DATA MANAGEMENT BEST PRACTICES WEBINAR

Data management practices to improve usability of data sets – now and forever
Workshop Information

Part –1: Planning Data collection (Today’s Webinar)

• Data Management Overview

• Best Practices for data collection
  – Site data
  – Field Data Collection
  – Model data

Part –2: Data Tools for Exploratory Data Analysis (EDA) (Future)

• Best Practices for getting most out of the data

• Tools for
  – Data Search
  – Visualization
  – Data Analysis
Speakers

**Suresh Vannan** is the manager of the Distributed Active Archive Center for Biogeochemistry at ORNL. Suresh has extensive experience with building data management plans and in handling data workflows from collection to archival.

**Giri Prakash** is the Architecture and Services Strategy Team Manager for the Atmospheric Radiation Measurement (ARM) Archive. Giri’s expertise include building end-to-end data management capabilities for climate and biodiversity projects.

**Terri S. Killeffer** is the Scientific Data Curator on the Data Team for the Next-Generation Ecosystem Experiments - Arctic (NGEE Arctic) project interacting with investigators to obtain data, assign metadata, and develop documentation.

**Yaxing Wei** is a scientist at the ORNL Distributed Active Archive Center. Wei has been working on several research projects to provide geospatial information management, analysis, visualization, and sharing.
About the Climate Change Science Institute

Advancing the Knowledge of Climate Change and Understanding its Consequences

- Formed in 2009 to integrate ORNL’s climate research programs
- 130 collocated scientists

- [https://ccsi.ornl.gov](https://ccsi.ornl.gov)
- Facebook, Twitter, YouTube
Webinar Notes

• Please fill out the survey at the beginning of this webinar
• Towards the end, there will be another quick survey
• Please mute your microphones
• Please send your questions via the Q&A box on your screen
  – Questions will be answered at the end of the webinar
• Slides will be made available on the CCSI website; link will be forwarded to your email address
Data Management Best Practices - Part 1

Workshop Goals and Data Management Overview

Suresh Vannan
Data Theme Lead, CCSI
santhanavans@ornl.gov
Provide data management practices that investigators should perform during the course of data collection to improve the usability of their data sets
Data Management – What?

20-year rule

The data set and accompanying documentation should be prepared for a user 20 years into the future

Prepare the data and documentation for a user who is unfamiliar with your project, methods, and observations
Data Management – Why?

- **Protection:** About 32 percent of computer users experience data loss each year.

- **Publication:** Agencies are now required to manage the digital data resulting from federally funded scientific research.

- **Provenance:** To support research findings we need to preserve data and its provenance to be able to trace and record the origins of the data.

- **Author Credit:** Recognize data creators for the value of their data.
Fundamentals of Data Management – Project Level

1. Define data workflow – Collection to Archive
2. Train staff for data/metadata handling
3. Communicate standards and metadata needs
4. Define Data Quality criteria
5. Establish a long-term data steward
6. Establish and communicate a Data use/Submission policy
Fundamentals of Data Management –
Data Set Level

1. Define the contents of your data files
2. Define the variables
3. Use consistent data organization
4. Use stable file formats
5. Assign descriptive file names
6. Preserve processing information
7. Perform basic quality assurance
8. Provide documentation
9. Protect your data
10. Preserve your data
Data Management Plan

• Information about the data
  – Description of data to be produced
  – How will it be managed in short-term?

• Description of Data
  – Format, number of files, approx. volume
  – Processing and quality

• Metadata Content & Format
  – Documentation about the data

• Policies for Access, Sharing, & Reuse

• Long-term Storage & Data Management
  – Where will data be archived?

Remember to include data management costs in Proposal Budget
Costs

At least **10% of total funding** is suggested to be devoted towards managing the data used and produced by a project. Cost will be relative to the size, complexity, length, and access needs for a project.

Another way to interpret this is that at least **10% of investigators time** should be spent on actively managing their data.

Data collection vs Data Management

The 1990 census:

- $2.6 billion for data Collection
- $433 million for data Processing
- $114 million for data Dissemination

Source: [http://www.nap.edu/read/4805/chapter/5](http://www.nap.edu/read/4805/chapter/5)
Staffing and Training

- Establish a Data Management plan
- Communicate plan
- Establish standards and conventions
- Embedded data management in the research costs
- Cultivate culture of data management to be done alongside research
- For large projects (multi-organization, multi-year, multi-levels of data) recruit a data officer
- Specialized technicians are needed for advanced tiers of services.
- Regularly communicate about data and metadata needs within the project

“A goal without a plan is just a wish.”
Standards - Metadata

<table>
<thead>
<tr>
<th>Type</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIS data? Raster/vector or point data</td>
<td>FGDC Content Standard</td>
</tr>
<tr>
<td>Data retrieved from instruments such as monitoring stations or satellites</td>
<td>ISO 19115</td>
</tr>
<tr>
<td>Ecological data</td>
<td>Ecological Markup Language</td>
</tr>
<tr>
<td>Specimen occurrence and observational records</td>
<td>Darwin Core</td>
</tr>
</tbody>
</table>

Most importantly …

- Data characteristics and description
  - When and how frequently the data were collected
  - Where and with what spatial resolution the data were collected
  - The name(s) of the data file(s) in the data set
  - Example data file records for each data type file
  - Special codes used, including those for missing values
  - Date the data set was last modified
  - English language translation of any data values and descriptors in another language
  - How each parameter was measured or produced (methods), units of measure, format, precision and accuracy, and relationship to other data in the data set

- Data acquisition
- Quality assessment
- Supplemental information

http://goo.gl/MJjXmt
Standards – File formats

Raster
- Geotiff
- netCDF
  - with Climate and Forecast convention preferred

Vector
- Shapefile
- KML

Tabular
- CSV
Resources


OSU Libraries Research Data Services: [http://guides.library.oregonstate.edu/managing-data](http://guides.library.oregonstate.edu/managing-data)


Agency Specific Data Management


Open access plans for all agencies: [http://guides.library.oregonstate.edu/federaloa](http://guides.library.oregonstate.edu/federaloa)

Data Management
Best Practices - Part 1

Site Data Management

Giri Prakash
Architecture and Services
Strategy Team Manager for
the Atmospheric Radiation
Measurement (ARM) Archive.

Palanisamyg@ornl.gov
Topics

• Instrument deployment
• Site data collection
• Data retention & transfer
• Data ingest & data flow monitoring
• Preparing and communicating data quality
• Data security and archival
• Metadata management
Site Data
Deployment of Site Instruments/Sensors

• Site visit & logistics
• Plan for a pre-operational period
  – Deployment and initial calibration
• Assign instrument experts
• Assess quality of pre-operational data and calibrate as necessary.
Data Collection

- Set up a site data system
- Establish network from instrument – data system – data center
- Train and mentor the site technician to manage the data system
- Establish instrument monitoring
Data Flow and Tracking

- Track the data through various phases – online, processed, stored and available
- Each phase has various checks to enable operations teams to monitor any data issues
Data Transfer

• Secured network-based access to the site data systems

• Data transfer from site to data center
  – Network based transfer
  – Disk shipments

• Confirmation of successful data transformation using checksums

• Setting up enough disk space until the data transfer is fully complete and verified
Data Processing

- **Unmodified from instrument**
  - **Ingests**
    - Engineering units
    - Quality flags
    - Updated hourly
  - **Value-added Products (VAP)**
    - Advanced algorithms
    - Multiple inputs
      - Principal investigators/Public access

“Raw”

1. **Standard Format**
2. **Value Added**
3. **Dissemination**

- **Fills some of the unmet measurement needs**
- **Improves the quality of existing measurements**
Data Quality

- Automation
  - Quality Control Processing
  - Data Assessment

- Human Input
  - Data Visualization
  - Problem Reporting
Data Quality

- Quality Control (QC) Processing
  - Automated Mentor QC (Included in data files)
    - MIN/MAX/DELTA; Many include additional QC tests
    - Library of functions to make these tests easily adaptable from one instrument to the next
  - Summarized into Easy to Read Metrics Tables

```c
float precip(time);
  precip:long_name = "Total precipitation over one minute sampling period";
  precip:units = "mm";
  precip:valid_min = 0.f;
  precip:valid_max = 100.f;
  precip:missing_value = -9999.f;
  precip:comment = "Sum of the 20 3-second observations during the sampling period.";
int qc_precip(time);
  qc_precip:long_name = "Quality check results on field: Total precipitation over one minute sampling period";
  qc_precip:units = "unitless";
  qc_precip:description = "This field contains bit packed values which should be interpreted as listed. No bits set (zero) represents good data.";
  qc_precip:bit_1_description = "Value is equal to missing_value.";
  qc_precip:bit_1_assessment = "Bad";
  qc_precip:bit_2_description = "Value is less than the valid_min.";
  qc_precip:bit_2_assessment = "Bad";
  qc_precip:bit_3_description = "Value is greater than the valid_max.";
  qc_precip:bit_3_assessment = "Bad";
  qc_precip:bit_4_description = "Difference between current and previous values exceeds valid_delta.";
  qc_precip:bit_4_assessment = "Indeterminate";
  qc_precip:bit_5_description = "Scans per min does not equal to 20, data set to -9999.";
  qc_precip:bit_5_assessment = "Bad";
  qc_precip:bit_6_description = "Error latch does not equal to 0, data set to -9999.";
```
Data Quality

• Quality Control Tests/Metrics
  – Some QC Tests:

  • Persistence (Flatlined data)
  • Grubbs Statistical Test for Outliers
  • Instrument Comparison
  • Derived Quantity Comparison
  • Time Drift
  • Power Outage Flagging

  • Probability Density Function
    o Cross Instrument and Cross Site Comparisons
  • Time Varying Limits
  • Specialized QC developed in conjunction with the Instrument Mentors
Data Quality

• Data Assessment
  – Review the data on a daily or weekly basis
  – Report findings to the infrastructure through a Data Quality Assessment Report
  – Make use of undergraduate student analysts
Data Quality

Problem Reporting

Problem Identification
Corrective Action
User Notification

Data Quality Problem Report
Provides a place for internal tracking of a problem to completion between the DQ Office, Site Operations, and Instrument Mentors

Data Quality Report (DQR)
Machine readable report that informs the end user of problems with the data
Displayed at the Archive, sent to end user with data, accessible through a web service
Data Security and Archival

• Verify data integrity during data transfer
• Archive every version of data files that are released from upstream process:
  – Rule of thumb: if you can’t reproduce, then it should be archived
• Archival needs to include backup strategy, both onsite (for immediate data recovery) and offsite (disaster recovery)
Reprocessing and Communications

- Changes to the published datasets
- Communicating users about any change in data quality and resulting new version of data
  - Archiving user order details
Metadata – Captured Throughout the Data Lifecycle

- **Metadata Management**
  - Define term
  - Example: Aerosol Observing System
  - Why, what, who, where, how

- **Processing data to add value**
  - Data products
  - VAP processing
  - Reprocessing
  - Data lifecycle and metadata overlay

Source: Alice Cialella, ARM
Data Management Best Practices - Part 1

Field Data Collection

Terri Killeffer
NGEE-Arctic Data Curator
killefferts@ornl.gov
What is Field Data?

Examples of Field Data Collections

- Organism observation/collection
- Soil samples
- Geophysics
- Sap flow
- Mineralization rates
- Soil incubation studies
- Geochemical and isotopic analyses
- Biodiversity inventory
- Images
Electrical Resistivity Methods Used to Explore Subsurface Properties in Ice-Rich Tundra. Stan Wullschleger. 2012-08-24. CC BY-NC 2.0

Proteus aircraft (426 004 001). Energy.gov. 2012-09-22. CC BY-NC-SA 2.0

USGS Hawaiian Volcano Observatory at Kilauea Volcano. Michael Poland, USGS. 2008-09-03. CC BY-NC 2.0

Modeling climate change. Argonne National Laboratory. 2013-07-25. CC BY-NC-SA 2.0

Ice Core. Kris McCraken. 2013-03-23. CC BY-NC-SA 2.0

Procambarus clarkii. Ryan Hagerty, USFWS. 2016-05-12

37 Data Management Best Practices

Before Collecting Data

While Collecting Data

After Collecting Data
Before Collecting Data

While Collecting Data

After Collecting Data
Don’t Work in Isolation

- Data will be less interoperable
- Potential duplication of effort
- Possibly miss collecting important data
Cooperate, Coordinate, and Collaborate

- Work within and across teams/disciplines
- Work across related projects
- Review and refine the Data Management Plan
Project Goals

Final Products

Data Goals
Items to Address Across the Project

- Naming of sampling locations
- Base maps
- Spatial Reference System and spatial coordinates
- Reporting time and date
- Identify data and sample archives
- Standardize variable names
- Instruments
- Collection techniques
- Sample and subsample labeling and tracking

http://www.colorado.edu/geography/gcraft/notes/mapproj/gif/threepro.gif
Plan to Capture Metadata at Every Stage

**What**
- Preferred Projections: UTM Zone 3N and Alaska Albers

**Why**
- Field Campaign: July 1-8, 2014
- With Jack Smith, Arnold Johnson, and Beth Allen

**Who**
- The vials with water samples from Doe River Mile 16 leaked.
- No samples exist for RM16 in 2010.
Data in Development

• Record the names of people participating in the field collection or laboratory analysis
• Take pictures and videos
• Follow best practices and record methodology
• Record method/instrument for capturing data and geolocation
Data in Development

• Label samples
• Note issues, lost samples/data, defective instruments, reason no data collected, etc.
• Develop well organized data files
• Record all data processing steps
• Backup data
Before Collecting Data

While Collecting Data

After Collecting Data
Metadata Collected

What

Why

Who

When

Where

How

Before Collecting Data
Preferred Projections: UTM Zone 3N and Alaska Albers

While Collecting Data
Field Campaign: July 1-8, 2014 With Jack Smith, Arnold Johnson, and Beth Allen

After Collecting Data
The vials with water samples from Doe River Mile 16 leaked. No samples exist for RM16 in 2010.

The vials with water samples from Doe River Mile 16 leaked. No samples exist for RM16 in 2010.
Content of the Data Files

- Measurement, Sample, Observation, Specimen, Analysis
  - Parameter name
  - Units
  - Formats
  - Dates
  - Time
  - Spatial coordinates
  - Missing values
  - Coded fields
  - Flags

ORNL DAAC, Define the contents of your data files.
## Data Dictionary Example

<table>
<thead>
<tr>
<th>column_name</th>
<th>units/format</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core_ID</td>
<td></td>
<td>Core collection: Hydraulic drill (Big Beaver) with a fitted liner was used to collect intact frozen soil cores to a maximum depth of 1m. See footnote 1.</td>
</tr>
<tr>
<td>Date_sampled</td>
<td>YYYY-MM-DD</td>
<td>Date core was collected.</td>
</tr>
<tr>
<td>Lab_processed</td>
<td></td>
<td>ORNL core processing: Soil cores were shipped frozen and returned to -20°C freezer and stored until further processing.</td>
</tr>
<tr>
<td>Date_processed</td>
<td>YYYY-MM-DD</td>
<td>Date core was processed at ORNL.</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td>Possible values: North Slope, Seward Peninsula</td>
</tr>
<tr>
<td>Locale</td>
<td></td>
<td>Possible values: Barrow, Council</td>
</tr>
<tr>
<td>Site</td>
<td></td>
<td>Possible values: Intensive Site 0, Intensive Site 1</td>
</tr>
<tr>
<td>Area</td>
<td></td>
<td>Possible values: A, B, C, D</td>
</tr>
<tr>
<td>Northing_UTM</td>
<td>meters</td>
<td>Core collection location. 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 4 North.</td>
</tr>
<tr>
<td>Easting_UTM</td>
<td>meters</td>
<td>Core collection location. 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 4 North.</td>
</tr>
<tr>
<td>Elevation</td>
<td>meters</td>
<td>Core collection location elevation.</td>
</tr>
</tbody>
</table>
# Data Dictionary Example

<table>
<thead>
<tr>
<th>Column_name</th>
<th>Units/format</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moisture_content_of_soil_layer</td>
<td>percent</td>
<td>Percent soil moisture during incubation.</td>
</tr>
<tr>
<td>pH_of_soil_layer</td>
<td>pH units</td>
<td>Initial pH of soils before incubation; Conducted on 2:1 distilled water to soil slurry (for mineral soil) or 4:1 distilled water to soil slurry (for organic soil).</td>
</tr>
<tr>
<td>Incubation_length</td>
<td>Days</td>
<td>Length of the incubation study in days.</td>
</tr>
<tr>
<td>Target_Incubation_Temperature</td>
<td>Degrees Celsius</td>
<td>Incubation temperature set on incubators (Thermo Scientific Precision Model 815 Incubator, Marietta, OH).</td>
</tr>
<tr>
<td>Anaerob_treat</td>
<td></td>
<td>Soil incubations - aerobic or anaerobic</td>
</tr>
<tr>
<td>Day_of_Incubation</td>
<td></td>
<td>Measurement day (if multiple measurements over incubation period).</td>
</tr>
<tr>
<td>CO2_per_g_dry_weight</td>
<td>mg C/ g dry weight</td>
<td>Cumulative CO₂ production per g dry soil over a given period of the incubation (e.g., between days 1 and 3). A value of “-9999” indicates missing data – see data quality flag column.</td>
</tr>
<tr>
<td>CO2_per_g_dry_weight_fl</td>
<td></td>
<td>Data quality flag for cumulative CO₂ production per g dry soil to indicate missing data. V0 is valid value, M1 is missing value because no value is available (a bad injection on the gas chromatograph).</td>
</tr>
</tbody>
</table>
## Data Quality Flag Table Example

<table>
<thead>
<tr>
<th>Flag Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0</td>
<td>Valid value</td>
</tr>
<tr>
<td>V1</td>
<td>Valid value but comprised wholly or partially of below detection limit data</td>
</tr>
<tr>
<td>V2</td>
<td>Valid estimated value</td>
</tr>
<tr>
<td>V3</td>
<td>Valid interpolated value</td>
</tr>
<tr>
<td>V4</td>
<td>Valid value despite failing to meet some QC or statistical criteria</td>
</tr>
<tr>
<td>V5</td>
<td>Valid value but qualified because of possible contamination (e.g., pollution source, laboratory contamination source)</td>
</tr>
<tr>
<td>V6</td>
<td>Valid value but qualified due to non-standard sampling conditions (e.g., instrument malfunction, sample handling)</td>
</tr>
<tr>
<td>V7</td>
<td>Valid value but set equal to the detection limit (DL) because the measured value was below the DL</td>
</tr>
<tr>
<td>M1</td>
<td>Missing value because no value is available</td>
</tr>
<tr>
<td>M2</td>
<td>Missing value because invalidated by data originator</td>
</tr>
<tr>
<td>H1</td>
<td>Historical data that have not been assessed or validated</td>
</tr>
</tbody>
</table>
Modified dataset from: http://dx.doi.org/10.5440/1235032

<table>
<thead>
<tr>
<th>CORE_ID</th>
<th>Date_sampled</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Region</th>
<th>Locale</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGADG0071</td>
<td>2012-04-16</td>
<td>-156.6046</td>
<td>71.2793</td>
<td>North</td>
<td>Slc Barrow</td>
<td>Intensive</td>
</tr>
<tr>
<td>NGADG0071</td>
<td>2012-04-16</td>
<td>-156.6046</td>
<td>71.2793</td>
<td>North</td>
<td>Slc Barrow</td>
<td>Intensive</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>METHYL_MERCURY</th>
<th>METHYL_MERCURY_SD</th>
<th>METHYL_MERCURY_DL</th>
<th>METHYL_MERCURY_FL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ng/g dwt</td>
<td>ng/g dwt</td>
<td>ng/g dwt</td>
<td></td>
</tr>
<tr>
<td>0.079</td>
<td>0.014</td>
<td>0.004</td>
<td>V0</td>
</tr>
<tr>
<td>-9999</td>
<td>-9999</td>
<td>0.004</td>
<td>M1</td>
</tr>
<tr>
<td>0.256</td>
<td>0.023</td>
<td>0.004</td>
<td>V0</td>
</tr>
<tr>
<td>-9999</td>
<td>-9999</td>
<td>0.004</td>
<td>M1</td>
</tr>
<tr>
<td>0.42</td>
<td>0.051</td>
<td>0.004</td>
<td>V0</td>
</tr>
<tr>
<td>-9999</td>
<td>-9999</td>
<td>0.004</td>
<td>M1</td>
</tr>
</tbody>
</table>
Standardizing the File Formats

- Save files in open software or open file format
- Proprietary software will most likely become obsolete or not back compatible
- Use text (ASCII) file formats for tabular data
  - e.g., .txt or .csv (comma-separated values)
- Utilize the standard formats established in the Data Management Plan

```
Core_ID, Date_sampled, Lab_processed, Date_processed, Region, Locale, Site, Area, Soil_type, upper_depth_of_soil_layer, lower_depth_of_soil_layer, Incubation_condition, Incubation_temp, Length_of_incubation, Microcosm_Replicate, Respiration_rate, Methane_production_rate
```

NGADG0017, 2012-04-12, ORNL, 2013-01-24, North Slope, Barrow, Intensive Site 1, A, organic, 0, 21.5, anoxic, -2, 2, 1, 0, 0
NGADG0017, 2012-04-12, ORNL, 2013-01-24, North Slope, Barrow, Intensive Site 1, A, organic, 0, 21.5, anoxic, -2, 2, 2, 0, 0
NGADG0017, 2012-04-12, ORNL, 2013-01-24, North Slope, Barrow, Intensive Site 1, A, organic, 0, 21.5, anoxic, -2, 2, 3, 0, 0

CSV example from dataset: http://dx.doi.org/10.5440/1109232
Help with Metadata Development

U.S. Geological Survey - Core Science Analytics, Synthesis, and Libraries - Online Metadata Editor (OME)

Online Metadata Editor (OME)

This tool will ask you simple, jargon-free questions about your dataset and produce a standardized metadata record. Using the Online Metadata Editor you can:

1. Log in and start the metadata records or upload and edit existing ones;
2. View all metadata records you have created or uploaded in the past;
3. Save metadata records and return later to complete them;
4. Save complete metadata records to your desktop.

Once your information is entered, the tool will output your record into a standard called the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata. The standard is widely used in Federal agencies for both geospatial and non-geospatial data.

The metadata record created will export in XML format, which is easily viewed in any web browser. The XML metadata record can then be submitted to metadata catalogs such as the USGS Science Data Catalog and data.gov.

This tool was developed through a partnership between the USGS Core Science Analytics and Synthesis (CSAS) Program and Oak Ridge National Laboratory.
You’ve Reached the Summit

✓ Determined what data to archive
✓ Described the content of the data
✓ Performed validation and quality control
✓ Saved in correct file formats
✓ Obtained a Digital Object Identifier (DOI)
✓ Submitted to an archive
✓ Shared with the public
Data Management

Best Practices – Part I

Model Outputs

Yaxing Wei
Data Scientist
weiy@ornl.gov
Why Model Outputs Management Matters?

• Consistency and Interoperability
  – To facilitate analysis, integration, and re-use

*Figure TS.14 in Intergovernmental Panel on Climate Change (IPCC) Assessment Report (AR5) WG1 Report*
What Makes Model Outputs Management Hard?

- **Spatial**
  - Site → Local → Regional → Global
  - Coarse resolution → Fine resolution

- **Temporal**
  - Paleo → Recent past → Future projections
  - Decadal → Annual → Monthly → Daily → Hourly

- **Voluminous**
  - 36 TB (CMIP3) x 50 → 1.8 PB (CMIP5) x 50 → 90 PB (CMIP6)*

- **Various Formats**
  - ASCII, Binary, netCDF, …

* CMIP: Coupled Model Intercomparison Project
  * Source: Michael Lautenschlager, CMIP5 Data Management, 2013
Best Practices for Managing Model Outputs

- Organize model outputs properly
- Name your data files and variables properly
- Accurately define spatial information
- Accurately define temporal information
- Provide accurate and rich metadata
- Leverage community conventions
- Effectively share data within team and beyond
Best Practices for Managing Model Outputs

• Organize model outputs properly
• Name your data files and variables properly
• Accurately define spatial information
• Accurately define temporal information
• Provide accurate and rich metadata
• Leverage community conventions
• Effectively share data within team and beyond
Organize Model Outputs Properly

• Balance the way you divide model outputs into files
  – Avoid too many small files
  – Avoid lumping everything together into huge files
  – Consider needs of target communities

• If your gridded model outputs are too big and need to split them into multiple files
  – First, consider separating independent variables
  – Then, consider splitting based on time, e.g. per year or decade
  – Splitting data spatially will always be your last choice

• Organize data files in hierarchical directory structure
  • e.g. project / model / simulation / variable / …
Model Outputs Split by Space

- **Daymet V3**
  - Provide gridded estimates of daily weather parameters for North America (including Puerto Rico) and Hawaii, on a 1-km grid, from 1980 to 2015.
  - Website: [https://daymet.ornl.gov](https://daymet.ornl.gov)
Original outputs of Daymet V3 were split onto 1060 2-deg x 2-deg tiles. There were 267120 individual files with this file organization scheme.
Daymet V3 outputs were reorganized by merging all tiles in each of the 3 regions together (Continental NA, Hawaii, and Puerto Rico), then split by year. This scheme yielded 756 individual files, with each of them under 5GB in size.
Best Practices for Managing Model Outputs

- Organize model outputs properly
- Name your data files and variables properly
- Accurately define spatial information
- Accurately define temporal information
- Provide accurate and rich metadata
- Leverage community conventions
- Effectively share data within team and with public
Name Data Files and Variables Properly

• Use descriptive file names
  – Model name
  – Simulation code
  – Version number
  – Variable name
  – Space info (e.g. place name and/or resolution)
  – Time info (e.g. range and/or resolution)

Example good filenames:

- BIOME-BGC_BG1_Monthly_GPP_V2.nc4
- rlds_Amon_CESM1-CAM5_historical_r1i1p1_185001-200512.nc
- daymet_v3_srad_2012_na.nc4
Name Data Files and Variables Properly

• Use unambiguous and “interoperable” variable names
  – Build a table that defines the “short name”→“full name” pairs for variables in your project

<table>
<thead>
<tr>
<th>Variable</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>tmax</code></td>
<td><code>land_surface_air__daily_time_max_of__temperature</code></td>
</tr>
<tr>
<td><code>srad</code></td>
<td><code>atmosphere_radiation~incoming~shortwave__transmitted_energy_flux</code></td>
</tr>
</tbody>
</table>

  – Consider “standard names” or common vocabularies used in the community
    • CMIP6 Variable Definitions (https://earthsystemcog.org/projects/wip/CMIP6DataRequest)
    • Climate & Forecast (CF) Standard Names (http://cfconventions.org/standard-names.html)
    • Community Surface Dynamics Modeling System (CSDMS) Standard Names (https://csdms.colorado.edu/wiki/CSDMS_Standard_Names)
  
  – Use standard data units
    • International System of Units (SI)
    • UDUNITS-2
Best Practices for Managing Model Outputs

• Organize model outputs properly
• Name your data files and variables properly
• Accurately define spatial information
• Accurately define temporal information
• Provide accurate and rich metadata
• Leverage community conventions
• Effectively share data within team and with public
Accurately Define Spatial Information

• Specify the grid space your model is running in
  – Type of grid space
    • Geographic lat/lon grid
    • Projected grid
  – Centers and borders of grid cells
  – Spatial Reference System (SRS)
Define the Grid Space For Model Outputs (1)

- NACP regional TBM output
  - Extent
    - West: -170.0
    - South: 10.0
    - East: -50.0
    - North: 84.0
  - Resolution
    - X-resolution: 1-degree
    - Y-resolution: 1-degree
  - SRS
    - Spherical Geographic Lat/Lon (R=6370997m)
Define the Grid Space For Model Outputs (2)

• SRS for Daymet data
  – Projection: Lambert Conformal Conic

  projection units: meters
datum (spheroid): WGS_84
1st standard parallel: 25 deg N
2nd standard parallel: 60 deg N
Central meridian: -100 deg (W)
Latitude of origin: 42.5 deg N
false easting: 0
false northing: 0
Best Practices for Managing Model Outputs

- Organize model outputs properly
- Name your data files and variables properly
- Accurately define spatial information
- Accurately define temporal information
- Provide accurate and rich metadata
- Leverage community conventions
- Effectively share data within team and with public
Accurately Define Temporal Information

• Calendar your model is using
• Overall start and end temporal representation of a data variable
• Time point/period that each data value represents
• Temporal frequency of a data variable
Calendar

• Determine how many days in each month/year

\begin{itemize}
  \item \textbf{julian}: one leap year in every 4 years
  \item \textbf{gregorian}: leap year if either (i) it is divisible by 4 but not by 100 or (ii) it is divisible by 400
  \item \textbf{proleptic\_gregorian}: gregorian calendar extended to dates before 1582-10-15
  \item \textbf{365\_day}: no leap year, Feb. always has 28 days
  \item \textbf{360\_day}: 30 days for each month
  \item \textbf{366\_day}: all leap years
\end{itemize}

\textbf{gregorian is the internationally used civil calendar}
Time Point and Period

- **ISO 8601:** *Data elements and interchange formats – Information interchange – Representation of dates and times*

- **Time point:** `YYYY-MM-DDThh:mm:ss.sTZD`

  - Example: `2010-03-22T18:00:00.00-06:00`


  - Example: `PT1H20M30S`
Best Practices for Managing Model Outputs

- Organize model outputs properly
- Name your data files and variables properly
- Accurately define spatial information
- Accurately define temporal information
- Provide accurate and rich metadata
- Leverage community conventions
- Effectively share data within team and with public
Provide Accurate and Rich Metadata (1)

• Make your model outputs to be easily found, understood, and re-used

• Key elements
  • What does the data set describe?
  • Why and how was the data set created?
  • Who produced the data set and Who prepared the metadata?
  • How reliable are the data?; what is the uncertainty, measurement accuracy?; what problems remain in the data set?
  • What assumptions were used to create the data set?
  • What is the use and distribution policy of the data set? How can someone get a copy of the data set?
  • Provide any references to use of data in publication(s)
Provide Accurate and Rich Metadata (2)

• Link your model outputs to the context they were created
  – Model codes
  – Input data (data files, parameters, initializations)
  – Configurations
  – Post-processing/distillation and analysis
  – Publication/presentation results

• Consider archiving the whole model package

  • [NGEE-Arctic Example] Addressing numerical challenges in introducing a reactive transport code into a land surface model: A biogeochemical modeling proof-of-concept with CLM-PFLOTRAN 1.0: Modeling Archive
Leverage code repositories (e.g., github.com and bitbucket.org) to actively manage model source codes.
Best Practices for Managing Model Outputs

• Organize model outputs properly
• Name your data files and variables properly
• Accurately define spatial information
• Accurately define temporal information
• Provide accurate and rich metadata
• Leverage community conventions
• Effectively share data within team and with public
Leverage Community Conventions

• One step forward for consistency and interoperability
• Choose proper data formats to maximize model outputs re-use
  – netCDF
• Follow conventions to create self-descriptive model output files
  – Climate & Forecast (CF) convention (http://cfconventions.org)
Best Practices for Managing Model Outputs

- Organize model outputs properly
- Name your data files and variables properly
- Accurately define spatial information
- Accurately define temporal information
- Provide accurate and rich metadata
- Leverage community conventions
- Effectively share data within team and beyond
Effectively Share Data Within Team and With Broader User Communities

• Set up team and public data repositories
  – Dedicated data storage and management servers
  – For non-sensitive data, commercial products like Dropbox provide out-of-box features, e.g. access control, version tracking, and backup, for setting up data repositories

• Set up services to support on-demand data access
  – Variable-, spatial-, temporal-subset
    • OPeNDAP: Open-source Project for a Network Data Access Protocol
    • NCSS: NetCDF Subset Service
Leverage Advanced Data Transfer Services

- Multistreaming transfer utilities
  - BBCP
  - GridFTP and Globus Transfer
Summary

• Follow best practices to create self-descriptive and interoperable data, your model outputs will be readily available for visualization, analysis, and re-use in future researches.

• Leverage proper tools and services to share data within team and/or with public so research can be conducted collaboratively and effectively within your team and usage of your model outputs can be maximized.
THANK YOU!

• Thanks for participating in this webinar
• Slides and a recording of this webinar will be available
  – Link will be sent to you by e-mail.
• Please fill out the survey at the bottom of your screen
• Send any additional feedback to sanseverinoj@ornl.gov
• Future webinars will be announced by e-mail