SIAM '04

Accelerating Scientific Discovery at NASA

Joseph C. Coughlan

NASA CICT/Intelligent Systems

Joseph.C.Coughlan@nasa.gov

Presentation to SIAM 2004

Outline

Van Gogh (1853-1890)
The Garden of Saint Paul's Hospital (1889)
Vincent van Gogh Foundation

- Credits
- Overview of data volume
- An Example: The Greening Earth
- A Response: Discovery systems proposal
- Summary
- NASA Research Announcement (NRA)

Presenting work from

Nemani, R.R., C.D. Keeling, H. Hashimoto, W.M. Jolly, S.C. Piper, C.J. Tucker, R.B. Myneni and S.W. Running.

Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science, 300, 1650 (2003).

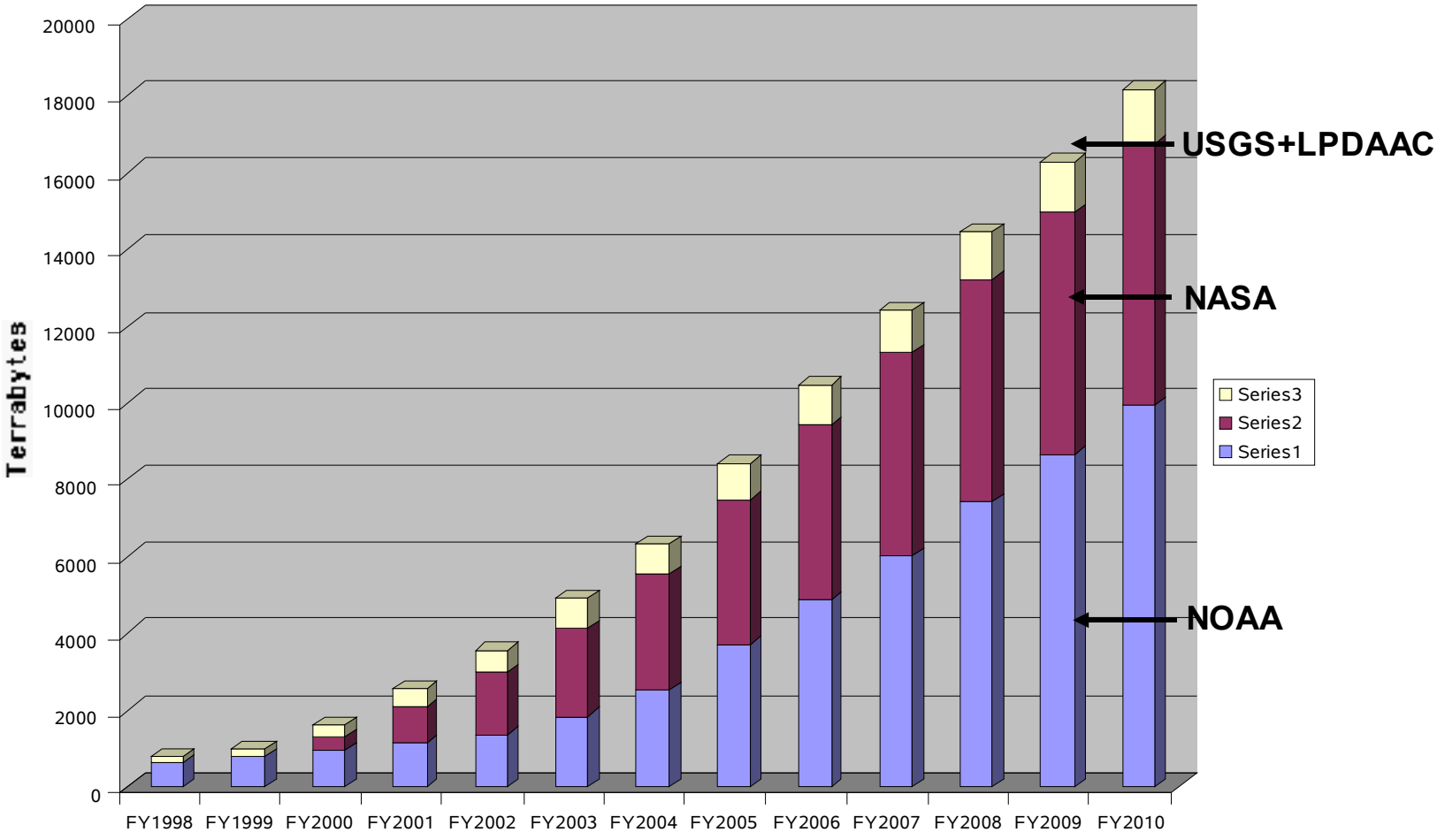
&

Barney Pell, co-lead and the Discovery Systems planning team

With reference to to

R. R. Nemani et al. 2001. Asymmetric warming over coastal California and its impact on the premium wine industry Climate Research Vol. 19: 25–34

National Agency Earth Science Data Holdings



Large Data Flows: Extracting Knowledge from Petabytes of Data

NASA Satellites

~1.5 TB/day

Climate/Weather Data

Ancillary Data

Topography, River Networks, Soils, Biodiversity . . .

10-100 MB/day

EOSDIS

250+ products,
> 2 Petabytes

.1-10 TB

Ecocast Architecture

IMAGEbot Planner

Optimizes data processing plan and retrieves select data for analysis

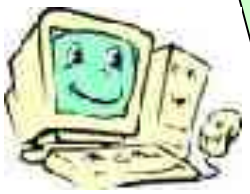
Causal Discovery

Autonomous analysis of data for discovery of novel causal models

TOPS

Biospheric models for ecological nowcasting / forecasting from data

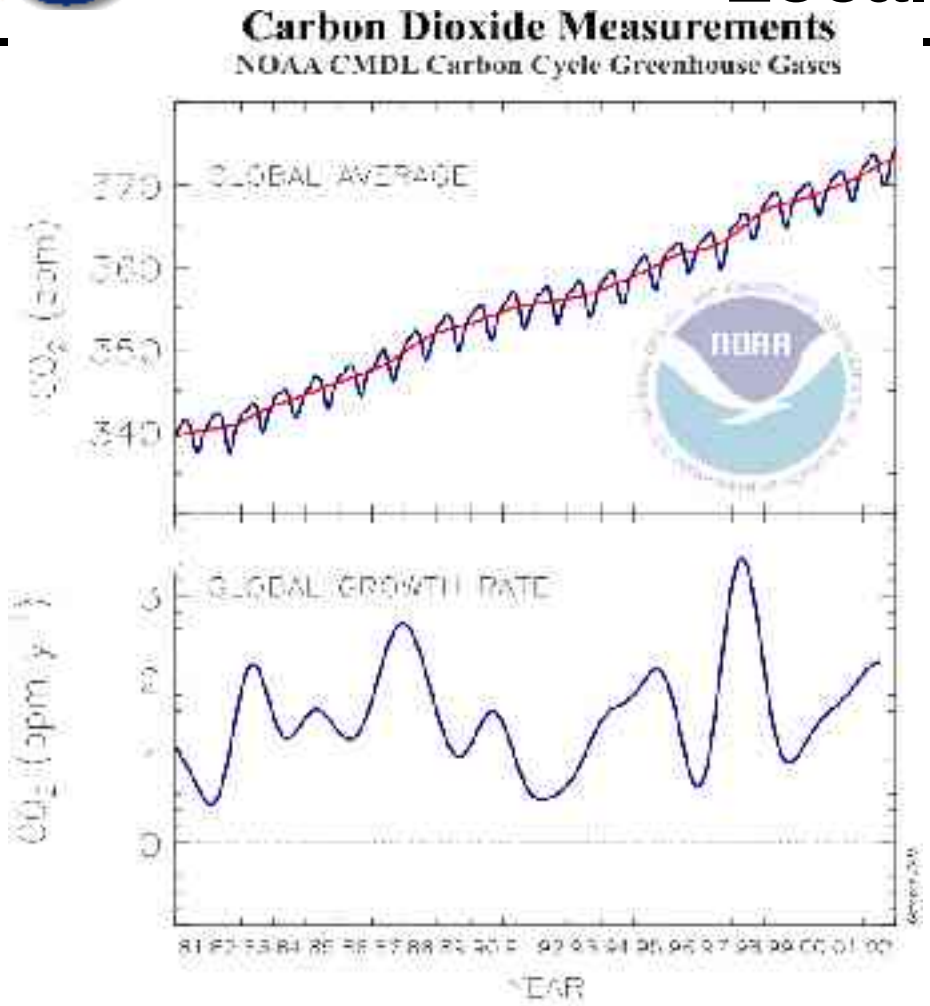
Data Overload:
1-100 TB
Massive data sets, multiple products, heterogeneous data types



Knowledge:
100K to 10 MB
Nowcast and forecast maps, integrated datasets, images, models

NASA Goal:Observation and Understanding and Prediction

Greening Earth: Evidence at Global and Local Scales



Top: Global average atmospheric carbon dioxide mixing ratios (blue line) determined using measurements from the NOAA CMDL cooperative air sampling network. The red line represents the long term trend. Bottom: Global average growth rate for carbon dioxide. Principal investigator: Dr. Pieter Tans, NOAA CMDL Carbon Cycle Greenhouse Gases, Boulder, Colorado, [date] (gmd.hotline@noaa.gov, http://www.cmdl.noaa.gov/ccgg)

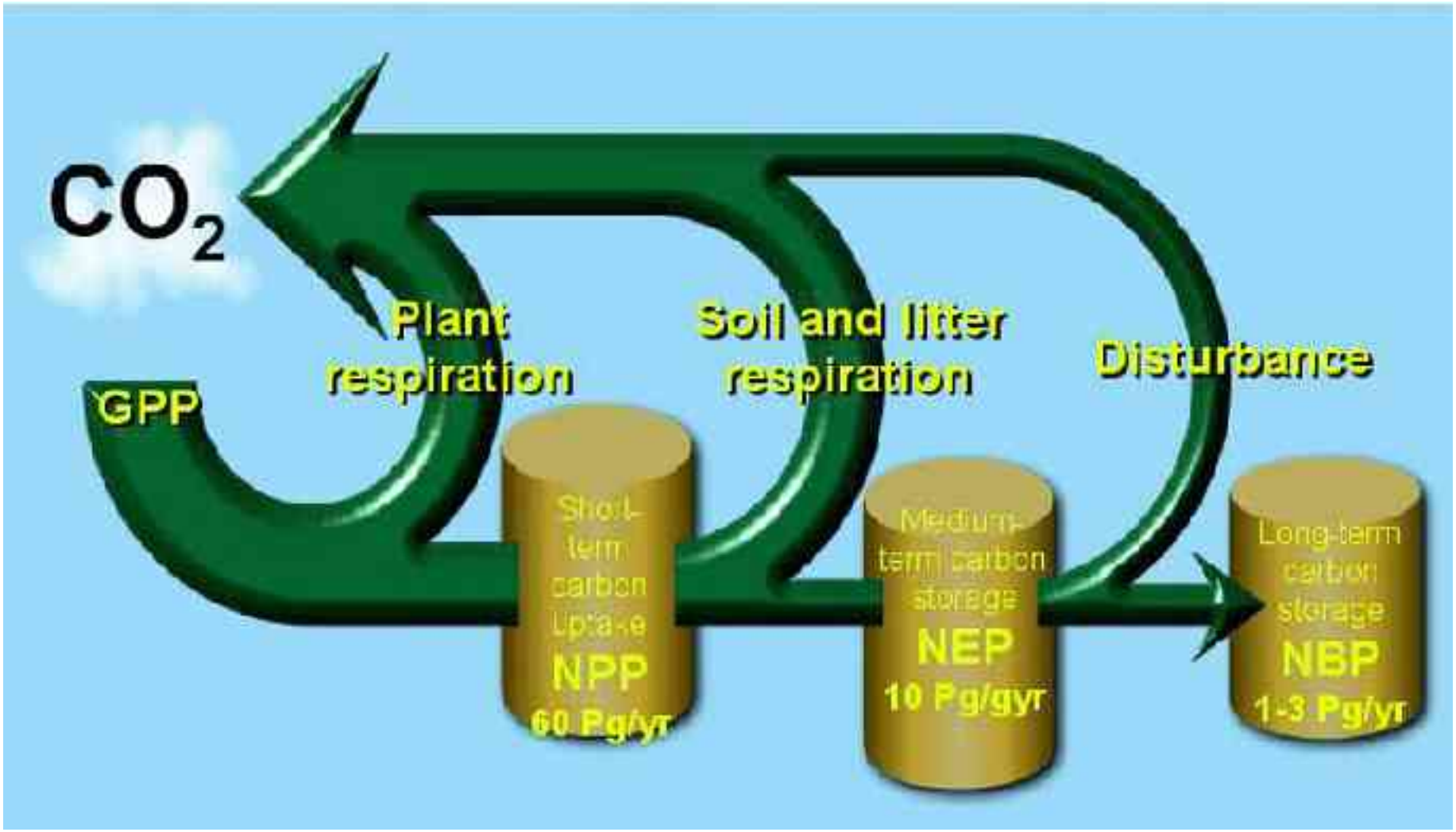
Claude Monet (1840-1926)
Peupliers au bord de l'Epte (1891)
The Tate Collection

Plot level biomass changes

Western Europe, Liski and Kauppi, 2000, UNEP
U.S., Pacala et al., 2001, Science
China, Fang et al., 2001, Science
South America, Phillips et al., 1998, Science

Nearly two decades of satellite observations

NPP and its relation to carbon cycling



Net primary production

Reflectance Spectrum of a Green Leaf

Pigments in green leaves (notably chlorophyll) absorb strongly at red and blue wavelengths. Lack of such absorption at near-infrared wavelengths results in strong scatter from leaves.

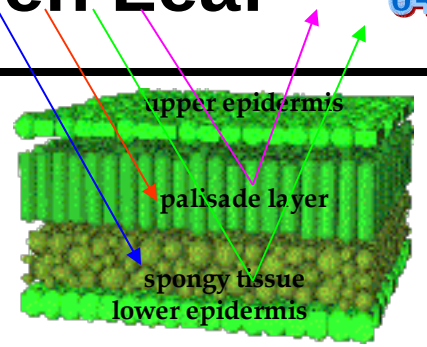
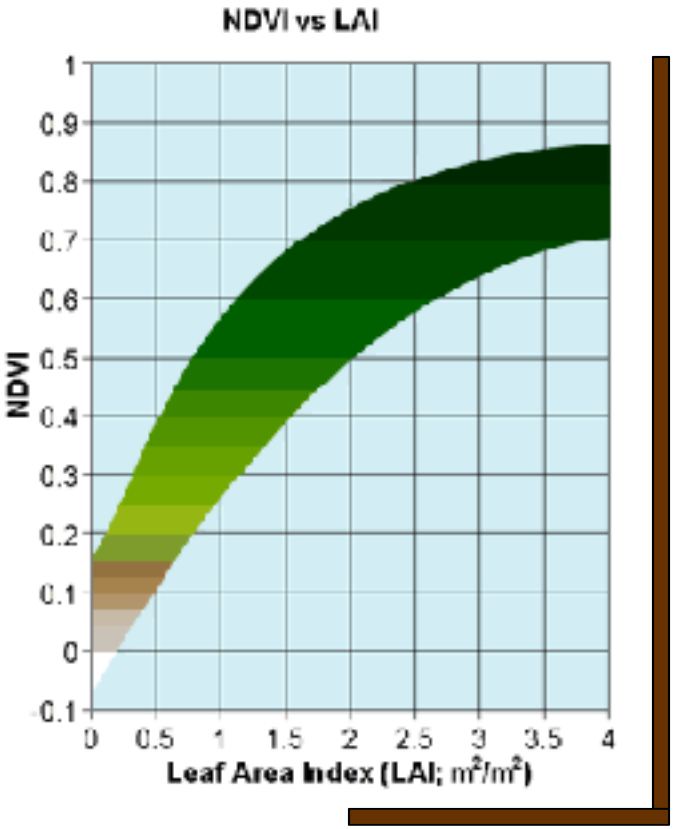
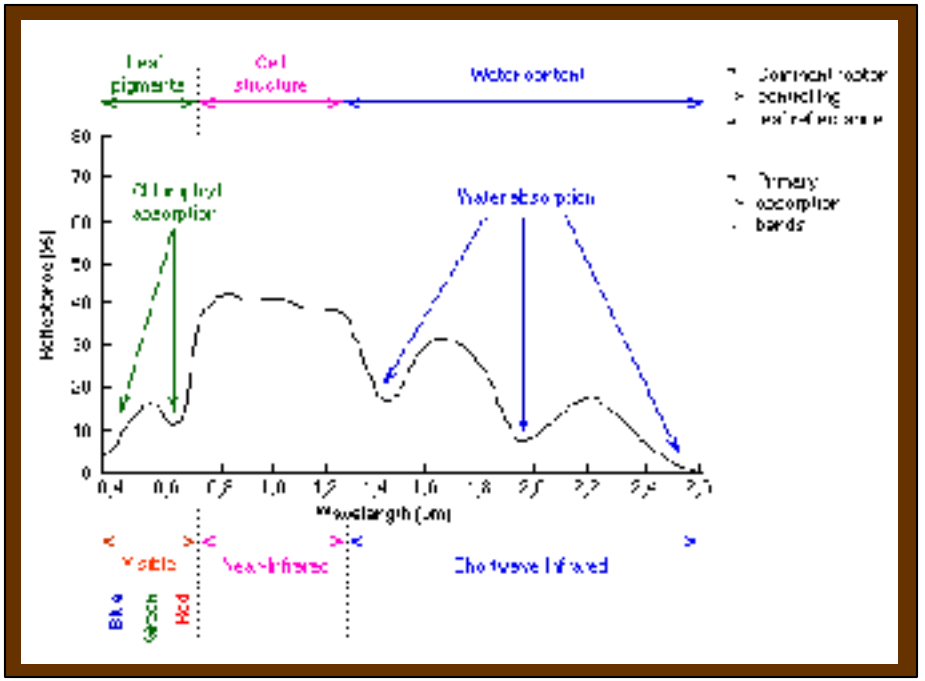


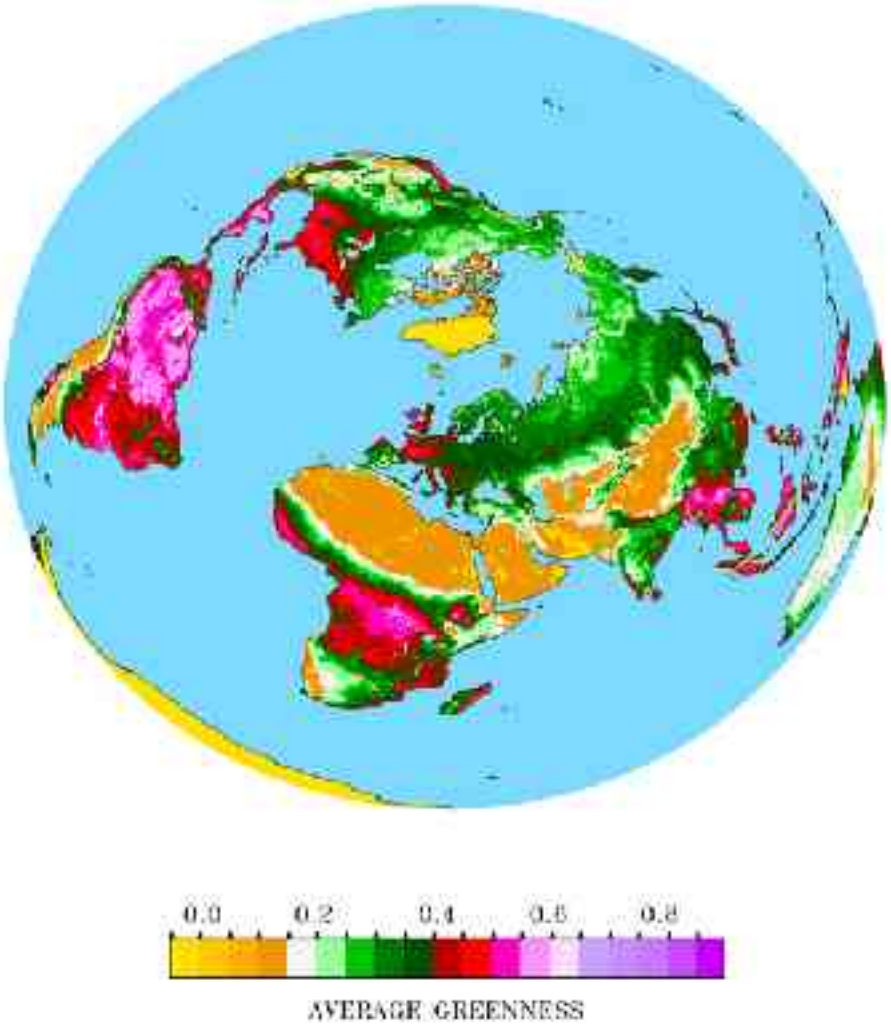
image credit: govaerts et al.





- background on NDVI and NDVI data sets

egetation Index



Rene Magritte (1989-1967)
The Beautiful Season

NPP Evaluation from Satellite Data

Step 1:
convert absorbed radiation to optimal gross production

Step 2:
downgrade by climate limiting factors to obtain gpp

Step 3:
subtract respiration to obtain npp

Joan Miro (1893-1983)
Le Chant De La Prairie (1964)
Miro Museum, Barcelona

Net primary production

Components of the NPP algorithm

satellite-derived vegetation properties: Land cover, Leaf Area Index (LAI) and fraction of absorbed photosynthetically active radiation (FPAR)

daily climate data: incident radiation (IPAR), minimum and average air temperatures and humidity

efficiencies: a biome specific parameterization to convert absorbed PAR to NPP

Methodology of the NPP algorithm

Application of the radiation conversion efficiency logic to predictions of daily gross primary production (GPP), using satellite-derived FPAR and independent estimates of PAR and other surface meteorological fields, and the subsequent estimation of maintenance and growth respiration terms that are subtracted from GPP to arrive at annual NPP.

Maintenance respiration and growth respiration components derived from allometric relationships linking biomass and annual growth of plant tissues to satellite-derived estimates of leaf area index.

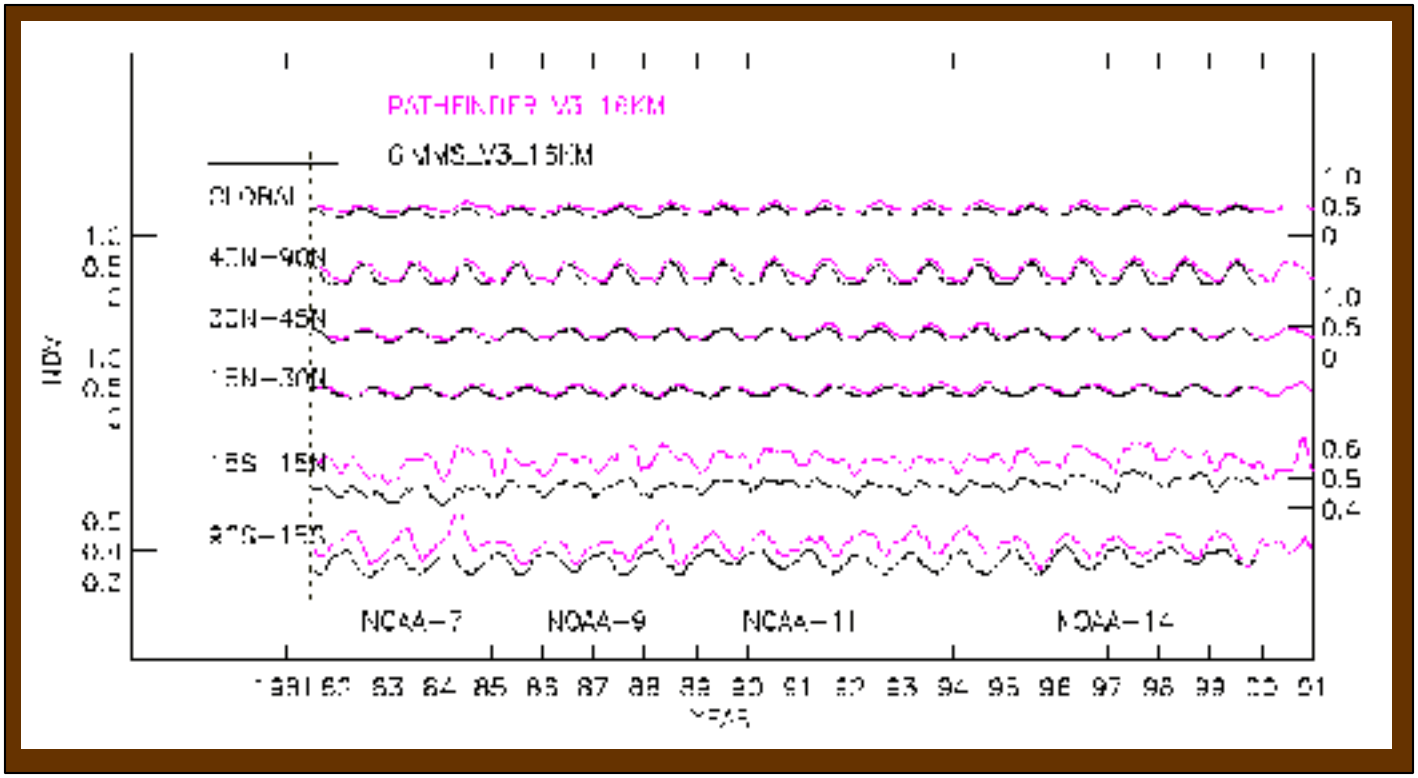
These allometric relationships have been derived from extensive literature review, and incorporate the same parameters used in the Biome-BGC ecosystem process model.

NDVI (Satellite) data set quality & improvement

version 1 PAL data set (as distributed by the daac) has known quality problems

version 1 GIMMS data set (produced in 2000) has similar problems outside of Africa, North America and Eurasia (residual atmospheric effects, orbital loss effects, volcanic aerosols, etc.)

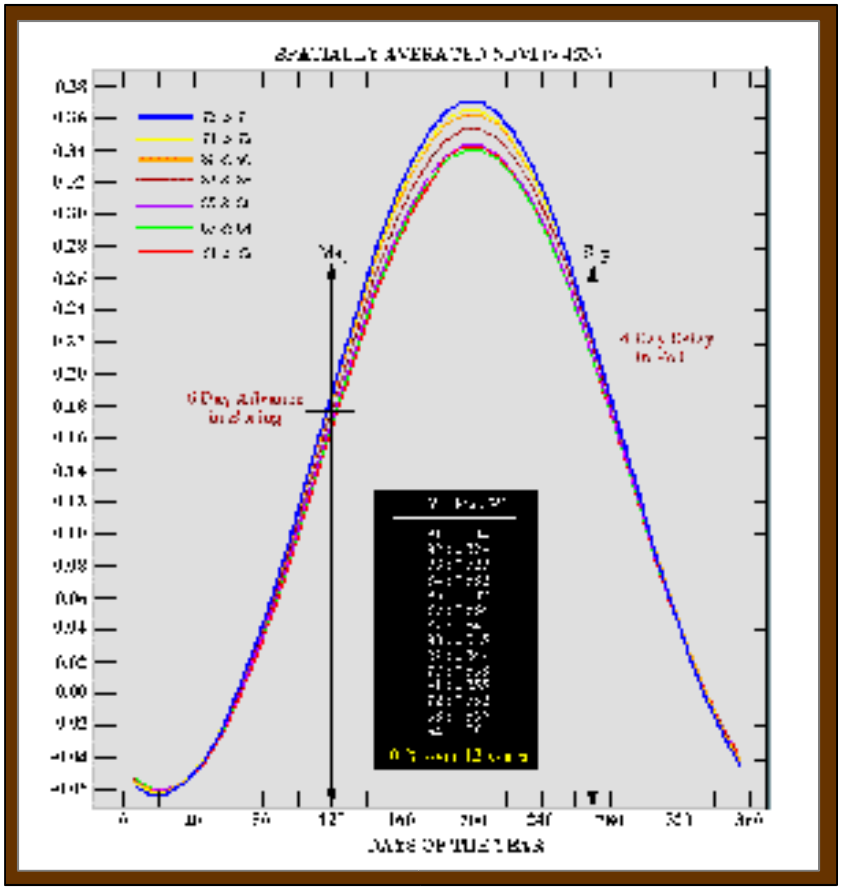
version 3 Data sets developed by us have temporal (15/10 day to 1 month) and spatial compositing (8 km to 16 km), and corrections for orbital loss.



How to proceed?

Guided by
structure and
function.

Longer growing season trend in the north (1980s)



From Myneni et al. (Nature, 386:698-701, 1997)

Analyses of both GIMMS (v0) and PAL (v2) data for the period 1981 to 1994 suggest that -

- NDVI averaged over the peak boreal growing season months of July and august increased by 10%

- The timing of spring green-up advanced by about 6 days

- The satellite data are concordant with an increase in the amplitude of the seasonal cycle of atmospheric CO_2 exceeding 20% since the early 1970s, and an advance in the timing of the draw-down of CO_2 in spring and early summer of up to 7 days (Keeling et al., Nature, 382:146-149, 1996)

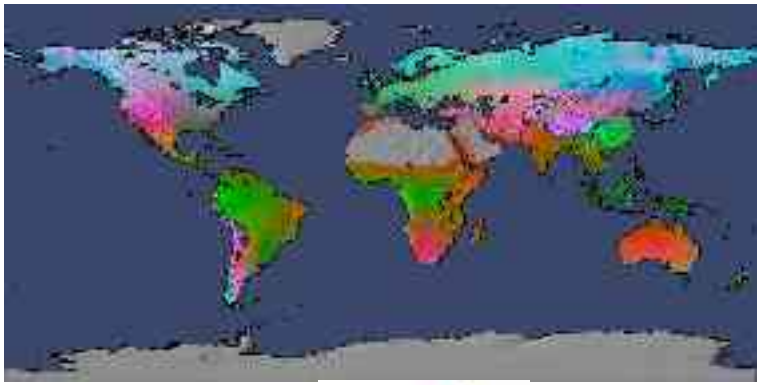
NDVI averaged over boreal growing season months of May to September increased by about 10%, the timing of spring green-up advanced by about 6 days

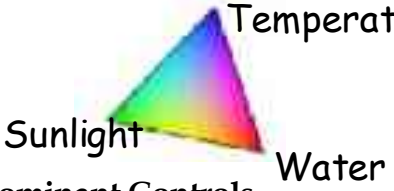
- greening in the north/the northern latitude greening trend during the 1980s & 1990s

Limiting Factors

Plant growth is assumed to be principally limited by sub-optimal climatic conditions such as low temperatures, inadequate rainfall and cloudiness (Churkina and Running, 1998). We used 1960-1990 average climate data (Leemans and Cramer, 1991) to develop scaling factors between 0 and 1 that indicate the reduction in growth potential.

Potential Climate Limits for Plant Growth





Dominant Controls

water availability 40%

temperature 33%

solar radiation 27%

total vegetated area 117 M km2

- black (no limits) and white (all at maximum limit)
- primary colors represent respective maximum limits
- cyan (temperate and radiation) represents cold winters and cloudy summers over eurasia
- magenta (water and temperature) represents cold winters and dry summers over western north america
- yellow (water and radiation) represents wet-cloudy and dry-hot periods induced by rainfall seasonality in the tropics
- these limits vary by season (e.g., high latitude regions are limited by temperature in the winter and by either water or radiation in the summer)

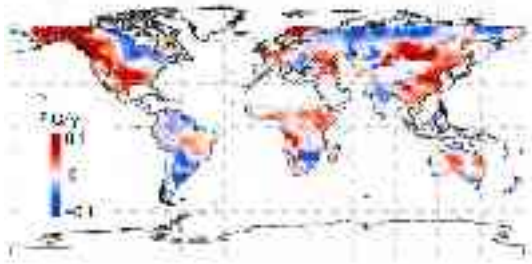
- the greening earth and increasing terrestrial npp

Trends in Climate Data

Potential Climate Limits for Plant Growth

Data

Reanalysis data (6-hourly 2 m height temperatures, 2-m height specific humidity, and incident solar radiation) from the National Center for Environmental Prediction (NCEP) to represent climate variability from 1982 to 1999

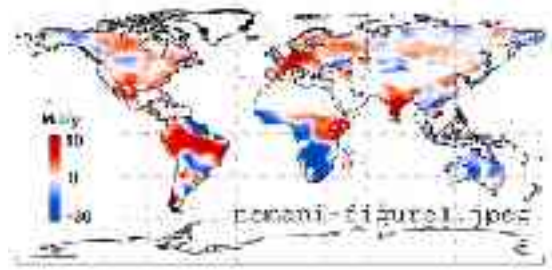


Interannual trends in daily average temperature 1982-99

Spring air temperatures that regulate the initiation of the growing season have increased over temperature-limited regions of North America and Northwest Europe promoting earlier plant growth and additional carbon sequestration.

Interannual trend in vapor pressure deficit 1982-99

Wetter rainfall regimes and the associated reduction in vapor pressure deficits in the 1990s are important for the water-limited ecosystems of Australia, Africa and the Indian sub-continent.



Interannual trend in solar radiation 1982-99

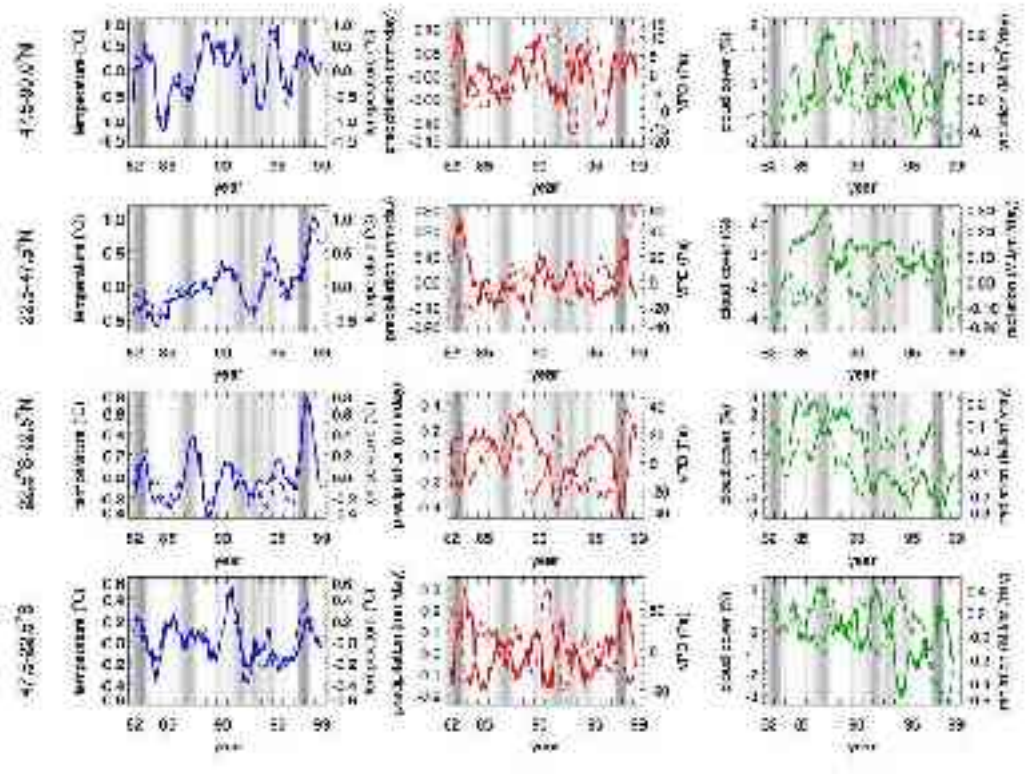
Significant increases in incident solar radiation are evident over radiation-limited regions of Eurasia and the equatorial tropics. While the increased solar radiation over Eurasia may be related to changes in the North Atlantic Oscillation, a decline in cloud cover along with increases in outgoing long-wave radiation, as a consequence of changes in tropical circulation patterns, was recently reported over tropical regions.

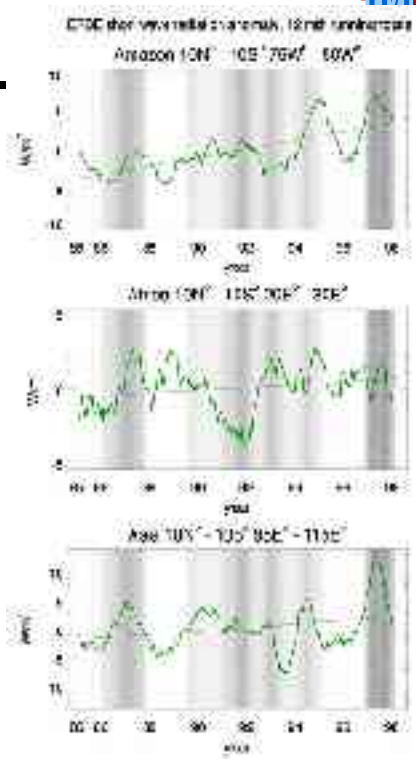
The observed climatic changes have been mostly in the direction of reducing climatic constraints to plant growth. Therefore, it seems likely that vegetation responded to such changes positively.

- the greening earth and increasing terrestrial npp

Trends in Climate Data (confirmation)

Comparisons between NCEP derived variables (dashed) with independent data (solid). NCEP temperatures reproduce the interannual variations in growing season average temperatures estimated from those of **UDEL** over all latitudinal zones. Similarly, higher NCEP derived VPDs correspond to low **CMAP** precipitation. A strong decline in cloud cover observed from **ISSCP** data over tropical regions corresponds to increases in NCEP solar radiation. Growing season is defined as all months with monthly average temperature above 0C.





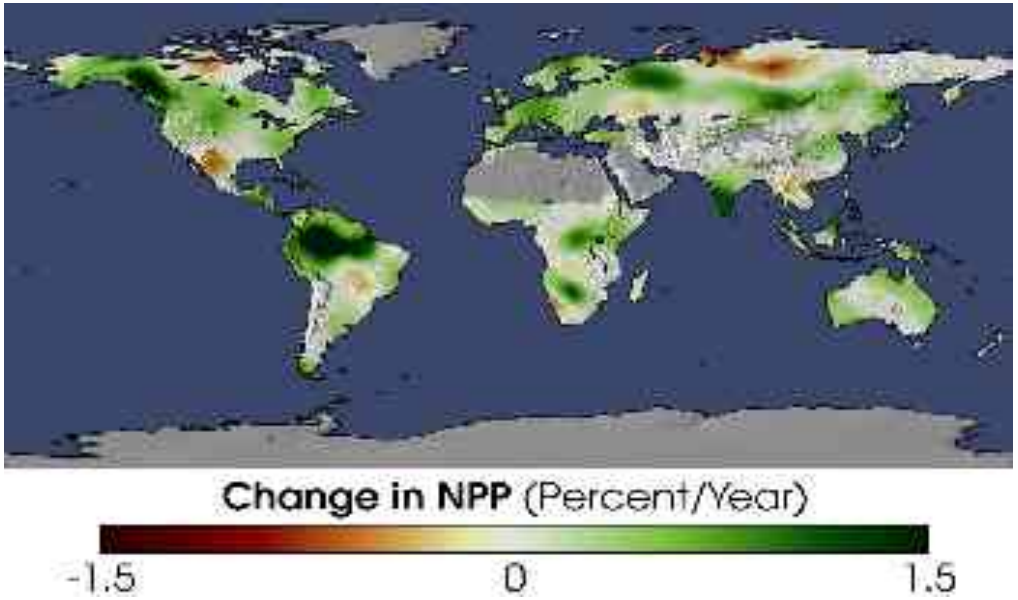
ERBE data, twelve month running totals, from 1985 to 1998 shows that solar radiation increased significantly over the Amazon and that similar increases are not evident in Africa and Asia. NCEP radiation trends show similar patterns.

The trends observed in NCEP data were confirmed with analyses of several other data climate data sets.

- the greening earth and increasing terrestrial npp

The NPP Algorithm

Step 1 convert absorbed radiation to optimal gross production → **Step 2** downgrade by climate limiting factors to obtain gpp → **Step 3** subtract respiration to obtain npp



Average of interannual trends (1982-99) in growing season NPP estimated with GIMMS and PAL (v3) FPAR

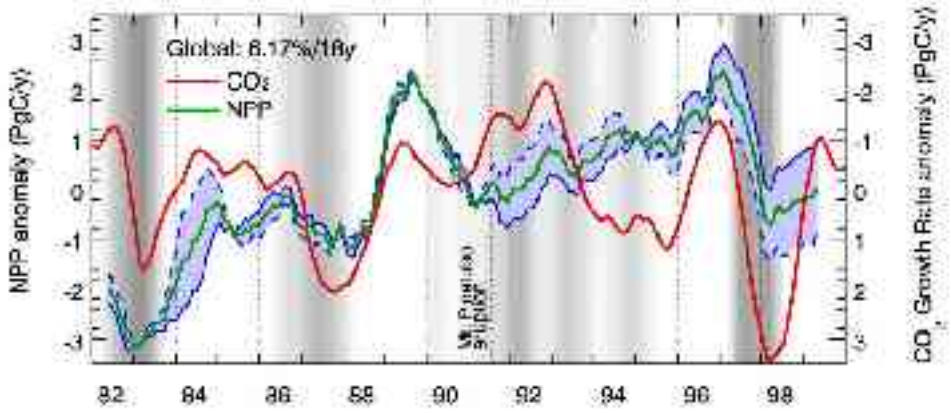
In water and radiation limited regions NPP showed the highest increase (6.5%) followed by those in temperature and radiation (5.7%), and temperature and water (5.4%) limited regions.

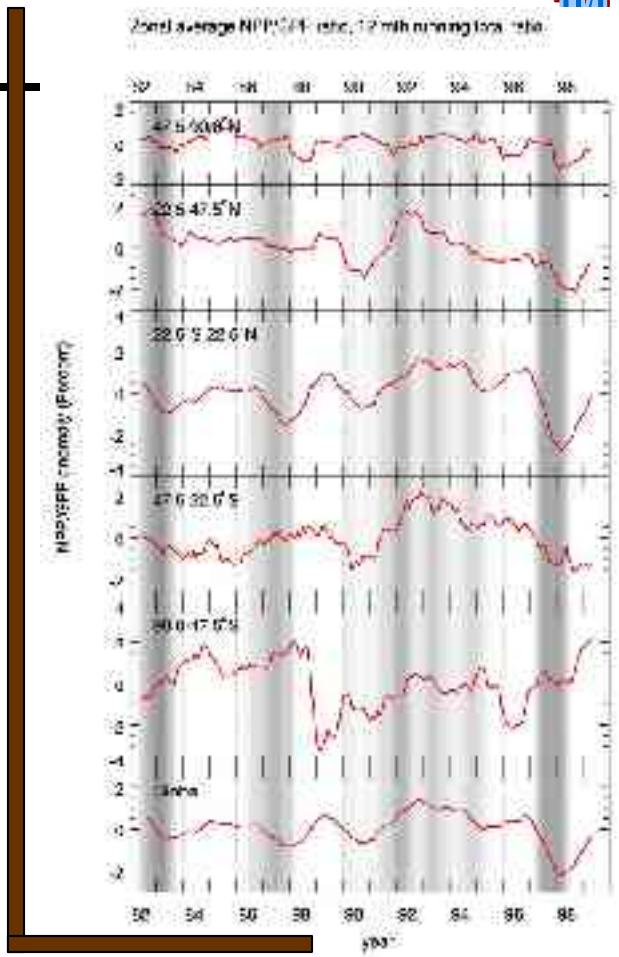
Globally all biomes, except open-shrubs, showed an increasing NPP trend from 1982 to 1999 with the largest increase in evergreen broadleaf forests.

Trends in NPP are positive over 55% of the global vegetated area and are statistically more significant than the declining trends observed over 19% of the vegetated area.

- the greening earth and increasing terrestrial npp

climate, NPP and atmospheric CO_2 growth rate





A moderately increasing trend (6% or 3.42 PgC/18yr, $p<0.001$) in global NPP was observed between 1982 and 1999, suggesting that the terrestrial biosphere may have been actively sequestering carbon in biomass.

Interannual variations in global NPP are correlated with global atmospheric CO_2 growth rates ($r = 0.70$, $p<0.001$).

NPP declined during all three El Niño events.

High CO_2 growth rates during El Niño years correspond to declines in global NPP.

Analyses of variation in the plant photosynthesis-respiration balance, expressed as NPP/GPP ratio (right panel), showed observed declines in NPP during El Niño years to be dominated by increases in respiration due to warmer temperatures.

Although the atmospheric CO_2 growth rate depends on the net air-sea and land-atmosphere exchanges, these Results highlight the preeminent role of plant growth in global carbon cycle.

- the greening earth and increasing terrestrial npp

NPP trends by latitude

Tropical ecosystems accounted for a large portion of both interannual variability and the increasing trend in global NPP.

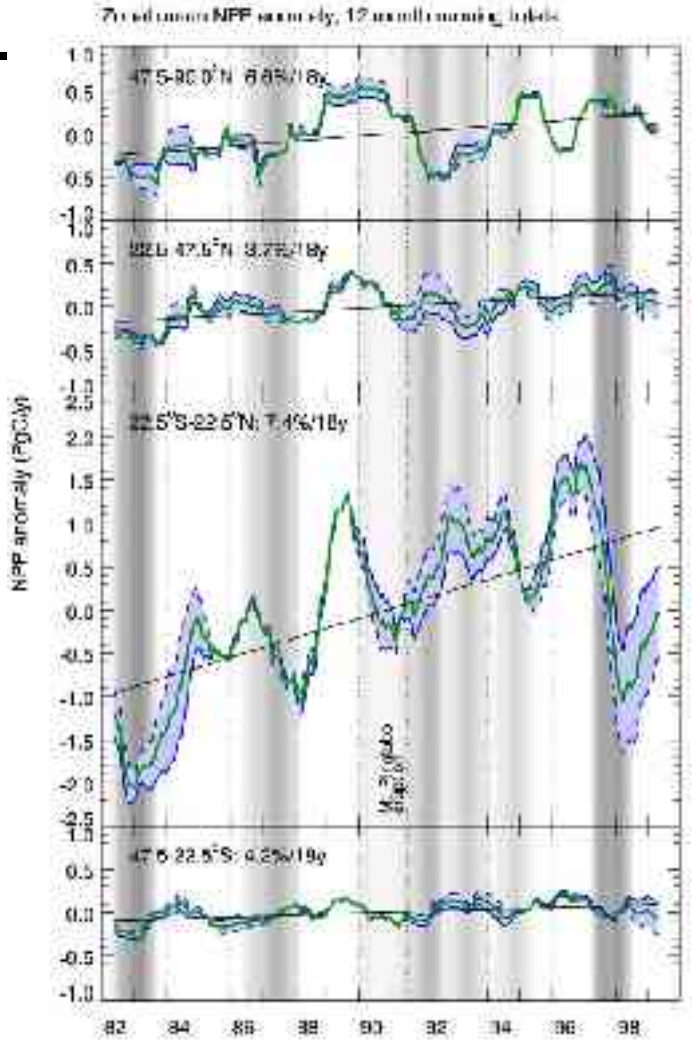
Ecosystems in all tropical regions and those in the high latitudes of the Northern Hemisphere accounted for 80% of the increase in global NPP between 1982 and 1999.

Variations in tropical NPP have the highest association with global CO2 growth rates (r = 0.75, p<0.001), as NPP and soil respiration in the tropics are more tightly coupled on an annual basis compared to ecosystems in other latitudes.

El Niño impacts are strong at low latitudes when compared to mid- and high latitudes.

A strong decline in NPP following the Mt. Pinatubo eruption (1991) was evident only at the high latitudes of the Northern Hemisphere. Cooler temperatures resulting from the eruption decreased the growing season length at high latitude.

The same cooling may have promoted plant growth in low latitude ecosystems by reducing the evaporative demand and respiration losses.



- the greening earth and increasing terrestrial npp

Trends & goals motivating NASA

- NASA conducts missions to take measurements that produce large amounts of data to support ambitious science goals
  - Too much work and expertise required to perform each of many steps in a discovery cycle to understand this data
- Socioeconomic demands are requiring increased use of federal sensors & data
  - NASA goal is to enable public access and benefit from the data to the same extent as the mission science team
- NASA's science questions are becoming increasingly large-scale and interdisciplinary.
  - Faster feedback loops in observing/simulation systems
  - Improvements in sensors, communications, and computing
- The discovery process requires specialized expertise which restricts the set of users and scientists using NASA products and slows the process for interdisciplinary research.

Discovery Steps and Gaps

- Current technology projects address difficulties of individual steps, typically in isolation
  - Impact: Machine-learning algorithms detect regularities in underlying phenomena but also artifacts of the data collection/processing system.
- NASA Data systems are constructed without rigorously characterizing the data stream to enable new users to analyze the data in unanticipated ways.
  - NASA Earth Observing System uses documents: Algorithm Theoretical Basis Documents or “ATBD’s”
- Experience has shown that simply automating existing methodologies and practices is not the most effective use of technology; it is necessary to fundamentally rethink how research is conducted in light of new technological capabilities.

Discovery Systems Proposed Overview

1. Objective
   - Develop technologies to scale up the scientific discovery process

2. Technology Categories
   - Distributed Data Search, Access, and Analysis
   - Machine Integration of Data & Quality Assurance
   - Machine-Assisted Model Discovery and Refinement
   - Exploratory Environments and Collaboration

3. Technology Infusion
   - Demonstrations, Applications, and co-funded Infusions

Distributed data search, access and analysis

- Produce customizable data products, data reprocessing and analysis for the wide variety of NASA stakeholders
- Allow seamless multidisciplinary access to and operation on massive, distributed archives of heterogeneous data, models, and data processing algorithms
- Evolve mission and archive data systems
- Instruments and platforms need to be integrated with large-scale computing and data systems.
- Data, models, and associated algorithms should be detailed, complete, easily located, catalogued, documented, and organized by content
- Efficiently use communication resources to get maximal value out of data stored in remote and distributed systems

- Framework to enable a self-managing network of data repositories, processes, and instruments that can support resource allocation to 100,000 simultaneous end-users and automated algorithms.
- Development and demonstration of virtual data systems that support user-defined computations and intermediate and deferred access, processing, and distribution.
- Methods for usable fine-grain resource access and policy enforcement based on security, trust, and commercial concerns
- Automatic and opportunistic optimization of data analysis methods for resource-effective execution over distributed architectures
- Algorithms and systems for search and analysis across diverse structured and unstructured data including semantic data, models, algorithms, documents, and conversations

Machine integration of data

- Data and algorithm interoperability - users do not have to cope with translations and interpolations of the data either across missions or disciplines.
- Data integration – heterogeneous data are automatically registered, reconciled and fused/merged prior to analysis for both real-time and retrospective studies.
- Data validation and annotation – primary and derived data products should include all information necessary for (re)analysis, including explicit representation of source, experimental context, uncertainty, and pedigree

- Formal process and framework for planning, integrating, and publishing metadata and semantics about a new resource so that new resources can be integrated in 1 week.
- Algorithms and processes to bridge different semantic ontologies and link diverse mission databases in 1 month
- Methods for data validation (quality-control) and annotation that integrate semantic models and external data sources, including explicit representation of source, experimental context, uncertainty and provenance (pedigree)
- Custom data product creation system that automatically integrates and reconciles heterogeneous data and models and republishes a self-annotated product

Machine-assisted model discovery and refinement

- Discover and understand complex behavior across vast heterogeneous data sets
- Automated methods to create explanatory, exploratory, and predictive models of complex data
- Automated methods to identify trends and events, track changes, summarize results, and identify information gaps
- Methods to effectively model complexity and co-variability of data at different spatial and temporal scales
- Use system simulations as predictive models in data analysis applications and closed-loop model-prediction-driven targeted data generation or requests
- Methods to improve evaluation and comparison of alternative models

- Algorithms to enable requests or collection of new data guided by automated analysis of model predictions and observations
- Methods to reduce massive high-dimensional databases by a factor of 100 while preserving information for task specific investigations.
- Data mining approaches that support discovery and analysis as volume, complexity, and diversity of data/models increases by a factor of 10 over a baseline capability
- Methods to effectively model complexity and co-variability of remote sensing and other data sources at different spatial and temporal scales.
- Method to exploit structures uncovered by machine learning algorithms to learn about the problem space in which you are trying to do optimization
- Methods to accurately encode domain expertise to enhance the efficiency, scalability, accuracy, and usability of knowledge discovery algorithms

Collaborative exploratory environments and knowledge sharing

- Visualize / navigate / explore / mine investigation results faster, on new data types, at higher volumes
- Support efficient interdisciplinary research and working groups, e.g., Astrobiology Institute, Instrument science teams, mishap investigations, integrated product development teams for aero
- Increase the value of data and knowledge about the data for multiple users across geographic and time barriers
- Increase diversity of participants across skill levels, and reduce organizational barriers to interdisciplinary collaboration
- Exploratory environments that support interaction with high-fidelity data and simulations

- Task-specific knowledge environments that can be specialized for multi-disciplinary investigations 1 week
- Integrated discovery assistants that assist in problem formulation, solution construction, and task execution, mining and understanding using distributed resources
- Methods to preserve 75% of the utility of PI generated data and knowledge about the data for multiple users across vast geographic distances over a period of 100 years.

Discovery Systems Before/After

Technical Theme	Current Capability	After 5 years
Distributed Data Search Access and Analysis	Answering queries requires specialized knowledge of content, location, and configuration of all relevant data and model resources. Solution construction is manual.	Search queries based on high-level requirements. Solution construction is mostly automated and accessible to users who aren't specialists in all elements.
Machine integration of data / QA	Publish a new resource takes 1-3 years. Assembling a consistent heterogeneous dataset takes 1-3 years. Automated data quality assessment by limits and rules.	Publish a new resource takes 1 week. Assembling a consistent heterogeneous dataset in real-time. Automated data quality assessment by world models and cross-validation.
Machine Assisted Model Discovery and Refinement	Physical models have hidden assumptions and legacy restrictions. Machine learning algorithms are separate from simulations, instrument models, and data manipulation codes.	Prediction and estimation systems integrate models of the data collection instruments, simulation models, observational data formatting and conditioning capabilities. Predictions and estimates with known certainties.
Exploratory environments and collaboration	Co-located interdisciplinary teams jointly visualize multi-dimensional preprocessed data or ensembles of running simulations on wall-sized matrixed displays.	Distributed teams visualize and interact with intelligently combined and presented data from such sources as distributed archives, pipelines, simulations, and instruments in networked environments.

Discovery Systems Summary

Goal:

Develop and demonstrate discovery and analysis technologies and integrated architectures to accelerate and scale up the scientific discovery process

Technology Categories

- Distributed Data Search, Access, and Analysis
- Machine integration of data & quality assurance (QA) Machine-Assisted Model Discovery and Refinement
- Exploratory Environments and Collaboration

SIAM
'04