Geoinformatics: Implications of Scientific Knowledge Discovery

Perspectives of an Atmospheric Scientist & a Provider of Atmospheric Data Services

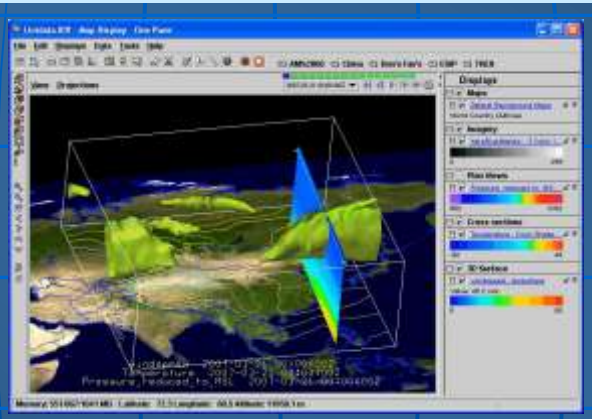
10 March 2011

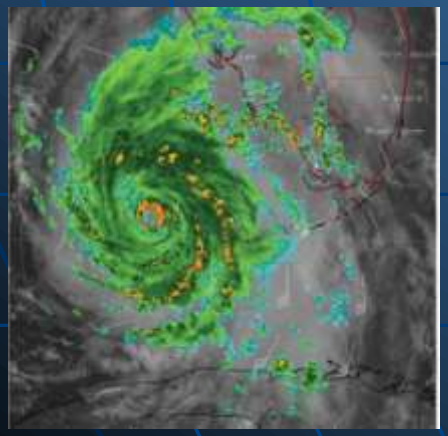
Dr. Mohan Ramamurthy
Director, Unidata
University Corporation for Atmospheric Research
Boulder, CO

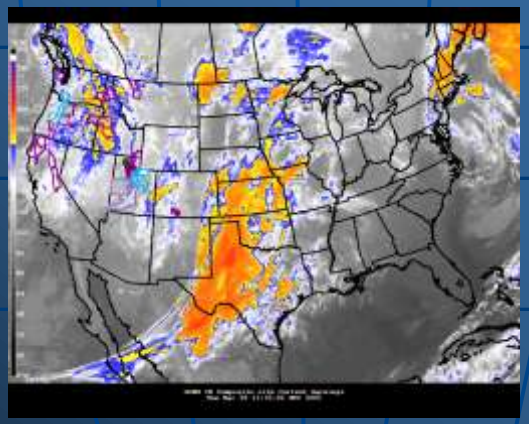
Unidata

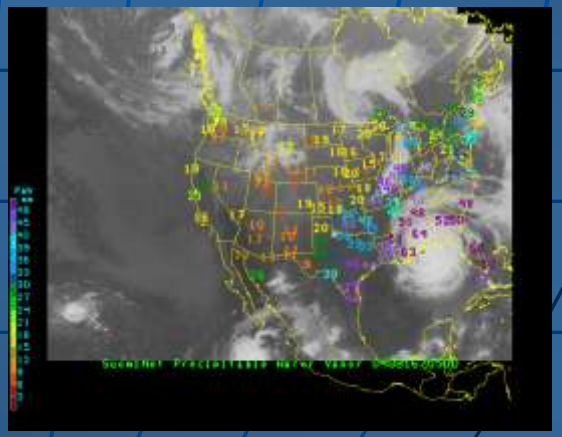
We are an NSF-funded program that provides data services for advancing geoscience education and research.

- Unidata has over 30,000 users from 1500 academic institutions and 7000 organizations in ~150 countries.



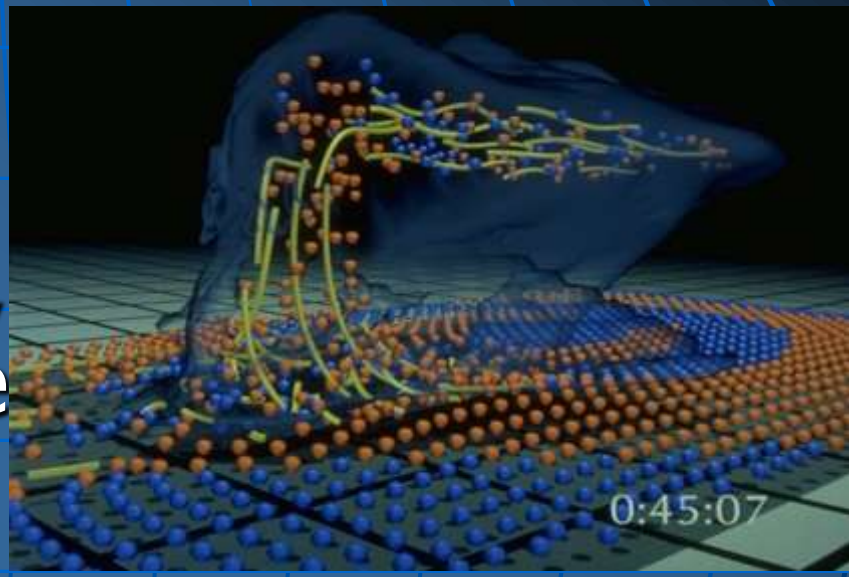






Geoinformatics & Science: A Classic Example

- High performance computing
- Cutting-edge Software, including a state-of-the art model
- Massive quantities of data

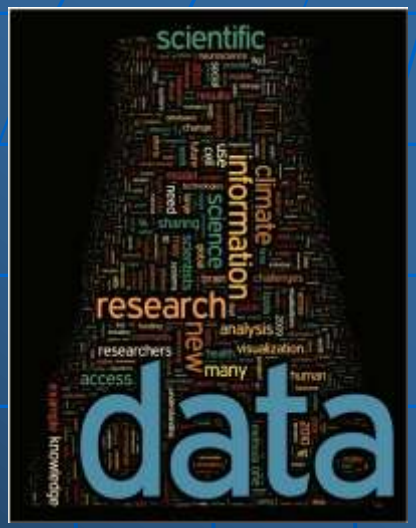


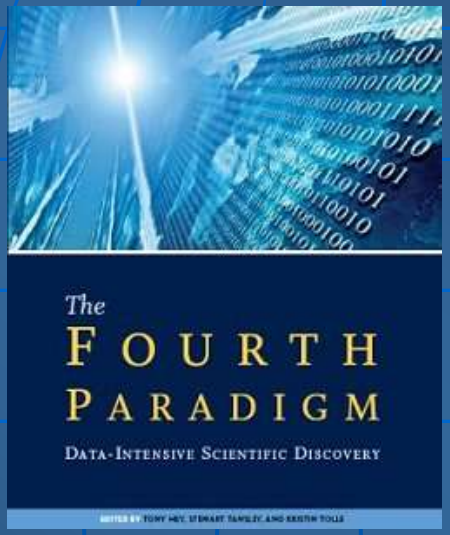
Source: Robert Wilhelmson, NCSA University of Illinois

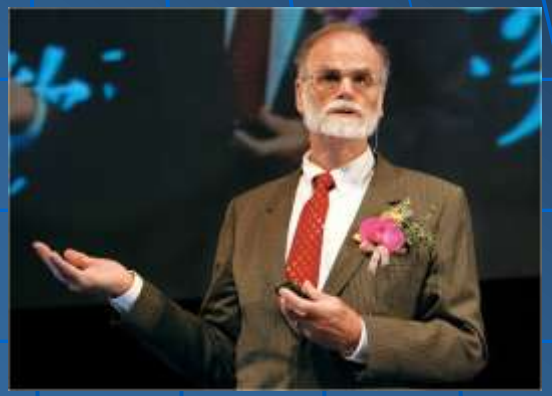
This visualization received the 1st Place Award at the London Computer Graphics Film Festival and it was subsequently submitted for an Academy Award.

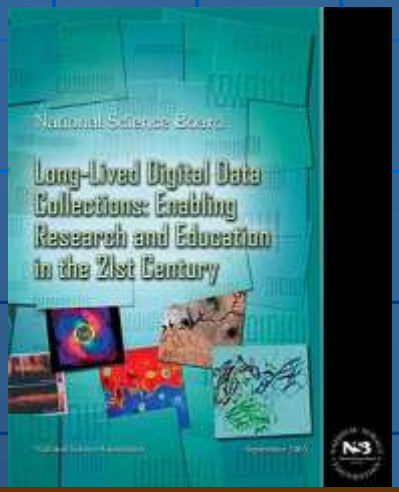
The Era of Data-Intensive Science

unidata









Data is the lifeblood of science, but we need to move from creating data to discovering knowledge.

A Provocative Suggestion

WIRED

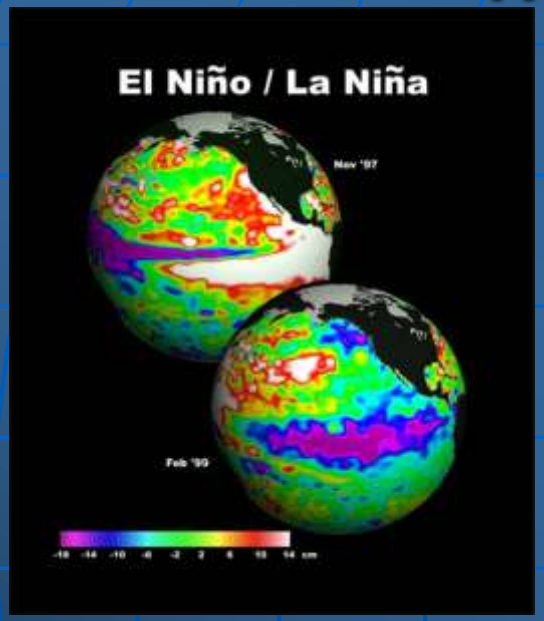
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

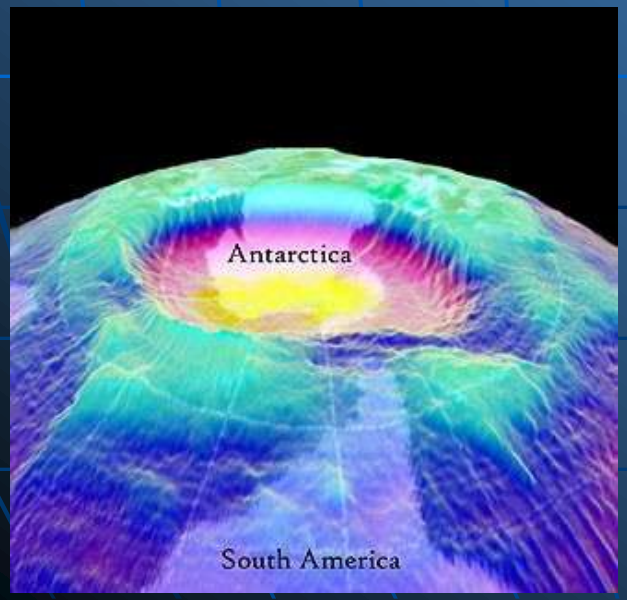
By Chris Anderson ✉ 06.23.08

Illustration: Marian Bantjes

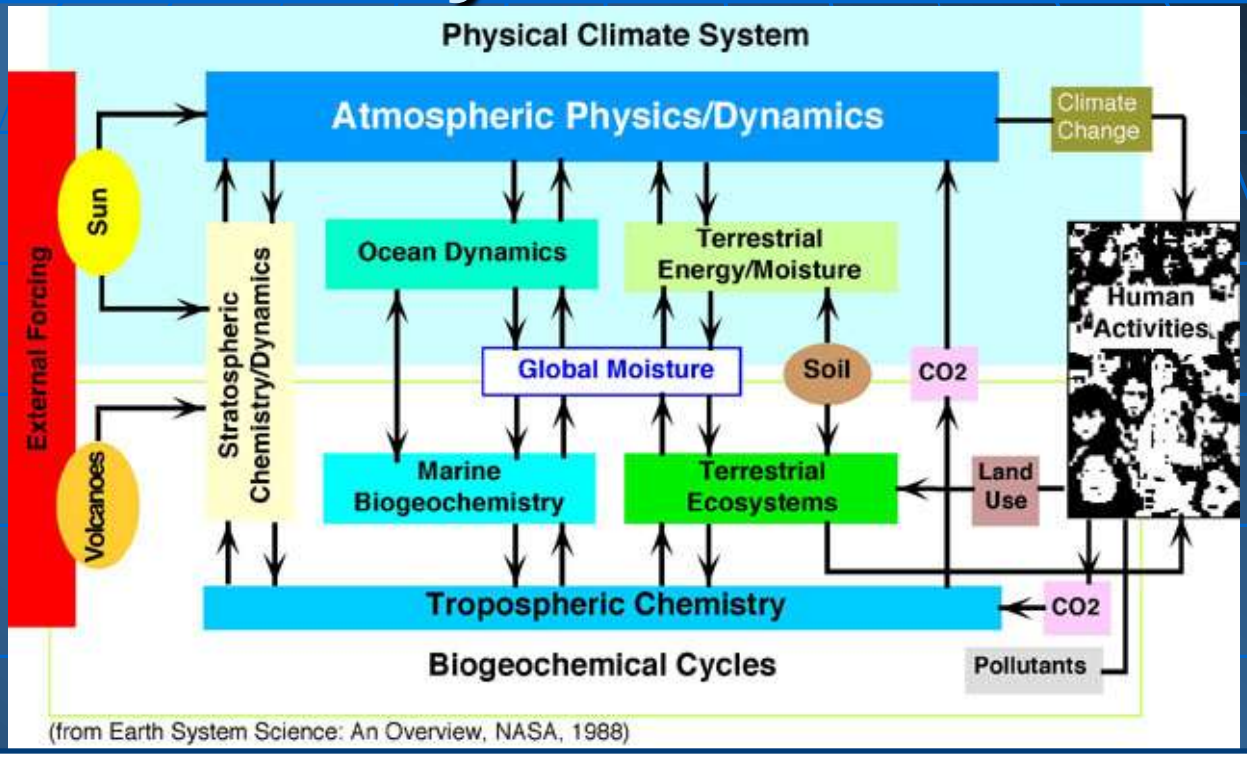
Wired, 23 June 2008 issue

Geosciences: Global & Multidisciplinary





Earth System Science



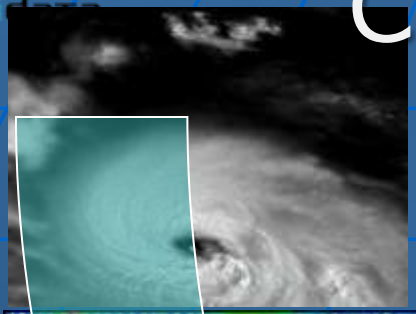
Need Geoinformatics to support ESS thinking.

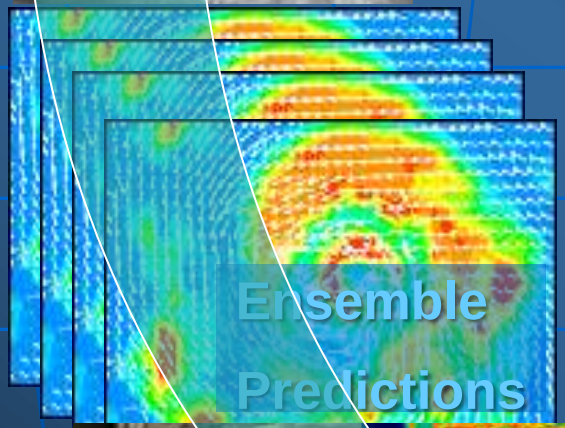
It requires data and information *integration* and knowledge *synthesis* across "systems" or domains.

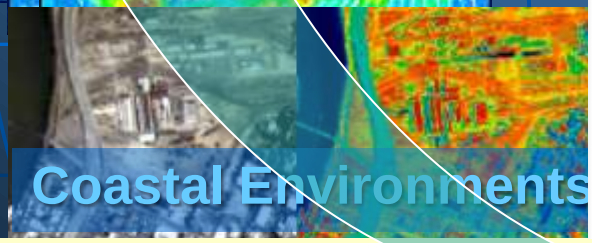
Challenge for Geoinformatics: Providing the right data, in the right format, to the right application.

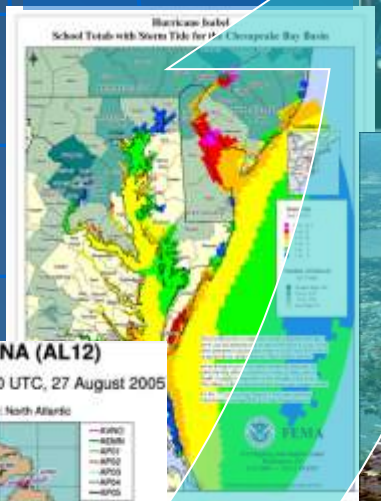
Key: Data interoperability

End-to-End Data Services Chained by Workflows





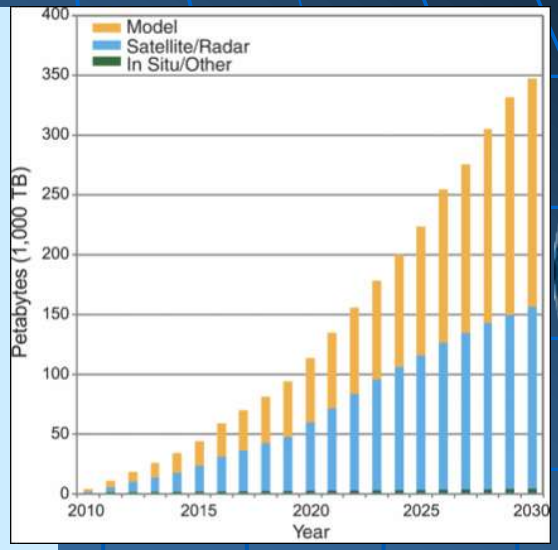


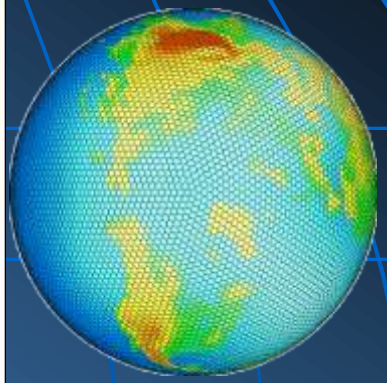


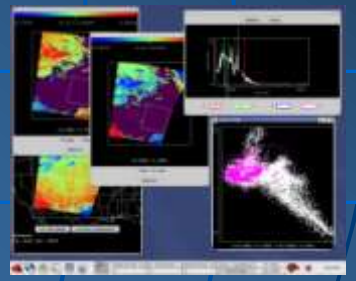
Understanding societal impact of flooding from hurricanes involves integrating data from atmospheric sciences, oceanography, hydrology, geology, and social sciences and interfacing the results with decision support systems.

“Sea of Data”

- GOES-R (2016)
  - *Hyperspectral Environmental Suite (~1600 channels)*
- JPSS (2014)
  - ~3 Tb of data/day
- Phased Array Radar, with 20 to 30-second volume scans, compared with 5-7 mins. with current radars.
- Global, high-resolution coupled models integrated in ensemble mode from days to decades



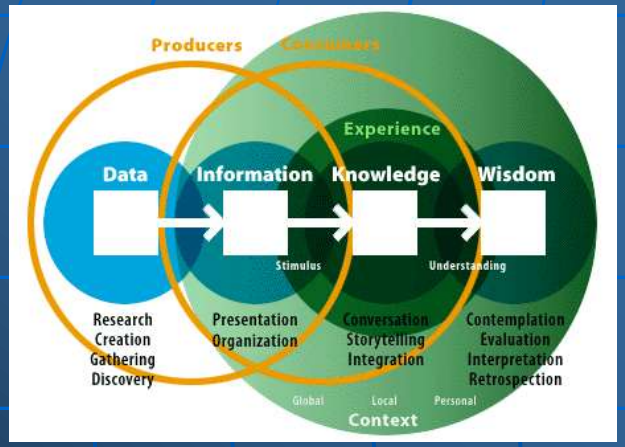


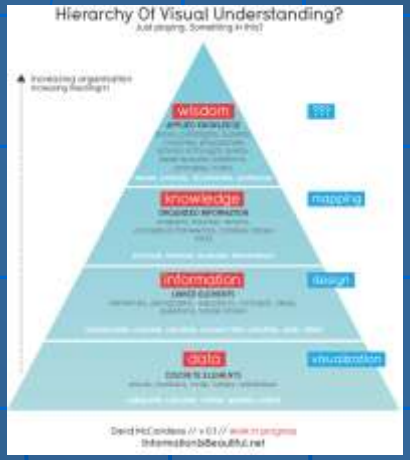


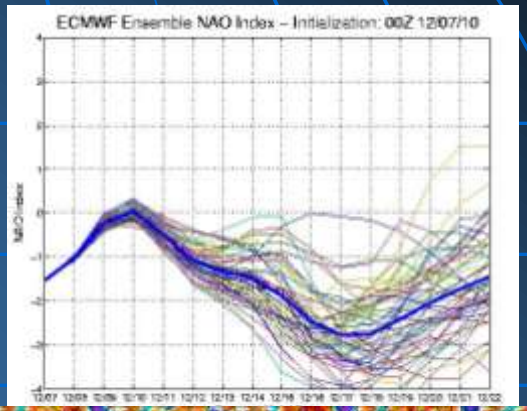
JOINT POLAR SATELLITE SYSTEM
NASA • NOAA
JPSS

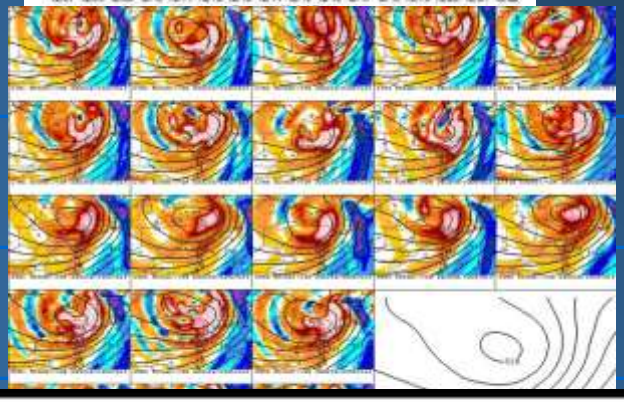
Knowledge Discovery

Hierarchy of Understanding



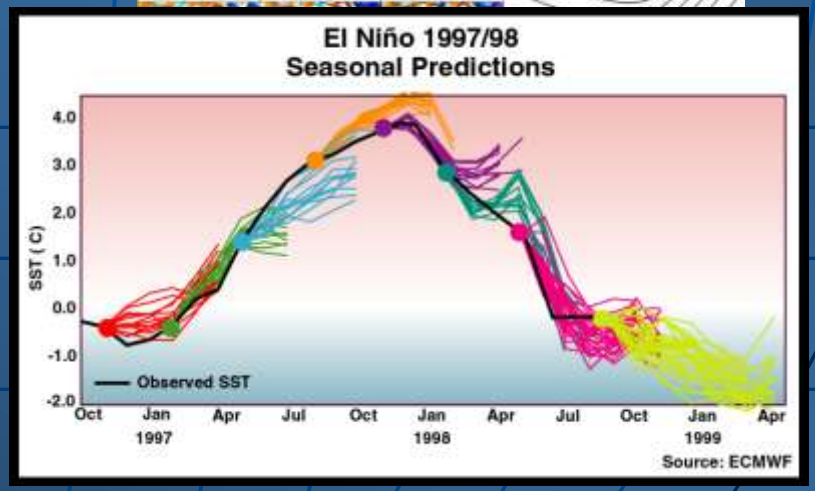






Knowledge Extraction:

- Analysis, Visualization, and Synthesis
- Data Mining (Pattern recognition, cluster analysis)



Impact on Science & Society

WMO

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE

UNEP

IPCC web sites

About IPCC

Activities

"Climate Change 2007"

The IPCC 4th Assessment Report

Click to activate and use this control

The Physical Science Basis of Climate Change the first volume of the new IPCC assessment report, has been released in Paris, 2 February 2007

PCMDI

Program for Climate Model Diagnosis and Intercomparison

Requirements for
IPCC Standard Output Contributed to the PCMDI Archive

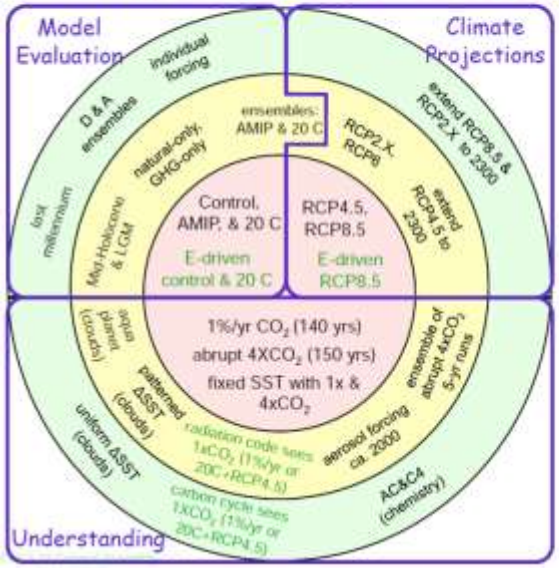
Data format, data structure, and file composition requirements:

- Data must be written through the netCDF [4] API (application program interface) and conform to the CF metadata standards [5]

IPCC Fifth Assessment Report

Introduction to CMIP5: The Experiments

Example: CMIP5 long-term suite of experiments



Green subset is for coupled carbon-cycle climate models only

Core:	≥1718 yrs
Tier 1:	≥1727 yrs
Tier 2:	≥2038 yrs

Science & Technology Facilities Council

CMIP5 in numbers

Simulations:

~90,000 years
~60 experiments
~20 modelling centres using
~30 major(*) model configurations
~2 million output datasets
~10's of petabytes of output
~2 petabytes of CMIP5 requested output
~1 petabyte of CMIP5 "replicated" output
~10 TB of land-biochemistry (from the long term experiments alone).

Of the replicants:

~ 220 TB decadal
~ 540 TB long term
~ 220 TB atmos-only

~100 TB of 3hourly atmos data!
~215 TB of ocean 3d monthly data!
~250 TB for the cloud feedbacks!

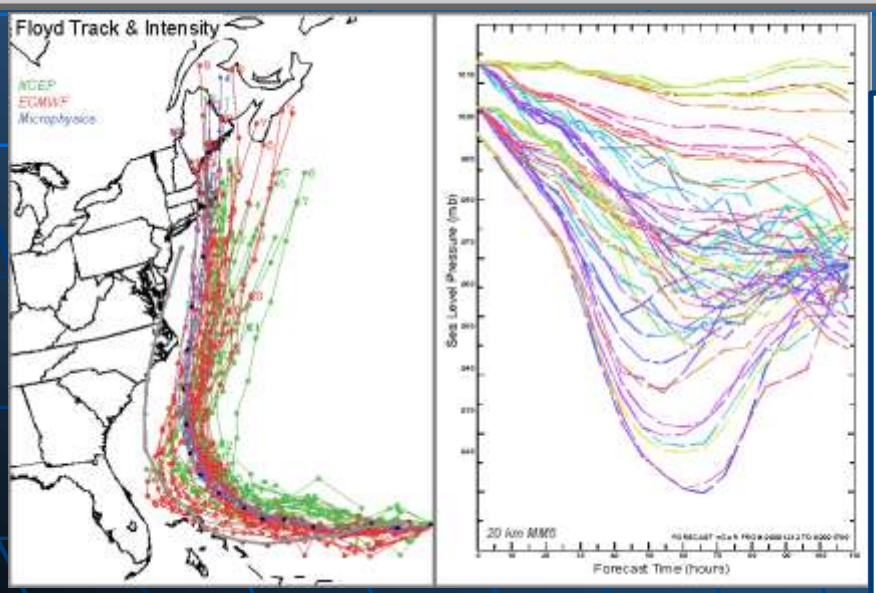
Expected Usage (@ BADC):

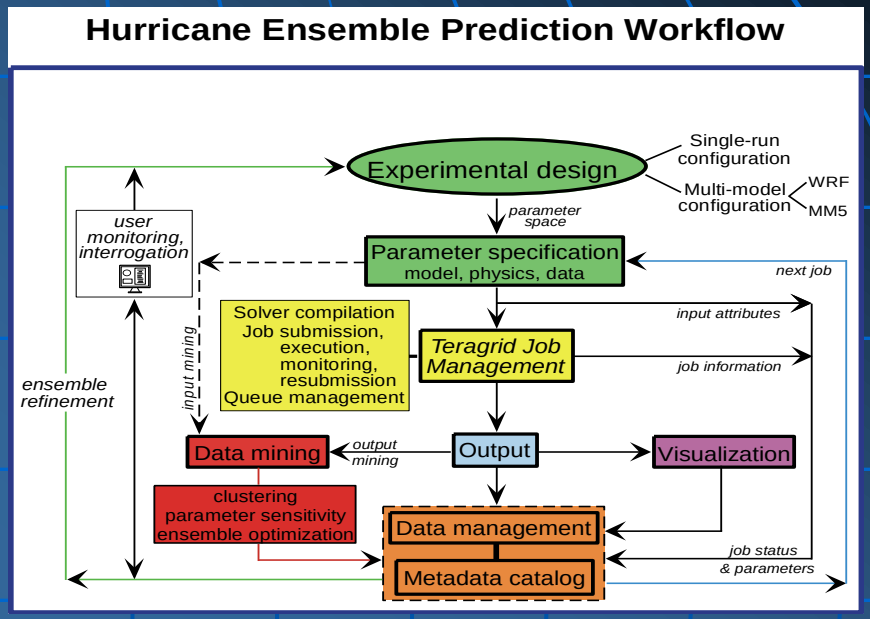
~ hundreds of users downloading at a sustained daily average rate of between 1 and 3 Gbit/s (or up to 35 TB/day from BADC ...)

Source: Bryan Lawrence, British Atmospheric Data Centre

Hurricane Ensemble Predictions

Initial condition Perturbations	*Model Physics Perturbations*	*Lateral Boundary Perturbations*	*Sea surface temperature*
Analysis fields (NCEP, ECMWF, NOGAPS)	Cumulus parameterization schemes	Perturbed outer grid (1-way nesting to inner grid)	Analysis fields (Navy, NCEP)
Bogus vortex (intensity, structure, location)	Boundary layer parameterization schemes		Perturbed analyses
BGM	Different Microphysical treatments		
Perturbed observations	Stochastic perturbations		





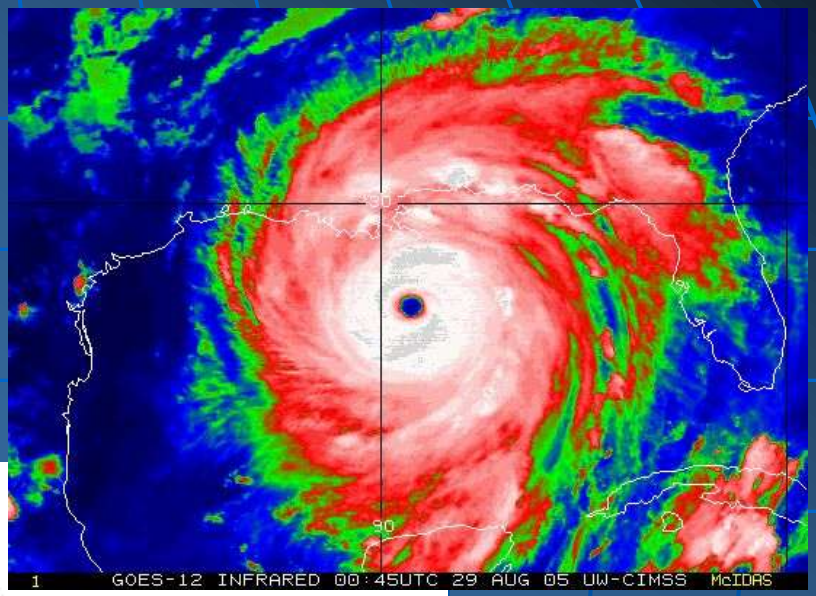
- Requires efficient, automated, end-to-end workflow systems that incorporates all of the necessary steps and components.
- Also, need an informatics system that will provide seamless access to and effective use of ALL OF THE ASSOCIATED DATA and INFORMATION.

Expected Impact on Hurricane Forecasts

Proposed Framework for Addressing the National Hurricane Research and Forecast Improvement Initiatives

NOAA's Hurricane Forecast Improvement Project

July 18, 2008



Specific metrics include:

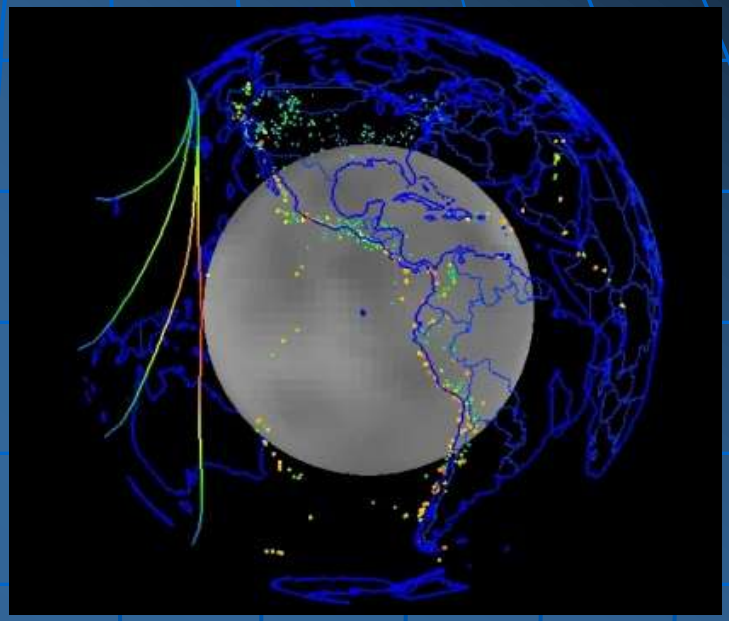
- Reduce average track error by 50% for Days 1 through 5.
- Reduce average intensity error by 50% for Days 1 through 5.
- Increase the probability of detection (POD) for rapid intensity change to 90% at Day 1 decreasing linearly to 60% at Day 5, and decrease the false alarm ratio (FAR) for rapid intensity change to 10% for Day 1 increasing linearly to 30% at Day 5.
- Extend the lead time for hurricane forecasts out to Day 7.

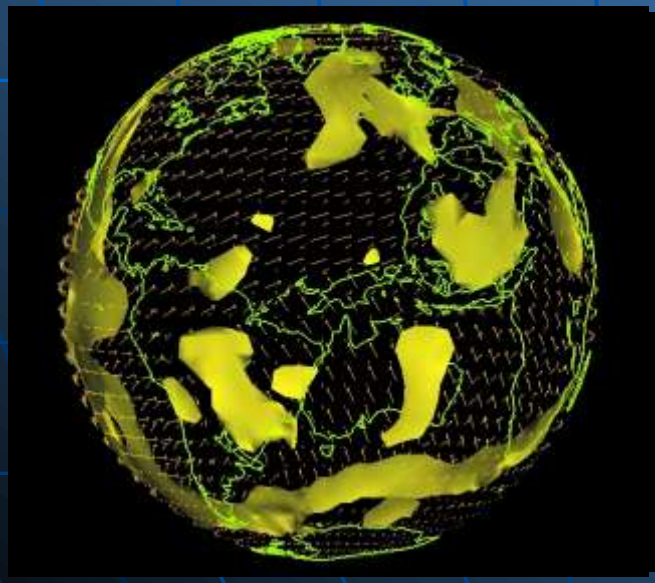
Furthermore, NOAA recognizes that addressing the broad scope of the research and technology challenges associated with improving hurricane forecasts requires interaction with, and support of, the larger research and academic community, including the open access to the data involved in this endeavor.

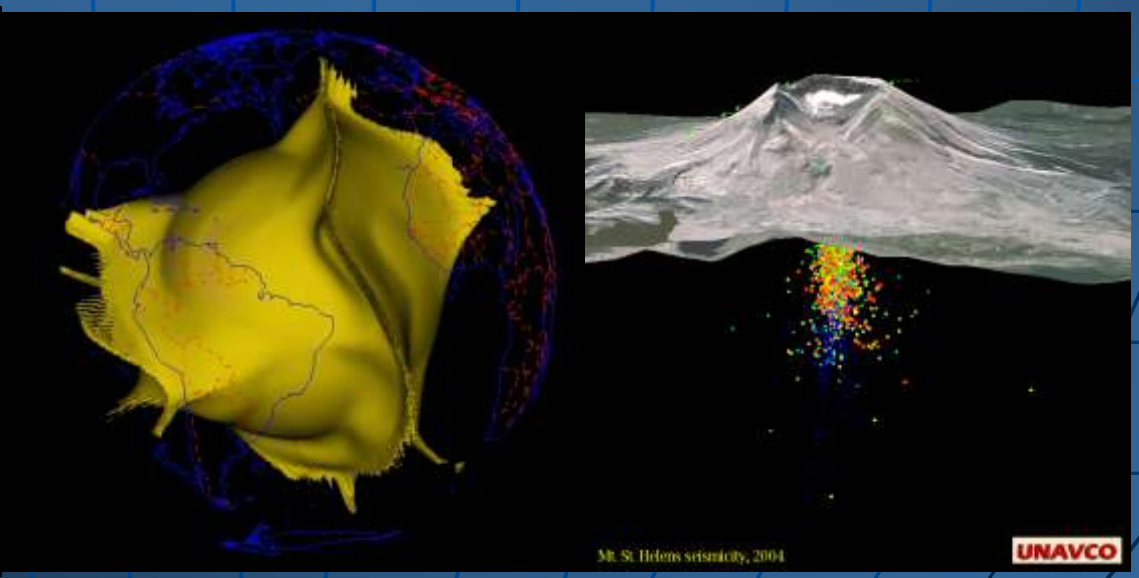
New Insights

The IDV is a multiplatform visualization and analysis tool that brings a wide range of data within a unified interface. These images were created using GEON-IDV, a version of the IDV.

Source: Unavco

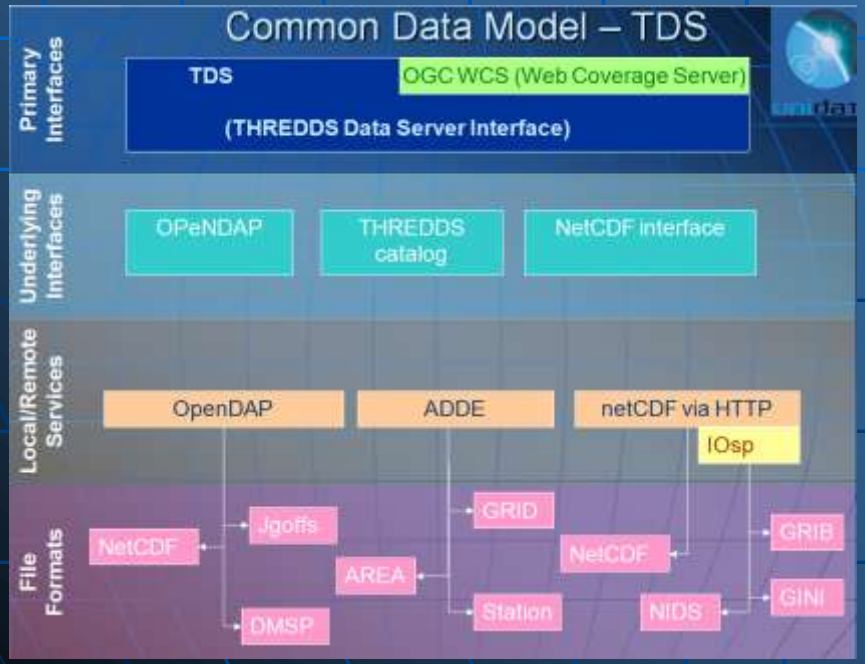


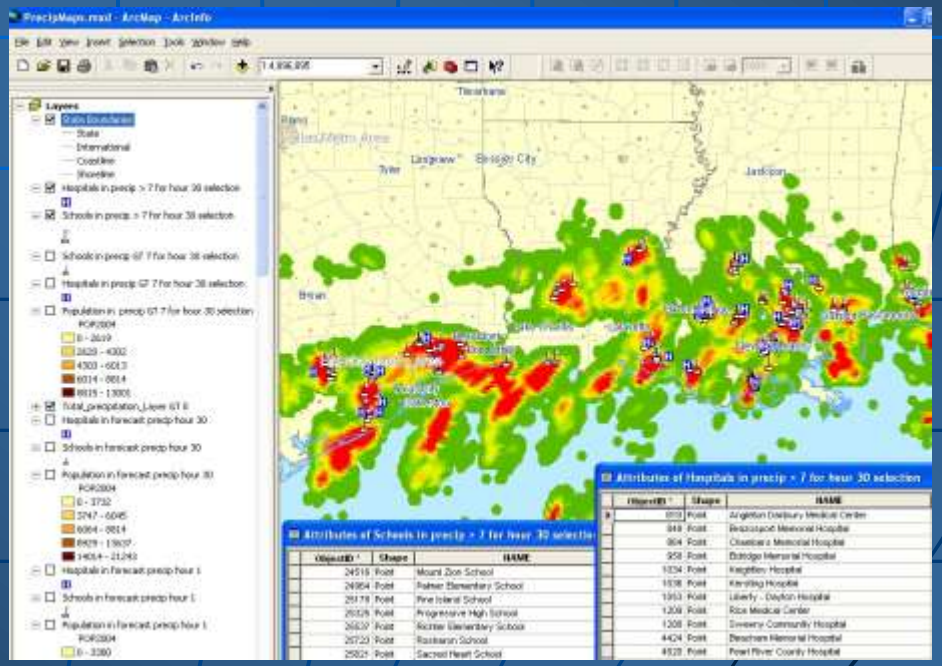




GIS Integration

- Need geospatially-enabled cyberinfrastructure so that information can be integrated for location-based understanding of events, processes, interactions, and impacts.
- GIS integration should not be an after thought. Scientific data systems need to directly enable GIS tools.





Data Citation: The Next Frontier?

- Scientific publications should be accompanied by data, algorithms, models, and parameters – need comprehensive data citation. Need transparency.
- This is not just a technical challenge, but it is also a major cultural and organizational challenge.

Proposed AMS Statement on the Importance of Data Availability in Support of Scholarly Publications

Repeatability of research results is one of the main tenants for the advancement of science; it is a benchmark upon which the reliability of results can be tested and faith placed in authors' conclusions. Some data have value that go beyond the study that generated them, and making available the data to those who can use them for other studies will further scientific advancement. Publication conclusions that cannot be replicated by independent researchers cast doubt on those conclusions and could lead others down false scientific avenues and waste time. The AMS supports a publication policy for its scholarly journals in which:

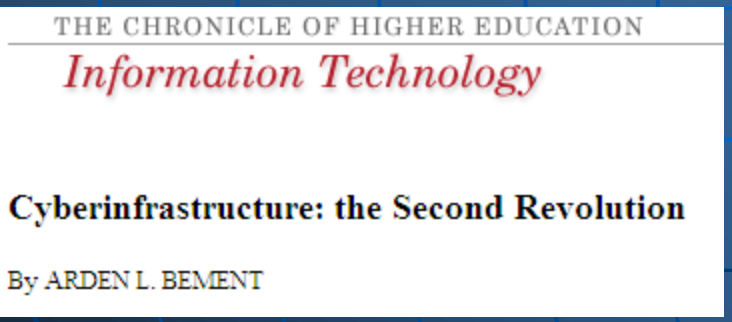
1. A condition for publication is that authors be required to agree to make data and software analysis techniques promptly available to readers upon request, subject to modest cost constraints. Exceptions may be appropriate in certain limited circumstances to preserve privacy, to assure patent protection, or for other legal reasons. Any restrictions on the availability of materials or information must be disclosed to the Editor at the time of manuscript submission.
2. Data sets must be made available to Editors and peer-reviewers if required at the time of review in order to ensure a comprehensive peer-review process. The AMS

This draft statement was summarily dismissed in less than two minutes

Networked Science

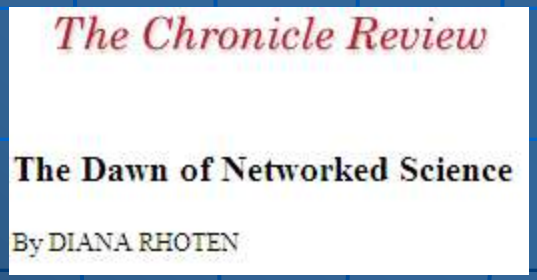
- Distributed knowledge communities working collaboratively in virtual organizations
- Networked science tackling problems never before possible before and helping to create new knowledge.

IPCC
INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE
WMO UNEP


THE CHRONICLE OF HIGHER EDUCATION
Information Technology

Cyberinfrastructure: the Second Revolution

By ARDEN L. BEMENT

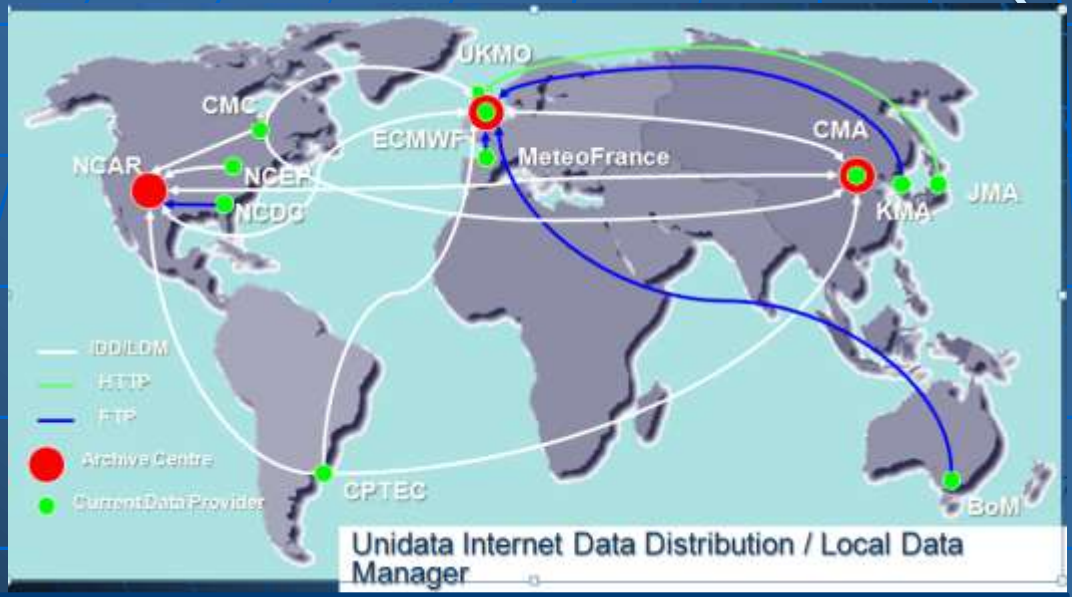

The Chronicle Review

The Dawn of Networked Science

By DIANA RHOTEN

THORPEX
A Global Atmospheric Research Programme

THORPEX Interactive Grand Global Ensemble (TIGGE) Project



Unidata Internet Data Distribution / Local Data Manager

Centre	*Ensemble members*	*Forecast length*	*Forecasts per day*
Bureau of (BoM) *	33	10 day	2
Meteorological Administration (CMA)	15	10 day	2
Meteorological Service of (MSC)	21	16 day	2
Centra de Previsao de Tempo e Estudos Climatico, Brazil (CPTEC)	15	15 day	2
European Centre for Medium-Range Weather Forecasts (ECMWF)	51	15 day	2
Meteorological Agency (JMA)	51	9 day	1
Meteorological Administration (KMA) *	17	10 day	2
Météo-France	35	4.5 day	2
National Centers for Environmental Prediction, (NCEP)	21	16 day	4
Met (UKMO)	24	15 day	2

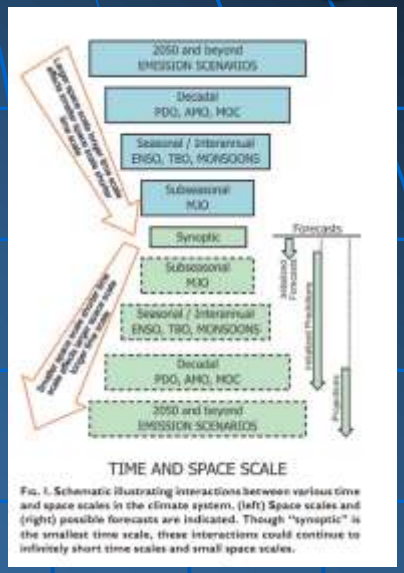
The TIGGE database now has ~0.5 Pb of data and it is growing at a rate of 500 Gb/day.

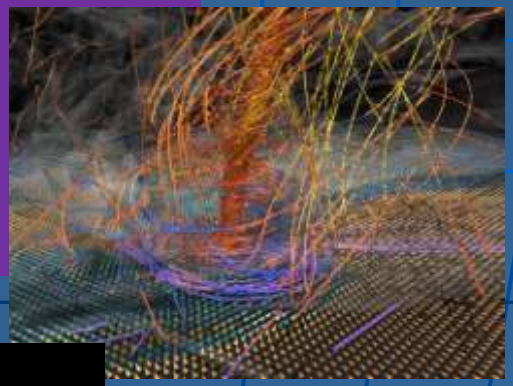
Users processed about 37 Tb and downloaded ~2.8 Tb of data in December 2010.

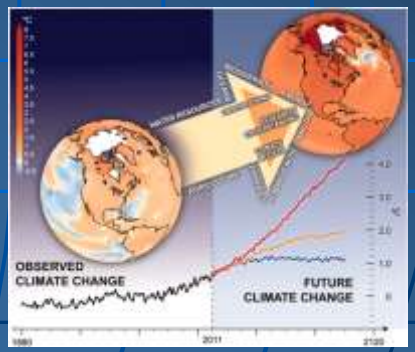
Geoinformatics Challenges

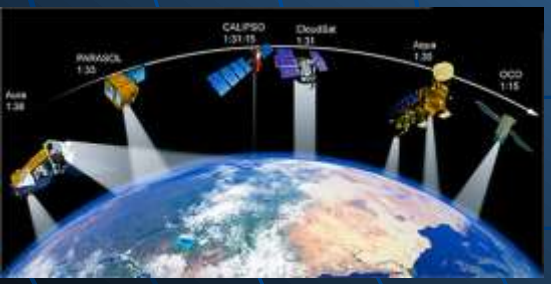
To solve grand challenge problems we need:

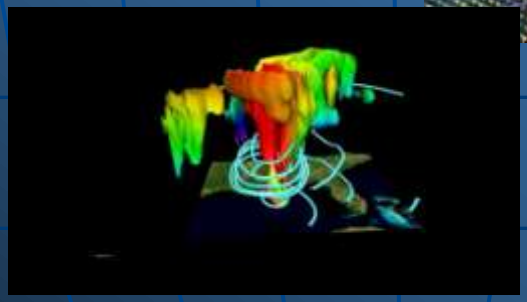
- Data intensive computing
- Systems for multiscale & multidisciplinary synthesis
- Integrative analysis tools
- A Skilled Workforce
- Collaborations

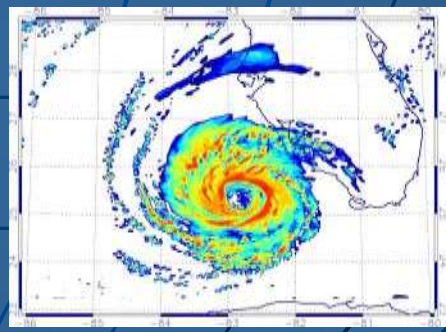








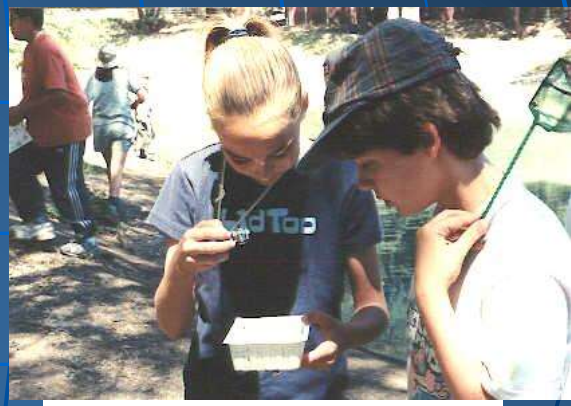




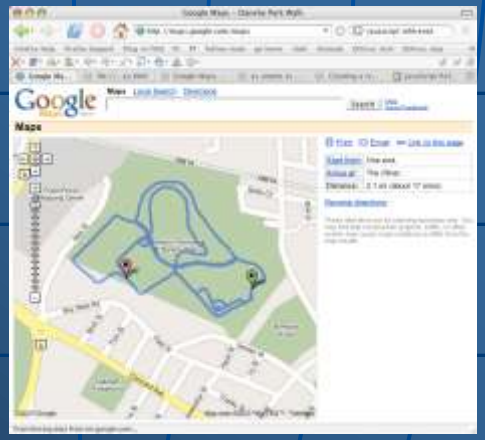
Opportunities

- GPS-enabled smartphones
- Mobile sensors
- Cloud Computing
- Social Networking

facebook



Unprecedented enablement of Citizen Science



Concluding Remarks

- We live in an exciting era in which advances in computing and communication technologies, coupled with a new generation of geoinformatics, are accelerating scientific research, creating new knowledge, and leading to new discoveries at an unprecedented rate.
- Meetings like this play an ever more important role in bringing people together to address opportunities and challenges and fostering new partnerships.

Thank you!