ABoVE Science Cloud:
An orientation for the ABoVE Science Team

Peter Griffith
Chief Support Scientist

Liz Hoy
Support Scientist

Carbon Cycle & Ecosystems Office

@NASA_ABoVE

Mark McInerney
Data Services Lead

Dan Duffy
NCCS HPC Lead

Computational and Information Sciences and Technology Office

Contributors

High Performance Computing

- Scott Sinno, System Architect and System Administrator
- Hoot Thompson, System Architect and System Administrator
- Garrison Vaughn, System Architect and Applications Engineer
- Brittney Wills, Computer Scientist
- Ellen Salmon, Computer Research and Development

Climate Model Data Services

- Laura Carriere, System Analyst
- Steven Ambrose, Principal Systems Engineer
- Julien Peters, Software Developer
- Eric Winter, Software Developer

Presentation Outline

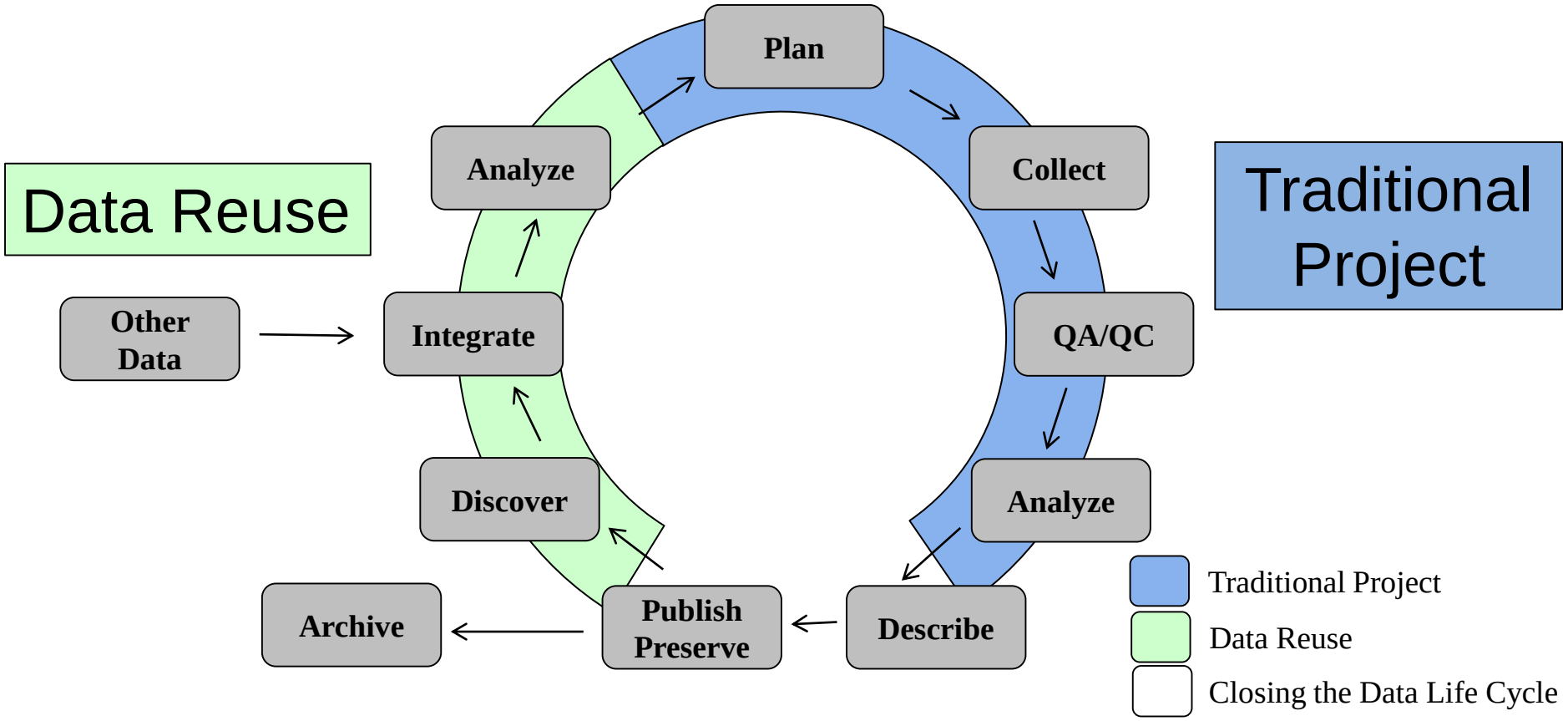
- Introduction (Peter)
- High Performance Compute Capabilities (Dan)
- Climate Data Services for ABoVE (Mark)
- Creating an Account (Liz)

The Carbon Cycle & Ecosystems Office is responsible for implementation and management of ABoVE

Science team members should plan to work closely with the CCEO and rely upon our guidance for field operations and safety, communications with local and regional stakeholders and authorities, and utilization of ABoVE cyberinfrastructure.

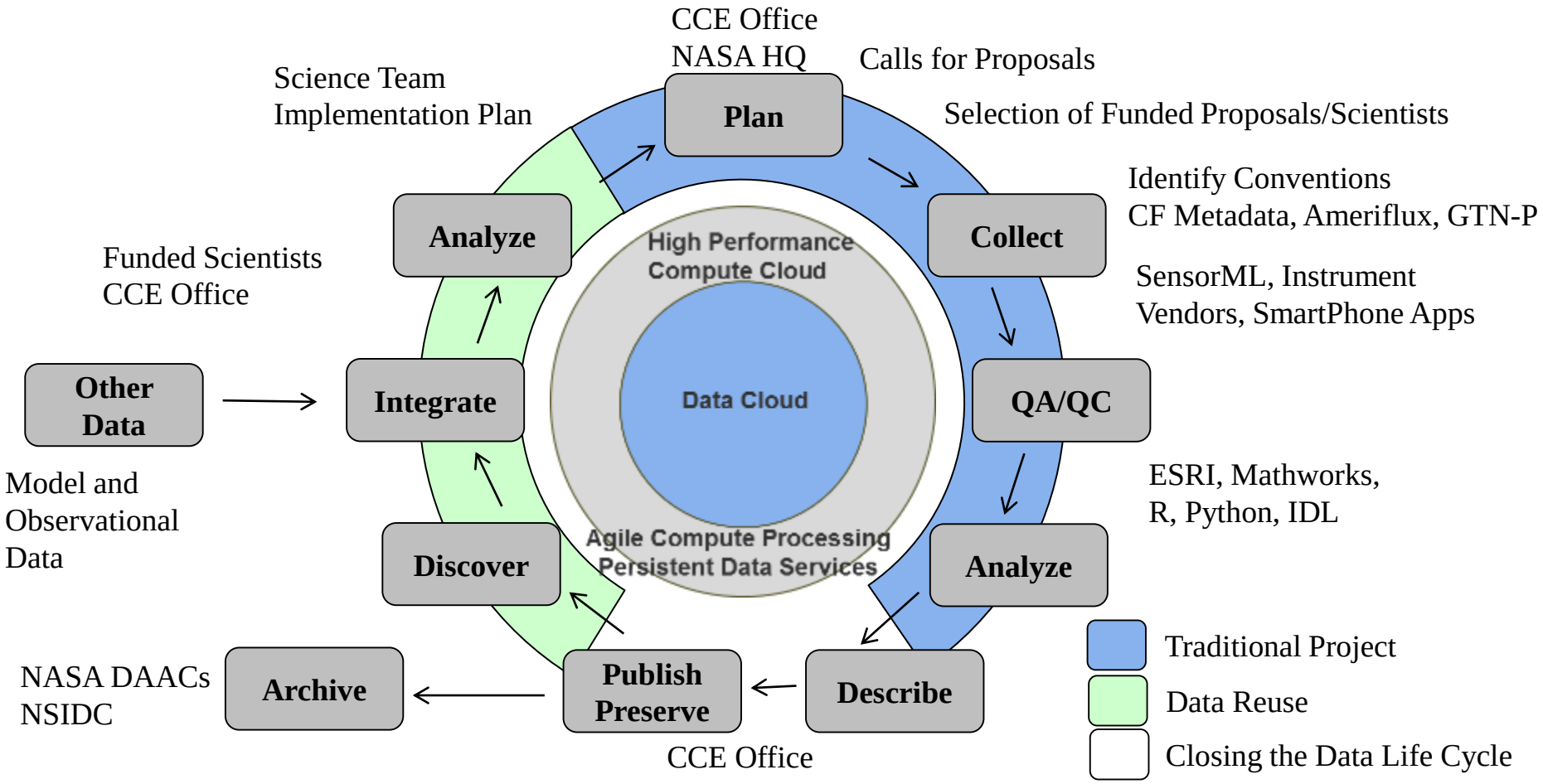
The ABoVE Science Cloud combines high performance computing with emerging technologies and data management with tools for analyzing and processing geographic information to create an environment specifically designed for large-scale modeling, analysis of remote sensing data, copious disk storage for "big data" with integrated data management, and integration of core variables from in-situ networks. The ABoVE Science Cloud is a collaboration that promises to accelerate the pace of new Arctic science for researchers participating in the field campaign. Furthermore, by using the ABoVE Science Cloud as a shared and centralized resource, researchers reduce costs for their proposed work, making proposed research more competitive.

The CCE Office will assist the Science Team throughout the Data Management Lifecycle.



Augmented from Rüegg et al 2014 in *Front Ecol Environ*

The ASC will surround aspects of the data lifecycle.



Augmented from Rüegg et al 2014 in *Front Ecol Environ*

Data Use on the ABoVE Science Cloud

- Researchers should anticipate sharing their data using ABoVE's cyberinfrastructure and/or partnering networks
- Storage in the ASC will be tailored to meet Science Team needs
- Using ORNL DAAC best practices will facilitate integration

NATIONAL WEATHER SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

AMERIFLUX

Geographic Information Network of Alaska

CALM

HADS Hydrometeorological Automated Data System

GTN-P Global Terrestrial Network for Permafrost

Imiq

Partnership between the High End Computing (HEC) Program and the Carbon Cycle and Ecosystems Office

Computational & Information Sciences and Technology Office (CISTO/606 – funded by HEC) will provide technical support and services for ABoVE

System Support / Science Cloud

Data Management / Application Support

NASA Center for Climate Simulation (NCCS/606.2)

- Daniel Duffy, Lead
- High performance computing solutions for NASA climate science

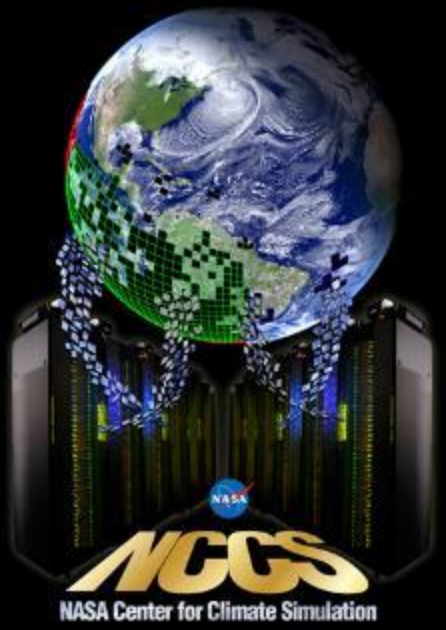
http://nccs.nasa.gov

https://twitter.com/NASA_NCCS

Climate Model Data Services (CMDS/606)

- Mark McInerney, Lead
- Tools and services to visualize, analyze, compare, and publish climate model data

http://cds.nccs.nasa.gov



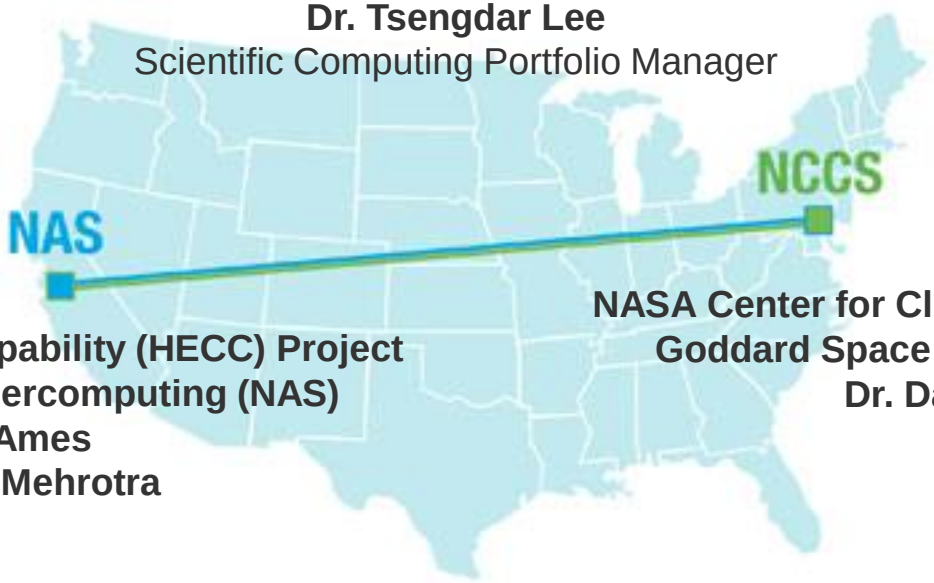
ABoVE Science Cloud
September 2, 2015

Daniel Duffy (daniel.q.duffy@nasa.gov and on Twitter @dqduffy)
High Performance Computing Lead
NASA Center for Climate Simulation (NCCS)
http://www.nccs.nasa.gov

NASA Goddard Space Flight Center
Sciences and Exploration Directorate
Computational and Information Sciences and Technology Office

NASA High-End Computing Program

HEC Program Office
NASA Headquarters
Dr. Tsengdar Lee
Scientific Computing Portfolio Manager



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.

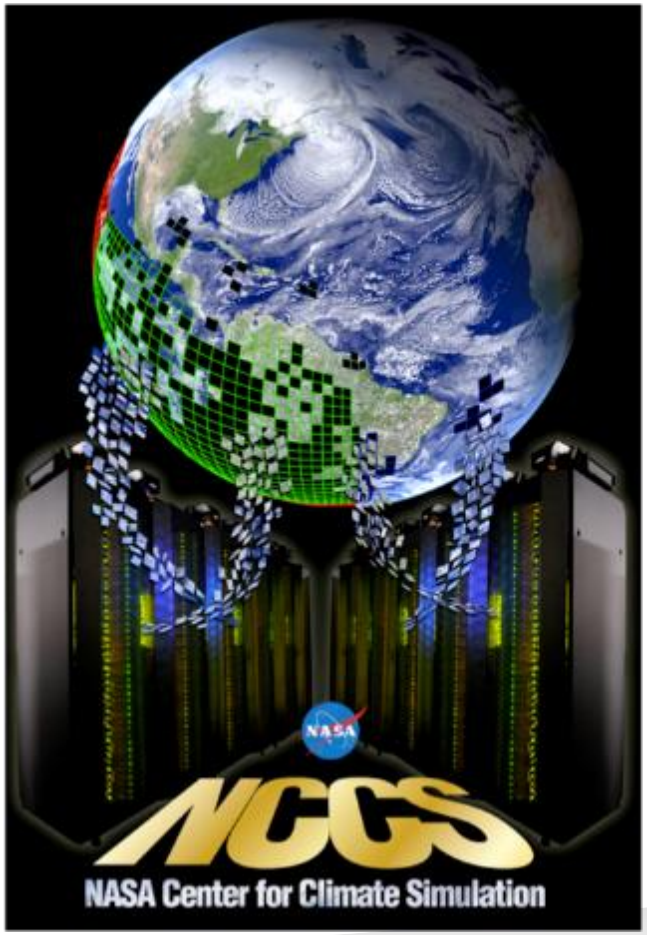
- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- Advanced Data Analytics Platform (ADAPT)

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
- Located in Building 28 at Goddard



Analysis is Different than HPC

High Performance Computing

Takes in small amounts of input and creates large amounts of output…

- Using relatively small amount of observation data, models are run to generate forecasts
- Tightly coupled processing requiring synchronization within the simulation
- Simulation applications are typically 100,000's of lines of code
- Fortran, Message Passing Interface (MPI), large shared parallel file systems
- Rigid environment – users adhere to the HPC systems

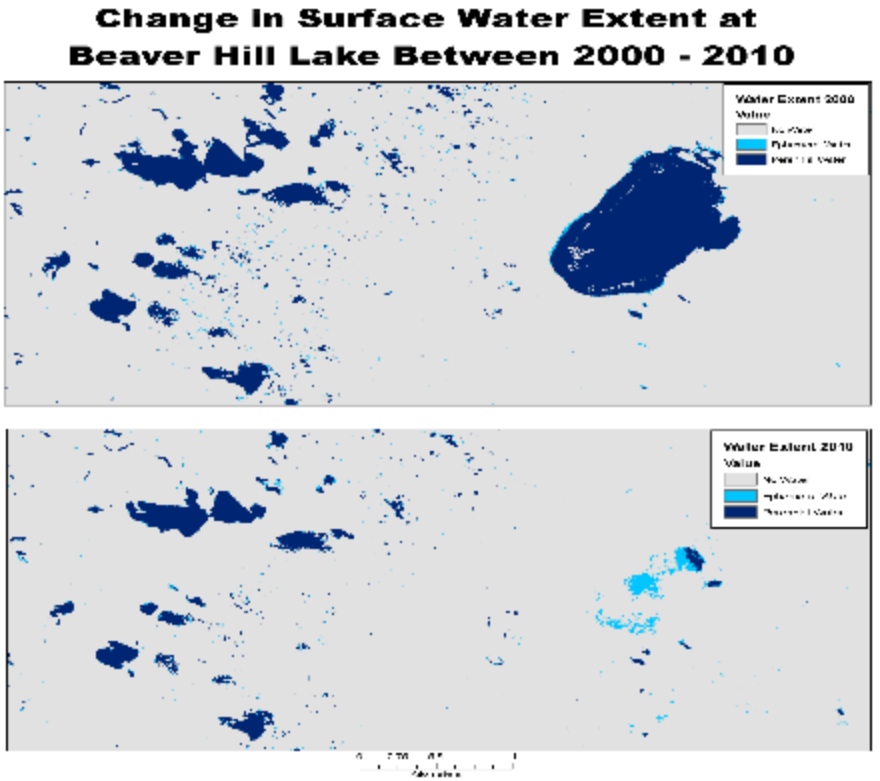
Data Analysis

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

- Use large amounts of distributed observation and model data to generate science
- Loosely coupled processes requiring little to no synchronization
- Analysis applications are typically 100's of lines of code
- Python, IDL, Matlab, custom
- Agile environment – users run in their own environments

Example Analysis Application Support

- Pre-ABoVE Project
  - Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign
- Principal Investigator
  - Mark Carroll
- Details
  - Decadal water predictions for the high northern latitudes for the past three decades
  - Requires 100,000+ Landsat images and about 20 TB of storage



M. Carroll

This project has been a pioneer user of the ABoVE Science Cloud for quite some time!

Advanced Data Analytics Platform (ADAPT) "High Performance Science Cloud"

High Performance Science Cloud is uniquely positioned to provide data processing and analytic services for NASA Science projects

Adjunct to the NCCS HPC environment

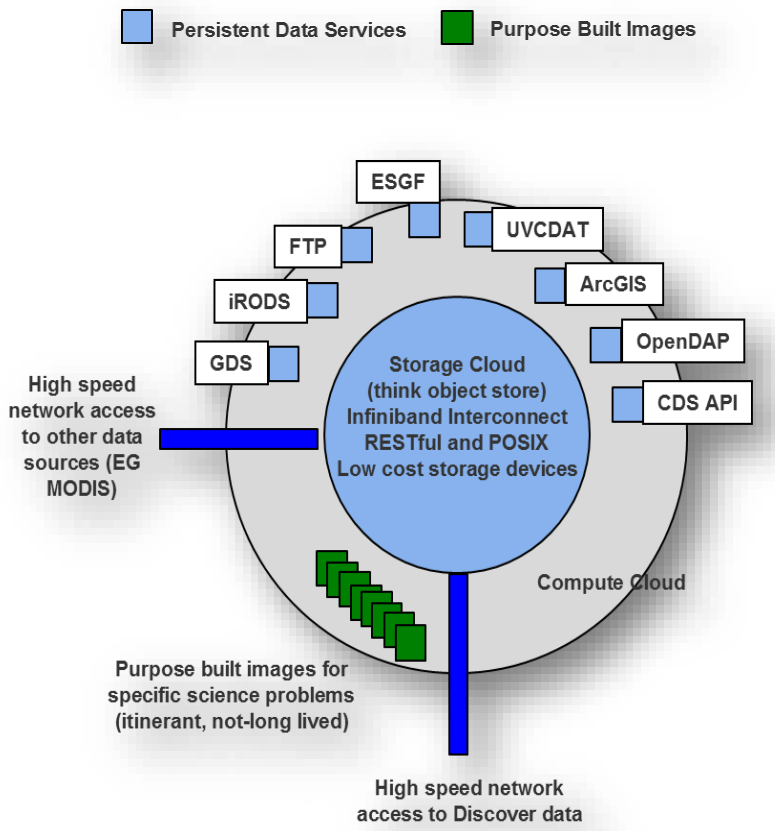
- Lower barrier to entry for scientists
- Customized run-time environments
- Reusable HPC/Discover hardware

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

- Platform-as-a-Service that comes close to matching HPC levels of performance
- Critical Node-to-node communication – high speed, low latency
- Shared, high performance file system
- Management and rapid provisioning of resources



High Performance Science Cloud Conceptual Architecture

ABoVE Science Cloud

To support the ABoVE Computing and Storage Requirements a portion of ADAPT has been dedicated for ABoVE research projects.

Partnership between the CCEO, HEC, CISTO, CDS, and NCCS

- Provide compute, storage, data management, and data publication for the ABoVE campaign using ADAPT resources
- 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

What applications will work best in the ASC?

- Not designed for Message Passing Interface (MPI)
  - These are highly coupled processes performing large amounts of data movement over high speed networks and synchronization
  - Recommend to use HPC systems for this
- ASC is designed more for inherently parallel processing of big data
  - Independent processes written to analyze large data sets
  - ASC has tools to assist users in submitting independent parallel scripts across multiple virtual machines
- Publishing of data
  - Persistent data services created to provide a capability for NASA scientists to share large data
  - Climate Model Data Services supports this

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.	Nodes with 128 GB of RAM, 10 GbE, and FDR IB
 **DataBase** High available database nodes with solid state disk.	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.	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 between 24 and 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.	Storage nodes configured with multiple PB's of RAW storage capacity

ASC Software Stack

External License Servers

Virtual machines can be set up to reach out to external license servers. However, additional time is needed to make requests to poke holes through various NASA firewalls.

Open Source Tools
Python, NetCDF, etc.

How to install open source tools:

- If the open source tool does not need elevated privileges to install, the user will be the responsible for installing this in their home or scratch directories.
- Commonly used tools may be installed in a shared directory for multiple users; the NCCS can assist with this as needed
- If the tool requires elevated privileges, users should submit a ticket to the NCCS for assistance. That tool will have to go through a security vetting process before it can be installed.

Commercial Tools
Intel Compiler (C, C++, Fortran), IDL (4 seats)

Operating Systems
Linux (Debian, CentOS) and Windows

Limited number of Windows systems for remote desktop and primarily for ArcGIS. All compute virtual machines will be Open Source Linux.

Long Term Data Storage

- ASC is not an archive
  - Users must adhere to their data management plan submitted with their proposals
- Large data sets relevant to multiple ABoVE research projects
  - Jointly curated in ASC by ABoVE, CDS, and NCCS team
  - Examples include LandSat, MODIS, NGA/DigitalGlobe, etc.
  - More about this later in this presentation
- Large data sets relevant to individual ABoVE research projects
  - Users will need to make a request to the NCCS
  - The NCCS will work with the ABoVE team to decide how to get the data and how long to retain the data
- Where do you archive your results?
  - NASA DAACs, other
  - Up to the researcher's data management plan

Moving Large Data

- The NCCS can help
  - High speed network connection into the ABoVE Science Cloud (10 GbE)
  - Will discuss this further during the initial requirements gathering with each research group
- Synchronizing data
  - Again, the NCCS can help
- Automated data transfers
  - Can set up automated (cron) jobs to move data into and out of ASC
- Public data only at this time!

NASA Climate Model Data Services

Building Capacity and Advancing Research and Applications

Mark McInerney
CISTO Data Services Lead
http://cds.nccs.nasa.gov
September 2, 2015

NASA Goddard Space Flight Center
Sciences and Exploration Directorate
Computational and Information Sciences and Technology Office

ABoVE Specific CDS Services: Summary

- **Data Staging**
  - Common datasets in the ASC Environment
  - Why Stage Data?
  - Folding ABoVE generated results into common ASC Environment
  - Locating staged data on ASC with the Data Services Manager ODISEA
- **NGA/DigitalGlobe High Resolution Satellite Data Imagery**
  - Goals / Summary
  - NGA based services, Digital Elevation Maps
  - Imagery access (EnhancedView, ESRI)
- **Distribution Services**

ABoVE Specific CDS Services: Staged / Common Data Sets in ABoVE Science Cloud

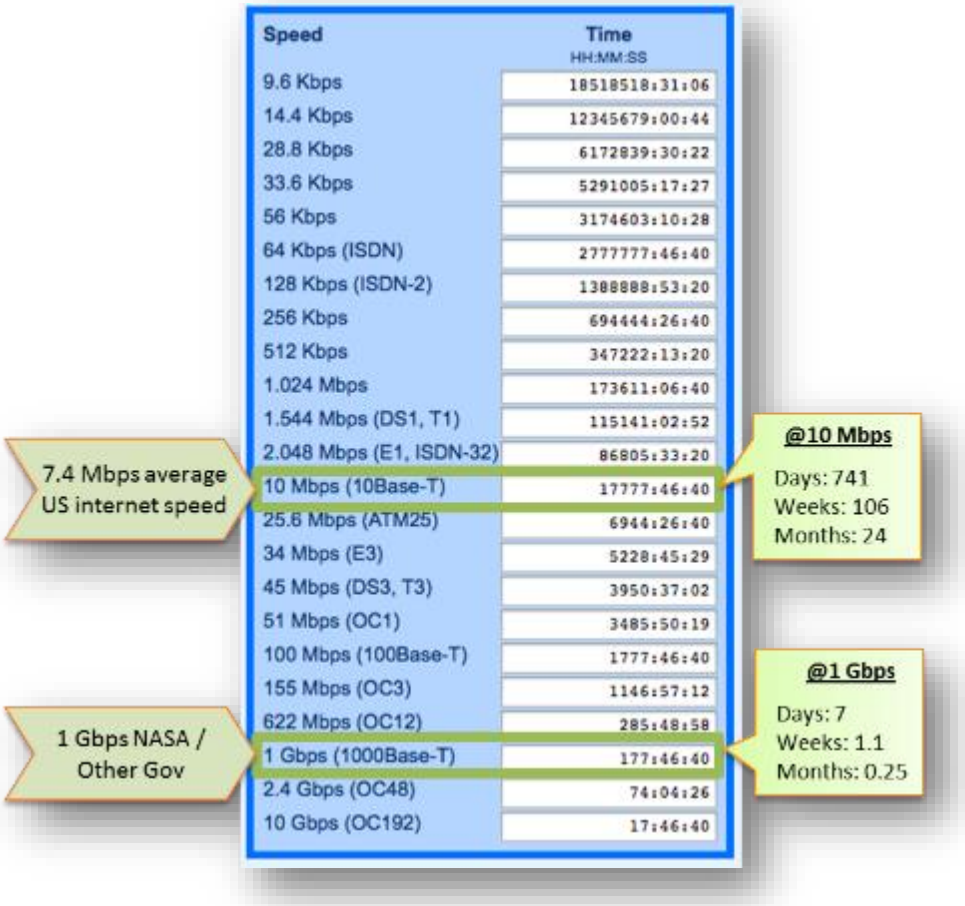
Common datasets "Staged" for ABoVE investigators in ABoVE Science Cloud

- Staged and available for direct use
- Individual investigators don't have to invest time to locate and transfer data into system
- Avoids duplications of large datasets on system
- Additional datasets can be added, including generated data from ABoVE PI
- Data Services Manager to locate data

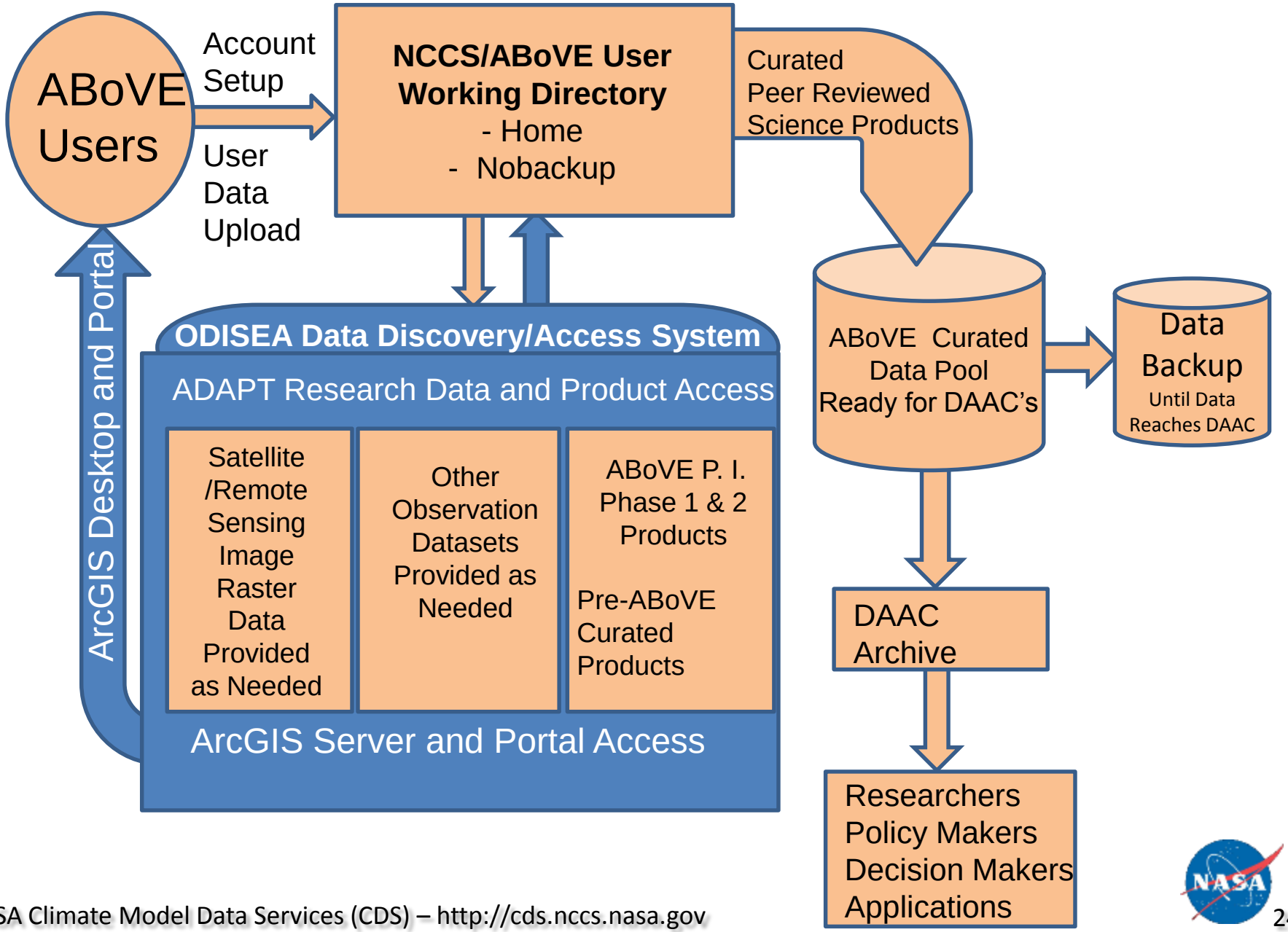
Example Staged datasets

- *Landsat*, Surface reflectance, 123 TB
- *MODIS*, Daily meter surface reflectance, 57 TB
- *MERRA* reanalysis, 80 TB

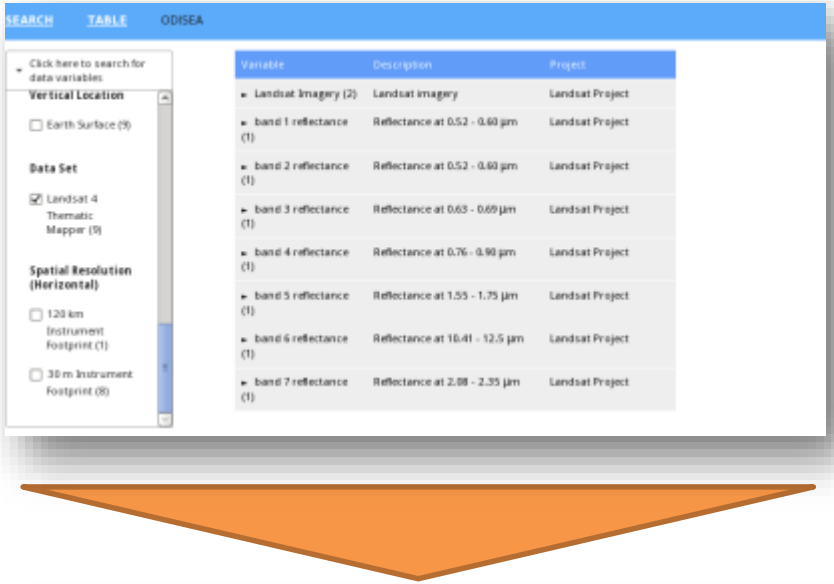
Example Download Times For 80TB



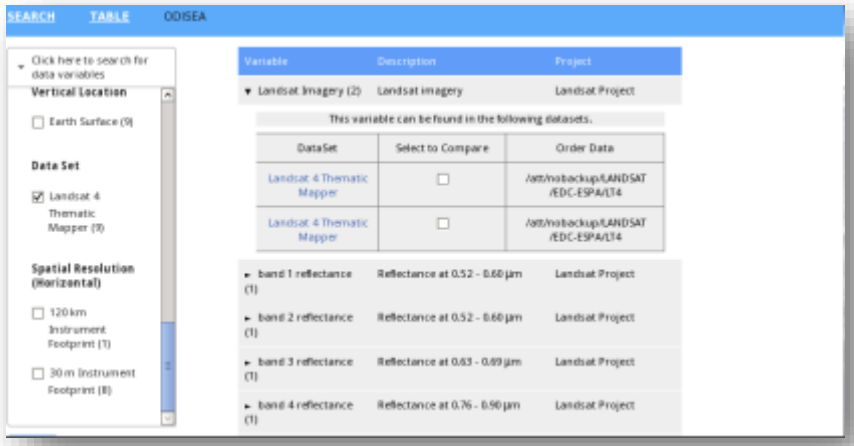
ABoVE Specific CDS Services: Data Lifecycle

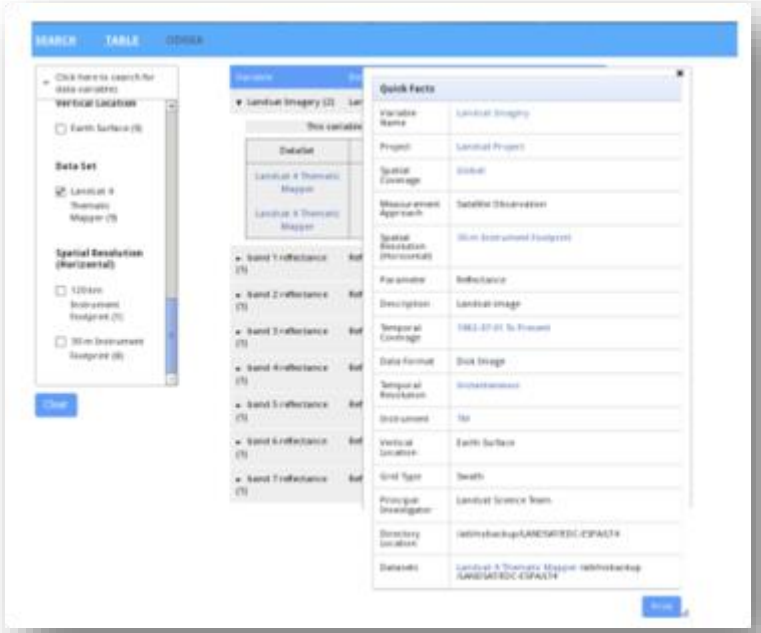


ABoVE Specific CDS Services: Data Services Manager (ODISEA)



ODISEA is a ASC system level tool for ABoVE Researchers to search for and locate system owned staged data





ABoVE Specific CDS Services: NGA/DigitalGlobe High Resolution Commercial Satellite Imagery

National Geospatial Agency (NGA) has licensed all DigitalGlobe ≥ 31 cm satellite imagery for US Federal use, i.e., NSF, NASA and NASA funded projects.

- Archive of 4.2 billion km^2 of data from 2000 to present
- Data from six different satellites: Worldview-1, 2 and 3; Ikonos; Quickbird; and Geoeye-1
- Access to NGA imagery (~3-4/ km^2) at no cost to NASA

Satellite	Bands	Nadir Panchromatic Resolution (m)	Nadir Multispectral Resolution (m)
Ikonos	Pan, R, G, B, Near IR	0.82	3.2
GeoEye	Pan, R, G, B, Near IR	0.41	1.65
Quickbird	Pan, R, G, B, Near IR	0.55	2.16
WorldView-1	Panchromatic only	0.5	N/A
WorldView-2	Pan, R, G, B, Near IR 1, Near IR 2, Coastal, Red Edge, Yellow	0.46	1.85
WorldView-3*	Same as WV-2 plus 8 SWIR bands and 12 CAVIS bands	0.31	1.24

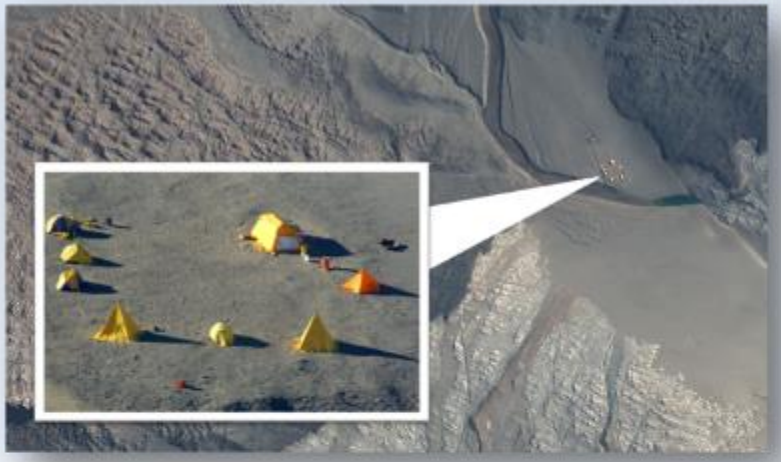
** Worldview 3 images will be offered once available*

DigitalGlobe Satellite Fleet

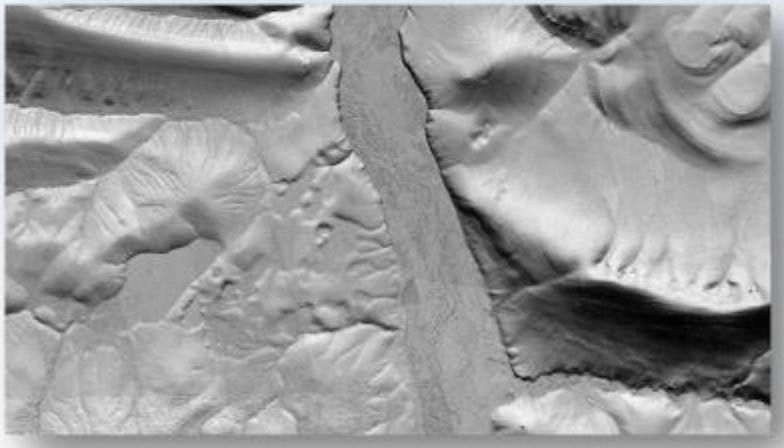
ABoVE Specific CDS Services: General NGA Data Imagery Goals and Support

- ✓ **GOAL 1: (OBTAIN)** Establish access to NGA/DigitalGlobe data through strong partnerships with NGA, Digital Globe, PGC, NASA HQ, and NSF
- ✓ **GOAL 2: (STAGE)** Collect and store domain specific NGA/DigitalGlobe data into the ASC

GOAL 3: (ACCESS) Provide ArcGIS based data access, to locate and access NGA data in ABoVE domain (in-work)



PGC camp in Bull Pass QuickBird-2 (January 2009)

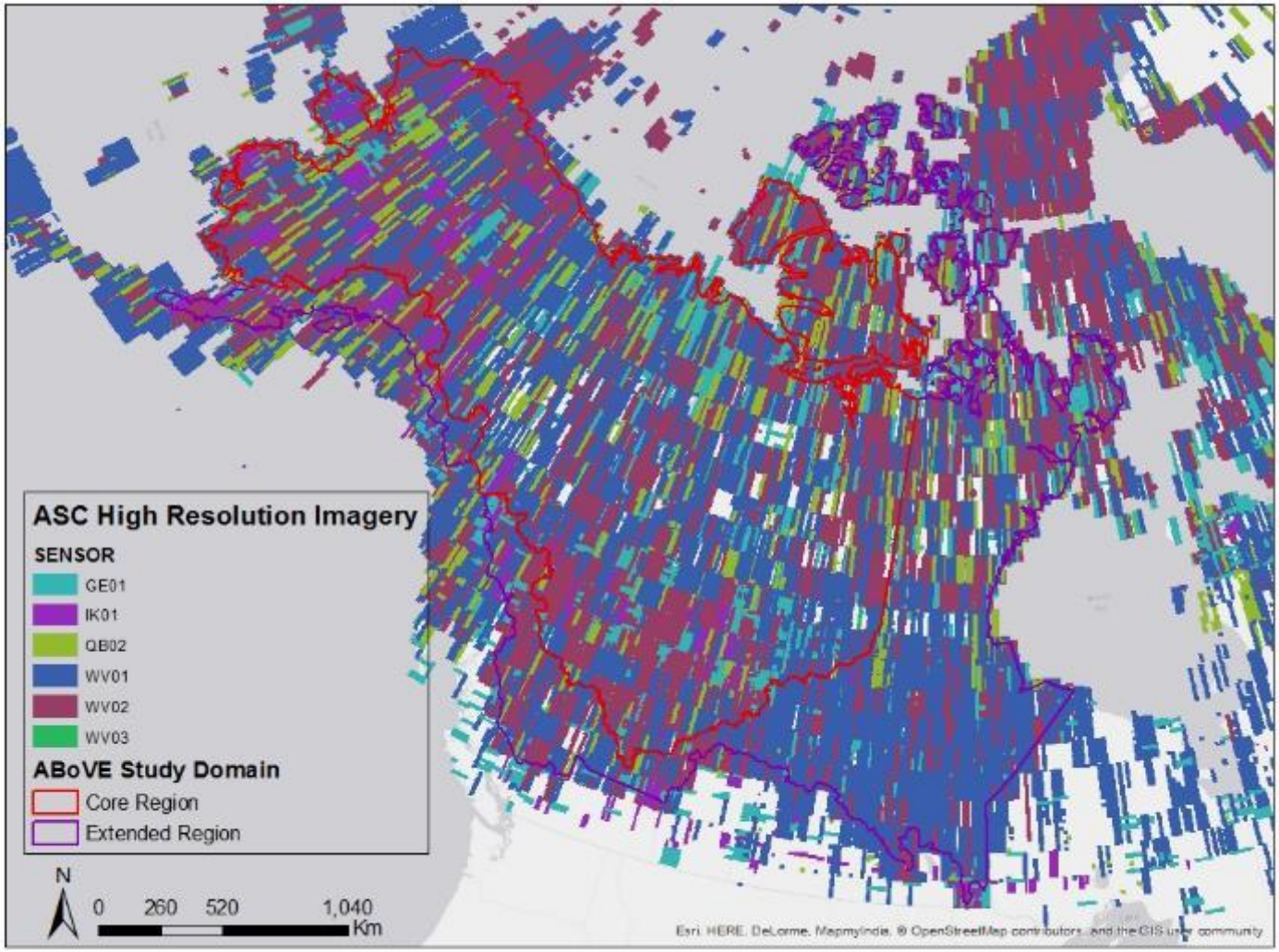


Shaded relief image of a 4m posting elevation model near the Toolik LTER on the north slope of Alaska; Polar Geospatial Center (PGC)

ABoVE Specific CDS Services: Summary of NGA/DigitalGlobe Data Services

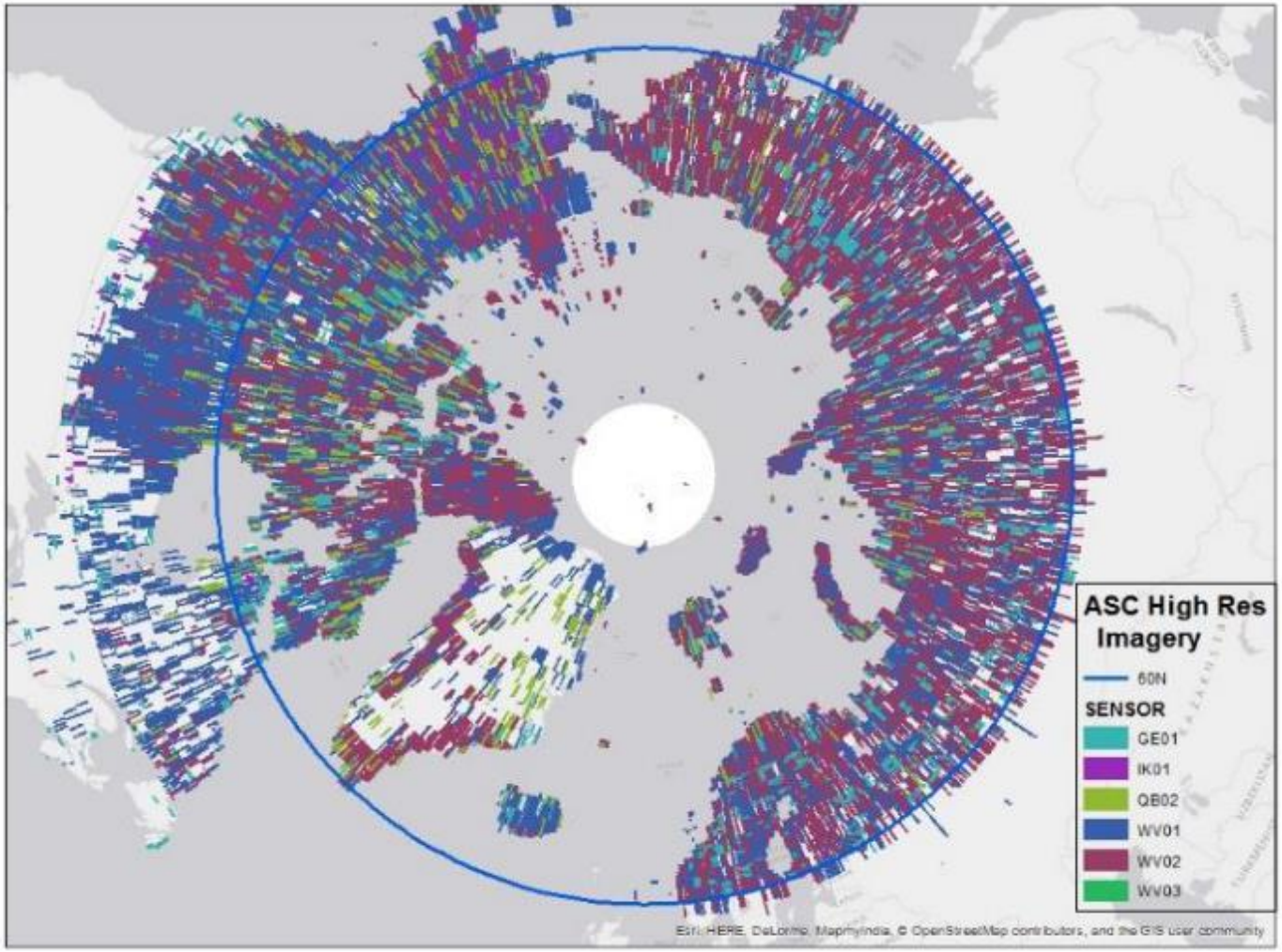
- **Obtain Existing Imagery**: Alaska and entire Arctic north of 60N being tasked by ABoVE via NGA in stereo-panchromatic and multispectral
- **Direct tasking**: for ABoVE core and extended domain below 60N (2-3 year activity)
- **Bulk transfer** and store raw DigitalGlobe (NITF format) images in ABoVE Science Cloud
  - Ordering process
  - Networked connection direct from ABoVE Science Cloud to NGA
  - Over 300 TB now in ASC
- **Value Added NGA/DigitalGlobe Imagery**: Create ~0.5m panchromatic, orthorectified Arctic Mosaic of ABoVE core and extended domain including south of 60N
- **Elevation model** of the ABoVE core and extended domain.
  - General DEMs created by PGC and delivered; Alaska March 2016, Remainder March 2017 from stereoscopic imagery using Ames Stereo Pipeline (ASP) / or Ohio State
  - ASP installed on ABoVE Science Cloud for small scale use by individual PIs

ABoVE Science Cloud DigitalGlobe Imagery: ABoVE Study Domain



Note: Imagery is included from all seasons and for years 1999-2015.

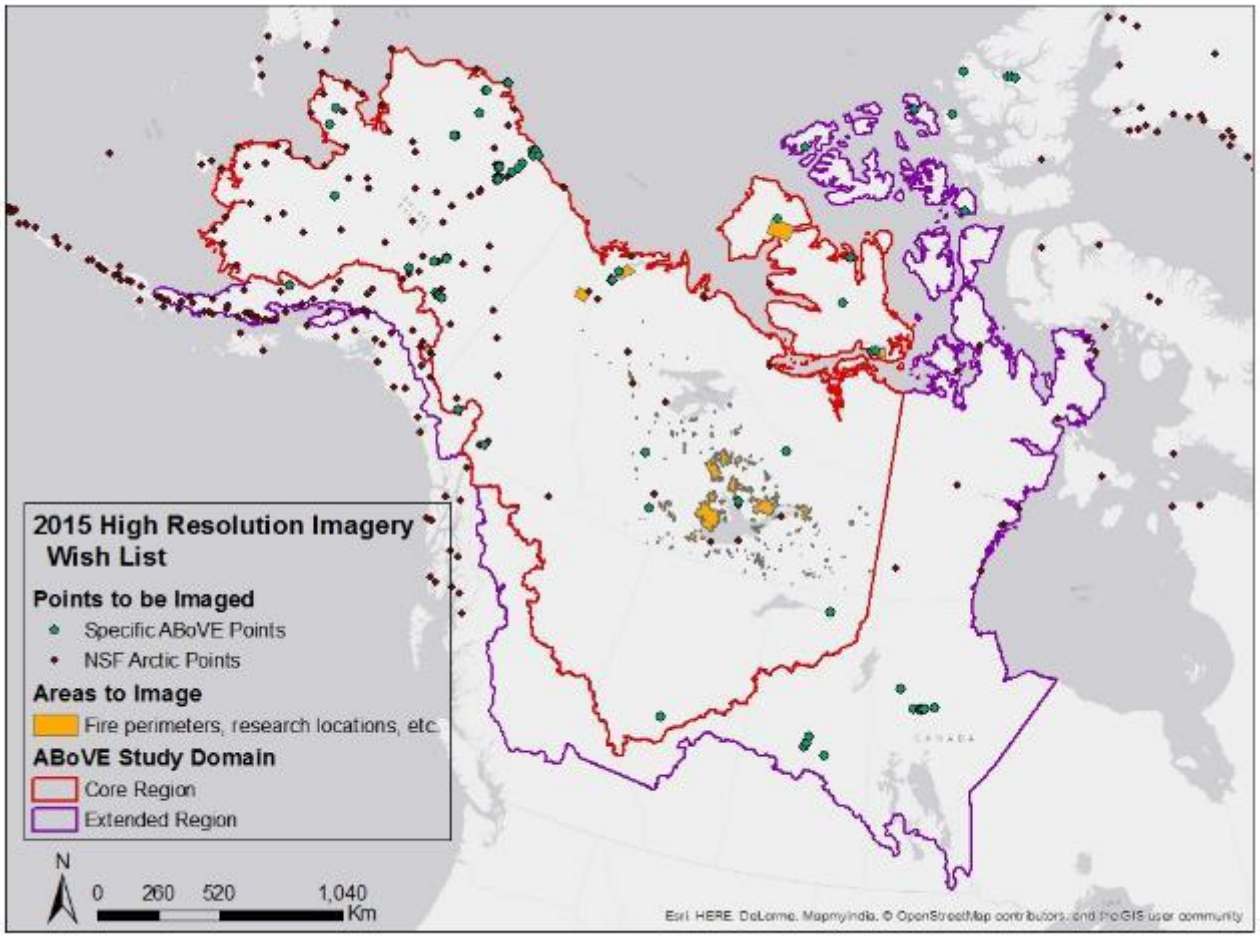
ABoVE Science Cloud DigitalGlobe Imagery: Circumpolar



Note: Imagery is included from all seasons and for years 1999-2015.

ABoVE NGA/DigitalGlobe Imagery Tasking for 2015

Tasking of new imagery coordinated through the CCE Office and PGC

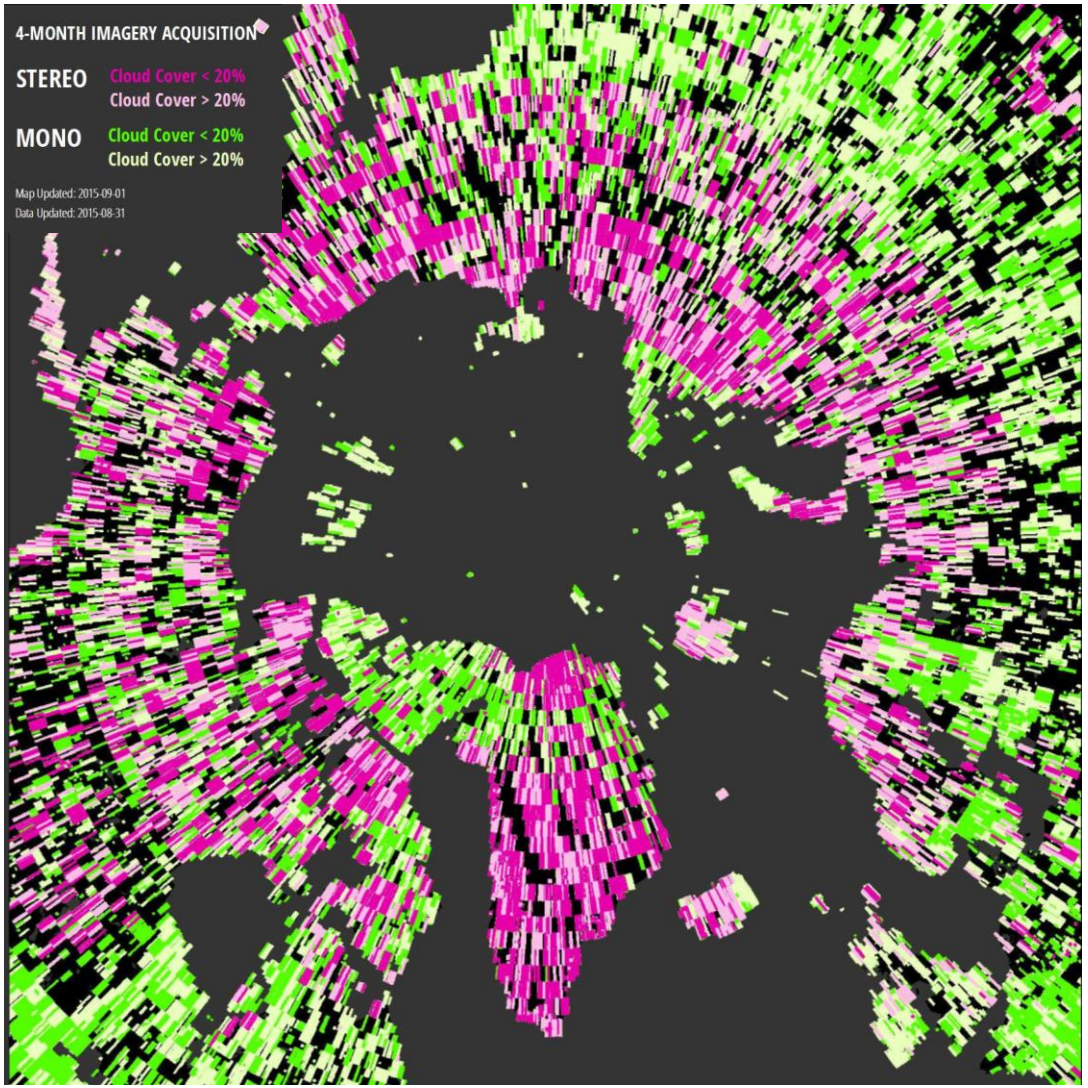


Tasking plan for 2016 will be developed with ABoVE Science Team support.

Liz will follow-up with interested research groups.

DigitalGlobe Arctic Imagery Acquisition: May through August 2015

- Time from acquisition to registry in ASC can take ~3 months or more
- Imagery needed sooner can be requested and prioritized (talk with Liz)
- EnhancedView Web Hosting Service as an alternative



DigitalGlobe EnhancedView Web Hosting Service

ABoVE researchers have online access to view imagery through the DigitalGlobe EnhancedView Web Hosting Service

- Ability to view imagery, however not a space to process imagery
- Imagery from ~2011 and forward is available within the web-based system
- Must adhere to the NASA-NGA data policy and the NextView License to use
- Liz will discuss access with individual research groups

ESRI ArcGIS Services In NCCS Science Cloud

- Science cloud does have NASA enterprise licensed ESRI products installed
  - ArcGIS for Desktop
  - ArcGIS for Server
  - ArcGIS for Portal

- USE 1: ArcGIS services to stage select data by Carbon Cycle & Ecosystem office
  - NGA/DigitalGlobe, other

- USE 2: ESRI application / system level support for ABoVE projects
  - EG. Grant scientist / project ArcGIS server account for data publication (scientist/project managed)

NASA Climate Model Data Services
Data Publication and Distribution Services

Data Publication Services	Protocol	Download	Subsetting	2D Visualization
Web Access For downloading small files	HTTP	✓		
File Transfer Protocol (FTP) Anonymous FTP supporting wget	FTP	✓		
GRads Data Server (GDS) Data subsetting and analysis services	OPENDAP	✓	✓	
Live Access Server (LAS) Data subsetting and analysis services	OPENDAP	✓	✓	
THREDDS Data Server (TDS) subsetting , analysis, & visualization	OPENDAP	✓	✓	✓
Earth System Grid Federation (ESGF) Data access to IPPC CMIP data	OPENDAP	✓		✓
Web Map Service (WMS) Data publication to IPPC CMIP Format	OPENDAP	✓	✓	✓

Questions

CISTO Data Services Lead

Mr. Mark A. McInerney
Phone: 301-286-1491
Email: Mark.McInerney@nasa.gov
URL: http://cds.nccs.nasa.gov

Accessing the ASC

- Requirements discussions with research groups (email coming within a week)
- Request NASA Credentials for ABoVE researchers
- Researchers complete NASA IT training
- Request NASA RSA authentication token
- Researcher completes NCCS paperwork

Requirements Discussion

Be prepared to answer these types of questions for the NCCS to prepare your virtual machines (VMs) in the ASC:

- What type of operating system is required (Linux, Windows)?
- How many compute cores or processors are required per VM?
- How much memory per VM is required?
- What large data sets will you be accessing?
- What software is required?

Getting Started

Once you have an account on the system, each user will initially get the following:

- 1 to 4 virtual machines
- Home Directory (15 GB)
  - This space is replicated within the system but not backed up
  - Recommended use of this space is for your user code and applications
- Scratch Space (2.5 TB)
  - Not backed up and not replicated (protected through hardware RAID)
  - Store working files, intermediate results, etc.
- Access to Large Data Sets
  - Shared storage environment (petabyte) for large datasets to be shared across the ABoVE research projects

Scaling Up

- Once a researcher's ASC account is set up, ABoVE researchers will work to test their code on 1-4 VMs
- Researchers should use these VMs to determine:
  - Representative benchmark timings of their application runs
  - Number of runs required for their analysis
    - For example, this may be the number of Landsat images that need to be processed.
  - Time frame for completion of the analysis
- This data will help the NCCS and ABoVE teams to calculate and prioritize the amount of ASC resources required
  - Discuss benchmarking with Liz or open an NCCS ticket when ready to request additional VMs

ASC Access Questions

- Contact Liz Hoy at elizabeth.hoy@nasa.gov
- ASC tickets can be sent to

  support@nccs.nasa.gov

  301-286-9120
- NCCS Weekly User Teleconference

  Tuesdays at 1:30 p.m. ET

  Meet me number:

  844-467-6272, passcode: 997370 followed by the pound sign (#)

@NASA_ABoVE