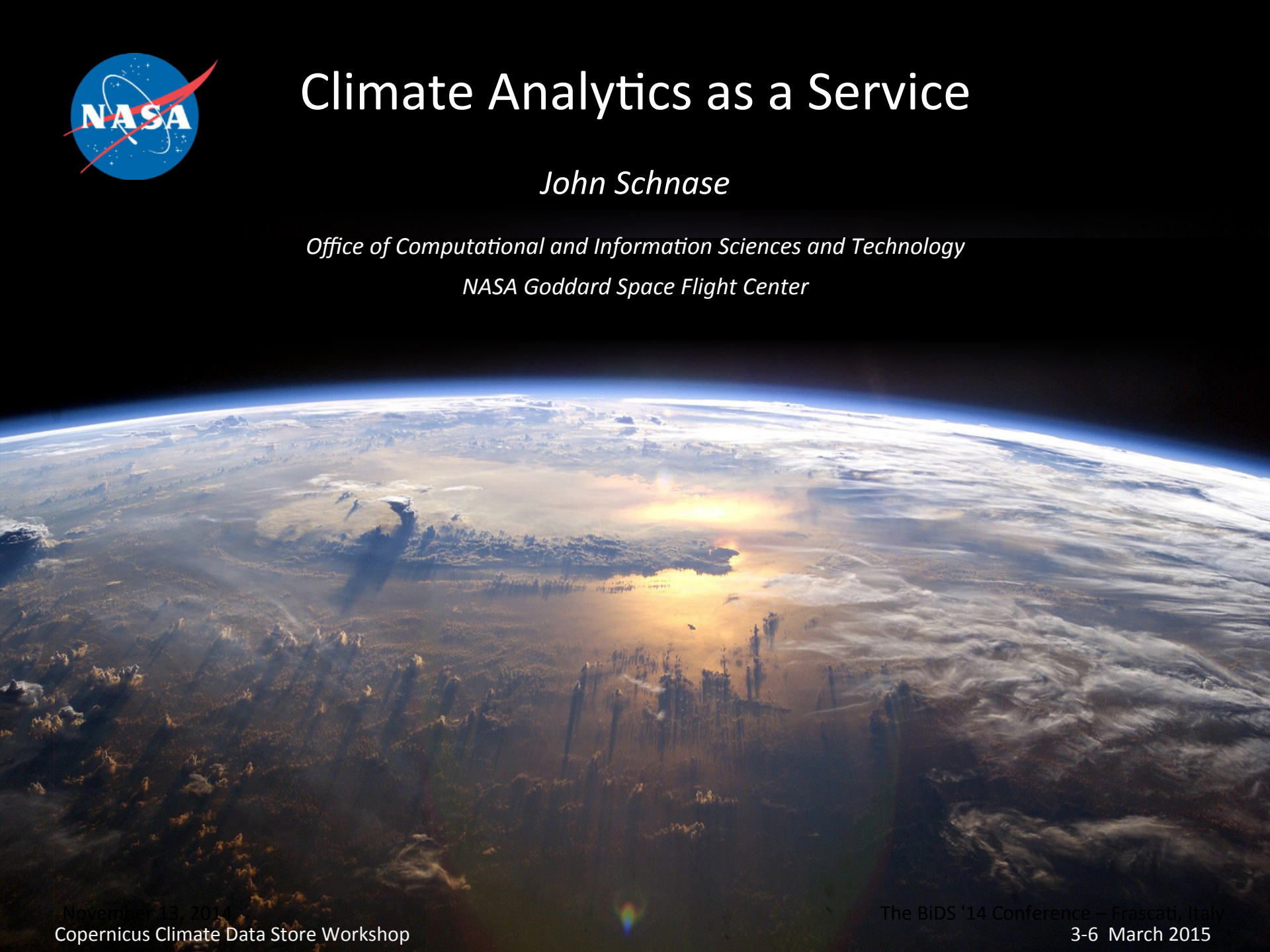
Climate Analytics as a Service

John Schnase

Office of Computational and Information Sciences and Technology

NASA Goddard Space Flight Center



November 13, 2014
Copernicus Climate Data Store Workshop

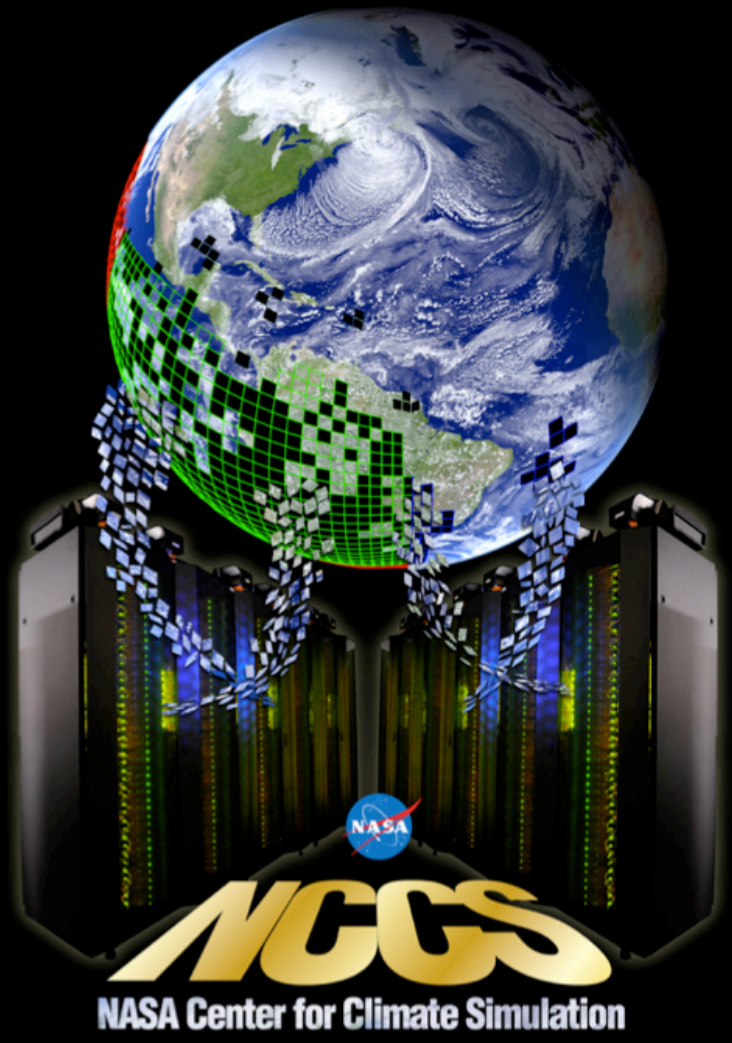
The BiDS '14 Conference – Frascati, Italy
3-6 March 2015

NASA Center for Climate Simulation

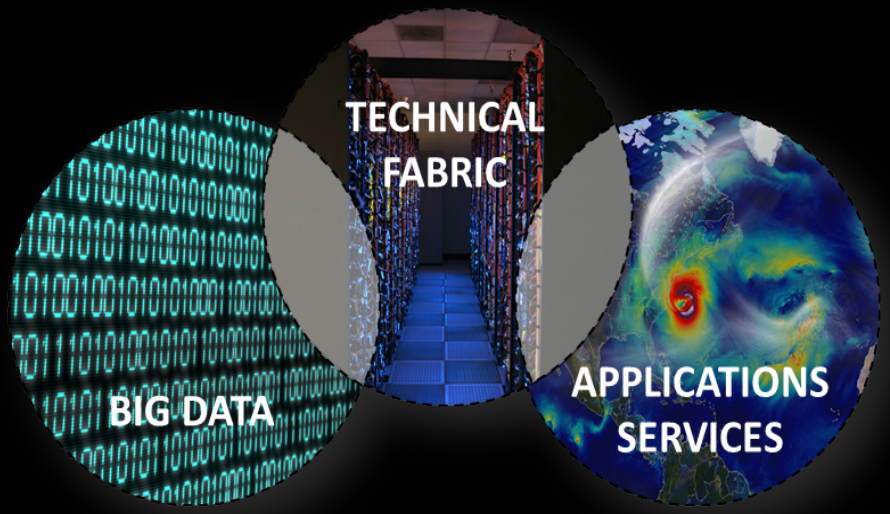
Provides an integrated high-end computing environment designed to support the specialized requirements of Climate and Weather modeling

- State-of-the-art high-performance computing, data storage, and networking technologies
- Advanced analysis and visualization environments
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services

http://www.nccs.nasa.gov

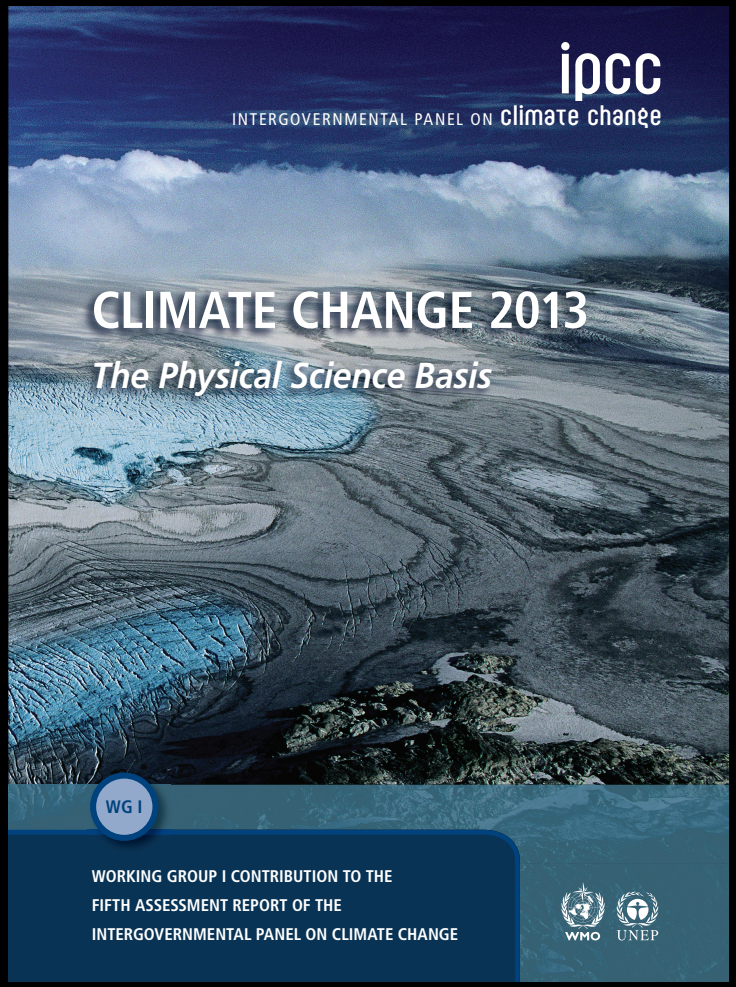


Climate Analytics-as-a-Service



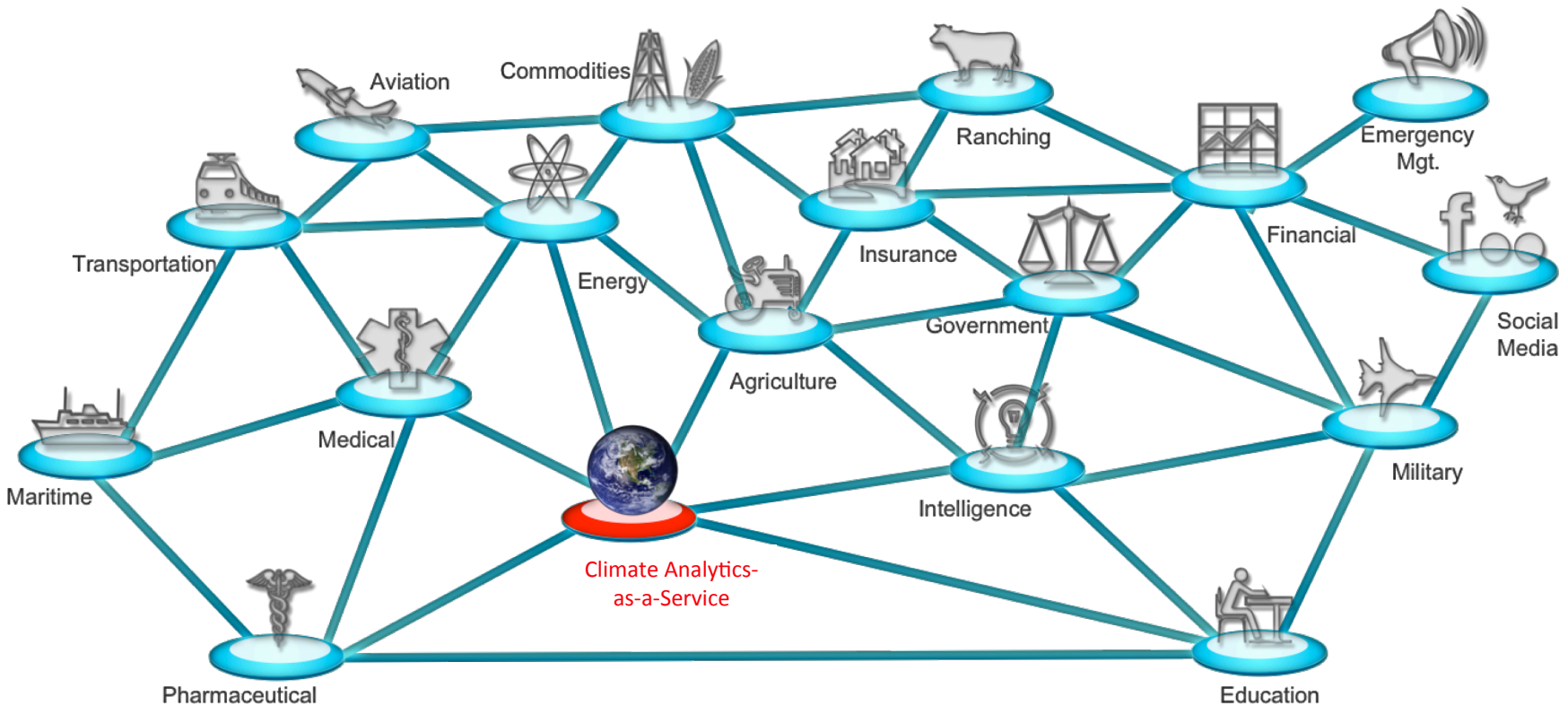
For research to be affordable, data analysis must increasingly be done where the data sets reside …

— Gordon Moore, 2009



Climate Analytics-as-a-Service

Climate Analytics-as-a-Service (CAaaS) is contributing to a global network of sector-specific data, driving innovation and discovery …



Climate Analytics-as-a-Service

What are the critical elements needed to deliver Climate Analytics-as-a-Service?

Data

Relevance
Co-location

Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …

Exposure

Convenience
Extensible

Capabilities need to be easy to use and facilitate community engagement and adaptive construction …

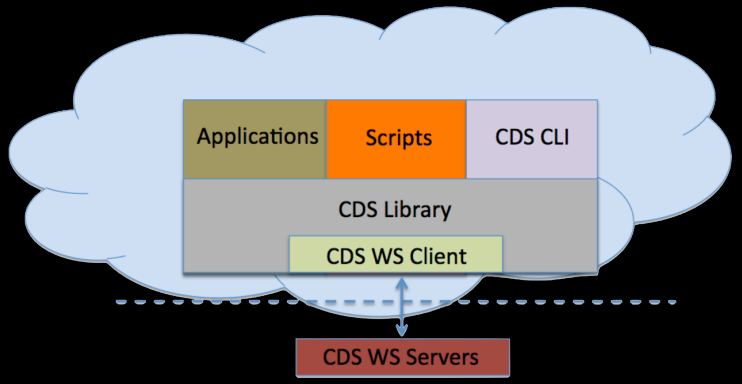
High-Performance Compute/Storage Fabric

Storage-proximal analytics
Canonical operations

Data can't move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …

MERRA Analytic Services

Climate Data Services API



Data

Relevance
Co-location

Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …

Exposure

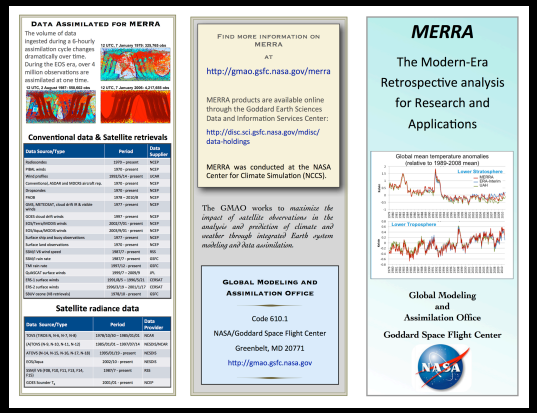
Convenience
Extensible

Capabilities need to be easy to use and facilitate community engagement and adaptive construction …

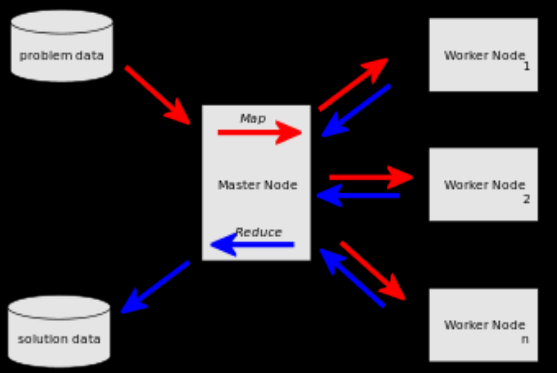
High-Performance Compute/Storage Fabric

Storage-proximal analytics
Canonical operations

Data can't move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …



MERRA Reanalysis

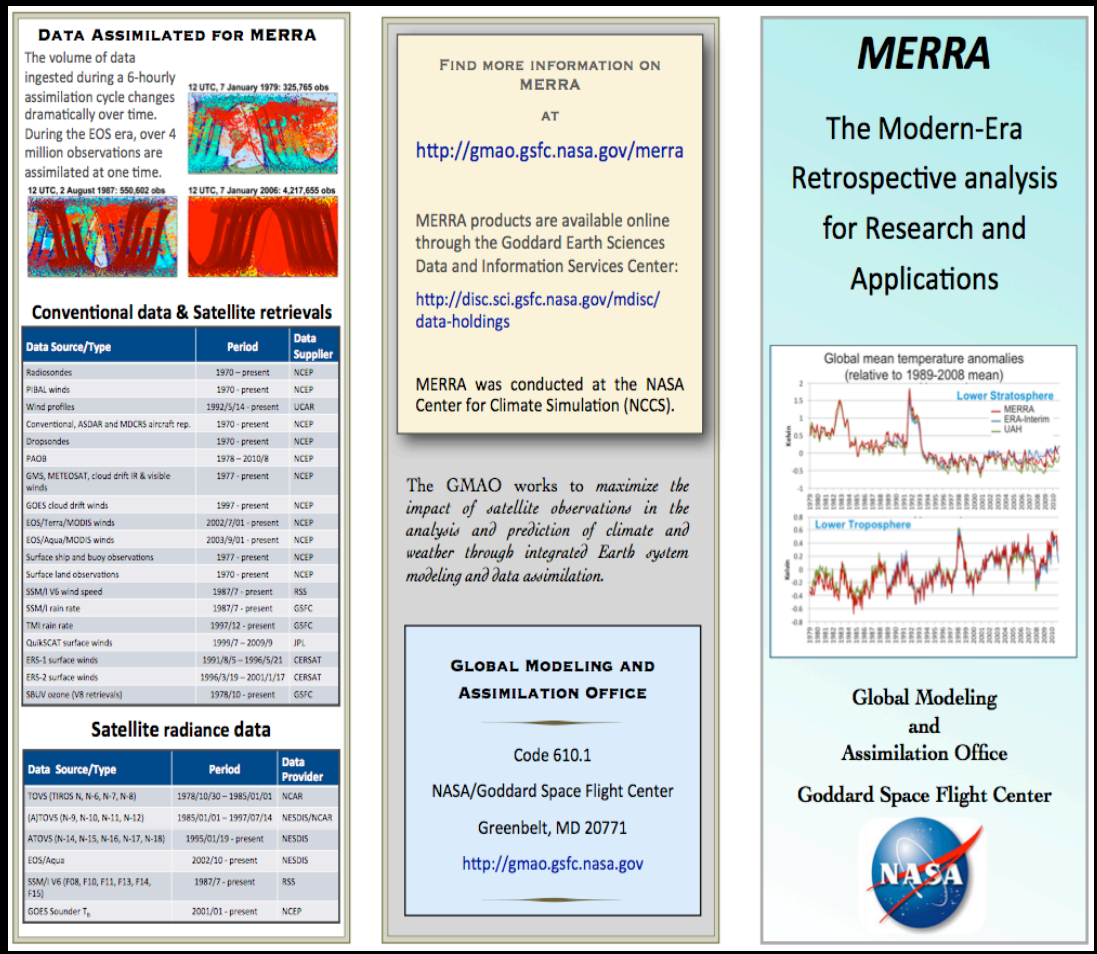


MapReduce

MERRA Analytic Services

Modern Era-Retrospective Analysis for Research and Applications

- Source: Global Modeling and Assimilation Office (GMAO)
- Input: 114 observation types (land, sea, air, space) into "frozen" numerical model.
  (~4 million observations/day)
- Output: a global temporally and spatially consistent synthesis of 26 key climate variables. (~418 under the hood.)
- Spatial resolution: 1/2° latitude × 2/3° longitude × 42 vertical levels extending through the stratosphere.
- Temporal resolution: 6-hours for three-dimensional, full spatial resolution, extending from 1979-Present.
- ~ 200 TB, but MERRA II is on the way …



DATA ASSIMILATED FOR MERRA

The volume of data ingested during a 6-hourly assimilation cycle changes dramatically over time. During the EOS era, over 4 million observations are assimilated at one time.

12 UTC, 7 January 1979: 325,765 obs

12 UTC, 2 August 1987: 550,602 obs

12 UTC, 7 January 2006: 4,217,655 obs

Conventional data & Satellite retrievals

Data Source/Type	Period	Data Supplier
Radiosondes	1970 – present	NCEP
PIBAL winds	1970 - present	NCEP
Wind profilers	1992/5/14 - present	UCAR
Conventional, ASDAR and MDCRS aircraft rep.	1970 - present	NCEP
Dropsondes	1970 - present	NCEP
PAOB	1978 – 2010/8	NCEP
GMS, METEOSAT, cloud drift IR & visible winds	1977 - present	NCEP
GOES cloud drift winds	1997 - present	NCEP
EOS/Terra/MODIS winds	2002/7/01 - present	NCEP
EOS/Aqua/MODIS winds	2003/9/01 - present	NCEP
Surface ship and buoy observations	1977 - present	NCEP
Surface land observations	1970 - present	NCEP
SSM/I V6 wind speed	1987/7 - present	RSS
SSM/I rain rate	1987/7 - present	GSFC
TMI rain rate	1997/12 - present	GSFC
QuikSCAT surface winds	1999/7 – 2009/9	JPL
ERS-1 surface winds	1991/8/5 – 1996/5/21	CERSAT
ERS-2 surface winds	1996/3/19 – 2001/1/17	CERSAT
SBUV ozone (V8 retrievals)	1978/10 - present	GSFC

Satellite radiance data

Data Source/Type	Period	Data Provider
TOVS (TIROS N, N-6, N-7, N-8)	1978/10/30 – 1985/01/01	NCAR
(A)TOVS (N-9, N-10, N-11, N-12)	1985/01/01 – 1997/07/14	NESDIS/NCAR
ATOVS (N-14, N-15, N-16, N-17, N-18)	1995/01/19 - present	NESDIS
EOS/Aqua	2002/10 - present	NESDIS
SSM/I V6 (F08, F10, F11, F13, F14, F15)	1987/7 - present	RSS
GOES Sounder T_b	2001/01 - present	NCEP

FIND MORE INFORMATION ON MERRA

AT

http://gmao.gsfc.nasa.gov/merra

MERRA products are available online through the Goddard Earth Sciences Data and Information Services Center:

http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings

MERRA was conducted at the NASA Center for Climate Simulation (NCCS).

The GMAO works to *maximize the impact of satellite observations in the analysis and prediction of climate and weather through integrated Earth system modeling and data assimilation.*

GLOBAL MODELING AND ASSIMILATION OFFICE

Code 610.1

NASA/Goddard Space Flight Center

Greenbelt, MD 20771

http://gmao.gsfc.nasa.gov

MERRA

The Modern-Era Retrospective analysis for Research and Applications

Global mean temperature anomalies (relative to 1989-2008 mean)

Lower Stratosphere

MERRA

ERA-Interim

UAH

Lower Troposphere

Global Modeling and Assimilation Office

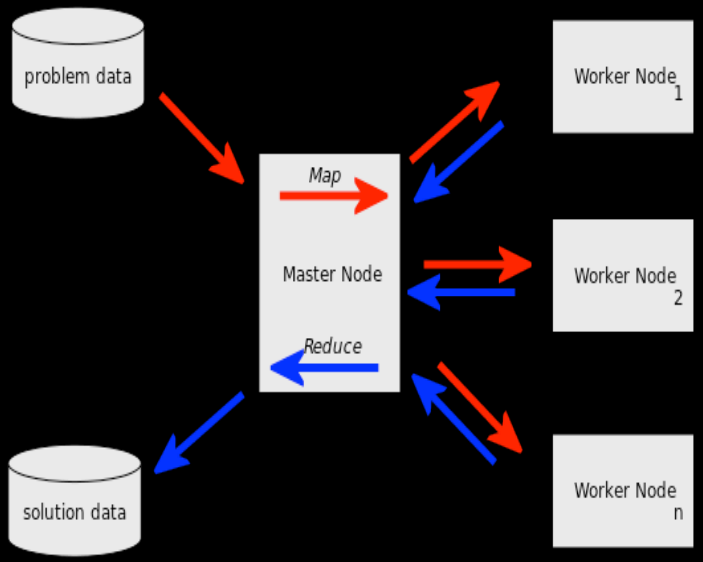
Goddard Space Flight Center

NASA

HP Data-Proximal Analytics

The Basic MapReduce Paradigm ...

- MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers.
- Computational processing can occur on data stored either in a filesystem (unstructured) or in a database (structured).
- MapReduce can take advantage of locality of data, processing data on or near the storage assets to decrease transmission of data.



- "Map" step: The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node.
- "Reduce" step: The master node then collects the answers to all the sub-problems and combines them to form the output – the answer to the problem it was originally trying to solve.

Adaptive Analytics

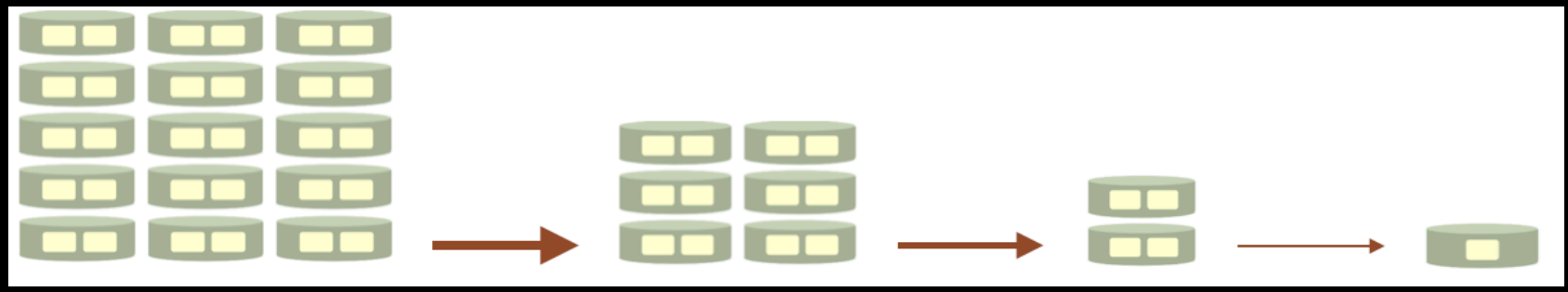
"Canonical Ops" ...

- We're building in to our analytic services near-storage, early-stage analytical operations that represent a common starting point in many analysis workflows in many domains. For example, average, max, min, standard deviation, sum, count, and difference operations of the general form:

$$result \leftarrow average(var, (t_0,t_1), ((x_0,y_0,z_0),(x_1,y_1,z_1))),$$

  that return, in this example, the average of a variable when given a variable name, temporal extent, and spatial extent.

- Built-in canonical ops exploit complexity stratification to optimize efficiencies along the workflow chain ...



- *Large amounts of unstructued data*
- *Simple, common, general-purpose operations*

- *Highly structured, tailored, reduced, refined analytic products*
- *Specialized tools, models, operations*

Domain Harmonized API

The Climate Data Services API...

- Based on the data flow interactions of the Open Archival Information System (OAIS) Reference Model
- Addresses climate science's "Big Data" challenge by integrating principals of archive data management with high-performance data analytics
  - – Makes it easier to integrate high-performance analytics into existing digital preservation systems
  - – Makes it easier to use high-performance analytics to create dynamically generated objects
- CDS API Methods

Ingest – Submit/register a Submission Information Package (SIP).

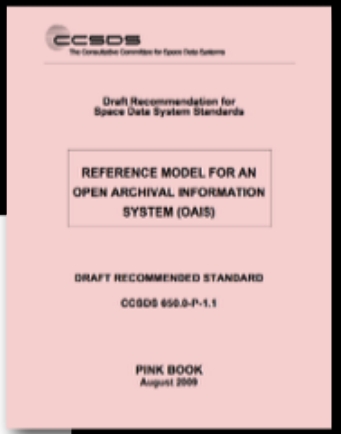
Query – Retrieve data from a pre-determined service request (synchronous).

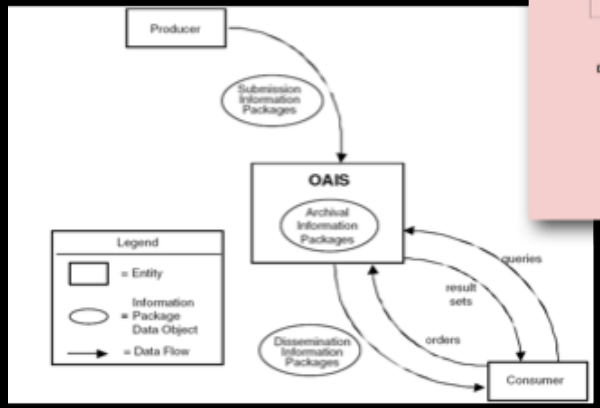
Order – Request data from a pre-determined service request (asynchronous).

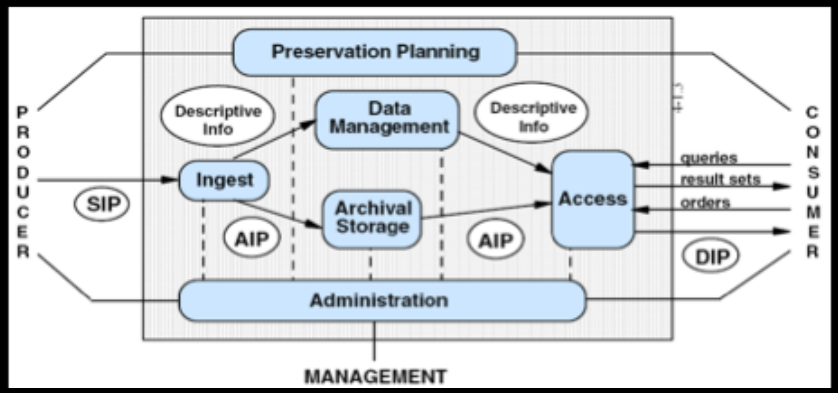
Download – Retrieve a Dissemination Information Package (DIP).

Status – Track progress of service activity.

Execute – Initiate a service-definable extension. Allows for parameterized growth without API change.







MERRA Analytic Services

MERRA/AS System

- Entire MERRA collection in a TRL 8 mission qualified analytic data service
- Virtual Hadoop Clusters
  - Three Hadoop clusters on the same hardware using containers: Test, Pre-Production, Production
  - Established agile protocols for testing new software and promoting the software changes into production
- Climate Data Services API
  - MERRA Analytic Service, Persistence Service, libraries, command interpreter, and client distributuion package
- Documentation / Administrative Infrastructure
  - Using standard NCCS practices for configuration management and authentication
  - Complete documentation and system administrative infrastructure
- Established beta test user community

Hardware Configuration

- 36 node Dell cluster (11.7 TF Peak)
- 576 total cores (Intel 2.6 GHz SandyBridge)
- 2,304 GB of RAM (64 GB per node)
- 1,296 TB of RAW storage (36 TB per node)
- FDR Infiniband

Wei Experiment

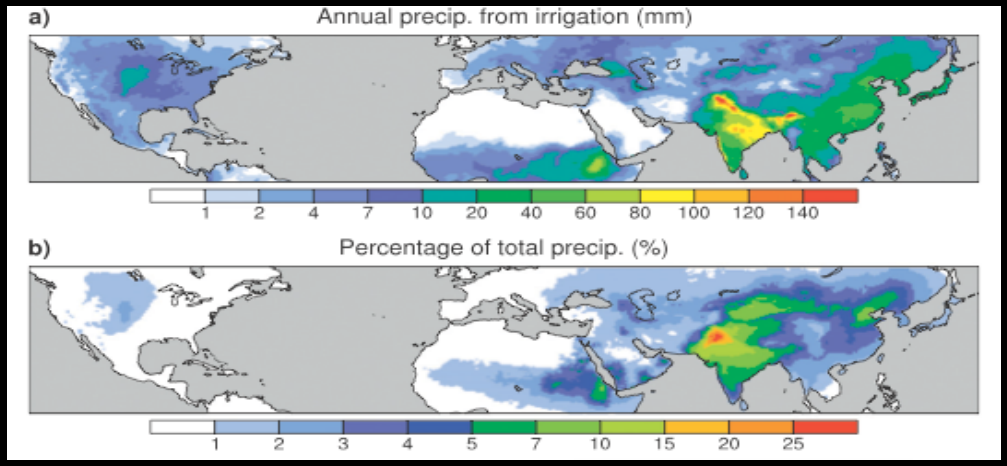
An Estimation of the Contribution of Irrigation to Precipitation Using MERRA

Study Areas
- Nile Valley
- North China
- California Central Valley
- Northern India/Pakistan

Other Requirements
- 1979 – 2002
- 6-hr time steps
- 18 atmospheric levels

Variables Needed
- Humidity
- Wind speed
- Temperature

Data Wrangled:
- 23 x 365 x 4 x 4 x 18 x 3 = 7,253,280 layers …



FEBRUARY 2013 WEI ET AL. 275

Where Does the Irrigation Water Go? An Estimate of the Contribution of Irrigation to Precipitation Using MERRA

JIANGFENG WEI*
Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

PAUL A. DIRMEYER
Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

DOMINIK WISSER
Department of Physical Geography, Utrecht University, Utrecht, Netherlands

MICHAEL G. BOSILOVICH
Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

DAVID M. MOCKO
SAIC and Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 24 May 2012, in final form 21 September 2012)

ABSTRACT

Irrigation is an important human activity that may impact local and regional climate, but current climate model simulations and data assimilation systems generally do not explicitly include it. The European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim) shows more irrigation signal in surface evapotranspiration (ET) than the Modern-Era Retrospective Analysis for Research and Applications (MERRA) because ERA-Interim adjusts soil moisture according to the observed surface temperature and humidity while MERRA has no explicit consideration of irrigation at the surface. But, when compared with the results from a hydrological model with detailed considerations of agriculture, the ET from both reanalyses show large deficiencies in capturing the impact of irrigation. Here, a back-trajectory method is used to estimate the contribution of irrigation to precipitation over local and surrounding regions, using MERRA with observation-based corrections and added irrigation-caused ET increase from the hydrological model. Results show substantial contributions of irrigation to precipitation over heavily irrigated regions in Asia, but the precipitation increase is much less than the ET increase over most areas, indicating that irrigation could lead to water deficits over these regions. For the same increase in ET, precipitation increases are larger over wetter areas where convection is more easily triggered, but the percentage increase in precipitation is similar for different areas. There are substantial regional differences in the patterns of irrigation impact, but, for all the studied regions, the highest percentage contribution to precipitation is over local land.

* Current affiliation: Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas.

Corresponding author address: Jiangfeng Wei, Jackson School of Geosciences, The University of Texas at Austin, 2275 Speedway C9000, Austin, TX 78712.
E-mail: jwei@utexas.edu

DOI: 10.1175/JHM-D-12-079.1

1. Introduction

Irrigation is an important human activity that has the potential to impact local and regional climate through the hydrological cycle and surface energy balance (e.g., Chase et al. 1999; Pielke et al. 2011). About two-thirds of the global freshwater withdrawals from surface and underground are used for agriculture (Shiklomanov 2000),

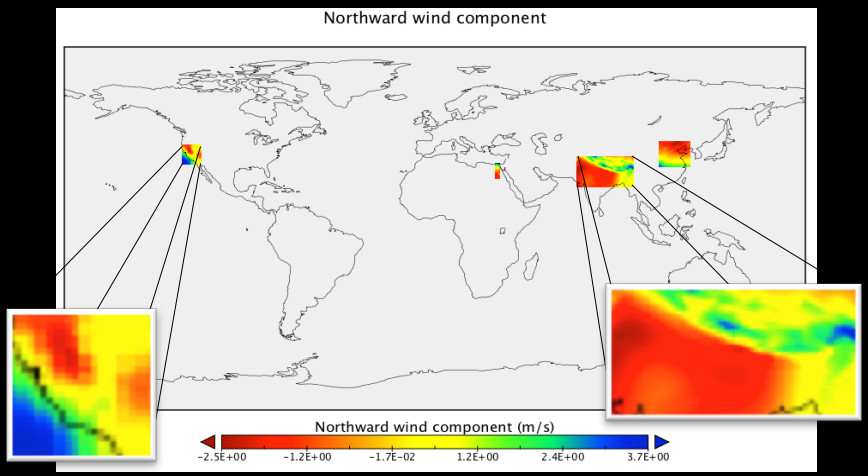
Wei Experiment

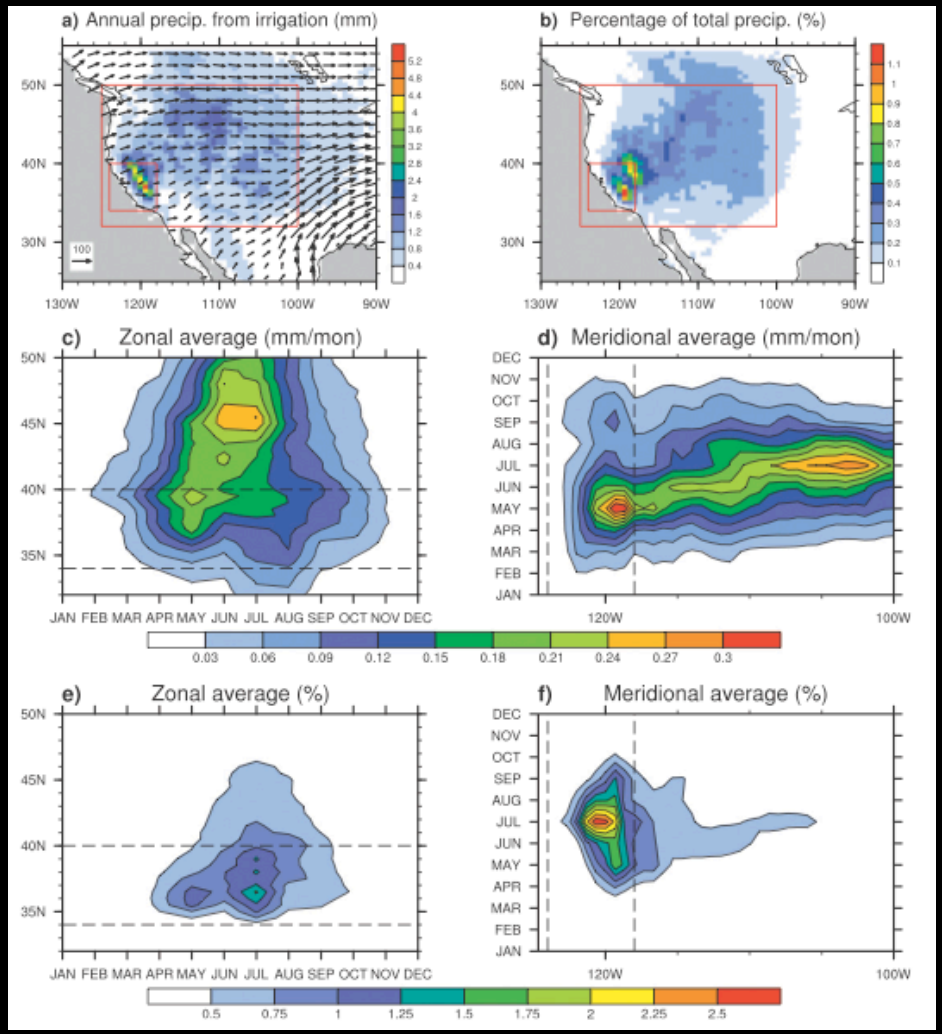
Traditional Approach

- 8.4 TB moved from archive (3 months)
- Clipping / averaging (days – weeks)

With MERRA/AS and the CDS API …

- Clipping / averaging (2.5 minutes)
- 500 MB of final product moved to local workstation (8 minutes)





Nadeau Experiment

Temperature Anomaly

- Coverage: Global
- Period: 1 month
- Collection: instM_3d_ana_Np
- Time span: January — December 2011
- Levels: 1 – 42 (0.1 hPa – 1000 hPa)

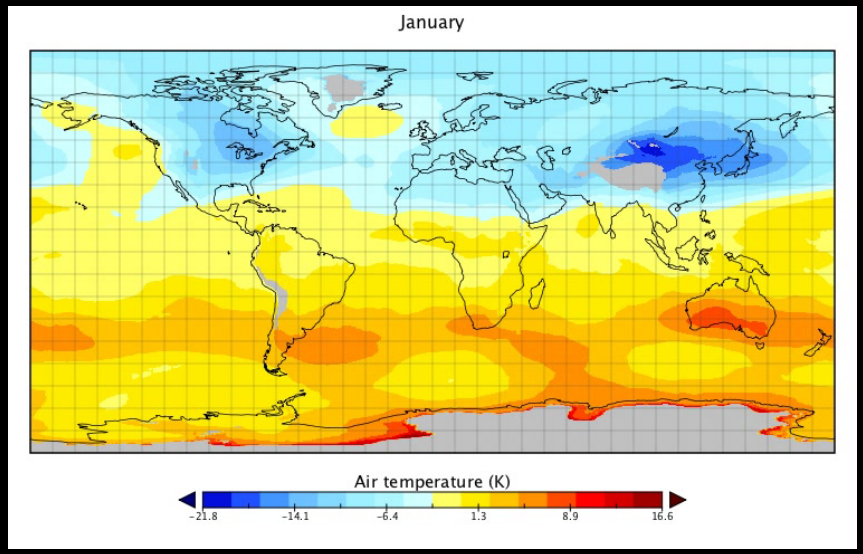
Traditional Approach

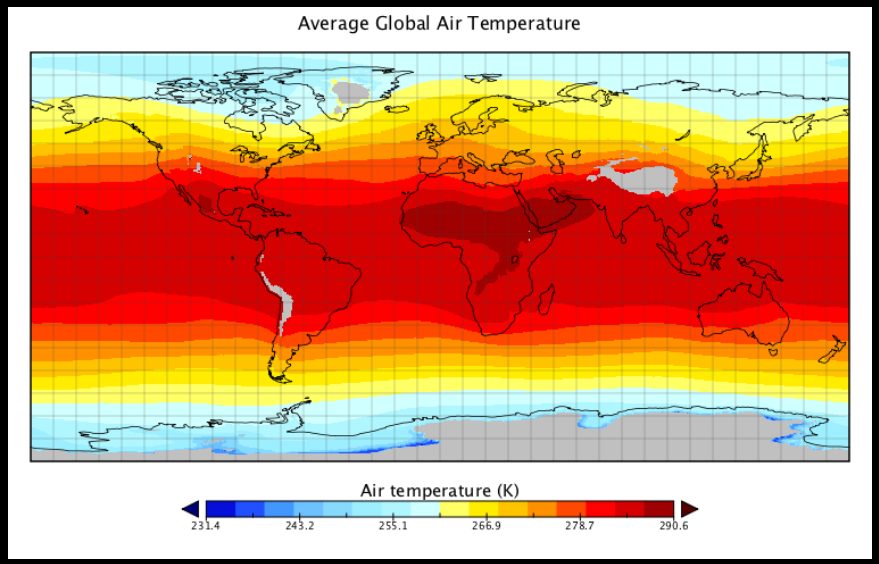
- Find and order from archive (hours – days)
- Transfer ~10 GB (~3 hours)*
- Client-side clip/compute, GrADS (~1.5 days)

With MERRA/AS and the CDS API …

- One line in a python script
- 3 minutes run time
- Final product ~0.5 GB (10 minutes to transfer)

** Assuming 10 Mbps average US internet speed with 25% overhead …*





MERRA/AS Beta Test Participants

- 21 individual testers across government, corporate, and university
- 10 projects using CDS API for access
  - NASA's ABoVE Campaign
  - iPlant Collaborative
  - Iowa State University/ClimateMonkeys
  - DataNet Federation Consortium
  - George Mason University
  - RECOVER Wildfire Decision Support
  - Invasive Species Forecasting System
- Usage Statistics
  - Total Orders: 19,067
  - Total Downloads: 35,673
  - Total Status Checks: 427,425

Example MERRA/AS Beta Test Use Cases (Research and Applications)

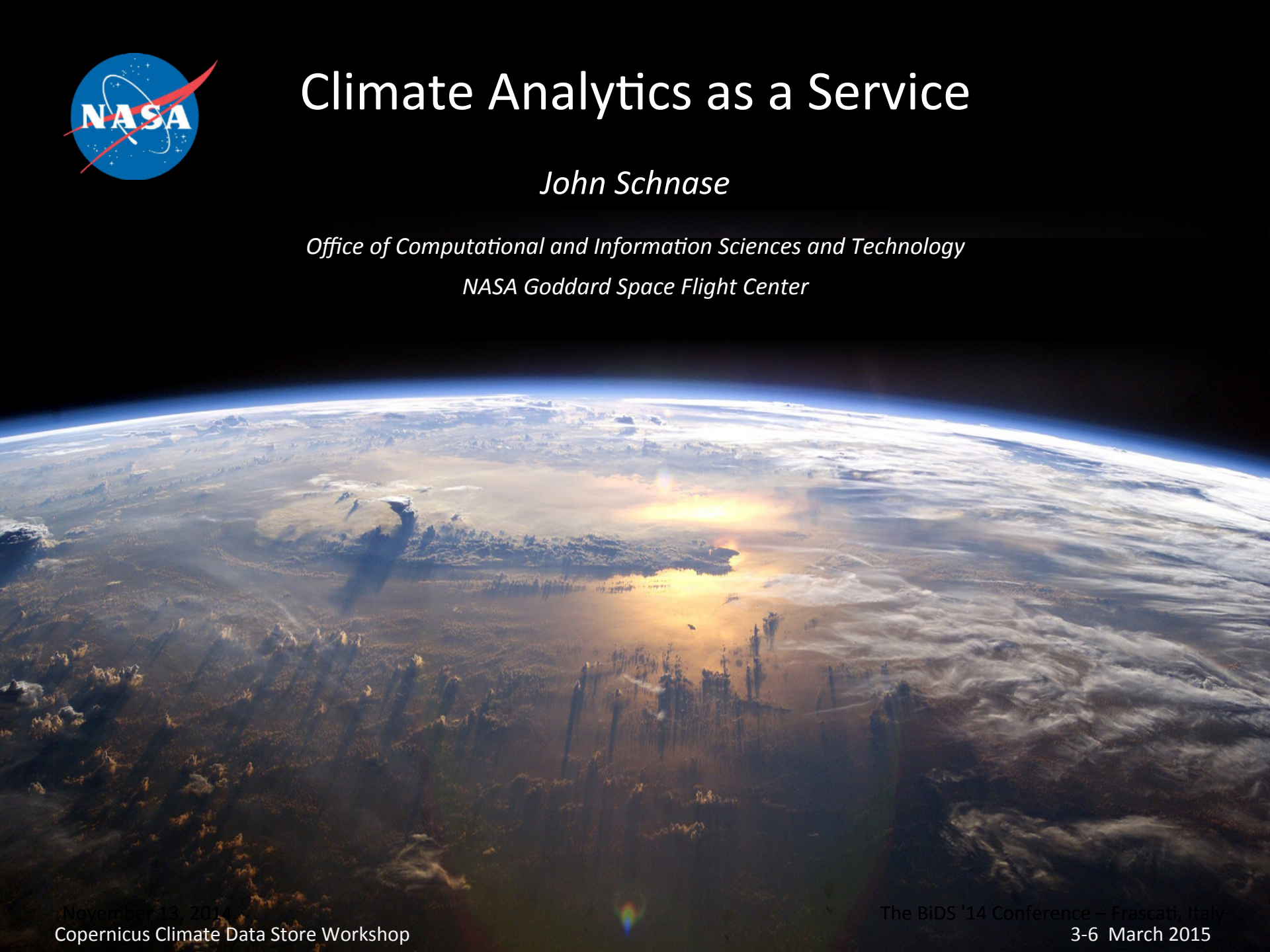
Project	Use Case/Application
NASA ABoVE Campaign	Using MERRA/AS to create historic climatology for the Arctic and Boreal region
NSF iPlant Collaborative	Leading collaborative effort between iPlant and MERRA/AS. Developer building an iPlant application within their Discover Environment that interfaces with our API
Iowa State University & ClimateMonkys	Using MERRA/AS to generate various averages pertaining to Brazil and Argentina
NSF DataNet Federation Consortium	NSF collaborative that's building the data grid infrastructure for data driven science (http://datafed.org) federation of MODIS data sets
George Mason University PhD Program	Uncertainty Quantification in Ensemble Atmospheric Reanalyses (C. Grieg)
Illinois Tech	Testing the interaction between Swift python script and CDS services
NASA / DOI RECOVER Project	Wildfire decision support system for BLM and USFS Burned Area Emergency Response (BAER) team post-wildfire remediation planning.
Invasive Species Forecasting System	Using MERRA/AS to generate habitat suitability maps based on 30+ key continental scale MERRA climatology variables

Climate Analytics as a Service

John Schnase

Office of Computational and Information Sciences and Technology

NASA Goddard Space Flight Center



November 13, 2014
Copernicus Climate Data Store Workshop

The BiDS '14 Conference – Frascati, Italy
3-6 March 2015

Backup Slides

Data Centric High Performance Computing

Data Sharing and Publication

- Capability to share data & results
- Supports community-based development
- Data distribution and publishing

Code Development

- Code repository for collaboration
- Environment for code development and test
- Code porting and optimization support
- Web based tools

User Services

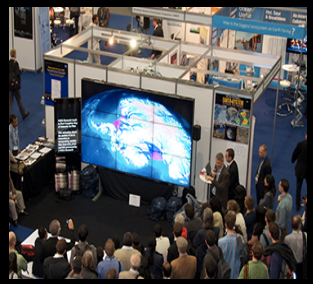
- Help Desk
- Account/Allocation support
- Computational science support
- User teleconferences
- Training & tutorials

DATA Storage & Management

Global file system enables data access for full range of modeling and analysis activities

High Performance Data Analytics

- Interactive analysis environment
- Software tools for image display
- Easy access to
- Specialized visualization support



Data Transfer

- Internal high speed interconnects for HPC components
- High-bandwidth to data center users
- Multi-gigabit network supports on-demand data transfers

Security

HPC Computing

- Large scale HPC computing
- Comprehensive toolsets for job scheduling and system monitoring

Data Archival and Stewardship

- Large capacity storage
- Tools to manage and protect data
- Data migration support

Discover Scalable Compute Unit 9

Summer 2013 Addition to the Discover Cluster (FY13)

- IBM iDataPlex
- 480 compute nodes
- Intel Xeon SandyBridge Processors
- 64 GB of RAM
- FDR Infiniband

Computational Capability

- Peak: 159,744 Gflops

Capable of additional Intel Phi coprocessors or NVIDIA GPUs

Discover Scalable Compute Unit 10

October 2014 Compute Upgrade

- SGI System
- Intel Xeon Haswell Processors
- 1,080 nodes; 30,240 cores
- 128 GB RAM per Node
- FDR Infiniband
- 1 PF peak Rmax Linpack

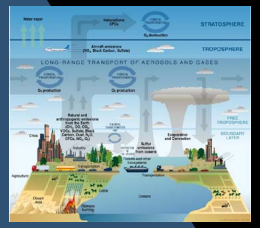
Storage Upgrades

- In procurement
- Target of 10 PB or more
- To be installed late 2014

Evolution to a Data Centric Environment

Data

HPC Models
- GEOS 5
- ModelE
- WRF



Observations
- Ground Based
- Satellite
- In Situ

Reanalysis
- MERRA
- NOAA
- Others

HPC Computing and Storage
- NASA NCCS
- NOAA
- Others

Analytics

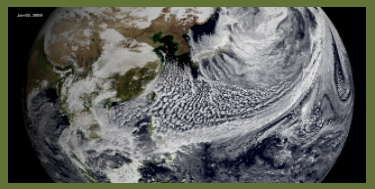
Data Services

Moving beyond just a file system and a storage repository.

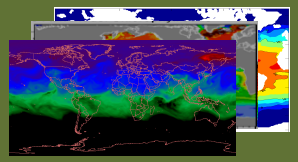
NCCS and Data Services Projects
- Dali Analysis Nodes
- vCDS
- Hadoop (HDFS)
- Merra Analytic Service
- Earth System Grid
- Web Portals

Discovery

Modelers/ Scientists



Downstream Users
- Agriculture
- Water Management
- Health
- Famine Prediction



Commercial
- Insurance/Reinsurance
- Commodity Trading

Public/Citizen Scientists

Data Management System
iRODS based management of federated data sets