Terra Level 1B Data Fusion Product

MuQun Yang[1], Landon Clipp[2], Hyo-Kyung Lee[1], Guangyu Zhao[2], and Larry Di Girolamo[2]
[1]The HDF Group [2]Dept. of Atmospheric Sciences and NCSA, Univ. of Illinois at Urbana-Champaign

What is Terra Data Fusion?

Overview

Goals

- Develop a system to efficiently generate mission-scale Terra data fusion products.
- Facilitate the use of Terra data fusion products.

Achievements

- Merging L1B radiance fields from all Terra five instruments
- Unpacking radiation data to physical units; and providing geolocation information at native radiance resolutions
- Generating metadata for each Terra orbit
- 2.4 petabytes fusion data generated by supercomputers at NCSA
- HDF5 format, netCDF4-CF compatible
- Tool to resample/reproject fusion data

Problems

- Large input datasets (~1 PB)
- Input data residing at different locations, in different file formats (HDF 4 & 5), with different granularities
- Inadequate cyberinfrastructure to tackle whole-mission data fusion and mining problems
- Difficult to transfer and manage input data
- Requiring a range of expertise from data scientists to software engineers
- HPC, Data format / interoperability / etc.
- Difficult to use by larger user communities
- Different methods to store radiance and geo-location data for each instrument
- Lack of common metadata conventions

These issues greatly limit and discourage scientists to utilize and analyze the whole Terra data set effectively.

Solutions

- Utilize expertise and computing infrastructure from different institutions to efficiently generate and deliver Terra data fusion products.
- Fuse radiance data from all instruments.
- Map granules from different instruments to a common granule that contains data for a single Terra orbit.
- Retrieve radiance data that correspond to physical units; data can be ready to use.
- Basic Fusion granules are HDF5 files converted from HDF4, HDF-EOS2, and HDF-EOS5 input files. They are netCDF-4 compatible and follow the CF conventions.

Basic Fusion (BF) – MODIS, MOPITT, CERES, ASTER, MISR

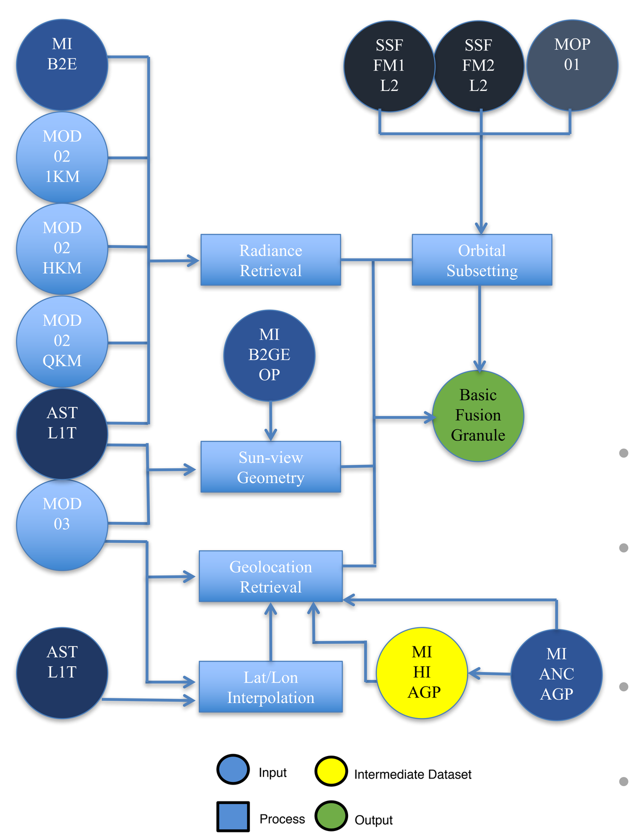
BF Granularity

- For the Basic Fusion product, **a granularity of one Terra orbit** was chosen.
- A Terra orbit represents a regular and repeating path that the satellite takes around the Earth and is nearly constant over the Terra mission.
- One granule ranges 10GB-70GB, ideal both for IO performance, processing speed, memory usage, and massive data transfer across both disk and tape systems.

BF File Organization

- The input granules are included if the start time is within the start/end time of the Terra orbit with two exceptions: MOPITT and CERES data are subsetted and then included if *any* radiance data falls within this orbit.
- Both radiance and geolocation data are included along with sun-view geometry and quality information.
- The BF files were generated in a format that follows CF conventions, such that they can be accessed by netCDF-4 enhanced data models.

BF Data Generation



Included Products in NASA CMR* Short Name and Version Notation	
MOPITT	MOP01_007
CERES	CER_SSF_Terra-FM1-MODIS_Edition4A_4A CER_SSF_Terra-FM2-MODIS_Edition4A_4A
ASTER	AST_L1T_003
MISR	MI1B2E_003, MIB2GEOP_002, MIANCAGP_1, MIHIAGP_1**
MODIS	MOD021KM_6, MOD02HKM_6, MOD02QKM_6, MOD03_6

*CMR: Common Metadata Repository
**275m resolution MISR geolocation fields produced by the Terra fusion project team; unavailable at NASA data centers.

- HDF5 chunking and compression are used to reduce the total BF file size.
- For MODIS, ASTER, and MISR, data converted from native integer (packed) to floating point (unpacked). Floating point data correspond to physical units.
- For MODIS, MISR, and ASTER, geolocation data are interpolated to native radiance resolution and merged into a BF file.
- MOPITT and CERES data are subsetted to span a Terra orbit.

BF Data Transfer

Instrument	Rate (MB/s)	Size (TB)	DAAC
MISR	160-210	240	ASDC
MODIS	0-100	331	LAADS
ASTER	0-20	400	LPDAAC
CERES	160-210	1.5	ASDC
MOPITT	160-210	0.8	ASDC

- Worked with NASA networking (CNOC), ASDC and LP DAAC teams to diagnose and remedy performance obstacles for data transfer out of the ASDC.
- Result a 2.5x increase in data transfer rates out of the ASDC

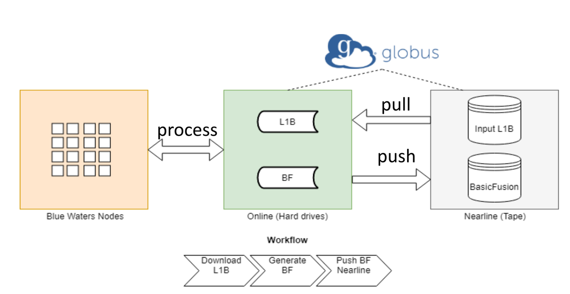
Lessons Learned

- Mission scale data transfers require active support from the DAACs.
- GridFTP provides much more efficient and reliable protocol than ftp and http for bulkd data transfer
- Managing 3+ petabytes of data requires systems specifically designed for big data processing.

BF Storage and Compute Infrastructure

- Both input and output data are staged on NCSA nearline tape archive system of the BlueWaters
- Input data moved to scratch storage via Globus when generating Basic Fusion files
- NCSA Blue Waters supercomputer used to generate BF data
- TORQUE job and resource management used

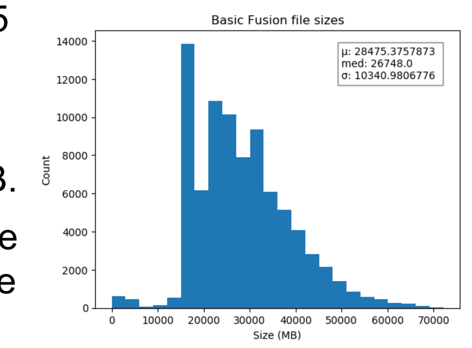
BF Mission Processing Workflow



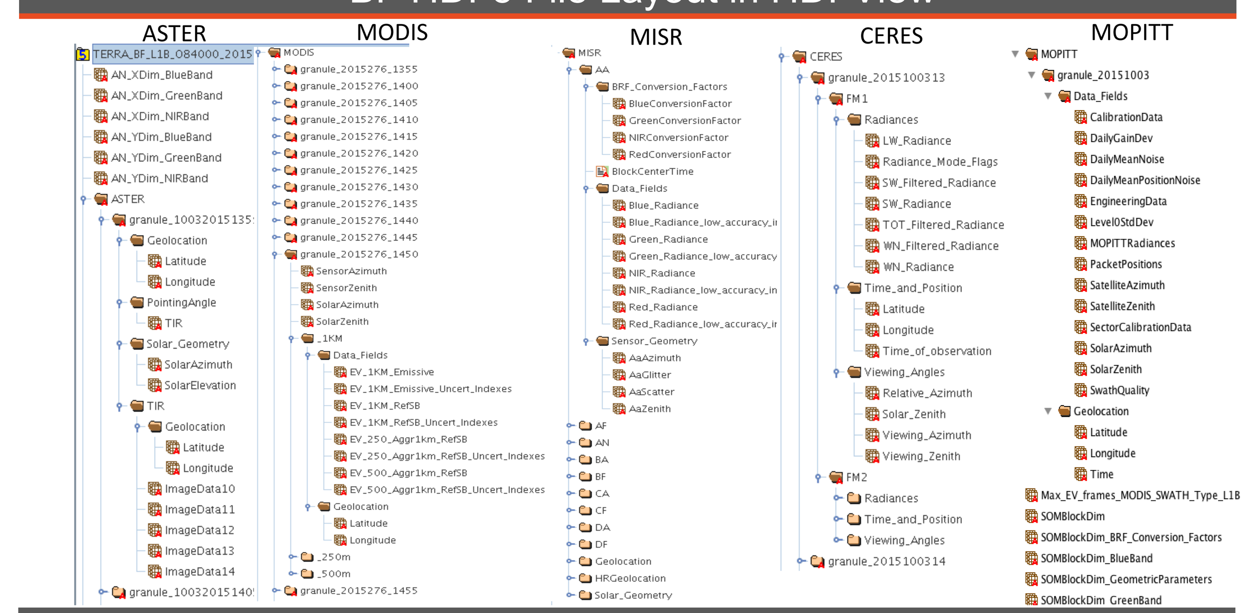
- 130,000 core CPU hours are used for actual mission processing.
- One granule takes 1-2 hours to produce.
- C is used for Basic Fusion software and Python for workflow.
- Granules are produced in batches of 2,000 to 3,000.
- Smaller NCSA cluster (ROGER) was used during code development and testing.
- SQLite database was used to query input granules by orbit.

BF File Statistics

- 84,303 files - from 02/25/2000 to 12/31/2015
- The total file size is 2.4 petabytes.
- Typical file sizes 15GB – 40GB. The largest file size is 68.7GB. Average file size is 26GB.
- In-memory compression reduces the total file size by 60%. Without using compression, the total size will be about 5.5~6 petabytes.
- The biggest BF file takes about 3 hours to finish at NCSA Blue Waters.



BF HDF5 File Layout in HDFView

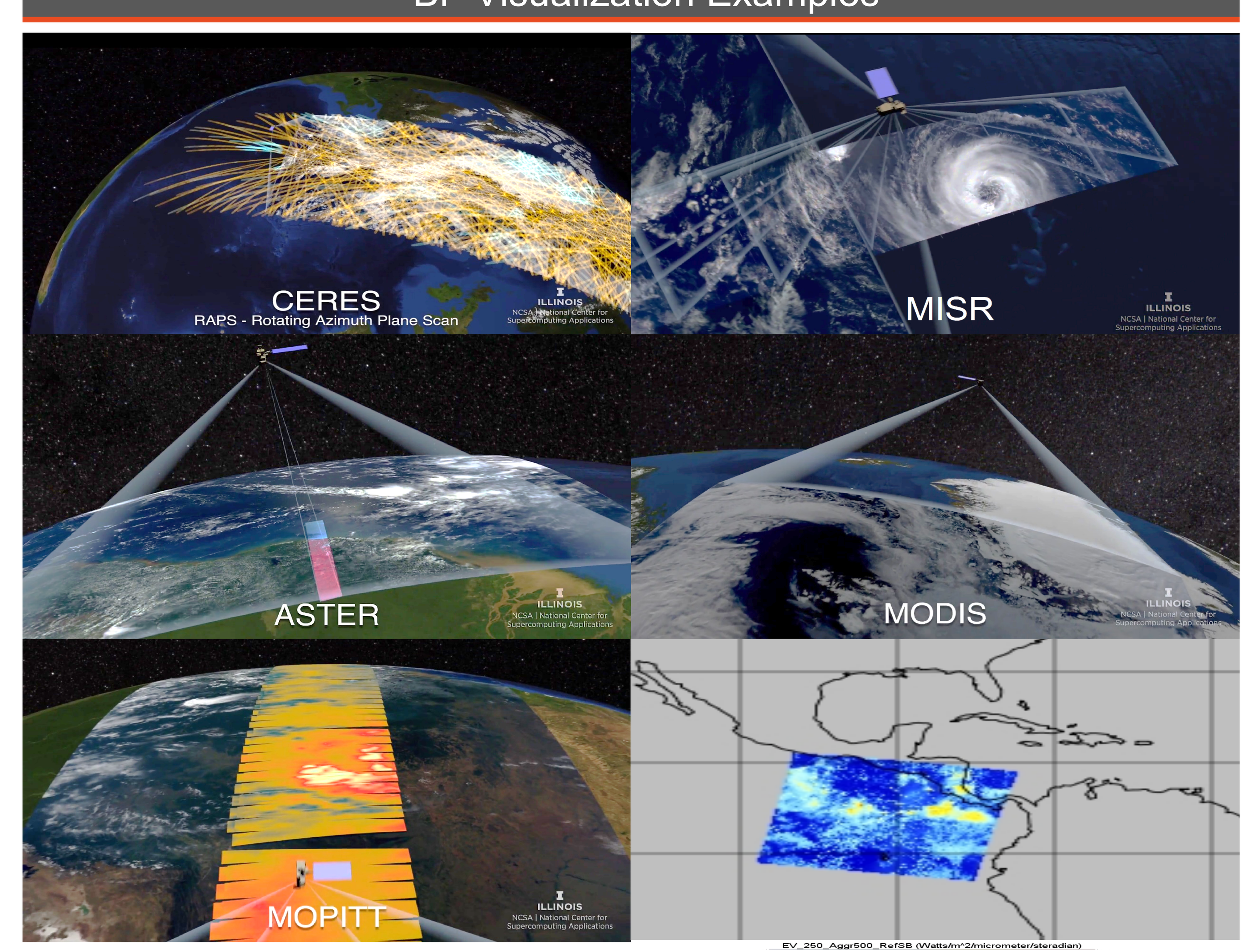


BF HDF5 File Layout in CDL

```
group: ASTER {
  group: TIR {
    variables:
      float ImageData10(ImageLine_TIR_Swath_g10222015151007,
ImagePixel_TIR_Swath_g10222015151007) ;
          ImageData10:units = "Watts/m^2/micrometer/steradian" ;
          ImageData10:_FillValue = -999.f ;
          ImageData10:valid_min = 0.f ;
          ImageData10:valid_max = 569.f ;
          ImageData10:coordinates =
"/ASTER/granule_10222015151007/TIR/Geolocation/Longitude
/ASTER/granule_10222015151007/TIR/Geolocation/Latitude" ;
  }
}
group: MISR {
  group: Data_Fields {
    variables:
      float Blue_Radiance(SOMBlockDim_BlueBand, XDim_BlueBand, YDim_BlueBand) ;
          Blue_Radiance:coordinates = "/MISR/Geolocation/GeoLongitude
/MISR/Geolocation/GeoLatitude" ;
          Blue_Radiance:units = "Watts/m^2/micrometer/steradian" ;
          Blue_Radiance:_FillValue = -999.f ;
          Blue_Radiance:valid_min = 0.f ;
          Blue_Radiance:valid_max = 800.f ;
  }
}
group: MODIS {
  group: _1KM {
    group: Data_Fields {
      variables:
          float EV_1KM_Emissive(Band_1KM_Emissive_MODIS_SWATH_Type_L1B,
_10_nscans_MODIS_SWATH_Type_L1B, Max_EV_frames_MODIS_SWATH_Type_L1B) ;
            EV_1KM_Emissive:valid_min = 0.f ;
            EV_1KM_Emissive:units = "Watts/m^2/micrometer/steradian" ;
            EV_1KM_Emissive:coordinates =
"/MODIS/granule_2015295_1425/_1KM/Geolocation/Longitude
/MODIS/granule_2015295_1425/_1KM/Geolocation/Latitude
Band_1KM_Emissive_MODIS_SWATH_Type_L1B" ;
            string EV_1KM_Emissive:band_names =
"20,21,22,23,24,25,27,28,29,30,31,32,33,34,35,36" ;
          float
EV_1KM_Emissive_Uncert_Indexes(Band_1KM_Emissive_MODIS_SWATH_Type_L1B,
_10_nscans_MODIS_SWATH_Type_L1B, Max_EV_frames_MODIS_SWATH_Type_L1B) ;
      } // group Data_Fields
  }
}

group: CERES {
  group: granule_2015102214 {
    group: FM1 {
      group: Radiances {
        variables:
        float LW_Radiance(Footprints_FM1_g2015102214) ;
            LW_Radiance:_FillValue = 3.402823e+38f ;
            LW_Radiance:valid_range = 0.f, 200.f ;
            LW_Radiance:units = "Watts per square meter per steradian" ;
      }
      group: Time_and_Position {
       variables:
          float Latitude(Footprints_FM1_g2015102214) ;
            Latitude:units = "degrees_north" ;
          float Longitude(Footprints_FM1_g2015102214) ;
            Longitude:units = "degrees_east" ;
        double Time_of_observation(Footprints_FM1_g2015102214) ;
            Time_of_observation:units = "day" ;
      }
    }
}
group: MOPITT {
  group: granule_20151022 {
    group: Data_Fields {
      variables:
        float MOPITTRadiances(ntrack_1, nstare, npixels, nchan, nstate) ;
          MOPITTRadiances:coordinates =
"/MOPITT/granule_20151022/Geolocation/Latitude
/MOPITT/granule_20151022/Geolocation/Longitude" ;
          MOPITTRadiances:_FillValue = -9999.f ;
          string MOPITTRadiances:units = "Watts/m^2/steradian" ;
      }
group: Geolocation {
      variables:
        float Latitude(ntrack_1, nstare, npixels) ;
            Latitude:units = "degrees_north" ;
        float Longitude(ntrack_1, nstare, npixels) ;
            Longitude:units = "degrees_east" ;
        double Time(ntrack_1) ;
            string Time:units = "seconds since 1993-01-01" ;
      } // group Geolocation
    }
```

BF Visualization Examples

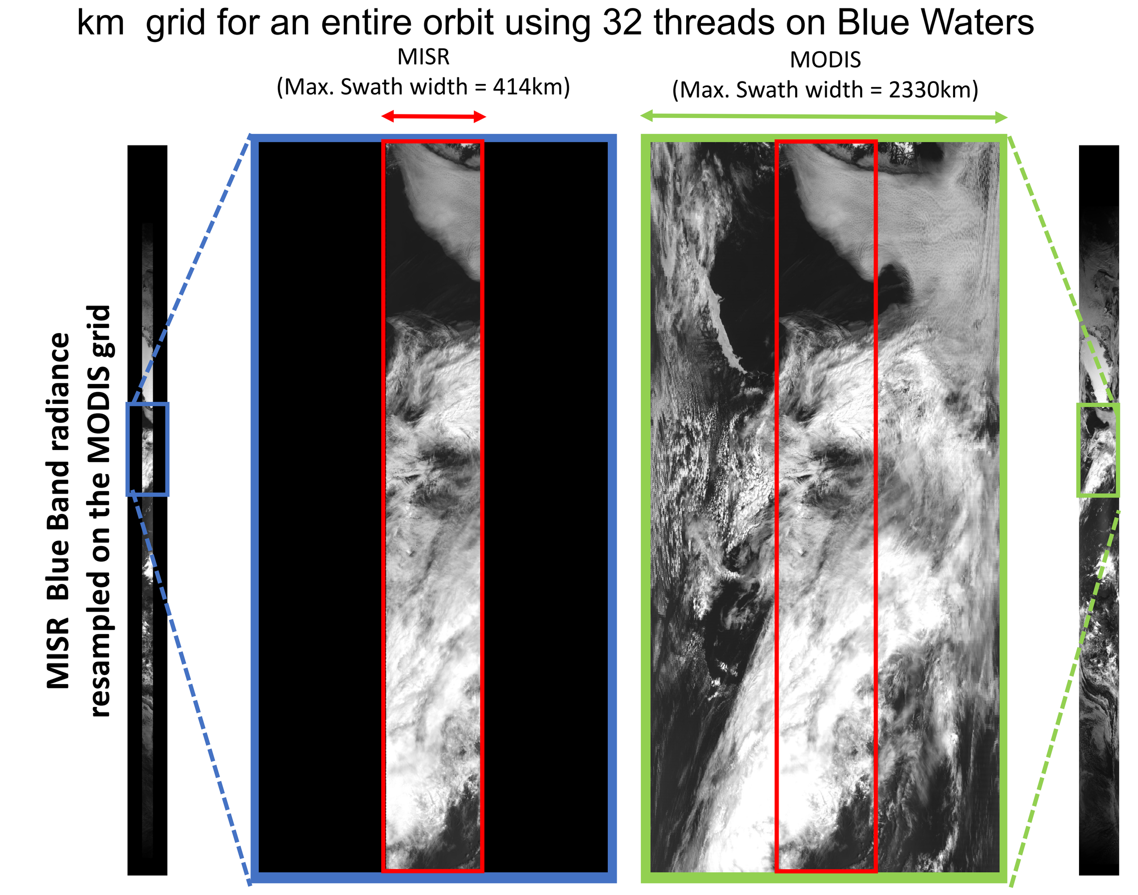


https://youtu.be/C2uyjRGwwOs
Credit: Advanced Visualization Lab, NCSA
Panoply

Advanced Fusion (AF)

AF Resampling and Reprojection Tool (in development)

- A community, open source tool
  - Resample / reproject the radiance fields for one Terra instrument onto the grid used by another Terra instrument, a regularly spaced grid, or any other user specified grid.
- Two major resampled strategies implemented (Nearest neighbor resampling & Statistical resampling)
- User-specified map projections and features
  - Regularly spaced grids defined by GDAL-supported map projections
  - MISR block unstacking
  - Common swath selection
- HDF5 and GeoTIFF output
- 40 seconds to resample MISR 1.1km radiance data to MODIS 1.0 km grid for an entire orbit using 32 threads on Blue Waters



TERRA_BF_L1B_O69626_20130119123228_F000_V000.h5

Other Related Work

Metadata Generation

- Collection and granule metadata in ECHO10 XML are available.
- Although input data are missing for some orbits, granule metadata are generated to help users locate these orbits through NASA CMR search.
- The metadata standards and format passed the CMR ingestion test

OPeNDAP Web Service

- BF files are tested and can be accessed with the OPeNDAP Hyrax server.
- OPeNDAP subset service may help users to retrieve data easily and efficiently.

Additional Tools

- A Globus Python interface data transfer tool
- A Python PBS workflow manager

Acknowledgements

Department of Atmospheric Sciences, University of Illinois

Larry Di Girolamo, Guangyu Zhao, Yizhe Zhan
Landon Clipp, Yat Long Lo, Dongwei Fu

Department of Geography and GIS, University of Illinois

Shaowen Wang, Yan Liu, Yizhao Gao

The HDF Group

Muqun (Kent) Yang, Hyo-Kyung (Joe) Lee

National Center for Supercomputing Applications

John Towns, Kandace Turner, Michelle Butler
Sean Stevens, David Raila, Jonathan Kim
Donna Cox, Stuart Levey, Robert Patterson
Andrew Christensen, Jay Roloff

This report was based upon work supported by the funding from NASA ACCESS Grant #NNX16AM07A. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NASA.