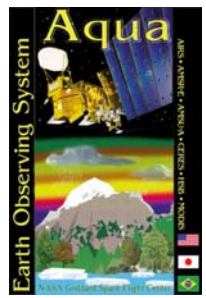
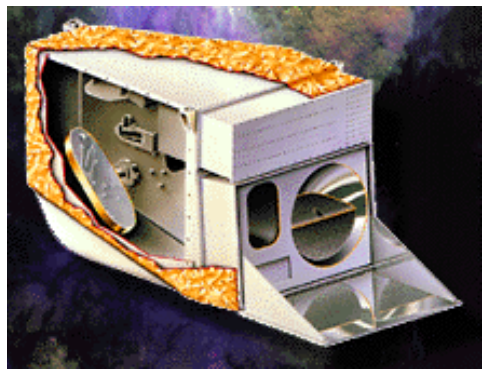
MODIS Land Long Term Archive

LP DAAC SAP Meeting
February 8, 2006

Robert Wolfe
Land Discipline Coordinator
NASA GSFC Code 614.5



Moderate Resolution Imaging Spectroradiometer (MODIS)



- Launched December 1999 on NASA EOS Terra Polar Orbiting Satellite
- Imaging since Feb. 25, 2000
- Whisk-broom scanner, 100° Field of View
- Orbit: Attitude 705 km, Inclination 98.2°, Mean-period 98.9 min., 16-day repeat cycle, 10:30 am equatorial crossing time
- Senses entire equator every two days, daily full coverage above 30° latitude
- 36 spectral bands: 29 - 1 km, 5 - 500 m and 2 - 250 m nadir spatial resolution
- Second MODIS on Aqua – launch planned for April 2002

MODIS Science Products

- Goals :
  - operationally produce terrestrial remotely sensed products that may be used by expert and non-expert community
  - establish a 10 year record that has continuity with precursor systems, e.g., AVHRR and SeaWiFS, and the future NPP and NPOESS VIIRS missions
- The MODIS products were developed primarily to serve the global change research community (MODIS has global, near daily coverage) and have many other potential applications
- Peer-reviewed product generation algorithms developed by MODIS Science Team
- Major contribution to 11 of the 24 EOS measurements and minor contribution to 2 others

Science Team – Phase 2

- Overall – Phase 1 Land Team intact and strengthened
  - Separate funding for Validation Coordination and Rapid Response
- Larger Science Team (27 total)
  - Algorithm Maintenance and Validation Team (10)
  - EOS Science – Data Users (17)
  - Four new/improved Land products: Plant Water Content, Evaporation, Improved Surface Reflectance, Physical Based Continuous Fields and Land Cover
- Discipline Leader: Dr. Chris Justice, UMD

MODIS Land Products

Energy Balance Product Suite

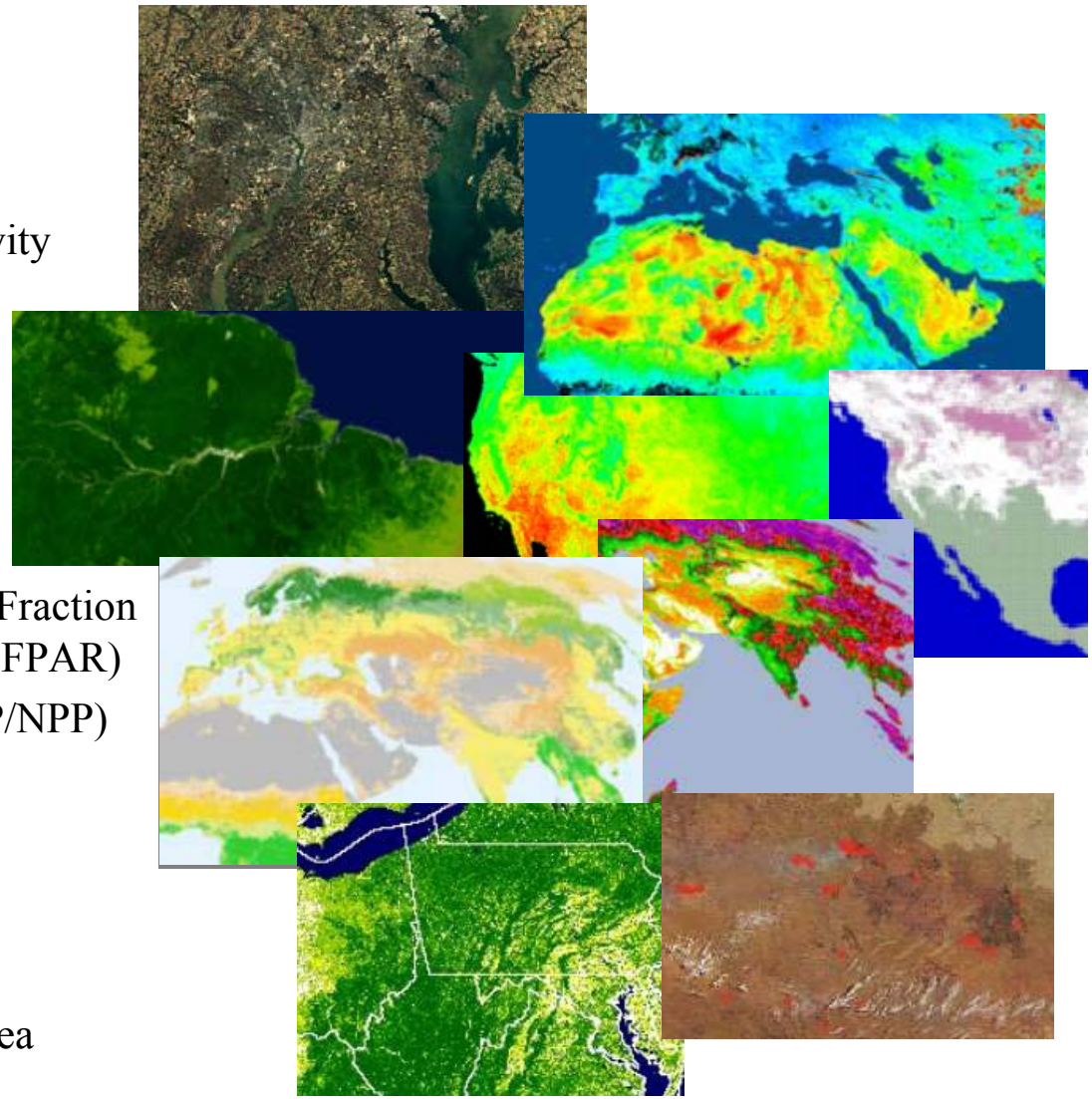
- Surface Reflectance
- Land Surface Temperature and Emmisivity
- Bi-directional Reflectance Distribution Function (BRDF) and Albedo
- Snow cover and sea-ice extent

Vegetation Parameters Suite

- Vegetation Indices
- Leaf area index and Fraction (LAI) and Fraction of Photosynthetically Active Radiation (FPAR)
- Gross and Net Primary Production (GPP/NPP)

Land Cover/Land Use Suite

- Land Cover/Vegetation Dynamics
- Vegetation Continuous Fields
- Vegetation Cover Change
- Fire/Thermal Anomalies and Burned Area



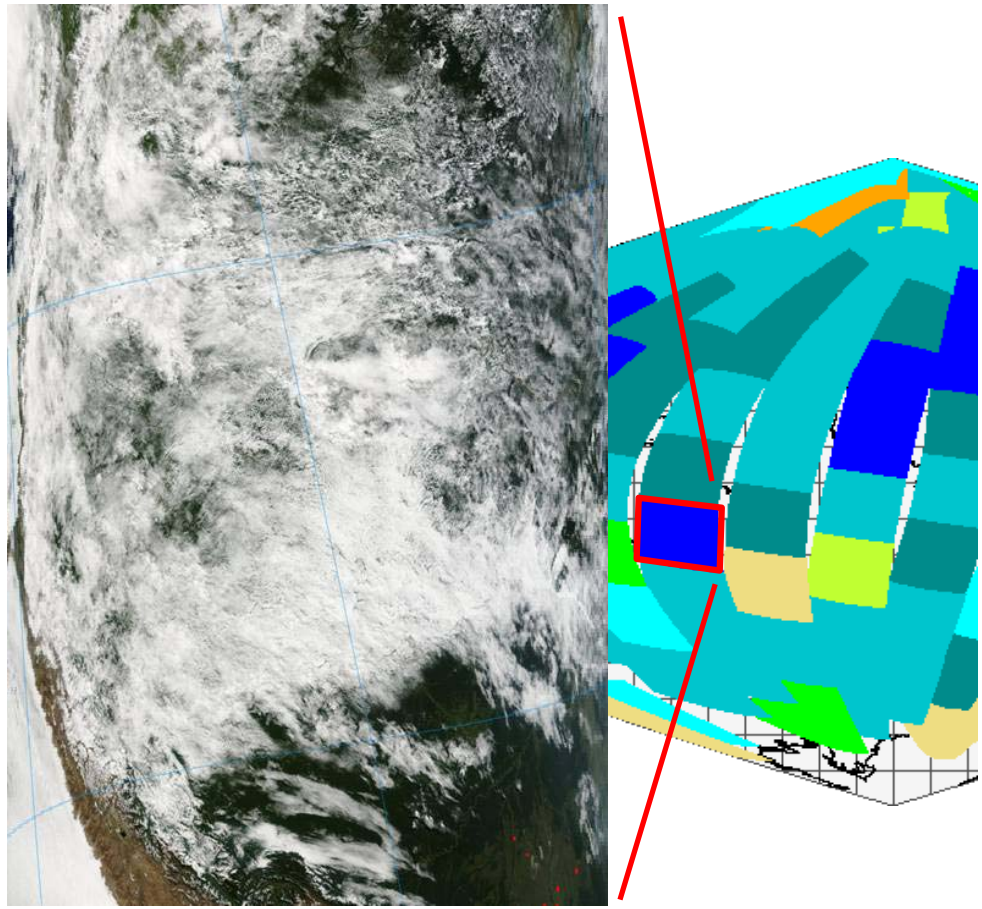
MODIS product hierarchy

- Level 1: geolocated and calibrated instrument data
- Level 2: retrieved geophysical parameters at the same location as the Level 1 data
- Level 2G/3: earth-gridded geophysical parameters
- Level 4: earth-gridded model outputs

Level 2 Swath

Retrieved geophysical parameters at same location and in same format as MODIS Level 1 instrument data

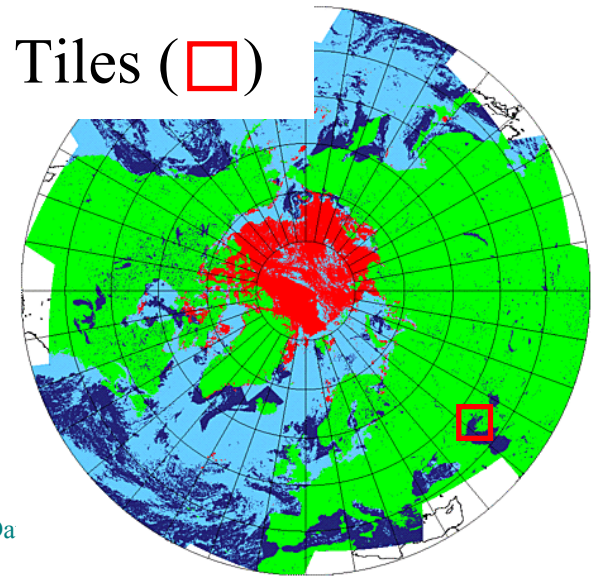
- 288 granules/day; 5 min.; approx. 2340 x 2030 km
- 250m, 500m and 1km nadir resolutions

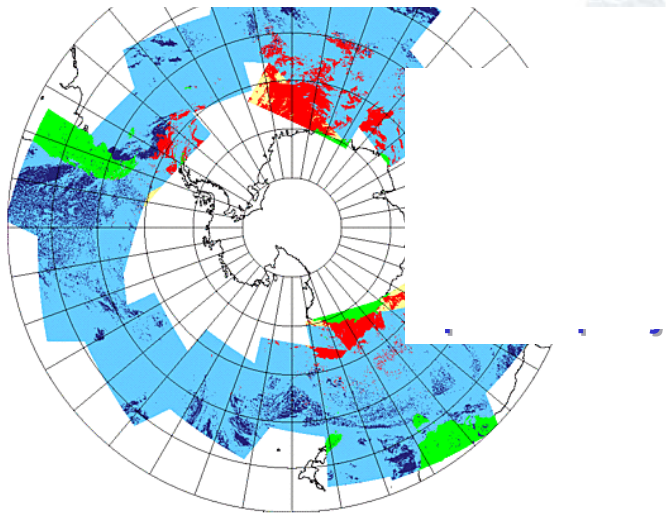


Level 2G, 3 and 4 Grids

- Level 2G/3: earth-gridded geophysical parameters
- Level 4: earth-gridded model outputs
- Daily, 8-day, 16-day, 32-day, monthly and yearly products
- 10º x 10º Tiles (□)

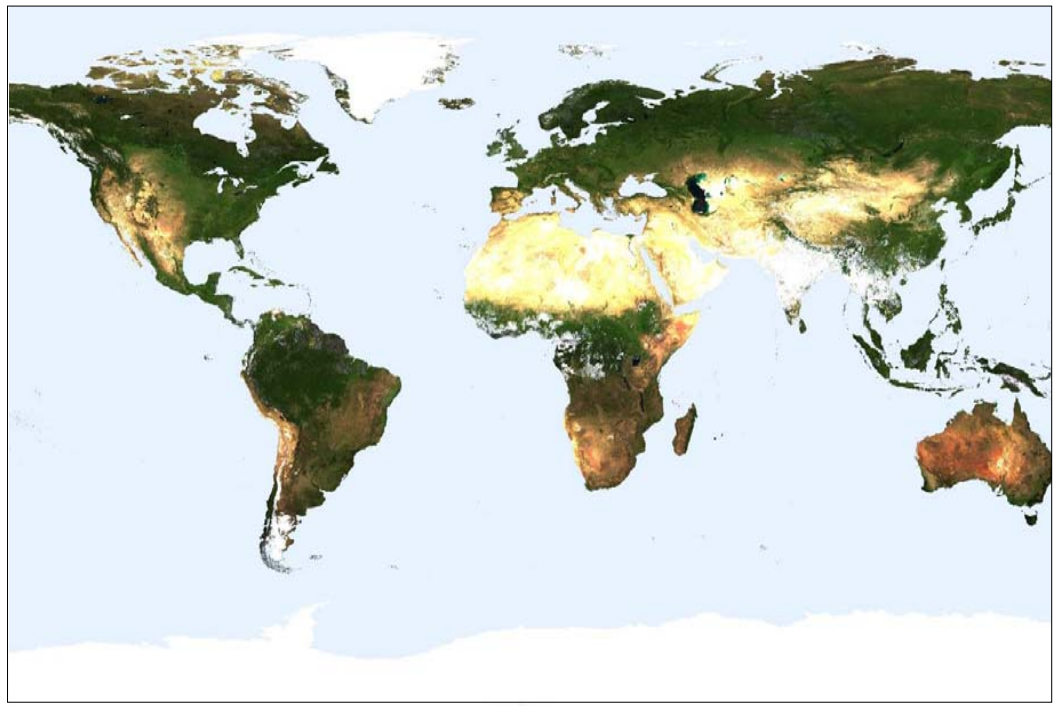
Sinusoidal (equatorial) – 7.5, 15 and 30 arcsec. resolution (roughly 250m, 500m and 1 km)





Climate Modeling Grids

- Global Level 3 Products designed for climate modelers
- 0.05º Resolution
- Almost all products are geographic (lat/long)
- Sea-ice is current exception – in polar grid



(from BU – NBAR CMG – days 193-208, 2001)

MODIS land products @ LP DAAC

	Products	Volume (GB/day)	Mission Total Volume (TB)	
			Full	Compressed
Land L2/L2G	4	119	413	207
Land L3+	31	80	304	152
Total Level 2+	**35**	**198**	**690**	**345**
Level 0	1	70	250	185

Notes:

- 10-years per mission (Terra, or Aqua with extension), double for both missions
- Uncompressed volume, expect compression factor of ~2 for L2+ products, 1.35 for L0 product
- ~2,300 files/day; 8.3 million files – per mission

Land ST Perspective on LTA

- USGS MODIS Long Term Land Archive planning underway
  - Science Team welcomes this institutional commitment to archiving land data
  - USGS should develop LTA user model – what functionality is needed
  - Strawman proposal of data types to be archived suggested by team
  - Individual archived products to be determined by measurement teams

From: Chris Justice/UMD, MODIS ST meeting, Jan. 2006

Strawman LTA Product Set

- Store all n-day L3+ products
- Store selected L2 and L3 daily products
  - L3 daily CMGs.
  - L2 and L3 daily fire – because they are very small when compressed
  - L2 and L3 daily LST – because they are difficult to reproduce and pretty small
- Produce remaining land L2 and L3 daily products on-demand
- Store Golden months (multiple version)

- Supplemental information that also should be archive:
  - Documentation
  - Metadata
  - Algorithm software
  - Data production software
  - QA tools and methods (browse, long-term trends, subsets, golden tiles, etc.)
  - Validation data (if not stored elsewhere)

User Model for High Level Products

- Issue: most models for LTA are for lower level products (L0 and L1 data)
- A good LTA user model is needed for higher level products, e.g., MODIS L2+ Land Products
- Are there similar Land data sets that could be used to help develop this model?
  - MRLC
  - Others?

Multiple version issues

- Pilot program
  - The archive should begin to store/process data before the end of mission to verify the capability will be available after the mission ends.
  - Also, this is needed to be sure that all of the information is captured.
  - This means that the first data stored may not be the definitive data set.
  - This could be done as a pilot program to verify all information (including supplemental) can be captured and transferred.
- Evolution of data sets
  - There will never be a "definitive" version of the products because algorithms will continue to evolve as algorithms and Earth Science advances.
  - Periodically, new versions of the archive products and algorithms will become available and the LTA should be updated appropriately.
  - When this occurs multiple version of the data will need to be stored.
  - Once older versions of data should be deleted after they are superseded and no longer needed.

Processing-on-demand Issues

- MODAPS will prove that Level 1 processing-on-demand (POD) is possible and is a quick, cost-effective and efficient way to deliver lower level Land products to users.
- The LTA could provide this POD capability to the LTA users to save archive space.
- Cost-benefit analysis needs to continue in the future to determine whether there are any cost savings from POD vs. storing products.

Levels of Service Detailed Definitions

- Detailed level of service definitions for:
  - User support
  - Ingest
  - Processing
  - Distribution
  - Archiving
  - Production Software Development
  - Documentation

- From: Ron Weaver, NSIDC data product review, Jan. 2006
  - Details extracted from the NSIDC Data Management Plan (7/15/2003 revision)
  - Also consulted the SEEDS Draft Recommendations (V1.31)

Levels of Service Summary (from a Data Center Perspective)

Level	1	2	3	4
User Services	None	Referral	General	Custom
Ingest	None	Ingest to archive	Ingest with QC	Custom
Processing	None	Simple monitor	Simple monitor	Custom
Distribution	None	Monitored	Manual	Custom
Archiving	None	Unsupported to minimal	Preservation of metadata only	Primary archive
Production Software Development	None	N/A	Internal standards for software followed	Custom
Documentation	None	Metadata only	Metadata and provider documentation	Compre-hensive

Details below

From: Weaver/NSDIC

Provide User Services

1. **None**

   User Services provides no support for the data set - typically used for disclaimered data that is not yet supported by NSIDC, or for brokered data sets.

2. **Referrals**

   User Services points customers to another source (e.g., data center or PI) for the data.

3. **General Support**

   User Services provides normal baseline support including responding to user inquiries, tracking requests, and access to broader data center resources (programmers, OPS, etc.).

4. **Custom Services**

   This may include things like reformatting or subsetting data, or creating non-standard or custom products. Examples of this include creating images for textbooks and creating a value-added product in direct response to a user request.

From: Weaver/NSDIC

Ingest Data

1. **None**

   The data is not held at NSIDC (e.g., brokered products)

2. **Ingest to Archive**

   The data is ingested and archived as is - It may take some setup to prepare the way for data archival (e.g., directory structures may need to be set up); but otherwise NSIDC's responsibility is to verify that the data is successfully archived.

3. **Ingest with Quality Control (QC) to Archive**

   Data goes through a QC process on its way to the archive. The QC may include such things as verifying that the data matches its descriptions, is named correctly, looks ok, is accompanied by complete metadata, etc.

4. **Custom Ingest**

   Indicates that due to high data volume, time criticality or other complexities, the ingest process is actively monitored and controlled (e.g., MODIS ingest into the ECS system).

From: Weaver/NSDIC

Data Processing

1. **None**

   No processing is done at NSIDC.

2. **Simple monitoring**

   Processing is automated; NSIDC's responsibility is to verify that the processing occurred successfully.

3. **(same as 2)**

4. **Custom processing**

   Significant manual intervention is required during the production process (e.g., to QC the results).

From: Weaver/NSDIC

Data Distribution

1. **None**

   NSIDC does not distribute the data (e.g., brokered products).

2. **Monitored**

   Distribution is automated; NSIDC's role is to verify successful distribution and to ensure that data access restrictions are followed (if any).

3. **Manual**

   Manual intervention is required to distribute the data (e.g., media distribution or the data needs to be staged).

4. **Custom Distribution**

   Indicates that due to high volume, time criticality, data restrictions or other complexities, the distribution process is manually intensive (e.g., media distribution from the ECS system, EASE-Grid CD production).

From: Weaver/NSDIC

Data Archiving

1. **None**

   NSIDC does not archive the data.

2. **Unsupported to minimal**

   NSIDC has the data, but it is not officially archived (i.e., the shoebox archive in somebody's office). This status should be avoided if at all possible, as there is absolutely no guarantee that the data will survive or even be known about in the future (i.e., if you would feel bad if the data were irretrievably lost, try to make sure the data does not stay in this state).

   One step up from unsupported. Minimal preservation metadata is maintained; whatever media, documentation, etc. exists is managed by the archive management group. Media refresh, offsite-backups, etc. occur only as time and resources permit.

3. **Standard support**

   A full set of preservation metadata is maintained, but NSIDC is not the primary archive for the data. Data recovery plans/agreements may exist, usually to recover data from the primary archive if the local copy is lost and/or destroyed. Media refresh and testing plans exist as appropriate. Provisions are made for data security.

4. **Primary archive**

   A full set of preservation metadata is maintained, as are off-site copies of the data (or other recovery processes). Media refresh and testing plans exist as appropriate. Provisions are made for data security.

From: Weaver/NSDIC

Data Set Production Software Development

1. **None**

   Production software was not required for the data set.

2. **Not applicable**

3. **Standard support**

   NSIDC follows its own internally established standards for software development. These standards include improvements or modifications to processing code or increased explanatory documentation when necessary. For in-house developed software, programmers review each other's code, and extensive testing is done to ensure the quality of the data set.

4. **Custom support**

   For data sets developed outside of NSIDC, a product team determines the appropriate level of development based on the available resources and the needs of the user community.

From: Weaver/NSDIC

Data Set Documentation

1. **None**

   No documentation support is provided. This status should be avoided since it does not guarantee enough information to ensure that the data can be usefully preserved into the long term.

2. **Metadata only**

   A standard form of metadata is created and/or captured for cataloging, tracking, or accessing. See the Catalog Team's level of catalog metadata.

3. **Metadata plus provider documentation**

   A standard form of metadata is created and/or captured for cataloging, tracking, or accessing. Additionally access is provided to original documentation.

4. **Comprehensive documentation**

   The preferred level of service for data to be maintained and/or presented to the public for the long term. Metadata is created or captured. Documentation is written or provider documentation has been reviewed and edited. Additional documentation services are provided by creating special web sites, expanded sets of user documentation or other value-added products as needed.

From: Weaver/NSDIC

Backup Slides

EOS SWGD LTA Questions, MODIS Answers

- MODIS answers to the EOS Science Working Group on Data LTA questions – presented at the LTA workshop (Jan. 2002)

1. A decision is needed on whether the LTA's ability to deliver higher-level products is based on archived higher-level products or the ability to generate them from lower-level data.

- The primary function of the EOS LTA is to provide long-term archive and data distribution for the standard EOS data products archived at the DAACs.
- On-demand processing is unworkable for climate data records (Level 3 time-series). Difficulties for MODIS include:
  - Maintenance of software/hardware
  - Complicated production rules/dependencies
  - Porting software to new hardware and software environments
  - New on-demand ordering interface would need to be developed for MODIS
  - Manually intensive quality assurance
  - Many ancillary datasets needed
  - Higher-level products require large amounts of lower-level products to be processed
  - Multiple orders for same data drive needed X-rate - which drive costs
- MODIS L1s and some L2s could possibly be done
  - Cost savings must be shown to be practical
  - Can only be reasonably done if very automated.
  - May best be done outside of archive

2. A decision is needed on which product levels should be archived.

- Product "levels" is too broad; what makes sense for each product must be assessed
- Earth science community expects data to be archived
- Who makes the decision?
  - Review of products needed - primarily by a high level committee; possibly a National Academy of Sciences Subcommittee; this is a national effort that crosses agency boundaries
- Possible international implications
- Criteria should be:
  - How well they address the 24 EOS measurement (and any new measurements)
  - Current usage should be a guide
- Archive vs. regenerate decisions on a per product basis have to be made (see #1)
- Other decisions needed on levels of service to provide
- L0 and associated data is mandatory

3. NASA investigators require continual, on-going access to the data sets, and without additional charges.

- Agreed.
- LTA should provide long-term easy access to the data.
- However, if charged:
  - Funding mechanisms like NASA data-buy is preferable to charging individual researchers.
  - Minimal charges for media and shipping are acceptable.
- Better interface needs to be developed – more automation is needed to allow volumes of data that investigators need
  - R&D for on-line access needed – should be "help yourself"
- Network bandwidth could be an issue: trades of proximity (collocation of investigators with dataset) vs. network bandwidth are needed
  - Media distribution likely to ramp down, but may still be required
  - Mirroring datasets at secondary sites may be needed

4. The LTA must cater for multiple instruments. Beware of concentrating the development on one or two instruments with the result of not addressing unique requirements of other instruments.

- What are the unique requirements for MODIS?
- MODIS has aspects of both an image archive (L1s and some L2s) and climate data record archive (time series)
  - On-demand does not work for MODIS climate data records (see #1)
- MODIS is a large data set: large number of products, large number of granules, large volume
- MODIS higher level products are spread across three DAACs
  - Bringing together inter-disciplinary datasets needs to be addressed

5. If the LTA is to have a production capability, it may have to operate without the Q.A. function provided by the current instrument teams. (Those teams may no longer exist.)

- Means QA has to be automated.
  - Currently manually intensive (costly).
  - Development should be done before teams disband.

6. Why not keep the existing DAACs, and if necessary transfer their ownership to USGS or NOAA?

- May be only way that makes sense fiscally (for archive)
- Difference in function (LTA vs. DAAC) need to be well defined.
  - Must determine mandatory levels of service vs. desirable
- Different mission end-dates may make transition from DAACs difficult (some data sets are serviced one way, others another?).
- Discussion needed of data transition (move from active to long-term archive) vs. migration (transfer to new technology)
- Documentation, metadata, (web pages?), must also make transition.
- Maintaining a processing capability outside the archive may still be needed (to generate L1s and L2s, for instance)

7. It is important for the LTA to handle multiple versions of products?

- The LTA primarily needs to handle the latest (final) products.
- However, limited sets of some key prior versions may need to be kept (in the mid-term?)
  - Rational: research results published on prior datasets
  - Can be limited: sampled spatially, temporally, geographically.

8. There should be an LTA science advisory board at a very high level within NASA/Code Y.

- Agreed.
- The function of the board is to decide: what products to archive; what level of services to provide; when data migration is needed
- Cross agency board also needed (see #2); NASA will likely need a parallel structure,
- Institutional stewardship of datasets needed:
  - Managed by scientist(s) who are funded to use it for research
  - Oversight by panel of users

9. There may need to be new role models, e.g. for climate data records, as distinct from traditional land or atmosphere data records.

- AVHRR may be good starting point for MODIS
- Secondary functions of the LTA are to provide services for subsetting, reprojeciton and mosaicing of the archived data products and to produce some well defined global data sets.

10. There are unique aspects of the short-term years and the long-term years that need to be addressed.

- Short-term is through mission life (including last reprocessing), mid-term is up to end-of-mission + 5 years and long-term is beyond end-of-mission + 5 years.
  - Final reprocessing would be done in the short-term years based on the combined knowledge of Terra and Aqua MODIS calibrations
  - For MODIS:
  - Level 0 and 1 products - archive long-term
  - Most Level 2 products - only archive for short-term
  - Selected Level 2, most daily and some 8-day Level 3 and higher products - only archive for mid-term
  - Selected daily, 8-day and the 16-day, monthly and quarterly products - archive long-term

11. How does the LTA handle the various extensive ancillary data sets needed for creating products.

- See #1
- Small ancillary data sets should be stored where needed
  - Multiple copies may be needed
- Some small datasets are currently copied into some MODIS products
- Critical ancillary data for MODIS processing: winds, atmospheric pressure, humidly, PAR, ozone
- Cross-instrument processing is another issue.

12. What are the prospects for truly low-cost storage devices that could be used to implement the LTA? What is the current state of the market? Is it sufficient to allow archiving of all product levels and all versions?

- Unknown. Moore's Law has helped us in the past and may save us in the future.

x. New issue(s)?

- How many copies of the data need to be kept if capability does not exist to regenerate the data products?
  - What is the failure rate of the archive?
  - How much data loss is acceptable?
  - Maybe a limited capacity to regenerate lost products is needed as opposed to keeping enough copies

Distribution activity levels

1. **None:** no requests. Data set is archived in a steady state
2. **Low:** between 1 – 5 requests per year, less than 5% of the total archive data volume
3. **Nominal:** greater than 5 requests per year
4. **High:** greater than 50 requests and/or greater than 100 GB per month

- Distribution impact on is driven more by number of requests that require user services interaction at the low end, but more by data volume at the high end.

From: Ron Weaver, NSIDC data product review, Jan. 2006