Data Processing Steps Between Observation Data, Model Data, and Live Access Server in the AOSN Program

The primary focus of the data processing steps between observation data, model data, and LAS at MBARI has been on getting the data into a consistent, well described format, and making the data accessible to researchers who can perform their own analyses.

A combination of Distributed Ocean Data Servers (DODS, also called OpenDAP) and Live Access Servers (LAS) have been pursued, along with an intense campaign of getting standard variable names and standard data preparation techniques refined. The LAS server allows users to generate side-by-side plots of model and observation data, and to download the data in a variety of formats for off-line viewing. Both the DODS and LAS servers will require that users provide identifying information to access the data. The identifying information combined with usage data will be used as a measure of success of the data sharing effort.

Over time, advanced data comparison, selection, visualization, and output capabilities (as well as statistical analyses) will be added to LAS server, as requested by users of the server. Some of the visualization capabilities already added have included stick plots for mooring winds and property-property plots for gridded data.

Data selection, visualization, and output capabilities envisioned for the near-term include

A) the ability to pick a 4D region, then have LAS report the data sets that contain data within that region;

B) the ability to select multiple scattered data sets at once, and provide merged output from the multiple data sets;

C) the ability to easily do side-by-side comparisons of gridded and scattered data;

D) the ability to plot curvilinear gridded data, and

E) the ability to create ODV-compatible output files on the fly.

As these capabilities are developed, they will be shared with the LAS development team so that they can be used throughout the environmental data management community.

Appendix A describes in detail the steps currently used to process the observation and model data at MBARI.

Appendix B describes the conventions used to generate consistently formatted data files, along with the metadata that is attached to processed data sets.

Appendix C describes the efforts to generate standard variable names.

Appendix A: Detailed Data Processing/Handling Steps

Step 1: Observation and model data are received at MBARI.

Investigators and modelers are responsible for sending the data via FTP to Polarbear. The exception in for large model runs that are impractical to send via FTP.

Investigators are also responsible for sending documentation that describes the data. This is usually sent to the 'doc' folder within the data folder on Polarbear.

In the case of very large model outputs that cannot be sent to Polarbear via FTP, investigators are expected to host the data on a DODS server accessible to the MBARI LAS server.

Step 2: Observation and model data are described at MBARI.

For 2003 the observation campaign, the las_info worksheet is filled in by MBARI staff based on the data and its description supplied by data generators. This las_info worksheet contains two tables: 'global', and 'variables'. The global table contains one row for each data set, while the variable table contains information about each variable in each dataset. For the 2005+ observation campaigns, it is envisioned that data generators will fill in the spreadsheet when the data is submitted to MBARI.

Together, the two tables contains adequate metadata to describe that dataset, both for metadata needs, and for running the las_convert MATLAB routine, which converts the data according to the "MBARI/AOSN" conventions. See Appendix B for the MBARI/AOSN conventions.

Step 3: Observation and model data are converted at MBARI.

This las_info worksheet is used to generate a text 'properties' file for each dataset, which contains all the data from the 'global', and 'variables' tables for each variable. The table is formatted like a java properties file, with a name=value pair for each bit of metadata.

The las_convert MATLAB macro is then run, using the text 'properties' file as an input. For most data sets (regular MATLAB, Text or NetCDF files), the las_convert macro can read in the data, based on the metadata. For more complex data set inputs (SIO gliders, ICON model, or Western Flyer CTD, for example), custom macros may be called by the las_convert macro.

For gridded data, outputs from the las_convert macro include A) a NetCDF file that follows the MBARI/AOSN conventions, and B) an xml file containing metadata adequate for hosting the dataset on the AOSN LAS server.

In the case of remotely stored data accessed through DODS, the las_convert macro will only generate the xml file for allowing the LAS server to access the data.

For scattered data, outputs from the las_convert macro include the above products, plus C) an ASCII table of the data, suitable for importing into an SQL database, D) a text file with the SQL commands for importing the ASCII table, E) a graphical trajectory map for inclusion in the LAS server, and F) another XML file that describes the graphical trajectory map for the LAS server.

Appendix B: MBARI/AOSN Conventions

(Adapted from EPIC netCDF General Conventions and GDT netCDF conventions)

Filenames: NetCDF files should have the file name extension ".nc"

Variable Names: See MBARI AOSN Variable Naming Convention, Appendix C

Variable Types: Normally, Float for data variables,

Double for coordinate variables

Unit Names: use SI names whenever possible, substituting "u" for micro
use space " " to indicate multiplication,
use numbers to indicate raising to a power use of "/" to indicate division is not recommended

Time: preferred time unit is "seconds since 1970-01-01 00:00:00" (EPIC 601)
(works well in a variety of processing and display software)
alternate time unit is "day as %Y%m%d.%f", (YYYYMMDD.ffffffff) (EPIC 620)

Global Data Attributes (Global Metadata) collected for each dataset:

Attribute	**Description**
Dataset_id	unique name used to identify variables and name .info file
Production	how the data was produced ("model_data" or "observations")
Platform	platform name (used to form data directory)
Title	display title for data set
CREATION_DATE	file creation date and time
History	data evolution history
Institution	who made or supplied the data
url	url link to institution
Doc	url link to information
Conventions	name of conventions used by the file,(eg. MBARI/AOSN)
INST_TYPE	Instrument type
source_path	Path to source file(s) (or regular expression with [ ] for path)
source_file	Source file, or regular expression with [ ] for source file(s)
source_type	Either "MATLAB,[arrayName]", "Text, #headerlines", "NetCDF", File,FileFunction, Custom,CustomFunction, or DODS
output_file_base	If indicated, used to concatenate multiple input files
filename_datetime	Format of Date/Time in FileName ([day as %Y%m%d]biolum.txt)
DATA_TYPE	Must be one of: "CTD", "TIME", "SCATTER", "TRACK" or "GRID"
DATA_SUB_TYPE	Description of DATA TYPE (see worksheet "data sub types")
DATA_ORIGIN	Data Origin Information
COORD_SYSTEM	Either "GEOGRAPHICAL", "ROTATED", or "CURVLINEAR"
Axes	Comma separated list of coordinate axis names
Associations	Comma separated list of non-coordinate axis names
Variables	Comma separated list of all variable names

Variable Data Attributes (Variable Metadata) collected for each variable:

var_dataset_id	Match to unique dataset_id in global worsheet. Can include wildcards ?, *, or []
type	variable type (axis, association, variable)
name	variable name (domain_param_qual)
skip	enter '_' to skip a variable
domain	variable domain
generic_name	variable generic name (parameter)
qualifier	variable qualifier(s)
old_name	original name in data source (use "_[name]" for dummy vars, or to convert pressure to depth) OR use name=value to assign, or .name for attributes
long_name	variable long name (Domain Param Qual)
old_long_name	original long name in data source
units	variable unit
old_units	original units in data source
symbol	MATLAB display unit (m^{-1})
epic_code	variable code; write only when it is defined in epic.key file
dimensions	dimensions (comma separated)
axis	Characters from TZYX- corresponding to dimensions of variable.
old_axis	Characters from TZYX- corresponding to old dimensions of variable.
_FillValue	substitute for invalid values (for AOSN use 1e35)
missing_value	substitute for missing values (=_FillVale for scatttered) (for AOSN use 1e34, 1e35 for scattered data)
valid_min	smallest valid value of a variable
valid_max	largest valid value of a variable
calendar	for time axis, Gregorian
positive	for depth axis, positive direction, either "up" or "down"
domain	for longitude axis = 0:360 or -180:180
modulo	for longitude axis = 360.0
topology	for longitude axis =circular
comment	Comment
associate	axes which data varies against, but which are not dimensions.
quantity	variable name with extra qualifiers
short_name	variable domain_ parameter
FORTRAN_format	data Fortran format "Flag[width.precision]"
C_format	data C language format "%[flag][width][.precision][char]"
regridX	Comma separated min, delta, max to regrid the X-axis (Longitude) data.
regridY	Comma separated min, delta, max to regrid the Y-axis (Latitude) data.
regridZ	Comma separated min, delta, max to regrid the Z-axis (Depth) data.
source_file	Source file, or regular expression for source file(s), If different than global attribute
institution	Who made or supplied the data, if different than global attribute
production	how the data was produced, if different than global attribute
serial_number	Serial # of instrument, if different than global attribute
add_offset	additive offset for packing data
scale_factor	multiplicative factor for packing data

Appendix C: MBARI/AOSN Variable Naming Convention

Variables are constructed from three components: “domain”, “parameter”, and “qualifier”.

“Domain” is always mandatory, unless it is explicitly implied. For example, time, latitude, and longitude have a geospatial domain, but since that domain is obvious for earth-based measurements, it may be omitted.

The lowest appropriate level in the following hierarchy of domains should be used. A lower level domain always implies the higher levels in the hierarchy:

- Biosphere
  - Atmosphere
    - Air
    - Wind
  - Hydrosphere
    - Ocean
      - Seasurface
      - Sealevel
    - Lake
    - Estuary
    - Geosphere
      - Soil
- Platform
  - Instrument
    - Sensor
- Deployment
  - Cast
    - Sample

“Parameter” (also called generic name) is always mandatory. It is the quantity being measured, and may be duplicated several times within a dataset. The standard list of parameter names is still under development, but tentatively includes:

altitude_above_sealevel
bin_length
bioluminescence
blank_length
carbon_dioxide
carbon_dioxide_dissolved
chlorophyll
conductivity
current_velocity_east
current_velocity_north
current_velocity_up
depth
diode
direction

echo_intensity
flow
fluorescence
heading
heading_bias
humidity
index
irradiance
irradiance_downwelling
latitude
logitude
longitude
map
nitrate
optical_backscatter
optical_particulate_backscatter
optical_particulate_volume_scattering_fraction
optical_volume_scattering_fraction
oxygen_dissolved
photosynthetically_active_radiation
pings
pitch
pressure
pump
radiance
radiance_downwelling
radiance_upwelling
roll
roll
salinity
speed
speed_of sound
temperature
time
time_between_pings
unused
velocity_east
velocity_north
voltage
wavelength

The optional "Qualifier" is used to differentiate between identical sets of domains and parameters. There will be no standard list of qualifiers.