

NGEE Arctic Data Management Guidance

Guidelines for collecting, reporting, processing, and archiving data for both project and public sharing.

For more information.

Please contact either Terri Killeffer (killefferts@ornl.gov) or Les Hook (hookla@ornl.gov) if you have any questions, comments, or need assistance.

DATA MANAGEMENT GUIDES

CONTENTS

Preparing Data for Project Data Sharing (UPDATED: November 5, 2013)	2
Data File Naming.....	4
Reporting Sampling and Measurement Dates and Times	7
Reporting Missing Data.....	8
Identifying Measurement Variables	9
Reporting and Flagging Measurement Values below Detection Limits (DRAFT: November 1, 2013)	11
Data Quality Flags for Qualifying Measurement Values (DRAFT: November 1, 2013)	12
Defining Locations of Sampling and Measurement Events (DRAFT January 08, 2015)	13
More Guides as Needed	19

What do we expect you to do for Project Data Sharing?

- Using the Metadata Entry and Data Upload Tool, enter the contact and descriptive metadata for the data files that you upload.

Project Data Sharing FAQs:

How do I organize my data files for sharing? What is a data set?

- Think of a data set as a collection of related data files – one-to-many data files. Seldom does a data set have only one data file. For example:
 - A data set can be a time series of measurements, even if the series is periodic rather than continuous. Soil moisture measurements made during several summers in Alaska would be considered one data set with a new data file added each year.
- A data set could be the result of a literature compilation task -- data values and locations compiled in a file with a citation code linking to a separate file with the full literature citations.
- A model data product would include the model code, installation/running instructions, model input/driver data files, and example output data files. If the products are directly related to a publication then the specific code, inputs, and outputs are needed.
- Data products that are NGEA Arctic subsets of existing larger regional or global data sets should also be archived and shared. These value-added products must include documentation of their source(s) and processing.
- GIS or remote sensing products could similarly represent a time series and be considered one data set.

How do I organize my data into files?

In general, keep similar information together:

- Keep a set of similar measurements together (e.g., same investigator, methods, time basis, and instruments) in one data file.
- Please do not break up your data into many small files, e.g., by month or by site if you are working with several months, years, or sites.
- Instead, make month or site a parameter and have all the data in one large file. Researchers who later use your relatively large data file won't have to process many small files individually.

Organize by data type:

- If you are collecting many observations of several different types of measurements at a site (e.g., leaf area index and above- and belowground biomass), place each type of measurement in a separate data file.
- For each data file, use similar data organization, parameter formats, and common site names, so that users understand the interrelationships between data files.
- Data types collected on different time bases (e.g., per hour, per day, per year) might be handled more efficiently in separate files.

Non-proprietary formats [added November 5, 2013]:

- If your collection operation has used proprietary file formats, create an export in a stable, well-documented, and non-proprietary format. It is important for maximizing others' abilities to use and build upon your data format is important for maximizing others' abilities to use and build upon your data.

Just the data please [added November 5, 2013]:

- Figures and analyses should be reported in companion documents - don't place figures or summary statistics in the data file.

DATA FILE NAMING

Provides guidance for naming tabular data files, image files, model products, and companion documentation files.

Following this syntax guidance will result in file names that will sort by Science Team and contents and will contain enough specific information to enable data providers to keep files distinct and to enable data users to find files of interest.

Data File Name Syntax:

Tabular Data Files: [task]_[file contents description]_[unique data file descriptor]_v1.extension

Where:

Task is the Science Team identifier, with or w/o the task_id number.

Science Team identifier:

Biogeochemistry (Task_IDs BGC*):

Hydrology and Geomorphology (Task_IDs HG*):

Integrated Model-Data Evaluation (Task_IDs I*):

Landscape Characterization (Task_IDs LC*):

Site Characterization and Design (Task_IDs S*):

Vegetation Dynamics (Task_IDs V*):

File contents description:

Could be same as 'Data Set Title' or 'Data Type' entered in Inventory. 20-30 characters. This general description should be repeated for files in a series or collection.

Unique data file descriptor:

Unique event, place, averaging time, gridded spatial resolution, date range, start date, start date and days of data in file (e.g., yyyyymmdd_nnn), quarter reported (e.g., Q1, Q1-Q2), data processing level, etc. 20-30 characters.

Version control:

Another underscore, a "v", and followed by a version number-- "_v1" (initially "one"). Data provider is responsible for incrementing the version number when an updated file is submitted. Documenting the changes is always a good idea.

File extensions:

The file extensions will follow standard conventions to identify the format of the file.

Use the file extension that best indicates the type of file and field delimiters, *.csv for comma-delimited text file and *.txt for a tabs or semicolons delimited text file. Do not use *.dat

Delimited text file formats ensure data are readable in the future. Use a consistent structure throughout the data set. Report figures and analyses in companion documents, not the data file. Use ASCII text encoding if possible, with UTF-8 or UTF-16 as secondary options.

Character set:

Filenames must use only alphanumeric characters, hyphen-minus, plus, or underscore characters. Use underscores rather than spaces.

Lower case please, unless uppercase uniquely defines a location or event (e.g., plot A4).

Length limit: There is no technical length limit. However, there is a practical limit of 70-80 characters.

Image Data Files:**Raster Image Data Files**

Although a variety of file formats are used for image data, descriptive file names as described above, should be used. Good raster file formats are open, non-proprietary, simple and commonly used. More importantly, they are self-descriptive, in other words, metadata are included inside the file. Recommended: GeoTIFF, NetCDF v3/v4, HDF-EOS.

Vector Data

Although a variety of file formats are used for image data, descriptive file names as described above, should be used. Good vector file formats are open, non-proprietary, simple and commonly used. More importantly, they are self-descriptive, in other words, metadata are included inside the file. Recommended: ESRI Shapefiles, KML.

Model Data Products:

Model input and output data files, when in common tabular or image formats, should be named as described above. Codes and outputs, specific to a particular model, should be named consistently with the naming convention and version controls of the modeling community and knowledgeable users.

Companion Documentation Files:

Documentation and supplemental information files (e.g., readme files) should be named as described above, with the same task “[task]_[file contents description]_” as the data product so the file(s) will sort with the data files. The “[unique data file descriptor]_” should be similar to the data file names and include “documentation” or “readme” as appropriate.

REPORTING SAMPLING AND MEASUREMENT DATES AND TIMES

Provides guidance for reporting sampling and measurement dates and times.

Because reporting dates and times is so important to the success of a project, we provide this guidance to prevent many of the reporting problems encountered by similar intensive monitoring projects.

Time Basis:

Investigators will report data on **AKST – Alaska Standard Time** basis.

AKST time zone offset: UTC - 9 hours. AKST is 9 hours behind Coordinated Universal Time (UTC)

Dates and Times to Report:

- Start date and time must be reported as time at the beginning of the sampling/measurement/averaging period.
- End date and time must be reported as the time at the end of the sampling/measurement/averaging period.
- For continuous processes, the end date and time of the preceding period may be the start date and time of the next period.
- There is no 24:00 time. 23:59, then 00:00 the next day.

Reporting Dates and Times:

- Local Time Zone is specified on every data record. Specify **AKST**.
- Sample dates and times must be reported in **AKST**.

AKST. Formats: 2003-02-28 (or 20030228) and 07:00. (Note leading zero.) Use 07:00:00, if seconds are important.

Note: AKST time zone offset: UTC - 9 hours. AKST is 9 hours behind Coordinated Universal Time (UTC)

Missing Date and Time fields:

- **There may not be blank date or time fields.**
- If a value for the time field is not reported, use 12:00 AKST as the time value.

REPORTING MISSING DATA

Guidance for reporting missing data.

- All data fields must have a value present, either the measured value or a missing value representation.
- There may not be blank data fields.

Numeric fields:

- Use '-9999' for missing Numeric fields.
- Be sure that the missing value is negative and large enough to be impossible as an actual data.

Character fields:

- Use 'None' as the missing code for Character fields.

Date and Time fields:

- If a value for the time field is not reported, use 12:00 AKST as the time value.

IDENTIFYING MEASUREMENT VARIABLES

These variable names should represent specifically what was measured, observed, and modeled.

It is a goal of the NGEE Arctic Data Archive to use variable names from a common vocabulary. This is something we will work towards. As you may already realize, a comprehensive list of variable names for the whole set of NGEE Arctic measurements and modeled parameters does not currently exist. We propose to start with these three generally accepted vocabularies, recognizing their limitations, and supplement as needed.

CF Standard Names:

NGEE Arctic would like data providers to use the CF Standard Names to identify parameters.

The CF Standard Names are part of the NetCDF Climate and Forecast (CF) Metadata Conventions.

The conventions for climate and forecast (CF) metadata are designed to promote the processing and sharing of files created with the NetCDF API. The CF conventions are increasingly gaining acceptance and have been adopted by a number of projects and groups as a primary standard.

The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities.

Access the CF Standard Names:

<http://cf-pcmdi.llnl.gov/documents/cf-standard-names>

<http://cf-pcmdi.llnl.gov/documents/cf-standard-names/standard-name-table/25/cf-standard-nametable.html>

Units are especially important to the “cf-standard-name” and data sharing. Note:

- Content - a quantity per unit area. The units will be “kg of measurand m⁻²”.
- Layer - any layer with upper and lower boundaries that have constant values in some vertical coordinate. There must be a vertical coordinate variable(s) indicating the extent of the layer(s).
- Amount - mass per unit area.

What if there is not a “cf-standard-name” that describes your variable?

Refer to the following guidelines and construct your “NGEE-cf-name”. For example, “leaf_carbon_content” is a “cf-standard-name” but there is no parallel name for leaf nitrogen. Use “leaf_nitrogen_content”.

Refer to the Guidelines for Construction of CF Standard Names (<http://cf-pcmdi.llnl.gov/documents/cf-standard-names/guidelines>) for information on how the names are constructed and interpreted and how new names could be derived.

Global Change Master Directory (GCMD) “VARIABLE” List:

Alternatively, GCMD variables may be selected. The GCMD variables are not as rigorously defined nor tightly coupled to units as the “cf-standard-name”. The GCMD variable list is also not complete. For example, you will not find LEAF AREA INDEX. It should be added. Also note that the sample matrix, e.g., VEGETATION, is separate from the measured CARBON, and will need to be provided as a separate descriptive data column. Units may be more flexibly applied.

Access the Global Change Master Directory (GCMD) “VARIABLE” List:

<http://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/>

<http://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/sciencekeywords.csv>

AmeriFlux Network Data Formats:

Investigators who are collecting data at flux towers that are part of the AmeriFlux Network should submit data to NGEE Arctic using the preferred AmeriFlux variable names and formats.

AmeriFlux website -- Submit/Upload data: <http://ameriflux.lbl.gov/HowTo/Data/SitePages/Home.aspx>

Guidance for reporting the measurement detection limits and data below the limit of detection, and how to flag these values with Data Quality Flags.

General guidance:

- Report detection limits in the data file and include in documentation. Clearly describe in the documentation the methodological determination of the detection limits.

(Preferred)

- Report the **actual measured value** even if the value is below the detection limit (including zero and negative values). Flag the value with the Data Quality Flag "**V1**" (Valid value but comprised wholly or partially of below detection limit data).

(Alternative)

- Substitute the detection limit (or other value) for the measured value. Flag the value with the Data Quality Flag "V7" (Valid value but set equal to the detection limit (DL) because the measured value was below the DL).
 - **We do not recommend substituting the detection limit, zero, ½ the detection limit, or any other value, for below detection limit data.**
- Estimated values below the detection limit
 - If a measured value is below what is considered to be the method detection limit but is nonetheless considered meaningful, suggesting where between zero and the DL the value lies, the measured value should be provided, and the Data Quality Flag "V2" ("Valid estimated value"), should be applied.
- Like noting in the documentation that there are no missing values in a data file, it is informative to add that "No below-detection-limit values are reported in this data file."

How do I name my "parameter Detection Limit"?

- Add the suffix "_dl" to the column name.

DATA QUALITY FLAGS FOR QUALIFYING MEASUREMENT VALUES (DRAFT: NOVEMBER 1, 2013)

A separate field with specified values may be used to provide additional information about the measured data value including, for example, quality considerations, reasons for missing values, or indicating below detection limit data.

Codes should not be parameter specific but should be consistent across parameters and data files. Definitions of flag codes should be included in the accompanying data set documentation.

DRAFT NGE- Arctic standard Data Quality Flag values and definitions: Comments?

Flag Value	Description
V0	Valid value
V1	Valid value but comprised wholly or partially of below detection limit data
V2	Valid estimated value
V3	Valid interpolated value
V4	Valid value despite failing to meet some QC or statistical criteria
V5	Valid value but qualified because of possible contamination (e.g., pollution source, laboratory contamination source)
V6	Valid value but qualified due to non-standard sampling conditions (e.g., instrument malfunction, sample handling)
V7	Valid value but set equal to the detection limit (DL) because the measured value was below the DL
M1	Missing value because no value is available
M2	Missing value because invalidated by data originator
H1	Historical data that have not been assessed or validated

How do I name my "parameter Data Quality Flag"?

- Add the suffix "_fl" to the column name.
- Add the suffix "_dl" to the column name.

The location types that characterize your sampling and measurement locations should be included as individual columns in your tabular data files. Use the location types as needed to describe GIS features, and model input and output data.

Defining Locations of Sampling and Measurement Events

The location types that characterize your sampling and measurement locations should be included as individual columns in your tabular data files.

Use the location “_types” as needed to describe transects, plots, and features. **“feature” is a new field.**

Seward Peninsula location questions remain.

North Slope updated January 8, 2015

Location Types (separate columns)	Required	Possible values	Possible values	Comments and ?	Defined Coordinates for Location*	Assigned by
region	yes	North Slope	Seward Peninsula	(SP) ?	Not applicable	NGEE Team
locale	yes	Barrow	Council, Kougarok, Teller	Nome? More?	Not applicable	NGEE Team
administrative_area	If applicable	BEO, NSB Conservation Area, International Biological Program (IBP)	???		Yes (bounding box) (optional)	NGEE Team
site	If applicable for Barrow.	Site 0 (Intensive Site 0) Site 1 (Intensive Site1)	Seward Peninsula highway designation	Other? Site must spatially	Yes (bounding box or point as applicable)	NGEE Team

	Required for SP.		See Business Rules	include all areas and transects. ???		
area	If applicable for Barrow. Area must be within a Site. SP ?	A,B,C,D,	???	watersheds? polygons?	Yes (bounding box)	NGEE Team
transect	If applicable for Barrow. May cross multiple sites and areas or be entirely outside of site. SP ?	Example?? DTLB_1 (transect across Intensive Site 1 & 0), DTLB_2 (transect and core locations south of road)???	???	Needs discussion	Yes (bounding box or two ends)	NGEE Team
transect_type	With transect	Geophysical, Thermal, Vegetation, Hydrology		More?	Not applicable	NGEE Team
plot_type	If applicable with area or transect.	Vegetation, BGC, more?		More? Do plots have to be associated with either an Area or a Transect?	Not applicable	NGEE Team

plot_ID	With plot_type	A1, etc	More? Defined surface area?	Yes (bounding box)	Task
feature_type	If applicable with area or transect.	Well, soil pit, tower, core, sampling site, more?	Single point. May or may not be in a plot. ? Do features have to be associated with either an Area or a Transect?	Not applicable	NGEE Team
feature_ID	With feature_type	??? See label examples.		Yes (point)	Task

*** Universal Transverse Mercator (UTM) Coordinate System

NGEE Arctic is working in the Universal Transverse Mercator (UTM) coordinate system. All coordinates are in northing and easting meters.

North Slope: We are using NAD 83 datum and UTM Zone 4North.

Seward Peninsula: We are using NAD 83 datum and UTM Zone 3North. ???

*It seems that we need both UTM and Lat/Long (both appropriately transformed) for every location, plot, and feature to support product use and also data search and visualization.

Landscape Characterization Task -- remote sensing preference (Alaska Albers Equal Area Conic projection) ???

Seward Peninsula Business Rules: (proposed)

1) TBD: Seward Peninsula Highway/Road and Site Designations

Accepted Site Designation:

Seward Peninsula Highway/Road and Site Designations

Kougarok Road is...

Kougarok_RD_MMxx or xx_x (e.g., 4.5 is 04_5)

Nome-Teller Highway is...

Teller_RD_MMxx or xx_x

Nome-Council Road is...

Council_RD_MMxx or xx_x

Site = Mileage delineated on road (Assigned by AK DOT). May be a trailhead or starting point for a longer transect.

PRESIOUS VERSION NOW SUPERCEDED.

Location Types (separate columns)	Required	Possible values	Comments?	Defined Coords - UTM? ***
region	yes	North Slope, Seward Peninsula		Not applicable
locale	yes	Barrow, Council	More?	Not applicable
administrative_area	If applicable	BEO		Yes
		NSB Conservation Area		???
site	If applicable	Intensive Site 0, Intensive Site 1	More?	Yes
area	With site	A,B,C,D		Yes
		Dry Lake Bed ?	More?	
			More?	
watershed ?				polygons?
plot_type	If applicable	Vegetation, BGC	More?	Not applicable
plot_ID	With plot_type	A1, etc	More?	Yes

transect_type	If applicable	Geophysical, Thermal	More?	Not applicable
transect_ID	With transect_type	Example: DTLB_1 (transect across Intensive Site 1 & 0), DTLB_2 (transect and core locations south of road)	More?	Yes

*** Universal Transverse Mercator (UTM) Coordinate System

NGEE Arctic is working in the Universal Transverse Mercator (UTM) coordinate system. All coordinates are in northing and easting meters.

We are using **NAD 83 datum** and **UTM Zone 4North**.

This link provides a conversion between geographic coordinates and Universal Transverse Mercator (UTM) coordinates: <http://home.hiwaay.net/~taylorc/toolbox/geography/geoutm.html>

MORE GUIDES AS NEEDED