

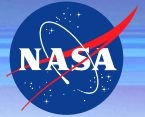
High Performance Science Cloud – Meeting the Big Data Challenges of Climate Science

Presentation at the 16th Workshop on High Performance Computing in Meteorology

Daniel Duffy¹, John Schnase², Phil Webster², and Mark McInerney³
NASA Center for Climate Simulation (NCCS)
Godard Space Flight Center (GSFC)

¹High Performance Computing Lead, NCCS, GSFC, daniel.q.duffy@nasa.gov
²Computational and Information Sciences and Technology Office (CISTO), GSFC
³Climate Model Data Services, GSFC

NASA High-End Computing Program



HEC Program Office

NASA Headquarters

Dr. Tsengdar Lee

Scientific Computing Portfolio Manager

NAS

NCCS

High-End Computing Capability (HECC) Project

NASA Advanced Supercomputing (NAS)

NASA Ames

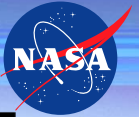
Dr. Piyush Mehrotra

NASA Center for Climate Simulation (NCCS)

Goddard Space Flight Center (GSFC)

Dr. Daniel Duffy

NASA Center for Climate Simulation (NCCS)



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

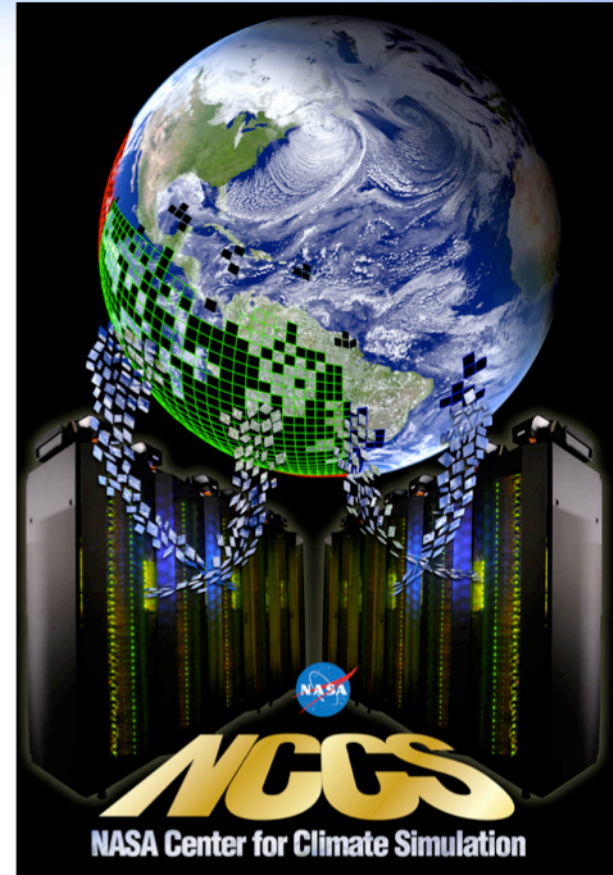
- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- **Advanced analysis and visualization environment – High Performance Science Cloud**

Primary Customers (NASA Climate Science)

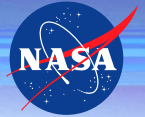
- Global Modeling and Assimilation Office (GMAO)
- Goddard Institute for Space Studies (GISS)

High-Performance Science

- <http://www.nccs.nasa.gov>



HPC Applications



Takes in small input and creates large output

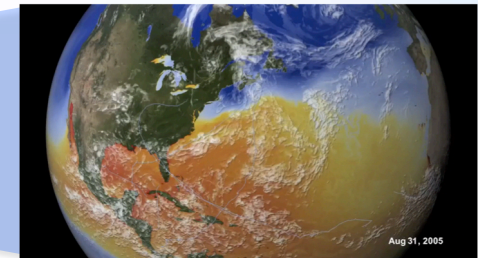
- Using relatively small amount of observation data, models are run to generate forecasts
- Fortran, Message Passing Interface (MPI), large shared parallel file systems
- Rigid environment – users adhere to the HPC systems

Example: GEOS-5 Nature Run (GMAO)

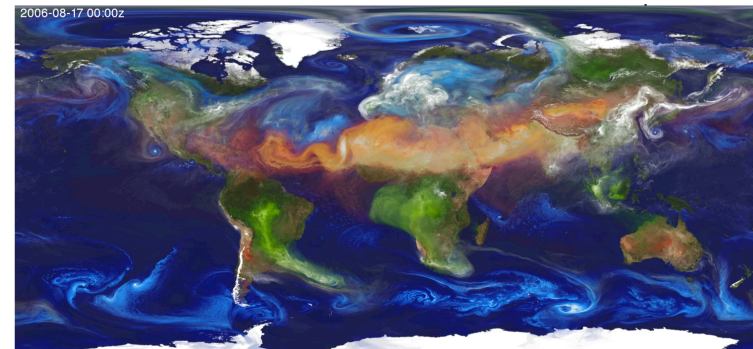
- 2-year Nature Run at 7.5 KM resolution
- 3-month Nature Run at 3.5 KM resolution
- Will generate about 4 PB of data (compressed)
- To be used for Observing System Simulation Experiments (OSSE's)
- All data to be publically accessible
 - <ftp://G5NR@dataportal.nccs.nasa.gov/>

Obs
Data

Model
(Many 100K
lines of code)

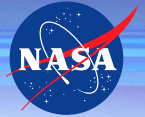


10-km GEOS-5 meso-scale simulation for Observing System Simulation Experiments(OSSEs)

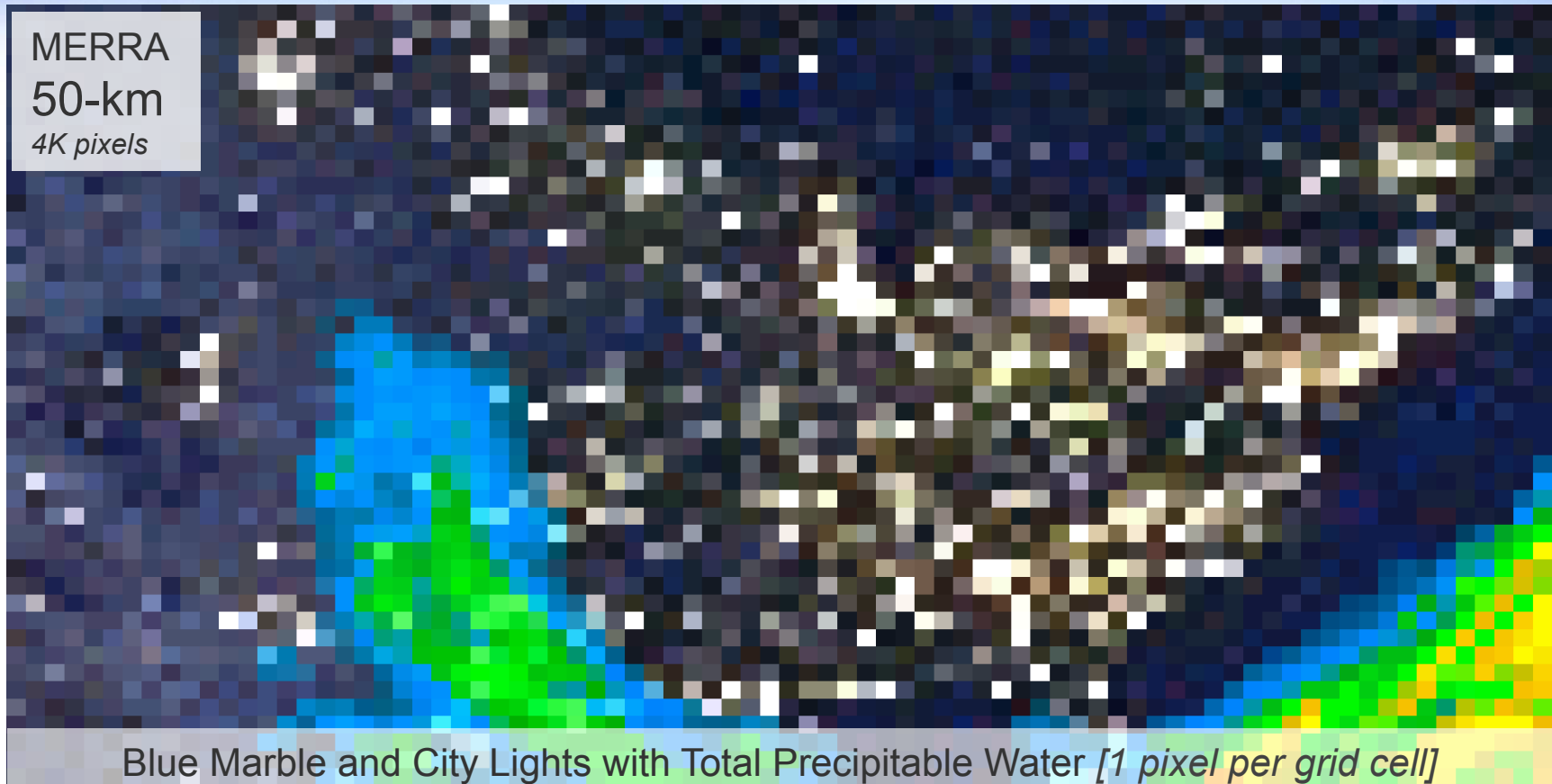


The Goddard Chemistry Aerosol Radiation and Transport (GOCART) model, Courtesy of Dr. Bill Putman, Global Modeling and Assimilation Office (GMAO), NASA Goddard Space Flight Center.

Reanalysis Resolution (50 KM)

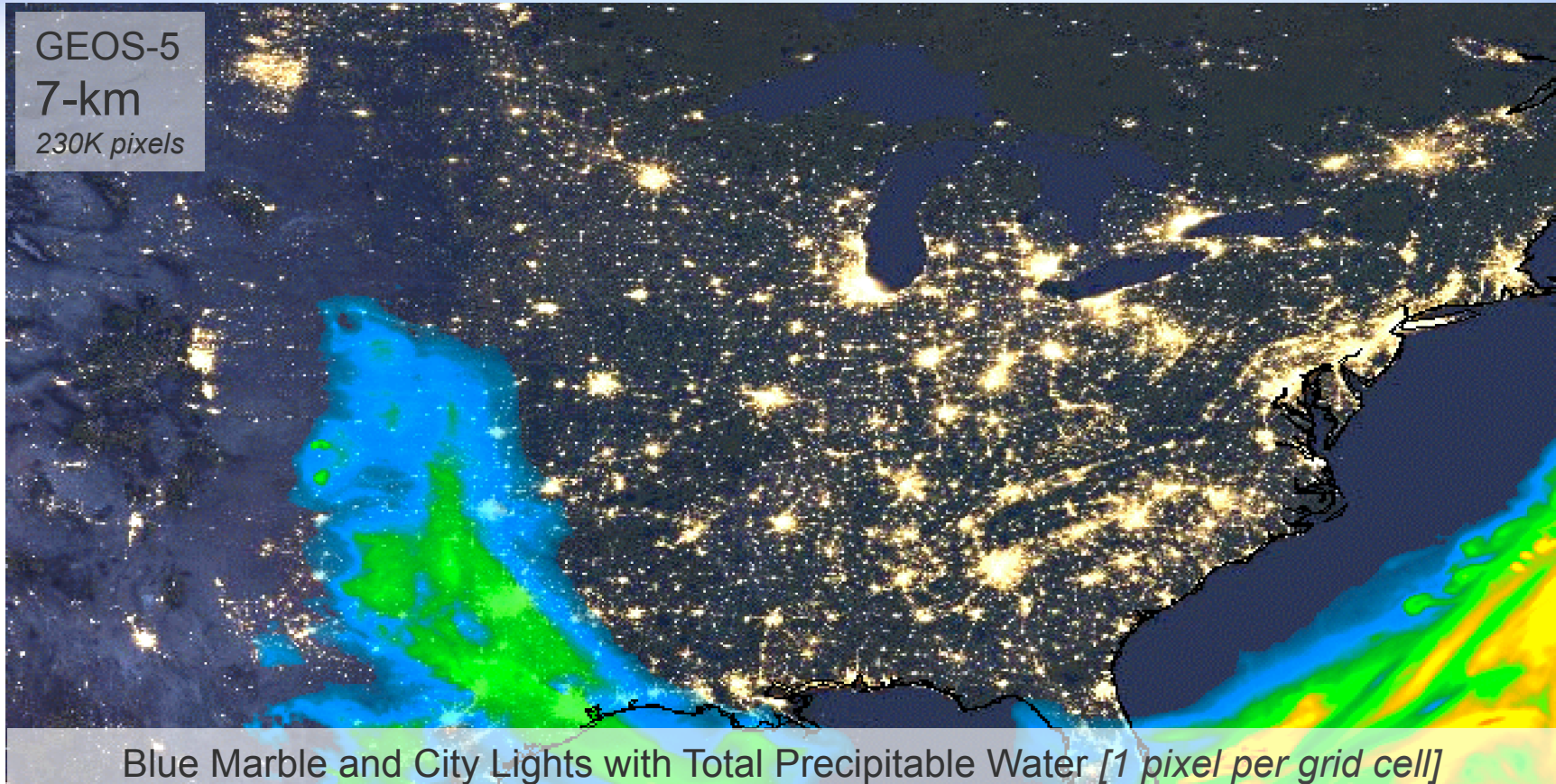
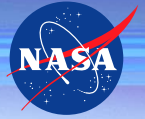


MERRA
50-km
4K pixels

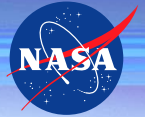


Blue Marble and City Lights with Total Precipitable Water [1 pixel per grid cell]

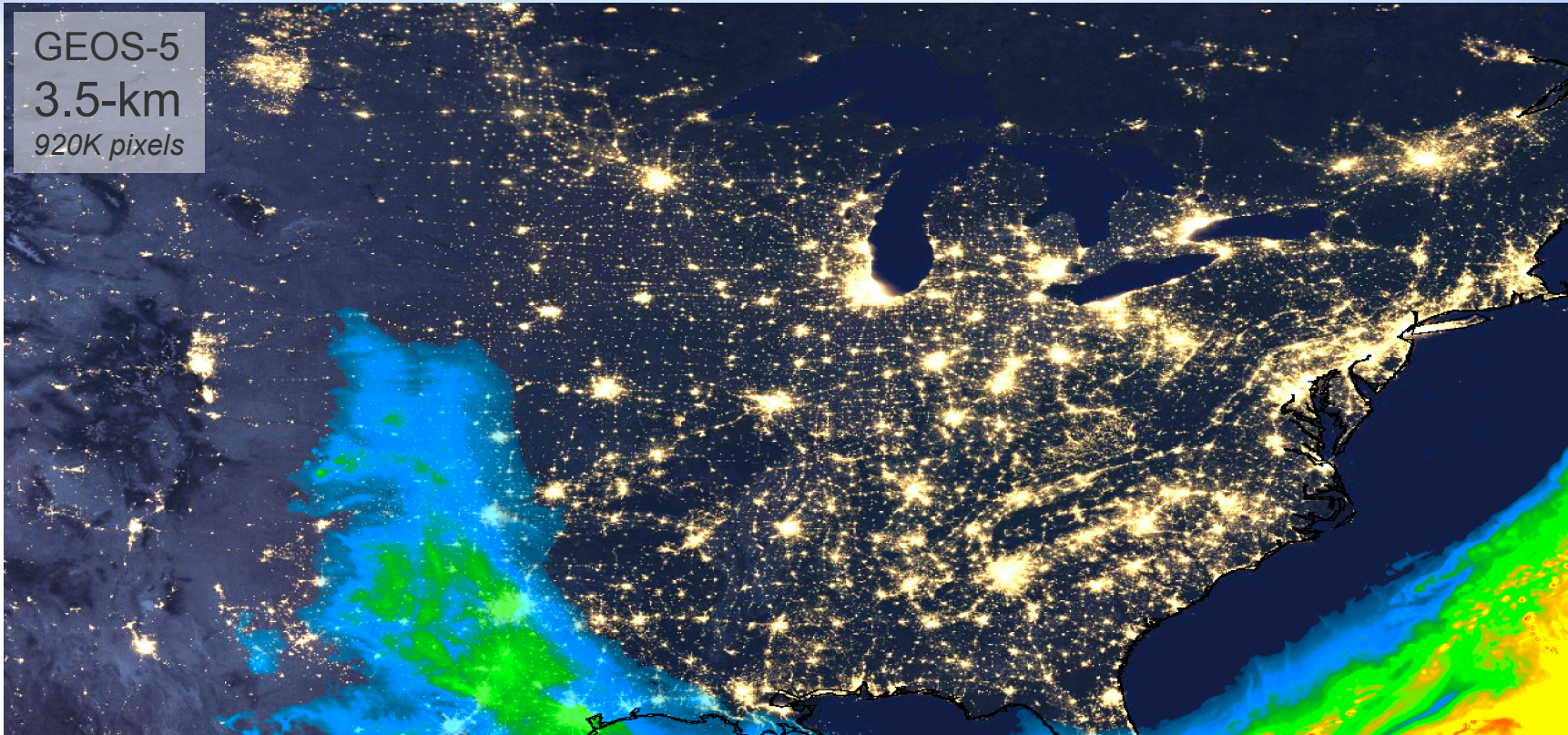
GEOS-5 Model Resolution (7 KM)



GEOS-5 Model Resolution (3.5 KM)

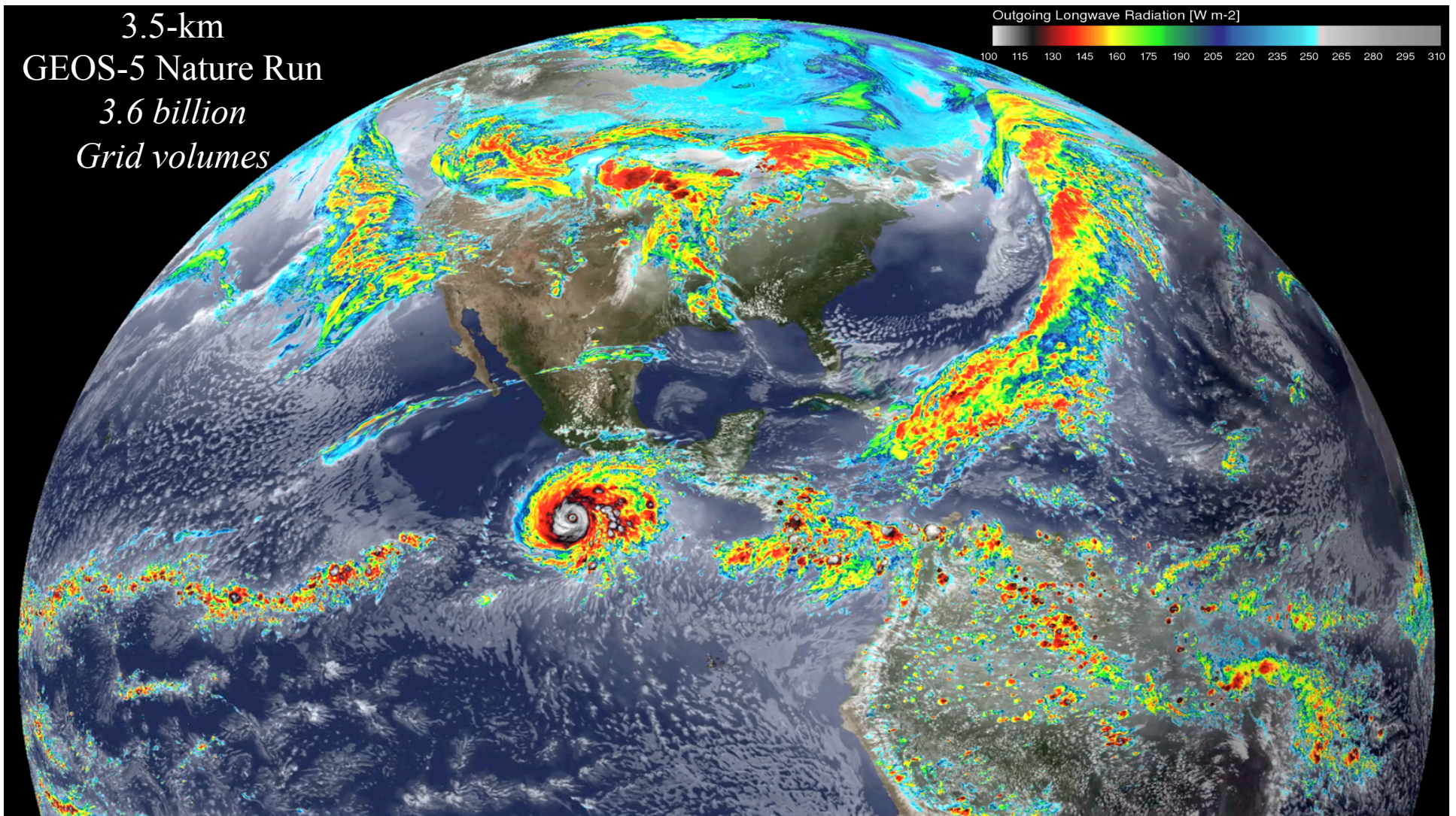


GEOS-5
3.5-km
920K pixels

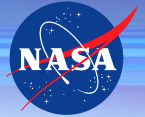


Blue Marble and City Lights with Total Precipitable Water [1 pixel per grid cell]

3.5-km
GEOS-5 Nature Run
3.6 billion
Grid volumes



Typical Analysis Applications



Takes in large amounts of input and creates a small amount of output

- Using large amounts of distributed observation and model data to generate science
- Python, IDL, Matlab
- Agile environment – users like to run in their own environments

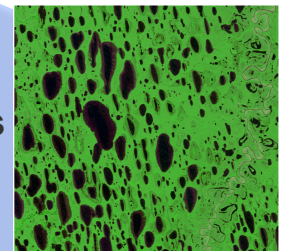
Examples

- Evaporative transport
 - Requires monthly reanalysis data sets for four different spatial extents
- Decadal water predictions for the high northern latitudes for the past three decades
 - Requires 100,000+ Landsat images and about 20 TB of storage



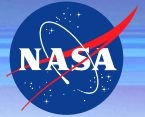
Yukon Delta Alaska; courtesy of Landsat
<http://landsat.visibleearth.nasa.gov/view.php?id=72762>

Analysis
(100's of lines
of code)



Representative Landsat image, false color composite, from near Barrow, AK; Courtesy of Mark Carroll (618).

Sign of Analysis Jobs on HPC Resources

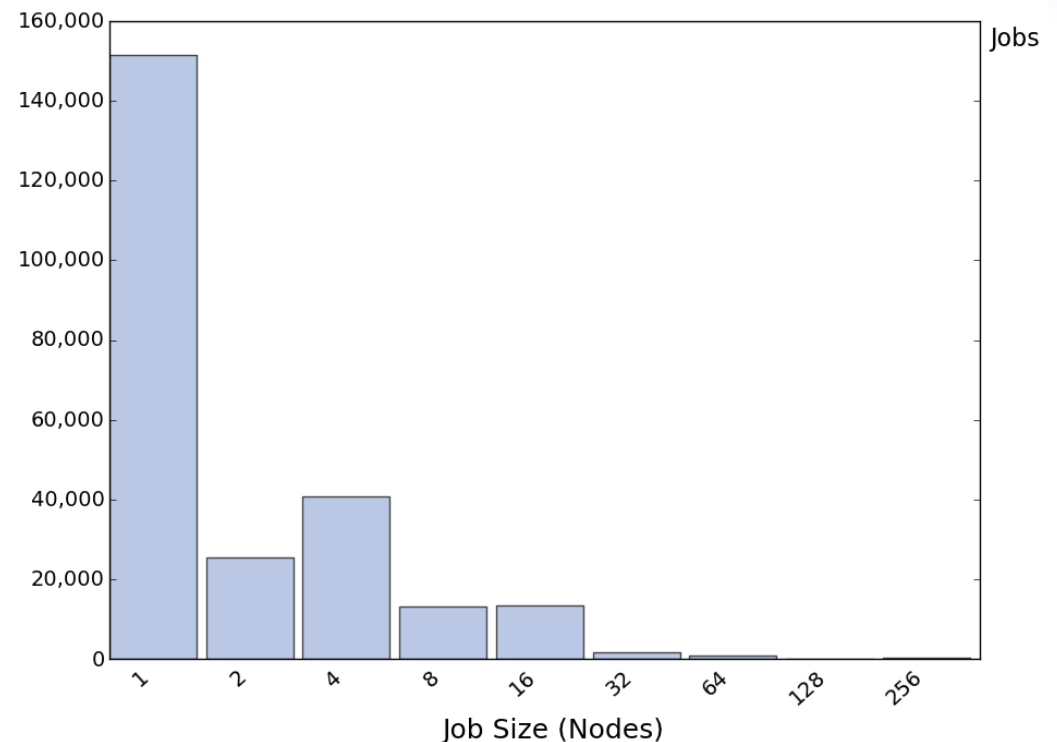


Large number of single node jobs submitted each month

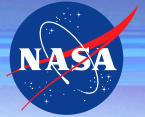
- Large amount of data analysis jobs
- While these jobs do not take up much overall cycles, they do cause issues with the batch scheduler
- These jobs also can slow down the overall system through their access patterns to data

Can we move these jobs off of the HPC system and into an environment designed for large-scale analytics?

Discover Jobs by Nodes September 2014



Analytics on Desktops



Scientists who are not using the HPC resources for analytics are trying to use their local resources, including their desktop systems.

Learning Curve of HPC

- Different operating environment (O/S and Tools)
- Rigid environment/slowly changes
- Batch queuing system
- Data transfers into and out of HPC
- Sharing data out of HPC

Scientist Desktops

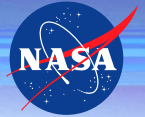
- Familiar environment
- Not enough capability
- Need more compute and storage
- Processing takes too long
- Slow data transfers
- Have to stage data into the system over time



Scientists are limiting their questions (and science) based on the IT resources of their desktops!

Can we move these jobs into an environment designed for large-scale analytics?

Data Analytics Technology Gap



Archive



Archive
~5 PB of Disk
~45 PB of Tape

Optimized for long term storage, typically slower storage designed for streaming reads and writes

Leads to Un-optimized Practices:

Users perform data analysis straight from the archive and complain that it is too slow.

Very Large
Performance Gap

Specifically for Data
Analysis, Analytics,
and Visualization of
large scale data

What technologies can
we use to help bridge
this gap?

Large Scale Compute



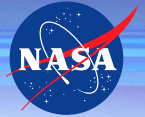
Discover Cluster
~2 PF Peak
~25 PB of Disk

Optimized for large scale simulations with fast storage designed for streaming applications

Leads to Un-optimized Practices:

Users analyze large data sets through a series of many small blocks reads and writes and complain that it is too slow.

Where do we look for help?



Google

- By 2012, Gmail had 425 million active users¹
- Each user gets 15 GB of storage for free
- $425,000,000 * 15 \text{ GB} = 6,375,000,000 \text{ GB} = 6,385,000 \text{ TB} = 6,375 \text{ PB} = 6.375 \text{ EB}$
- Assuming about 6% of the email is spam², Gmail carried around 382.5 PB of spam!



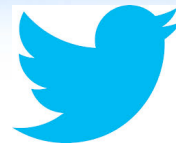
Facebook

- By 2012, Facebook was processing 500 TB of data per day³
- 2.7 billion Like actions and 300 million photos per day; Facebook scanned about 105 TB every 30 minutes⁴



1. <http://venturebeat.com/2012/06/28/gmail-hotmail-yahoo-email-users/>
2. <http://krebsonsecurity.com/2013/01/spam-volumes-past-present-global-local/>
3. http://news.cnet.com/8301-1023_3-57498531-93/facebook-processes-more-than-500-tb-of-data-daily/
4. <http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/>

Compare HPC to Large Scale Internet



High Performance Computing

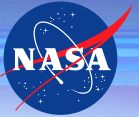
- Shared everything environment
- Very fast networks; tightly coupled systems
- Cannot lose data (POSIX)
- Big data (100 PBs)
- Bring the data to the application
- Large scale applications (up to 100K cores)
- Applications cannot survive HW/SW failures
- Commodity and non-commodity components; high availability is costly; premium cost for storage

Object Storage
MapReduce
Hadoop
Cloud
Open Stack
Virtualization

Large Scale Internet

- Examples: Google, Yahoo, Amazon, Facebook, Twitter
- Shared nothing environment
- Slow networks
- Data is itinerant and constantly changing (RESTful)
- Huge data (Exabytes)
- Bring the application to the data
- Very large scale applications (beyond 100Ks)
- Applications assume HW/SW failures
- Commodity components; low cost storage

Why not just let scientists use public clouds?



They can!

There are a number of reasons why a NASA managed cloud makes sense...

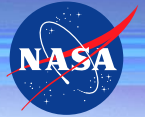
- Data in Public Clouds
 - Transferring data into the cloud is slow and expensive
 - Storing data in the cloud is expensive
 - Data access in the cloud is not high performance
 - Many copies of data could end up in the cloud multiplying expenses
- Scientists Administering Systems
 - Asking scientists to be their own system administrators
 - Scientists must be knowledgeable in operating systems, tools, license management, file systems, security, etc.
- Low Support
 - There is little to no support in public clouds for science processing
 - Do we really want scientists Google'ing how to install an NFS server in the cloud?

This takes away time and funding from scientists.

With a NASA managed cloud, more science can be done!



Genesis of the Science Cloud



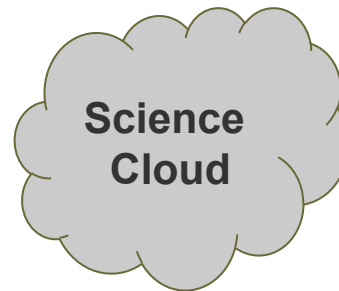
2010 Virtualized Infiniband – Implementation SR-IOV technology to enable extremely high speed Infiniband interconnects (HPC speeds) to virtual systems

2013 Virtual Machines and Containers – Comparison of virtual machines versus containers for application performance; use of virtual machines to support the HPC environment

2011 Object Storage Environments – Analysis of alternatives of various data environments for cloud computing, including Amazon S3, EBS, RedHat Gluster, IBM GPFS, Hadoop, LLNL ZFS, etc.

2013 Remote Visualization Solutions – Proof-of-concept and representative architecture to enable remote visualization of large-scale data sets

2011 Exploration of Cloud Computing for HPC – Benchmarking large scale applications in Amazon and NASA Cloud (Nebula) to compare to HPC performance

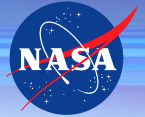


2014 Cloud Software Stacks – Proof-of-concept implementations and analysis of alternatives of RedHat Enterprise Virtualization (RHEV), RedHat OpenStack, OpenStack, Linux and KVM

2012 Cloud Storage Servers – In house storage servers built from commodity components designed for the science cloud.

2014 Science Proof of Concept – Test case for the Arctic Boreal Vulnerability Experiment campaign.

NCCS High Performance Science Cloud



Adjunct to the NCCS HPC environment

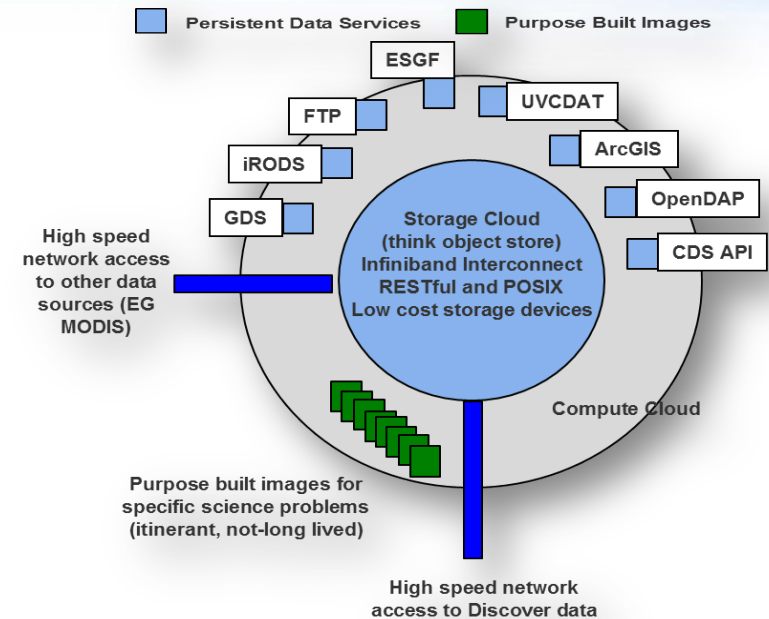
- Lower barrier to entry for scientists
- High level of support – help desk, system administration, security, applications, etc.
- Agile and customized run-time environments purpose built for specific science projects
- Low cost compute – reusable HPC/Discover hardware

Expanded customer base

- Scientist brings their analysis to the data
- Extensible storage; build and expand as needed
- Persistent data services build in virtual machines
- Create purpose built VMs for specific science projects

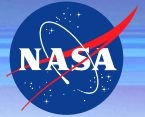
Difference between a commodity cloud






- Close to matching HPC levels of performance
- Critical Node-to-node communication – high speed, low latency
- Shared, high performance file system
- The system owns the data
- Management and rapid provisioning of resources



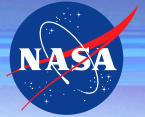
High Performance Science Conceptual Architecture *Platform-as-a-Service*

System Components/Configuration



Capability and Description	Configuration
 Persistent Data Services Virtual machines or containers deployed for web services, examples include ESGF, GDS, THREDDS, FTP, etc.	8 nodes with 128 GB of RAM, 10 GbE, and FDR IB
 DataBase High available database nodes with solid state disk.	2 nodes with 128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB
 Remote Visualization Enable server side graphical processing and rendering of data.	4 nodes with 128 GB of RAM, 10 GbE, FDR IB, and GPUs
 High Performance Compute More than 1,000 cores coupled via high speed Infiniband networks for elastic or itinerant computing requirements.	~100 nodes with 32 to 64 GB of RAM, and FDR IB
 High-Speed/High-Capacity Storage Petabytes of storage accessible to all the above capabilities over the high speed Infiniband network.	10 storage nodes configured with a total of about 3 PB of RAW storage capacity

Additional Details About System Capabilities



High Performance Compute

- Using decommissioned HPC servers; still very capable for analytics
 - » 30 nodes with 12 cores and 24 GB of RAM
 - » 80 nodes with 20 cores and 64 GB of RAM
 - » Coming Soon
 - 30 nodes with 12 cores and 48 GB of RAM and NVIDIA M2070's
 - ~200 nodes with 12 cores and 24 GB of RAM

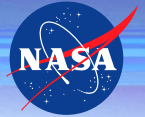
Storage

- Large storage space is available for shared data sets
- Small storage space is available for scratch

Network

- External networks at 1 GbE now; moving to 10 GbE next year
- Internal networks have 1 GbE, 10 GbE, and even FDR Infiniband (56 Gbps)

Current Status and Schedule

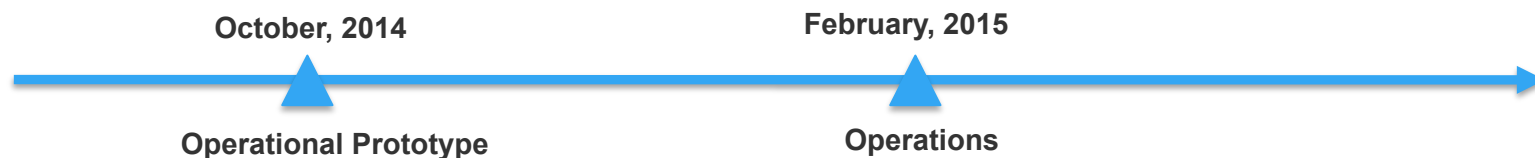


The current science cloud is an *operational prototype*

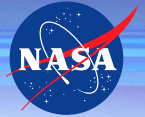
- Not production yet
- Currently supporting a number of science projects

Full operational production is scheduled for February of 2015

- What does that mean?
- Documentation will be created (some, not all; there is never enough)
- User account process will be completed (and hopefully simplified)
- User interfaces to be able to launch and manager virtual machines



What works best in the cloud?



Not designed for MPI

- Highly coupled processes performing large amounts of message passing
- Though it could be used with MPI

Designed more for inherently parallel processing of big data

- Independent processes written to analyze large data sets

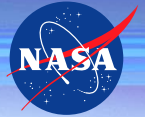
Publishing of data

- Persistent data services created to provide a capability for NASA scientists to share large data

Analytics as a Service

- Creation of storage proximal processing of large data sets through standard application programming interfaces

Science Cloud In Action



Arctic Boreal Vulnerability Experiment (ABOVE)

- Comparing multiple, time displaced images of the same geographical area looking for changes in surface water extent using LANDSAT data.
- Long-Term Multi-Sensor Record of Fire Disturbances in High Northern Latitudes using LANDSAT and MODIS

Multi-Angle Implementation of Atmospheric Correction (MAIAC)

- Improve accuracy of cloud detection, aerosol retrievals and atmospheric correction using new advanced algorithm based on time series analysis and a combination of pixel- and image-based processing of MODIS

National Geospatial Agency (NGA) High Resolution Image Processing

- Converting native NGA NTF to GTIFF (GeoTIFF)
- Calculate vegetation index by counting trees, shrubs, etc.

Asteroid Hunter

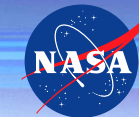
- Evaluating design of next generation space-based telescope using a combination of asteroid projection modeling and simulated telescope

Level 2 Aerosol data

- Multiple sources (~10) extracting relevant aerosol data from the observations for about 800 locations around the globe.

Climate Model Data Post Processing

ABoVE Science Cloud (ASC)



The screenshot shows the ABoVE website interface. At the top, there is a NASA logo and the text 'National Aeronautics and Space Administration'. Below this is a banner for the 'Arctic-Boreal Vulnerability Experiment' with a scenic background of mountains and trees. A navigation menu on the left includes links for Home, About, Timeline, Concise Experiment Plan, SDT Activities, Study Domain, Pre-ABOVE Projects, Funding Resources, Field Operations, Documents, Contacts, Acronyms, and Calendar. The main content area features a paragraph about climate change in the Arctic and Boreal regions, followed by a box titled 'Concise Experiment Plan released' dated 'June 23, 2014'. Below this are two columns of announcements and a 'Where Are We Now?' section.

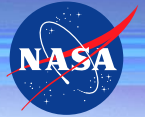
Partnership between the CCE, CISTO, and NCCS

- Subset of the science cloud prioritized for the ABoVE Campaign
- Provide compute, storage, data management, and data publication
- Reduces technical overhead for ABoVE scientists
- Allows scientists to focus on science in a optimized computing environment

The Conceptual Architecture:

- Data analysis platform collocating data, compute, data management, and data services
- Ease of use for scientists; customized run-time environments; agile environment
- Data storage surrounded by a compute cloud
- Large amount of data storage
- High performance compute capabilities
- Very high speed interconnects

Mixture of Observations and Model Data



ABoVE Observations

- CF Conventions, Ameriflux, CALM, SensorML, Instrument Vendors, SmartPhone Apps

Other Observations (as needed)

- Landsat, MODIS
- Derived geospatial products

Assimilations

- SMAP L4

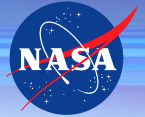
Reanalysis Data

- MERRA from GMAO
- Other reanalysis data sets from ECMWF, NCEP, and more

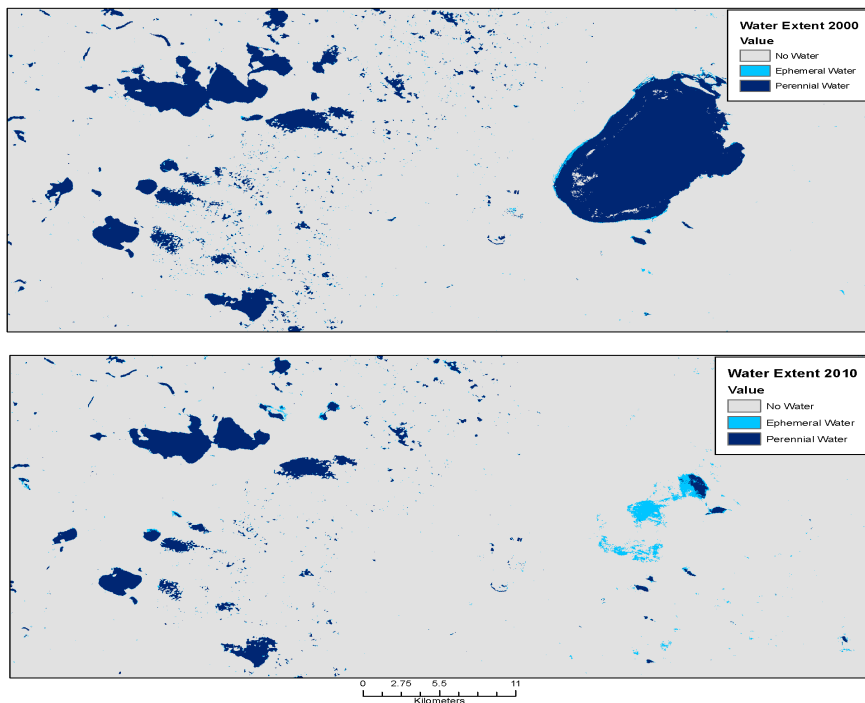
Data Set	Description	Total Data (compressed)	Total Data (uncompressed)
Landsat	Surface reflectance	19 TB	123 TB
MODIS	Daily 500 meter surface reflectance	57 TB	57 TB
MERRA	GEOS-5 reanalysis	89 TB	192 TB
Total		195 TB	372 TB

Data holdings relevant to the ABoVE mission currently in the science cloud. Downloading this data to users' workstations takes weeks to months!

Pre-ABoVE Example

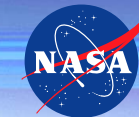


**Change in Surface Water Extent at
Beaver Hill Lake Between 2000 - 2010**



- **Tracking the change in surface water over the study region using Landsat**
 - Average across three epochs (1990, 2000, 2010)
 - 25,000 Landsat scenes/~7 TB of data
 - Projected time 9 months
- **Using the science cloud**
 - 48 virtual machines
 - 6 weeks of processing
- **Opened up the opportunity to do more processing**
 - Explore the complete Landsat record
 - 100,000 scenes > 20 TB of data

Next Steps for the Science Cloud



NCCS Science Cloud

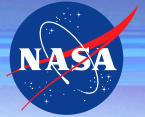
- Production by March 2015
- More compute nodes
 - Reuse of NCCS hardware that is being upgraded
- Bring more service offerings on-line
 - Remote Visualization, Databases, Self Service Portals, GPUs
 - Climate-Analytics-as-a-Service

Science Support

- NCCS Data Portal
- ABoVE Mission Support
Data as a Service (e.g., Hadoop)
- Nature Run Data Processing
- Looking for Other Science Opportunities



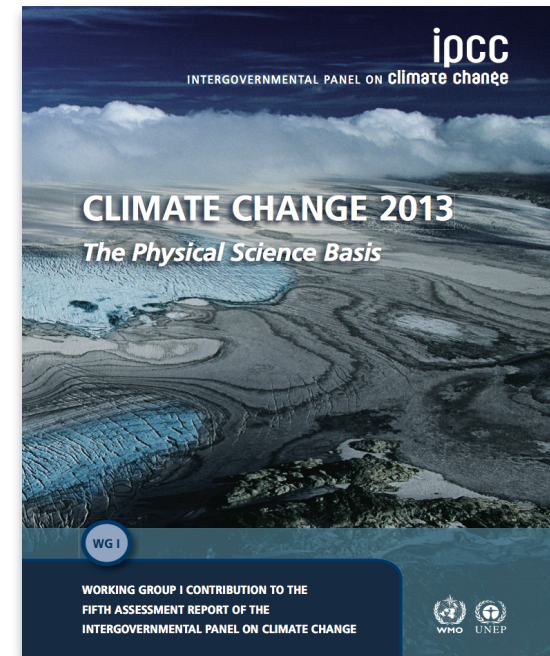
Climate-Analytics-as-a-Service (CAaaS)



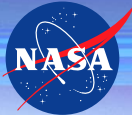
New methods for climate analytics are needed – access to large distributed data sets through APIs for storage-proximal processing. Work currently being done at GSFC and in collaboration with ESGF.

How much climate data?

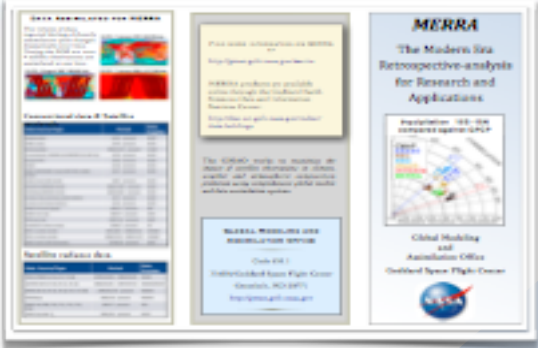
- MERRA Reanalysis Collection ~200 TB
- Total data holdings of the NASA Center for Climate Simulation (NCCS) is ~40 PB
- Intergovernmental Panel on Climate Change Fifth Assessment Report ~5 PB (data on line now)
- Intergovernmental Panel on Climate Change Sixth Assessment Report ~100 PB (to be created within the next 5 to 6 years)



Climate Analytics as a Service



MERRA Reanalysis



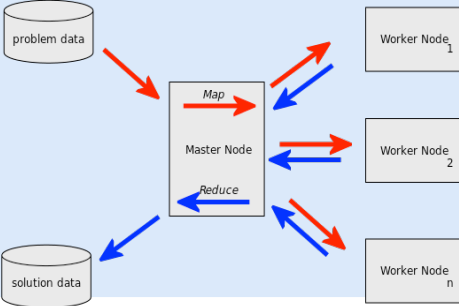
Data

High-Performance Compute/Storage Fabric

Storage-proximal analytics with simple canonical operations

Data do not move, analyses need horsepower, and leverage requires something akin to an analytical assembly language ...

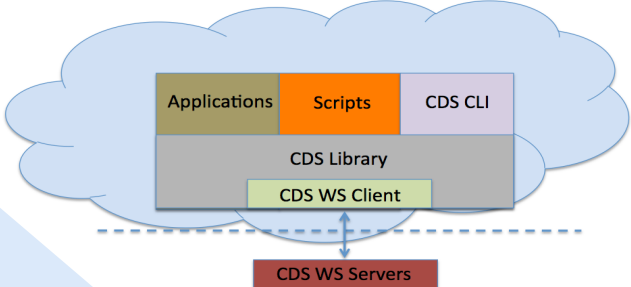
MERRA Analytic Services



Relevance and Collocation

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

Climate Data Services API



Exposure

Convenient and Extensible

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



Simple ABoVE Related Example

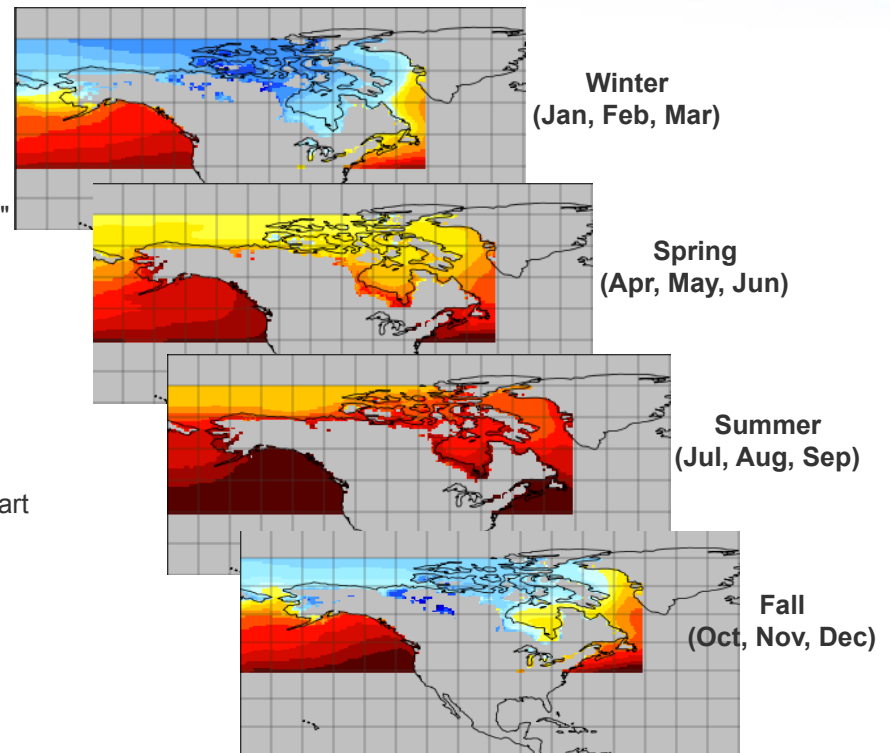
```
#!/usr/bin/env python
import sys
from CDSSLibrary import CDSApi
cds_lib = CDSApi()
service = "MAS"

name = "above_avg_seasonal_temp_1980_instM_3d_ana_Np"
job = "&job_name=" + name
collection = "&collection=instM_3d_ana_Np"
request = "&request=GetVariableBy_TimeRange_SpatialExtent_VerticalExtent"
variable = "&variable_list=T"
operation = "&operation=avg"
start = "&start_date=198001"
end = "&end_date=198012"
period = "&avg_period=3"
space = "&min_lon=-180&min_lat=40&max_lon=-50&max_lat=80"
levels = "&start_level=1&end_level=42"
file_job_epoch1_aveT = "." + name + ".nc"
above_job_epoch1_aveT = job + collection + request + variable + operation + start
+ end + period + space + levels

class UserApp(object):
    if __name__ == '__main__':

        cds_lib.avg(service, above_job_epoch1_aveT, file_job_epoch1_aveT)
```

QUESTION: Extract the average temperature by season for the year 1980 for the ABoVE region at every level in the MERRA reanalysis data.



Thank You and Thanks to the All People That Make This Work!

