Extending CF Conventions to Enhance data FAIRness for Atmospheric Composition Observations

Sean Leavor, Megan Buzanowicz, John Kusterer, Gao Chen, Morgan Silverman

ADNET Systems Inc., Bethesda, MD; NASA Langley Research Center, Hampton, VA

HDF and NetCDF File Format

- Useful for handling large volumes of data
- Able to host extensive metadata
- Frequently used with remote sensing data
  - Has been used to support numerous satellite, airborne, and ground-based measurements
- Climate and Forecast (CF) conventions give a structure for data to follow
  - Aim for self-describing data that can be read by both humans and software

Airborne HDF and NetCDF Files

- Amount and content of metadata can vary wildly between campaigns or principal investigators
  - Current guidelines are not always specific enough about what should be included with the metadata
- The Measurements of Aerosols, Clouds, and their Interactions for ESMs (MACIE) group has started an effort to build upon existing metadata conventions
  - Would help to ensure airborne data is better able to support their use for research

FAIR Principles

- Findable
  - Metadata and data should be easy to find for both humans and computers
- Accessible
  - Once found, the user needs to know how data can be accessed
- Interoperable
  - Vocabulary describing the data allows it to be integrated with other data or applications for analysis, storage, or processing
- Reusable
  - Metadata and data should be well-described so that it can be easily understood or replicated for future reuse

Goals

Primary Goal: Build upon FAIR principles to enhance the interoperability and reusability of data.

This can be achieved by:

- Building upon current Climate and Forecast (CF) conventions
- Leveraging attributes to improve metadata representation
- Adding domain relevant metadata attributes to support data (re)use
- Making sure variables are dimension scaled
- Ensuring data use standardized units
- Making arrangements for standardized uncertainty reporting

Airborne HDF/NetCDF Metadata Issues

- Metadata content can be inconsistent between campaigns or even between principal investigators (PIs) of a single campaign
- Data is not interoperable because it does not have a predictable structure
- File metadata can be missing
  - Descriptive information – PI Information, Measurement Location/Time
  - Global attributes or variable information – Missing time variable
  - Variable dimensions – Time, altitude, etc.

File "discoveraq-HSRL2_B200_20140812_R0_F1_.h5"

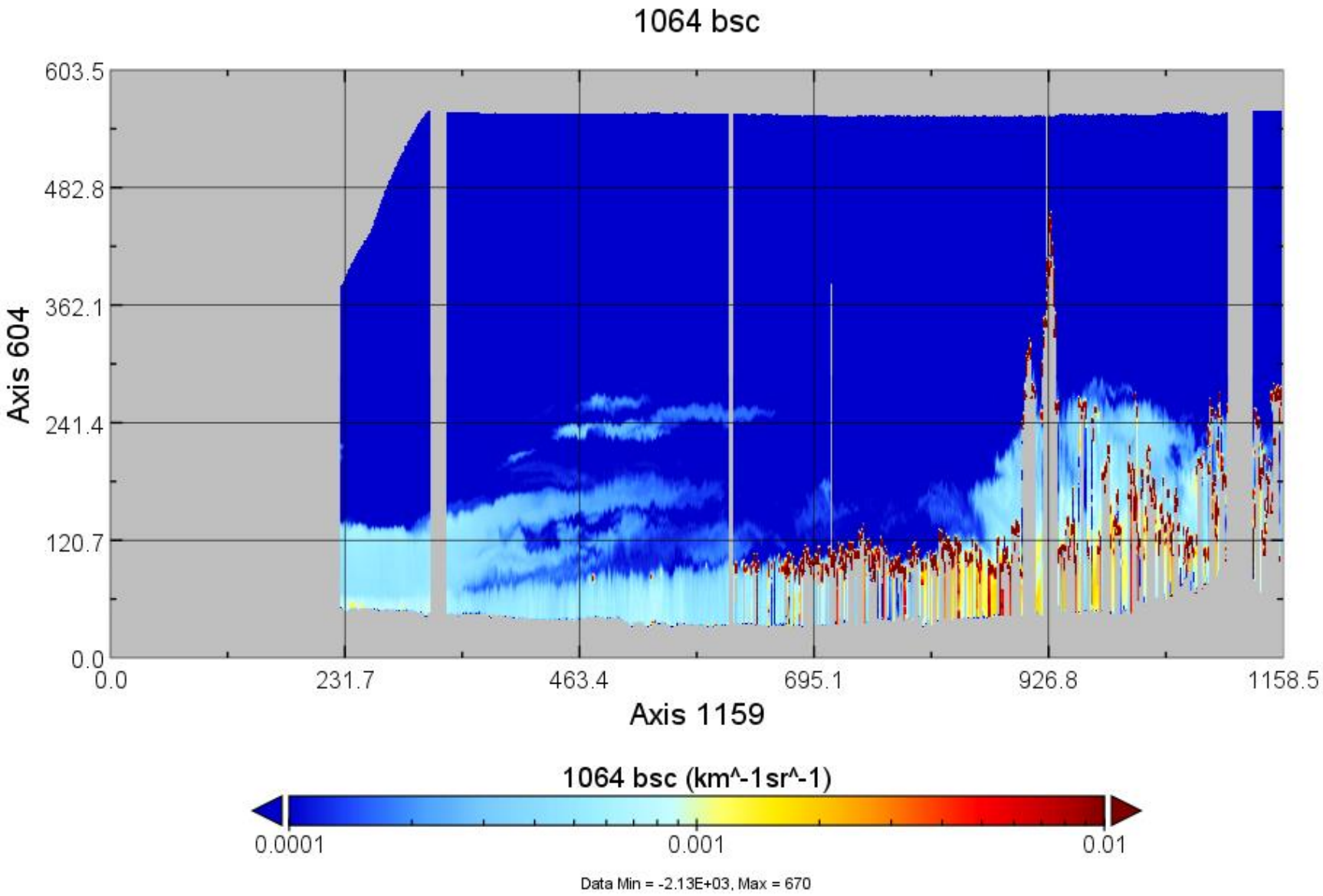
File type: Hierarchical Data Format, version 5

```
variables:
  char 000_Readme(18, 335);

  String Process_Log(158);
    :_ChunkSizes = 1U; // uint

  double 1064_bsc(1159, 604);
    :Description = "1064 nm aerosol backscatter coefficient";
    :Units = "km^-1sr^-1";
    :units = "km^-1sr^-1";

// global attributes:
:data_collection_start_date = "20140812";
:file_creation_date = "02-Feb-2015 16:56:01";
```


1064 bsc
603.5
482.8
362.1
241.4
120.7
0.0
Axis 604
0.0
231.7
463.4
695.1
926.8
1158.5
Axis 1159
1064 bsc (km^-1sr^-1)
0.0001
0.001
0.01
Data Min = -2.13E+03, Max = 670

Proposed Guidelines

- Standardize file structure based upon CF conventions
- Use global attributes to host required general metadata information
- Require use of CF global and variable attributes for domain relevant metadata
- Indicated linked variables using CF attributes
- Recommend using standardized variables Ensure data variables are properly attached to dimensions
- Allow for guidelines to be expanded upon to cater to specific types of data

File "traceraq-HSRL2_GV_20210901_R1.h5"

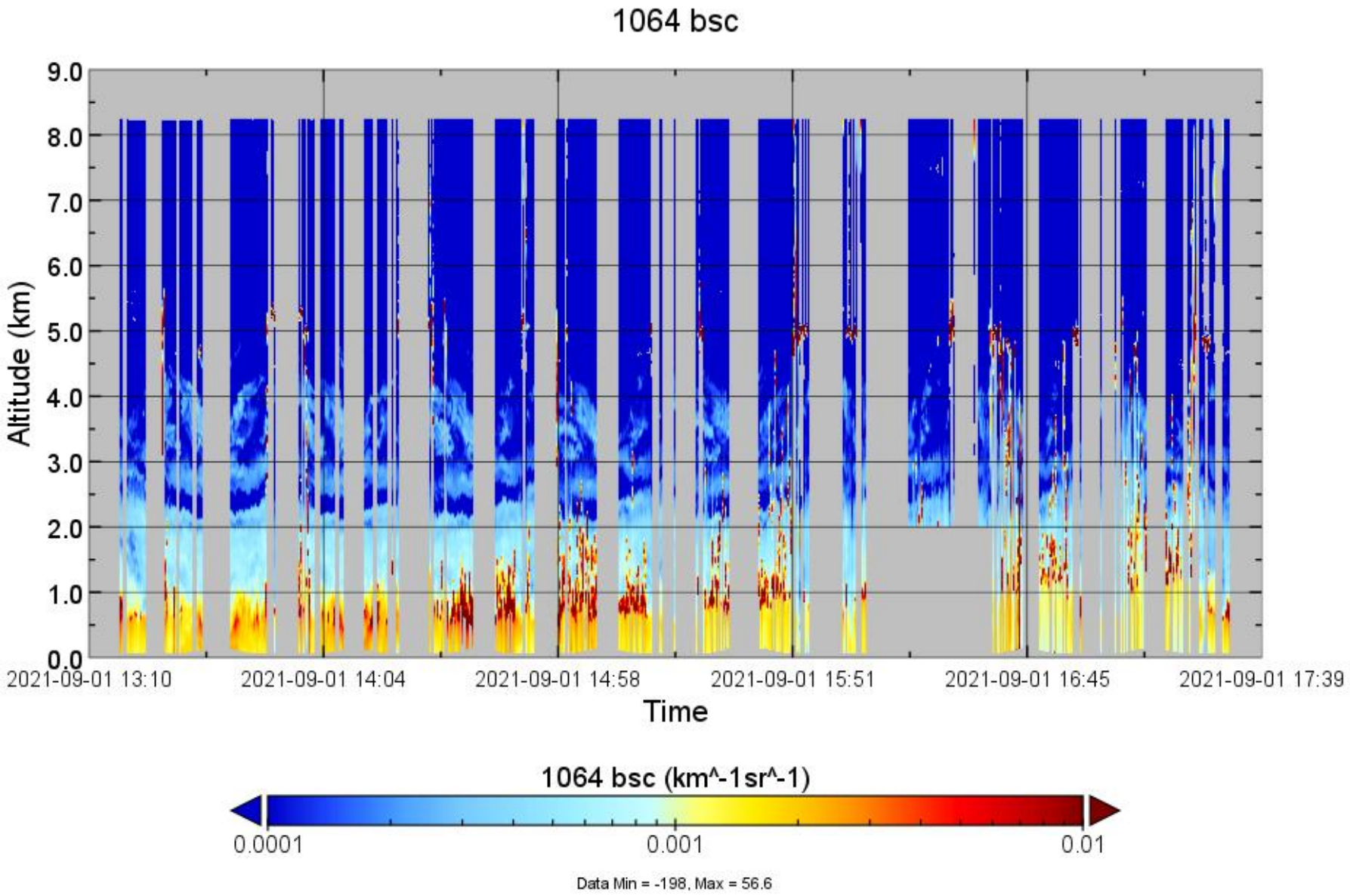
File type: Hierarchical Data Format, version 5

```
double 1064_bsc(time=1603, z=643);
  :Description = "1064 nm aerosol backscatter coefficient";
  :Description2 = "None";
  :Horizontal_resolution_(seconds) = 10.0; // double
  :Vertical_resolution_(meters) = 15.0; // double
  :chi = 1.0; // double
  :molecular_depolarization = 0.0037; // double
  :units = "km^-1sr^-1";
  :code_date = "24-Feb-2023 12:46:53";
  :run_date = "21-Mar-2023 09:35:05";
  :StandardName = "AerOpt_BackScattering_Profile_IR_RHa_Bulk_AMB";
```

File "traceraq-GCAS-NO2_GV_20210830_R2.nc"

File type: Hierarchical Data Format, version 5

```
// global attributes:
:DATA_INFO = "see traceraq-GCAS-NO2_GV_readme_R2.txt";
:DataProvider = "Scott Janz";
:DataProviderAffiliation = "NASA GSFC";
:DataProviderContact = "scott.j.janz@nasa.gov";
:DataUseGuideline = "For responsible scientific use of the data sets
:Description = "GCAS NO2 Columns and ancillary information from the
:FileOriginator = "Laura Judd";
:FileOriginatorContact = "laura.m.judd@nasa.gov";
:FileType = "PI";
:Instrument = "GCAS";
:InstrumentDescription = "Ultraviolet-visible-near infrared airborne
:LastModifiedDate = "2023 Nov 21 ";
:MACIEStandardNameURL = "https://www-air.larc.nasa.gov/missions/etc/
:MeasurementPlatform = "NASA JSC G-V";
```


1064 bsc
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Altitude (km)
2021-09-01 13:10
2021-09-01 14:04
2021-09-01 14:58
2021-09-01 15:51
2021-09-01 16:45
2021-09-01 17:39
Time
1064 bsc (km^-1sr^-1)
0.0001
0.001
0.01
Data Min = -198, Max = 56.6

Summary

- Metadata for airborne HDF and NetCDF files can vary greatly between campaigns or PIs
- Additional guidelines for the metadata structure will make data more interoperable and reusable
  - Can aid data users in their use of measurements for research
  - Allows data to be more easily ingested for archival
- Updated standards may need to be catered to individual types of measurements
- Proposed guidelines have been tested for several instruments
  - High Spectral Resolution Lidar (HSRL)
  - GEOstationary Coastal and Air Pollution Events (GEO-CAPE) Airborne Simulator (GCAS)
  - High Altitude Laser Observatory (HALO)
  - Research Scanning Polarimeter (RSP)