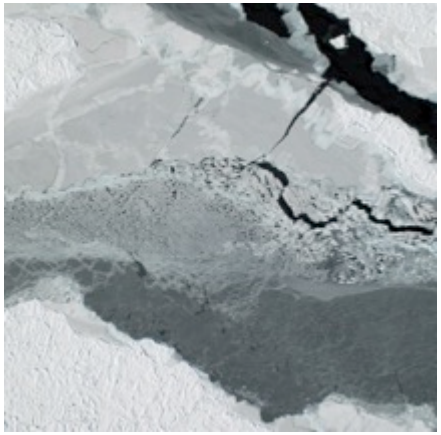
Data File Requirements & Recommendations

1. **Data formats**
2. **Filenames**
3. **File structure and content**

   Variables → What?

   Time → When?

   Geolocation → Where?
4. **General ASCII/CSV structure**



Required vs. Recommended

While there are few items noted as "required", all recommendations are key to enabling long-term usability of data products

- System requirement – Required by NASA data system
- User requirement – Essential to data usage

NSIDC DAAC
Distributed Active Archive Center

1. File Formats

- Avoid proprietary formats
  - Challenging for users now
  - May not be readable in the future
  - NASA standards allow proprietary for raw/Level-0 data
- NetCDF4 & HDF5
  - Good for multidimensional data
  - Capable of holding rich metadata
  - Interoperable with variety of computational platforms and protocols (e.g., OPeNDAP)
- GeoTIFF
  - User friendly format; most requested format by NSIDC users
  - Widely interoperable with GIS, image processing, and map server applications
- Shapefile
  - Good for feature data (e.g., points, lines, polygons)

NSIDC DAAC
Distributed Active Archive Center

NASA Earth Science Format Standards

ASCII
https://earthdata.nasa.gov/standards/ascii-file-format-guidelines-for-earth-science-data

HDF5
https://earthdata.nasa.gov/standards/hdf5

HDF-EOS5
https://earthdata.nasa.gov/standards/hdf-eos5

NetCDF-4/HDF5
https://earthdata.nasa.gov/standards/netcdf-4hdf5-file-format

NSIDC DAAC
Distributed Active Archive Center

1. File Formats

Type of data	Recommended formats
Tabular or site-based data	Delimited ASCII/CSV
	HDF5
	NetCDF4
Raster	GeoTIFF
	HDF5
	NetCDF4
	Delimited ASCII/CSV
Vector	LAS 1.2 (LAZ)
	Shapefile
	Delimited ASCII/CSV
Photos/Movies	JPEG/MPEG

2. Filenames

- Must begin with data product identifier (NSIDC provides) in order to capture metrics*
- Filenames must be unique*
- No spaces; ASCII characters only*
- File extension indicates data format*
- Include descriptive information such as:
  - Campaign
  - Instrument
  - Measurement
  - Spatial (site, resolution)
  - Temporal (date, time)
  - Processing version

<ProductID>_<ProductSpecificDescriptiveInfo>.<extension>

Icebridge example:
ILATM2_20090422_131620_smooth_nadir3seg_50pt.csv

NSIDC DAAC
Distributed Active Archive Center

* System requirement

3. File structure and content: Variables

Standard variable names & units

ISO 8601 — Representation of dates and times

International System of Units

UDUNITS — Unit database and conversion between units

CF Standard Name — Climate Forecast (CF) standards promote sharing

NSIDC DAAC — Distributed Active Archive Center

- Define unique, interpretable names for each variable
  - Full interpretation can come from a mapping of short variable names to description
- Define units for each variable
- Define a value for missing data and use consistently
  - E.g., -9999

3. File structure and content: Time*

- Define time standard and time zone used
  - Recommend using UTC
  - Timestamps may be reported as UTC decimal seconds from the time at which measurements began (commonly as seconds past midnight)
- Use standard date/time formats
  - Recommend using yyyymmdd, hhmmss, or yyyymmddTHHMMSS.SSSZ (ISO 8601 standard)

NSIDC DAAC
Distributed Active Archive Center

* User requirement

3. File structure and content: Geolocation*

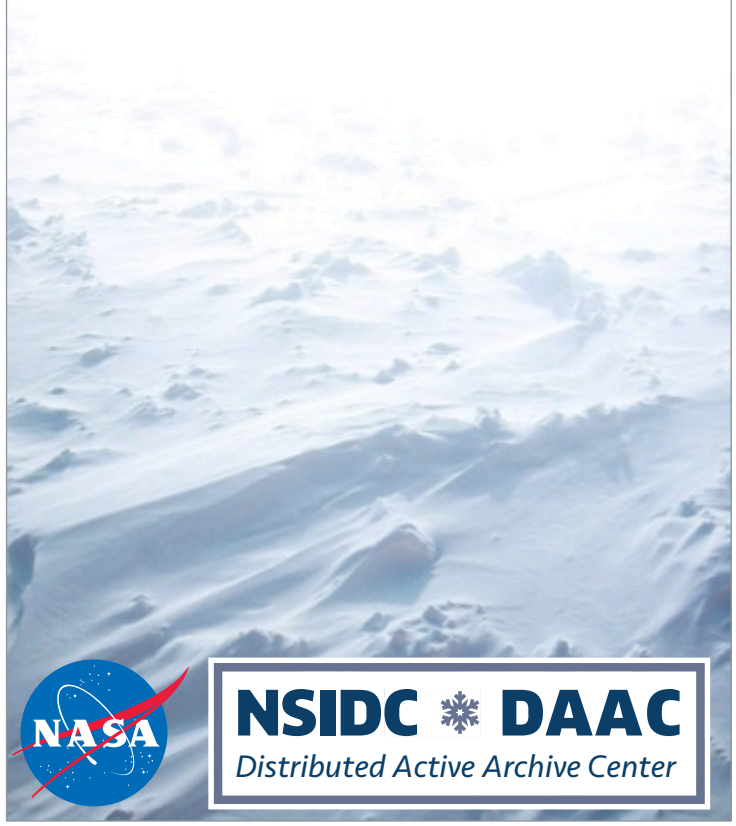
- Include geographic coordinates for measurements in the data file
  - Latitude, longitude, and altitude (where applicable)
- Use a consistent coordinate format
  - E.g., lat/lon decimal degrees, UTME/UTMN
  - Recommend SnowEx select standard format(s) for delivered data products
- Provide coordinate reference system and horizontal and vertical datum

NSIDC DAAC
Distributed Active Archive Center

* User requirement

Full ASCII Earth Science Recommendations

https://earthdata.nasa.gov/standards/ascii-file-format-guidelines-for-earth-science-data



NSIDC DAAC
Distributed Active Archive Center

4. General ASCII/ CSV Structure

- Include separate header and data sections within file
  - Header needs to be clearly delineated from data rows (e.g., begin with #)
- Use consistent delimiter between data values
  - Visible characters are preferred (e.g., comma, semi-colon, colon, |)
- Separate rows with end-of-line character
  - Mac: CR
  - Unix: LF
  - Windows: CR/LF
- Do not use empty lines or rows

Data File Checklist

- Data formats
  - Use nonproprietary formats
  - Review the NASA Earth Science format recommendations
- Filenames
  - Use descriptive, unique names
  - Product identifier begins each filename*
  - No spaces; ASCII characters only*
  - File extension indicates data format*
- File contents/structure: Variables
  - Unique, interpretable variable names
  - Define units for each variable
  - Consistently use a missing data value
- File contents/structure: Time*
  - Define time standard and time zone used: recommend UTC
  - Use standard date/time formats
- File structure/content: Geolocation*
  - Include geographic coordinates
  - Use a consistent coordinate format
  - Provide coordinate reference system and horizontal and vertical datum
- General ASCII/CSV Structure
  - Include and delineate header from data section
  - Use consistent delimiter between data values: visible characters preferred
  - Separate rows with EOL
  - Do not use empty lines or rows

NSIDC DAAC
Distributed Active Archive Center

*System requirement
*User requirement