ABoVE Science Cloud: *An orientation for proposers to the 2014 A.4 Research Announcement*

During the solicitation please address all questions about the Science Cloud to peter.griffith@nasa.gov

Peter Griffith
Chief Support Scientist,
Carbon Cycle & Ecosystems Office

Mark McInerney
CISTO Data Services Lead

Dan Duffy
NCCS HPC Lead

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

Climate Model Data Services

- Steven Ambrose, Principal Systems Engineer
- Julien Peters, Software Developer
- Eric Winter, Software Developer

Carbon Cycle & Ecosystems Office

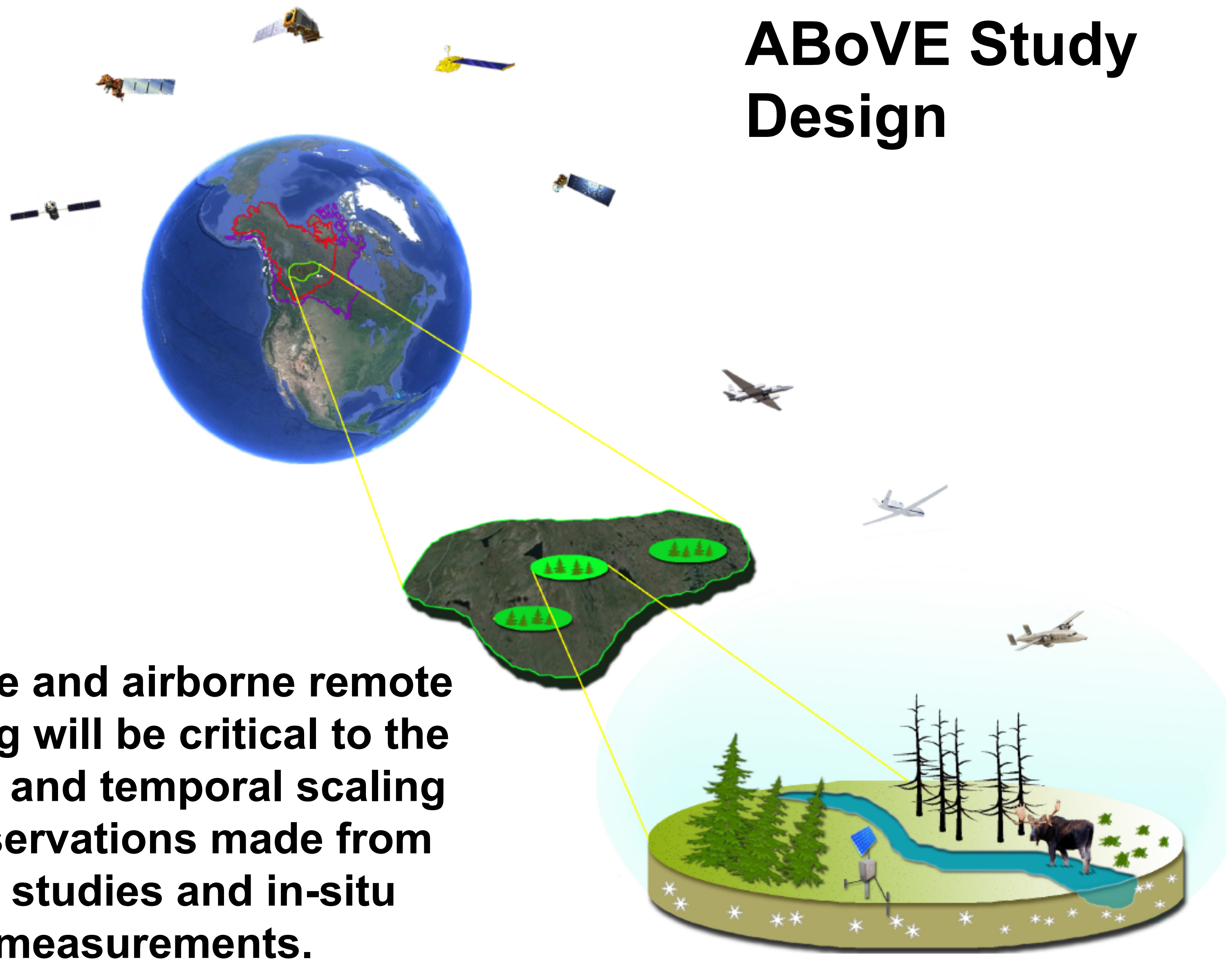
- Liz Hoy, Support Scientist

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

Selected 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.

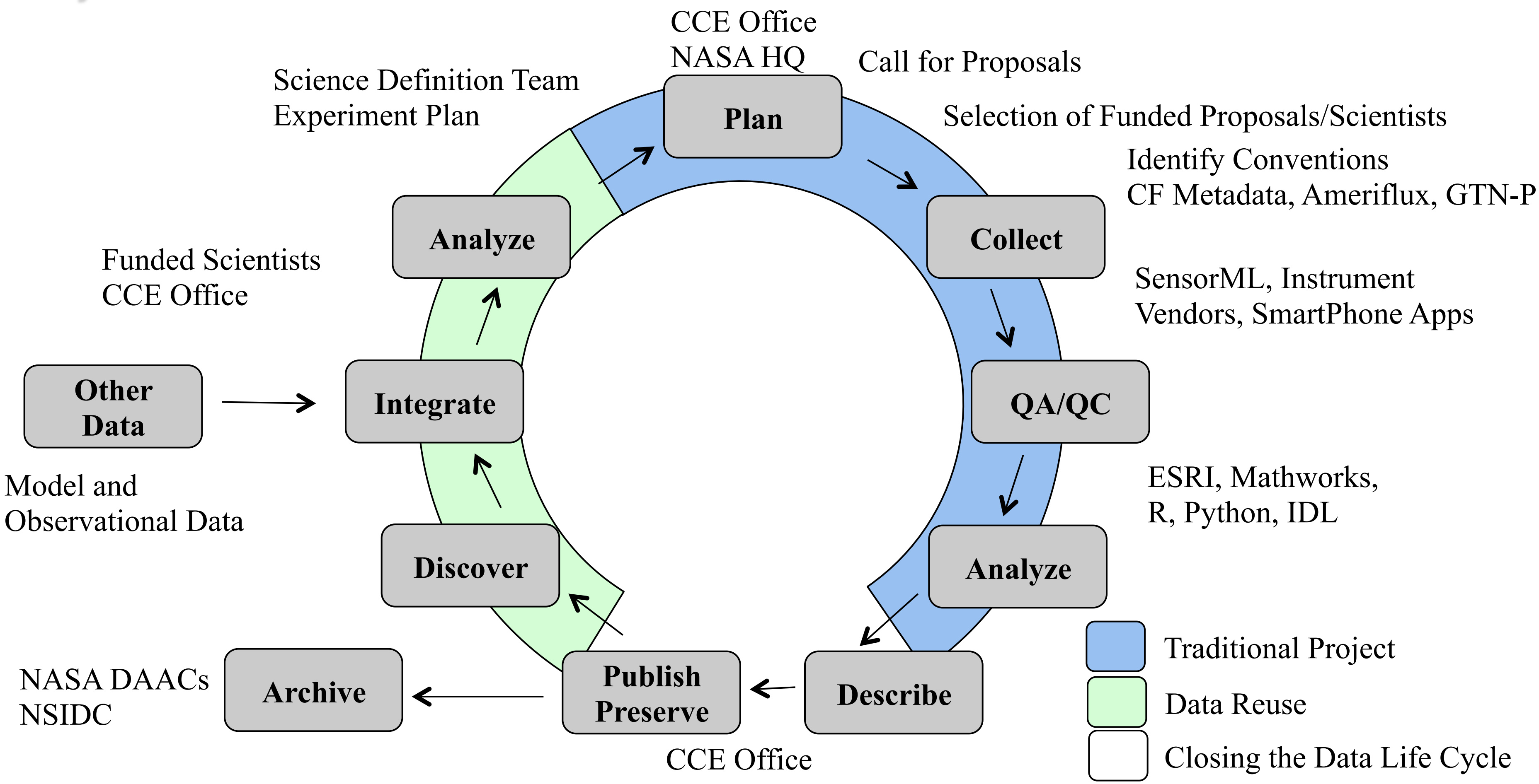
The ABoVE Science Cloud combines high performance computing with emerging technologies and data management with tools for analyzing and processing geographic information to create an environment specifically designed for large-scale modeling, analysis of remote sensing data, copious disk storage for "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.

ABoVE Study Design



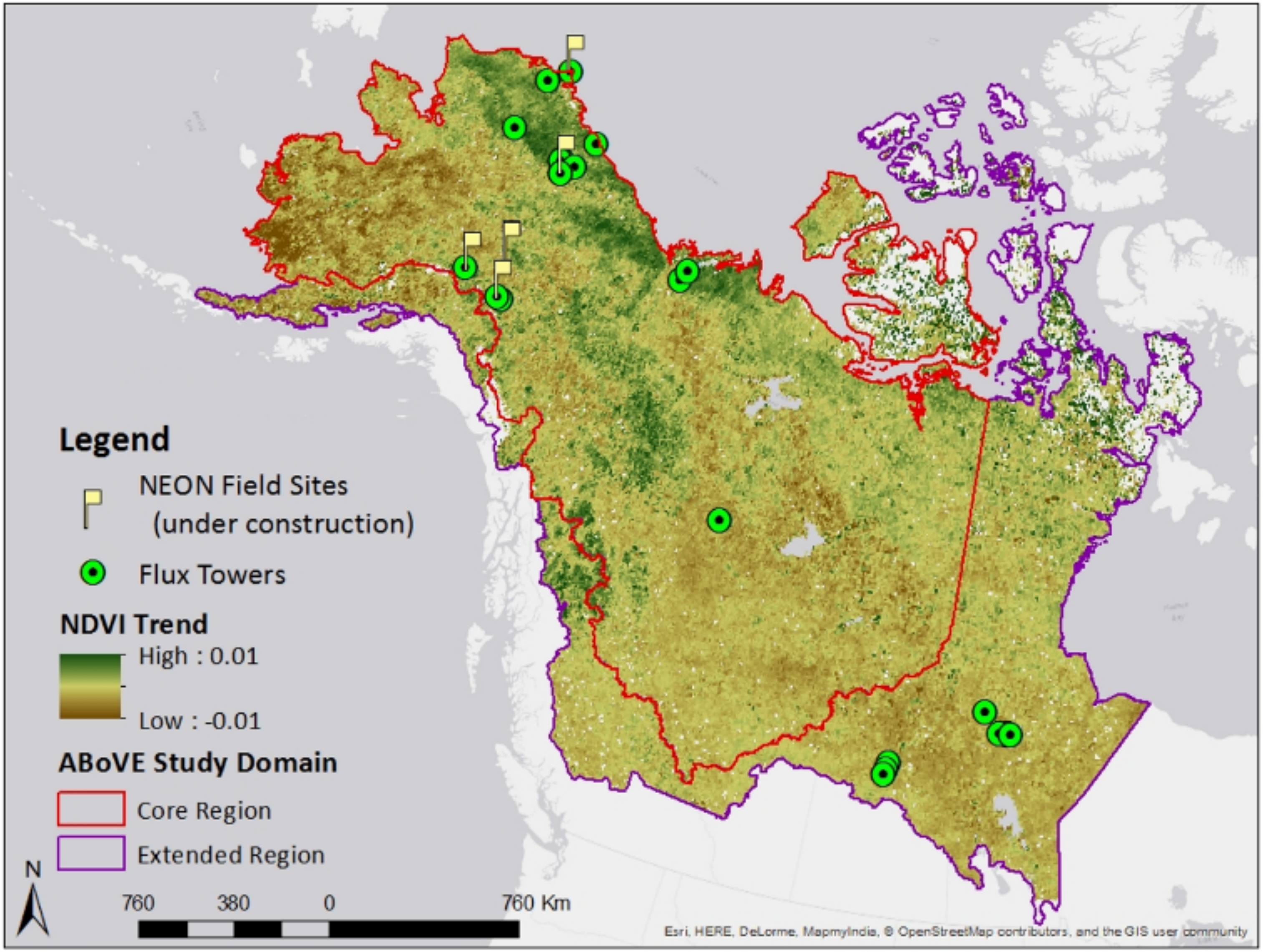
Satellite and airborne remote sensing will be critical to the spatial and temporal scaling of observations made from field studies and in-situ measurements.

CCEO will assist the Science Team throughout the Data Management Lifecycle



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

ABoVE will contribute to and extend in-situ networks



- 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

CALM

AmeriFlux

NATIONAL WEATHER SERVICE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Imiq

HADS
Hydrometeorological Automated Data System

GTN-P
Global Terrestrial Network for Permafrost

Geographic Information Network of Alaska

EARTHDATA Data Discovery Data Centers Community Science Disciplines

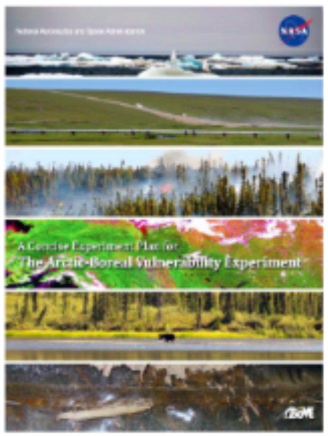
ORNL DAAC Distributed Active Archive Center for Biogeochemical Dynamics

About Us Products Data Tools Help

ABoVE Concise Experiment Plan

The Arctic-Boreal Vulnerability Experiment (ABoVE) Science Definition Team has completed the ABoVE Concise Experiment Plan. This report refines the science questions and rationale for ABoVE, poses a next tier of science questions, and lays out a top-level study design. The study design defines the ABoVE study domain and research areas, identifies the types of studies and observations needed to address the ABoVE science questions, and provides guidance for study implementation.

NASA Terrestrial Ecology will solicit for ABoVE research in 2014 through NASA ROSES Appendix A.4 TERRESTRIAL ECOLOGY.

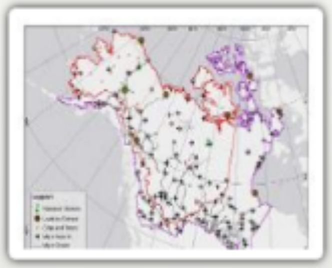


Concise Experiment Plan

ABoVE Concise Experiment Plan (ACEP) Figures as Web Maps

ACEP Figure 1.1: ABoVE Study Domain

ACEP Figure 4.1: Spatial Hierarchy



ACEP Figure A01: Research stations and logistics centers

- *ACEP Erratum, November 2014
- **More About Research Areas for ABoVE**
  The ABoVE Science Definition Team (SDT) identified 15 conceptual Research Areas within the Study Domain that offer the range of environmental conditions necessary to address most of the research questions and objectives for ABoVE. more...
- ABoVE Concise Experiment Plan (10MB pdf)
- Figures and Appendices as interactive web maps

Plan to use data products from pre-ABoVE investigations

Project Lead	Project Title (view profile)
Bourgeau-Chavez, Laura	**Bourgeau-Chavez-02:** Planning and Collection of Data on Boreal Wildfire Effects: Studies of broad-scale 2014 Wildfires in NWT, Canada (view)
Carroll, Mark	**Carroll-01:** Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign (view)
Loboda, Tatiana	**Loboda-01:** Long-Term Multi-Sensor Record of Fire Disturbances in High Northern Latitudes (view)
Munger, J. (Bill)	**Munger-03:** Development of a Data-Assimilation Framework for Integrating 25 Years of Surface and Airborne observations to assess patterns of net CO2 Exchange from Arctic Ecosystems (view)
Schaefer, Kevin	**Schaefer-03:** Remotely-Sensed Active Layer Thickness (ReSALT) Product Derived from InSAR Data Over North American Arctic Regions (view)
Walker, Donald (Skip)	**Walker-01:** Recovery and Archiving of Key Arctic Alaska Vegetation Map and Plot Data for Long-Term Vegetation Analyses (view)

Some ABoVE data products available now: Schaefer

Contribution	Available by: Feb 2015
Product Title: Active Layer Thickness derived from InSAR for the Barrow, Alaska	
Description: We evaluated the Remotely Sensed Active Layer Thickness (ReSALT) product that uses the Interferometric Synthetic Aperture Radar technique to meas seasonal surface subsidence and infer ALT around Barrow, Alaska. We compare ReSALT with ground-based Active Layer Thickness (ALT) obtained using probing calibrated, 500 MHz Ground Penetrating Radar at multiple sites around Barrow. V used a total of 9 PALSAR scenes between June and September and formed 20 interferograms with good coherence and without contamination by obvious ionospheric artifacts. ReSALT accurately reproduced observed ALT within uncert of the GPR and probing data in ~76% of the study area. However, ReSALT was l than observed ALT in ~22% of the study area with well-drained soils and in ~1% the area where soils contained gravel. ReSALT was greater than observed ALT ir some drained thermokarst lake basins representing ~1% of the area. This datase include Subsidence trends, ALT estimates, and the corresponding uncertainty ma	
Expected Users: scientists and researchers interested in environmental change Barrow, Alaska and the use of InSAR for Active Layer Thickness	
Metadata URL(s):	
Data Server URL(s):	
Archived Data Citation: Liu, L., K. Schaefer, A. Chen, A. Gusmeroli, E. Jafarov, Panda, A. Parsekian, T. Schaefer, H. A. Zebker, T. Zhang. 2015. Pre-ABoVE: Remotely Sensed Active Layer Thickness (ReSALT) Barrow, Alaska. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1266	

Contribution	Available by: Feb 2015
Product Title: Active Layer Thickness derived from InSAR for the Prudhoe Bay Region of Alaska	
Description: The InSAR Prudhoe Bay product includes 1992–2000 average Active Layer Thickness (ALT) for an 80x100 km study area of continuous permafrost on the North Slope of Alaska near Prudhoe Bay. ALT is about 30–50 cm over moist tundra areas, and a larger ALT of 50–80 cm over wet tundra areas. The InSAR estimated ALT values match in situ measurements at Circumpolar Active Layer Monitoring (CALM) sites within uncertainties. This dataset include Subsidence trends, ALT estimates, and the corresponding uncertainty map.	
Expected Users: scientists and researchers interested in environemntal change to Alaska's North Slope and the use of InSAR for Active Layer Thickness	
Metadata URL(s):	
Data Server URL(s):	
Archived Data Citation: Liu, L., K. Schaefer, A. Chen, A. Gusmeroli, E. Jafarov, S. Panda, A. Parsekian, T. Schaefer, H. A. Zebker, T. Zhang. 2015. Pre-ABoVE: Estimating 1999-2000 Average Active Layer Thickness, Alaskan North Slope. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA. http://dx.doi.org/10.3334/ORNLDAAC/1267	

Some ABoVE data products available now: Walker

Contribution	Available by:
Product Title: Arctic Alaska Geoecological Atlas	
Description: Abundant ground-based information will be necessary to inform the planned Arctic-Boreal Vulnerability Experiment (ABoVE) activities. The Atlas is comprised of archives of maps and plot-based vegetation data, and associated information. The Map Archive contains map products at several scales and numerous themes. The maps range from detailed geoecological maps, which are polygon-based integrated terrain maps at relatively fine scales, to raster-based map products derived from satellite data and digital elevation models. The Vegetation Plot Archive contains vegetation-plot data, associated environmental data, and other related information from over 3,000 plots in Arctic Alaska.	
Expected Users: ABoVE Science Team	
Metadata URL(s):	
Data Server URL(s): http://geobotanical.portal.gina.alaska.edu/search	
Archived Data Citation:	

Plan to use data products from pre-ABoVE investigations

Project Lead	Project Title (view profile)
Bourgeau-Chavez, Laura	**Bourgeau-Chavez-02:** Planning and Collection of Data on Boreal Wildfire Effects: Studies of broad-scale 2014 Wildfires in NWT, Canada (view)
Carroll, Mark	**Carroll-01:** Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign (view)
Loboda, Tatiana	**Loboda-01:** Long-Term Multi-Sensor Record of Fire Disturbances in High Northern Latitudes (view)
Munger, J. (Bill)	**Munger-03:** Development of a Data-Assimilation Framework for Integrating 25 Years of Surface and Airborne observations to assess patterns of net CO2 Exchange from Arctic Ecosystems (view)
Schaefer, Kevin	**Schaefer-03:** Remotely-Sensed Active Layer Thickness (ReSALT) Product Derived from InSAR Data Over North American Arctic Regions (view)
Walker, Donald (Skip)	**Walker-01:** Recovery and Archiving of Key Arctic Alaska Vegetation Map and Plot Data for Long-Term Vegetation Analyses (view)

Next up:
High Performance Computing at Goddard Space Flight Center

Computational and Informational Science and Technologies Office (CISTO)
http://science.gsfc.nasa.gov/sed/index.cfm?fuseAction=home.main&&navOrgCode=606

CISTO

Climate Model Data Services (CMDS)

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

http://cds.nccs.nasa.gov

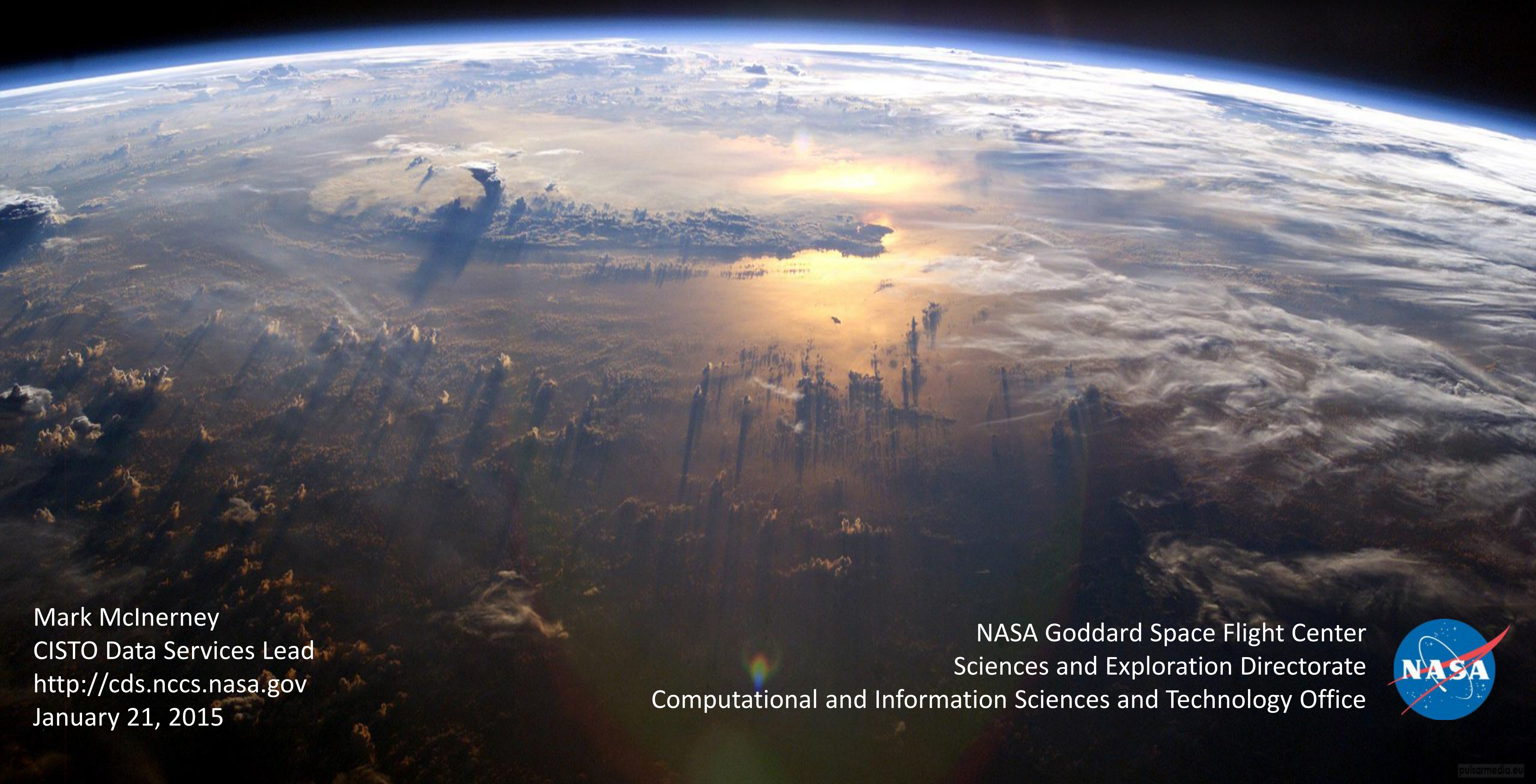
NASA Center for Climate Simulation (NCCS)

- Daniel Duffy, Lead
- High performance computing solutions for NASA science, specifically engineered to meet the needs of climate science

http://nccs.nasa.gov
https://twitter.com/NASA_NCCS

NASA Climate Model Data Services

Advancing Research and Applications with NASA Climate Model Data



Mark McInerney
CISTO Data Services Lead
http://cds.nccs.nasa.gov
January 21, 2015

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

Overview

Climate Model Data Service (CDS) Support for the Arctic-Boreal Vulnerability Experiment (ABoVE)

Background

ABoVE Specific Services

CISTO (NCCS/CDS) Support for ABoVE?

Computational & Information Sciences and Technology Office (CISTO / 606)
Dedicated Technical Support for ABoVE

System Support / Science Cloud

Data Management / Application Support

Climate Model Data Service (CDS) is a team of dedicated "science technical" experts providing data access, data visualization, data distribution, and data management tools and services for ABoVE Science Cloud users

CISTO Background: Conceptual Service Layers

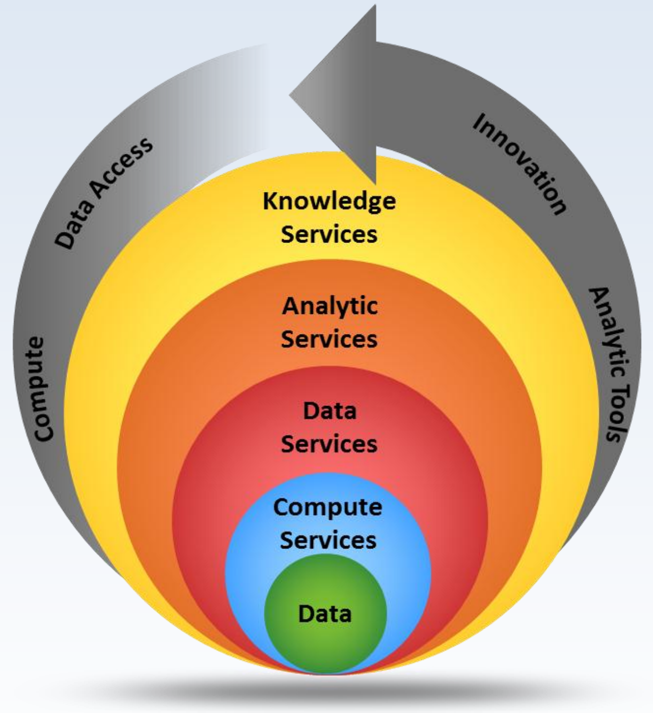
Computational and Information Sciences and Technology Office's (CISTO) Service Layers...

Brings together the tools, data storage and high-performance computing to for timely analysis over large-scale data sets, where the data resides, to ultimately produce societal benefits

Services layers managed by CISTO Units:

- Climate Informatics (CI)
- NASA Center for Climate Simulation (NCCS)
- Climate Model Data Services (CDS)

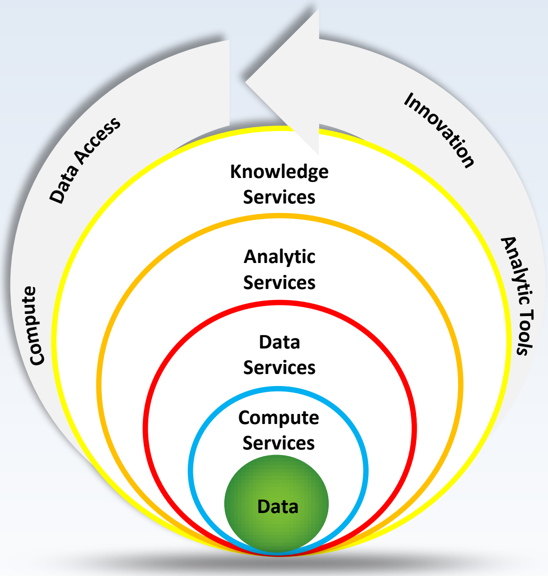
CISTO Conceptual Service Layers



Serving the Earth Science Community

NASA Climate Model: Data

Data To Knowledge Services



Data

Continuous rapid growth of climate model data

Climate Model Data

Climate science data is a Big Data domain that is experiencing unprecedented growth

The size of NASA's space-based observational data is growing dramatically as new missions come online

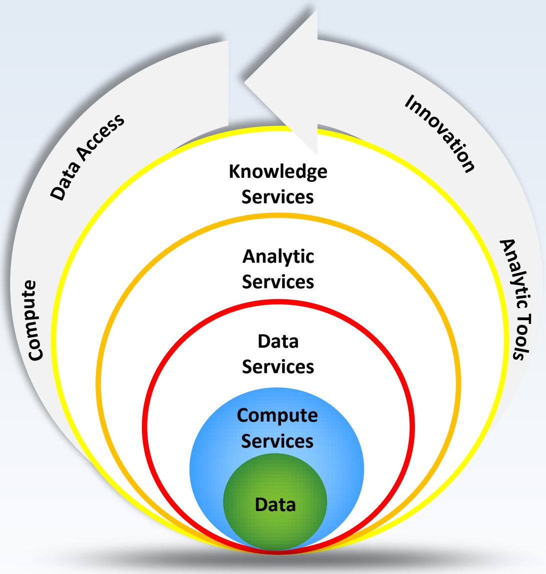
The bigger challenge is the work of data intensive climate scientists whose models regularly ingest observations, produce dynamic petabyte size data sets, and require more efficient data access to enable scientific discovery. CDS is working to fill this existing gap.

Sample Data Sets:

- Nature Run Simulation (GMAO): 5 PB
- Climate Model Output for Intercomparisons, published in Earth System Grid Federation (ESGF): 94.2 TB
  - IPCC CMIP5 Data (GISS ModelE, GEOS5 Decadal): 68 TB
  - NEX-DCP30 Downscaled Climate Projections: 13 TB
  - GeoMIP Geoengineering Data: 13 TB
  - Reanalysis (e.g. MERRA, CFSR): 200 GB
- MERRA/AS: 80 TB (or 160 TB uncompressed, 480 TB in Hadoop)
- NASA Land & Cryosphere Science Very-High Resolution Data Delivery: 1 PB

NASA Climate Model: Compute Services

Data To Knowledge Services



Compute Services
NASA's existing high-end computing environment

Climate Model Compute Services

CDS is partnered with NASA Center for Climate Simulation (NCCS) which provides high performance computing services for climate scientists on "Discover" and where generated model output is then accessed and published through CDS.

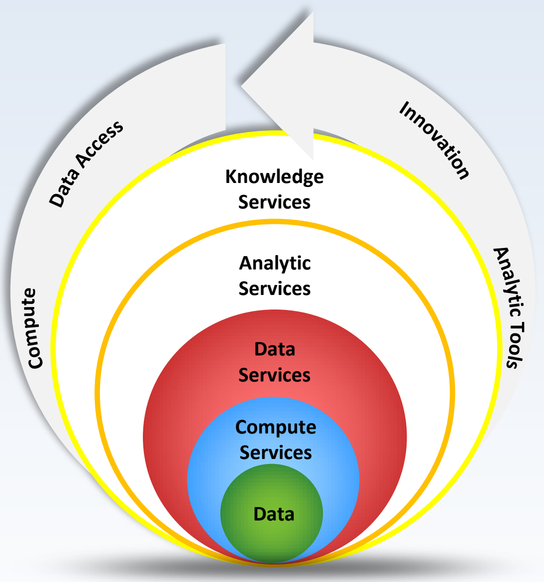
<u>Discover Compute Summary</u>:

- 1.121 petaflops peak = 1,121,000,000,000,000 floating-point operations per second
- 3,354 compute nodes
- 43,048 Intel Xeon processor cores (603 teraflops)
- 64 NVIDIA Tesla M2070 GPUs with 28,672 "streaming GPU" CUDA cores (33 teraflops)
- 480 Intel Phi Many Integrated Core (MIC) co-processors with 28,800 cores (485 double precision teraflops)
- 102.088 terabytes main memory
- InfiniBand interconnect: 20, 40, or 56 gigabits per second

HEC *computing services* generate the data for *Data Services* ….

NASA Climate Model: Data Services

Data To *Knowledge Services*



Data Services
Currently being operationalized

Climate Model Data Services

NASA's Climate Model Data Services (CDS) provides IT infrastructure, tools and services, for climate science community to visualize, analyze, compare, and publish large volumes of climate model data.

Allows climate scientists to focus on science and not scientific data management and supporting technology infrastructure.

Efficient access to climate model data supporting broader use. E.g., Understanding multiple data formats or location of data not required

Providing end-to-end data management of large-scale climate model repositories

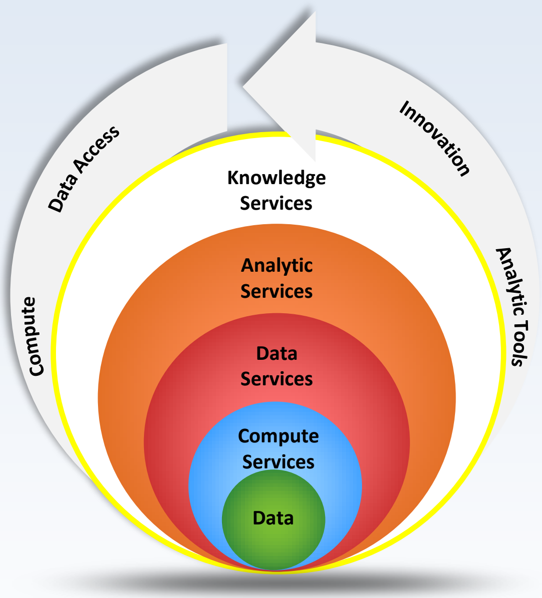
<u>Sample Services:</u>

- CDS API for direct access to climate model data
- Large-scale data federation
- Large-scale publication and distribution Services
- 3D Data Visualization and Analysis
- Digital Object Identifier (DOI) Services
- Climate Ontology Services

Data Services infrastructure then feeds *Analytic Services* ….

NASA Climate Model: Analytic Services

Data To *Knowledge Services*



Analytic Services
In BETA testing

Climate Model Data Analytics Services

Climate Analytics-as-a-Service (CAaaS) brings together the tools, data storage and high-performance computing to perform timely analysis over large-scale data sets, where the data resides, to ultimately produce societal benefits

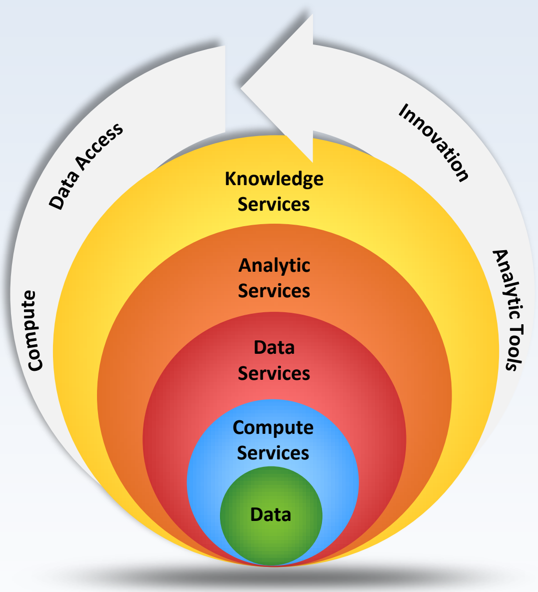
Example CDS CAaaS: MERRA Analytic Services (MERRA/AS):

- Enabling MapReduce analytics over NASA's Modern-Era Retrospective Analysis for Research and Applications (MERRA) data collection
- MERRA reanalysis integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of hundreds of climate variables
- MERRA/AS brings together end-to-end CAaaS Capability:
  - High performance, data proximal analytics
  - Scalable data management
  - Software appliance virtualization
  - Adaptive analytic
  - Domain-harmonized API

Analytic Services infrastructure then *feeds Knowledge Services…*

NASA Climate Model: Knowledge Services

Data To *Knowledge Services*



Knowledge Services
Future?

Climate Model Knowledge Services

Tight integration of knowledge management and organization to share and enable broader climate science discovery …….

Data	Compute Services	Data Services	Analytic Services	Knowledge Services
Scientists Develop algorithms Create climate models	Climate models executed on Supercomputers Climate model data files	Model Data… • Indexed • Cataloged • AS Formatted Data Sources Federated Publication & Access	Open services that perform timely analysis over large-scale data sets, where the data resides	?

Are there knowledge services to develop with the defined tighter integration of climate science model **data**, **compute services**, **data services**, and **analytic services**….?

Knowledge Management: Achieving objectives for the climate science community by acquiring, creating, and sharing climate model data, information, and knowledge across multiple disciplines and a supportive technical foundation

Knowledge Organization: Activities that document, index, and classify climate model data , by subject matter experts, with supportive computer algorithms and infrastructure

Overview

Climate Model Data Service (CDS) Support for the Arctic-Boreal Vulnerability Experiment (ABoVE)

Background

ABoVE Specific Services

Access to Common Datasets in Science Cloud

Common Datasets "Staged" for ABoVE Investigators in NCCS Science Cloud

- Staged and available for direct use
- Individual investigators don't have to locate and download
- Additional datasets can be added, requirements driven by ABoVE
- Data Management Services
  - Location service, metadata support

Existing Staged Datasets (Pre-ABoVE)

- *Landsat (study domain)*, surface reflectance, 50 TB (250 uncompressed) [temporal coverage 1990-2012 by time projects are funded]
- *MODIS (circumpolar)*, daily surface reflectance (MOD/MYD09GA) & 8-day surface reflectance composites (MOD09A1) 248TB
- *MERRA*, GEOS-5 reanalysis, 81 TB
- *Hourly Weather Observations* (METAR), 350 stations, ~38 MB per week
- *NGA (study domain),* Very-High Resolution Commercial Satellite Imagery, E1.8 PB

Speed	Time HH:MM:SS
9.6 Kbps	18518518:31:06
14.4 Kbps	12345679:00:44
28.8 Kbps	6172839:30:22
33.6 Kbps	5291005:17:27
56 Kbps	3174603:10:28
64 Kbps (ISDN)	2777777:46:40
128 Kbps (ISDN-2)	1388888:53:20
256 Kbps	694444:26:40
512 Kbps	347222:13:20
1.024 Mbps	173611:06:40
1.544 Mbps (DS1, T1)	115141:02:52
2.048 Mbps (E1, ISDN-32)	86805:33:20
10 Mbps (10Base-T)	17777:46:40
25.6 Mbps (ATM25)	6944:26:40
34 Mbps (E3)	5228:45:29
45 Mbps (DS3, T3)	3950:37:02
51 Mbps (OC1)	3485:50:19
100 Mbps (100Base-T)	1777:46:40
155 Mbps (OC3)	1146:57:12
622 Mbps (OC12)	285:48:58
1 Gbps (1000Base-T)	177:46:40
2.4 Gbps (OC48)	74:04:26
10 Gbps (OC192)	17:46:40

7.4 Mbps average US internet speed

@10 Mbps
Days: 741
Weeks: 106
Months: 24

1 Gbps NASA / Other Gov

@1 Gbps
Days: 7
Weeks: 1.1
Months: 0.25

Example Download Times for 80TB

Investigator Generated Data

- Short-term storage and distribution services provided
- Folded into science cloud staged or common datasets for use by other investigators
- Long-term archive and distribution outside the Science Cloud (DAAC/ORNL, other)
  - CDS data management services assistance provided

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	✓	✓	✓

Climate Model Data Services

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, other

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

NGA High Resolution Commercial Satellite Imagery Support for ABoVE: Overview / Background

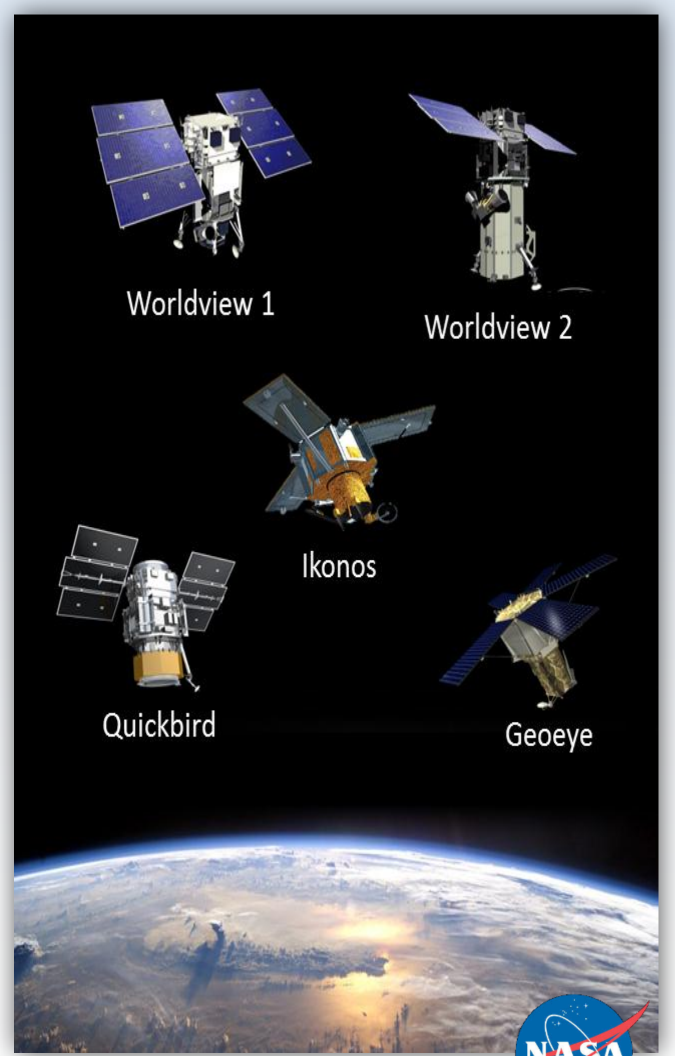
National Geospatial Agency (NGA) has licensed all DigitialGlobe > 0.5 m satellite imagery for US Federal use, i.e., NSF, NASA and NASA funded projects…and so ABoVE.

- Archive of 4.2 billion km^2 of data from 2000 to present
- Data from five different satellites; Worldview-1, Worldview-2, 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	1.0	4
GeoEye	Pan, R, G, B, Near IR	0.5	2.0
Quickbird	Pan, R, G, B, Near IR	0.6	2.4
WorldView-1	Panchromatic only	0.5	N/A
WorldView-2	Pan, R, G, B, Near IR 1 Near IR 2,Coastal, Red Edge, Yellow	0.5	2.0

** Worldview 3 images will be offered once available*

NGA (DigitialGlobe) Satellite Fleet

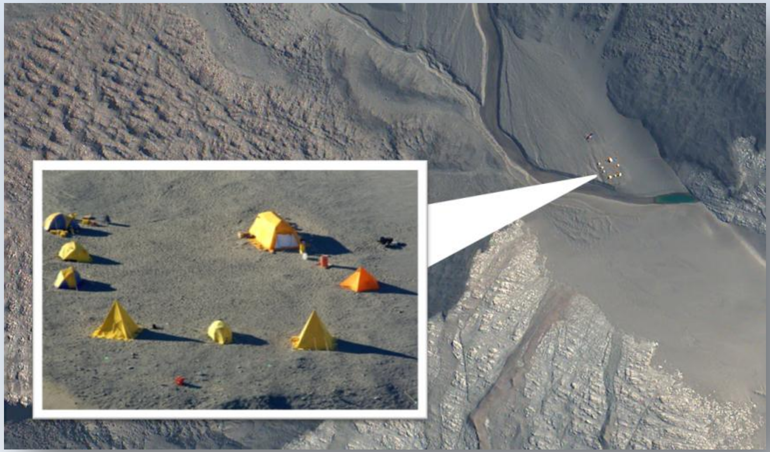


NGA High Resolution Commercial Satellite Imagery: High Level Goals for ABoVE

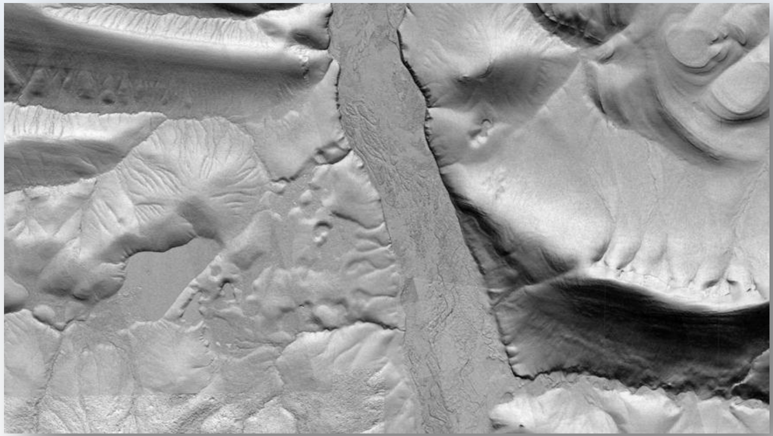
GOAL 1: (OBTAIN) Establish access to NGA data through strong partnerships with NGA, Digital Globe, PGC, and NSF

GOAL 2: (STAGE) Collect and store ABoVE domain specific NGA data into the NCCS Science Cloud

GOAL 3: (ACCESS) Provide short-term data access, system compute, processing and analytics services for ABoVE scientists



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

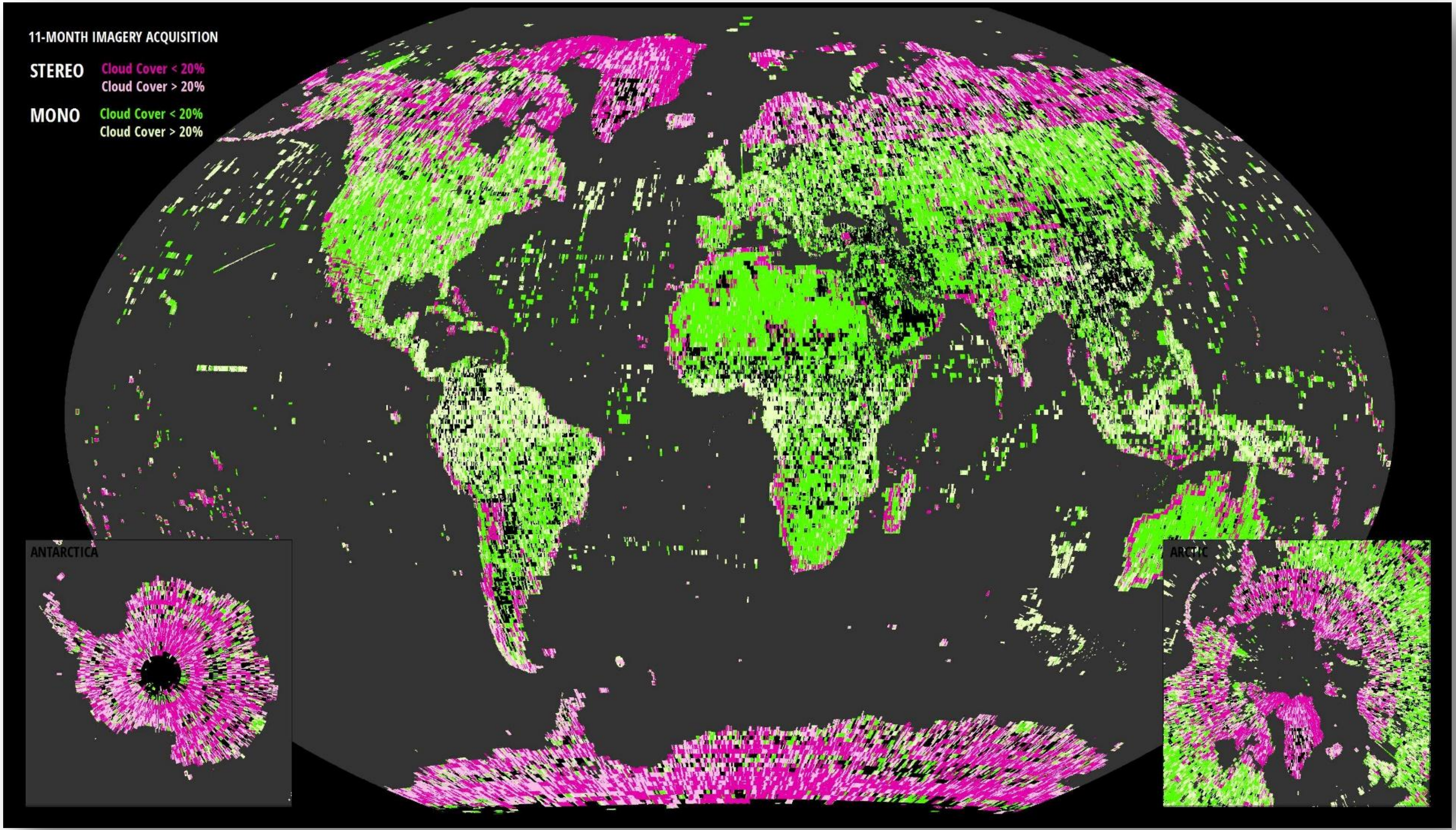


Shaded relief image of a 4m posting elevation model near the Toolik LTER on the north slope of Alaska; Polar Geospatial Center (PGC)

ABoVE / NGA Data Specifics

- **Obtain Existing Imagery**: Alaska and entire Artic north of 60N already tasked by NGA in stereo-panchromatic and multispectral, obtain from PGC
- **Direct tasking**: For ABoVE core and extended domain below 60N and 60N circumpolar (2-3 year activity)
- **Bulk transfer** and store raw NGA (NTF) images in ABoVE Science Cloud
  - Ordering process
  - Networked connection direct from Science Cloud to NGA
- **Value Added NGA Imagery**: Create ~0.5m panchromatic, orthorectified Arctic Mosaic of ABoVE core and extended domain including south of 60N
- **Elevation model** “first pass” of the ABoVE core and extended domain
  - “first pass” meaning models will not be corrected for vegetation and built structures and so should be used as first pass product
  - DEMS from stereoscopic imagery using Ames Stereo Pipeline (ASP)

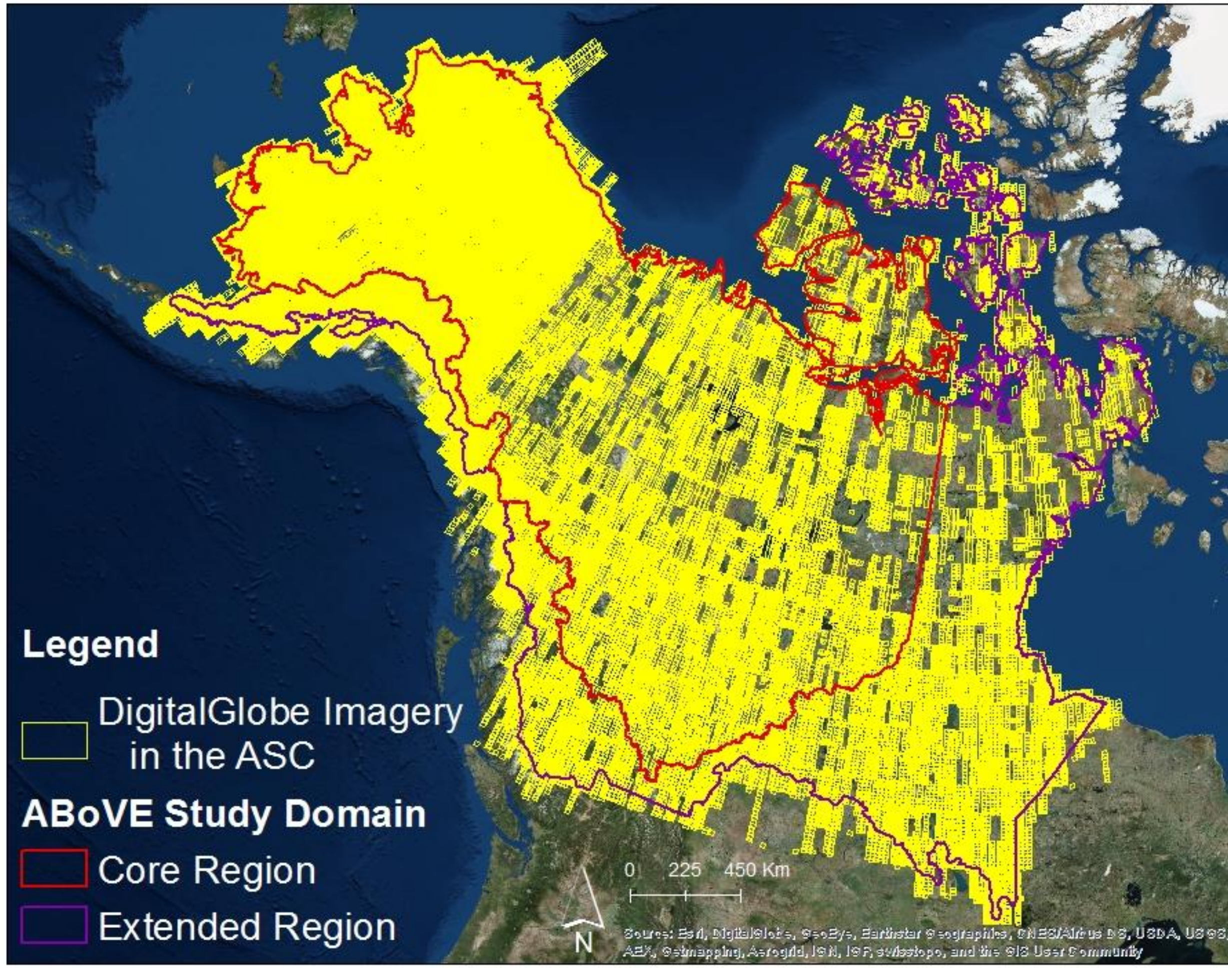
11 Month Collection of NGA/DG Images at PGC



Current NGA Data Holdings in ABoVE Science Cloud

Current NGA Data holdings in ABoVE Science Cloud as of January 2015

- > 100 TB
- Image is zoomed in, numerous gaps remain (EG Alaska)



** as of January 2015*

Climate Model Data Services

Questions

CISTO Data Services Lead

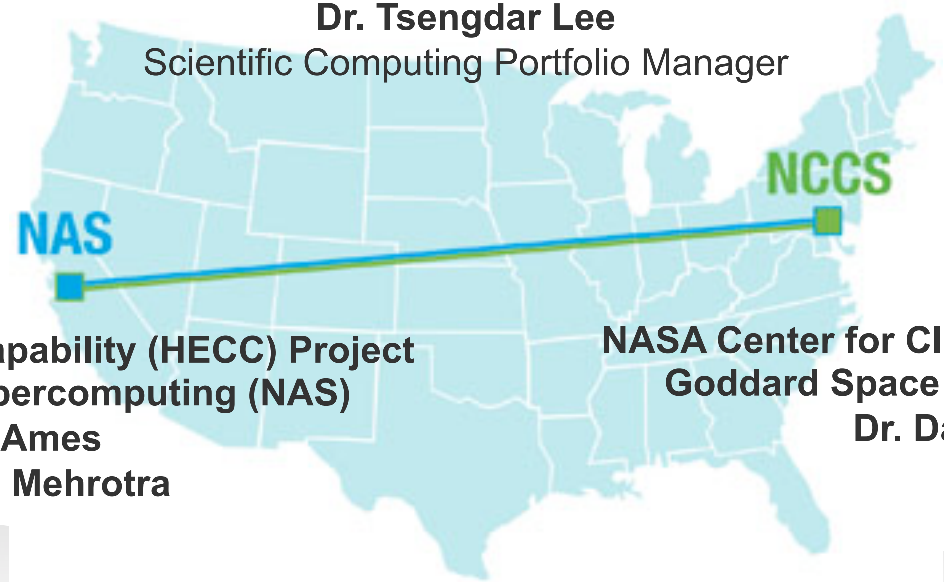
Mr. Mark A. McInerney
Phone: 301-286-1491
Email: Mark.McInerney@nasa.gov
URL: http://cds.nccs.nasa.gov

ABoVE Science Cloud
January 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 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.

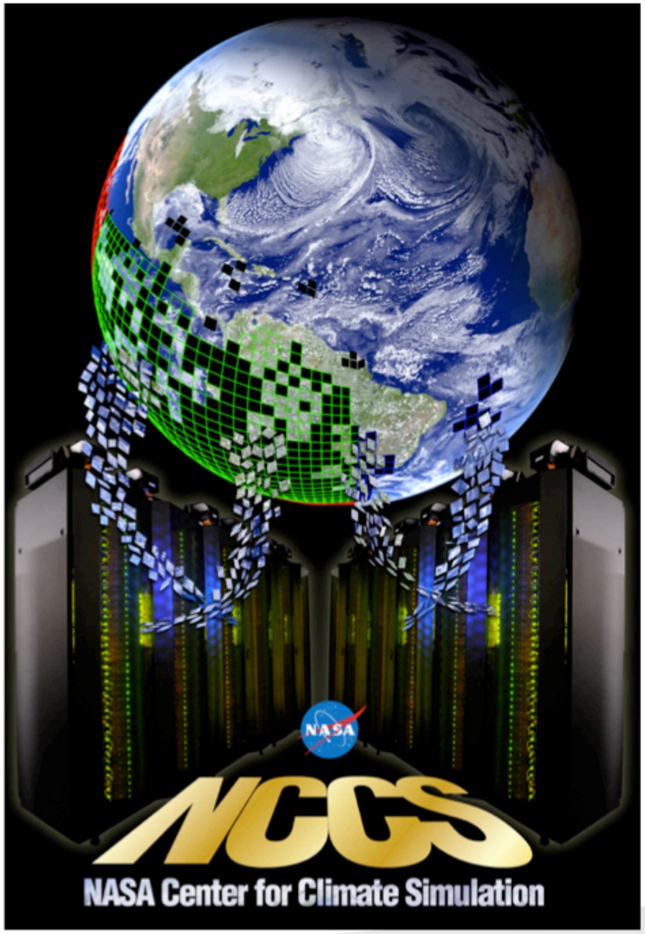
- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- **Advanced Data Analytics Platform (ADAPT) – High Performance Science Cloud**

Primary Customers (NASA Climate Science)

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

High-Performance Science

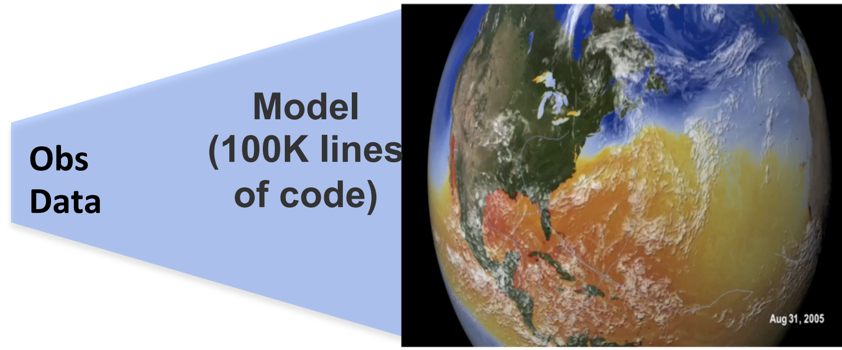
- http://www.nccs.nasa.gov
- Located in Building 28 at Goddard



HPC Applications Support

Takes in small input and creates large output

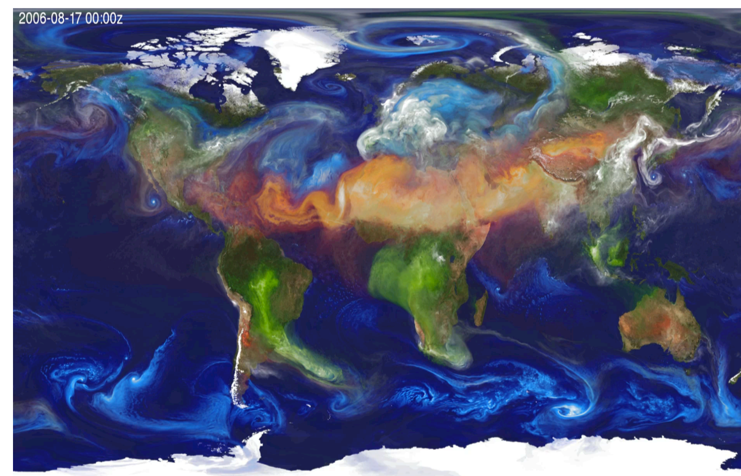
- Using relatively small amount of observation data, models are run to generate forecasts
- Fortran, Message Passing Interface (MPI), large shared parallel file systems
- Rigid environment – users adhere to the HPC systems



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

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/



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.

High Performance Solutions

- High performance computing capabilities within the NCCS are available for ABoVE research projects.
- The NCCS currently has a large amount of high performance computing and storage available for large-scale simulations
  - Petaflops of peak computing capability with 1,000's of compute nodes
  - Petabytes of high performance storage
  - Help desk, ticketing system, applications support, etc.
- For more information about the high performance computing capabilities, please go to our web site
  - www.nccs.nasa.gov

Analysis Application Support

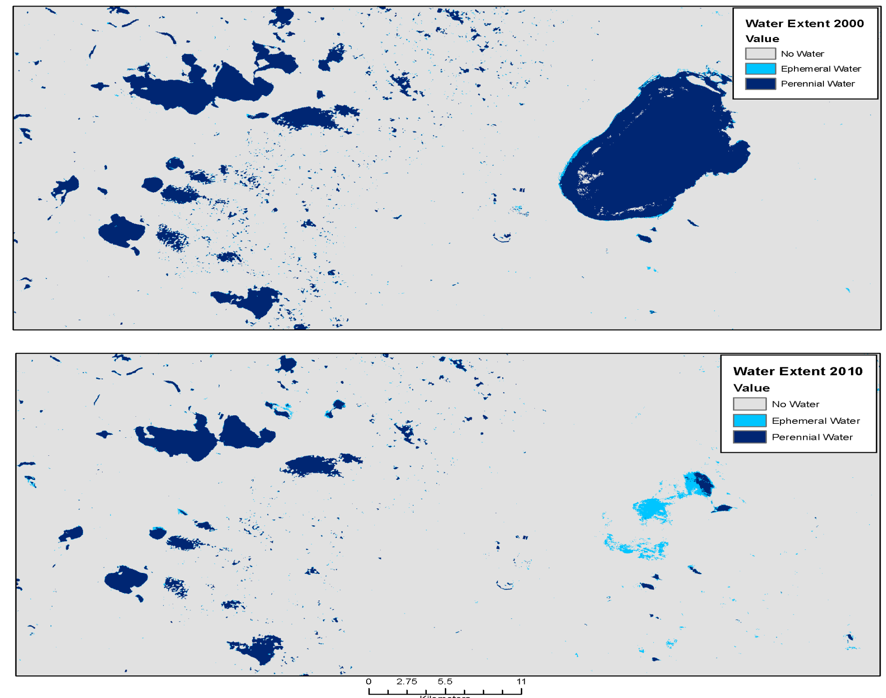
Takes in large amounts of input and creates a small amount of output

- Use large amounts of distributed observation and model data to generate science
- Analysis applications are typically 100's of lines of code
- Python, IDL, Matlab, custom
- Agile environment – users run in their own environments

Examples

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

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



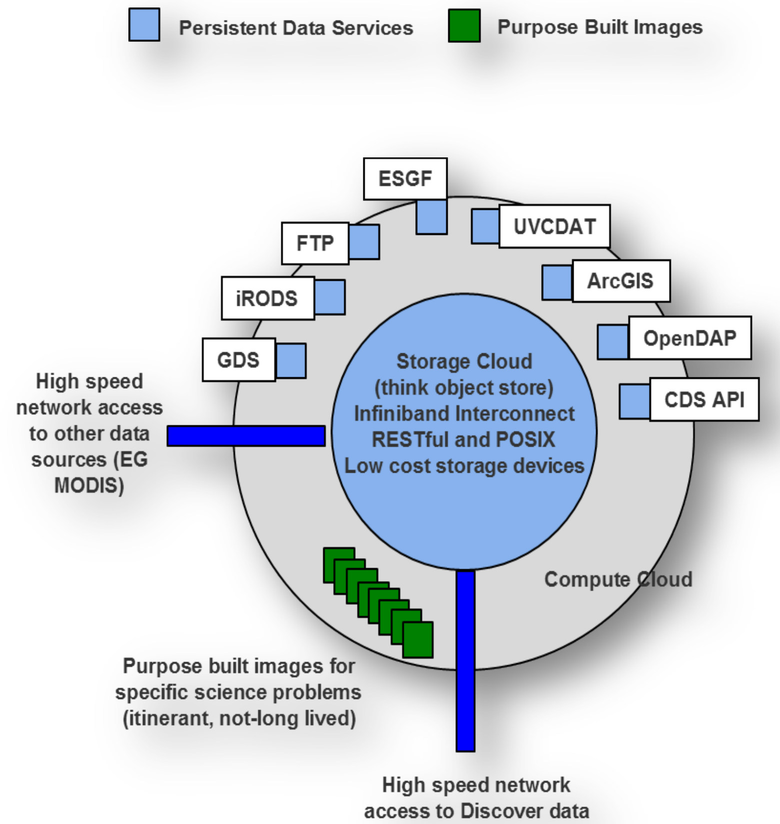
Advanced Data Analytics Platform (ADAPT) "High Performance Science Cloud"

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

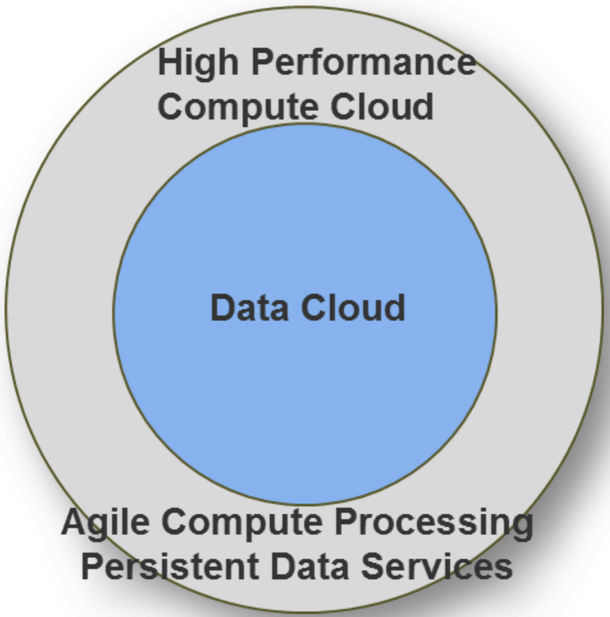
Difference between a commodity cloud
- Platform-as-a-Service that comes close to matching HPC levels of performance
- Critical Node-to-node communication – high speed, low latency
- Shared, high performance file system
- Management and rapid provisioning of resources



High Performance Science Cloud Conceptual Architecture

ABoVE Science Cloud

Solution for the ABoVE Computing and Storage Services Requirement is the innovative *High Performance Science Cloud*



Partnership between the CCE, CISTO, and NCCS

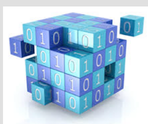
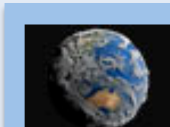
- Provide compute, storage, data management, and data publication for the ABoVE campaign using the HPSC
- Reduces technical overhead for ABoVE scientists
- Allows scientists to focus on science in a optimized computing environment

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

*** Repurposes NCCS HPC Discover Compute for routine growth*

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 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.	Storage nodes configured with a total of about 3 PB of RAW storage capacity

Software

- Operating Systems
  - Debian Linux
  - CentOS Linux
  - Windows (4 licenses)
- Commercial Tools
  - Intel Compilers (C, C++, FORTRAN)
  - IDL (4 licenses)
- Open Source Tools
  - Python
  - NetCDF
  - Almost anything you need
- How to install software?
  - Users can install software in your home and nobackup directories
  - Users can request support to help install and configure software on your virtual machines
- Can reach out to external license servers
  - We can do this and make external software licenses available to your virtual machines; takes time to poke holes through firewalls

```
#!/usr/bin/env python
'''
Created on 03 June 2014

@author: dqduffy
'''
import sys
from CDSLibrary 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__':

        # exercise all canonical operations
        print(above_job_epoch1_aveT)
        cds_lib.avg(service, above_job_epoch1_aveT, file_job_epoch1_aveT)
GSSLA40029712:python dqduffy$
```

Data Currently In the Cloud

Data Set	Amount	Path	Environment Variable
Landsat – Surface reflectance	123 TB	/att/hdvault/LANDSAT	$LANDSAT
MODIS – Daily 500m surface reflectance	57 TB	/att/nobackup/MERRA	$MODIS
MERRA – GEOS-5 reanalysis	192 TB	/att/nobackup/MERRA	$MERRA
Total	**372 TB**		

NGA Data

- NGA data is available only on specific virtual machines
- Contact Mark McInerney and the Climate Model Data Services for access

Additional data sets to be moved in the cloud will be dependent on the funded science research.
Ownership and management of these large data sets will be a shard responsibility between the system (NCCS) and the project (ABoVE).

User Requirements Table
What we will ask you before you get access

Category	Options	Requirement
Operating System	Debian Linux, Centos Linux, Windows	
Number of Cores per Virtual Machine	1 to 20	
Amount of Memory per Virtual Machine	24 GB to 64 GB	
Number of Virtual Machines	1 to 48	
Data Sets	LANDSAT, MERRA, MODIS, NGA, Other	
Data Storage	Depends on request	
Software	Compilers (gnu, Intel), IDL, python, NetCDF, perl, etc.	

What works best in the cloud?

- Not designed for MPI (use HPC systems for that)
  - Highly coupled processes performing large amounts of message passing
- Designed more for inherently parallel processing of big data
  - Independent processes written to analyze large data sets
  - We have 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 (Mark) supports this

Not an Archive

- The system is not an archive
  - Users will have to adhere to their data management plan
- Large data stored in the system will be stored for the life of the project
  - ABoVE related data, such as Landsat and NGA, will be there for a long time
  - Other data may only be there for months and can be deleted after the project is completed
    - Dependent on the project and research
    - Will work with the ABoVE team and the researchers to decide how best to do this
- Where do you archive your results?
  - Outside the cloud, NASA DAACs, other

Moving Large Data

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

How Much Storage Can You Use

- Home Directories
  - Quota of 15 GB
  - Replicated within the system, but not backed up
  - User code, applications, etc.
- Scratch Space
  - Quota of 2.5 TB
  - Not backed up
  - Temporary files, larger scratch space
  - Not a long term archive
- Large Data Sets
  - Shared storage environment (petabyte) for large datasets to be shared across the ABoVE research

The Future

- More capabilities
  - More compute nodes
- More service offerings to come on line
  - Remote Visualization
  - Database
  - Remote Desktop (maybe)
  - More Windows licenses
- Self Service Portal
  - Users will be able to launch their own virtual machines

The Team – The people that do the real work!

- System Engineers/System Administrators
  - Hoot Thompson
  - Scott Sinno
  - Garrison Vaughan
  - Ramon Rameriz-Linan
  - Brittney Wills

- NCCS Help Desk Contact Information
  - support@nccs.nasa.gov
  - 301-286-9120