## An Overview of Ecological Modeling and Machine Learning Research Within the U.S. National Aeronautics and Space Administration

Fourth International Workshop on Environmental Applications of Machine Learning September 27 - 29, 2004, Bled, Slovenia

Joseph.C.Coughlan@nasa.gov

Earth Science http://is.arc.nasa.gov



# **Machine Learning**

- Machine Learning is the study of computer algorithms that improve automatically through experience.
  - Machine Learning, Tom Mitchell, McGraw Hill, 1997.
- Machine Learning is ... research on computational approaches to learning.
  - Machine Learning, Kluwer
  - Learning Problems: Classification, regression, recognition, and prediction; Problem solving and planning; Reasoning and inference; Data mining; Web mining; Scientific discovery; Information retrieval; Natural language processing; Design and diagnosis; Vision and speech perception; Robotics and control; Combinatorial optimization; Game playing; Industrial, financial, and scientific applications of all kinds.
  - Learning Methods: Supervised and unsupervised learning methods (including learning decision and regression trees, rules, connectionist networks, probabilistic networks and other statistical models, inductive logic programming, case-based methods, ensemble methods, clustering, etc.); Reinforcement learning; Evolution-based methods; Explanation-based learning; Analogical learning methods; Automated knowledge acquisition; Learning from instruction; Visualization of patterns in data; Learning in integrated architectures; Multistrategy learning; Multi-agent learning.



# **Environmental; Ecology**

Environmental: Relating to or being concerned with the ecological impact of altering the environment

Ecology: The science of the relationships between organisms and their environments

Environment The totality of circumstances surrounding an organism or group of organisms, especially:

- The combination of external physical conditions that affect and influence the growth, development, and survival of organisms: "We shall never understand the natural environment until we see it as a living organism" (Paul Brooks).
- The complex of social and cultural conditions affecting the nature of an individual or community.

Dictionary.com



# **USA funding of Earth Science**

Who pays for federal Earth science research and applications in the United States?

Definition: Any of the sciences that deal with the Earth or its parts.

- NSF National Science Foundation
- DOE Department of Energy (CO<sub>2</sub>)
- EPA- Environmental Protection Agency
- Dept of Commerce / NOAA National Oceanic and Atmospheric Administration
- Dept Agriculture / USFS US Forest Service
- Dept of Interior / NPS National Park Service; BLM Bureau Land Management / USGS - US Geologic Survey
- NASA National Aeronautic and Space Administration
  - NASA Funds Earth science because of NASA's unique vantage point from space. NASA develops technology to make measurements (observations) and build predictive models of the Earth system.
- No single agency is assigned the lead role for research and monitoring of climate or global environmental change.



# Why NASA

The NASA Vision
 To improve life here,
 To extend life to there,
 To find life beyond
 ...as only NASA can.

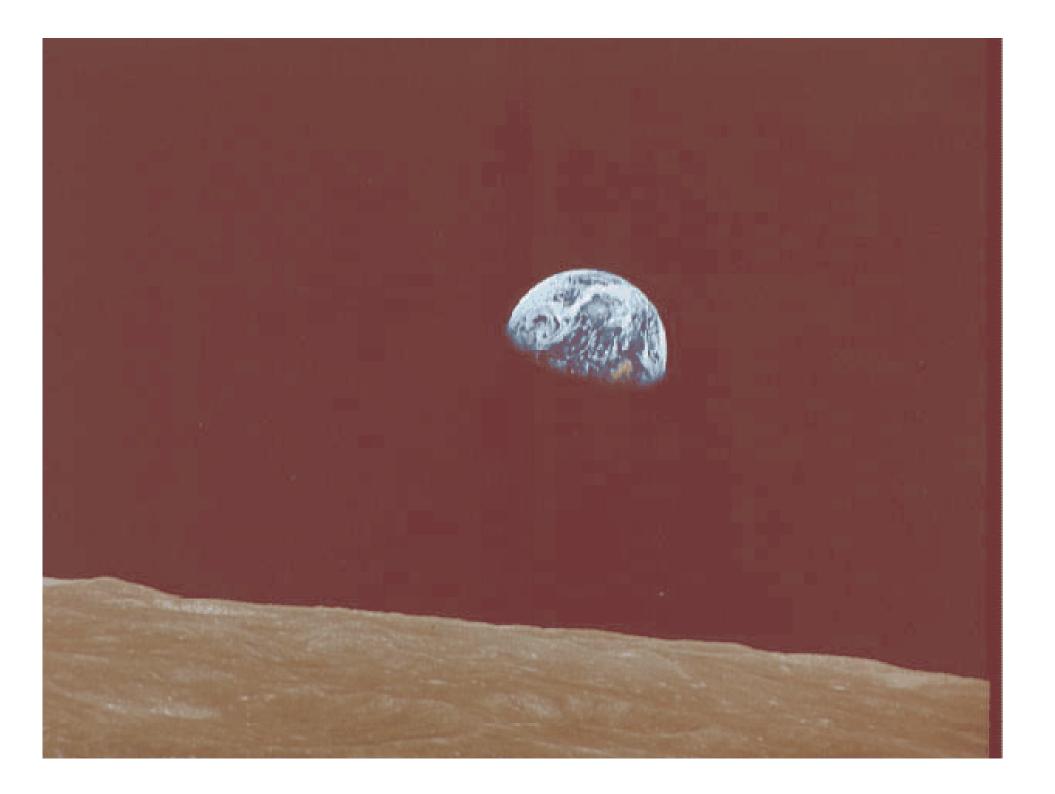
#### - NASA Mission Statement:

To understand and protect our home planet To explore the universe, and search for life To inspire the next generation of explorers, as only NASA can.

#### – 1958 NASA Mission Statement

To advance and communicate scientific knowledge and understanding of the Earth, the solar system, and the universe and use the environment of space for research. To explore, use, and enable the development of space for human enterprise. To research, develop, verify, and transfer advanced aeronautics, space, and related technologies.



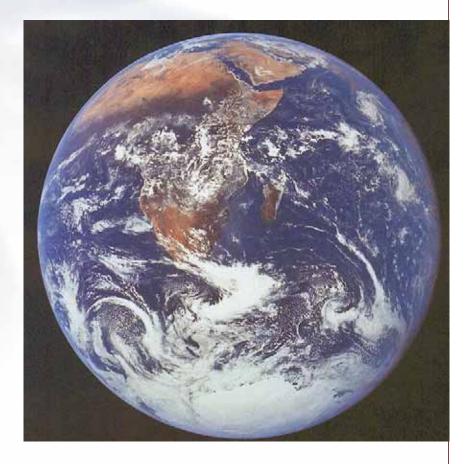




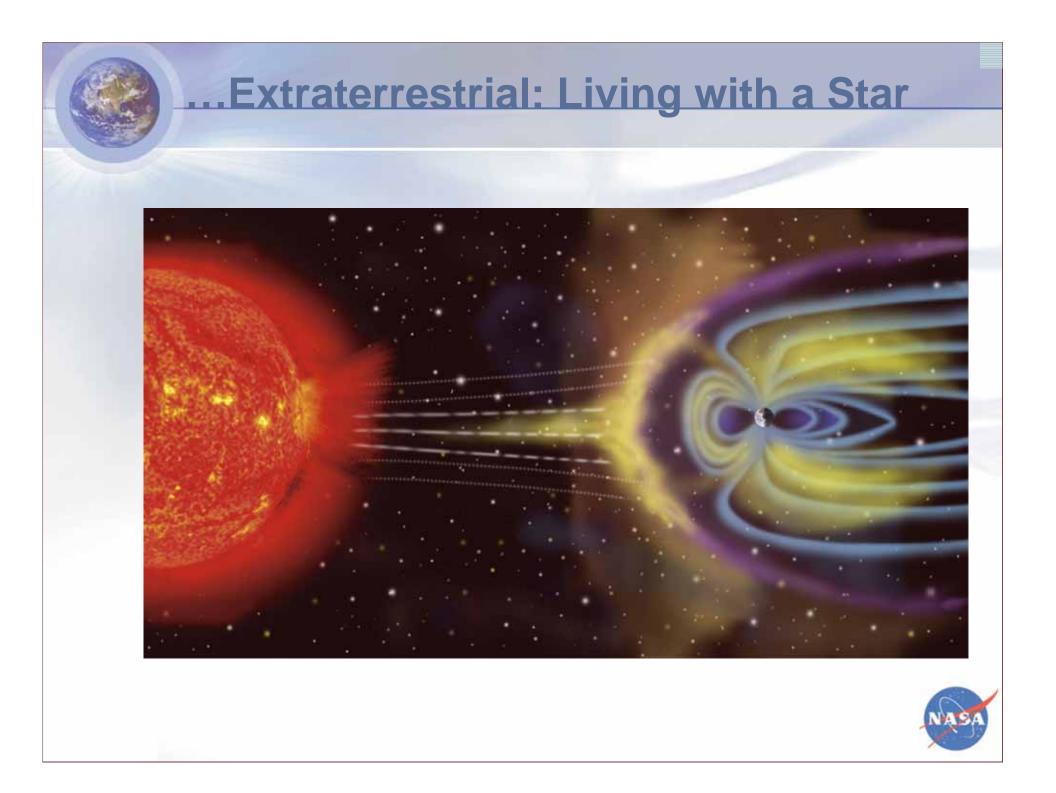
# Relationship Between Local, Regional, Global and...

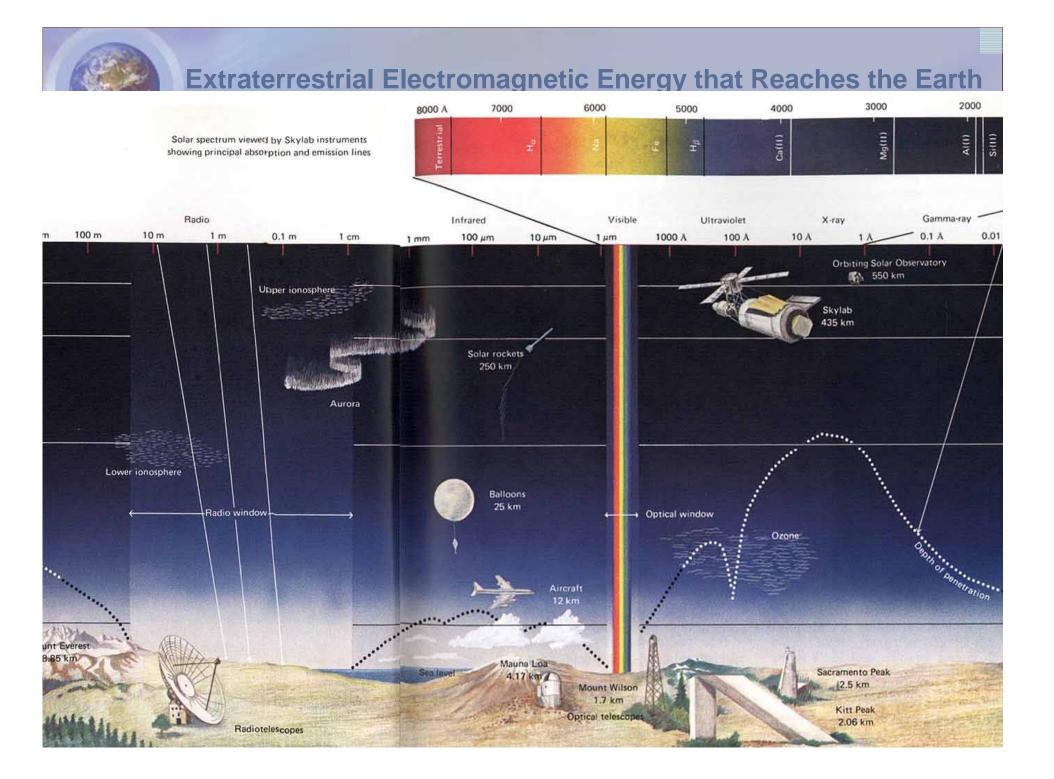




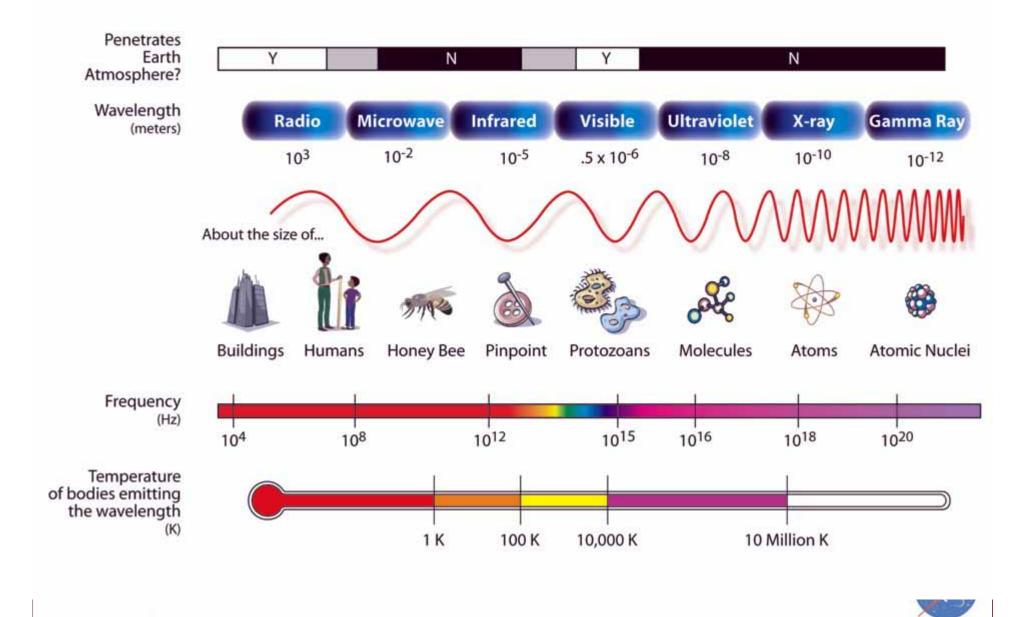




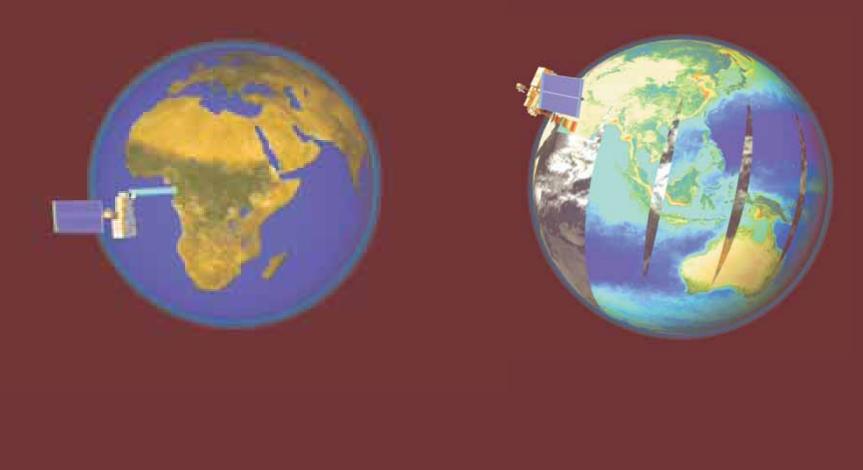


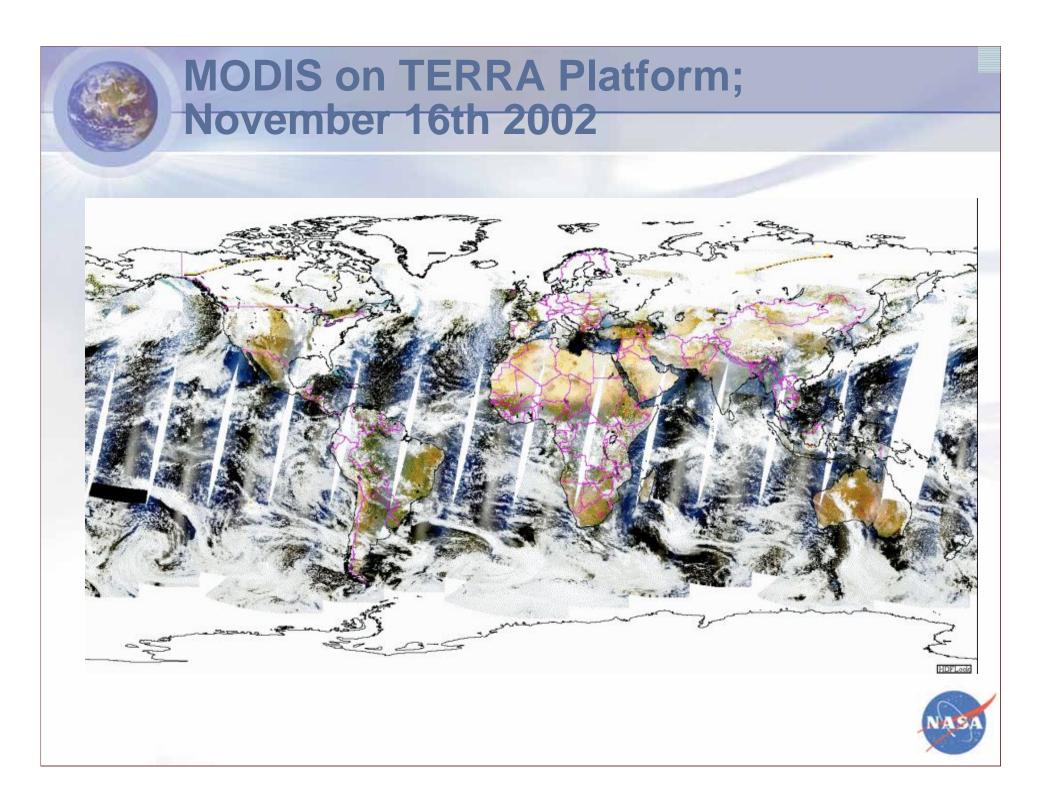


#### THE ELECTROMAGNETIC SPECTRUM



# TERRA (PM) & AQUA (AM) Platforms; MODIS Sensor Scanning the Earth

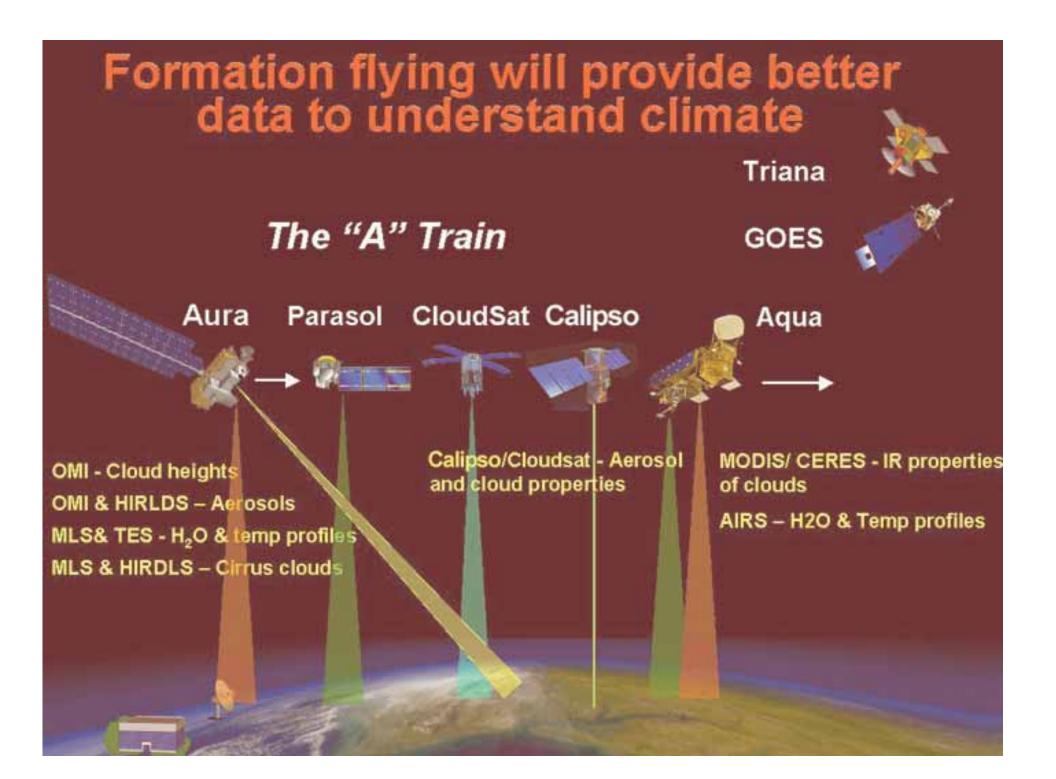




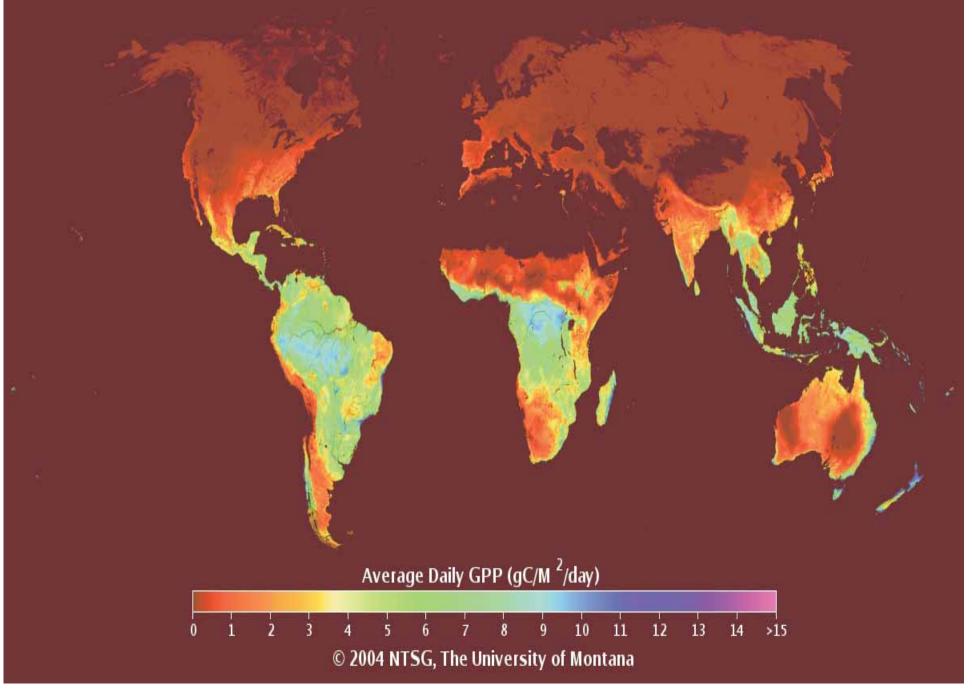


MODIS True-Color Composite Composite Period: June-September 2001 1 kilometer resolution



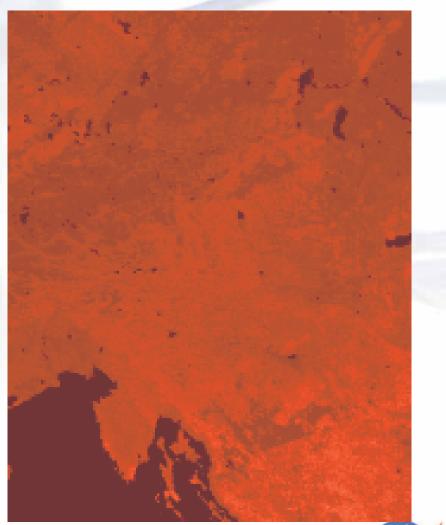


#### MOD17A2 v105 (Enhanced GPP) over the Globe, December 27 - December 31, 2003

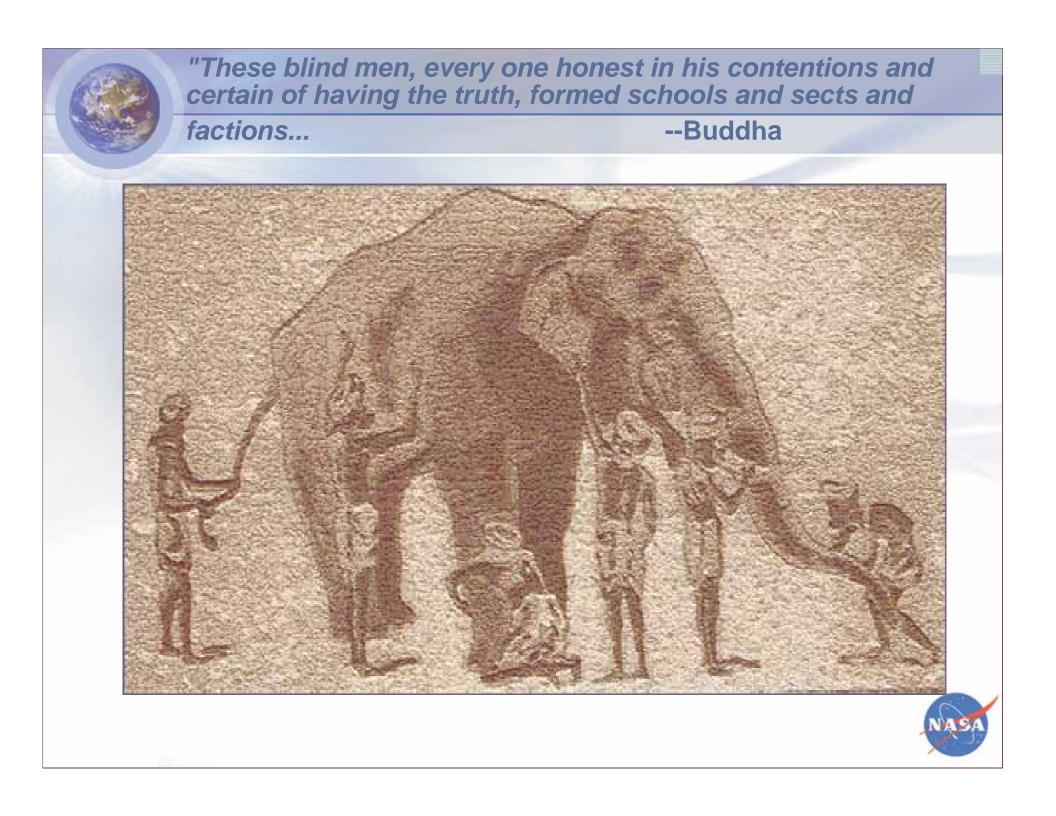


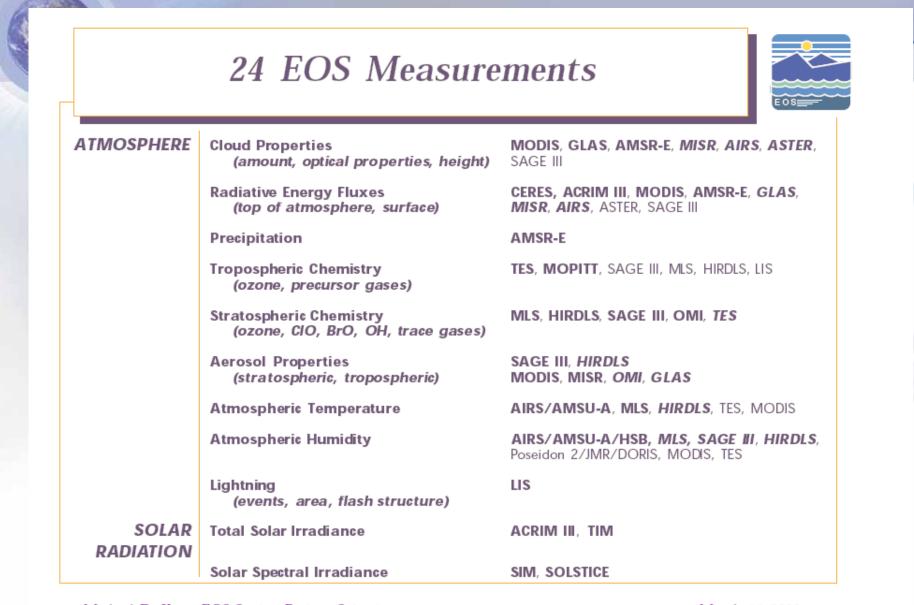
## Example of NASA Data Product, Gross Primary Production 27-31, 12, 2003











Michael D. King, EOS Senior Project Scientist

1

March 14, 2000



|       | 24 EOS Measure  | ements                                 |
|-------|---|--|
| LAND  | Land Cover & Land Use Change  | ETM+, MODIS, ASTER, MISR               |
|       | Vegetation Dynamics   | MODIS, MISR, ETM+, ASTER               |
|       | Surface Temperature   | ASTER, MODIS, AIRS, AMSR-E, ETM+       |
|       | Fire Occurrence<br>(extent, thermal anomalies)                              | MODIS, ASTER, ETM+                     |
|       | Volcanic Effects<br>(frequency of occurrence, thermal<br>anomalies, impact) | MODIS, ASTER, ETM+, MISR               |
| OCEAN | Surface Wetness   | AMSR-E                                 |
|       | Surface Temperature   | MODIS, AIRS, AMSR-E                    |
|       | Phytoplankton & Dissolved<br>Organic Matter                                 | MODIS                                  |
|       | Surface Wind Fields   | SeaWinds, AMSR-E, Poseidon 2/JMR/DORIS |
|       | Ocean Surface Topography<br>(height, waves, sea level)                      | Poseidon 2/JMR/DORIS                   |

Michael D. King, EOS Senior Project Scientist

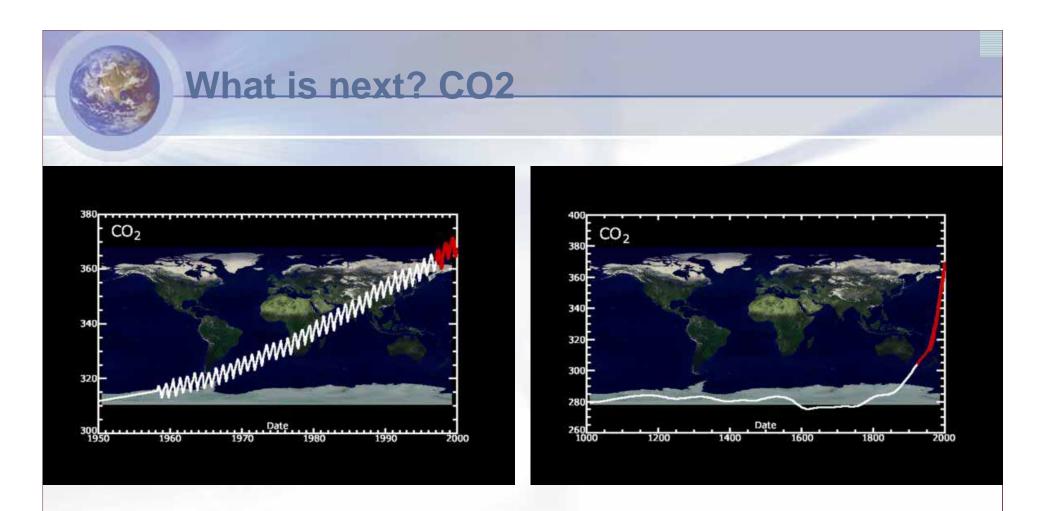
2



# 24 EOS Measurements CRYOSPHERE Land Ice GLAS, ASTER, ETM+ (ice sheet topography, ice sheet volume change, glacier change) Sea Ice AMSR-E, Poseidon 2/JMR/DORIS, MODIS, (extent, concentration, motion, ETM+, ASTER temperature) Snow Cover MODIS, AMSR-E, ASTER, ETM+ (extent, water equivalent)

3



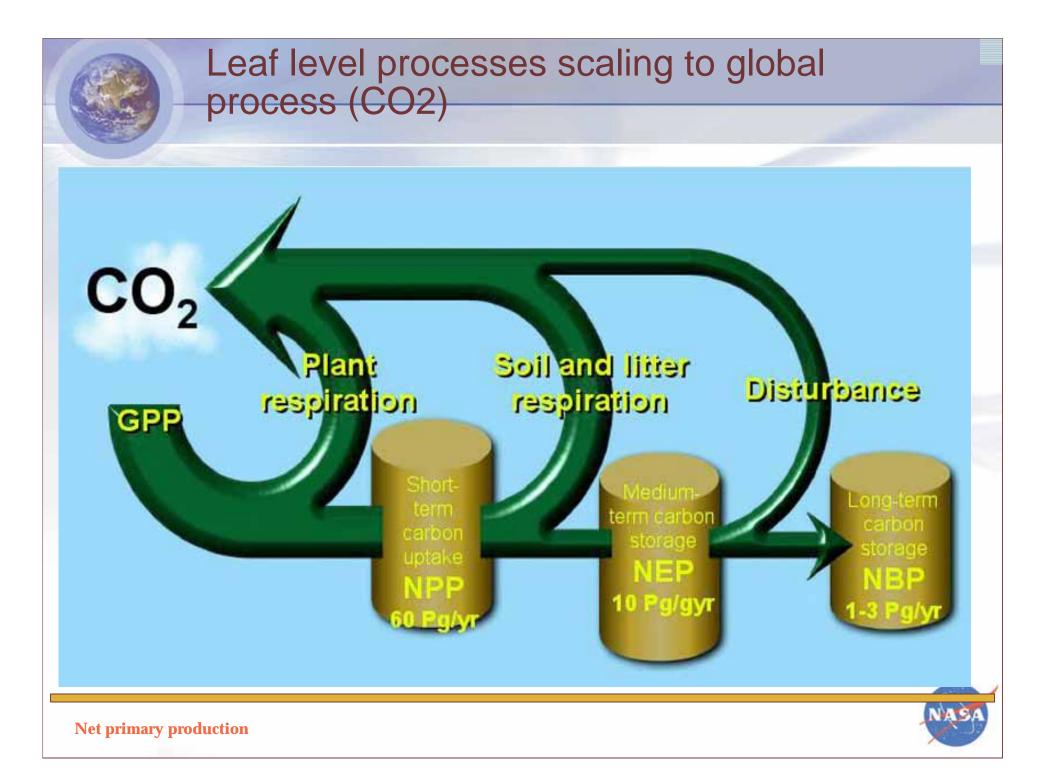


#### Carbon Dioxide (CO<sub>2</sub>)Concentration

- (a) 1950 to 2000 Mona Loa, HI record and (b) 1000 to 2000
- The Mauna Loa record shows a 16.6% increase in the mean annual concentration, rising from 315.83 ppmv (parts per million by volume) of dry air in 1959 to 368.37 ppmv in 1999. Between 1997 and 1998, the record shows the single largest one-year increase: 2.9

ppmv.



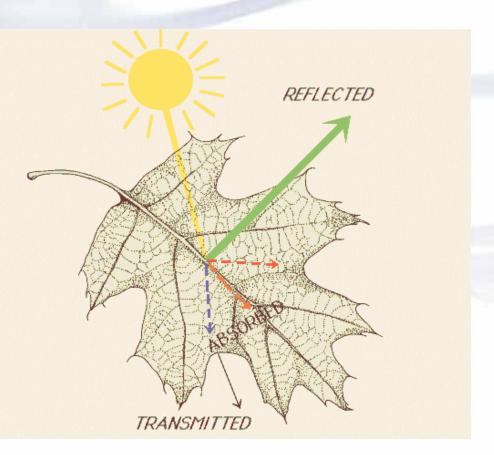


## **Definitions**

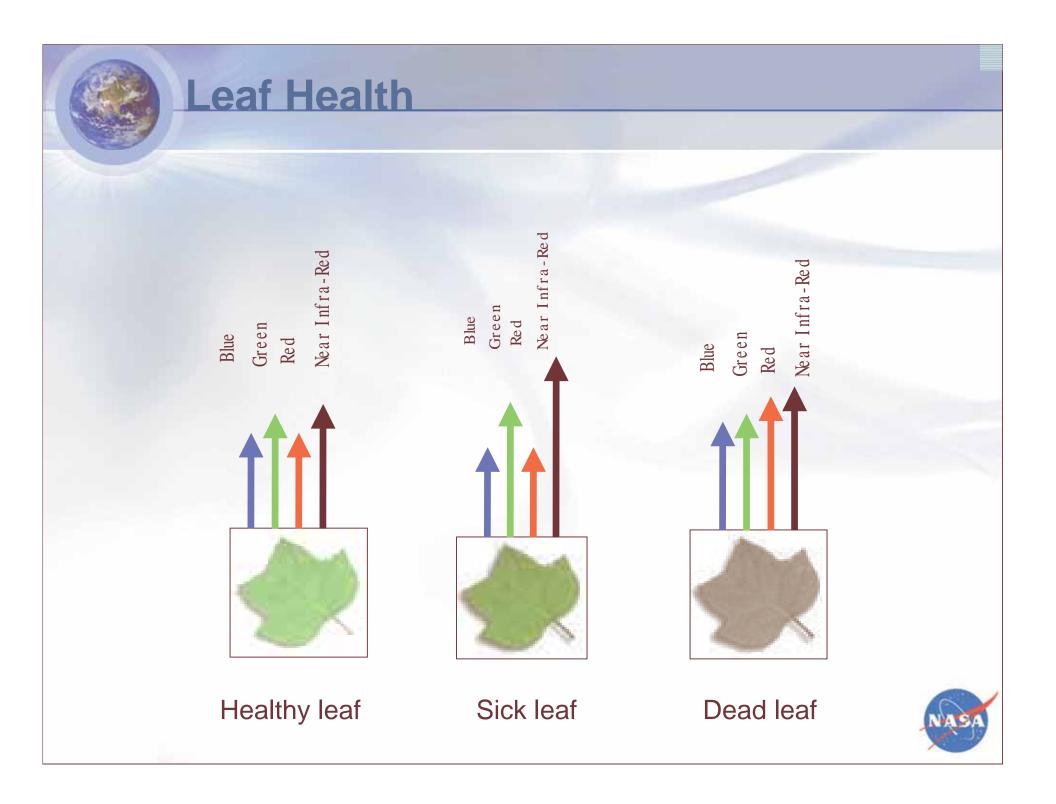
Electromagnetic radiation (light) occurring in the visible portion of the spectrum (i.e. that portion detectable to the human eye), roughly 400-75 nm (nanometers).

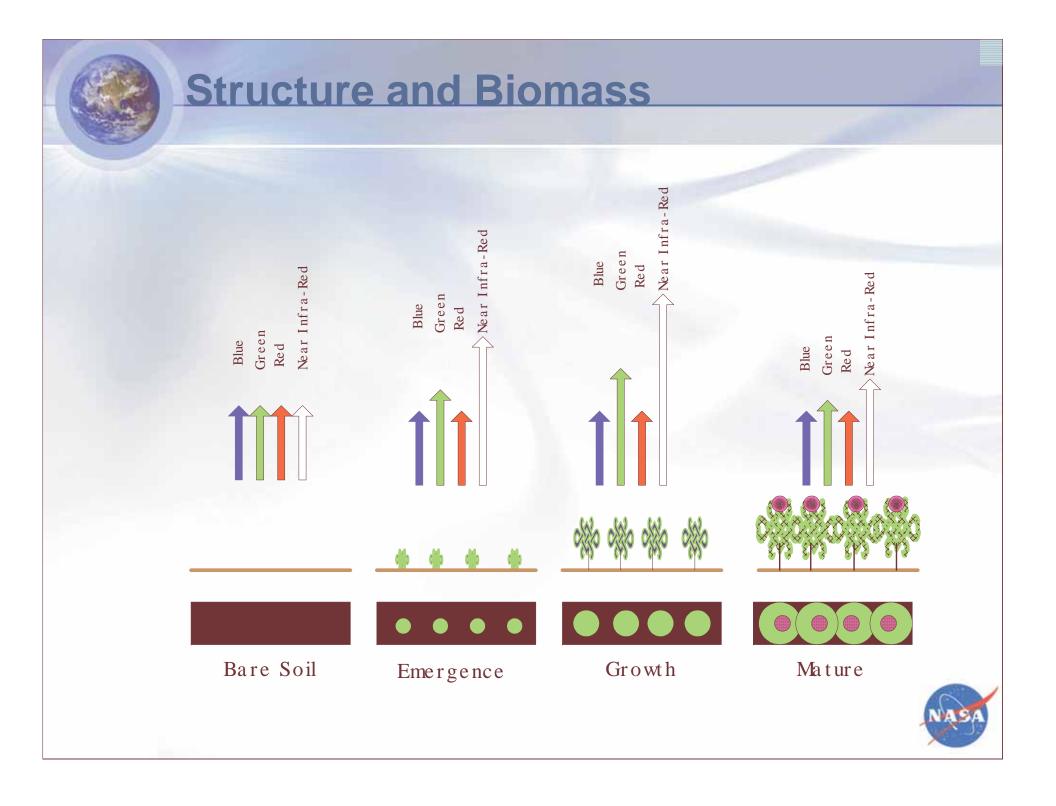
Absorption: An imbibing or reception by molecular or chemical action; as, the absorption of light, heat, electricity, etc.

Reflectance: The ratio of the total amount of radiation, as light, reflected by a surface the total amount of radiatior incident on the surface.

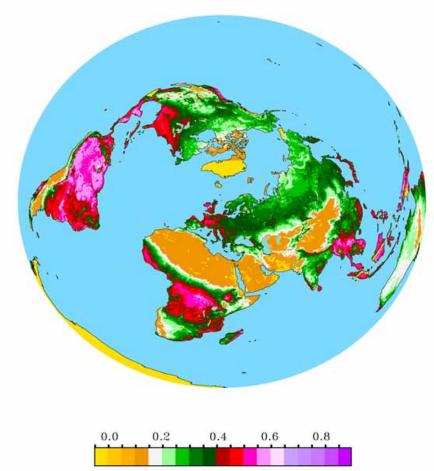




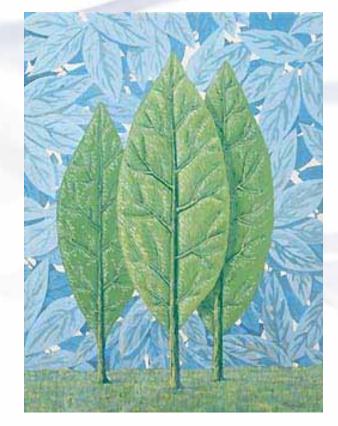




## A Satellite Based Vegetation Index (NDVI) (1981 to present)



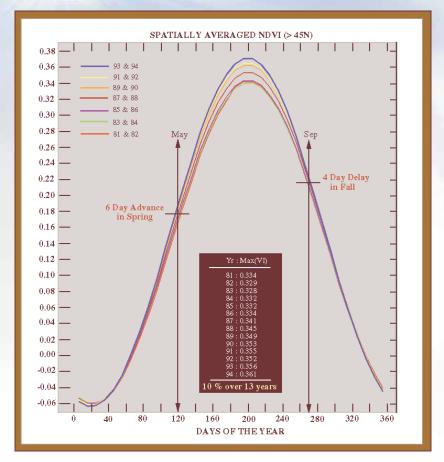
AVERAGE GREENNESS



Rene Magritte (1989-1967) The Beautiful Season



# Satellite Observed: Longer growing season trend in the north (1980s)



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

Analyses of two independent data sets for the period 1981 to 1994 suggest that -

- Vegetation greenness 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

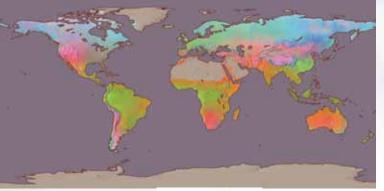
- The satellite data are concordant with an increase in the amplitude of the seasonal cycle of atmospheric  $CO_2$  exceeding 20% since the early 1970s, and an advance in the timing of the draw-down of  $CO_2$  in spring and early summer of up to 7 days (Keeling et al., Nature, 382:146-149, 1996)

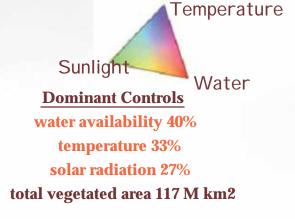
Greenness during the boreal forest growing season months of May to September increased by about 10%, the timing of spring green-up advanced by about 6 days

# **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). Using 1960-1990 average climate data (Leemans and Cramer, 1991) to develop scaling factors between 0 and 1 that indicate the reduction in growth potential. (0 is low and 1 high)

**Potential Climate Limits for Plant Growth** 





- 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)

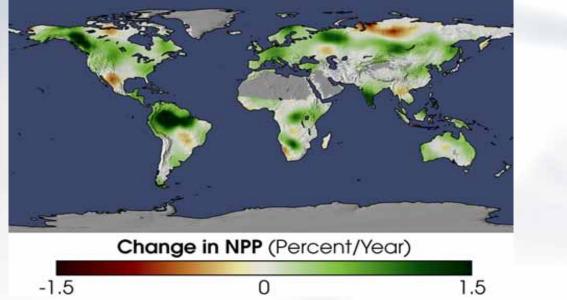
- the greening earth and increasing terrestrial npp



# **NPP: Net Primary (Plant) Production**

#### 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

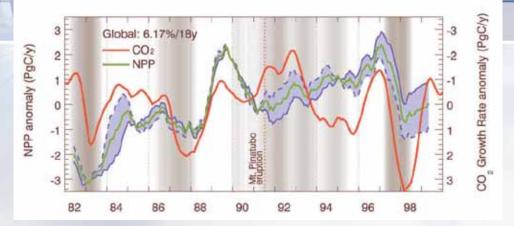
In water and radiation limited regions NPP showed the highest increase (6.5%) followed by those in temperature and radiation (5.7%), and temperature and water (5.4%) limited regions.

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<sub>2</sub> 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).

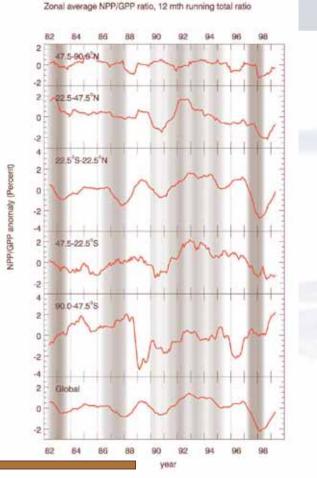
NPP declined during all three El Niño events.

High CO<sub>2</sub> 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<sub>2</sub> 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

# **Technology to-date**

- Examples assume reliable facts and outcomes are made within a multi-month to year time frame. Current systems exchange human labor in place of machine automation.
- Large teams of scientists working for years building customized but brittle systems EOS science teams have ~10 year time frame to develop maps and maps are of recent past.
- Combining diverse data types, sources with computer models for forecasting is very difficult so it isn't done often and is done at "high cost".
- Social-economic needs can require faster timeframes for data utilization e.g. Hurricane landfall. (Evacuation costs \$1,000,000 per mile)
- How do we fully utilize and exploit these data?



## **Motivations**

NASA is a "data collection" agency

- NASA collects unique data:
  - The data describes unique operational craft and platforms.
  - NASA is the lead agency for collecting space borne imagery data.
  - NASA maintains extensive, distributed data holdings.
  - Many missions require a virtual presence and automated understanding.



## Background

NASA science missions are beginning to automatically *Transform* raw data into "geophysical parameters" for users.

Interpretation of all data is still labor intensive.
 Moore's Law: For remote sensing, ~10<sup>3</sup> drop in system costs since 70's. Increased capabilities is driven by more users producing more results, not process efficiency.

Data streams and data archives are distributed, transactional systems.

Socioeconomic questions and safety require better data use and data reuse.



# **Background**

#### **NASA** Projects

- Autoclass Bayesian unsupervised classification. Found new classes in IRAS catalogue.
- Intelligent Flight Control Neural network that learns the flight characteristics of a damaged aircraft. Successful L1 milestone flight test.
- Skicat Supervised clustering of celestial objects using decision trees.
- Quakefinder Automated discovery of fault lines from satellite images.
- *OPAD* Optical Plume Anomaly Detection. Bayesian on-line analysis of shuttle main engine plume.
- SuperRes Fusing of information from multiple images into one sharper image.
- Muscle Talk Real-time embedded Hidden Markov Models for gesture recognition.
- *QUORM* Incident analysis modeling and ranking of text bodies, (FAA incident narratives, reports) and groups according to relevance to any topic of interest.

#### University/Government projects

- Sloan Digital Sky Survey CMU project to analyze huge amounts of astrophysical data.
- Potter's Wheel Berkeley project for on-line data conditioning and reformatting.

#### Commercial projects

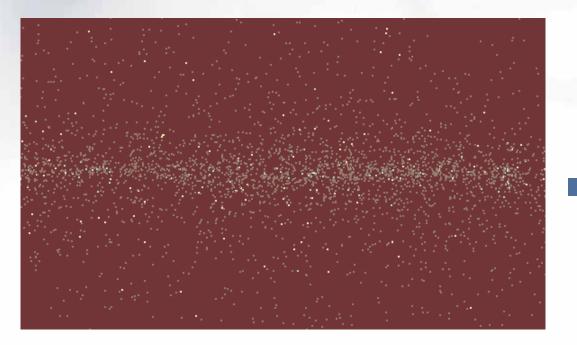
- Google Mines the link structure of the web to improve search relevance.
- *Clementine/Darwin/MineSet/Intelligent Miner* Various algorithms (notably Association Rules) for pattern discovery in transaction databases.





## Background: Autoclass IRAS Class Discovery

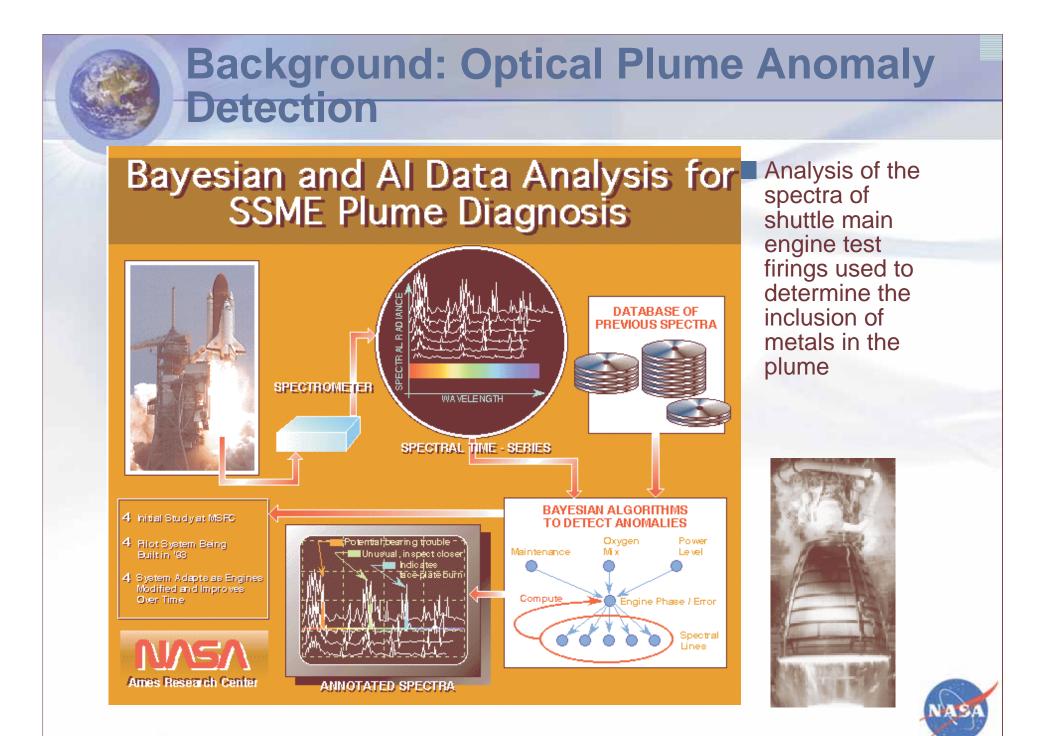




Autoclass - uses a restricted class of Bayes Nets to autonomously discover classes in data.

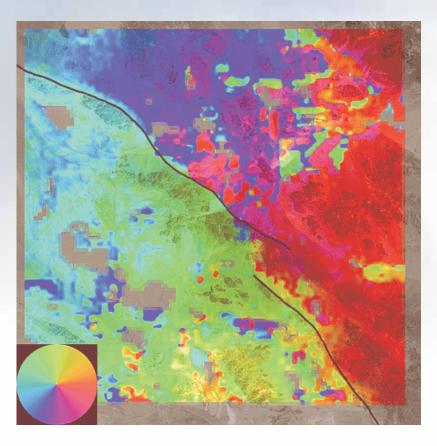
When applied to the IRAS catalogue, from subtle differences between their infrared spectra, two subgroups of stars were distinguished, where previously no difference was suspected.
 Published in Astronomy & Astrophysics 1989





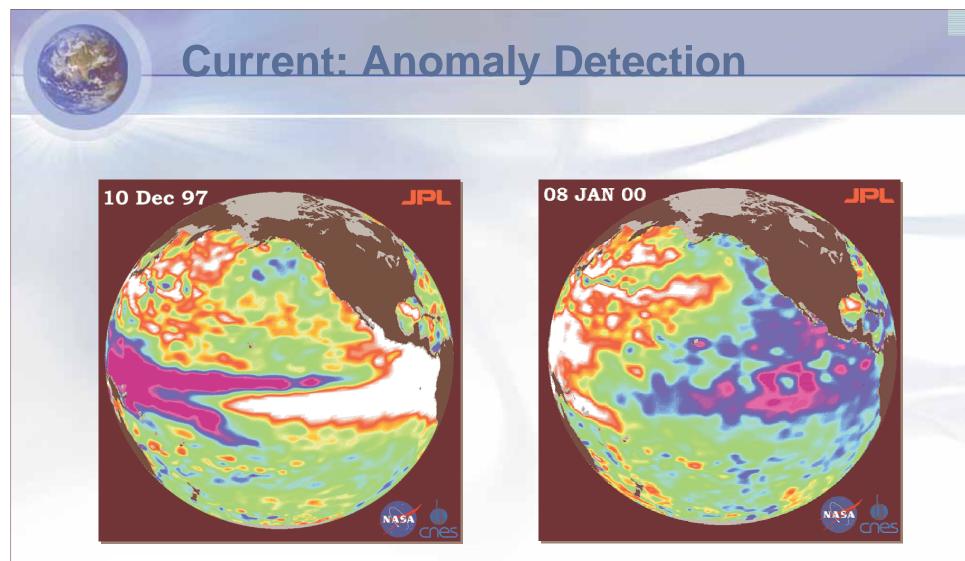


# Background: Quakefinder: Earthquakes from Space



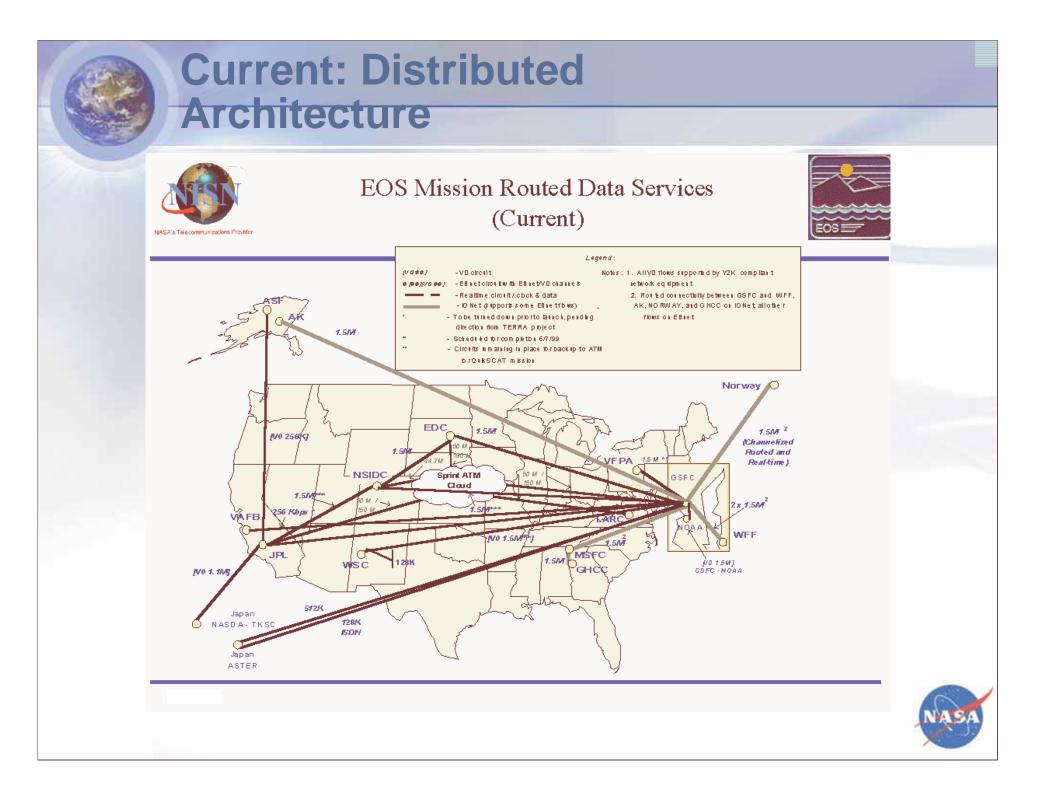
- Applies machine learning techniques to the measurement of very small ground displacements due to earthquake activity
- Used to analyze the 1992 Landers earthquake in Southern California.
- Ground motions as small as 50 cm were then measured.
- Colours on the image represent the direction of the ground motion

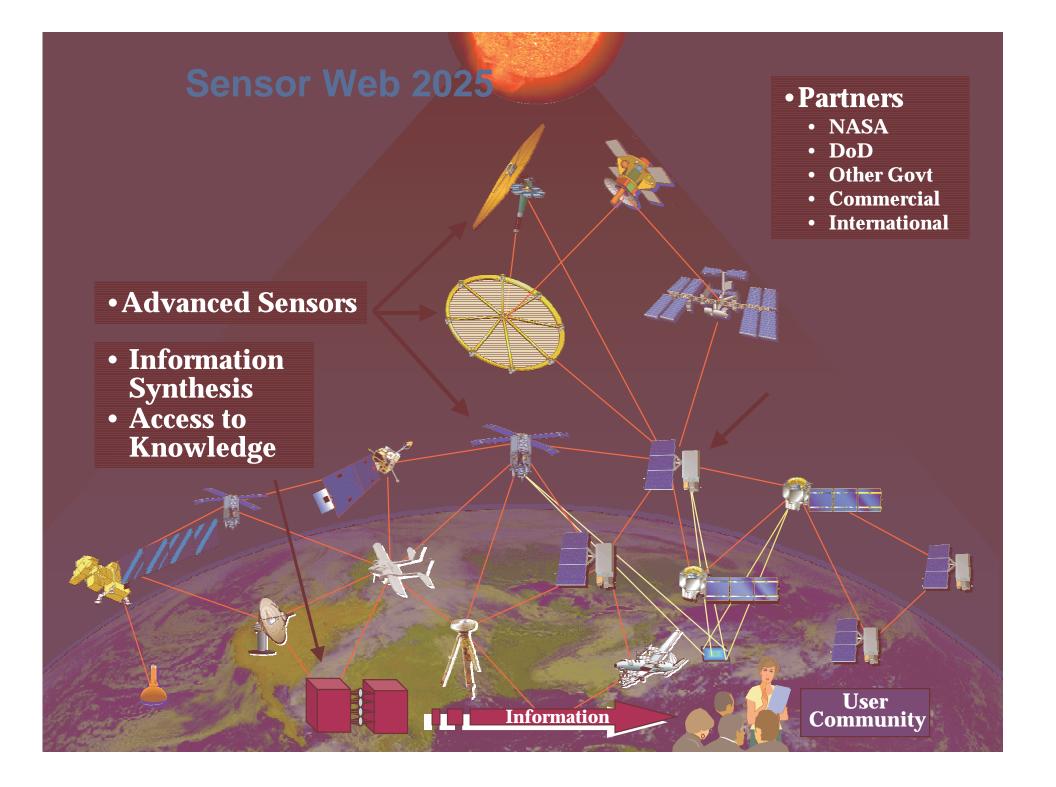




**"1997 El Niño Rhythm"** Sea surface height (mm) TOPEX/Poseidon "La Niña's persistence may be part of larger climate pattern"







### **Technical Objective**

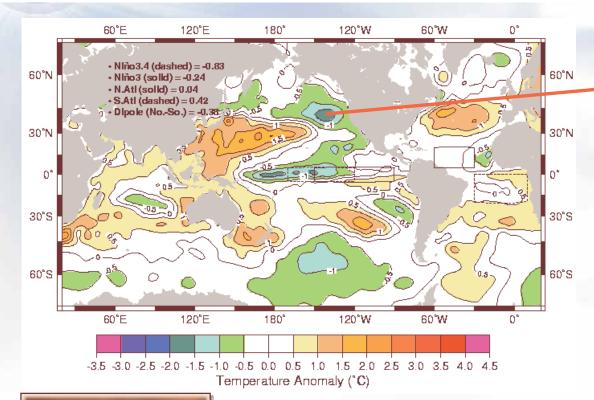
Develop the ability to autonomously transform data to generate knowledge and take actions

Develop the capability to exploit NASA's distributed science and engineering databases by developing automated methods to discover and determine "causation" of novel features in large heterogeneous datasets.



# **Technical Objective: Illustration**

January Sea Surface Temperature in N Pacific (PDO) strongly correlated to winter night warming and wine quality & yield. Predictive model uses Sea Surface Temperature of Pacific Decadal Oscillation





1) Industry does not do long term trend analysis – We did.

2) Uncovered <u>asymmetrical</u> warming trend 1976+. Early spring months warmed during night.

Ocean



3) Mining NASA and NOAA data show spring warming was caused by increase water vapor producing cloudy nights, with less radiative cooling at night – fewer frosts. Positive trend since 1976.



# **Technical Approach**

|                     | Data Mining  | Knowledge<br>Discovery<br>Understanding<br>and Analysis | Machine Learning<br>for Decisions and<br>Actions |
|---------------------|--|---|--|
| Question Answered   | What   | How and Why   | What and When                                    |
| Model and Use       | Batch  | Batch/On-line   | On-line  |
| Interactive         | No   | Yes   | Sometimes  |
| Data Set Size       | Any/Large  | Any   | Small and<br>Progressive                         |
| Example Application | Probability of<br>Cause and Effect.<br>What to Observe.<br>How to Summarize. | Hypothesis<br>Formulation and<br>Rejection              | On-line Prediction                               |



### **Technical Approach - Questions Answered**

### Machine Learning for Decision-making and Actions

*"Where and When"* is more important than *"How and Why"* 

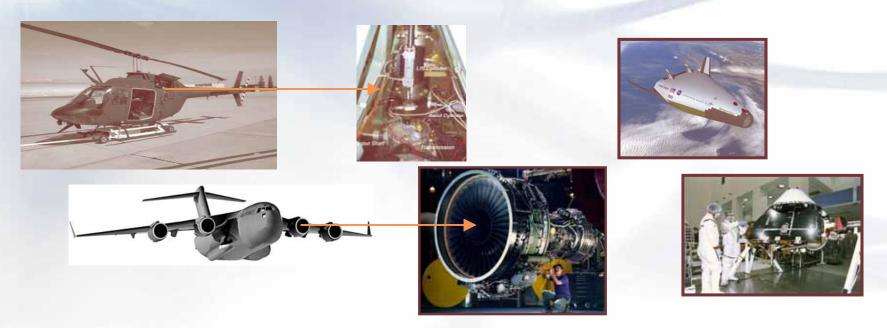


- Capability to perform automated association and "causation" from heterogeneous data.
- Build near real-time and on-line machine learning methods.
- Develop approaches for generalizing and learning with skewed and biased data.
- Provide models for prediction of component failures, degradation rates, and real-time troubleshooting and failure identification.



### **Detect and Predict Faults**

#### Anomaly Detection and Failure Prediction for Aerospace Vehicles



#### **Systems and Data:**

Integrated vehicle health monitoring systems to be installed in aerospace vehicles to monitor health and safety, detect component damage (in gears, bearings, etc.) Reliability and repeatability for onboard usage completely dependent on reducing high rates of false alarms and missed detections

#### **NASA** Relevance:

Increase flight safety mission/reduce accident rates by reducing missed detection rates Decrease maintenance costs/unscheduled overhaul by reducing false detection rates

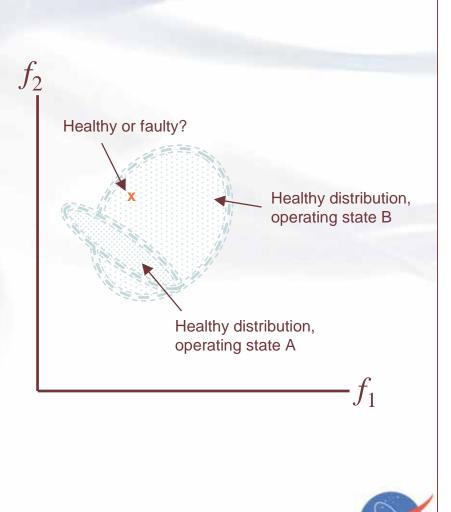
# (ML) Step III: Novelty/Anomaly Detection

#### Problem:

Outlier detection requires accurate density models of healthy baseline
Accurate model captures characteristics of healthy (nonstationary) operating states

•Should lower false alarm rate

Should raise detection rate
If operating state is known, robust fault detection should be possible



## EOSDIS: EOS Information System – Facts for 2003

- In 2003 NASA EOS Program processed nearly three terrabytes a day. (375 DVD movies a day [8 gig]).
- Currently holds over 2 petabytes 2.2 x 10<sup>15</sup>
- Grew 8 fold in volume since 1998.
  - Continues growth ~2-3 terrabytes a day
- Ingests 393 gig / day of raw data

- 1 day = 2 years of HUBBLE Space Telescope
- 1 day = more than 3 years of a legacy Earth Science satellites
- 2002 EOS sent out 12 million data products
  - Users by Type:
    - 20% Foreign
    - 12% Educational
    - 7% Foreign Commercial
    - 3% Foreign Educational
    - 2% US Government
    - 55% US Commercial
- College/Univ: 16,400 distinct users at 1529 institutions.

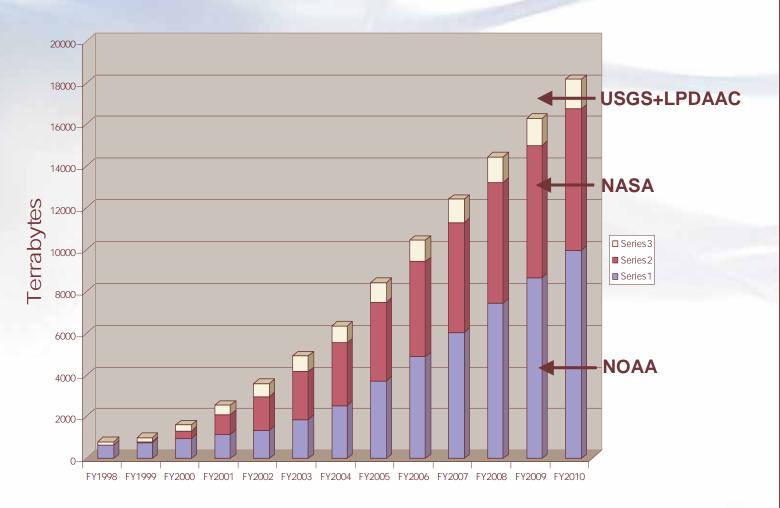


### What are some problems?

- Data volume and complexity overwhelms scientists capability to and best practices.
- Computer models of the Earth system are not integrated, no universal model exists (probably not desirable).
- Need to make informed decisions to support social-economic policy.
- Need to be more flexible and faster producing data products and information.



### National Agency Earth Science Data Holdings Size Is Increasing Faster than we can Utilize Them





### Land Cover Land Use Change & Ecology

#### Land Cover Land Use Change

#### http://www.usgcrp.gov/usgcrp/ProgramElements/land.htm

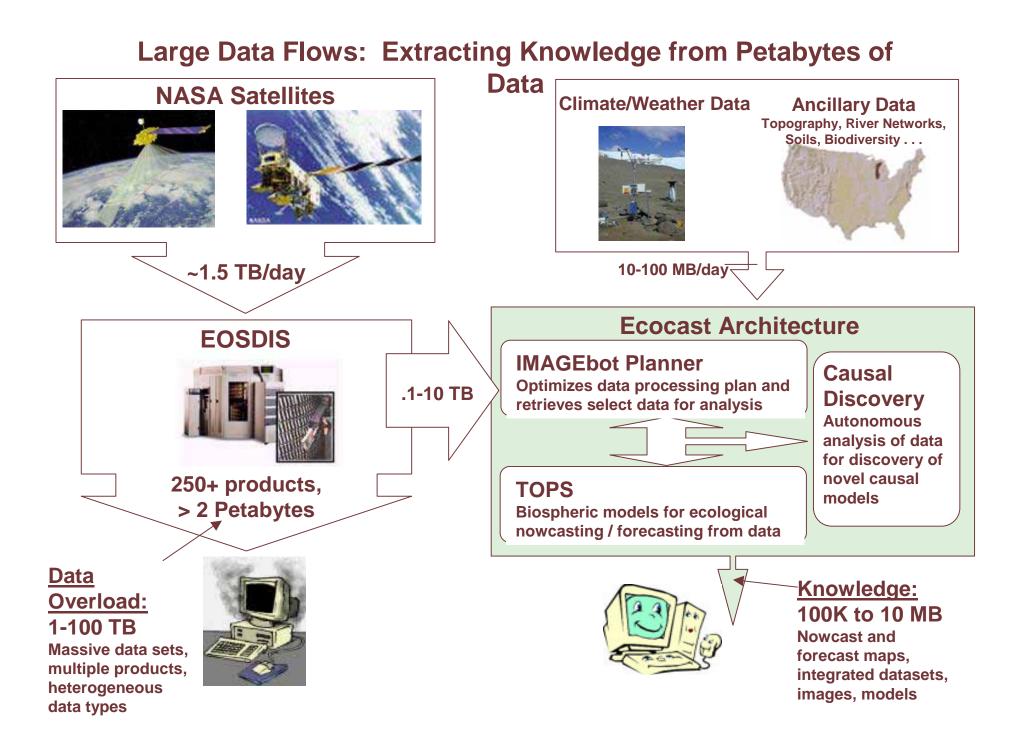
- What tools or methods are needed to better characterize historic and current land-use and land-cover attributes and dynamics?
- What are the primary drivers of land-use and land-cover change?
- What will land-use and land-cover patterns and characteristics be 5 to 50 years into the future?
- How do climate variability and change affect land use and land cover, and what are the potential feedbacks of changes in land use and land cover to climate?
- What are the environmental, social, economic, and human health consequences of current and potential land-use and land-cover change over the next 5 to 50 years?

#### Ecology

#### http://www.usgcrp.gov/usgcrp/ProgramElements/bio.htm

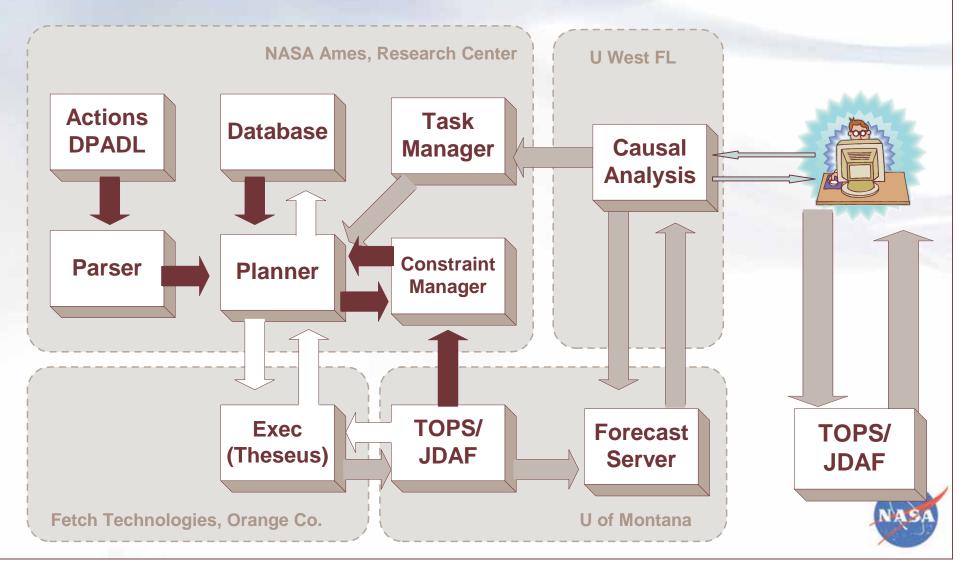
- What are the most important feedbacks between ecological systems and global change (especially climate), and what are their quantitative relationships?
- What are the potential consequences of global change for ecological systems?
- What are the options for sustaining and improving ecological systems and related goods and services, given projected global changes?



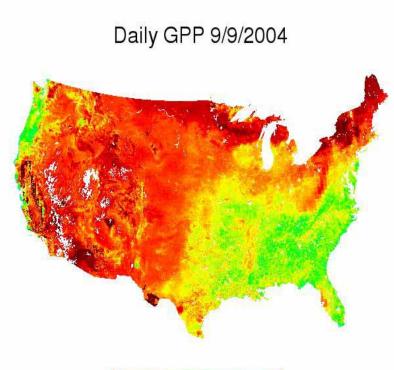


# **Ecological Forecasting Diagram**

Distributed system with commercial, academic, and Earth science partners



### Current, Cutting-edge Capability: Daily, "Live" Estimates of Global Primary (Plant) Production

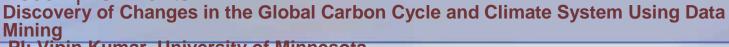


0.00 3.74 7.48 11.21 gC/m^2

- Gross Primary Production (GPP) (measures plant production). Units are grams of carbon / square meter / day.
- 1 km resolution for the conterminous USA (no HI and AK etc.)
- Daily observations
- New capability combining Earth science and Aerospace technologies to rapidly bring observed data and Earth science models together to estimate plant production. <u>http://geo.arc.nasa.gov/sge/ecocast</u>



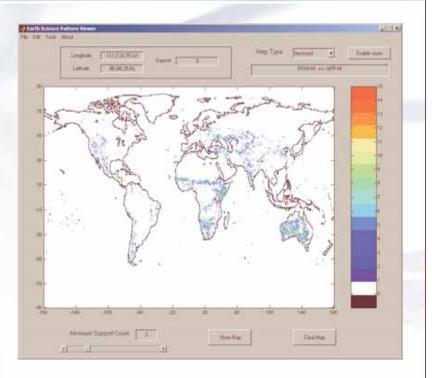
#### Accomplishments:



**PI: Vipin Kumar, University of Minnesota** 

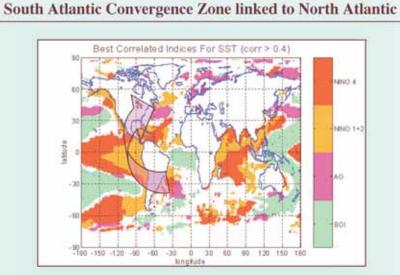
**Co-Investigators:** 

- S Klooster, Cal.St.U. Monterey Bay; S. Shekhar, G. Karypis, U MN; C.Potter NASA.
- Often data and sensors are repurposed to resolve important, unanticipated questions. Must detect subtle features in old, noisy data satellite record.
   Spatiotemporal data ill-suited for classic mining methods and solutions may be computationally complex.
- Approach: Mining the historical record to detect local ecosystem disturbance events on a global scale
  - Global analysis of 18 years of 8 KM resolution data, monthly.
  - Filter out "noise" and signals in data due to seasonality, sensor drift and etc.
- Identified ecosystem events indicated by sustained declines in light absorption in plants.
- Verified through comparison with global records of large-scale wildfires by ES collaborator using interactive tool (shown on right).
- 9 Peta-grams of carbon could have been lost from the terrestrial biosphere to the atmosphere as a result of large scale ecosystem disturbance over an 18 year period.





#### Accomplishments: Discovery of Climate Indices

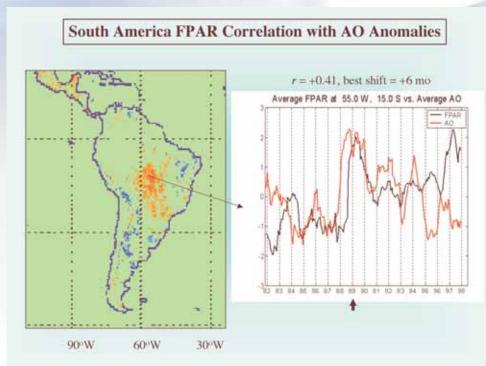


Circulation and heating associated with the South American Summer Monsoon (SASM) may exert an important influence on the boreal winter subtropical jet over eastern North America through changes Amazon rainfall and regional Hadley circulations. (Sources: Nogues-Paegle et al., 1998; Robertson et al., 2000; Cassou and Terray, 2000). New clustering method detects known indices around the globe and discovered interesting clusters that may be associated with previously unknown indices.

- The technique clusters based on similarity and on variability in similarity (i.e., uniform density)
- This clustering technique is suited for high-dimensional data, many of the known techniques break down for high dimension data in a variety of domains, (in particular, Earth Science data)
- This technique is also powerful at eliminating points with noise and detecting points with structure, whereas current algorithms would cluster noisy data and structured data together, and would lose weak patterns in the process.



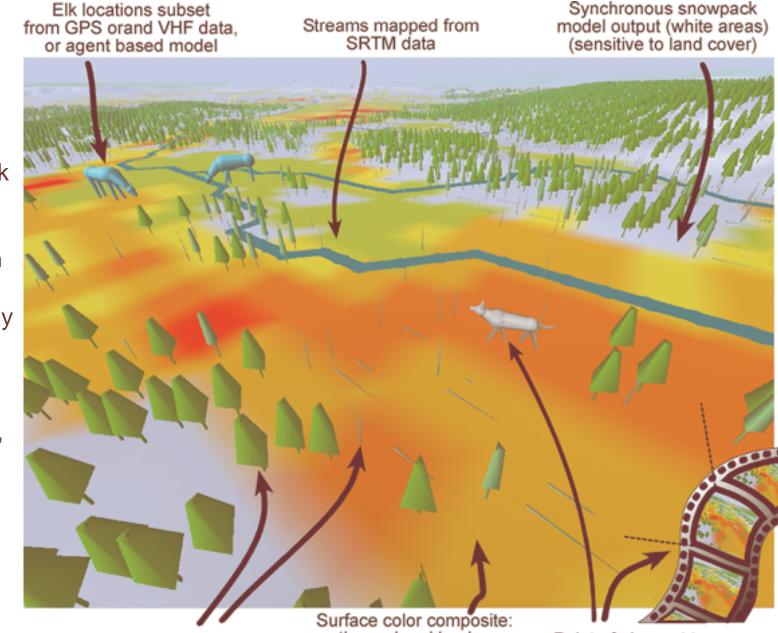
#### Accomplishments: Discovery of link between climate and terrestrial carbon fluxes



Discovered strong relationships between climate variables and terrestrial carbon fluxes over large parts of the land surface.

- On a global level, combined climate indices can be used to predict net ecosystem carbon fluxes over nearly 60 percent of the non-desert/ice covered land surface with a lead period of 2-6 months.
- Discovered, for the first time, a strong connection South American Monsoon system and terrestrial greenness of the southern Amazon region.





Yellowstone National Park USA

Fred Watson CSU Monterey Bay

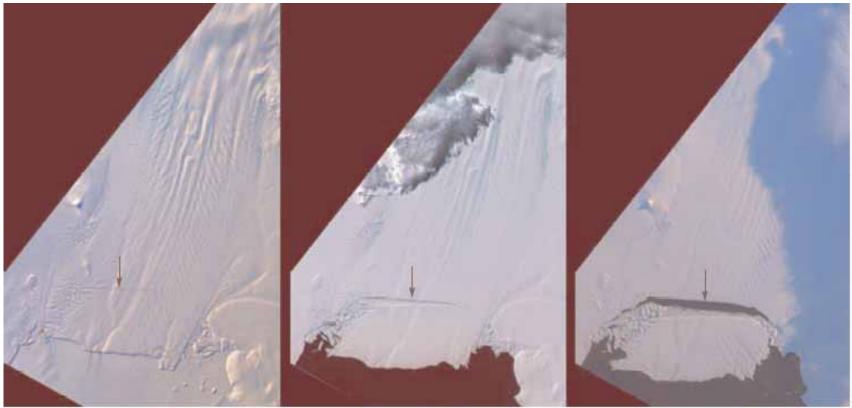
Effect of snowpack, geothermals, And wolf predation on Elk population.

> Symbolic land cover representation (trees, logs, etc.) based on remote sensing data

Surface color composite: geothermal and land cover underlay (red, green, yellow) & snowpack overlay (white)

Point-of-view set to follow movements of wolf pack over time as video progresses

# Sea ice 42 x 17 km iceberg



September 16, 2000

November 4, 2001

November 12, 2001



# **Experiments and Simulation Output**

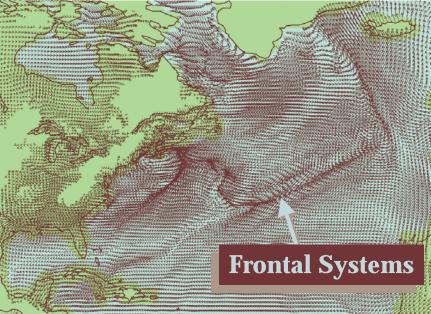
#### 100 km resolution



GEOS-2, GCM January 22, 1994

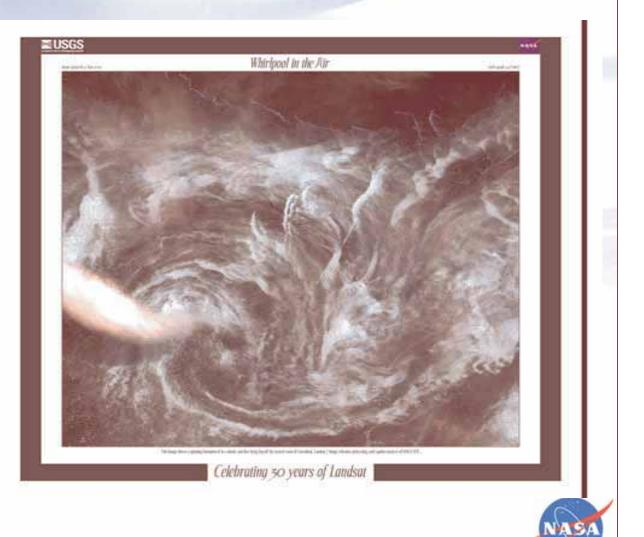
Increased computational power produces better results but simulated frontal systems are identified by human experts. Simulation overwhelms human capabilities to understand model behavior.

#### 60 km resolution



### Future Challenges International Geosphere-Biosphere Programme: IGBP

Carbon Cycle
Regional assessm
Food Provision
Water resources
Human Health



# **IGBP Carbon Cycle**

- Patterns and variability: what are the geographical and temporal patterns of carbon sources and sinks?
- **Processes, controls and interactions:** what are the controls and feedback mechanisms - natural and anthropogenic that determine the dynamics of the carbon cycle on scales of years to millennia?
- Management of the carbon cycle: what are the future dynamics of the carbonclimate system and what are the points of intervention and windows of opportunity for managing this system?





### Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA)

- 80 research groups 600 scientists including NASA
- how does Amazonia function as a regional entity (e.g. water, energy, aerosol, carbon, nutrient and tracegas cycles)?
- how will changes in land use and climate affect the biological, chemical and physical functioning of Amazonia, including its sustainability and influence on global climate?





### International Geosphere-Biosphere Programme: IGBP Food Provision

Food demands are changing:

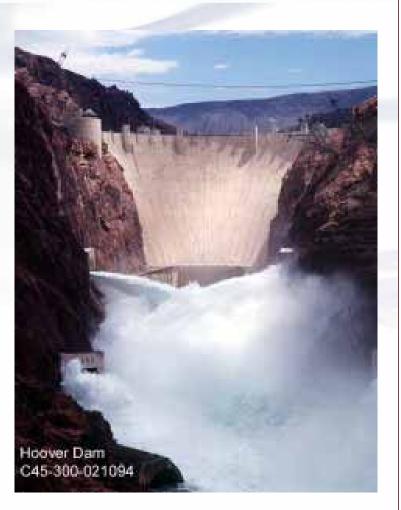
- How will global environmental change (GEC) affect food provision and vulnerability?
- How might societies and producers adapt their food systems to cope with GEC?
- What would be the environmental and socioeconomic **consequences** of such adaptations?





# **IGBP** Water Resources

- What are the relative magnitudes of changes in the global water system (GWS) due to human activities and environmental factors?
- What are the social and Earth System **feedbacks** of human-driven change to the global water system?
- To what extent is the GWS resilient and adaptable to global change?













# **IGBP Human Health**

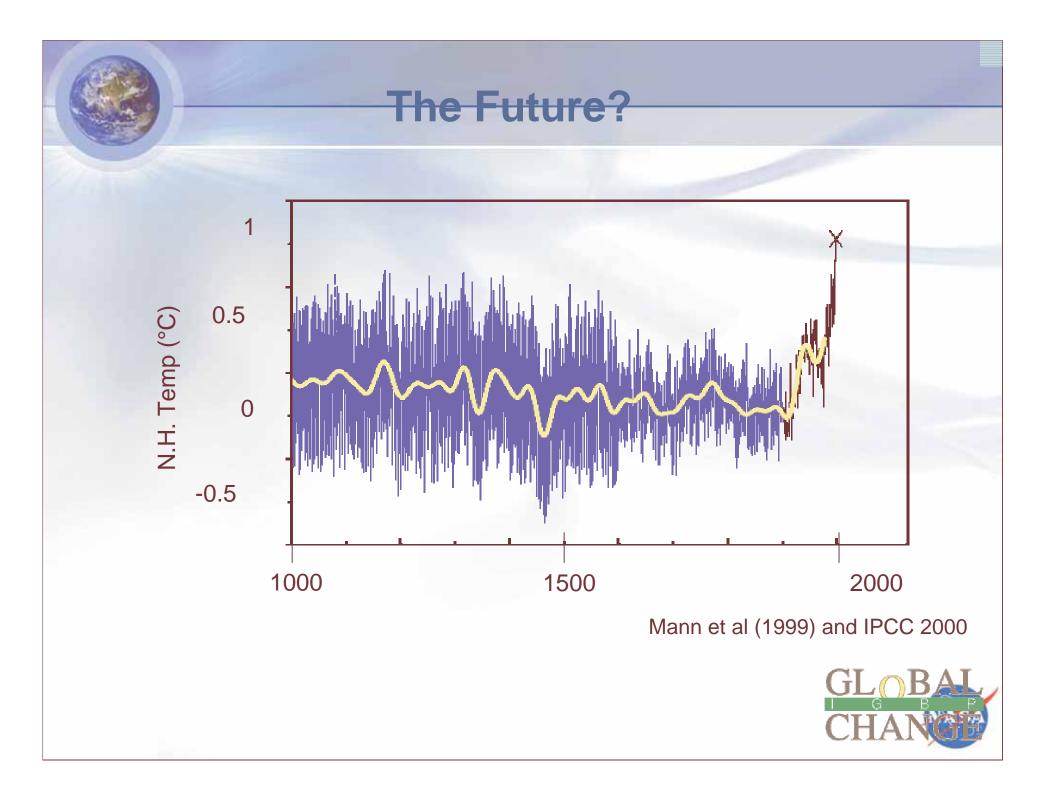
**Under Development** 

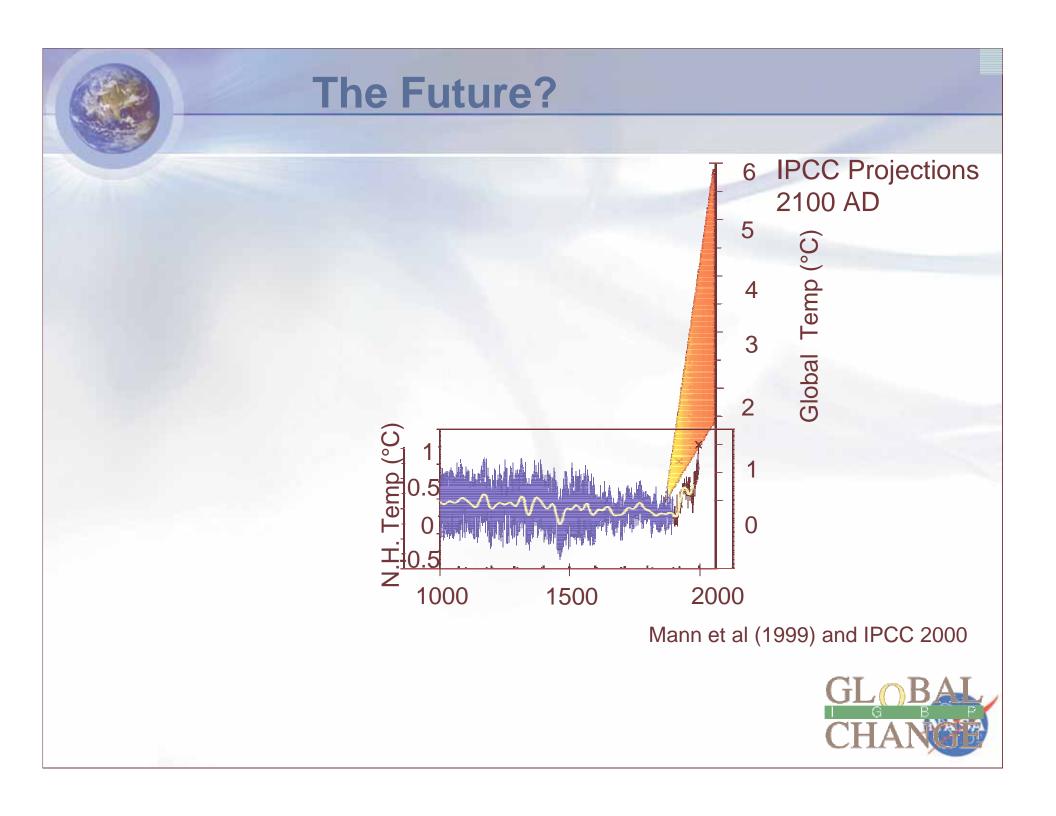
### **Project Goals:**

- To determine the past, current, and future health impacts of global environmental change.
- To enrich the policy discussion about mitigation and adaption from a human health perspective.









### **More Reading**

OUR CHANGING PLANET The U.S. Climate Change Science Program for Fiscal Years 2004 and 2005 A Report by the Climate Change Science Program and the Subcommittee on

Global Change Research A Supplement to the President's Fiscal Year 2004 and 2005 Budgets

http://www.usgcrp.gov/usgcrp/Librar y/ocp2004-5/default.htm OUR CHANGING PLANET

The LLS: Climate Charge Science Program for Fiscal Years 2004 and 2005

in the Research Design in State New York

