Ocean Color Data Formats and Conventions:

NASA's perspective

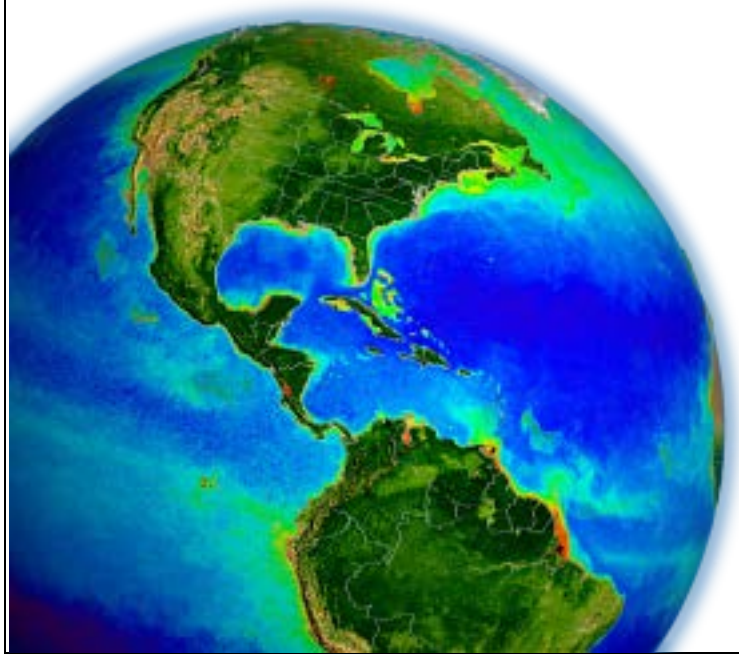
Sean Bailey

NASA Goddard Space Flight Center

07 May 2013

International Ocean Color Science Meeting

Darmstadt, Germany



The Big Picture

- The NASA Science Directorate has a *primary* goal of contributing to fundamental advances in scientific understanding of the Earth system
- To further that goal, the Earth Observing System Data and Information System (EOSDIS) is actively working toward improving the interoperability of its observations (typically in HDF format) with other observations and models (often in netCDF format)
- This is partially assured by adhering to metadata conventions, such as the Climate and Forecast (CF), International Standards Organization (ISO), and Attribute Conventions for Data Discovery (ACDD) metadata conventions.
- There is an active Earth Science Data System Working Group (ESDSWG) tasked with identifying best practices to bridge or reduce gaps between NASA-produced data and data from the broader community, and to ensure NASA data discoverability, maintainability and extensibility using CF, ISO, and ACDD conventions.

Formats*…an Alphabet Soup

- HDF (aka HDF4)
- HDF-EOS
- HDF5
- HDF5-EOS
- netCDF-Classic
- netCDF4
- GeoTIFF
- ICARTT
  - International Consortium for Atmospheric Research on Transport and Transformation
  - ASCII

* *approved for use and officially endorsed by the NASA Earth Science Division (ESD) for use in Earth science data systems*

HDF5 – A Contender

- The Ocean Biology Processing Group (OBPG) has long used HDF4…but…
- The once beautiful carriage is turning into a pumpkin:
  - "HDF (also known as HDF4) is a library and multi-object file format for storing and managing data between machines. There are two versions of HDF: HDF4 and HDF5. HDF4 is the first HDF format. *Although HDF4 is still funded*, new users that are not constrained to using HDF4, should use HDF5 ." (taken from http://www.hdfgroup.org/products/hdf4/)

netCDF4 – The Best of Both

- NetCDF-4 was NASA-funded effort to improve:
  - interoperability among scientific data representations
  - integration of observations and model outputs
  - I/O for high-performance computing.
- NetCDF-4 combines the netCDF-3 and HDF5 data models, taking the desirable characteristics of each, while taking advantage of their separate strengths:
  - NetCDF-3 is popular and easy to use, and includes many tools and multiple implementations.
  - HDF5 is powerful, has high-performance, is efficient for storage and extensible.
- The goal of netCDF-4 is to make netCDF more suitable for high-performance computing and large datasets, and to provide a simple high-level application programming interface (API) for HDF4
- (from http://www.hdfgroup.org/projects/netCDF-4/)

Standards…

- ISO-19115
  - ISO 19115 is a standard of the International Organization for Standardization (ISO). It defines how to describe geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use.
  - The objective of this International Standard is to provide a clear procedure for the description of digital geographic datasets so that users will be able to determine whether the data in a holding will be of use to them and how to access the data. By establishing a common set of metadata terminology, definitions and extension procedures, this standard will promote the proper use and effective retrieval of geographic data
- FGDC-CSDGM
  - The Federal Geographic Data Committee (FGDC) is an interagency committee that promotes the coordinated development, use, sharing, and dissemination of geospatial data on a national basis. The current Federal standard for geospatial data is the Content Standard for Digital Geospatial Metadata (CSDGM). The standard provides a common set of terminology and definitions for the documentation of digital geospatial data.

…and Conventions

- Climate and Forecast (CF)
  - The netCDF-CF (Climate and Forecast) conventions are a set of codified recommendations for practices built around published specifications. While CF is a convention rather than an established metadata standard, CF is a critically important step towards better interoperability
  - CF used the FGDC-CSDGM as a guide in choosing the values for and the attribute names of the parameters describing map projections.

The OBPG Adoption

- Migrating to netCDF4 for all products Level 2 and above
- Following CF convention
  - as closely as possible…
- Level 2 products are already available!
  - …if you make them yourself
  - l2gen code released with SeaDAS 7 includes the ability to output netCDF4 files (fmtofile=NCDF)
- Code has written to read/write netCDF4 files for the Level 3 processing – but not yet in the wild…
- …on a related note, we're taking the opportunity to modify the L3 bin file structure…

Side-by-Side

```
netcdf A2007081050500.L2_LAC.nc {
dimensions:
        Number_of_Scan_Lines = 500 ;
        Pixels_per_Scan_Line = 1354 ;
        Number_of_Pixel_Control_Points = 1354 ;
        total_band_number = 24 ;
        band_number = 16 ;

// global attributes:
                        :Title = "HMODISA Level-2 Data" ;
                        :Sensor\ Name = "HMODISA" ;

…                       :Conventions = "CF-1.6" ;
group: Sensor\ Band\ Parameters {
  variables:
        int wavelength(total_band_number) ;
                        wavelength:long_name = "Wavelengths" ;
                        wavelength:units = "nm" ;
…

group: Geophysical\ Data {
  variables:
        short Rrs_443(Number_of_Scan_Lines, Pixels_per_Scan_Line) ;
                        Rrs_443:long_name = "Remote sensing reflectance
        at 443 nm" ;
                        Rrs_443:slope = 2.e-06f ;
                        Rrs_443:intercept = 0.05f ;
                        Rrs_443:units = "sr^-1" ;
                        Rrs_443:solar_irradiance = 188.7541f ;
                        Rrs_443:bad_value_scaled = -32767s ;
                        Rrs_443:bad_value_unscaled = -0.015534f ;
...

  } // group Geophysical\ Data
```

```
netcdf A2007081050500.L2_LAC {
dimensions:
        Number of Scan Lines = 500 ;
        Number of Pixel Control Points = 1354 ;
        Pixels per Scan Line = 1354 ;
        total band number = 24 ;
        band number = 16 ;

variables:
        long year(Number of Scan Lines) ;
                        year:long_name = "Scan year" ;
                        year:valid_range = 1996, 2038 ;
                        year:units = "years" ;
        short Rrs_443(Number of Scan Lines, Pixels per Scan
        Line) ;
                        Rrs_443:long_name = "Remote sensing
        reflectance at 443 nm" ;
                        Rrs_443:slope = 2.e-06f ;
                        Rrs_443:intercept = 0.050000001f ;
                        Rrs_443:units = "sr^-1" ;
                        Rrs_443:solar_irradiance = 188.75414f ;
                        Rrs_443:bad_value_scaled = -32767s ;
                        Rrs_443:bad_value_unscaled = -0.015534002f ;
…
// global attributes:
                        :Title = "HMODISA Level-2 Data" ;
                        :Sensor Name = "HMODISA" ;
                        :Product Name = "A2007081050500.L2_LAC" ;
                        :Software Name = "l2gen" ;
                        :Software Version = "6.6.7" ;
                        :Processing Version = "Unspecified" ;
                        :Conventions = "CF-1.6" ;
```

Redesigning the Bin Files

There are a number of reasons to consider redesigning the bin files. They fall into one of four categories:

1. Data-day determination
2. Binning strategy
3. Statistics
4. File format.

Data-day determination

The current implementation uses a temporal/geographical scheme based on the nominal orbit node crossing time.

Pros:

- Done, and working

Cons:

- There are a series of complex decision blocks for each sensor l2bin accommodates. requires "tweaking" hard-coded parameters to select correct file set
- Numerous mission specific sections that are not all using consistent logic.

Suggested solution:

Implement a scheme similar to that used by the browse quick look code. This code is mostly mission independent. It accepts the node crossing time as input, and determines the disposition of each L2 pixel based on time and location.

Pros:

- Simplified logic, less prone to mission specific quirkiness Can handle the orbit drift for SeaWiFS which is not addressed by the current method

Cons:

- Will need heavy testing to ensure there are no gotchas

Binning strategy

The current bin strategy uses an integerized sinusoidal grid.

Pros:

- The devil you know.
- Efficient storage of sparse data
- Equal Area
- Allows reasonably arbitrary resolution selection
- Can be easily displayed "as is" (almost)

Cons:

- Not "nested", so identifying nearest neighbors for spatial statistics is VERY difficult
- Changing resolutions will implement statistical (and geographic) artifacts (a 9.2km bin created by "down-res"-ing a 4.6km bin will NOT be exactly the same as creating a native 4.6km bin)

Suggested solution:

- HEALPix (Hierarchical Equal Area isoLatitude Pixelization of a sphere)

Pros:

- Retain the benefits of efficient sparse data storage
- Retain the benefits of Equal Area Nested. (See cons above…)
- Allows for easy spatial outlier rejection (the SeaWiFS speckling problem).
- Code exists for easy implementation
- Code to interpret the data exists for a number of high-level programming languages: IDL, Matlab, python

Cons:

- New. Change is difficult for end users
- Not easily represented "as is"

Alternatives:

- Quadsphere
- Hierarchical Triangular Matrix
- Equal Angular Cylindrical

Statistics

- The current bin files contain only the sum of squares and weights which allow for the estimation of the sample variance (and standard deviation).
- This is simply a population statistic on the data that went into a bin.
- The l2gen program has the ability to produce an uncertainty for a number of products. This should be added to the bin files.

Pros:

- Uncertainties with the products!

Cons:

- Increases the bin file size
- Not all products have a corresponding uncertainty
- The determination of the uncertainty is less than complete/perfect
- The current binning code treats each L2 pixel as a point and bins based on the pixel center lat/lon.
- It would be better to treat the pixels as areas, and bin based on its areal weighted coverage of bins.

Pros:

- A better representation of the coverage of the data

Cons:

- Computationally complicated.

File Format & Structure

- The current format uses HDF4 VGroups/Vdata structures. These are no longer supported in HDF5, and were never well adopted in HDF4.
- There are also a number of vestigial blocks of information that can be eliminated. These include:
  - the SEAGrid structuretime_rec (part of the BinList structure)
  - sel_cat (part of the BinList structure)
  - In place of the time_rec field, we propose the addition of a mean (local) time of observation field.

Pros:

- Provide only data that is actually useful in the files :)

Cons:

- None, really.

- As we are transitioning the OBPG data products to netCDF4, we should adopt netCDF for the bin files.
- We should also make the metadata compliant with the CF standard

Pros:

- Moving to a format that will be compatible for the long term.
  - HDF4 is no longer supported by the HDF Group,
  - VGroup/Vdata structures were never well adopted anyway.
- Using a metadata standard will improve usability of the files

Cons:

- Its different, so existing code that reads these will need to be modified