CERES Data Management Activity

Presented to
CERES Science Team

NASA Langley Research Center
November 3, 2009

Jonathan Gleason
jonathan.l.gleason@nasa.gov
Topics to be Covered

- CERES Background/Overview
- CERES Processing Approach
- NPP as it compares to Terra/Aqua
- Terra/Aqua Edition 2 & Edition 3 status
- Development and Production Platforms
- AMI Migration Status
- Production optimization effort
- DMT Operations and Documentation
CERES Statistics
(from the Terra & Aqua Senior Review)

- High level of data fusion
  - 11 Instruments on 7 satellites
- 25 unique input data sources
- 18 CERES data products
- Over 90% of CERES data product data volume involves 2+ instruments
- Individual data products include up to 260 unique parameters
- Approximately 1.7 million lines of QC and validation codes
- Approximately 0.85 million lines of production codes
CERES Top Level Data Flow Diagram

**Instrument**
- CERES Data
  - Geolocate And Calibrate
    - BDS
    - IES
  - Bidirectional Scans (BDS)
    - Erbe Monthly Gridded (ES4/ES4G)
    - Erbe Monthly Regional (ES9)
    - Erbe Footprints (ES8)
  - Erbe Monthly Monthly Averaging
    - Time Space Averaging
    - EDDS
  - Erbe Like
    - Inversion
    - MOA Data

**Empirical Surface Estimates**
- Single Scanner Footprint (SSF)
  - Grid Fluxes 9
    - Empirical Surface Fluxes (SFC)
    - Compute Surface Averages
      - Monthly Surface Averages (SRBAVG)
      - MOA Data

**Clouds and Radiation / Atmosphere and Surface**
- Clouds and Radiative Swath (CRS)
  - Grid Fluxes and Cloud Props 6
    - Grided Fluxes And Cloud Props (FSW)
    - Time Interpolate
      - Intermediate Synoptic Fluxes and Clouds (SYNI)
    - Compute Spatial Averages
      - Average Fluxes And Clouds (AVG,ZAVG)

**Data Product Size**
- Key
  - Algorithm
  - Data Product
  - Ancillary Data Set
  - 1-Hour
  - 6-Hour
  - Daily
  - Monthly

**MOA Data**
- MOA: Meteorological Ozone & Aerosol
CERES Organization

Science
- Derives & refines algorithms
- Validates algorithms
- Validates CERES data sets
- Writes Quality Summary

Data Management Team (DMT)
- Implements algorithms
- Maintains software
- Verifies data
- Assists in validation
- Provides CM and documentation support

Atmospheric Sciences Data Center (ASDC)
- Ingests data
- Places operational software in production
- Produces data sets
- Distributes data sets
- Archives data
- Provides User Services

Algorithm Development - Algorithm Implementation - Data Processing
CERES Subsystems

• CERES is made up of 7 Working Groups
  - Instrument
  - ERBElike
  - Clouds
  - Inversion or ADM
  - SOFA
  - SARB
  - TISA

• Code organized into 12 Subsystems
  - Each subsystem tied to 1 or more working groups

• Each Subsystem made up of 1 or more Product Generation Executives (PGEs)
  - Currently there are about 70 active PGEs
Data from other Instruments used by CERES

• CERES Instrument/ERBElike only subsystems that can process when only CERES data available.

• CERES directly uses the following MODIS data sets:
  • MYD02SS1/MOD02SS1* (19 channel radiance subset of every other pixel every other scanline)
  • MYD03/MOD03* (geolocation)
  • MYD04_L2/MOD04_L2 (5 min 10 km aerosol swath)
  • MYD08_D3/ MOD08_D3 (daily 1 deg aerosol)
     • Critical data sets; must have matched pairs to process.

• Additionally CERES uses, Geostationary satellite data:
  • MET-5, MET-6, MET-7, MET-8, MET-9
  • GOES-8, GOES-9, GOES-10, GOES-11, GOES-12
  • GMS-5, MTSAT-1R
## CERES Processing Software

<table>
<thead>
<tr>
<th>Subsystem Number</th>
<th>Subsystem Name</th>
<th>LOC (to nearest 1K)</th>
<th>Publicly Available Date Products</th>
<th>Product Frequency</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERESlib</td>
<td>115K</td>
<td></td>
<td></td>
<td></td>
<td>All Satellites</td>
</tr>
<tr>
<td>1</td>
<td>Instrument/Pre-Processor</td>
<td>4K</td>
<td></td>
<td></td>
<td>NPP only</td>
</tr>
<tr>
<td>1</td>
<td>Instrument</td>
<td>110K</td>
<td>BDS</td>
<td>1/day</td>
<td>All Satellites</td>
</tr>
<tr>
<td>2</td>
<td>ERBElike/ Inversion</td>
<td>33K</td>
<td>ES-8</td>
<td>1/day</td>
<td>All Satellites</td>
</tr>
<tr>
<td>3</td>
<td>ERBElike/ TSA</td>
<td>16K</td>
<td>ES-9, ES-4</td>
<td>1/month</td>
<td>All Satellites</td>
</tr>
<tr>
<td>12</td>
<td>MOA</td>
<td>10K</td>
<td></td>
<td></td>
<td>Run monthly</td>
</tr>
<tr>
<td>4.1 – 4.4</td>
<td>Clouds</td>
<td>231K</td>
<td></td>
<td></td>
<td>All Satellites</td>
</tr>
<tr>
<td>4.5 – 4.6</td>
<td>Inversion</td>
<td>26K</td>
<td>SSF</td>
<td>1/hour</td>
<td>All Satellites</td>
</tr>
<tr>
<td>5</td>
<td>SARB</td>
<td>51K</td>
<td>CRS</td>
<td>1/hour</td>
<td>All Satellites</td>
</tr>
<tr>
<td>6 &amp; 9</td>
<td>TISA-Gridding</td>
<td>31K</td>
<td>FSW, SFC, ISCCP-D2like-Day/Nit</td>
<td>60/month, 36/month, 1/month</td>
<td>All Satellites</td>
</tr>
<tr>
<td>11</td>
<td>GGEEO</td>
<td>50K</td>
<td>ISCCP-D2like-GEO</td>
<td>1/month</td>
<td>Geostationary</td>
</tr>
<tr>
<td>7.2</td>
<td>Synoptic SARB</td>
<td>10K</td>
<td></td>
<td></td>
<td>All Satellites</td>
</tr>
<tr>
<td>7.1 &amp; 8 10</td>
<td>TISA-Averaging</td>
<td>164K</td>
<td>SYN, AVG, ZAVG, SRBAVG</td>
<td>1/day, 1/month, 1/month 5/month</td>
<td>All Satellites</td>
</tr>
<tr>
<td><strong>TOTAL LOC</strong></td>
<td><strong>851K</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Current CERES Terra/Aqua Processing Approach

CERES processes data 3 times:

<table>
<thead>
<tr>
<th>Baseline1-QC</th>
<th>Edition1-CV</th>
<th>Edition2</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Processed daily</td>
<td>• Processed monthly</td>
<td>• Processed in blocks of 4+ months at a time</td>
</tr>
<tr>
<td>• Run Instrument &amp; ERBElike Inversion subsystems</td>
<td>• Run Instrument and ERBElike subsystems</td>
<td>• Run all CERES subsystems as inputs become available</td>
</tr>
<tr>
<td>• Use Composite Snow Map</td>
<td>• Use actual Snow map and wait for all expected instrument inputs</td>
<td>• All primary archival products made publicly available</td>
</tr>
<tr>
<td>• Not publicly available</td>
<td>• CV stands for “Calibration/Validation”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Primary Instrument &amp; ERBElike products made publicly available</td>
<td></td>
</tr>
</tbody>
</table>
## Edition2 Processing

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Latency</strong></td>
<td>6 – 24 months (Wait for Gains and SRF based on Edition1-CV)</td>
<td>~6 months after SFC (Wait for Cal Coef for geo-sat)</td>
<td>Wait for aerosol inputs (MATCH-like data)</td>
</tr>
<tr>
<td><strong>Processing</strong></td>
<td>Instrument (BDS, IES)<em>&lt;br&gt;ERBElike Inv (ES8)</em>&lt;br&gt;ERBElike TSA (ES4, ES9)&lt;br&gt;MOA (MOA)&lt;br&gt;Clouds (Temp)<em>&lt;br&gt;Inversion (SSF)</em>&lt;br&gt;TISA-gridding (SFC)</td>
<td>G GEO (G GEO)&lt;br&gt;TISA-Averaging (SRBAVG)</td>
<td>SARB (CRS)*&lt;br&gt;TISA-Gridding (FSW)</td>
</tr>
<tr>
<td>* Instantaneous</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CERES Data Flow Diagram
Relevant NPP Issues

• Imager input data required for Climate Data Record (CDR) must be of climate quality and consistently calibrated over entire period.
  – In NPP era, Land PEATE provides CERES aggregated radiance and geolocation files and sub-sampled data files using CERES provided code. Land PEATE also provides AOT files that correspond to sub-sampled radiance/geolocation.
  – For Terra/Aqua, MODAPS provides radiance, geolocation, and aerosol files from a collection that begins at covers open.

• NPP CERES made use of already existing interfaces.
  – Cost savings by using existing infrastructure.
  – Land PEATE already getting VIIRS data. Agreed to also obtain CERES RDRs.
  – Network between Land PEATE and ASDC exists for Terra/Aqua.
<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Parameter Description</th>
<th>Terra/Aqua Freq &amp; Source</th>
<th>NPP Freq &amp; Source</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CERES L0 files</td>
<td>Instrument level 0 data</td>
<td>3/day; EDOS @ GSFC</td>
<td>~131/day; Land PEATE</td>
<td>In case of NPP, RDRs also contains spacecraft diary</td>
</tr>
<tr>
<td>Attitude</td>
<td>Attitude</td>
<td>12/day; GSFC Flight Dyn Facility</td>
<td></td>
<td>included in RDR</td>
</tr>
<tr>
<td>Ephemeris</td>
<td>Ephemeris</td>
<td>12/day (Terra); GSFC Flight Dyn Facility</td>
<td></td>
<td>included in RDR</td>
</tr>
<tr>
<td>Imager Calibrated Data, Instantaneous</td>
<td>Imager Radiances &amp; Geolocation Aerosols</td>
<td>288/day; MODAPS ~144/day; MODAPS</td>
<td>288/day; Land PEATE~144/day; Land PEATE</td>
<td>CERES provided code to sub-sample radiance files at GSFC</td>
</tr>
<tr>
<td>Aerosol data</td>
<td>Aerosol (Coln) Optical thickness, type/size</td>
<td>1/day; MODAPS</td>
<td></td>
<td>For Terra/Aqua using MODIS MOD08 and MATCH. Plan to do same for NPP</td>
</tr>
<tr>
<td>Meteorological and Ozone data</td>
<td>3-D Met Data</td>
<td>4/day; GMAO</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-D atmospheric data</td>
<td>24/day; GMAO</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-D constants</td>
<td>1; GMAO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precipitable Water</td>
<td>2-D constants</td>
<td>2/day; Global Hydrology Resource Center (GHRC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geostationary data</td>
<td>MCIDAS data from 5 geostationary satellites per month</td>
<td>120/day; University of Wisconsin Space Science and Engineering Center (SSEC)</td>
<td>Only every 3rd hour is used for production</td>
<td></td>
</tr>
<tr>
<td>SURFMAP(Snow/Ice)</td>
<td>Snow/Ice Map</td>
<td>4/day; NCEP/NESDIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SURFMAP(Snow/Ice)</td>
<td>Snow/Ice Map</td>
<td>1/day; NSIDC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Shared NPP and Terra/Aqua Data Source
Simplified CERES Processing Flow

Autonomous Subsystems affected by changing NPP inputs:

- CERES RDRs
- SSI Instrument
- IES

Autonomous CERES Subsystems:

- BDS Instantaneous ERBElike
- EID6
- SS3 ERBElike TSA
- ES-4
- ES-9

VIIRS Subsystems affected by NPP input:

- SS4 Compute Clouds, Convolve with CERES, compare TOA/surface fluxes
- SS5 Compute fluxes at levels
- SS6 TISA Gridding
- SS7 & 8 TISA Averaging

Subsystems affected by NPP input:

- SS9 TISA Gridding
- SS10 TISA Averaging
- SRBAVG
- AVG
- ZAVG
- SYN

ES-8
NPP CERES Operational Data Flow

- **C3S**
  - Commands, Loads and requests
  - Mission Notices and Data

- **CERES Instrument Ops Team**
  - Commands, Loads and requests
  - Mission Notices and Data

- **SD3E**
  - RDRs to PST
  - All RDRs

- **SDS**
  - RDRs, SDRs, EDRs, IPS
  - CERES RDRs, VIIRS Sub-samples & aerosols

- **L-PEATE**
  - Data Distribution
  - Data Users

- **ERB CARS**
  - Data and Science Operations

- **C3S**
  - Mission Notices and Data requests

- **IDPS**
  - RDRs to PST
  - All xDRs

- **ADS**
  - Reuse existing systems and interfaces
  - System enhancements for NPP CERES

- **Svalbard**
  - Cmd & Tlm
  - Mission Data

- **Ancillary Data Providers**
  - Existing data, agreements

- **NPOESS Systems**
  - NPP Systems
  - NOAA Systems
  - CERES Systems
FM5 Code Development

• Instrument only subsystem modified
  – Preprocessor will convert data to format consistent with Terra and Aqua
  – Instrument subsystem current implemented with Ada
  – Convert Ada to C++ and deliver preprocessor for Ada as schedule risk mitigation

• Six total code deliveries
  – Deliver Ada FM5 ready December 4 (AMI x86)
  – Deliver C++ FM5 ready March 12 (AMI P6)
## Main Terra and Aqua Edition2 Data Sets

<table>
<thead>
<tr>
<th>Product</th>
<th>Latest Edition</th>
<th>Data available through</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>BDS, ES8, ES4, ES9</td>
<td>Edition2 (T, A)</td>
<td>Dec‘08</td>
<td>Waiting on gains/SRF</td>
</tr>
<tr>
<td>SFC</td>
<td>Edition2F (T) Edition2C (A)</td>
<td>Dec’07</td>
<td></td>
</tr>
<tr>
<td>SRB AVG</td>
<td>Edition2D (T) Edition2A (A)</td>
<td>Oct’05</td>
<td>Waiting on MTSAT coefficients</td>
</tr>
<tr>
<td>FSW</td>
<td>Edition2F (T) Edition2C (A)</td>
<td>Dec’07</td>
<td></td>
</tr>
<tr>
<td>ISCCP-D2like-Day, ISCCP-D2like-Nit</td>
<td>Beta1</td>
<td>Aug’07</td>
<td>Awaiting Edition 2 Code delivery this month</td>
</tr>
<tr>
<td>ISCCP-D2like-Geo</td>
<td>Beta1</td>
<td>Oct’05</td>
<td>Awaiting Edition 2 code delivery this month</td>
</tr>
</tbody>
</table>
Terra and Aqua Edition 3

- Terra Edition 3 instrument ready to process this week
- Aqua Edition 3 data sets are delayed until Gains and Spectral Response Functions (SRF) available
  - Once Science approves Gains & SRF, production expected to begin within 4-6 weeks
  - Other CERES Edition3 data sets will follow as code deliveries are made
- CERES will reprocess and forward process all data
- Product parameters may be added or changed
  - SSF size to increase over 33%! (over 50 parameters added)
- Edition2 will extend until Edition3 catches up
  - Expect letter change in Edition2 data set names starting 2008 due to switch from GEOS-4 to G5-CERES
<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Data Start</th>
<th>Data End</th>
<th>Processing Complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T) Instrument</td>
<td>Mar’00</td>
<td>Dec’04</td>
<td>1/30/2010</td>
</tr>
<tr>
<td>(T) ERBE-like</td>
<td>Mar’00</td>
<td>Dec’04</td>
<td>3/5/2010</td>
</tr>
<tr>
<td>(A) Instrument</td>
<td>June’02</td>
<td>Mar’09</td>
<td>5/21/2010</td>
</tr>
<tr>
<td>(A) ERBE-like</td>
<td>June’02</td>
<td>Mar’09</td>
<td>5/28/2010</td>
</tr>
<tr>
<td>(T) Instrument</td>
<td>Jan’05</td>
<td>Mar’09</td>
<td>3/26/2010</td>
</tr>
<tr>
<td>(T) ERBE-like</td>
<td>Jan’05</td>
<td>Mar’09</td>
<td>3/26/2010</td>
</tr>
</tbody>
</table>
Development and Production Platforms

- Most efficient to develop code on machine that is identical to production platform
  - Same environment including Toolkit, operating system, compiler, HDF, and library versions
  - Science approves for delivery to production platform newly developed code by examining associated data runs
    - Science shouldn’t have to repeat exercise on production platform
    - Unless production environment itself introduced a change, data management can quickly compare files created in development and production environments

- Access to production output products from development machine improves efficiency
  - No need to order data, no duplicate copies of products
  - Faster to evaluate and use data sets
Lessons Learned When Delivering Code and Testing

• Data Management personnel create expected output on target production machine
  – Verify that expected output looks as expected on target machine

• CM untars delivery and compiles source code on target machine, runs to reproduce expected output
  – Verify that delivery tar file includes all necessary components prior to turning delivery over to ASDC for testing

• Once delivery is in production do not immediately begin running an Edition data set
  – Run ValRx for all instrument/input combinations
    • Ensure production environment not altering output
    • Ensure correct files were delivered
    • Ensure scripts set up correctly
Platform Migration

• Codes typically must be modified to work on new platform
  – Extensive updates may be needed
  – Takes time, may not be highest priority
  – Currently migrating last 4 CERES subsystems off SGI and onto IBM cluster

• Because CERES produces Climate Data Records, must verify that output is scientifically equivalent regardless of production platform
  – Can’t upgrade algorithms as part of migration
# AMI Transition

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>Edition</th>
<th>Target Delivery to AMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERBE-like</td>
<td>Ed2 &amp; Ed3</td>
<td>6/25/2010</td>
</tr>
<tr>
<td>Clouds Convolution</td>
<td>Ed3</td>
<td>6/11/2010</td>
</tr>
<tr>
<td>Inversion SOFA</td>
<td>Ed2</td>
<td>8/20/2010</td>
</tr>
<tr>
<td>Instantaneous SARB</td>
<td>Ed2</td>
<td>3/5/2010</td>
</tr>
<tr>
<td>Synoptic SARB</td>
<td>Ed2</td>
<td>5/14/2010</td>
</tr>
<tr>
<td>TISA G GEO</td>
<td>Ed3</td>
<td>4/9/2010</td>
</tr>
<tr>
<td>TISA SRBAVG, AVG, ZAVG, SYN</td>
<td>Ed2</td>
<td>4/30/2010</td>
</tr>
<tr>
<td>TISA SFC &amp; FSW</td>
<td>Ed2</td>
<td>3/19/2010</td>
</tr>
<tr>
<td>MOA</td>
<td>Ed3</td>
<td>12/14/09</td>
</tr>
<tr>
<td>PMOA</td>
<td>Ed3</td>
<td>2/19/2010</td>
</tr>
</tbody>
</table>
Production Processing Optimization

- Increased capacity with AMI environment: Utilize with optimized code and production scheduling
- Code optimization
  - Plan to work with ODU beginning in November to explore parallelization techniques to implement within source code.
  - TISA software is first test case (AVG, ZAVG, SYN)
  - Working Groups will decide maintainability of suggested source code changes
Production Processing Optimization

- Production requests, job submittal, production monitoring and reporting currently independent
- Migrate PRs to relational database
  - Dependencies for PGE represented
- Develop tools to retrieve and display production status
  - Currently manually search archive for output
- Ideally store production status in database for PR generation
- Maximize scheduling with Sun Grid Engine
CERES uses CMMI Approved Processes

- The CERES DM task was successfully appraised at CMMI Maturity Level 2. All individual Process Areas were appraised at Capability Level 3.

- CERES DM task CMMI processes are described in 9 process plans (http://science.larc.nasa.gov/ceres/DMP_Plans/index.html)
  - Data Management Process Plans
  - Configuration Management Plan
  - Data Management Plan – DRAFT
  - Measurement and Analysis Plan
  - Process and Product Quality Assurance Plan
  - Requirements Management Plan
  - Risk Management Plan
  - Software Development Plan
  - Software Management Plan
  - Training Management Plan

- During the recent (Aug 12th) ISO9100 internal process audit of the (entire) Science Directorate, CERES was audited as a representative project. There were NO non-conformances identified.
Documentation Overview
(http://eosweb.larc.nasa.gov/PRODOCS/ceres/table_ceres.html)

• Data Quality Summaries
  – Detailed information about a particular data set
  – Living Document; most up-to-date
  – Always consult Data Quality Summary prior to using data or publishing research

• Data Products Catalog
  – Parameter lists for each data product
  – Version of pages that apply to data set included with order

• Collection Guides
  – User Guide for data product

• Description/Abstract
  – Record of differences between data sets and configuration codes
Questions about Data Sets??

- Look over Data Products Catalog pages
- Reread Data Quality Summary
- Consult Collection Guide, if available
- Specific science questions may be sent to Contact Scientist listed in Section 2.2 of Collection Guide or in Description/Abstract
- All other questions should be sent to User Services larc@eos.nasa.gov
- For data products for which no Collection Guide or Description/Abstract is available, send all questions to User Services
Science Data Product URLs and Contacts

• Ordering Data
  – https://wist.echo.nasa.gov/api
  – EOS Data Gateway was decommissioned Feb. 27, 2009.

• Subsets of SSF, CRS, and ES8 are available
  – Order data using Java version of Langley Ordering Tool
  – Can subset by parameters or latitude/longitude box

• Contact Points
  – All questions regarding production data products and their use
    • E-mail: larc@eos.nasa.gov
    • Langley ASDC Customer Service

• CERES News (e-mail)
  – Subscribe from CERES Data Products webpage
  – All new public datasets are announced soon after public release
  – Mechanism for distributing CERES information
Questions and Comments