Metadata Standards for In-situ Observational datasets

Giri Palanisamy, Misha Krassovski, Luca Cinquini, Sig Christensen, Renata McCoy, Tom Boden, Raymond McCord and Galen Shipman

GO-ESSP, May 10, 2011
Inclusion of Observations in CMIP5/IPCC-AR5

• In the past 6 months, the community has engaged in a collaborative effort to include observational datasets into CMIP5
  – Two NASA-PCMDI workshops held in Fall 2010 with participation from ORNL, NOAA etc.
  – Continued community engagement on email, conf calls, wiki w/ IPSL, BADC etc.
• Goal: increase usage of observations in climate change research, especially for IPCC Assessment Reports
  – Observations are critical to evaluating, improving and scoring models
• Strategy: make it possible for users to find, access and analyze observations with the same processes and tools they use for models
  – Structure all observations homogenously (across instruments, platforms, agencies…)
  – Make observational datasets “look” just like models
Working w/ PCMDI, providers of observational datasets for CMIP5 have finalized a common specification for data and metadata conventions, and for the observations publishing process:

- Data format: NetCDF/CF, single variable files
- Capture search & discovery metadata as NetCDF global attributes
  - Harvested by publisher, encoded in THREDDS, used as facet values
  - Use same categories as models: realm, CF variable, ...
- Common convention for directory structure and filenames
  - “DRS specification for observations”
- Must pass CF and CMOR checkers
- Each dataset must be accompanied by Technical Note
- Publish onto Earth System Grid Federation

http://esgf.org
NASA Observations

• NASA is working towards including satellite data into CMIP5
  – Level 3 datasets (uniformly gridded, not swath)
  – Especially prepared by science teams for CMIP5 (not original products stored at DAACs)
  – “Best” products selected across missions:
    • AIRS, MLS: Air Temperature & Specific Humidity
    • CCMP, QuickSCAT: Ocean winds
    • Top Of the Atmosphere radiation: CERES
    • Total Cloud Fraction: MODIS
    • Sea Surface Height: TOPEX/JASON
    • Sea Surface Temperature: AMSR-E
    • Precipitation: TRMM, GCPC
    • Sea Ice Fraction, Thickness, Albedo: NSIDC
  – Products to be published @ JPL and propagated throughout ESG Federation
Provide the national and international scientific community with the infrastructure needed for scientific research on global change.

Research includes the study of alterations to climate, land productivity, oceans, water cycle, atmospheric chemistry, and ecological systems.

The goal of ASR, is to quantify the interactions among aerosols, clouds, precipitation, radiation, dynamics, and thermodynamics to improve fundamental process-level understanding, with the ultimate goal to reduce the uncertainty in global and regional climate simulations and projections.

http://www.arm.gov
ARM Research Sites and Measurements

Measurements and Instruments:

- Cloud profiles: millimeter radar and lidar
- Temperature/relative humidity/wind profiles: radiosondes
- Column water: microwave radiometer
- Column aerosol: solar spectral radiometer
- In situ aerosol optical and cloud nucleation properties
- Surface radiation budget: solar and terrestrial IR radiometers
- Surface meteorology: T/RH/wind
CMIP5 variables from ARM Data

Initial Mapping of ARM variables

- Based on CMIP5 Categories:
  - Cloud diagnostic fields from specified locations
  - Monthly Mean Atmospheric Fields and Some Surface Fields

- Using ARM Climate Model Best Estimate Product

- Data from 3 ARM fixed sites
  - SGP: 1993 to 2008
  - NSA: 2001 to Jan 2010
  - TWP: 1996 to Jan 2010 (3 locations)

NetCDF CF Datasets available for:

- Air Temperature
- Relative Humidity
- Surface Air Pressure
- Wind Speed
- Eastward Near-Surface Wind
- Northward Near-Surface Wind

Next release will include:

- Cloud fraction
- Longwave Radiation (upwelling and downwelling)
- Shortwave Radiation (upwelling and downwelling)
AmeriFlux Network

142 sites in 5 countries; 94 active sites, 48 inactive sites

Participation requirements:

- Make year-round core measurements using the eddy-covariance technique
- Submit data to the Carbon Dioxide Information Analysis Center (CDIAC) within 1 year of collection
- Participate in AmeriFlux Science Meetings and synthesis & modeling activities

http://public.ornl.gov/ameriflux/
AmeriFlux Data Levels (Products)

For meteorological data:

Level 1 — Processed Data Provided by the Site Investigators: Data files provided by the site measurement teams are processed by CDIAC to produce Level 2 data products (see below). Level 1 data files are archived long-term by CDIAC. Level 1 data are available to users in their native submission formats but users are encouraged to use Level 2, 3 or 4 AmeriFlux data products.

Level 2 — Data Checked & Formatted by CDIAC: Data received from individual sites are reviewed and incorporated into a network-wide AmeriFlux database. The review process includes checks for consistent units, naming conventions, and reporting intervals and reformatting is often necessary to maintain consistency within the larger network-wide database.

Level 3 — Processed Data With Quality Flags Assigned and NEE calculated using standardized techniques: AmeriFlux Level 2 files are processed by the European flux data activity to produce Level 3 and 4 files identical to the European regional network.

Level 4 — Gap-filled & Adjusted Data Files with GEP & Re Estimates: Level 4 files contain gap-filled (ANN and MDS techniques) and $\ddot{u}t$ar filtered records, complete with calculated gross productivity and total ecosystem respiration terms on varying time intervals including hourly, daily, weekly, and monthly with flags regarding the quality of the original and gap-filled data.

For biological data:

Level 1 — Processed Data Provided by the Site Investigators: Site investigators provide ecological data and disturbance information in various ways including submission of the Law et al. Biological-Ancillary-Disturbance-Methodology (BADM) template. Level 1 biological data are available to users in their native submission formats.

Level 2 — Uniform Law et al. Biological-Ancillary-Disturbance-Methodology (BADM) Templates: The original BADM templates submitted by the site measurement teams are processed by the Microsoft/LBL data team to produce improved templates with uniform fields (e.g., consistent date representations) and better suited for automated processing.
Directory structure convention for datasets included in CMIP5

/<activity>/<product>/<realm>/<variable>/<frequency>/<obs_structure>/<institution or agency>/<obs_project>/<instrument>/<version>/filename

<activity> = "obs4cmip5" (fixed)
<product> = "observations" (fixed)
<realm> = the physical realm where the data is defined, chosen from the CMIP controlled vocabulary, same as for models (example: "atmos", "ocean", "land", "landIce", "seaIce", "aerosol" "atmosChem", "ocnBgchem")
<variable> = the measured physical quantity, set to a value from the CMIP controlled vocabulary, same as for models (example: "ta", "hus", etc.)
<frequency> = the temporal frequency of the data, chosen from the DRS controlled vocabulary, same as for models (example: "1hr", "day", "mon" etc.)
<obs_structure> = a keyword that reflects the internal organization of the data, with the following possible values:
- grid
- station
- trajectory
- swath
<institution or agency> : the funding agency or institution, chosen from a controlled vocabulary
<obs_project> : the funded project, chosen from a controlled vocabulary
<instrument> : the instrument that took the measurements (or possibly the data product providing the measurements), chosen from a controlled vocabulary
<version> : the dataset version, encoded according to the same specification as models: vYYYYMMDD

Examples:

ARM:/obs4MIPS/observations/atmos/uas/1hr/in_situ_sites/usdoe/arm/armbe/v120110314/files

Ameriflux:/obs4MIPS/observations/atmos/ta/1hr/in_situ_sites/usdoe/ameriflux/tower/v120110101/files
File naming convention for in-situ datasets included in CMIP5

<variable>_<project>_<location>_<instrument>_<processing_level_and_product_version>_<start_date>_<end_date>.ext

<variable> = measured variable
<project> = project short name (ameriflux, arm..)
<location> = site name
<instrument> : the instrument that took the measurements (or possibly the data product providing the measurements), chosen from a controlled vocabulary
<processing_level_and_product_version> = data processing level and version
<start_data> = observation start date in format YYYYMMDD
<end_data> = observation end date in format YYYYMMDD

Examples:

Ameriflux: ta_ameriflux_barrow_tower_L2v1_20110101_20110209.nc
ARM: ps_arm_darwin_cmbe_v1p1_20020101-20101231.nc
Global Attributes for in-situ datasets included in the NetCDF files

Required Global Attributes:

- cmor_version: 2.5.7
- contact: renata@llnl.gov or armarchive@ornl.gov
- Conventions: CF-1.4
- creation_date: 2010-03-23-T05:56:23Z
- data_structure: station
- frequency: 1hr
- institute_id: ARM_DOE

Institution: Atmospheric Radiation Measurement Program - U.S. Department of Energy

- mip_specs: CMIP5
- product: observations
- project_id: Obs4MIPS
- realm: atmos

- references: published web references

- source: ARMBE-Atm-1.0 2010 Atmospheric Radiation Measurement Program Best Estimates, atmospheric state profiles, Version 1.0
- source_id: ARMBE-ATM-1.0
- table_id: obsSites-ARMBE
- source_type: in-situ value added
- tracking_id: 02c0e6c5-9467-382e-8f9b-9300a64ac3cd

Optional Global Attributes:

- Comment
- history
- Title
- Location
- Experiment_id

https://oodt.jpl.nasa.gov/wiki/display/CLIMATE/Data+and+Metadata+Requirements+for+CMIP5+Observational+Datasets
Common Metadata Standards Used in Observational Data

Popular Metadata Standards:
- FGDC, FGDC BDP, EML, Dublin Core, Darwin Core, ISO 19115, ISO 19139, DIF

Advantages of using these metadata standards:
- Metadata fields to support Data Discovery:
- Hierarchical metadata fields
- Data Quality information
- Data Provenance
- Scientific keywords
- Data Citation
- Service level information
Thanks!