OceanSITES Data Users' Guide

Version 1.0
Feb. 05, 2019

Document History

Version	Date	Comment
0.0	03/05/2010?	TT,IS: originalHow to Access OceanSITES Data
0,1	10/02/2014	Galbraith: new document compiled from several sources
0.2b	03/16/2016	Galbraith: Just cleaning up track changes
0.3	08/10/2016	Lankhorst: added quick-start guide, minor edits
0.4	08/11/2016	Lankhorst: added FAQ, edits up to and including "how to access".
0.5b	01/17/2019	Galbraith: reformat slightly, add some standard elements, discuss FTP site
1.0	02/05/2019	Lankhorst: adjudicate comments and finalize for public release

1. OceanSITES Overview

The OceanSITES program is the global network of open-ocean sustained time series sites, being implemented by an international partnership of researchers. OceanSITES provides fixed-point time series of various physical, biogeochemical, and atmospheric variables at different locations around the globe, from the atmosphere and sea surface to the seafloor. The program's objective is to build and maintain a multidisciplinary global network for a broad range of research and operational applications including climate, carbon, and ecosystem variability and forecasting and ocean state validation, and to provide data from all those stations.

All OceanSITES data are publicly available. More information about the project is available at: http://www.oceansites.org, where you can find the OceanSITES mission statement, governance documents, and general contact information.

Technical documentation

Please see the Quick-Start Guide, below.

OceanSITES Data Users' Guide (this document)
For data users, formerly called How to Access OceanSITES Data, this document contains an outline of GDAC data directory structure and ftp/opendap access, and basic information on how to get and use OceanSITES data. It should be used in conjunction with the Data Format Reference Manual.
Please see the Quick Start Guide, below

OceanSITES Data Format Reference Manual
For data producers and users, formerly called the OceanSITES Users Manual, this document contains a description of the OceanSITES NetCDF specification, code lists for required variables, and OceanSITES Data Management terms of reference.

OceanSITES Data Providers' Guide
For data producers (DACs and PIs), this document contains guidelines for collecting and providing metadata and data, the file naming scheme, and how to upload data to the OceanSITES GDACs. It should be used in conjunction with the Data Format Reference Manual.

2. Quick-Start Guide

You want to get started with OceanSITES data quickly, without reading any further? Here's what you need to do. If you cannot figure it out, please do read on, as the rest of this document is meant to fill you in on the details!

1. Locate the anonymous FTP server nearest to you. One is in France, the other in the US. There are other data access routes, but FTP is the standard method.
    a. ftp://ftp.ifremer.fr/ifremer/oceansites/
    b. ftp://data.ndbc.noaa.gov/data/oceansites/
2. Locate the data files you want. At each of the two servers, there is an "index" file called **oceansites_index.txt**, which contains a list of all files and their contents (time, latitude, and longitude ranges, as well as data types). The index file is a comma-separated list. You need to identify from this list which files you want.
3. Realize that OceanSITES data come as individual files in **NetCDF format**. Most data are available as deployment-by-deployment data files at the native instrument resolution in time. Some data are available in a form that concatenates multiple deployments, and/or contains data that are interpolated or gridded in depth and/or time. That said, download the files that you identified in the previous step.
4. Understand that the NetCDF files contain both the data and metadata. **You need to read at least some of the metadata** to understand the file contents. Metadata is contained in fields called "attributes". These apply either to a particular data variable in the file, or to the whole file. The latter are called "global attributes". OceanSITES follows existing naming conventions with most of these attributes, most notably the **CF and ACDD** conventions; see the Links section for more information.
    Examples:
    a. A file might contain a data variable for temperature. It will likely also contain an attribute to that variable called "units", which may read "degree_Celsius".
    b. The same file may hold a global attribute called "principal_investigator", which shows that person's name.
5. Read the global attributes "title" and "summary", if they exist, to understand what the data are. **Read the global attributes that describe the license** under which you may use the data, and be sure to abide by that. These are typically in attributes called "license", "acknowledgement", and "citation". Proper etiquette is to always acknowledge the source(s) of the data you used whenever you publish anything based on these data.
6. Locate the data variables in the file that you need. Use the variable attributes "**standard_name**" or "**long_name**" to figure out what each variable means.
    Example: if you need water temperature, look for standard_name "sea_water_temperature".
7. Understand that there may be ancillary data that you need to understand the data themselves. These are usually listed in attributes called "**ancillary_variables**".
    Example: the temperature data from a mooring may include bad data such as outliers, and there may be a separate variable with quality control flags that identify which data points are good and which ones are bad. In that case, you will need the quality control flags to extract only those data points that are flagged good.
8. **That's it, you're good to go!**

3. Obligations of Users of OceanSITES Data

Respect distribution and citation requirements

The overall OceanSITES guidance is to follow the CLIVAR (Climate and Ocean: Variability, Predictability and Change) data policy, which calls for free, unrestricted, and timely exchange of data. However, this does not necessarily mean that nothing is asked of the user. Typically, the user will be asked to give proper credit to the data providers for any use of the data. Such distribution and citation requirements are usually set out in the "license", "citation", and "acknowledgement" attributes in each NetCDF data file. If you use OceanSITES data, you must read these metadata, and find and follow all requirements that are asked of you as part of your license to use the data.

At a minimum, if no more specific information is included in the data file, OceanSITES asks all users to include the following standard citation with any publications arising out of the use of OceanSITES data:

> **"These data were collected and made freely available by the international OceanSITES project and the national programs that contribute to it.".**

If you wish to include more information, OceanSITES would appreciate it if you could cite the following publication as a reference to OceanSITES:

Send, U. & Co-Authors (2010). "OceanSITES" in Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society (Vol. 2), *Venice, Italy, 21-25 September 2009, Hall, J., Harrison, D.E. & Stammer, D., Eds., ESA Publication WPP-306, doi:10.5270/OceanObs09.cwp.79.*

Likewise, OceanSITES encourages display of its logo on any and all graphical presentations that use OceanSITES data. The logo is available in several formats on the OceanSITES website: http://www.oceansites.org.

Recognise the DISCLAIMER that applies to all OceanSITES data

OceanSITES data are published without any warranty, express or implied. The user assumes all risk arising from the use of OceanSITES data. OceanSITES data are intended to be research-quality and include estimates of data quality and accuracy, but it is possible that these estimates or the data themselves contain errors.

It is the sole responsibility of the user to assess if the data are appropriate for his/her use, and to interpret the data, data quality, and data accuracy accordingly.

OceanSITES welcomes users to ask questions and report problems to the contact addresses listed in the data files or on the OceanSITES web page.

4. How to Access OceanSITES Data

Data Access Pathways

- **Anonymous FTP**: Two Global Data Access Centers (GDACs) provide access points for OceanSITES data via ftp servers. One is in France at the Coriolis center (http://www.coriolis.eu.org), the other is in the US at the NOAA National Data Buoy Center (NDBC, http://www.ndbc.noaa.gov). The two servers at the GDACs are synchronized at least daily to provide the same OceanSITES data redundantly. The user can access the data at either GDAC's ftp site:
  - **ftp://data.ndbc.noaa.gov/data/oceansites** or
  - **ftp://ftp.ifremer.fr/ifremer/oceansites**

- **OPeNDAP/THREDDS**: The GDACs also provide OceanSITES data via a THREDDS:
  - **http://dods.ndbc.noaa.gov/thredds/dodsC/data/oceansites/DATA/**
  - **http://tds0.ifremer.fr/thredds/CORIOLIS-OCEANSITES-GDAC-OBS/CORIOLIS-OCEANSITES-GDAC-OBS.html**

Data Organization

OceanSITES measurements are collected at locations called "sites". Each site is occupied by at least one "platform", and over the lifecycle of a project, there may be multiple subsequent "deployments" of a given platform. The underlying organization of OceanSITES data follows the concept of sites, platforms, and deployments: the data directories on the FTP servers reflect the sites, and within each directory, the file names of the NetCDF data files contain both platform and deployment identifiers. A list of the site names (full names and abbreviated names) along with maps showing their locations are provided on the OceanSITES website.

There are two main directories on the GDAC ftp sites. The DATA directory contains edited data as close to the original observed data as possible. Some files are merged across depths to present data in its most useful form.

The DATA _GRIDDED directory contains gridded data such as aggregated long time series files that have been merged across deployments, and may contain higher-level products derived from the observed variables.

The project attempts to prevent duplicate data points in the DATA directories, however care must be taken because, in some cases, certain observations may appear in more than one file. For example, temperatures may be presented in a 2 dimensional file of all temperature values, and again in a file that contains all the salinity measurements. The same data points will appear in the DATA _GRIDDED directory.

The exact naming conventions for files in the DATA directory tree are specified in the OceanSITES Data Format Reference Manual, but generally, the deployment-by-deployment data files are named as follows:

OS_XXX_YYY_Z_PDF.nc
where: OS - stands for OceanSITES
                XXX - platform code
                YYY - deployment code (unique code for deployment - date or number)
                Z - data mode (e.g. "R" or "D" for real-time or delayed-mode data)
                PDF - provider-defined field, used for identification of data subsets

Additional subdirectories on the FTP server, and naming conventions for aggregated data, interpolated data, and derived data products are specified in the OceanSITES Data Format Reference Manual.

Example: all data from the Bermuda Atlantic Time Series site are contained in the folder "BATS" at either GDAC:
          ftp://ftp.ifremer.fr/ifremer/oceansites/DATA/BATS/
          ftp://data.ndbc.noaa.gov/data/oceansites/DATA/BATS/

A particular data file in these folders is OS_BATS-1_BATSCR-0068-1_D_BTL.nc, which indicates that it was collected from a platform called "BATS-1" with a deployment (mission) identifier "BATSCR-0068-1", providing delayed-mode data ("D").

5. Data discovery

OceanSITES data files should be completely self-describing and easily discovered in multiple ways, from using the GDACs' index files, to THREDDS and DODS servers, to an interactive map at http://dods.ndbc.noaa.gov/oceansites/.

Data discovery via indexing

To allow for data discovery without downloading the data files themselves, an 'index file' is created by each of the GDACs. The index file is a comma-separated-values text file named oceansites_index.txt, inthe root (OceanSITES) directory of each GDAC. It contains a list of the files on the server, and metadata extracted from those files. The file contains a header section, lines of which start with # characters, the list of all data files available on the GDAC, and their descriptions. Each line contains the following information:

- file: the file name, beginning from the GDAC root directory
- date_update: the update date of the file, YYYY-MM-DDTHH:MI:SSZ
- start_date: first date for observations, YYYY-MM-DDTHH:MI:SSZ
- end_date: last date for observations, YYYY-MM-DDTHH:MI:SSZ
- southern_most_latitude, northern_most_latitude, western_most_longitude, eastern_most_longitude, geospatial_vertical_min, geospatial_vertical_min
- update_interval: M monthly, D daily, Y yearly, V void
- size: the size of the file in megabytes
- gdac_creation_date: date of creation of the file on the GDAC
- gdac_update_date: date of update of the file on the GDAC.
- data_mode: R, P, D, M (real-time, provisional, delayed mode, mixed; see Data Format Reference Manual, (table 5)
- parameters: blank-separated list of parameters available in the file.

6. OceanSITES NetCDF Data Files

OceanSITES uses NetCDF (Network Common Data Form) files. These are 'self-describing' files that contain both data and metadata. OceanSITES implements existing standards such as NUG (NetCDF Users Guide), CF (Climate and Forecast), and ACDD (Attribute Convention for Data Discovery). These standards are extended in some areas to maximize data usefulness and to permit smooth operation of the GDACs.

For example, oxygen concentrations from a moored sensor are accompanied by details of the sensor, its calibration, data quality flags, and other information.

The **OceanSITES Data Format Reference Manual** provides details of the conventions used to construct the OceanSITES NetCDF files. The fullest information about the data in each file is contained within it, and must be extracted.

Many software packages are available to extract data and metadata from NetCDF files within programming environments (e.g. Matlab and Python), or to view contents using Internet browsers. Online resources to learn about NetCDF files, and obtain tools to extract them are available from many sites, including:

http://www.unidata.ucar.edu/software/netcdf/software.html#nctoolbox
This site provides an extensive list of NetCDF access tools

http://www.epic.noaa.gov/java/ncBrowse/
this site allows contents to be viewed in a browser

7. Links

OceanSITES:
Home: http://www.oceansites.org
Data Format Reference Manual:
http://www.oceansites.org/docs/oceansites_data_format_reference_manual.pdf
Index file: ftp://data.ndbc.noaa.gov/data/oceansites/oceansites_index.txt

External:
NetCDF: http://www.unidata.ucar.edu/software/netcdf/
CF: http://cfconventions.org/index.html
ACDD: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery
CLIVAR: http://www.clivar.org/

8. Frequently Asked Questions (FAQ)

Data Access and GDAC (Global Data Access Center):	
Does it matter which GDAC FTP server I use?	No. The data holdings are synchronized between the two servers at least daily. The recommendation is that you use the one that delivers data fastest for you (usually the one closer to you). If you ever notice that data files are not identical on the two servers, even after 24 hours have passed, please contact the GDACs and let them know!
Data Content:	
There are data values in the file that make no sense (outliers etc.)! What should I do?	1. Read the "units" attribute for the data variable in the NetCDF file. Are the data presented in the units you are expecting? 2. Read the quality control information in the metadata of the NetCDF file. Are there QC flags in the file that you should be using to discard flagged data, or is the entire dataset listed as problematic or having no QC applied? 3. Check if there are attributes to the data variable in the NetCDF file called "add_offset" and "scale_factor". Sometimes, these are used to pack the data inside the files. Some, but not all, NetCDF software will automatically apply these when reading the files. Do the data make sense when you multiply them with that factor and add that offset manually? 4. Read all metadata attributes. If this does not explain the situation, feel free to contact the person(s) listed in the global attributes. Try to direct your request to a more technical or a more scientific person, based on the situation, if multiple contacts are listed.
What units can the data possibly be in?	There should be an attribute called "units" for each variable inside the NetCDF file. This shows the situation for a given file. Originally, OceanSITES had a specific unit for every parameter. However, now all units are allowed which the UDUNITS software can convert to the "canonical units" that the CF metadata conventions list for the "standard_name" attributed to the variable.
The data variables are presented as multi-(2-, 3-, 4-)dimensional arrays. How do I know which dimension is which?	Read the metadata in the NetCDF file. Each dimension of a variable has a name that is listed, which should answer that question. In addition, check if the variable has an attribute called "coordinates", which may have additional information. If you really care about the details, read the documentation about "discrete sampling geometries" and the attribute "featureType" in the CF metadata conventions.
What does OceanSITES do about data provenance? How do I know which version of a dataset I am using?	The OceanSITES data holdings at the GDACs only include the most current datasets. Older versions are not retained. If you might require access to the exact same file later, be sure to keep your own copy. That said, there are global attributes ("history", "date_created") inside the files that may hold useful information about file versions. In the future, OceanSITES data holdings will likely be preserved as e.g. quarterly "snapshots" in a permanent archive, but this is not yet operational (as of this edition of the manual).
MatLab-Specific:	
How do I read OceanSITES NetCDF files into MatLab?	Current MatLab versions have NetCDF capability built in. The "ncdisp" command displays the content of a file on the screen, including the

	metadata attributes. Use "ncread" and "ncreadatt" to read variables and attributes into the MatLab work space. The regular MatLab documentation ("help …") shows the correct syntax. For older MatLab versions, you have to install external toolboxes to access NetCDF files. A simple internet search should show several available options.
I have loaded data into MatLab, but later routines return error messages when I operate them on those numbers! What should I do?	Some NetCDF files contain data other than double-precision floating-point numbers (e.g. integers or single-precision to save disk space). Use MatLab's "double" command to convert them to double-precision, which is usually what other operations expect.
Data Use:	
Am I required to acknowledge OceanSITES data in my publications? Am I even allowed to use the data for my specific purpose?	If nothing else, etiquette requires you to always give proper credit to OceanSITES and the data generators. To see what you are required to do for a particular OceanSITES data file, read the metadata fields inside that file (esp. attributes "license", "citation", and "acknowledgement"). The general guidance is that OceanSITES data are freely available, and any exceptions or requirements beyond that should be listed in the aforementioned metadata.
What sort of guarantee can OceanSITES make about how good their data are?	OceanSITES and its data providers make no warranties of any kind about their data, express or implied. It is solely the users' responsibility to assess whether the data are appropriate for their particular use.
What is the primary purpose of OceanSITES data anyways?	This is listed in the official mission statement of OceanSITES. Bottom line is that the data are usually meant for oceanographic research in academia.