NODC's Automated Archive and Access Strategy For GHRSST Sea Surface Temperature Data at The GHRSST Long Term Stewardship and Reanalysis Facility

Kenneth S. Casey, Sheri A. Phillips, and John Relph
Last updated: October 02, 2008

Introduction

This document describes the approach that has been developed to manage data acquired by the NOAA National Oceanographic Data Center (NODC) on a daily basis as part of the Group for High Resolution SST (GHRSST) Program [note: formerly known as the Global Ocean Data Assimilation Experiment (GODAE) High Resolution Sea Surface Temperature Pilot Project (GHRSST)]. These data will be automatically archived and made accessible at NODC through its GHRSST Long Term Stewardship and Reanalysis Facility (LTSRF). This automated procedure is made possible largely by the formalized procedures developed within the GHRSST, especially those pertaining to data management, metadata, and file naming conventions. The basic strategy is to bring in data files and corresponding FGDC metadata in XML format on a daily basis, then use a combination of static information, information from the XML filenames, information from within the XML metadata records, and information generated by the NODC archive system to automatically create an Accession Tracking Data Base (ATDB) entry and Archival Information Package (AIP, also known as an NODC "accession") for each logical grouping of data (there may be several each day) and then move the relevant data files (there may be several per AIP) and FGDC record into the NODC archive file systems. On a routine basis, a user-friendly http/ftp/OPeNDAP hierarchy will then be constructed of symbolic links pointing to the data and metadata residing in the formal archive file system.

The GHRSST data streams consist of several types of data and metadata. Three of them will be handled by the procedures outlined here. Two other types, High Resolution Diagnostic Data Sets (HR-DDS) and Matchup Database (MDB) records, are expected to be handled similarly and will be addressed in an update to this document later in FY07. The three types of data managed using the procedures documented here make up the core of the GHRSST system. These data types, in netCDF format, are:

1. Level 2 Preprocessed (L2P): Individual satellite sensor observations processed into geophysical units (SST in degrees C) and formatted according to the GHRSST Data Processing Specifications (GDS v1.5, GHRSST/17).
2. Gridded Level 2 Preprocessed (L2P_GRIDDED_xx): Like other L2P products, but these are for individual satellite sensor observations that have been mapped onto a regular grid with a spatial resolution in kilometers specified by the "_xx".
3. Level 4 Analyzed (L4): These are uniformly gridded and gap-free products created by optimally merging multiple L2P inputs.

Both L2P data types and L4 data are stored in netCDF format following the procedures detailed in the GDS.

Overview of Relevant GHRSST Data Management Structure

The details of the overall GHRSST data management structure are discussed in other GRHSST documentation and will not be addressed here. However, it is important to note a few key elements of the GHRSST system:

1. Data processing is managed by a distributed system with international partners at Regional Data Assembly Centers (RDACs) around the world.
2. The RDACs generate data and GCMD-style Document Interchange Format (DIF) metadata for L2P and L2P_GRIDDED data streams as well as L4 products, along with MDB and HR-DDS information.
3. All RDAC data streams are sent first to the Global Data Assembly Center (GDAC) at the NASA JPL/Caltech Physical Oceanography Distributed Active Archive Center (PO.DAAC). The PO.DAAC, in addition to numerous critical data management functions, provides user access to these real-time data, holds them in a 30-day rolling store, and constructs an XML-formatted FGDC-compliant metadata record for each logical grouping of those data (an AIP) based on the GCMD-DIF style metadata mandated as part of the GRHSST system.
4. The PO.DAAC GDAC system makes these data available to the GHRSST LTSRF at NODC 30 days after receipt at the GDAC.

Overview of NODC Data Management Steps

At this point the GHRSST data and associated metadata are ready for acquisition by NODC. An overview of the automated procedures is provided here, followed by a detailed step-by-step procedure and relevant information tables.

1. The PO.DAAC makes available at ftp://melia.jpl.nasa.gov (username: ghrsst-nodc) an index file called "index.txt" which lists the files over 30 days past date of observation along with their MD5 checksums.
2. NODC accesses this file once per day and uses it determine which files are to be acquired by an FTP-pull. Those files are transferred and checksums generated to verify the transmissions. File transferred include the individual netCDF data files, corresponding GCMD-style DIF formatted XML metadata File Records (FRs) mandated by the GHRSST system, and a single baseline FGDC record for each AIP created by the GDAC.
3. Based on the filenames of the XML-formatted FGDC records, information contained within the FGDC and FR records, some static parameters, and information determined by the NODC archive system, an ATDB entry is created along with an accession number (an NODC unique reference number for the AIP) and directory in the archive file system. Additionally, the baseline FGDC record provided by the GDAC is augmented and all changes noted in the "journal.txt" file. The FGDC record and associated data files are moved into the AIP's accession directory tree within the archive files system. The individual FRs are also placed into the accession directory under a subdirectory called "FileSpecificMetadata". If the NODC system determines that any of these files are part of an existing AIP, then they are moved into that existing directory structure as a new version. This step is detailed in the following section, "Automated ATDB Entry and Accession Assignment".

4. On a daily basis, the NODC will run through the NODC GHRSST AIPs and build a user-friendly directory structure for http/ftp/OPeNDAP data access, based on symbolic links that point into the archive directory structures.
5. Throughout this process, automated checks are performed and failures reported via email messages to Kenneth.Casey@noaa.gov, Sheri.Phillips@noaa.gov, and John.Relph@noaa.gov. These automated tests are listed in the "Automated Tests and Alerts" section.
6. Each day, the automated routines also update an RSS feed which is published to http://www.nodc.noaa.gov/SatelliteData/ghrsst/LTSRF_OpStatus.xml. This feed indicates the number of new Archival Information Packages that have been accepted into the archive that day. The automated routine also updates a small red/yellow/green stoplight graphic to indicate overall GDAC-LTSRF operating status.

Automated ATDB Entry and Accession Assignment

The algorithm for automatically creating an AIP's ATDB entry and Accession Number based on a given FGDC metadata record is documented in this section. The pieces of information needed to fill out the optional and required ATDB elements come from four sources: information in the FGDC XML record filename, information within the FGDC XML record and FRs, static information, and information provided by the NODC archive system. The logical grouping of metadata and data files into a single NODC AIP is based on specific GHRSST datasets, which are defined by Processing Level (L2P, L2P_GRIDDED, or L4), Instrument/Platform, and RDAC. For example, European RDAC (EUR) L2P AVHRR-LAC data from NOAA-17 would be placed in a different AIP than either BlueLink L2P AVHRR-LAC data from NOAA-17 or EUR L2P AATSR on ENVISAT. Each AIP represents one day of data from one particular type of GHRSST dataset.

Filename convention for XML-Formatted FGDC-compliant GHRSST Metadata:

L2P and L2P_GRIDDED:
FGDC-<Date Valid>-<L0_ID>-<Processing_Centre_Code>-<L2_product_code>-<Processing_Model_ID>.xml

L4:
FGDC-<Date Valid>-<Processing_Centre_Code>-<L4_product_code>-<area>-<Processing_Model_ID>-<Identifying_Characteristic>.xml

Where:

<L2_product_code> = L2P | L2P_GRIDDED_xx (where _xx can vary and refers to the spatial resolution in kilometers of the gridded data).

<L4_product_code>= SSTFND | SSTFNDUHR | L4UHFnd | L4LR1m | L4LRblend

<area> = area code for L4 products as specified in the GDS. Currently in use include GLOB, MED, AUS

<Identifying_Characteristic> = a characteristic code that helps distinguish different datasets when one RDAC produces more than one analysis product. For example, NCDC's daily L4 products use "AVHRR_OI" and "AMSR_AVHRR_OI".

An example for the FGDC metadata record describing L2P files:

FGDC-20060621-AVHRR17_L-EUR-L2P-v01.xml

And, an example for an L4 metadata record would be:

FGDC-20070519-UKMO-L4HRfnd-GLOB-v01-OSTIA.xml

Codes for the GHRSST defined <Procesing_Centre_Code> and their translations into NODC ATDB Institution codes are provided below in *Table 1: GHRSST-NODC Institution Conversions*. Codes for converting between the GHRSST defined <L0_ID> and NODC ATDB Instrument codes and Platform codes are provided below in *Table 2: GHRSST-NODC Platform and Instrument Conversions*.

Detailed Algorithm for Automated AIP ATDB Entry and Accession Assignment

1. Set ATDB elements using FGDC record filename:
    a. Use <Date Valid> to create ATDB "start date" and "end date". Since each NODC GHRSST AIP is based on data from a single day, these two elements should be equal.
    b. Take <Processing_Centre_Code> and convert it using *Table 1: GHRSST-NODC Institution Conversions* to create ATDB "Collecting Institution"
    c. Next, use the processing level:
        i. If L4, set ATDB "Instrument" to "Satellite sensor – General [36]" and ATDB "Platform" to "Satellite [3811]" [note: eventually it may be possible to retrieve specific platforms and sensors used by extracting from L4 metadata]
        ii. Else if L2P, use <L0_ID> and convert it using *Table 2: GHRSST-NODC Platform and Instrument Conversions* to create ATDB "Platform" and ATDB "Instrument"
2. Set ATDB elements using FGDC XML or FR XML tags:
    a. Search through all the FR XML bounding coordinate tags, finding the maxima and minima, to create ATDB bounding coordinates. These also replace the baseline FGDC bounding coordinates, which are based on general ranges and not the specific set of files in the AIP.
    b. Use FGDC XML title tag to create ATDB “Title”. Be sure to append the NODC accession number to this title as well as the date as established under NODC procedures.
    c. Use FGDC XML Supplemental Info tag to create ATDB "Supplemental Info"
3. Set Static ATDB elements:
    a. ATDB “Submitted By” = “Edward Armstrong [2255]”
    b. ATDB "Contains Data Types" = "Sea Surface Temperature [319]"
    c. ATDB "Contains Observation Types" = "Satellite Data [20]"
    d. ATDB "Includes Sea Areas" = "World-Wide Distribution [179]"
    e. ATDB "Submitting Institution" = "JPL PODAAC [1222]"
    f. ATDB "Availability Date" = Leave this field blank
    g. ATDB "Number of Observations" = "Unknown" – this is the only ATDB element that is not defined
    h. ATDB "NODC Contact" = "Dr. Kenneth Casey [2224]"
    i. ATDB "Contributing Projects" = "GHRSST [385]"
    j. ATDB "Requested Action" = "close"
    k. ATDB "Disposition" = "online"
4. Set NODC system-determined ATDB and corresponding FGDC elements:

   a. ATDB "Size in Mbytes" and FGDC "Transfer Size" [The volume element is left blank now since this element is supposed to be removed from the ATDB list.]
   b. ATDB "Incoming Directory and File"
   c. ATDB "Date Received" and FGDC "Metadata Date"
5. The above steps should yield a complete ATDB entry and a more robust FGDC record. Compare this ATDB entry to previous ATDB entries. If this AIP already exists, then move the data and metadata into it as a new version under *n*-version where *n* is one greater than the previous version available. If this is a new AIP, create a new accession number, append it to the title in both the FGDC record and the ATDB entry, create the accession directory structure and move the data and associated metadata into it:
   a. Determine which data files belong to the accession by pulling their file names out of the Entities and Attributes section of the FGDC record, where each data file is listed as a separate Entity.
   b. Move those files into 01-version/data/0-data of the appropriate AIP. Also, move to this location any browse images that accompany the data files
   c. Move the FGDC XML-formatted record into the 01-version/data/0-data directory
   d. Move the FR XML-formatted records into the 01-version/data/0-data/FR directory
   e. Move any other information (logs, anomaly reports, etc.) into the appropriate 01-version/about directory

Automated Tests and Alerts

During the management of GHRSST data, several automated procedures must be conducted. Upon failure of any of these, an email alert must be automatically sent out to Kenneth.Casey@noaa.gov, Sheri.Phillips@noaa.gov, and John.Relph@noaa.gov. While that alert is being addressed, the relevant data and metadata must remain in the temporary holding area where the data are held after ftp from the PO.DAAC. These tests include:

1. Check FGDC-compliance of metadata records using *mp* software.
2. Check build of ATDB entries. If, for example, any of the conversions from GHRSST <Processing_Centre_Code> to NODC Institution code fails due to no corresponding NODC Institution in the ATDB list, then a new list entry would be made manually as needed.
3. Others as needed.

Automated User-Friendly FTP/HTTP/OPeNDAP Directory Creation and Update

On a routine basis, a user-friendly ftp/http/OPeNDAP directory hierarchy must be created/updated to link to any new or updated AIPs. Only the most recent version of any given AIP should be included in this hierarchy, which will be created using symbolic links pointing to the archive file systems. The structure should follow this sequence, from top down:

Processing Level => <L0_ID | area> => Processing_Centre_Code => <Identifying_Characteristic> => Year (YYYY) => Day of Year (DDD)

Where:

Processing Level = L2P | L4 | L2P_GRIDDED
For L2P and L2P_GRIDDED: L0_ID = the platform/instrument code, see Table 2 below

For L4: area code as specified in the GDS. Currently in use include:
GLOB, MED, AUS, and NSEABALTIC
Processing_Centre_Code = the code for the producing center, see Table 1 below
<Identifying_Characteristic> = for L4 only (see description of FGDC filenames above)

For example, L2P/AMSRE/EUR/2005/107 would contain L2P AMSRE data produced by the European RDAC for day 107 of 2005. L4/GLOB/UKMO/OSTIA/2007/139 would contain L4 OSTIA data from the UK Met Office RDAC for day 139 of 2007.

Note that when the product code = L2P_GRIDDED_xx (where "xx" refers to the grid resolution in kilometers), the "_xx" is dropped from the ftp/http/OPeNDAP directory structure and the Processing Level represented simply as "L2P_GRIDDED". For all L4 product codes (e.g., "L4UHFnd") the directory hierarchy will use "L4" as the processing level.

Other Possible Steps

At this point, GHRSST data is flowing in to the NODC archives as formally archived data and available online through the NODC Ocean Archive System. It is also available through more intuitive ftp/http/OPeNDAP directory paths. These steps satisfy NODC's commitment to the international GHRSST. However, other steps could be implemented in the future to more fully optimize the GHRSST archive at NODC.

One step would be to develop consistent browse imagery for all incoming GHRSST data. Some RDACs might be providing browse graphics, but they are not mandated to do so. Experience with AVHRR Pathfinder reveals that browse graphics are often the most downloaded parts of the archive by users.

Another step would be to implement a granule-level metadata database containing information about the actual SST values contained within the data files. Calculations like mean, minimum, maximum, number of observations, and standard deviations can be made on each incoming data file and stored as a searchable database capable of being graphically displayed. A test of this type of system has been implemented with Ted Haberman of NGDC and the AVHRR Pathfinder archive.

Table 1: GHRSST-NODC Institution Conversions				
GHRSST Processing Center Code	**NODC ATDB Institute ID**	**NODC ATDB Acronym**	**NODC ATDB Institution Name**	**GHRSST Data Centre Name (for the GDS)**
EUR	1238	MEDSPIRATION	Medspiration	European RDAC
USGODAE	1236	NRL-MRY	US Navy; Naval Research Laboratory – Monterey, CA	US-GODAE
REMSS	1226	REMSS	Remote Sensing Systems	Remote Sensing Systems, CA, USA
RSMAS	717	RSMAS	University of Miami; Rosenstiel School of Marine and Atmospheric Science	University of Miami, RSMAS
JPL	1222	JPL PODAAC	US National Aeronautic And Space Administration; Jet Propulsion Laboratory Physical Oceanography Distributed Active Archive Center	JPL Physical Oceanography DAAC
OSISAF	1232	OSI-SAF	Ocean and Sea Ice Satellite Application Facility	EUMETSAT Ocean and Sea Ice Satellite Applications Facility
ABOM	86	ABOM	Australian Bureau of Meteorology	Australian RDAC
MEDS	953	MEDS	Marine Environmental Data Service	MEDS Data Centre, Ontario, Canada
UKMO	1188	UKMO	British Meteorological Office	UK Meteorological Office
NOCS	1170	NOCS	National Oceanography Centre, Southampton	National Oceanography Centre, Southampton
MERSEA	1233	MERSEA	Marine Environment and Security for the European Area	Marine Environment and Security for the European Area
NAVO	267	NAVOCEANO	US Navy; Naval Oceanographic Office	Naval Oceanographic Office
NODC	295	NODC	US DOC; NOAA; NESDIS; National Oceanographic Data Center	NOAA National Oceanographic Data Center
NCDC	1258	NCDC	US DOC; NOAA; NESDIS; National Climatic Data Center	NOAA National Climatic Data Center
OSDPD	1259	OSDPD	US DOC; NOAA; NESDIS; Office of Satellite Data Processing and Distribution	NOAA Office of Satellite Data Processing and Distribution
GOS	1266	GOS	Gruppo Di Oceanografia Da Satellite; Istituto Di Scienze Dell Atmosfera E Del Clima - Rome [Satellite Oceanography Group; Institute Of Atmospheric And Climate Sciences]	Gruppo di Oceanografia da Satellite
DMI	1358	DMI	Danish Meteorological Institute	Danish Meteorological Institute
NEODAAS	1392	NEODAAS	Natural Environment Research Council Earth Observation Data Acquisition and Analysis Service	NERC Observation Data Acquisition and Analysis Service
NCOF	1235	NCOF	National Centre for Ocean Forecasting	Added for possible future use, but use UKMO now
IFREMER	806	IFREMER	Institut Francais de Recherche pour L'exploitation de la Mer	Here as parent institution only
EUMETSAT	1231	EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites	Here as parent institution only

ESA	1237	ESA	European Space Agency	Here as parent institution only
JAP			Add this institution in future updates	Japanese RDAC
SEASNET			Add this institution in future updates	SEASnet Tropical coverage RDAC
JAXA			Add this institution in future updates	National Space Development Agency (of Japan)
TOHOKU			Add this institution in future updates	University of Tohoku, Japan
NOAA		**NOAA**	**Add this in update. Currently no NOAA (only, lots of NOAA offices) in ATDB**	**National Oceanic and Atmospheric Administration**

Note: Items in grey are not currently expected to appear in GHRSST filenames as Processing Center Codes. Some of these were required additions to the NODC parameter tables since they serve as parent institutions and others may appear in the future.

Table 2: GHRSST-NODC Platform and Instrument Conversions					
GHRSST <L0_ID>	**NODC Platform Acronym**	**NODC Platform ATDB ID**	**NODC Instrument Acronym**	**NODC Instrument ATDB ID**	**NODC ATDB Description**
NAR18_SST	NOAA18	**10633**	AVHRR-HRPT	**126**	Advanced Very High Resolution Radiometer High Resolution Picture Transmission
NAR17_SST	NOAA17	**10614**	AVHRR-HRPT	**126**	Advanced Very High Resolution Radiometer High Resolution Picture Transmission
NAR16_SST	NOAA16	**10468**	AVHRR-HRPT	**126**	Advanced Very High Resolution Radiometer High Resolution Picture Transmission
ATS_NR__2P[1]	Envisat	**10630**	AATSR-NR	**129**	Advanced Along Track Scanning Radiometer Near Real time
ATS_MET_2P	Envisat	**10630**	AATSR-MET	**130**	Advanced Along Track Scanning Radiometer real time METeorological
AVHRR16_G	NOAA16	**10468**	AVHRR-GAC	**121**	Advanced Very High Resolution Radiometer Global Area Coverage
AVHRR16_L	NOAA16	**10468**	AVHRR-LAC	**120**	Advanced Very High Resolution Radiometer Local Area Coverage
AVHRR17_G	NOAA17	**10614**	AVHRR-GAC	**121**	Advanced Very High Resolution Radiometer Global Area Coverage
AVHRR17_L	NOAA17	**10614**	AVHRR-LAC	**120**	Advanced Very High Resolution Radiometer Local Area Coverage
AVHRR18_G	NOAA18	**10633**	AVHRR-GAC	**121**	Advanced Very High Resolution Radiometer Global Area Coverage
AVHRR18_L	NOAA18	**10633**	AVHRR-LAC	**120**	Advanced Very High Resolution Radiometer Local Area Coverage
SEVIRI_SST	MSG	**10637**	SEVIRI	**98**	Spinning Enhanced Visible and Infra-Red Imager
GOES12[2]	GOES-12	**10616**	GOES Imager	**99**	Geostationary Operational Environmental Satellite Imager
GOES11	GOES-11	**10826**	GOES Imager	**99**	Geostationary Operational Environmental Satellite Imager
GOES10[3]	GOES-10	**10615**	GOES Imager	**99**	Geostationary Operational Environmental Satellite Imager
GOES9	GOES-9	**10624**	GOES Imager	**99**	Geostationary Operational Environmental Satellite Imager
AMSRE	Aqua	**10617**	AMSR-E	**100**	Advanced Microwave Scanning Radiometer-EOS
TMI	TRMM	**10620**	TMI	**101**	Tropical Rainfall Measuring Mission (TRMM) Microwave Imager
TMI_VIRS	TRMM	**10620**	VIRS	**102**	Visible and Infrared Scanner
MODIS_A	Aqua	**10617**	MODIS	**103**	Moderate Resolution Imaging Spectroradiometer
MODIS_T	Terra	**10622**	MODIS	**103**	Moderate Resolution Imaging Spectroradiometer
AVHRRMTA_G	MetOp-A	**10949**	AVHRR-GAC	**121**	Advanced Very High Resolution Radiometer Global Area Coverage
WindSAT	Coriolis	**10623**	WindSAT	**104**	Multi-frequency polarimetric microwave radiometer on Coriolis
MTSAT_1R	MTSAT-1R	**10626**	MTSAT Imager	**127**	Multi-functional Transport Satellite Imager
AIRS	Aqua	**10617**	AIRS	**107**	Atmospheric Infrared Sounder

[1] Note this GHRSST acronym, ATSR_NR__2P contains TWO underscore characters between the NR and 2P, not just one.

[2] Note that this L0_ID may include the additional characters “SAF”. See GDS Table A3.2.1 for an explanation of the differences.
[3] Note that this L0_ID may include the additional characters “NAV”. See GDS Table A3.2.1 for an explanation of the differences