

# Data Reorganization for Optimal Time Series Data Access, Analysis, and Visualization

Hualan Rui<sup>1,2</sup>, William L Teng<sup>1,3</sup>, Richard Strub<sup>1,2</sup>, and Bruce Vollmer<sup>1</sup>

<sup>1</sup>Goddard Earth Sciences Data and Information Services Center (GES DISC), NASA, Greenbelt, MD 20771 USA

<sup>2</sup>ADNET Systems, Inc., Rockville, MD 20852, USA

<sup>3</sup>Wyle Information Systems, Inc., McLean, VA 22102, USA

Hydrology Data and Information Services Center (HDISC)

Data and Information Services Center (DISC)

NASA Goddard Earth Sciences (GES)

Help Desk: [gsfc-help-disc@lists.nasa.gov](mailto:gsfc-help-disc@lists.nasa.gov)

AGU Fall December 2012 (IN23B-1500)

Email: [Hualan.Rui@nasa.gov](mailto:Hualan.Rui@nasa.gov)

## Introduction

The way data are archived is often not optimal for their access by many user communities (e.g., hydrological), particularly if the data volumes and/or number of data files are large. The number of data records of a non-static data set generally increases with time. Therefore, most data sets are commonly archived by time steps, one step per file, often containing multiple variables. However, many research and application efforts need time series data for a given geographical location or area, i.e., a data organization that is orthogonal to the way the data are archived. The retrieval of a time series of the entire temporal coverage of a data set for a single variable at a single data point, in an optimal way, is an important and longstanding challenge, especially for large science data sets (i.e., with volumes greater than 100 GB). The approach that we took to meet this challenge is summarized in this poster. This work is part of a larger project to enhance access to and use of NASA and other data by the hydrological community (See poster IN31A-1494).

## Large Science Data Sets at NASA GES DISC

### Two examples:

- North American Land Data Assimilation System (NLDAS, Phases 1 and 2)
- Global Land Data Assimilation System (GLDAS, Versions 1 and 2)

For NLDAS and GLDAS data access, please see poster H21F-1239, "Analysis of Water and Energy Budgets and Trends Using the NLDAS Monthly Data Products."

Data Sets	Temporal		Spatial		Dimension	Total Grids	# Files per Data Set	Total Volume
	Resolution	Coverage	Resolution	Coverage				
NLDAS	hourly	1979-present	0.125x0.125	N. America	224x464	103936	289872	~ 4.8 TB
GLDAS	3-hourly	1948-present	0.25x0.25	Global	600x1440	864000	96360	~ 1.6 TB

### The "Digital Divide" problem

- The data are archived at NASA GES DISC in the form of all variables one time step per file.
- Users often need long time series for single variables at single grid points, so they would have to get all the files of the data set (> 289,872 files) and then process through all the data (> 1 TB) for parameter and spatial subsetting.
- This is the "Digital Divide," described by Maidment et al. (2010).

### Bridging the "Digital Divide"

- NASA GES DISC has a long history of continual efforts to bridge the gap between NASA data and end user communities.
  - Giovanni online visualization and analysis system that provides Time Series plot and ASCII outputs without users needing to download the entire data set.
  - Mirador subsetting service that provides parameter & spatial subsetted files on-the-fly.
  - GrADS Data Server (GDS) that provides parameter and spatial subsetting service and outputs data in binary and ASCII (good for short time ranges).
- However, all these methods are still accessing the data in a way that is orthogonal to how the data are archived, so they are not effective and very costly.
- Two successful prototypes of integrating NASA GES DISC data into end user community tools and environment:
  - Integrating NLDAS precipitation data into EPA BASINS (via partially reorganized data)
  - Integrating NASA GES DISC hydrologic data into CUAHSI-HIS (via GDS)
- Optimal reorganization of NASA GES DISC data for access and use by hydrologic user community.

## Optimal Reorganization of Large Data Sets

Although GES DISC data, as is, can be and are served as time series, they are limited to relatively short time periods. For time series of long durations, with current technology, only a reorganization of the data can enable the serving of time series, on the fly, with satisfactory performance.

Figure 2. Converting from time-stepped continuous spatial fields to single "point" time series ("data rods").

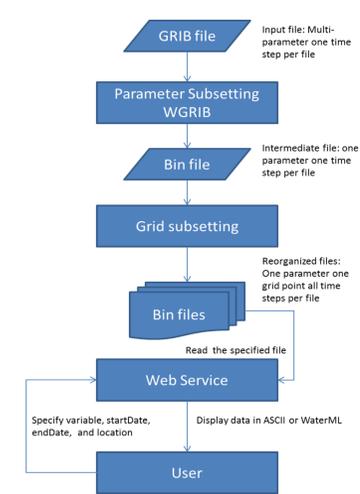
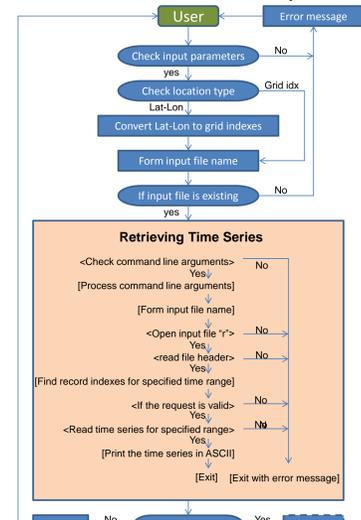


Figure 3. A web Service that serves single point time series in ASCII and time series plot.



### Directory Structure and file name convention for reorganized time series files ("data rods")

One directory per parameter  
One file per grid point (time series)

#### Directory Name Convention:

prod\_name.vid.param\_short\_name

#### File Name Convention:

prod\_name.vid.param\_short\_name.Y###-X###.bin

#### Example for NLDAS Precipitation (APCpsfc):

Directory Name:  
NLDAS.FORA012\_H.002.APCpsfc

File Names:  
NLDAS.FORA012\_H.002.APCpsfc.Y000-X001.bin  
NLDAS.FORA012\_H.002.APCpsfc.Y000-X002.bin  
.....  
NLDAS.FORA012\_H.002.APCpsfc.Y080-X300.bin  
.....  
NLDAS.FORA012\_H.002.APCpsfc.Y223-X463.bin

### Time Series File Structure

**File header (300 bytes)**

```

char prod_name[50];
char param_short_name[20];
char param_name[50];
char unit[20];
int grid_y, grid_x;
char begin_time[14];
char end_time[14];
unsigned int tot_record;
float min,max,mean,std;
int dt;
int ydim, xdim;
float start_lat, start_lon;
float dat, dlon;
char last_upd[20];
char spare[60];

```

**Data records**  
N = tot\_record  
written in the file header

```

float data1
float data2
.....
float dataN

```

- Time Series files are written in direct access binary file.
- Direct access makes the processing and reprocessing simple, fast, and less expensive.
- Direct access makes the time series temporally searchable and easy to retrieve.
- The Grid indexes in the file names "Y###-X###" make the time series spatially searchable.

## Application

### Texas Natural Resources Information System (TNRIS) Drought monitoring

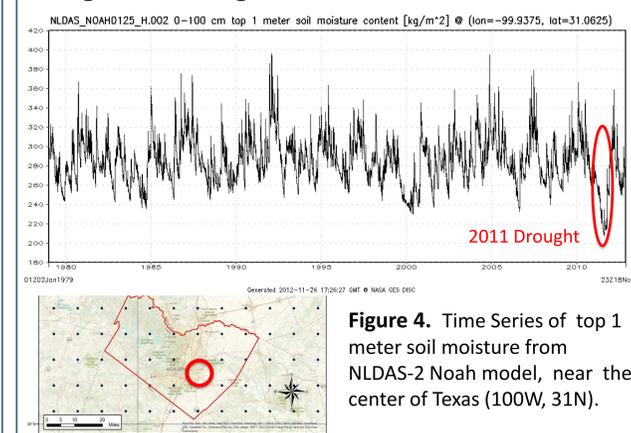


Figure 4. Time Series of top 1 meter soil moisture from NLDAS-2 Noah model, near the center of Texas (100W, 31N).

### Hourly precipitation from Hurricane Sandy in New Jersey

### Case study for extreme weather event

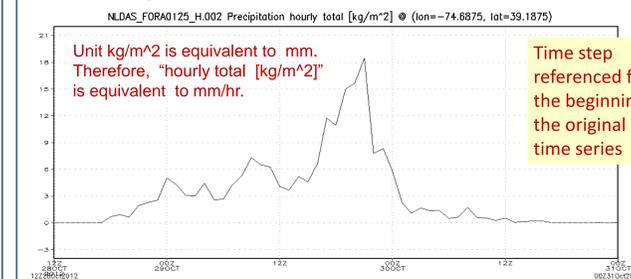


Figure 5. Time Series of hourly precipitation from NLDAS-2 Noah model, showing the passage of Hurricane Sandy over a location in New Jersey (74.7W, 39.2N).

Table 2. NLDAS parameters with time series access available.

Short name	Parameter Name	Unit	Begin Date
APCpsfc	Precipitation hourly total	kg/m <sup>2</sup>	1979-01-01T13
TMP2m	2-m above ground temperature	K	1979-01-01T13
UGRD10m	10-m above ground zonal wind speed	m/s	1979-01-01T13
VGRD10m	10-m above ground meridional wind speed	m/s	1979-01-01T13
PEVAPsfc	Potential evaporation	w/m <sup>2</sup>	1979-01-01T13
SOILM0-100cm	0-100 cm top 1 meter soil moisture content	kg/m <sup>2</sup>	1079-01-02T01
TSOIL0-10cm	0-10 cm soil temperature	K	1079-01-02T01
SSRUNsfc	Surface runoff (non-infiltrating)	kg/m <sup>2</sup>	1079-01-02T01
EVPSfc	Total evapotranspiration	kg/m <sup>2</sup>	1079-01-02T01

Metadata of the Time Series file:  
 prod\_name=NLDAS\_FORA012\_H.002  
 param\_short\_name=APCpsfc  
 param\_name=Precipitation hourly total  
 unit=kg/m<sup>2</sup>  
 begin\_time=1979/01/01/13  
 end\_time=2012/11/19/12  
 start\_lat=31.1975  
 start\_lon=-99.9375  
 xdim=1  
 ydim=1  
 tot\_record=297024

Metadata for the original time series

Metadata for the requested time series

Time step referenced from the beginning of the requested time series

Both Figs. 4 and 5 are generated by the NASA GES DISC Web Service.

## Time Series Data Access

- Integrating time series data into Hydrology community tools, CUASHI HIS and EPA BASINS

See IN31A-1494, Bridging the Digital Divide between Discrete and Continuous Space-Time Array Data to Enhance Accessibility to and Usability of NASA Earth Sciences Data for the Hydrological Community.

- Web Service:** A simple Web Service at NASA GES DISC serves time series data in:
  - ASCII format, along with complete metadata
  - Time series plot ( add "&type=plot" to the sample URLs )

Sample URLs for accessing a time series of a point location near central Texas (100W, 31N)

**Specify location by grid point indexes (ASCII)**  
[http://hydro1.sci.gsfc.nasa.gov/daac-bin/access/timeseries.cgi?variable=NLDAS:NLDAS\\_NOAH0125\\_H.002:SOILM0-100cm&startDate=1979-01-02T01&endDate=2012-11-18T23&location=NLDAS:X200-Y048](http://hydro1.sci.gsfc.nasa.gov/daac-bin/access/timeseries.cgi?variable=NLDAS:NLDAS_NOAH0125_H.002:SOILM0-100cm&startDate=1979-01-02T01&endDate=2012-11-18T23&location=NLDAS:X200-Y048)

**Specify location by longitude and latitude (plot)**  
[http://hydro1.sci.gsfc.nasa.gov/daac-bin/access/timeseries.cgi?variable=NLDAS:NLDAS\\_NOAH0125\\_H.002:SOILM0-100cm&startDate=1979-01-02T01&endDate=2012-11-18T23&location=GEOM:POINT\(-99.9375,31.0625\)&type=plot](http://hydro1.sci.gsfc.nasa.gov/daac-bin/access/timeseries.cgi?variable=NLDAS:NLDAS_NOAH0125_H.002:SOILM0-100cm&startDate=1979-01-02T01&endDate=2012-11-18T23&location=GEOM:POINT(-99.9375,31.0625)&type=plot)

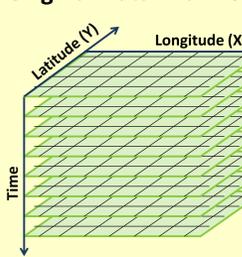
- Direct ftp** [ftp://hydro1.sci.gsfc.nasa.gov/data/Time\\_Series/](ftp://hydro1.sci.gsfc.nasa.gov/data/Time_Series/)

## Summary

- The way data are archived is often not optimal for their access by many user communities (e.g., hydrological), particularly if the data volumes and/or number of data files are large. This is the "Digital Divide," described by Maidment et al. (2010).
- NLDAS and GLDAS data at NASA GES DISC, with volumes in the multi-TB per data set, are two examples of these large data sets. NASA GES DISC has a long history of continual efforts to bridge the "Digital Divide" between NASA data and end user communities.
- For time series of long durations, with current technology, only a reorganization of the data can enable the serving of time series, on the fly, with satisfactory performance.
- A set of NLDAS parameters (Table 2) has been re-organized and archived as time series files in direct access binary format. More parameters will be made available later.
- The time series data have been integrated into CUASHI HIS and EPA BASINS.
- A Web Service has been implemented to serve the time series data in ASCII and as time series plots. The time series data can also be accessed via direct ftp.

**Acknowledgment:** This work is supported by NASA ROSES NNH11ZDA001N-ACCESS.

### Original Data Archive



### Reorganized Data Archive

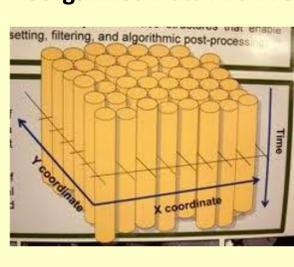


Figure 1. A schematic diagram for data reorganization for optimal time series access.