



NASA
Science
Mission
Directorate

From Data Curation to Data Analytics: A Big Data Experiment in NASA Earth Science

Tsengdar J. Lee, Ph.D.
Weather Data Analysis Program Manager
High-End Computing Manager

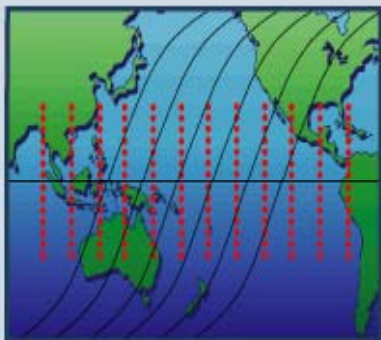
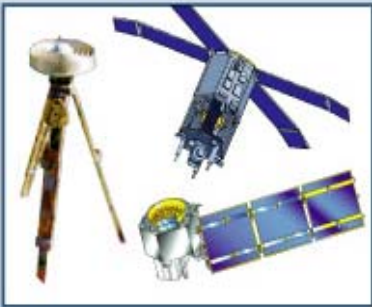
Turning Observations into Knowledge Products

Downlink Speed

Petabytes 10^{15}

Multi-platform, multiparameter, high spatial and temporal resolution, remote & in-situ sensing

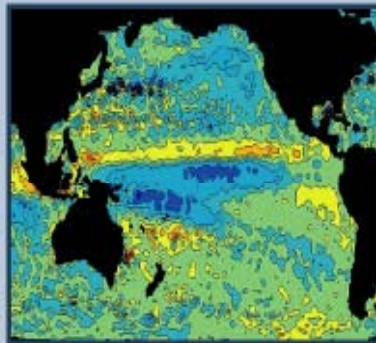
Advanced Sensors



Terabytes 10^{12}

Calibration, Transformation To Characterized Geo-physical Parameters

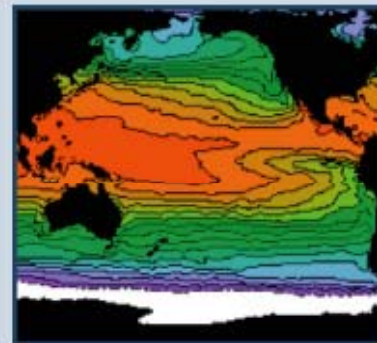
Data Processing & Analysis



Gigabytes 10^9

Interaction Between Modeling/Forecasting and Observation Systems

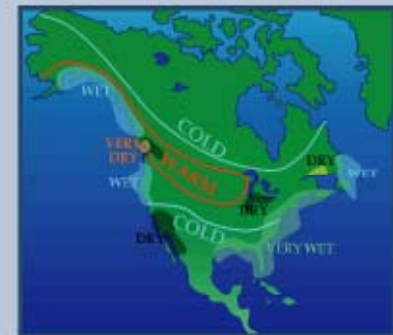
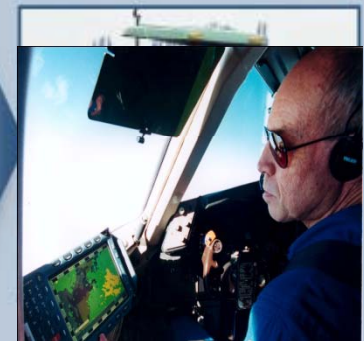
Information Synthesis



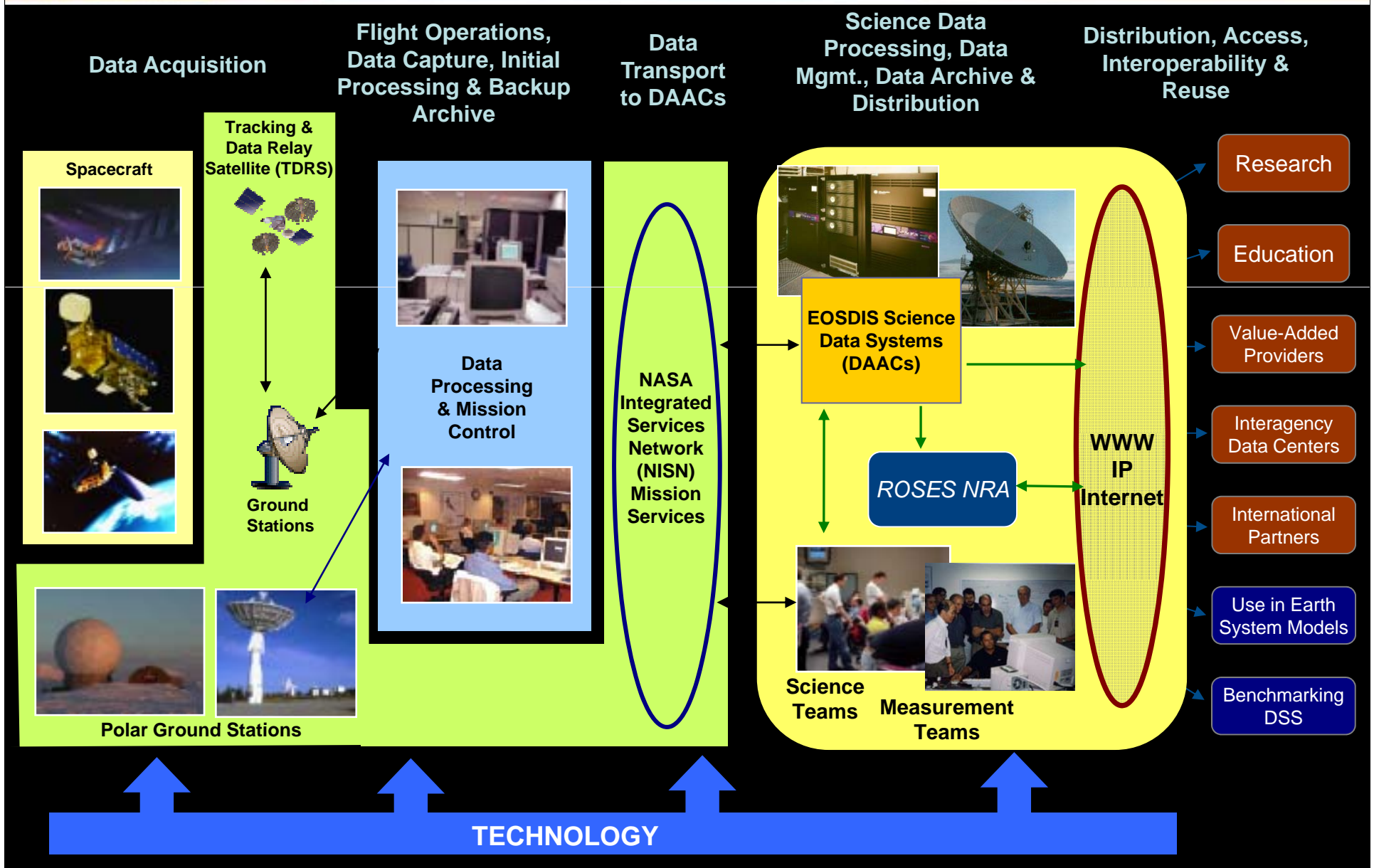
Megabytes 10^6

Interactive Dissemination and Predictions

Access to Knowledge



Data Acquisition to Data Access



Where Can We Research/Play with Big Data Problems?

- Ian Foster talked about data processing workflow all the way to data distribution but why stop there?

Small science is struggling



More data, more complex data
Ad-hoc solutions
Inadequate software, hardware
Data plan mandates

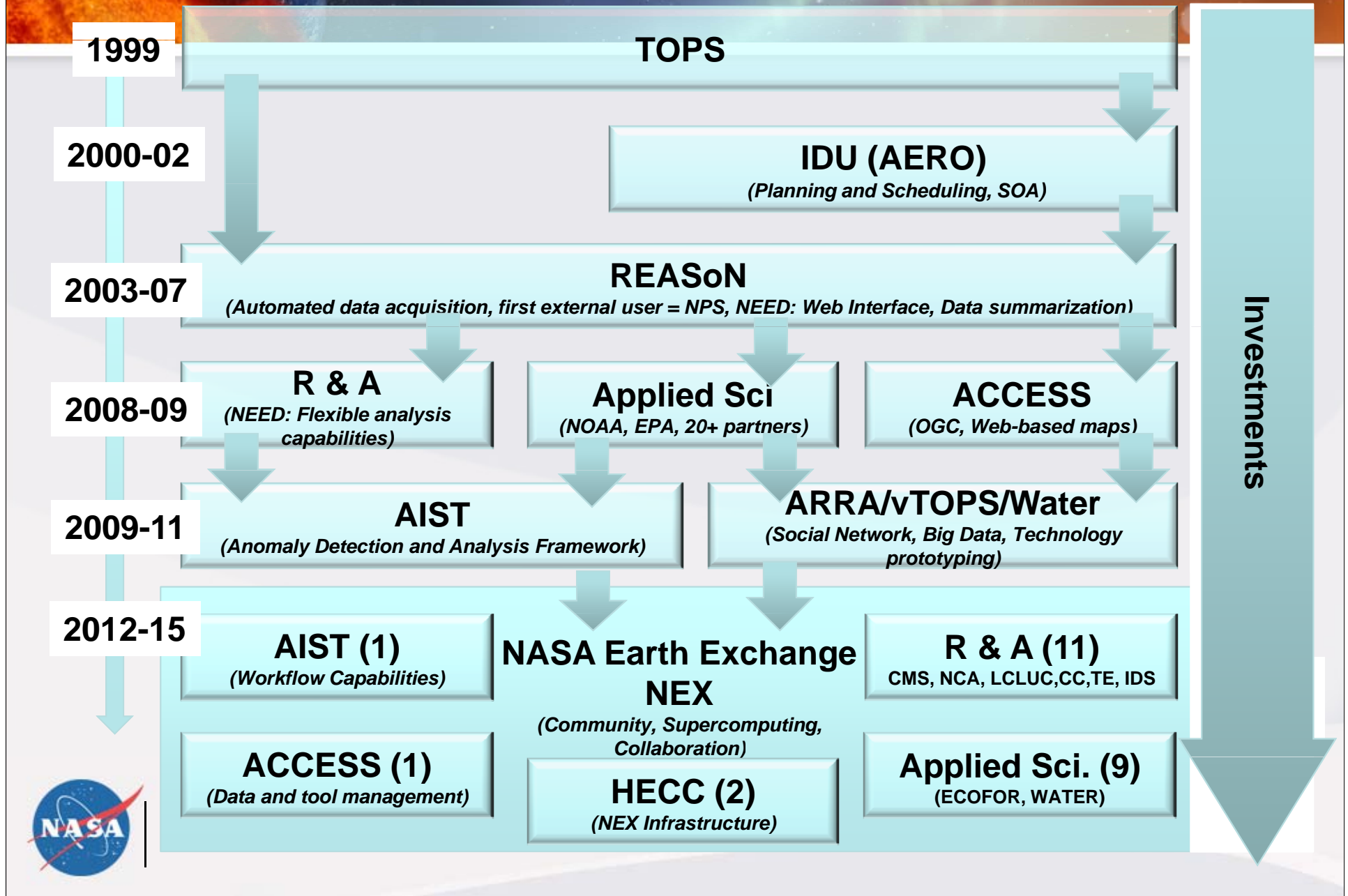
9

Argonne
THE UNIVERSITY OF CHICAGO
www.slac.stanford.edu
www.luchicago.edu

- Michael Stonebraker talked about moving compute to data but there is a tremendous challenge. *(I can't even do it at one NASA center).*



NEX Precursors and Development History



NEX: the Big Data experiment and work completed to date

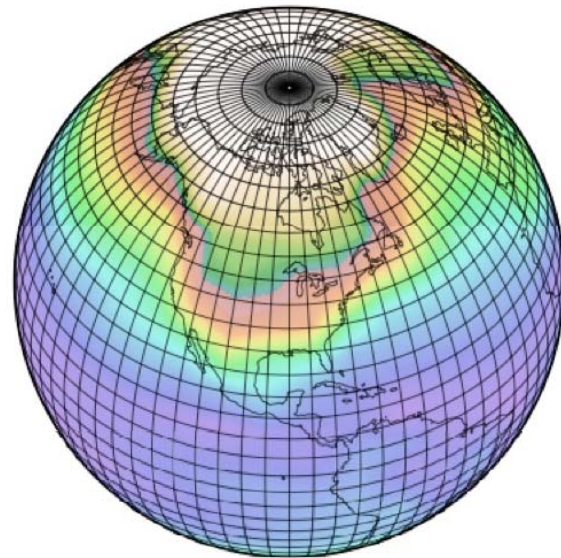
Access to community/knowledge (240 members)

Access to ready-to-use data (24 products, 350TB)

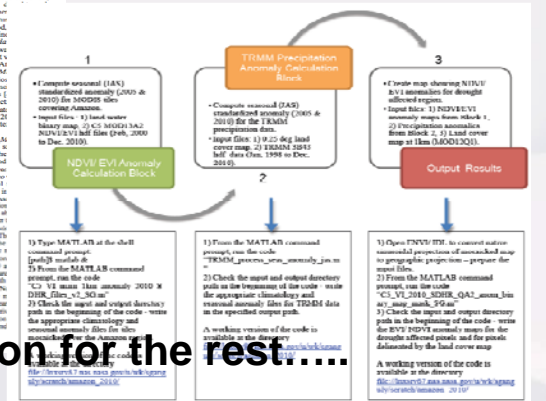


Access to models/analysis tools

Climate, weather, carbon, hydrology



Access to workflows to build upon (VisTrails)



Ready for sophisticated users, needs proper integration for the rest....

EARTH SCIENCE COMMUNITY USERS

- Access to Portal

NEX SCIENCE USERS

- Access to compute resources

NEX HPC USERS

- Access to NAS
supercomputing resources

Portal

Point of entry to NEX collaborative environment

- Project Information
- Collaboration and Social Networking
- Document Publication
- Resource Requests
- Data Discovery

Runs on NAS web servers

Sandbox

Virtualized NEX compute environment

Domain Platform

- Workflow Management
- Provenance
- Rich semantic search
- Data/Model/Tool access (API)

Infrastructure Platform

- Virtualization Support
- Model and Analytic Tool Execution

Data Management

- Data acquisition and pre-processing
- Data storage

Runs on dedicated NEX servers and storage

NAS HPC

Environment for Computing at scale

- Execution of Jobs migrated from Sandbox
- Storage of results in NEX Data Management environment

Runs on NAS supercomputers and storage

**Component
Architecture**



NEX Portal

The screenshot shows the NEX NASA Earth Exchange website. At the top, there is a navigation bar with the NEX logo, the text "NASA Earth Exchange", a NASA logo, a search bar, and a "Search" button. Below the navigation bar are tabs for "HOME", "RESEARCH AREAS", "PROJECTS", "RESOURCES", and "ABOUT". A secondary navigation bar includes links for "Add New Resources", "Create New Project", "Invite Others", and "Feedback".

The main content area features a large banner with the text "Dashlink is a public collaborative tool for data mining communities who are interested in aviation systems health". Below this banner are several sections:

- Explore DASHLink:** A section with sub-sections for "Groups", "Projects", "Algorithms", and "Research Areas".
- how do I start?:** A list of instructions: "Register or login using your nasa account.", "Browse and Download free algorithms, data sets, or publications.", "Find projects to work on.", and "Connect with people."
- how can I contribute?:** A section explaining that users don't need to be registered to browse or download publicly available data, but registration allows for uploading their own.
- who is DASHLink for?:** A section stating that it is for everybody interested in data mining from any scientific community.
- Latest NEWS:** A section titled "Power Planning and Analysis Tool (PLATO) Project Prototype Release" with a "read more" link.
- Latest Activities:** A list of recent events: "John Doe just joined DashLink (March 10, 2010)", "New data set added by Tyler D (Feb 15, 2010)", and "New algorithm added by Newton (March 3, 2010)".
- Most Popular:** A section listing popular items like "TEXT MINING 2007 SIAM 2007 Text Mining Competition Data Sets" and "Orca".
- Featured:** A section with two featured items, both titled "LI-ion Battery Aging Datasets", each with a thumbnail image, creator information, and a short description.
- Sister Sites:** A section listing "EcoCast".

At the bottom of the page, there is a NASA logo and a footer with links for "Privacy Setting", "FAQ", and "Contact Us".

Search capabilities

Who is doing what where
Science network through
abstracts and papers

Who's the expert

Workflows

Archived seminars

Reporting capabilities

Annual reports

Highlights

Publications

Spatial distribution of funding

Virtual Institute



- Following the lead from Astrobiology and NASA Lunar Science Institutes to create a virtual institute, and offer
 - Summer short courses
 - Seminars
 - Conferences
 - Presentations
 - Have each funded project do one or two seminars that can be archived



Science 2007

BREVIA

Amazon Forests Green-Up During 2005 Drought

Scott R. Saleska,^{1*} Kamel Didan,^{2*} Alfredo R. Huete,² Humberto R. da Rocha³

Large-scale numerical models that simulate the interactions between changing global climate and terrestrial vegetation predict substantial carbon loss from tropical ecosystems (1), including the drought-induced collapse of the Amazon forest and conversion to savanna (2).

Resolution Imaging Spectroradiometer (MODIS) content that does not saturate, even over dense forests. Properly filtered to remove atmospheric aerosol and cloud effects, EVI tracks variations in canopy photosynthesis, as confirmed by ecosystem flux measurements on the ground (3, 4).

and C). Much of the smaller area exhibiting decline is heavily affected by human activity or consists of different vegetation types (fig. S2).

Increased greenness is inconsistent with expectation if trees are limited by water but follows from increased availability of sunlight (due to decreased cloudiness) when water is not limiting—if, for example, trees are able to use deep roots and hydraulic redistribution to access and retain water availability.

These observations suggest that Amazon forests may be more resilient to short-term climate perturbations than previously thought. However, the extent to which Amazon forests are resilient to longer-term drought (5), such as may be induced by strong El Niño events or longer-term climate change.

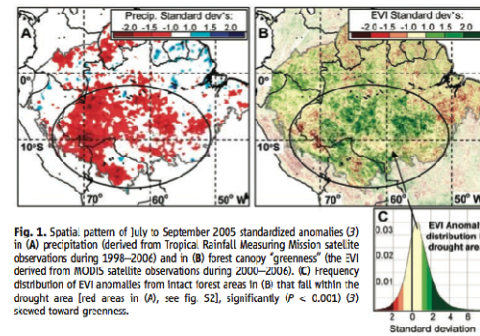


Fig. 1. Spatial pattern of July to September 2005 standardized anomalies (3) in (A) precipitation (derived from Tropical Rainfall Measuring Mission) satellite observations during 1998–2006 and in (B) forest canopy 'greenness' (the EVI derived from MODIS satellite observations during 2000–2006). (C) Frequency distribution of EVI anomalies for intact forest areas in (B) that fall within the drought area [red areas in (A), see fig. S2], significantly ($P < 0.001$) (3) skewed toward greenness.

Model-simulated forest collapse is a consequence not only of climate change-induced drought but also of amplification by the physiological response of the forest: Water-limited vegetation responds promptly to initial drought by reducing transpiration (and photosynthesis), which in turn exacerbates the drought by interrupting the supply of water that would otherwise contribute to the recycled component of precipitation (2). This physiological feedback mechanism should be observable as short-term reductions in transpiration and photosynthesis in response to drought under current climates.

We used satellite to observe whether an Amazon drought in fact reduced whole-canopy photosynthesis (3). The enhanced vegetation index (EVI) from the Terra satellite's Moderate

A widespread drought occurred in the Amazon in 2005 (5), the first such climatic anomaly since the launch of the Terra MODIS sensor in 1999, providing a unique opportunity to compare actual forest drought response to expectation at large scales.

Through intensity peaked during dry season onset (July to September), primarily in southwest and central Amazonia (Fig. 1A) [the drought's temporal evolution is depicted in (5)]. If drought had the expected negative effect on canopy photosynthesis, it should have been especially observable during this period, when anomalous interannual drought coincided with the already seasonally low precipitation. The observations of intact forest canopy 'greenness' in the allusion areas, however, are dominated by a significant increase ($P < 0.001$) (3) not a decline (Fig. 1, B

- References and Notes
1. P. Friedlingstein et al., *J. Climate*, **19**, 3589 (2006).
 2. R. A. Betts et al., *Theor. Appl. Climatol.*, **7**, 200 (2004).
 3. Materials and methods are available on the Online.
 4. R. Huete et al., *Geophys. Res. Lett.*, **33**, L20403 (2006).
 5. S. E. O. C. Aragão, Y. Malhi, R. M. Román-Collado, S. Saatchi, Y. E. Shimoda, *Geophys. Res. Lett.*, **34**, L07101 (2007).
 6. D. C. Nepstad et al., *Nature*, **372**, 644 (2005).
 7. N. M. Coariza, V. G. Garbajosa, *Hydrobiol.*, **564**, 11–10 (2007).
 8. D. C. Nepstad, I. G. Sobrinho, B. Ray, P. M. S. Cardozo, *Ecology*, **88**, 2259 (2007).
 9. Supported by NASA grants NA00OAR04 and Large-Scale Biosphere–Atmosphere Experiment in Amazonia (Biological and NS044920C (00000000)).
 10. We thank M. Keller, S. C. Wright, N. R. Coariza, S. Cardozo, and the anonymous reviewers for comments.

Supporting Online Material
www.sciencemag.org/cgi/content/full/318/5963/3061
Materials and Methods
Figs. S1 to S3
18 June 2007; accepted 30 August 2007
Published online 20 September 2007;
10.1126/science.1141663
Include this information when citing this paper.

¹Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA; ²Department of Soil, Water, and Environmental Science, University of Arizona, Tucson, AZ 85721, USA; ³Department of Atmospheric Science, University of São Paulo, São Paulo, SP, Brazil, CEP 05389-000.
*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: saleska@emlab.arizona.edu

Would have taken less than a week in NEX, Instead of one year it took to repeat the analysis

Efficient use of resources
Lowering barriers to entry
Interdisciplinary work
Transparency, repeatability, extensibility
Cost reduction
Travel
Hardware
IT personnel
Data acquisition
Network costs

Click Here for Full Article

GEOPHYSICAL RESEARCH LETTERS, VOL. 37, L05401, doi:10.1029/2009GL042154, 2010

GRL 2010

Amazon forests did not green-up during the 2005 drought

Arindam Samanta,¹ Sargram Ganguly,² Hirofumi Hashimoto,³ Sadashiva Devadiga,⁴ Eric Vermote,⁵ Yuri Knyazikhin,¹ Ramakrishna R. Nemani,⁶ and Ranga B. Myneni¹

2009; accepted 26 January 2010; published 5 March 2010

Amazon rainforests to dry-season poorly understood, with reports of mortality and forest fires on one hand, and greening on the other. Here, we report results of large-scale greening of the Amazon forest from an earlier version of satellite-derived greenness data - Collection 4 (C4) Enhanced Vegetation Index (EVI), with version as well as the improved, current version to inclusion of atmosphere-corrected data. We find no evidence of large-scale greening at Amazon Forests during the 2005 drought. Approximately 11%–12% of these drought-stricken forests display greening, while, 28%–29% show browning or no-change, and for the rest, the data are not of sufficient quality to characterize any changes. These changes are also not unique - approximately similar changes are observed in non-drought years as well. Changes in surface solar irradiance are contrary to the speculation in the previously published report of enhanced sunlight availability during the 2005 drought. There was no co-relation between drought severity and greenness changes, which is contrary to the idea of drought-induced greening. Thus, we conclude that Amazon forests did not green-up during the 2005 drought. Citation: Samanta, A., S. Ganguly, H. Hashimoto, S. Devadiga, E. Vermote, Y. Knyazikhin, R. R. Nemani, and R. B. Myneni (2010), Amazon forests did not green-up during the 2005 drought, *Geophys. Res. Lett.*, **37**, L05401, doi:10.1029/2009GL042154.

The availability of a new and improved version of satellite-derived vegetation greenness data set - Collection 5 (C5) Enhanced Vegetation Index (EVI) - facilitates a reconciliation of the aforesaid reports for two reasons. First, the C5 EVI data were generated from significantly improved algorithms and input-data filtering schemes related to clouds and aerosols that otherwise corrupt EVI data [Dian and Huete, 2006] - aerosols from biomass burning are widespread in the Amazon during the dry season [e.g., Eck et al., 1998; Schafer et al., 2002], and aerosol loads were significantly higher, compared to other years, during the dry season of 2005 [Koren et al., 2007; Bevan et al., 2009]. Second, this data set spans a longer time period (2000–2008). Our analysis here is focused on answering the following five questions: (a) are the results published by SDHR07 reproducible with both the current and previous versions of EVI data? (b) What fraction of the intact forest area impacted by the drought exhibited anomalous greening in year 2005? (c) Is there evidence of higher than normal amounts of sunlight during the 2005 drought, which may have somehow caused the forests to green-up, as speculated by SDHR07? (d) If drought caused the forests to green-up, is there a relationship between the severity of drought and the spatial extent or magnitude of greening? (e) Are greenness changes during the 2005 drought unique compared to changes in non-drought years?

1. Introduction

The Amazon forests store significant amount of carbon, by some estimates as much as 100 billion tons [Malhi et al., 2006], in their woody biomass. Should these forests die due to moisture stress in a progressively warming climate and savannas replace them, as some studies have suggested [e.g., Cox et al., 2004; Salazar et al., 2007; Huntingford et al., 2008], the carbon released to the atmosphere will act to accelerate global climate changes significantly [Cox et al., 2000]. However, the drought sensitivity of these forests is poorly understood and currently under debate. Extreme droughts such as those associated with the El Niño Southern Oscillation (ENSO), when the plant-available soil moisture stays below a critical threshold level for a prolonged period, are known to result in higher rates of tree mortality and increased forest flammability [Nepstad et al., 2004, 2007]. The drought of 2005, however, was unlike the ENSO-related droughts of 1983 and 1998 - it was especially severe during the dry season in southwestern Amazon but did not impact the central and eastern regions [Marengo et al., 2008]. There are varying reports of forest response to this drought - higher tree mortality and decline in tree growth from ground observations [Phillips et al., 2009] and more biomass fires [Aragão et al., 2007], on the one hand, and excessive greening from satellite observations [Saleska et al., 2007, hereafter SDHR07], on the other. Reconciling these reports remains a priority.

¹Department of Geography and Environment, Boston University, Boston, Massachusetts, USA.
²BAERI, NASA Ames Research Center, Moffett Field, California, USA.

³Department of Science and Environmental Policy, California State University, Monterey Bay, Seaside, California, USA.

⁴Sigma Space Corporation, at NASA Goddard Space Flight Center, Greenbelt, Maryland, USA.

⁵Department of Geography, University of Maryland, College Park, Maryland, USA.

⁶Biospheric Science Branch, NASA Ames Research Center, Moffett Field, California, USA.

Copyright 2010 by the American Geophysical Union.
0094-8276/10/2009GL042154\$5.00

sphere will act to accelerate global climate changes significantly [Cox et al., 2000]. However, the drought sensitivity of these forests is poorly understood and currently under debate. Extreme droughts such as those associated with the El Niño Southern Oscillation (ENSO), when the plant-available soil moisture stays below a critical threshold level for a prolonged period, are known to result in higher rates of tree mortality and increased forest flammability [Nepstad et al., 2004, 2007]. The drought of 2005, however, was unlike the ENSO-related droughts of 1983 and 1998 - it was especially severe during the dry season in southwestern Amazon but did not impact the central and eastern regions [Marengo et al., 2008]. There are varying reports of forest response to this drought - higher tree mortality and decline in tree growth from ground observations [Phillips et al., 2009] and more biomass fires [Aragão et al., 2007], on the one hand, and excessive greening from satellite observations [Saleska et al., 2007, hereafter SDHR07], on the other. Reconciling these reports remains a priority.

The availability of a new and improved version of satellite-derived vegetation greenness data set - Collection 5 (C5) Enhanced Vegetation Index (EVI) - facilitates a reconciliation of the aforesaid reports for two reasons. First, the C5 EVI data were generated from significantly improved algorithms and input-data filtering schemes related to clouds and aerosols that otherwise corrupt EVI data [Dian and Huete, 2006] - aerosols from biomass burning are widespread in the Amazon during the dry season [e.g., Eck et al., 1998; Schafer et al., 2002], and aerosol loads were significantly higher, compared to other years, during the dry season of 2005 [Koren et al., 2007; Bevan et al., 2009]. Second, this data set spans a longer time period (2000–2008). Our analysis here is focused on answering the following five questions: (a) are the results published by SDHR07 reproducible with both the current and previous versions of EVI data? (b) What fraction of the intact forest area impacted by the drought exhibited anomalous greening in year 2005? (c) Is there evidence of higher than normal amounts of sunlight during the 2005 drought, which may have somehow caused the forests to green-up, as speculated by SDHR07? (d) If drought caused the forests to green-up, is there a relationship between the severity of drought and the spatial extent or magnitude of greening? (e) Are greenness changes during the 2005 drought unique compared to changes in non-drought years?

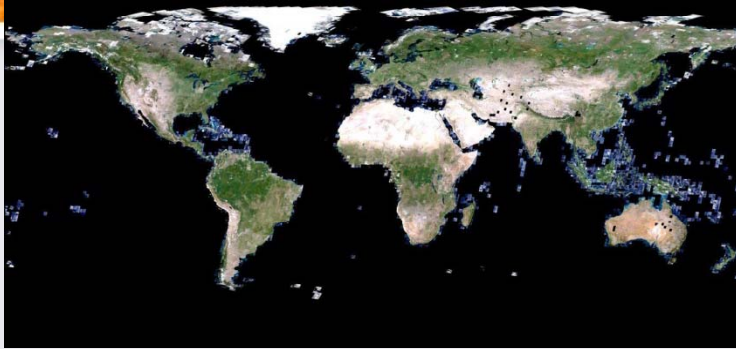
2. Data and Methods

[a] Detailed information on data and methods is provided in the auxiliary material.⁷ “Amazon forests” in this report

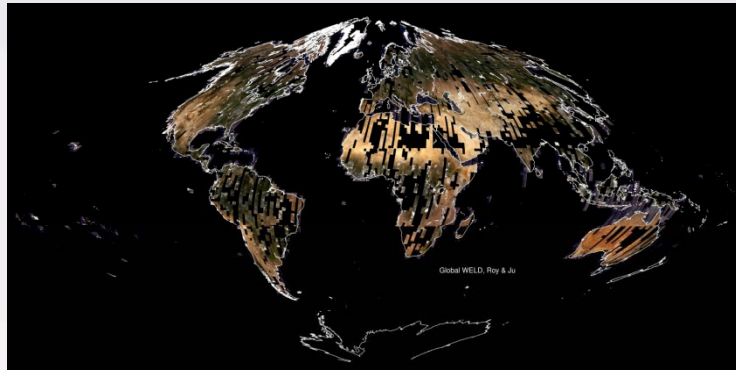
⁷Auxiliary materials are available in the HTML. doi:10.1029/2009GL042154.

Benefits (Unprecedented Opportunities)

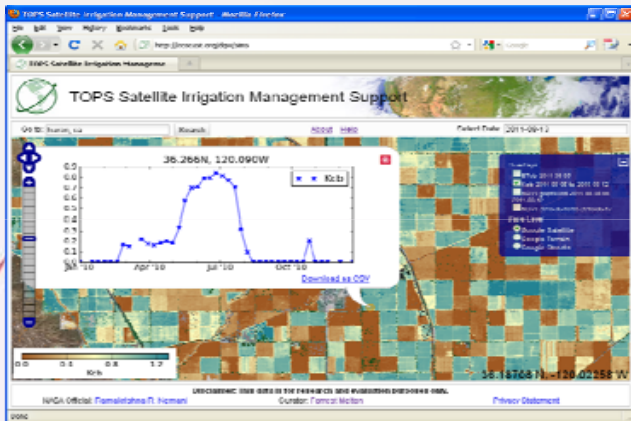
Global Land Survey from Landsat (LCLUC)



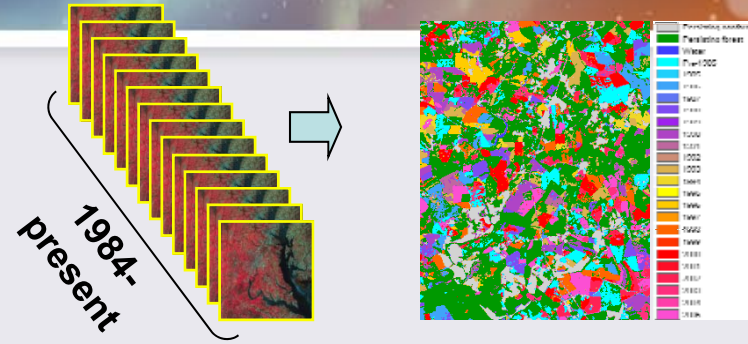
Global monthly Landsat (WELD, MEASURES)



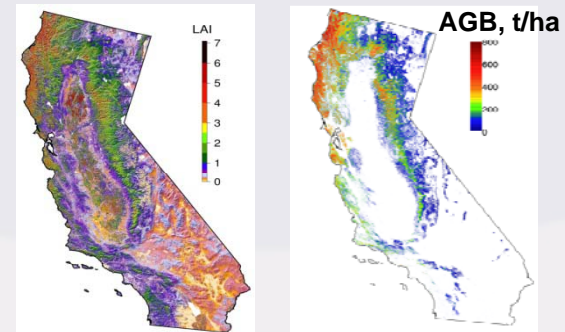
Crop water management with Landsat



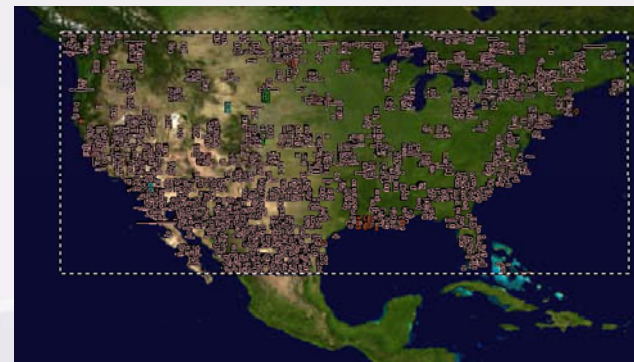
30 years of change analysis (CC)



Biophysical products from Landsat (CMS)

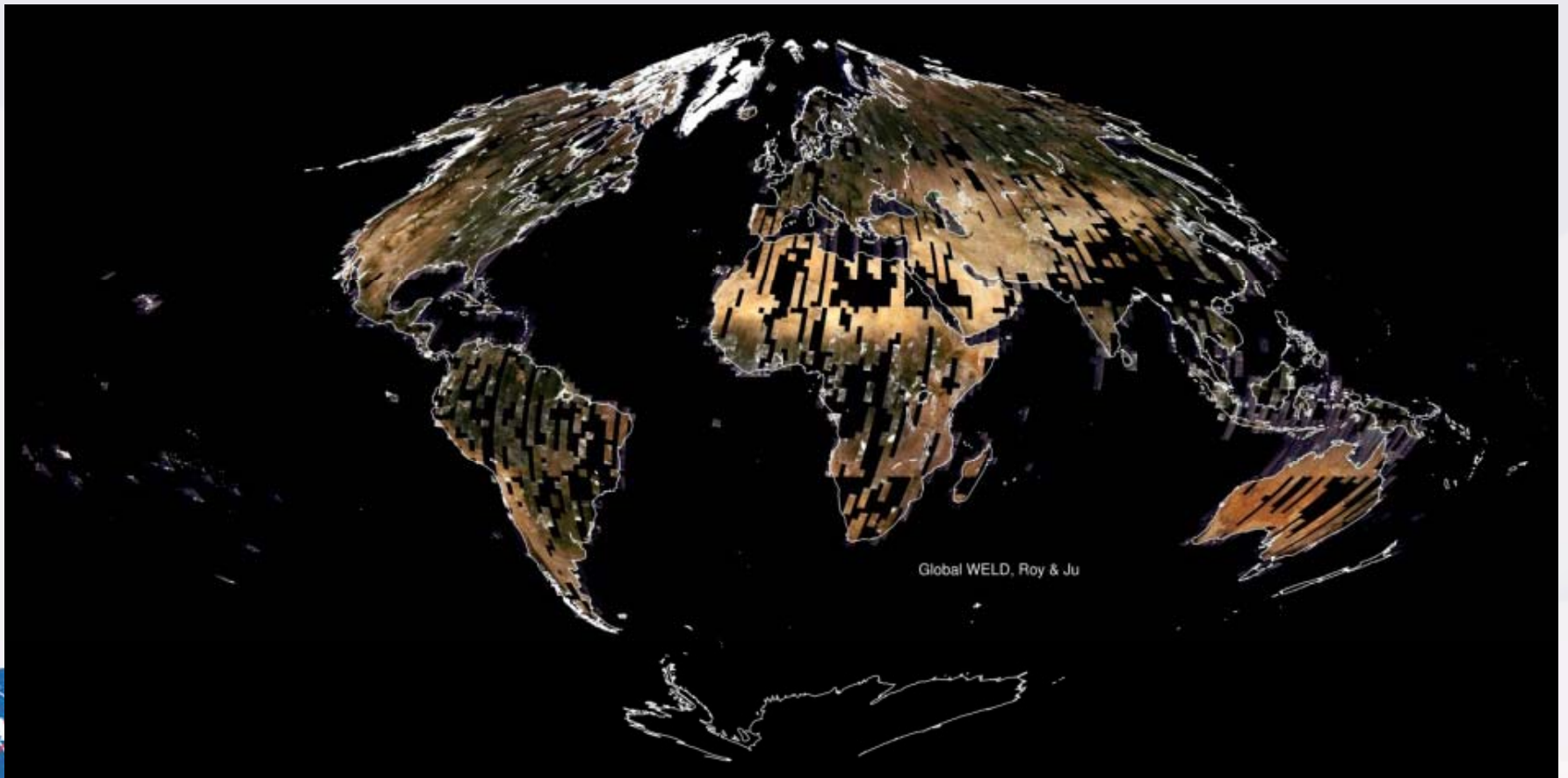


WorldView-2, 50cm, 8 bands (NGA)



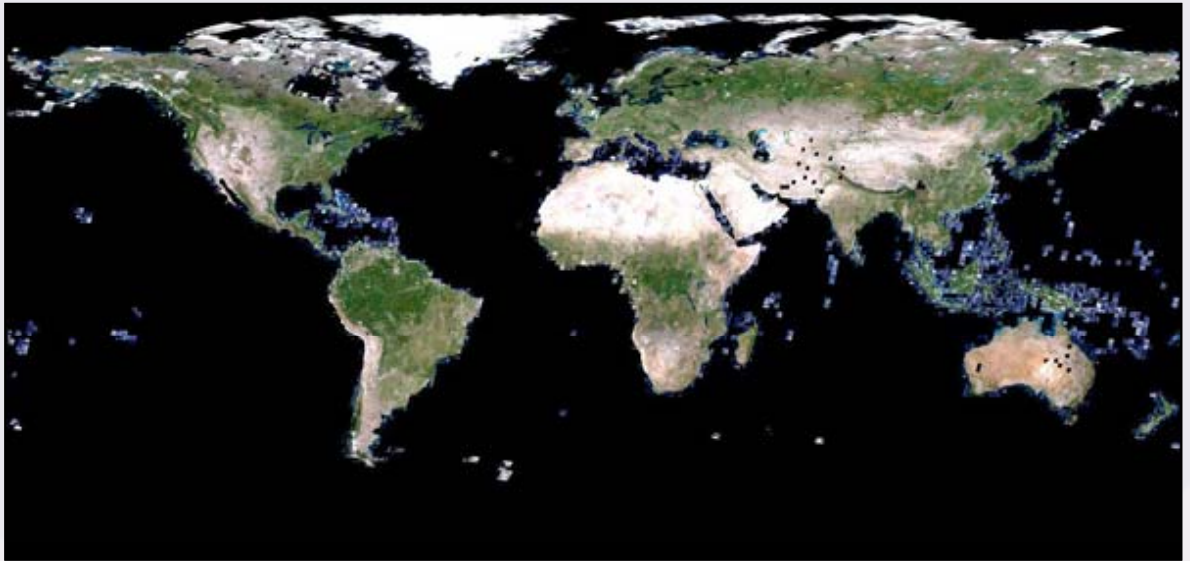
Big Data Metric?

***Space-time-resolution metric
(higher spatial resolution, larger extent,
longer time-series studies)***



NEX Experiments Demonstrate Value of Collaborative Environment

- In a first application of NEX, a research team from around the U.S. used the environment to adjoin and atmospherically correct a mosaic of 9,000 Landsat Thematic Mapper scenes and retrieve global vegetation density at a 30-meter resolution.
- The entire processing of the nearly 340 billion pixels in the composite took just a few hours on the Pleiades supercomputer, allowing the team to experiment with new algorithms and products within just a few days.



- NEX's collaboration and knowledge-sharing platform for the Earth science community combines supercomputing, Earth system modeling, workflow management, and NASA remote sensing data feeds to deliver a complete work environment for users to explore/analyze large datasets, run modeling codes, collaborate, and share results.

*POC: Petr Votava, petr.votava@nasa.gov, (650) 604-4675;
Ramakrishna Nemani, ramakrishna.nemani@nasa.gov,
(650) 604-6185, NASA Ames Research Center*



- Data Management
 - How to load and offload data in a multi-tier storage environment?
 - Distributed Multi-site Analytics?
- Workflow Reuse
 - Why create your own if a similar analytics was done before?
 - Use workflow to capture the data provenance?
- Workflow Discovery
 - Why not mine the literature to uncover the workflow?

