MODIS processing: lessons learned and future plans

Edward Masuoka and Robert Wolfe

NASA Goddard Space Flight Center, Greenbelt, MD, USA

Outline

- MODIS Lessons Learned
- MODAPS and LAADS processing/distribution
- Quality Assessment and Validation

MODIS Lessons Learned

Team Leadership and Communication

- Good leadership and regular communication are needed
- Regular forums for information exchange
  - Weekly Technical team meeting keeps everyone on same page
    - consists of team leader, discipline leads, instrument/calibration lead and data system lead
  - Weekly calibration and data system meetings
  - Full Science Team meetings twice per year early in mission (once per year later)
- Science Team members funded before and through the life of the instrument – a long term commitment and experience base is critical
- Disciplines – Discipline Leads and a core full-time group built around the data systems with weekly discipline telecons
  - Close working relationship between data team and science team members is important
- Calibration – core calibration group working with discipline representation
  - Discipline representatives provide important feedback because they understand calibration and its impact on fundamental (upstream) science products (e.g. water leaving radiance, land surface temperature, land surface reflectance, aerosols)

Data Systems

- Reprocessing: It should be possible do a minimum of entire mission in 1 year
  - Ideally should be able process mission in 3 months or less
  - Having Level 0 data on-line makes science testing and reprocessing much easier
  - Disciplines will have their own reprocessing schedules
  - So the reprocessing data system needs to be able to handle this
  - Multiple strings may be needed for individual disciplines
- Science test and integration systems should be able to finish major tests in 1-3 months
- Science testing
  - Don't bundle too many changes together – impacts of upstream algorithms are hard to untangle
  - Many iterations of testing are needed
  - Storage is needed for multiple baselines and changes
  - Carefully control for ancillary data dependencies
  - Maintain a basic set of science test days
- Over the life of the mission, porting to new systems may be necessary, but should be kept to a minimum
  - Build portability into software
  - Poor programming practices in the science software are often found during porting
  - Changes to compilers, libraries and data system may also need to be tested
  - Porting requires careful evaluation and key human resources – but it is not of primary interest to the science team

Algorithm Development, Integration and Testing

- ST members should play a key role in development and maintenance of code – remain responsible
  - It is better to produce some experimental products at the Scientist Team Member's Computing Facility – but a method is needed for eventually bringing code into the primary production facility
- Science team led processing is needed to ensure that the best products are generated and reprocessed with a workable schedule
- Stable calibration and algorithms are needed to produce a good time series
  - Don't make changes in middle of reprocessing
  - Minimize changes in forward processing (after calibration and algorithms stabilize)
- Aqua had much better algorithms after Terra – end-to-end flow already tested
- Key Aqua instrument issues (no instrument is a perfect clone):
  - Band 6 dead detectors caused changes in some algorithms (e.g. cloud mask, snow)
  - Band to band registration problem had an impact on some algorithms (e.g. aerosols)
- Improved Aqua instrument characteristics made some products possible, e.g. Ocean Color, Deep Blue Aerosols
- Process for bringing in new products needs to be streamlined because of the current 3-year proposal cycle
  - Ocean group has developed a well-defined process that allows for community feedback

Distribution

- Key communities will need specialized products
  - Products in Climate Modeling Grid for climate modelers
  - Near Real Time products for applications users
  - Plan for additional formats and tailored products
  - One map projection (grid spacing, etc.) will not work for all users (or even a majority)
- Self document format (HDF) is useful – but additional formats are needed
  - Geographic information system users have community format standards (not HDF)
- Always have a full collection going forward
  - Reprocess the full mission in background
  - Give users time to switch over to new dataset
- On-line products (including Level 0) greatly increased distribution
  - Made getting data much easier and increased speed of fulfilling orders
  - Enables processing-on-demand for tailoring products as well as web services
  - Need to perform trade study between processing-on-demand vs. storing products on a product by product basis

Quality Assurance

- Centralized QA activities for each discipline is invaluable
- Land group has the following functions:
  - Global browse and golden tiles
  - Subsets and QA database
  - Known issues pages
  - Initial point of contact for user's quality concerns
  - Help in tracking quality issues to source (sometimes in the upstream algorithm)
  - Coordination and feedback on science algorithm testing
- Atmosphere team makes available:
  - Projected granule browse of L1B and key atmosphere products
  - Images of hundreds of L3 global parameters including min, max, standard deviation fields
  - Collection highlights and known issues
  - Standard visualizations of all test data for the science team
- Ocean team offers similar capabilities

MODAPS and LAADS: Evolution, Current Systems and Future Plans

Introduction Terra – The Early Years

- EOSDIS as initially conceived consisted of a common suite of software and hardware running at 5 data centers (DAACs) which would handle the processing, archiving and distribution of products from EOS satellites
- In an effort to make the data system more distributed and reduce its components to more manageable “chunks”, the generation of standard products was moved, in most cases, from the EOSDIS Core System (ECS) to Science Investigator-led Processing Systems (SIPSs) developed and operated by the respective instrument teams
- The approach to the development of ECS was also modified to result in more frequent releases of its Science Data Processing Segment based on priorities expressed by the science community
- These steps led to the successful completion of all subsystems needed to support Landsat-7 (launched in April 1999) and Terra

Data system evolution (or Mouse to Elephant)

- For MODIS/Terra (launched in 1999), biggest data processing system challenge was to retrieving terrestrial geophysical values at 250m, 500m and 1km resolutions on a daily basis over the global land surface (using 1990 tech.)
  - Data processing system was primarily I/O constrained
  - Only 1TB of production disk and 10TB of fast intermediate tape storage
  - More than 1TB of data a day had to be written and retrieved from this tape system to produce the multi-day products
- To overcome this problem, two approaches were used:
  1. Reduce the volume by producing products at the 250m resolution only over the continental US and small areas outside of the US – **science team set priorities**
  2. Acquire more disk storage so that the amount of I/O to tape was minimized
- For MODIS/Aqua (launched in 2002), the data system had improved so much that the system was capable of producing the full set of Land products over the global Land surface for both instruments
- Since that time, investments were made in commodity hardware so that the systems are now capable of producing the MODIS products at a rate of more than 50 times the real-time rate of each instrument

Production – Driven by Science

- Over the Terra mission lifetime, better calibration and characterization of the instruments have been performed and new and improved algorithms have been developed
- This means that to extract the maximum value from the investment in the Terra research mission, multiple reprocessing campaigns are needed
- For MODIS, reprocessing is also driven by the complex interdependencies between the algorithms
- Terra, like many of the NASA missions, has lasted well beyond its design lifetime of 5 years, and is expected to survive long enough to produce a total of more than 15 years of high quality data
- These reprocessing activities are typically driven by the science team and community
- The reprocessing cycle involves three phases
  - Development of the algorithm improvements
  - Testing the improvements
  - (Then) the actual reprocessing

Production – Reprocessing Rates

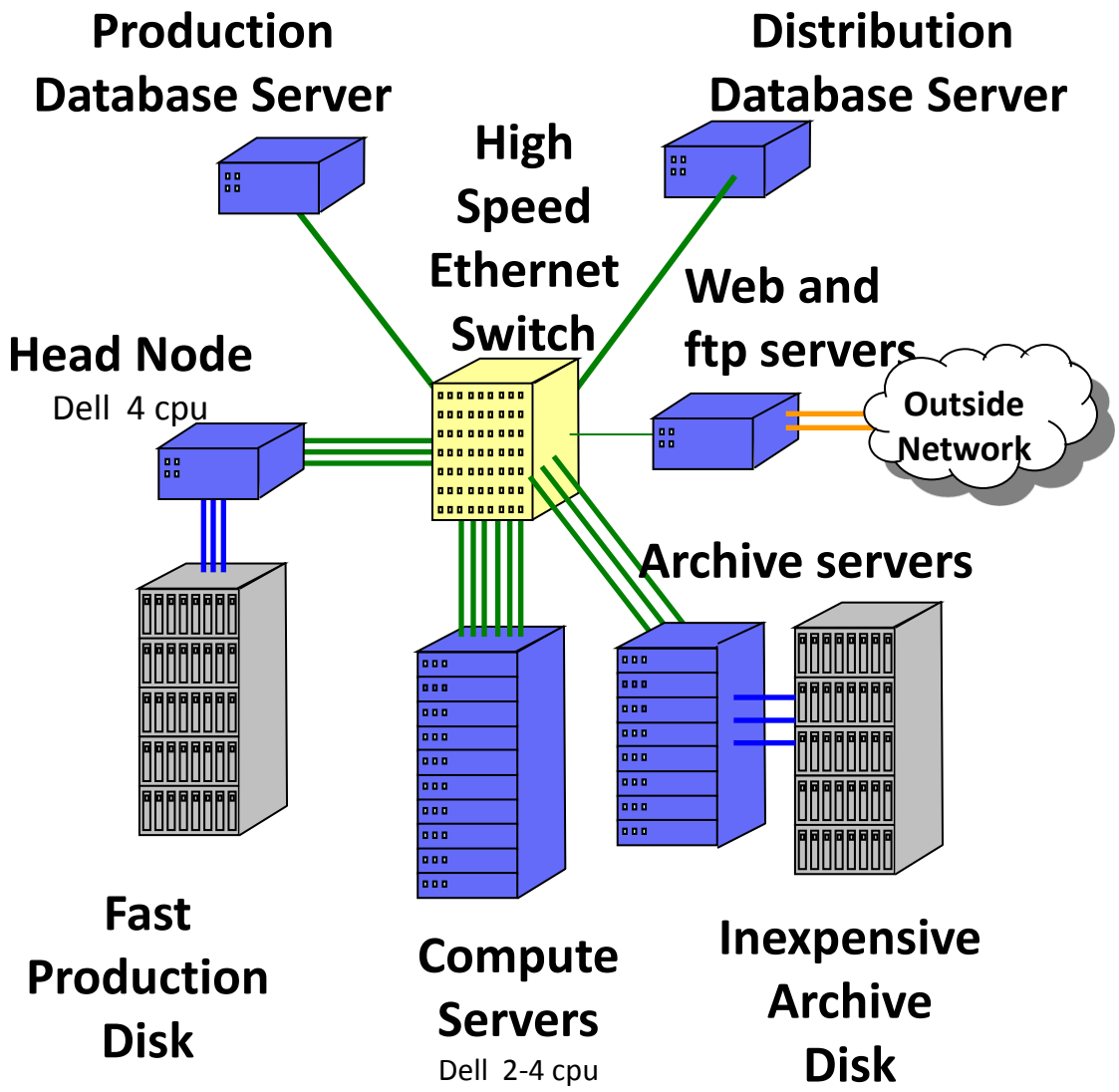
- Scaling the data production system to meet these reprocessing needs is a major challenge for the EOS data systems
- MODIS example:
  - MODIS team is in the process of preparing for the fourth major reprocessing for which the production phase is schedule to begin later this year, about 11 years after launch
  - Average time between reprocessing is about 3 years with the first one taking place after the algorithms were stabilized after launch
  - With a three year cycle, the actual reprocessing needs to be done in less than a year so as much time as possible is available for the important algorithm development and testing phases
  - So, needed reprocessing resources are more than 1X per mission-year
  - For current reprocessing campaign of 11 mission-years of MODIS/Terra data, the reprocessing capacity is 50X (a conservative estimate based on ingest capabilities at the archives) and network bandwidth

MODIS Adaptive Processing System

- Scalable system for MODIS, AVHRR, VIIRS and Landsat processing
  - Designed so that processing resources can be easily moved where needed
- Built with commodity hardware
  - Easier to scale, cost savings and easier technology refresh
- Built with open source components
  - Linux, Apache, Perl, Postgres, Subversion, FUSE
- All data products are online
  - Facilitates reprocessing of Level 2 and Level 3 products
- Designed to run with limited staff
  - “Lights out” processing outside of normal business hours
  - Months of processing can be queued up and execute w/o human intervention
  - Alerts emailed to system administrators when hardware components generate warnings or fail
  - Easy to use tools for monitoring the system (Ganglia) and investigating failed jobs
- Rapid updating or provisioning of servers with science processing software, the operating system and applications (Depot and SATE)
- Capability to separate data products into archive sets to accommodate storing a large variety of test results in an online archive

MODIS and OMI Data Systems

- Middle-tier servers and disk storage
- Open-source software
- Uniform system configurations
- Automated updates daily from Depot
- Automated H/W problem reporting
- Moderate Systems under Flight Projects Security Plan
- Currently 1,000 servers, 5PB of storage
- Maintained and enhanced by 6 staff members (Security, Database, Linux Sys Admin, Facility Management, Documentation, Property Management and Purchasing)



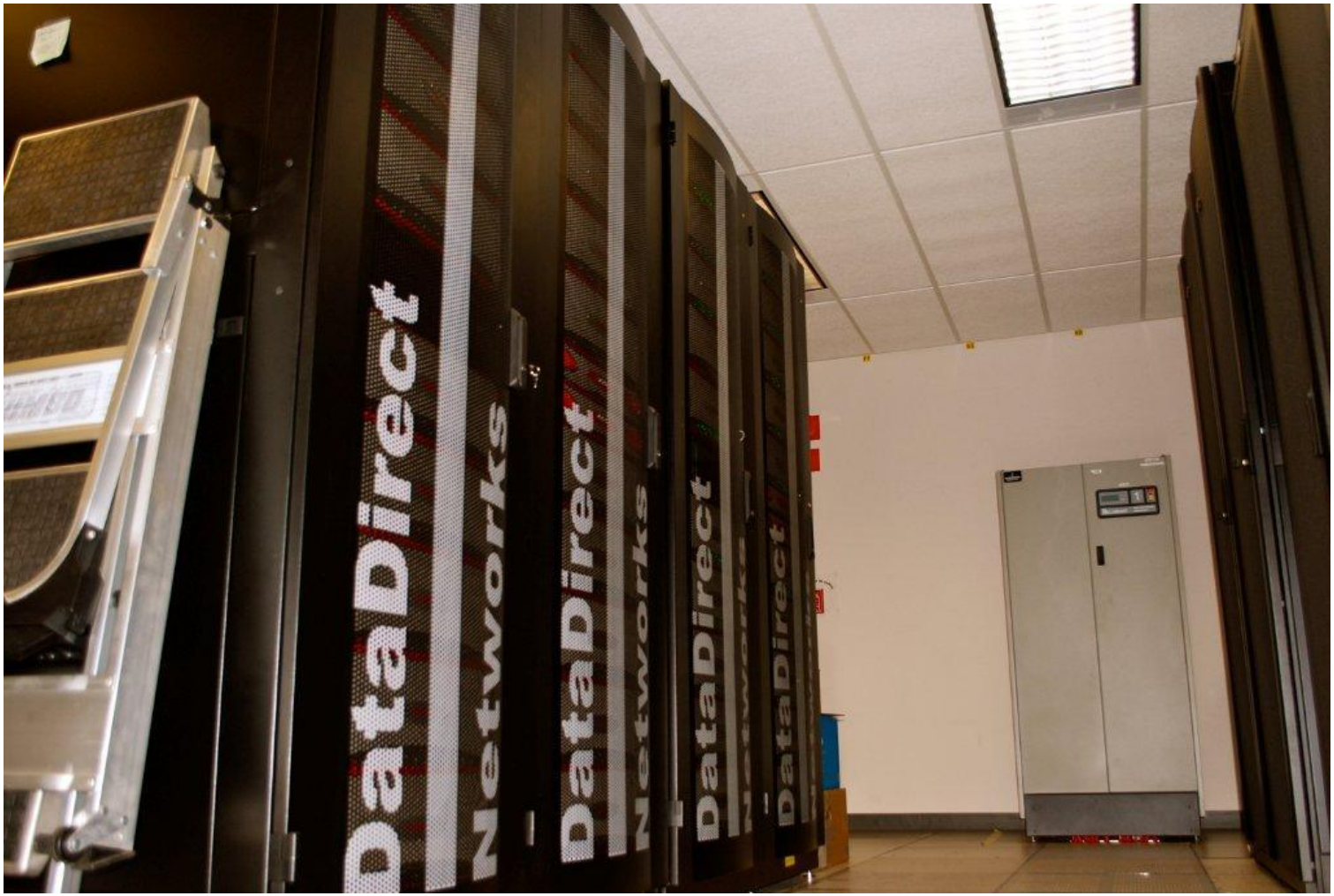
Some of our Dell servers

Level 1 and Atmosphere Archive and Distribution System

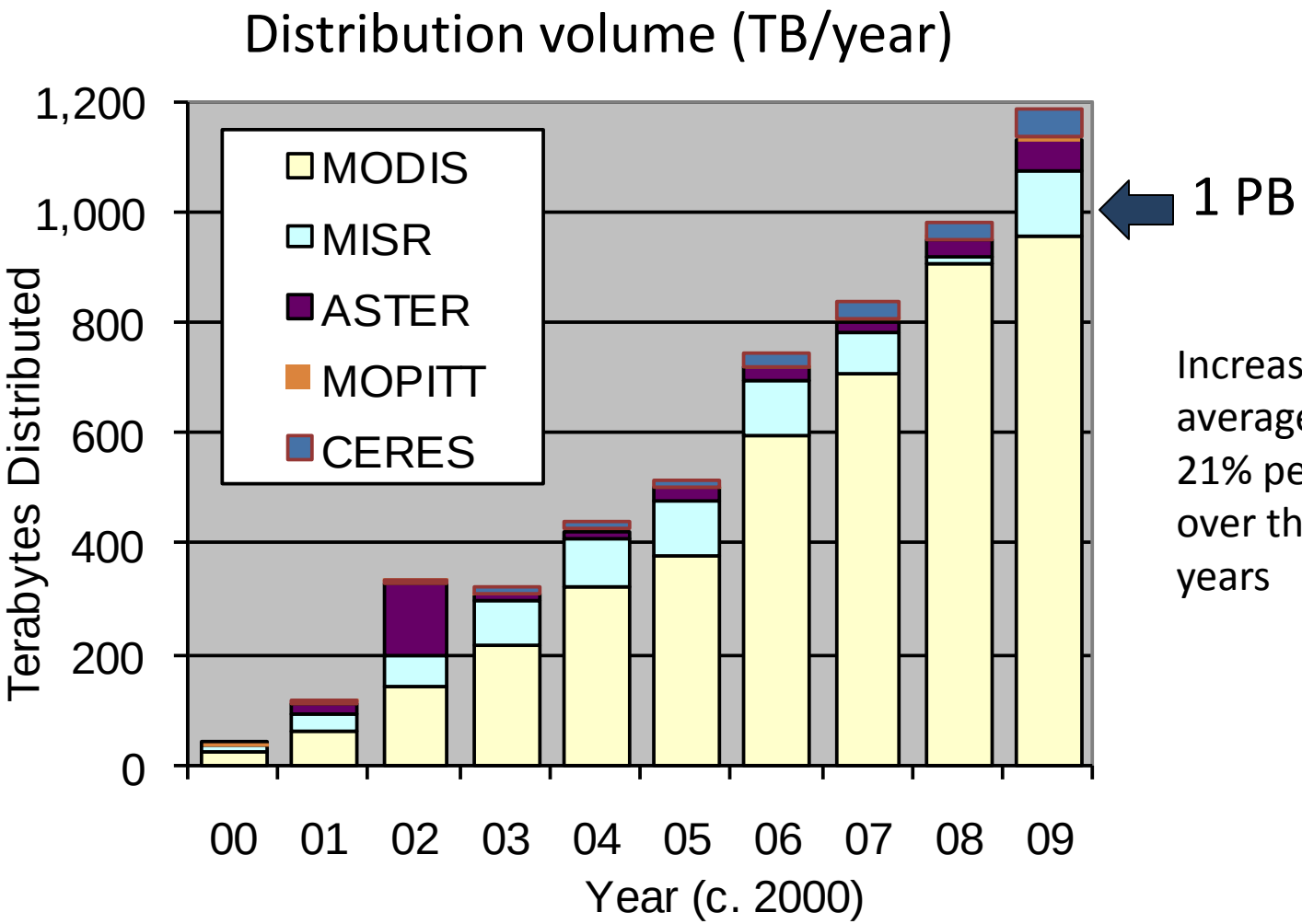
- Web-based search and order at http://ladsweb.nascom.nasa.gov/
- Online archive of MODIS Level 1, Atmosphere and Land products can also be accessed using ftp though a directory structure organized as:

  /allData/Reprocessing Collection #/Data Product/Year/Day
- Built upon the MODAPS framework to support post processing of products including the following operations: subset, sub-sample, mosaic, mask, parameter selection, geographic reprojection and format conversion)
- The archive also includes on-demand products which are produced when ordered by the end-user
- Web services allow machine to machine access for functions on web site
- Separating data products into archive sets to accommodate storing the large number of test results in the online archive as well as multiple reprocessing campaigns
- Other interfaces to the online archive include: an iRODS server providing access to the MODIS atmosphere products for users of the NASA Center for Climate Simulation (NCCS) and a server for EPA scientists that works in concert with visualization and analysis software running on their desktop systems to subset, sub-sample and combine MODIS atmosphere products.

2.5 PB of online archive



Terra data volume distributed



Increased an average of 21% per year over the last 7 years

Looking for Fast, Cheap **and** Reliable

The MODIS guys at the National Marathon not very fast and not very cheap

Greyhound at the Knoxville Marathon, very fast and very cheap but can he run 26.2 miles?

Data System Evolution: What's Next

- **Cloud Computing** (Tilmes, Golpayegani) in Nebula OMI and MODIS
  - OMI SIPS instance in Nebula generating OMI products on 100 virtual machines
  - Archiving data products in the cloud is also under evaluation
  - Assess performance, ease of use, data transfer and cost
- **More cores per processor** (Masuoka) MODIS
  - MODAPS generates all the products it can from 5 minutes of data (1 granule) sent to a single core, multiple granules are being processed in a single server
  - Current processing rates (50x atmos., 35x land) are based on 2 racks of 20 Dell servers with 8 cores per server
  - Exploring how to maximize performance for Dell servers with 24 cores each
- **Graphical Processing Units many-cores** (Tilmes) OMPS Limb
  - nVIDIA "Fermi" GPUs, 480 cores, 500Gflops (peak)
  - Apply in compute intensive applications, e.g. hyperspectral instruments.
  - Learning curve for parallel code, upside is low cost, low power
- **Data provenance and reproducibility** (Tilmes) Ozone, Land LTDR
  - How to capture everything needed to reproduce a product in the product
  - How to reproduce a product on a variety of platforms (platform independent PGE)
  - How to easily determine where products differ and what is close enough

MODIS Land Q/A and Validation

Summary of MODIS Land Lessons Learned

- Close interaction between the science team and the Science Data Support Team (SDST) allowed for timely decisions to be made on the production priorities and adjustments to be made to the production schedule so that early understanding of the product quality could be made while the data system was being improved.
- Centralized QA group expedited handling a large number of interdependent algorithm changes by investigating quality problems and coordinating the geographically dispersed science team members.
- Recognized the importance of validation and set up a centralized activity to coordinate validation of the Land products.

MODIS Land Products

Energy Balance Product Suite

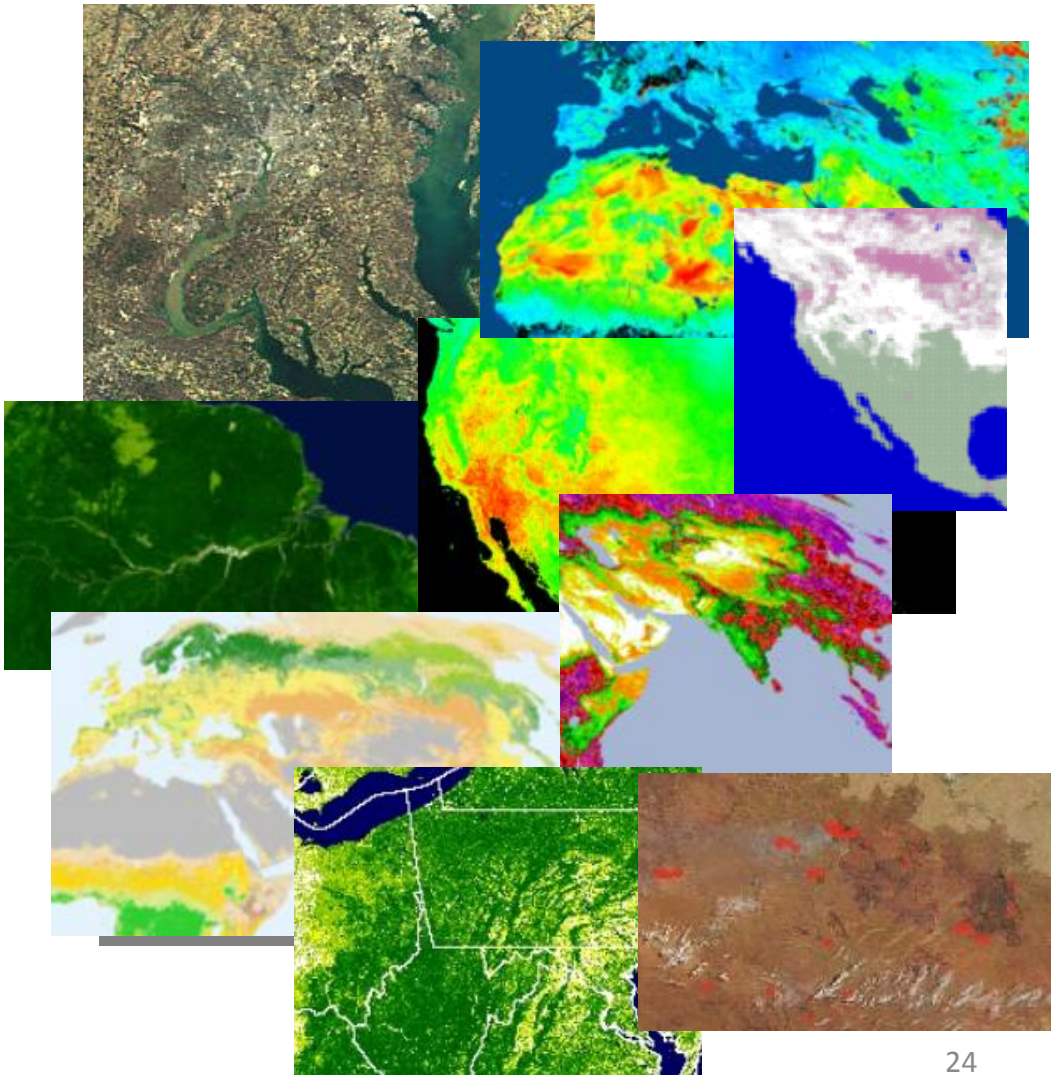
- Surface Reflectance
- Land Surface Temperature, Emmisivity
- BRDF/Albedo
- Snow/Sea-ice Cover

Vegetation Parameters Suite

- Vegetation Indices
- LAI/FPAR
- GPP/NPP

Land Cover/Land Use Suite

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



MODLAND QA

http://landweb.nascom.nasa.gov/cgi-bin/QA_WWW/newPage.cgi

QA Summaries

Known Issues

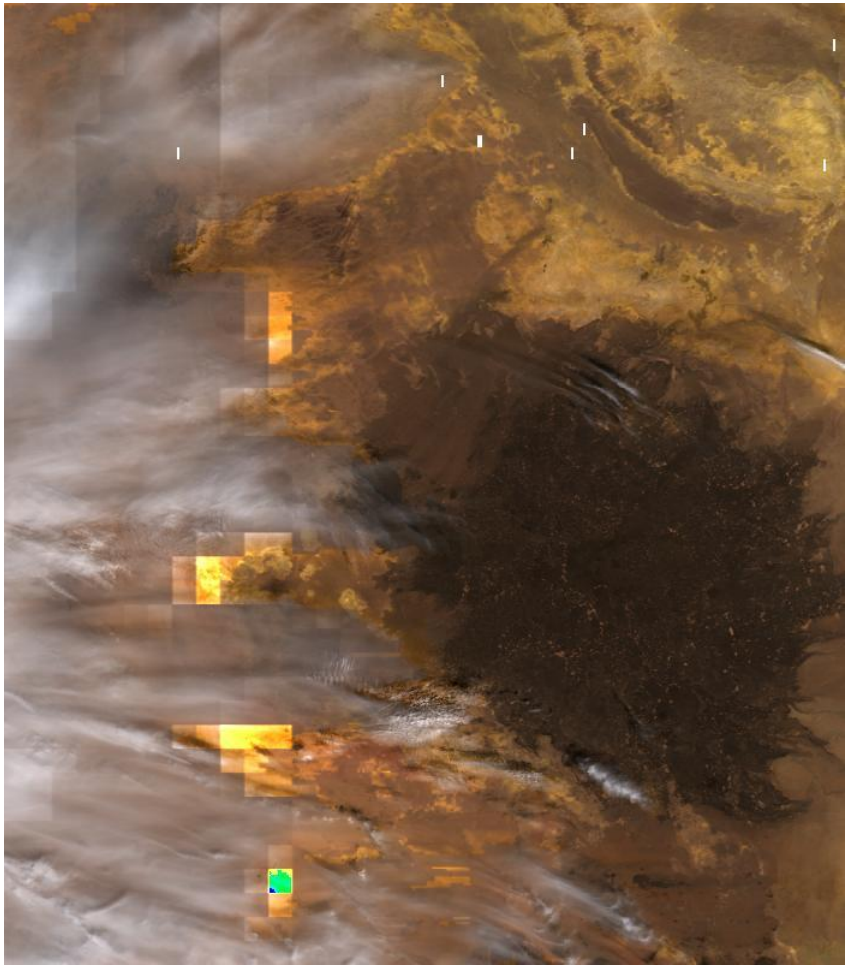
Global Browse

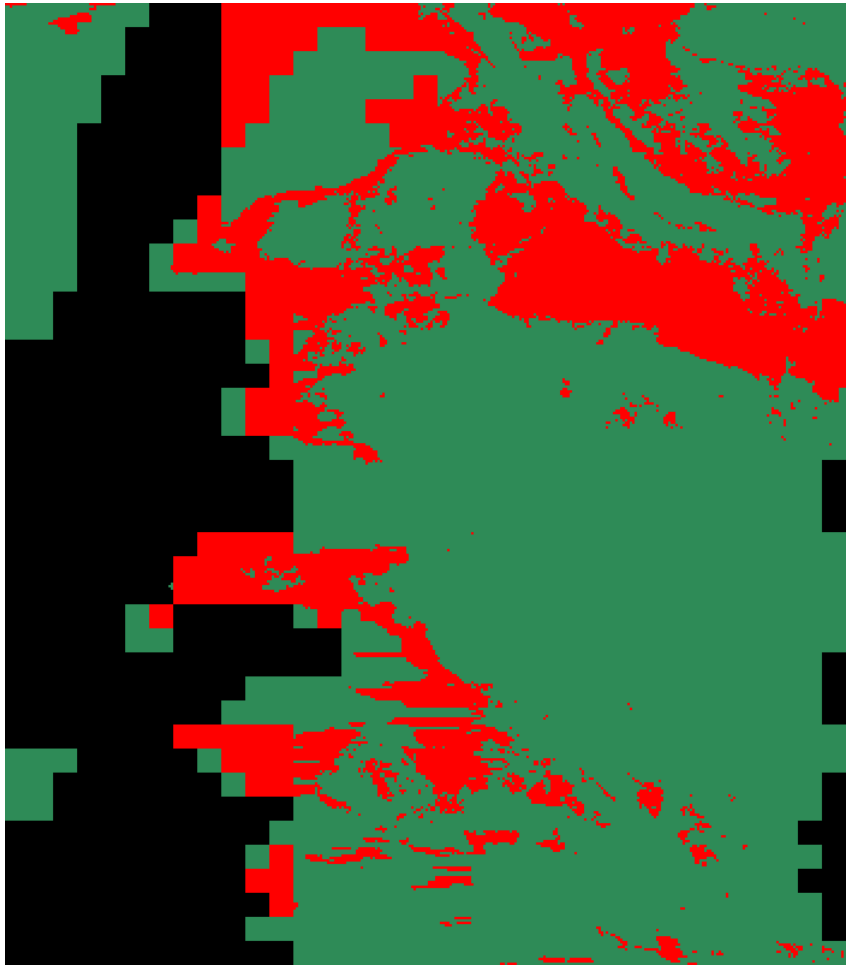
Time Series

Communicate Known Issues to team and community

Color Key	Case pending	Case closed	Case reopened	QA note

Large Image





MODLAND Validation
http://landval.gsfc.nasa.gov/

Val Summaries

EOS Core Sites

Major Campaigns

Val Metadata w. ORNL

MODIS Validation Data sub sets w. EDC

GODDARD SPACE FLIGHT CENTER | + NASA Homepage

MODIS land team

Home | Core Sites | Val Status | Campaigns | Documentation

Announcements:

- View and comment on the ESDR White Papers developed by the NASA Land Measurement Team
- MODIS Science Team Meeting, Oct. 31-Nov. 2, 2006, at the University of Maryland.
- *Announcing....* MODIS Land Collection 5 Workshop, Jan. 17-18, 2007, University of Maryland, see Draft Agenda.
- Validation of global vegetation indices and their time series, a one-day workshop, August 7, 2006, Missoula, MT
- Global Vegetation Workshop 2006, August 8-10, Missoula, MT
- TGARS Special Issue on Global Land Product Validation now available

MODIS News

- Terra
- Aqua

The Committee on Earth Observation Satellites (CEOS), defines validation as the process of assessing, by independent means, the quality of the data products derived from the system outputs.

The MODIS Land Validation Strategy

MODIS Land (MODLAND) product quality is ensured by Calibration, Quality Assurance (QA) and Validation. The MODIS land validation effort will contribute to and leverage off of international validation activities, helping to establish standards and protocols through close coordination with the CEOS Land Product Validation (LPV) subgroup, under its Working Group on Calibration and Validation (WGCV).

MODLAND uses several validation techniques to develop uncertainty information for its products. These include comparisons with in situ data collected over a distributed set of validation test sites, comparisons with data and products from other sensors (e.g., ASTER, AVHRR, MISR, TM/ETM+), intercomparison of trends derived from independently-obtained reference data, and analysis of process model results.

MODLAND's primary validation technique includes the collection of field and aircraft data, and comparison with these and with products from other satellites. The infrastructure for these efforts has resulted in the establishment of a semi-permanent array of EOS Land Validation Core Sites, most of which include a flux tower for extended temporal

Validation

- Should be funded as an integral part of the product development
- Some centralized validation coordination is needed
  - Makes effort more cost effective
- Five stage product maturity and validation approach works well
  - Beta, Provisional, Validated Stage 1 to 3
  - Helpful for provisional data to be available for community feedback
- Land: high resolution imagery and subsets for core validation sites were valuable
- Learned from Terra: validation should not be too soon after launch was applied to Aqua
  - Products need to be stable
  - Aqua products were more mature before validation occurred
- Useful to have Near Real Time images to support validation campaigns