National Aeronautics and Space Administration

ATMOSPHERIC SCIENCE DATA CENTER

‘Best’ Practices for Aggregating Subset Results from Archived Datasets

Walter E. Baskin[1], Jennifer Perez[2]

(1) Science Systems and Applications, Inc., Hampton, VA, USA. (2) NASA Langley Research Center, Hampton, VA,USA.

Introduction

In response to the exponential growth in science data analysis and visualization capabilities, data centers have been developing new processes to package and deliver large volumes of aggregated subsets of archived data. New standards are evolving to help data providers and application programmers manage the growing needs of the science community. These standards evolve from the best practices gleaned from new products and capabilities. The NASA Atmospheric Sciences Data Center (ASDC) has developed and deployed production provider-specific search and subset web applications for the CALIPSO, CERES, TES, and MOPITT missions.
(URL: https://subset.larc.nasa.gov)

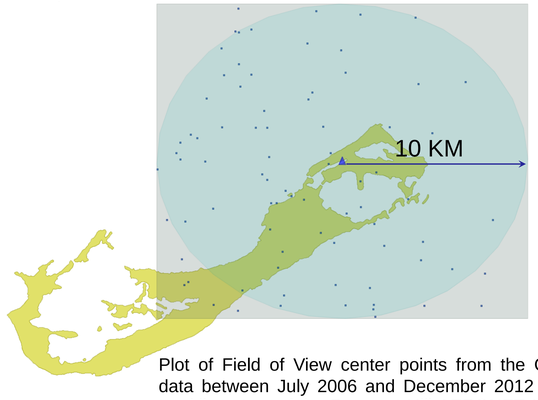
This presentation explores a CERES CCCM (CALIPSO, CloudSat, CERES, MODIS) data validation use case that leverages aggregated subset results from CERES CCCM (Level2), CERES SSF (Level2), and CALIPSO LIDAR (Level1) datasets. Additionally, it examines the standards and formats that ASDC developers have applied to the delivered files as well as the implementation strategies for subsetting and processing the aggregated products.

Key Components

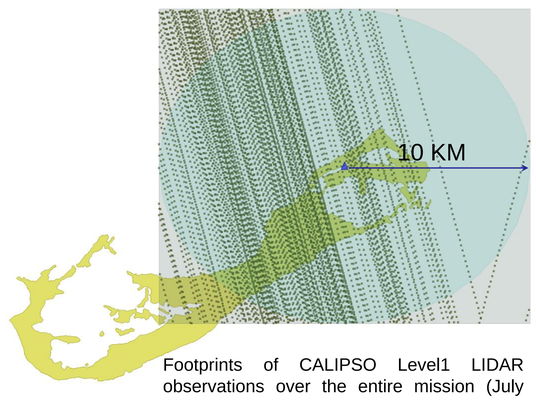
- Application of NetCDF Climate Forecast (CF) conventions to aggregated subsets of archived level 2 satellite data products.
- Data-Provider-Specific format requirements vs. generalized standards.
- Organization of the file structure of aggregated NetCDF subset output.
- Global Attributes of individual subsetted files vs. aggregated results.
- Specific applications and framework used for subsetting and delivering derivative data files.

Example Use Case Scenario:
Compile an Aggregated Subset of A-Train Integrated CALIPSO, CloudSat, CERES, MODIS merged product (CCCM or C3M) Satellite Observations Within a Ten Kilometer Radius of a surface observation site

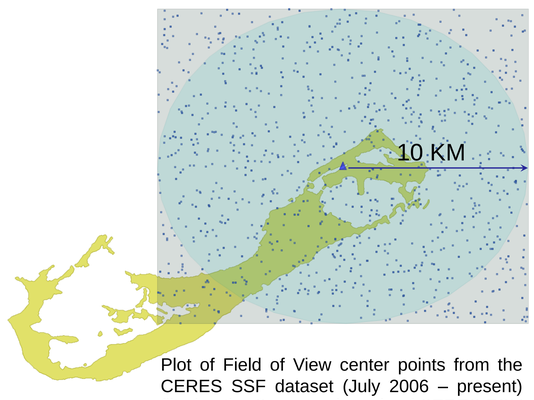
Prospect Hill AERONET Site on Bermuda Island



Plot of Field of View center points from the CCCM data between July 2006 and December 2012 within ten kilometers of the Prospect Hill AERONET Site on Bermuda Island.



Footprints of CALIPSO Level1 LIDAR observations over the entire mission (July 2006 – present) within area of interest.

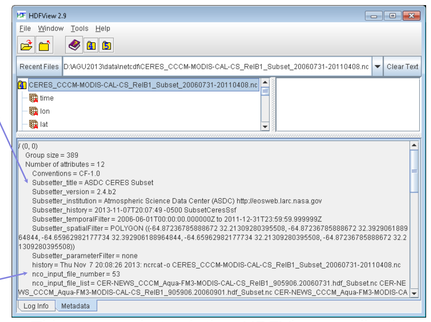


Plot of Field of View center points from the CERES SSF dataset (July 2006 – present) filtering the Cross-track angle of CERES FOV at surface values between -2.03 and 2.03 degrees.

CERES SSF and CALIPSO LIDAR components of the CCCM product were subsetted and aggregated from their original archived datasets. Each of these results are at the full resolution of their respective sensors enabling additional insight into the state of the environment that could not be inferred at a specific CCCM data point.

The ‘Subsetter_’ prefix was added to standardized Climate Forecast (CF) compliant attribute names to distinguish subsetter-specific metadata from metadata inherited from metadata in the original data granules and metadata generated by the aggregator.

Default NCO ncrcat utility attribute population:
- history - append the invocation command
- nco_input_file_number - number of files
nco_input_file_list – names of input subsets



This screenshot of the HDFView application shows the root level attributes of the resulting aggregated subset request result .

Aggregation process for CAL_LID_L1: NCML – Unidata NetCDF ToolsUI


```
<?xml version='1.0' encoding='UTF-8'?>
<netcdf xmlns='http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2'>
<aggregation dimName='time' type='joinExisting'>
<scan location='D:/AGU2013/data/netcdf' suffix='Subset.nc' />
</aggregation>
</netcdf>
```

Aggregation process for CERES SSF and CERES CCCM: NCO ncrcat

```
ls *_Subset.nc > dir_listing_of_netcf.txt
cat dir_listing_of_netcdf.txt | ncrcat –o aggregatedoutput.nc
```

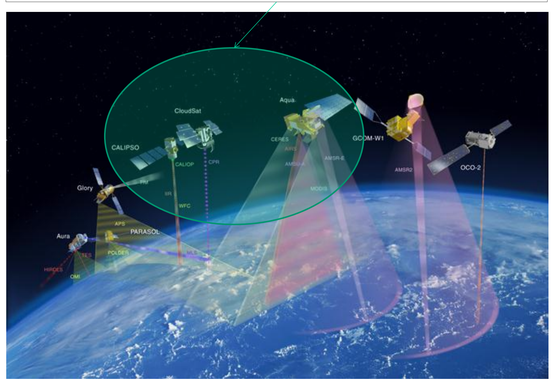
CALIPSO, CloudSat, CERES, MODIS merged product (CCCM)

The CCCM satellite dataset is a product of the NASA CERES Science Team. CCCM data is archived at the ASDC and can be ordered from the Reverb service discovery portal.
(URL: http://reverb.echo.nasa.gov)

The ASDC development team added high resolution spatial metadata harvesting, subsetting, and NetCDF aggregation capabilities for this data product in November 2013. Currently the CERES Science development team expects to expand their web interface to support CCCM dataset discovery and subset ordering within the next year.

(URL: http://ceres.larc.nasa.gov/order_data.php)

Science data parameters in the CCCM dataset are derived from data products from instruments on the CALIPSO, CloudSat, and Aqua satellites.



Artist's Concept of the A-Train constellation of satellites. **Credit:** NASA

Attribute Convention for Data Discovery (ACDD)

The Federation of Earth Science Information Partners (ESIP) Documentation Cluster publishes a wiki that organizes recommended global and variable attributes based on several metadata standards implementations (THREDDS, ISO 19115-2, CF, and OGC CSW). In addition to providing specific real world examples of data products with conformant attribute naming, the attribute descriptions are organized into levels of priority:

- Highly Recommended
- Recommended
- Suggested

This site is a very useful resource for data providers and archive centers developing new data products, or augmenting metadata from legacy archived science datasets.

(URL: http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery_(ACDD))

Conclusion

The capability of Earth Scientists to interactively discover, filter, and aggregate subsetted results from archives of entire satellite missions is a reality.

Science data standards and compliance organizations are better able to define best practices by exploring specific examples of data processing, formatting, and attribution of aggregated dataset results. By embracing and implementing these standards and preserving data provenance, science data producers will simplify access to original observations. This will significantly reduce the complexity and expense of data analysis and visualization, expanding the understanding of Earth’s geophysical processes.