1]
[1]The HDF Group,
[2]National Snow and Ice Data Center (NSIDC), University of Colorado

1. INTRODUCTION

The number and volume of remote sensing data products generated per day can be very large, and the data must be accessible for generations to come. Several challenges must be addressed to ensure the long-term accessibility of these data products.

Due to the complicated internal data structures for many of the data products, the I/O libraries used to create and read the products are quite complex. This complexity, in combination with the rapid pace of change in computing technologies such as compilers and hardware, make long-term maintenance of the I/O libraries challenging and expensive.

The metadata needed to describe the remote sensing data products are also complicated and diverse. Transferal of the archive responsibility from one agency to another and lack of standard metadata also make it challenging to manage, distribute, and archive these data.

In this paper we present an alternative approach for archiving remote sensing data that does not require the long-term maintenance of multiple versions of multiple I/O libraries. Our approach is to migrate the data to a single standards-based archive format, which also helps mitigate the transfer of custody problem. Data from NASA's Earth Observing System (EOS) Data Centers formatted in Hierarchical Data Format (HDF) (1) are used.

The paper also discusses technical challenges in the access of NASA HDF data. Many widely used data access tools — such as Unidata's Integrated Data Viewer (IDV) (2), the NOAA Ferret data visualization tool (3), and the Grid Analysis and Display System (GrADS) (4) follow the Network Common Data Form (netCDF) data model (5) and Climate and Forecast (CF) Metadata Conventions (6). Full support for direct access to NASA HDF-EOS (7) data via these tools is desired throughout the NASA Earth Science user community.

*Corresponding author's address:
MuQun Yang, The HDF Group,
1901 S. First St., Suite C-2
Champaign, IL 61820
E-mail: myang6@hdfgroup.org

We hope our discussions will benefit other developers who would like to support access to NASA HDF data via their tools.

2. BACKGROUND

EOS is an ambitious NASA program to enable an improved understanding of the Earth as an integrated system by providing long-term global earth observations. The program includes a variety of sensors on several satellites; a distributed data system that processes, archives, and distributes data products about fundamental properties of the land surface, atmosphere, and oceans; and a globally distributed science community fully vested in the creation and use of EOS products (8).

Since the launch of the first NASA EOS satellite, Terra, in 1999, upwards of three terabytes of EOS earth science data have been produced and archived by the EOS Data and Information System (EOSDIS) (9) every day. Substantial amounts of these data are in HDF-EOS2 (based on the HDF4 library) or HDF-EOS5 (based on the HDF5 library) formats, and are expected to eventually transfer into NOAA's Comprehensive Large Array-data Stewardship System (CLASS) (10).

HDF is a general-purpose format and library, designed to support scientific data management and high-performance computing. It was developed at the National Center for Supercomputing Applications at the University of Illinois, Urbana-Champaign, and is now owned and maintained by The HDF Group. HDF version 4 (HDF4) (11) has been widely used to store and distribute NASA Earth Science Data. The HDF-EOS2 profile and library were built on HDF4, and HDF-EOS2 is the standard format for many of the products from EOSDIS.

HDF5, introduced in 1998, represents an important technological leap that replaced the aging HDF4 technology with a new format and library offering a more flexible data model, improved scalability, and high-performance features including support for parallel I/O (12).

The HDF4 and HDF5 libraries are quite different from each other, and both are complex, requiring special expertise to maintain. In order to fully support HDF-EOS data, the higher-level HDF-EOS2 and HDF-EOS5 libraries must also

be supported. In other words, in order to support NASA EOS data in its native form, a total of four libraries must be maintained indefinitely, a task that NASA continues to support. However, once EOS data migrates to NOAA, it is unlikely that NASA will continue to fund the library support. The alternative, demonstrated by this paper, is to convert these data into a standard archive format during migration to the NOAA CLASS system.

NetCDF-4 (13) is a new version of netCDF that includes an option to use HDF5 as its data storage layer within the existing netCDF3 interfaces. NetCDF-4, developed under the sponsorship of NASA, was designed to combine the widespread use and simplicity of netCDF with the generality and performance of HDF5. The netCDF-4 library is compatible with existing netCDF programs, has flexible data-modeling abstractions, and has features for performance and storage efficiency such as chunking, compression, and parallel I/O. NetCDF-4 interfaces can read and write HDF5 files, with some restrictions. Many NOAA data products, such as the NOAA NCEP Reanalysis products, monthly soil moisture, etc., are stored in netCDF.

Common Data Model (CDM) (14), which was developed at Unidata, is a unification of the data models of the Open-source Project for a Network Data Access Protocol (OPeNDAP) (15), netCDF, and HDF5. CDM provides an abstract data model and interface for applications that may need to deal with a wide range of data formats (e.g., HDF4, HDF5, netCDF, GRIB, BUFR, etc.) and access protocols (e.g., OPeNDAP). One implementation of CDM uses netCDF-3/netCDF-4 interfaces to access HDF5 and OPeNDAP files.

A major contribution of our work is the conversion of HDF-EOS2 data into HDF5 data that can be accessed using netCDF-4, which forms the backbone of the current CDM.

The Reference Model for an Open Archival Information System (OAIS) (16) has been accepted as an ISO standard and widely adopted by archives of all kinds. Two conditions are required for an archive to be considered OAIS compliant. First, the archive must assume the responsibilities of an OAIS, and second, the archive must support the OAIS information model. A major component of the information model is the Archive Information Package (AIP). It defines the additional information that needs to be archived if an object is to be understandable to its user community both now and in the future. This additional information includes representation information that defines the structural and semantic content of the object, along with its preservation description. The PREservation Metadata Implementation Strategies Working Group (PREMIS) (17) released a model and schema for preservation metadata based on OAIS.

There are two categories of metadata for each remote sensing data set: catalog metadata, which is used to store information about a data set as a whole, and inventory metadata, which is used to store and provide access to information about individual files. The responsibility for transferring data from one institution to another should include the transfer of the data itself, as well as its catalog and inventory metadata. Moderate Resolution Imaging Spectroradiometer (MODIS) HDF-EOS2 data sets provide an excellent test of what it would take to transfer archival responsibility from one organization to another, and were used in our work.

Developed by The HDF Group for the purpose of archiving HDF5 data, the HDF5 Archival Information Package (HDF5-AIP) (18) consists of a metadata file that complies with the Metadata Encoding & Transmission Standard (METS) (19) and an HDF5 file. METS provides a coherent overall structure for encoding all relevant types of metadata. It is an internationally accepted standard in the digital library world, and a variety of XML-based tools and software used to create and administer METS files are freely available. Interestingly, it can also incorporate metadata using other standards, in particular PREMIS metadata. The intent is that the resulting HDF5-AIP files meet the needs of long-term preservation at the file level and also be easily accessible through netCDF-4 interfaces.

3. PROJECT DESCRIPTION

This project is comprised of two main activities: 1) to develop or modify prototype software to produce HDF5-AIP files that are also CDM compliant, and 2) to independently test the tools developed for both correctness and performance using a number of types of EOS data products archived at the NSIDC. These activities are described in further detail in the next sections.

3.1 Prototype and Development Activities

The prototype software consists of three components. First, The HDF Group currently maintains an HDF4-to-HDF5 conversion utility based on a default mapping between HDF4 and HDF5 data objects. This utility is unaware of the extra layers of meaning and association present in files created using HDF-EOS2. The current conversion utility uses object reference types to represent the relations among HDF5 datasets and HDF5 dimension scales. The current netCDF-4 library can only read HDF5 files that

describe all dimensions with dimension scales that conform to the HDF5 dimension scale specification (20). Consequently, the HDF5 files converted from HDF4/HDF-EOS2 files by the current utility cannot be read using the current netCDF-4 library. An enhanced HDF4-to-HDF5 conversion tool that can generate netCDF-4-compliant HDF5 files is necessary.

Second, it is the intent of this project to demonstrate whether the recently defined HDF5-AIP is up to the challenge of preserving the complexity in metadata that is necessary for the successful migration of archived NASA EOS Core System (ECS) data to other archival systems. Towards this end, a tool is being developed that will extract the metadata from an ECS system and format it according to the METS standard.

Third, the NSIDC data catalog currently contains metadata about each of NSIDC's publicly available datasets and is the source of the dataset information for NSIDC's website. NSIDC developed an extraction tool that converts entries in the NSIDC catalog into a format that follows the Content Standard for Digital Geospatial Metadata (CSDGM) (21). However, this tool does not capture catalog-level metadata from the data sets Earth Science Data Types (ESDTs) described in NSIDC's ECS system or the plethora of ancillary information necessary to provide a complete Preservation Information Package. Moreover, the international equivalent to the CSDGM standard, ISO 19115(22), was recently approved, and, as a result, a tool needs to be developed to extract catalog-level information from the ECS and NSIDC systems in a format compatible with ISO 19115.

3.2 Testing and Demonstration Activities

Once development is complete, this suite of tools will be used to convert a variety of HDF-EOS2-formatted products to HDF5-AIP files. NSIDC currently archives 20 different MODIS Level 2 and 3 data sets, as well as another 16 HDF-EOS2 data sets from Advanced Microwave Scanning Radiometer – Earth Observing System (AMSR-E). This variety of data products covers a wide range of product sizes and HDF file configurations, and thus represents not only a good first test of the tools developed but also an AIP for a remote sensing data set.

A comprehensive test plan is being implemented to verify that the conversions have been properly accomplished without a loss of information, that the resulting files are accessible using standard HDF5 and netCDF-4 user interfaces, that the resulting files are compliant with the relevant standards, and that the conversion performance is acceptable. Testing will begin with a single MODIS ESDT, and then continue with a wide variety of products to ensure broad applicability of the tools.

4. PROJECT REPORT

Of the three development activities discussed previously, only the work related to enhancing the HDF4-to-HDF5 converter is complete. In this section, we describe this work and the challenges that we encountered during the process of enhancing the HDF4-to-HDF5 converter.

We present an overview of challenges in section 4.1, and in section 4.2 we address the details of challenges related to handling HDF-EOS2 objects. In section 4.3 we address the challenges related to making HDF5 files accessible by netCDF-4 APIs. Section 4.4 gives the current status of the enhanced HDF4-to-HDF5 converter. In section 4.5, we conclude with a discussion of the conversion challenges that we hope will benefit other developers who would like to access HDF-EOS2 files directly.

4.1 Overview of Challenges

To provide a physically meaningful interpretation of remote sensing data, one not only needs to have the physical field data, such as radiant temperature, but also needs to know where those data are located, that is, the geolocation information of each data point, such as latitude and longitude.

Although most I/O packages store physical field data as multi-dimensional arrays, the geolocation information may be stored differently. The complete discussion on how geolocation information is stored by different I/O packages is beyond the scope of this paper.

In particular, the geolocation information, such as latitude and longitude, of a physical data field stored in an HDF-EOS2 file is difficult to retrieve without using HDF-EOS2 APIs. Many popular visualization tools, such as IDV and GrADS, follow the netCDF/CF model and explicitly need the geolocation information to be saved as dimensional variables in netCDF. The original HDF4-to-HDF5 converter can convert any HDF-EOS2 file to an HDF5 file by following the default HDF4-to-HDF5 mapping document (23), which doesn't consider the specialty of HDF-EOS2 data objects at all. One challenge for us is how to retrieve geolocation information from an HDF-EOS2 file. This challenge is discussed in section 4.2.

Another challenge is related to making the converted HDF5 file conform to the netCDF-4

data model. This process involves the adjustment of the file structure inside the HDF5 file, and the inclusion of the geolocation information in a way that allows applications to retrieve it. Details can be found in section 4.3.

4.2 Handling HDF-EOS2 Objects

Two types of HDF-EOS2 objects, swath and grid, are widely used to store NASA HDF-EOS2 data. More than 99% of NASA HDF-EOS2 data are either swaths or grids. In most cases, an HDF-EOS2 file includes either a grid object or a swath object but not both.

An HDF-EOS2 grid is a data object that is organized by a geographic spacing that can be computed through projection information. An HDF-EOS2 grid consists primarily of physical data fields and attributes describing those fields. The physical data fields are usually multidimensional data arrays, and are stored as HDF4 Scientific Data Sets (SDSs) or Vdata. The geolocation information is always provided implicitly by projection parameters contained in a special file attribute called StructMetadata. At times, the geolocation data is also stored as multidimensional data arrays inside an HDF-EOS2 grid. The HDF-EOS2 library, through its use of the General Cartographic Transformation Package (GCTP) library (23), allows an application to store the grid data in one of more than a dozen projections.

Figure 1(a) shows an HDF-EOS2 grid using geographic projection and Figure 1(b) shows a grid using polar stereographic projection. Network Common Data Form Language (CDL)-like descriptions of the grid structure of these HDF-EOS2 grid files are also inserted in the figure.

The example data field shown in Figure 1(a) is CloudCover. It is a 14-by-8, two-dimensional array represented by red dots. All data points along each horizontal line (parallel to dim1) share the same latitude, and all data points along each vertical line (parallel to dim2) share the same longitude. Therefore, a one-dimensional array with 14 elements can be used to describe the longitude of this 2-D data field. A one-dimensional array with 8 elements can be used to describe the latitude of this 2-D data field.

The example data field shown in Figure 1(b) is SnowDepth over the North Polar Region. It is a 9-by-7, two-dimensional array. For this projection, neither latitude nor longitude at each data point (red dots) along the grid line (parallel to dim1 or dim2) shares the same value. Therefore, a 9-by-7, two-dimensional array is needed to describe latitude and longitude for each data point. To compute the value of latitude and longitude at each data point, different formulas must be used for different projections, which is certainly not a trivial task. For most projections, it will be extremely inconvenient for each application to retrieve geolocation information without using the HDF-EOS2 library. In fact, our experience shows that only latitude and longitude values for geographic projection can be easily retrieved without the HDF-EOS2 and GCTP libraries. Therefore, in our enhanced HDF4-to-HDF5 conversion tool, we added a configuration option to compile the converter with the HDF-EOS2 library so that we can easily retrieve geolocation information by calling HDF-EOS2 APIs.

The concept of an HDF-EOS2 swath is based on a typical satellite swath, where an instrument takes a series of scans perpendicular to the ground track of the satellite as it moves along that ground track (Figure 2). An HDF-EOS2 swath also consists of physical data fields and attributes that describe those fields. Like an HDF-EOS2 grid, the data are stored as HDF4 SDS or Vdata. Unlike an HDF-EOS2 grid, geolocation fields are stored as HDF4 SDS.

In addition, HDF-EOS2 defines a concept called a *dimension map*. When dimension maps are used, the geolocation fields do not necessarily have the same extent as the physical data fields. The dimension map provides sufficient information to allow the geolocation information at each data point to be derived from the geolocation fields by interpolation.

HDF-EOS2 applications can optionally use a dimension map to create a swath file. In our analysis, we did find a substantial number of NASA HDF-EOS2 swath files using dimension maps. The size of geolocation fields in most of these files is less than the size of the corresponding physical data fields. Because of this, the overall swath file size is reduced. Dimension map information is also stored in StructMetadata. When the HDF-EOS2 APIs are not used, it can be difficult and time-consuming to retrieve the dimension map information from StructMetadata.

HDF-EOS2 swath examples are illustrated in the graphical portions of Figures 3(a) and (b). Figure 3(a) shows an HDF-EOS2 swath without a dimension map. Both physical data and geolocation data values are provided at data points represented by red dots. The geolocation fields are latitude and longitude. The latitude and longitude are both 11-by-3 two-dimensional arrays, the same extent as physical data field RainRate. In figure 3(b), the latitude and longitude are 11-by-2 arrays shown by red dots. However, the physical data field, RainRate, is an

11-by-3 array shown by both red and white dots. Latitude and longitude in the mid-column have to be interpolated based on information provided by the dimension map.

In summary, it will be difficult to obtain geolocation information at each point for grids and swaths without using the HDF-EOS2 library.

4.3 Conforming to NetCDF-4

NetCDF-4 can read HDF5 files with some restrictions of which there are two types. The netCDF data model imposes one type of restrictions. The second type of restrictions is related to HDF5 features that are not currently supported by the netCDF-4 library. Some restrictions may be lifted in future releases as netCDF-4 evolves (24).

A thorough discussion of these two types of restrictions is beyond the scope of this paper. Our focus is to report some of the challenges that we encountered when converting NASA HDF-EOS2 files to HDF5 files that can be accessed by netCDF-4 APIs. A summary of these challenges follows.

As previously described, HDF-EOS2 swath data and some types of gridded data contain two-dimensional geolocation information. However, the current netCDF-4 implementation only supports one-dimensional geolocation fields. Fortunately, the CF conventions allow the applications to have 2-D dimension scale data (26). For the initial implementation, the choice was made to use the Two-Dimensional Latitude, Longitude, Coordinate Variables convention (27), allowing applications to retrieve the two-dimensional geolocation fields following this convention. The inclusion of the complete two-dimensional geolocation fields may significantly increase the converted file size.

The HDF5 files produced by the new HDF4-to-HDF5 conversion tool need to match the data and programming models underlying the netCDF-4 interface. For example, the original HDF4-to-HDF5 conversion tool stored all dimensions in a predefined group, which the netCDF-4 interface could not read. NetCDF-4 requires that dimension information be stored in either the same HDF5 group as the data or in the parent group.

4.4 Current Status of Converter

To date, more than 45 data files from a total of 33 different AMSR-E and MODIS snow and ice products have been successfully converted using the enhanced HDF4-to-HDF5 conversion tool. These products cover both HDF-EOS2 grid and swath files. Five different geospatial projections have been found for HDF-EOS2 grid files.

To verify the correctness of the enhanced HDF4-to-HDF5 conversion tool, a verification tool was developed. This tool compares the converted HDF-EOS2 grid data retrieved using the netCDF-4 APIs with the original grid data retrieved using the HDF-EOS2 APIs.

In addition, a netCDF-4 tool, *ncdump*, was used to check that all objects in the generated HDF5 file could be successfully accessed via netCDF-4 APIs.

Preliminary performance results indicate that HDF-EOS2 files up to 100 megabytes in size could be converted in less than ten seconds on an Intel Xeon 3.2 GHz CPU Linux machine with 8GB of memory.

4.5 Notes to netCDF/CF Tool Developers

The experience of converting HDF-EOS2 files to netCDF-4 files exposed three hurdles that netCDF/CF tools need to overcome in order to directly access HDF-EOS2 files.

The first hurdle is the calculation of geolocation information based on different projections for HDF-EOS2 grid data. The HDF-EOS2 library, or another geographic projection library, can be used to calculate the geolocation information. Without one of these libraries, retrieval of the geolocation information can be very unwieldy.

The second issue is the interpolation or shrinking of geolocation fields for HDF-EOS2 swaths so that the geolocation information (latitude, longitude, etc.) at each data point can be derived. The use of the HDF-EOS2 library to retrieve and apply the dimension map information addresses this problem.

The third obstacle is rooted in the differences between the netCDF/CF and HDF-EOS2 data models. An HDF-EOS2 file can have multiple grids, each with its own geolocation information, such as latitude and longitude, resulting in multiple sets of geolocation information in the file. In contrast, the netCDF-3 data model only allows one set of geolocation fields in a file which must be shared by all variables in the file. Although netCDF-4 allows multiple sets of geolocation fields in a file, many netCDF/CF tools still follow the netCDF-3 data model and therefore assume one set of latitude and longitude fields in a file for all variables. Adoption of the netCDF-4 model would allow the full set of geolocation information to be accessed.

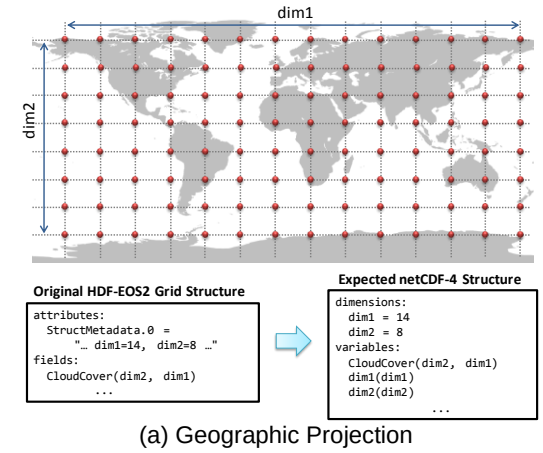
REFERENCES

1. *The HDF Group.* [Online] http://www.hdfgroup.org/.
2. *Integrated Data Viewer (IDV).* [Online] http://www.unidata.ucar.edu/software/idv/.
3. *Ferret.* [Online] http://ferret.pmel.noaa.gov/Ferret/home.
4. *Grid Analysis and Display System (GrADS).* [Online] http://www.iges.org/grads/.
5. Rew, Russ**.** NetCDF: Data Model,APIs, and Format. [Online] June 8, 2004. http://www.unidata.ucar.edu/presentations/Rew/esri-netcdf.pdf.
6. CF Metadata. *NetCDF Climate and Forecast Metadata Convention.* [Online] http://cf-pcmdi.llnl.gov/.
7. *HDF-EOS.* [Online] http://hdfeos.org/.
8. NASA's Earth Observing System. *NASA's Earth Observing System.* [Online] http://eospso.gsfc.nasa.gov/.
9. *ESDIS.* [Online] http://esdis.eosdis.nasa.gov/.
10. *NOAA's Comprehensive Large Array-data Stewardship System.* [Online] http://www.class.noaa.gov/.
11. *HDF4.* [Online] The HDF Group. http://www.hdfgroup.org/products/hdf4/.
12. *HDF5.* [Online] The HDF Group. http://www.hdfgroup.org/HDF5/.
13. *NetCDF-4 (network Common Data Form, version 4).* [Online] www.unidata.ucar.edu/software/netcdf/netcdf-4/.
14. *Common Data Model.* [Online] http://www.unidata.ucar.edu/software/netcdf/CDM/.
15. *OPeNDAP.* [Online] http://opendap.org/.
16. *Reference Model for an Open Archival Information System (OAIS).* Washington, DC. : Consultative Committee for Space Data Systems, 2002. CCDSD 650.0-B-1.
17. *PREMIS: PREservation Metadata Implementation Strategies.* [Online] http://www.oclc.org/research/projects/pmwg/.
18. *HDF5 AIP Project.* [Online] http://www.hdfgroup.org/projects/hdf5_aip/index.html.
19. *Metadata Encoding and Transmission Standard (METS) Official Web Site.* [Online] http://www.loc.gov/standards/mets/.
20. HDF5 Dimension Scale Specification and Design Notes. [Online] March 1, 2005. http://www.hdfgroup.org/HDF5/doc/HL/H5DS_Spec.pdf.
21. *Content Standard for Digital Geospatial Metadata.* [Online] http://www.fgdc.gov/metadata/csdgm/.
22. ISO 19115. [Online] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020.
23. Folk, Mike, McGrath, Robert E. and Yang, Kent**.** Mapping HDF4 Objects to HDF5 Objects. [Online] http://www.hdfgroup.org/HDF5/doc/ADGuide/H4toH5Mapping.pdf.
24. *General Cartographic Transformation Package (GCTP).* [Online] http://gcmd.nasa.gov/records/USGS-GCTP.html.
25. NetCDF-4.3 Requirements. *Unidata.* [Online] http://www.unidata.ucar.edu/software/netcdf/docs/reqs_4_3.html.
26. NetCDF Climate and Forecast (CF) Metadata Conventions. *CF Metadata.* [Online] http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/cf-conventions.html#coordinate-system.
27. Two-Dimensional Latitude, Longitude, Coordinate Variables. *CF Metadata.* [Online] http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.4/cf-conventions.html#id2680719.

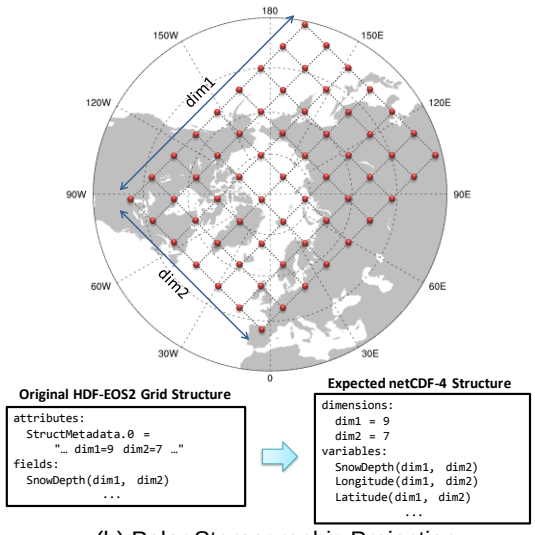
ACKNOWLEDGEMENTS

The authors would like to thank Ms. Ruth Aydt and Mr. Herb Morgan, who provided editorial assistance.

This work was supported under NOAA Scientific Stewardship Program grant number NA07OAR4310286 and under NASA grant NNX08A077A. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NOAA and NASA.



(a) Geographic Projection



(b) Polar Stereographic Projection

Figure 1. An HDF-EOS2 Grid

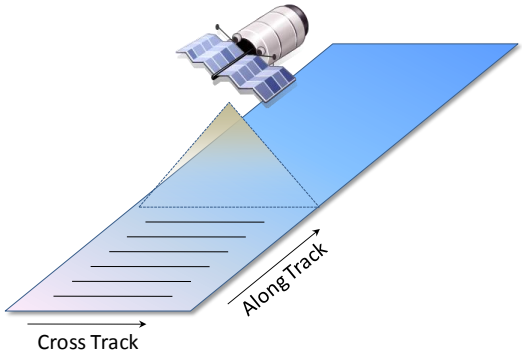
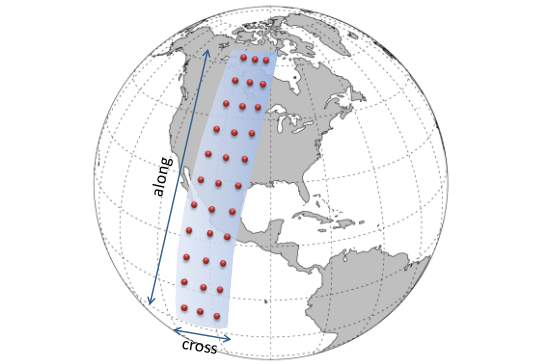
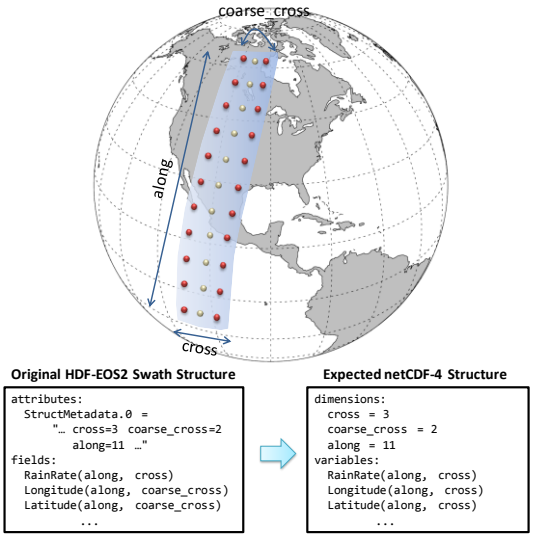


Figure 2. A Typical Satellite Swath



(a) Without dimension maps



(b) With dimension maps

Figure 1. An HDF-EOS2 Swath. Both geolocation and data values are provided at red dots; only data values are provided at white dots.