

A sunset over a body of water with clouds reflected in the water. The sun is low on the horizon, creating a bright glow and long shadows. The sky is filled with scattered clouds, and the water below mirrors the scene.

# How to Cloud for Earth Scientists: The ABoVE Science Cloud on ADAPT

Peter Griffith (SSAI) and Elizabeth Hoy (GST, Inc.)  
NASA Carbon Cycle and Ecosystems Office

# Contributors

- Dan Duffy, NCCS High Performance Computing Lead
- Scott Sinno, System Architect and System Administrator
- Hoot Thompson, System Architect and System Administrator
- Garrison Vaughn, System Architect and Applications Engineer
- Ellen Salmon, Computer Research and Development
- Laura Carriere, System Analyst
- Julien Peters, Software Developer
- Others.....

# ABoVE is a large-scale study of environmental change in Arctic and boreal regions and the implications for ecological systems and society

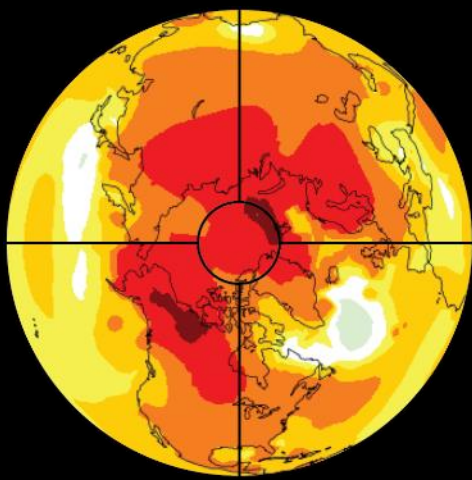
## Overarching Science Question:

*How vulnerable or resilient are ecosystems and society to environmental change in the Arctic and boreal region of western North America?*





# Resilience Framework



Causes of Change



Changes to Ecosystems

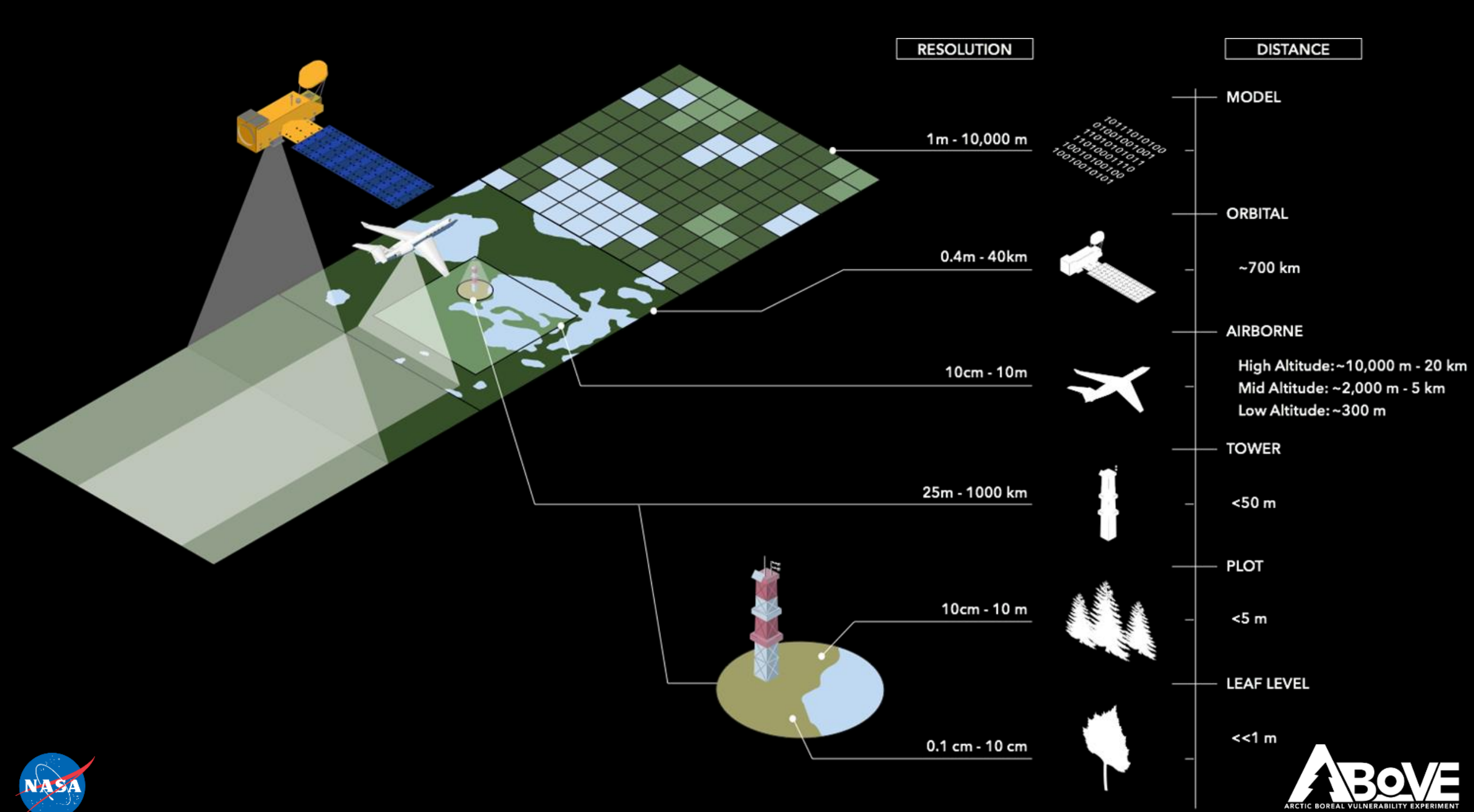


Social Systems

Ecosystem Services

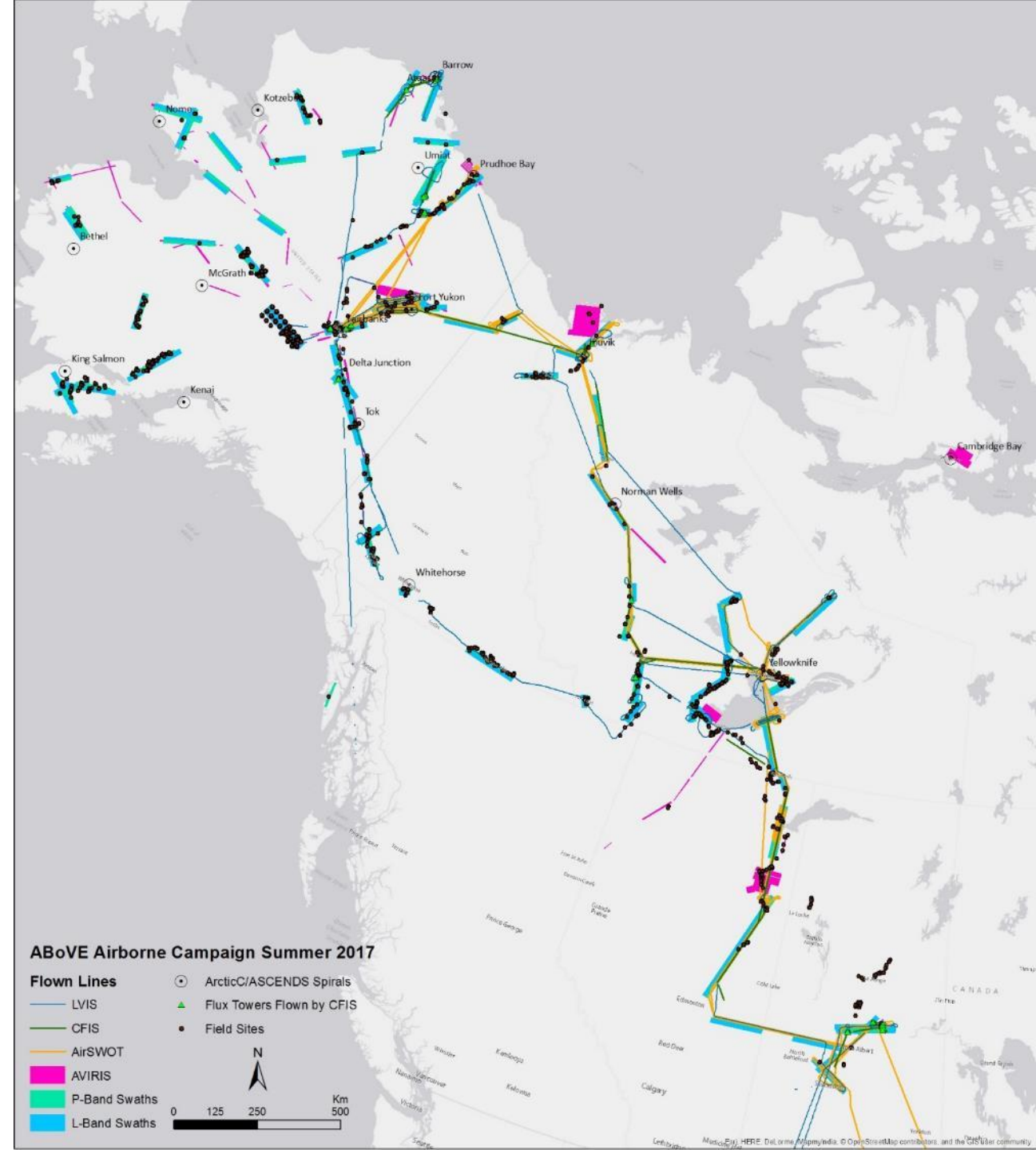


A3



# Where are we working?

- 80 total projects (including NASA funded, Partner, Affiliated)
- 550 participants from universities, national agencies/labs, state/provincial/territorial groups, private, and native/aboriginal organizations
- Summer airborne campaign:
  - 10 aircraft, 20 deployments, and 200 science flights
  - April to October 2017
  - 4 million km<sup>2</sup> in Alaska and Canada





# Why do we need a new approach?

- Science datasets are becoming larger, with intensive computation needed for data processing
- And collaboration across diverse research groups is essential,
- But it is often time consuming and expensive to transfer, download, process and share data with others
- Therefore the ABoVE Science Cloud (ASC) was created to meet the needs of ABoVE investigators and encourage collaboration within the field campaign.

# NASA Center for Climate Simulation (NCCS)

Integrated high-end computing environment supporting the specialized requirements of Climate and Weather modeling.

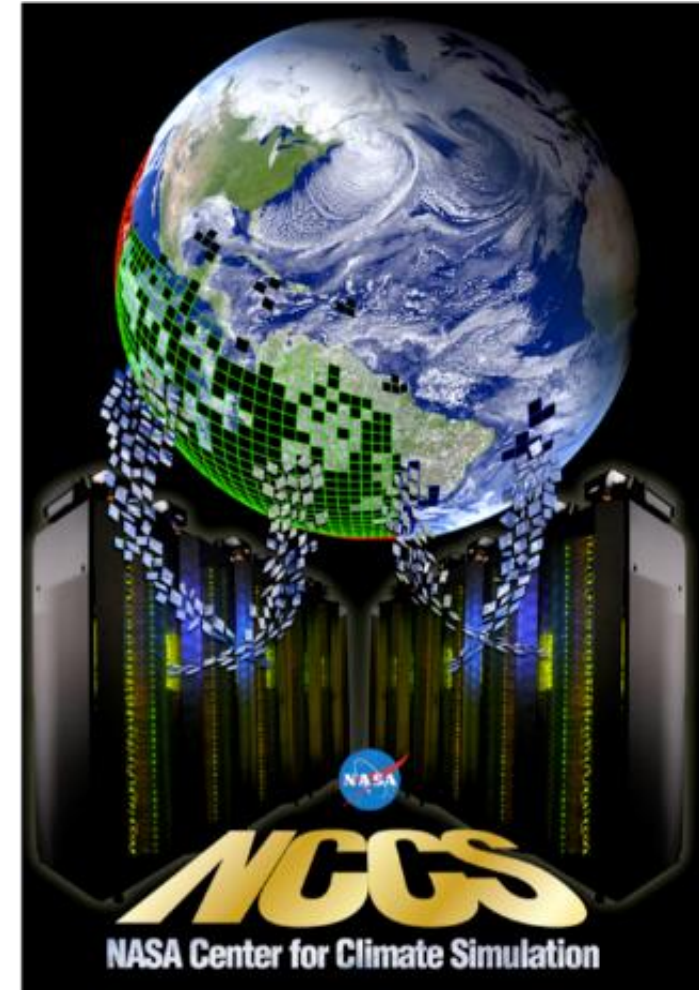
- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- Advanced Data Analytics Platform (ADAPT)

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
- Dan Duffy, High Performance Computing Lead (Code 606.2)





# Analysis is Different than HPC

## High Performance Computing

*Takes in small amounts of input and creates large amounts of output...*

Relatively small amount of observation data, models generate forecasts

Tightly coupled processes require 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...*

Large amounts of observational/model data 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

# 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. A portion of ADAPT is dedicated to ABoVE (the ABoVE Science Cloud).

### Adjunct to the NCCS HPC environment

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

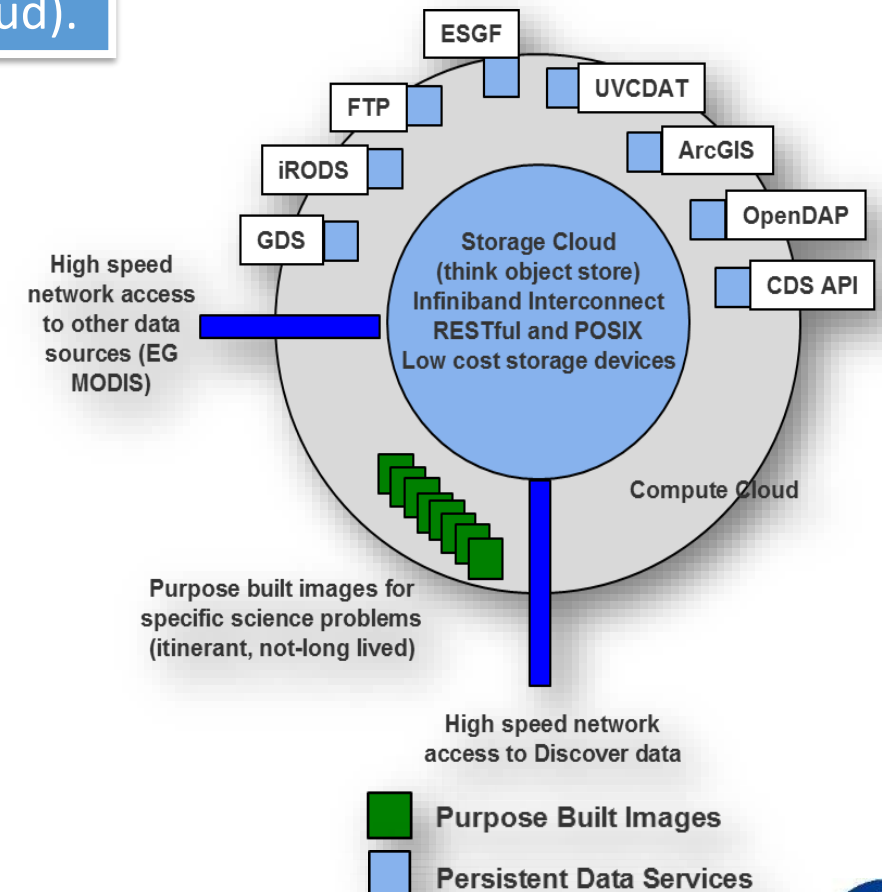
### Expanded customer base

- 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
- Critical Node-to-node communication – high speed, low latency
- Shared, high performance file system
- Management and rapid provisioning of resources

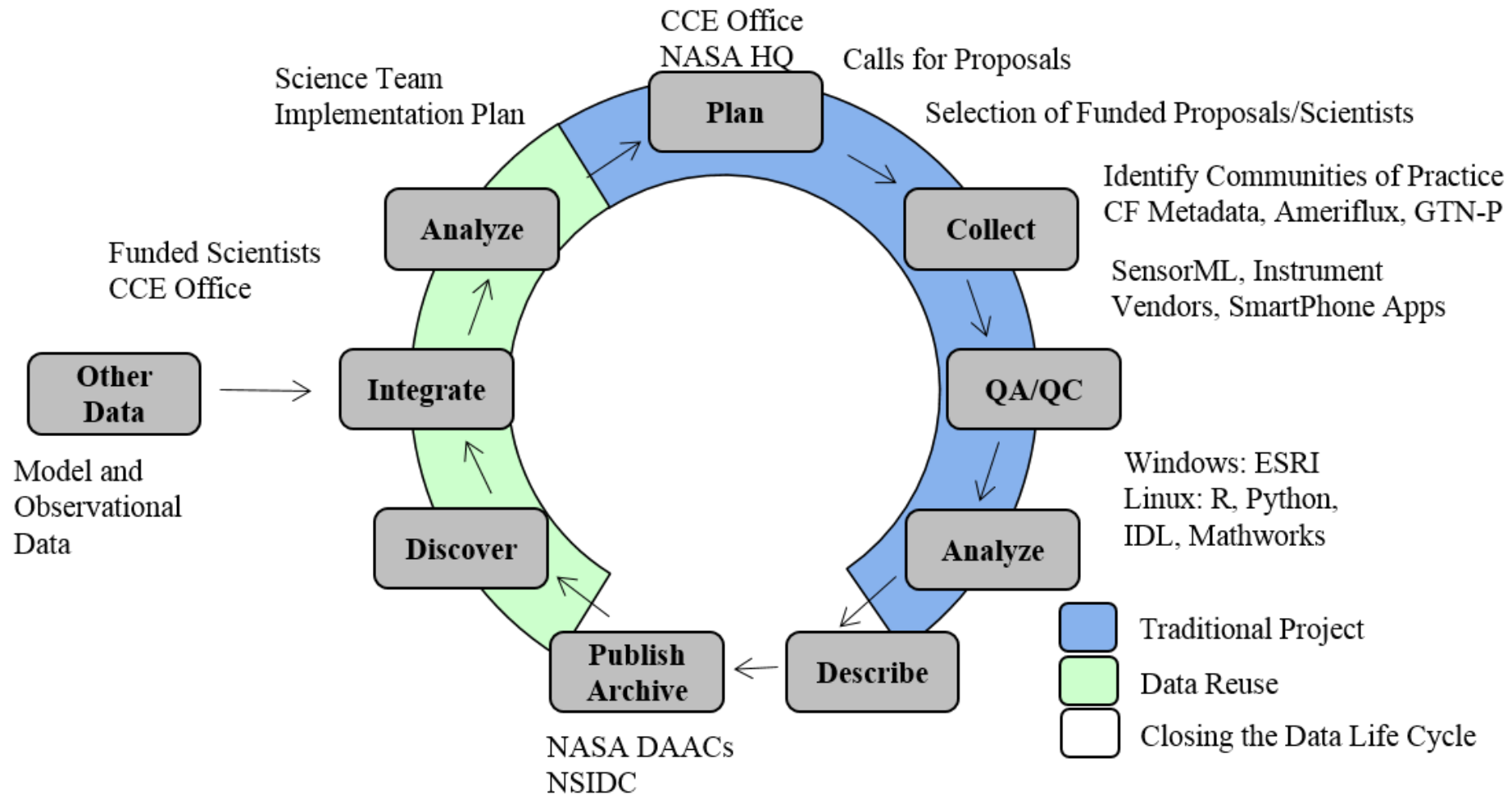
### Conceptual Architecture



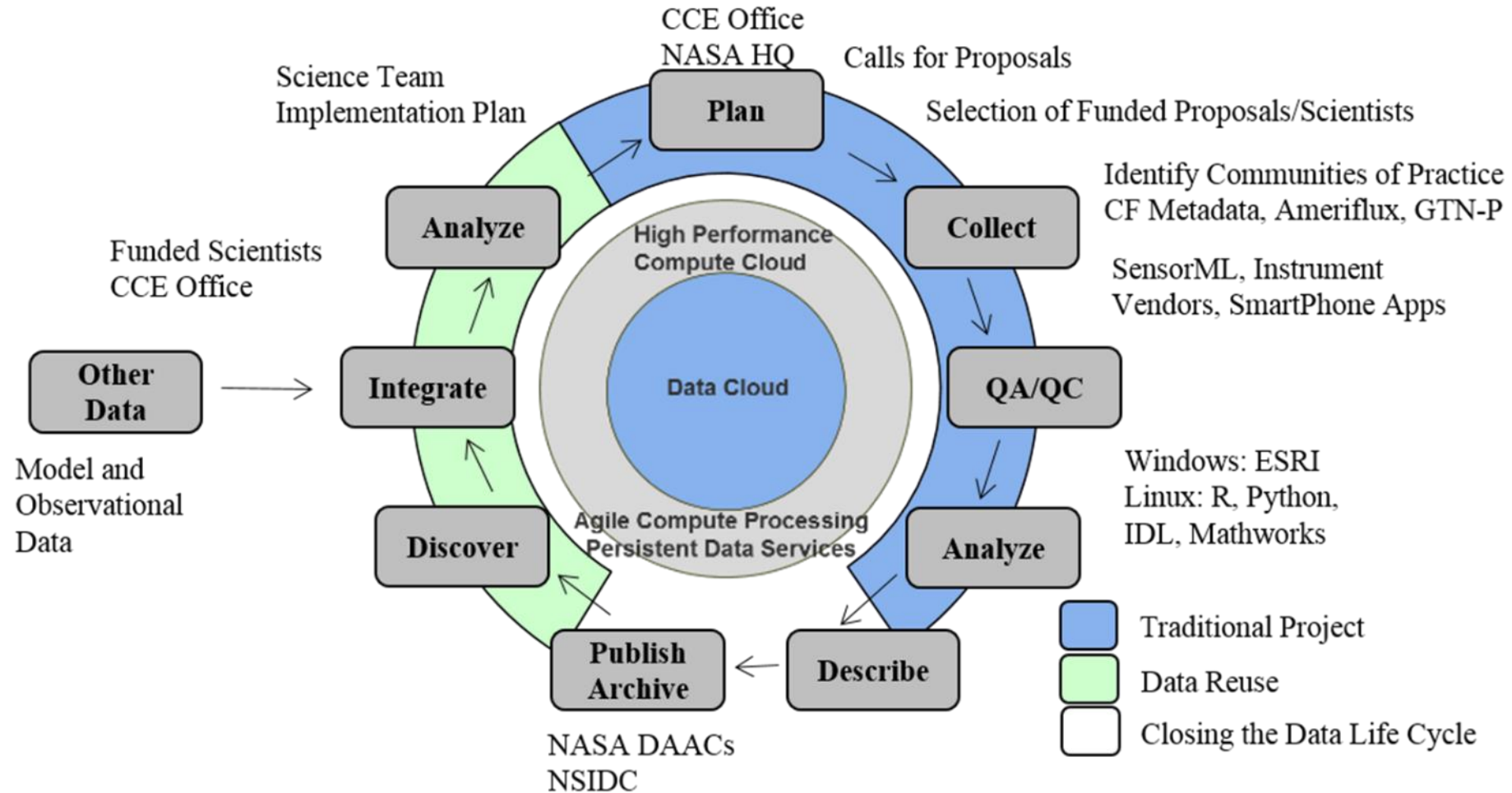









# The ABoVE Science Cloud is the center of the data lifecycle.



# The ABoVE Science Cloud is the center of the data lifecycle.



# System Components/Configuration

Capability and Description	Configuration
 <p><b>Persistent Data Services</b> Virtual machines or containers deployed for web services, examples include ESGF, GDS, THREDDS, FTP, etc.</p>	Nodes with 128 GB of RAM, 10 GbE, and FDR IB
 <p><b>DataBase</b> High available database nodes with solid state disk.</p>	Nodes with 128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB
 <p><b>Remote Visualization - planned</b> Enable server side graphical processing and rendering of data.</p>	Nodes with 128 GB of RAM, 10 GbE, FDR IB, and GPUs
 <p><b>High Performance Compute</b> More than 1,000 cores coupled via high speed Infiniband networks for elastic or itinerant computing requirements.</p>	~100 nodes with between 24 and 64 GB of RAM and FDR IB
 <p><b>High-Speed/High-Capacity Storage</b> Petabytes of storage accessible to all the above capabilities over the high speed Infiniband network.</p>	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.

Open Source Tools  
Python, NetCDF, etc.



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



Operating Systems  
Linux (Debian, CentOS)  
and Windows



# 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 Download Times For 80TB

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

1 Gbps NASA / Other Gov

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

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

# ABOVE Science Cloud Data Holdings

Large Collections	Amount
Landsat	186 TB
MODIS	MODAPS collection remotely mounted
MERRA & MERRA2	406 TB
DigitalGlobe Imagery	2.8 PB
Total	> 3 PB

## Other Data Sets

- Elevation datasets: ArcticDEM, CDEM, ASTER GDEM, etc.
- Vegetation products
- Land cover products
- Products generated by the science team

*\*Others as the team requests...*

Find a list of all common datasets available on the ASC here [>>](#)



## NGA/DigitalGlobe High Resolution Commercial Satellite Imagery

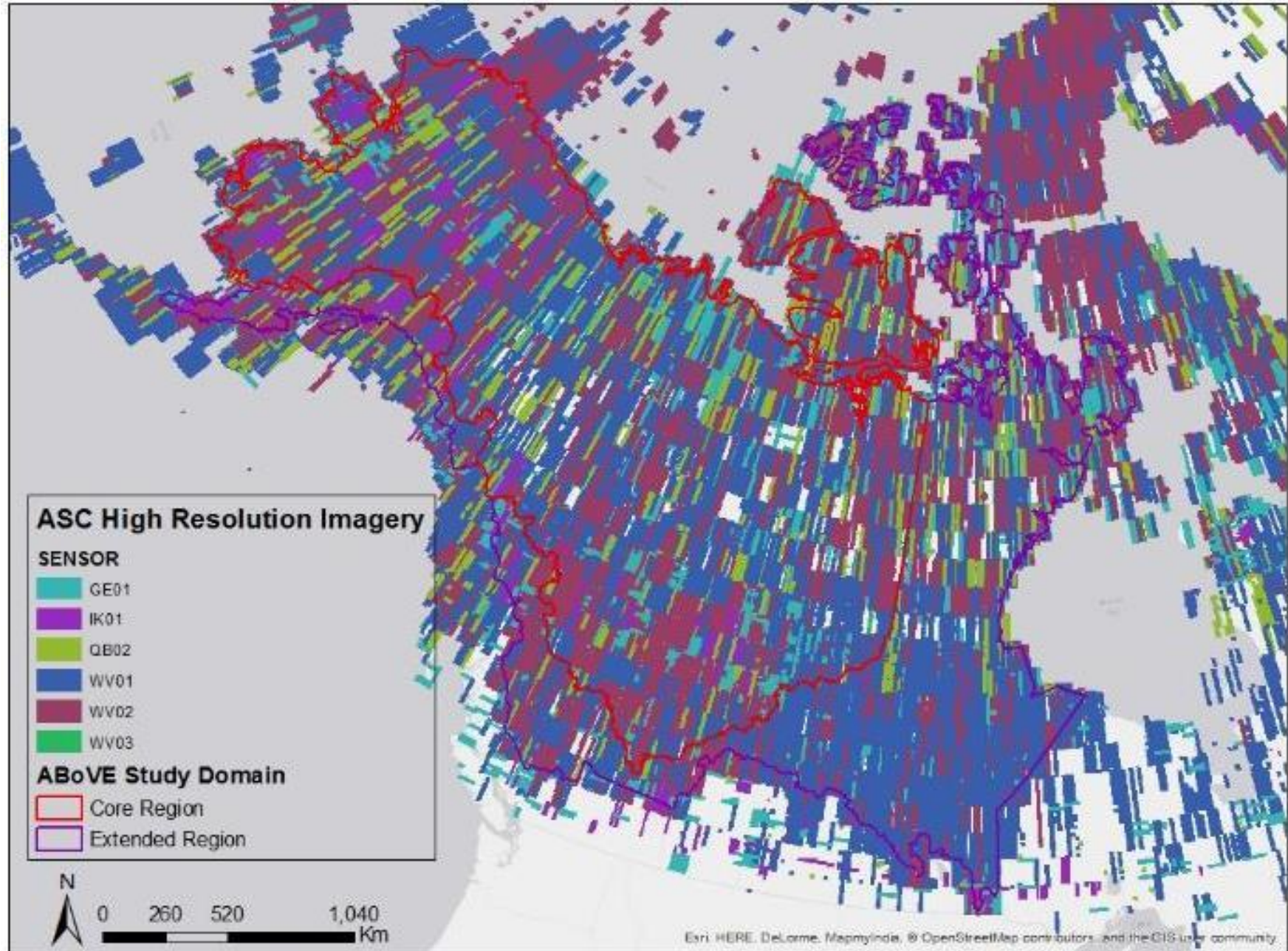
**National Geospatial Intelligence Agency (NGA) has licensed all DigitalGlobe  $\geq 31$  cm satellite imagery for US Federal use, i.e., NSF, NASA and NASA funded projects.**

- Archive of  $>4.2$  billion  $\text{km}^2$  of data from 2000 to present
- Data from six different satellites: Worldview-1, 2 and 3; Ikonos; Quickbird; and Geoeye-1

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

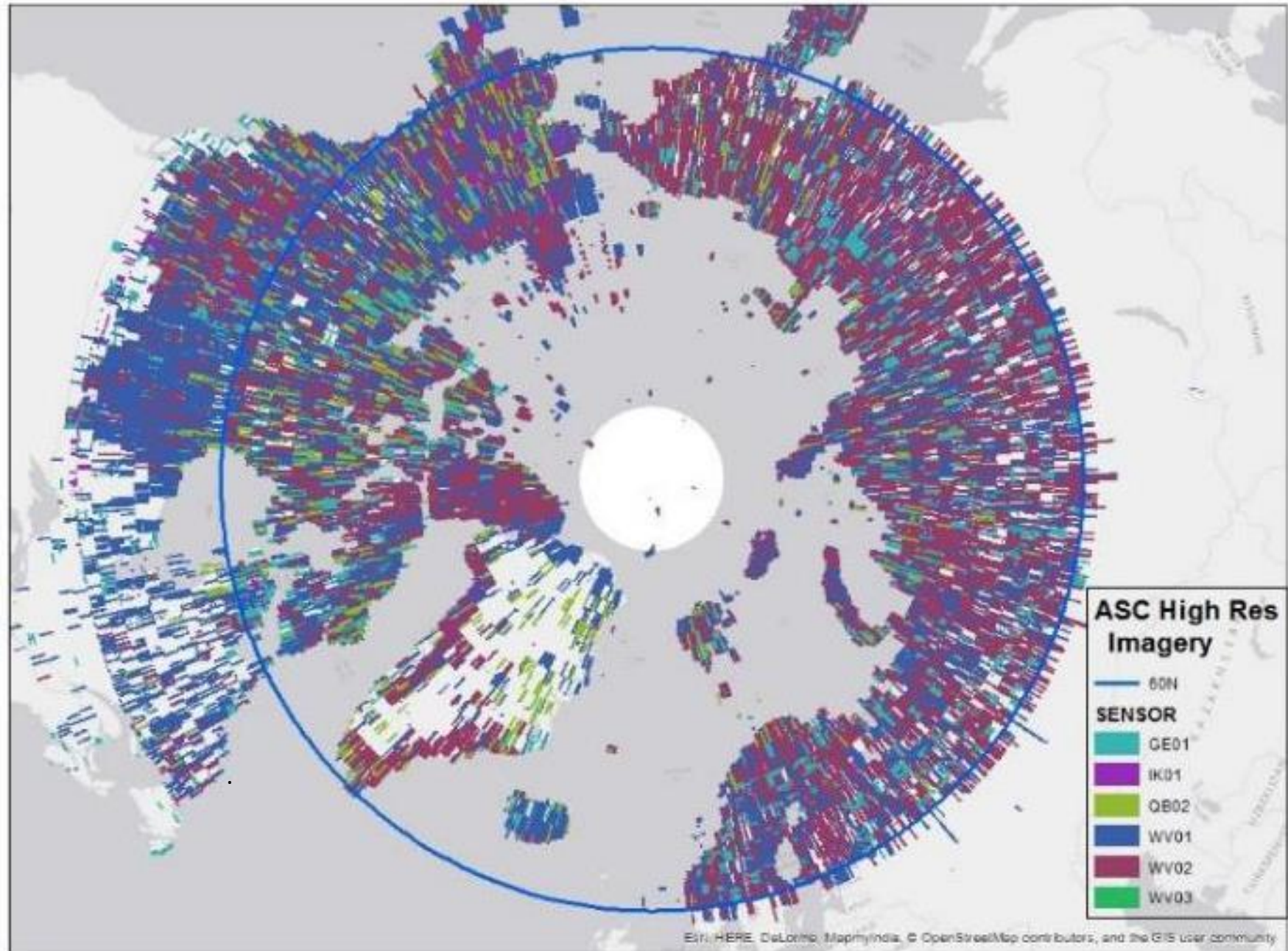


# ABOVE Science Cloud DigitalGlobe Imagery: Study Domain





# ABOVE Science Cloud DigitalGlobe Imagery: Circumpolar

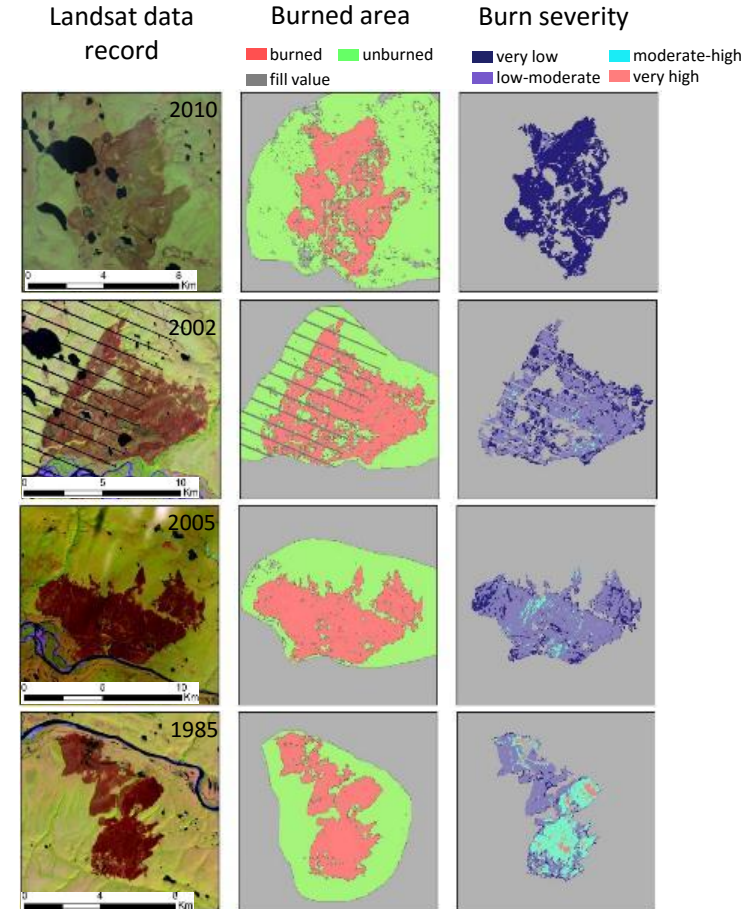




# Examples of the ASC In Action

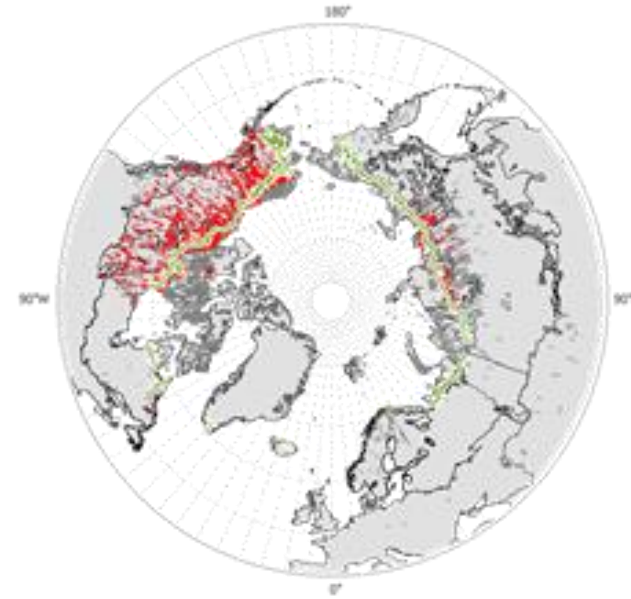
# Fire History for ABoVE – T. Loboda & M. Miller

- Fire history across the ABoVE study region is compiled from available and new (Miller et al. in prep) data products and enhanced
- Multiple VMs on the ASC are used to process Landsat and MODIS data to develop the burn severity characterization



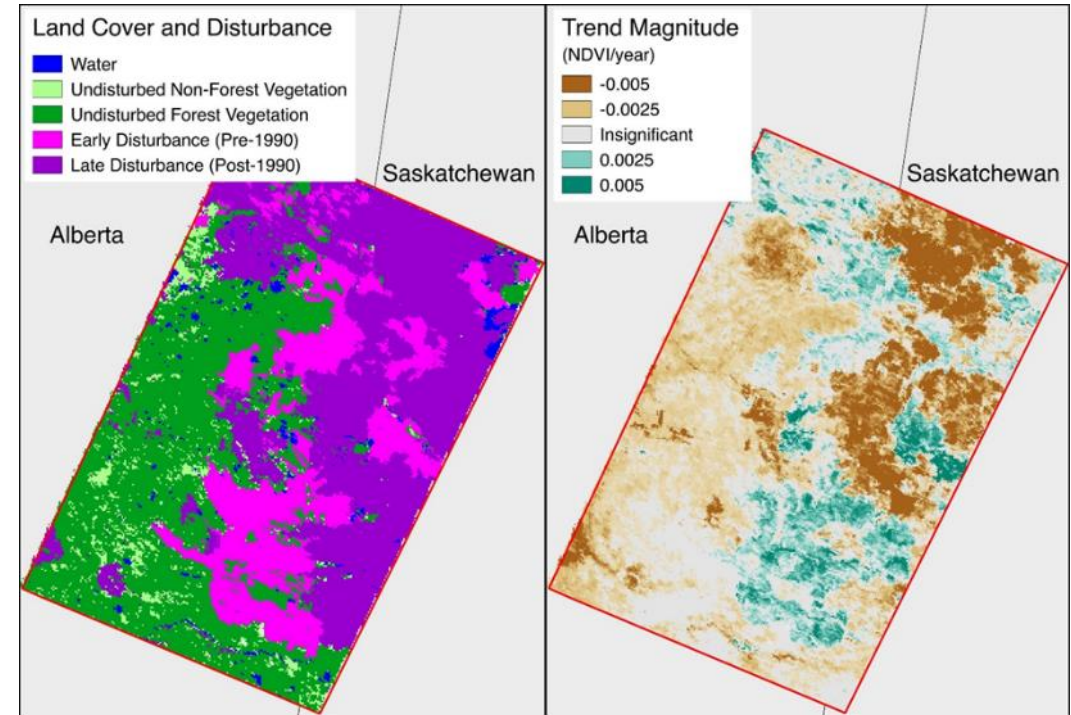
# Forest Canopy Surface Elevations – C. Neigh & P. Montesano

- Understanding forest patterns using DigitalGlobe high-resolution satellite imagery
- Using multiple VMs and Ames Stereo Pipeline (ASP) on the ASC to process Digital Elevation Models



# Landscape-Scale Histories of Disturbance, Seasonality and Greenness Trends - C. Woodcock & D. Sulla-Menashe

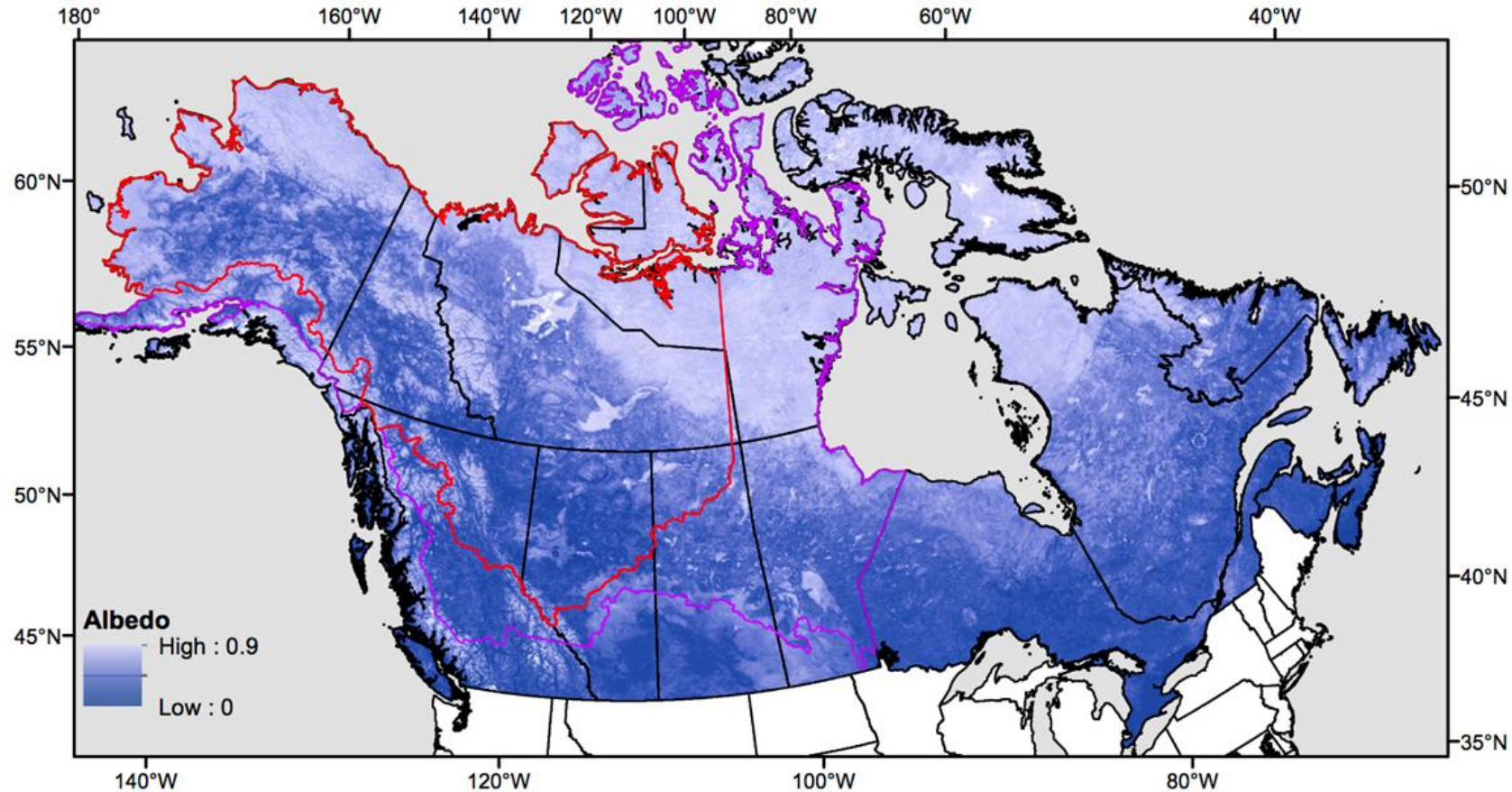
- 30+ year historical record and ongoing characterization of disturbance events and phenology across the ABoVE study domain
- Using multiple VMs to move Landsat data into the ABoVE grid and then develop the landscape histories





# Understanding the Causes and Implications of Enhanced Seasonal CO<sub>2</sub> Exchange in Boreal and Arctic Ecosystems – B. Rogers

- Modeling driving factors of post-fire albedo trajectories
- Creation of mean albedo maps
- Fire combustion mapping





# ABOVE

ARCTIC BOREAL VULNERABILITY EXPERIMENT

<http://above.nasa.gov/sciencecloud.html?>

[Sign In](#) | [My Account](#) | [Sign Out](#)

[Home](#)

[About](#)

[Science Team](#)

[Meetings & Events](#)

[Publications](#)

[Data](#)

[Safety & Logistics](#)

[Funding](#)

[Jobs](#)

## The ABoVE Science Cloud (ASC)

Referenced on page A.4-8 in NASA Research Announcement for Terrestrial Ecology: Airborne Campaign For ABoVE [NNH16ZDA001N-TE](#)

[Science Cloud Setup Instructions](#)

[About the Science Cloud](#)

[Webinar](#)

The NASA Center for Climate Simulation (NCCS) has partnered with the NASA Carbon Cycle and Ecosystems Office (CCE Office) to create a high performance science cloud for this field campaign. 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

# Seeing the ABoVE Science Cloud in Action

## Presentation by Mark Carroll



[Above.nasa.gov](http://Above.nasa.gov) @NASA\_ABoVE



# Transitioning from workstation to cloud computing

Mark Carroll

Biospheric Sciences Lab

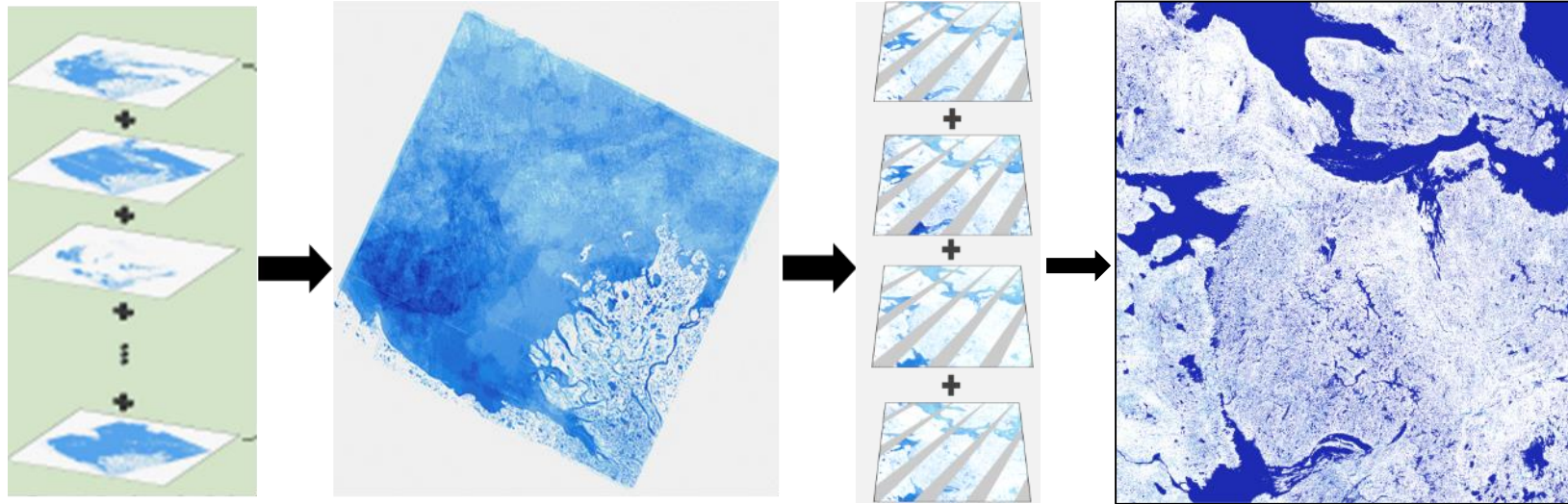
Science Systems and Applications Inc.



# Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign

- Utilize the dense time series of Landsat data in the North American Arctic to create a time series of surface water maps
- We use the full available time series to minimize the impact of anomalous weather events (drought, flood) in individual scenes
- Maps will represent surface water extent for 3 epochs 1990 – 1992, 2000 – 2002, and 2010 – 2012
- These maps can be used to identify hotspots of change and to identify field sites for study during the ABoVE campaign

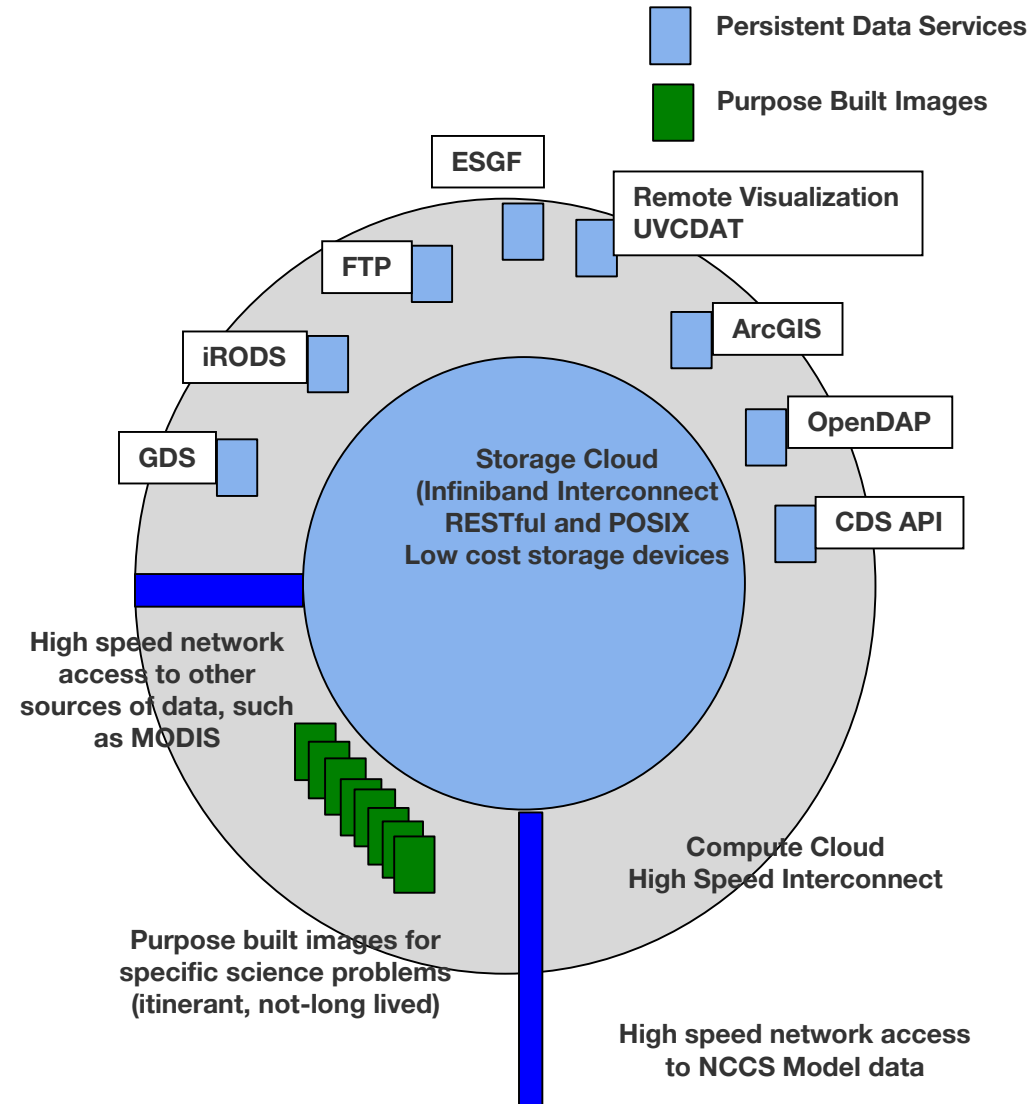
# Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign



- Decision tree classification on each Landsat scene
- Extract theme (water) from each date, build data stack
- Sum water observations for entire epoch
- Mosaic each themed scene into ABoVE tile (no overlap)
- Sum mosaicked tiles to create a total per theme for each ABoVE tile (utilizes all available observations including overlap)

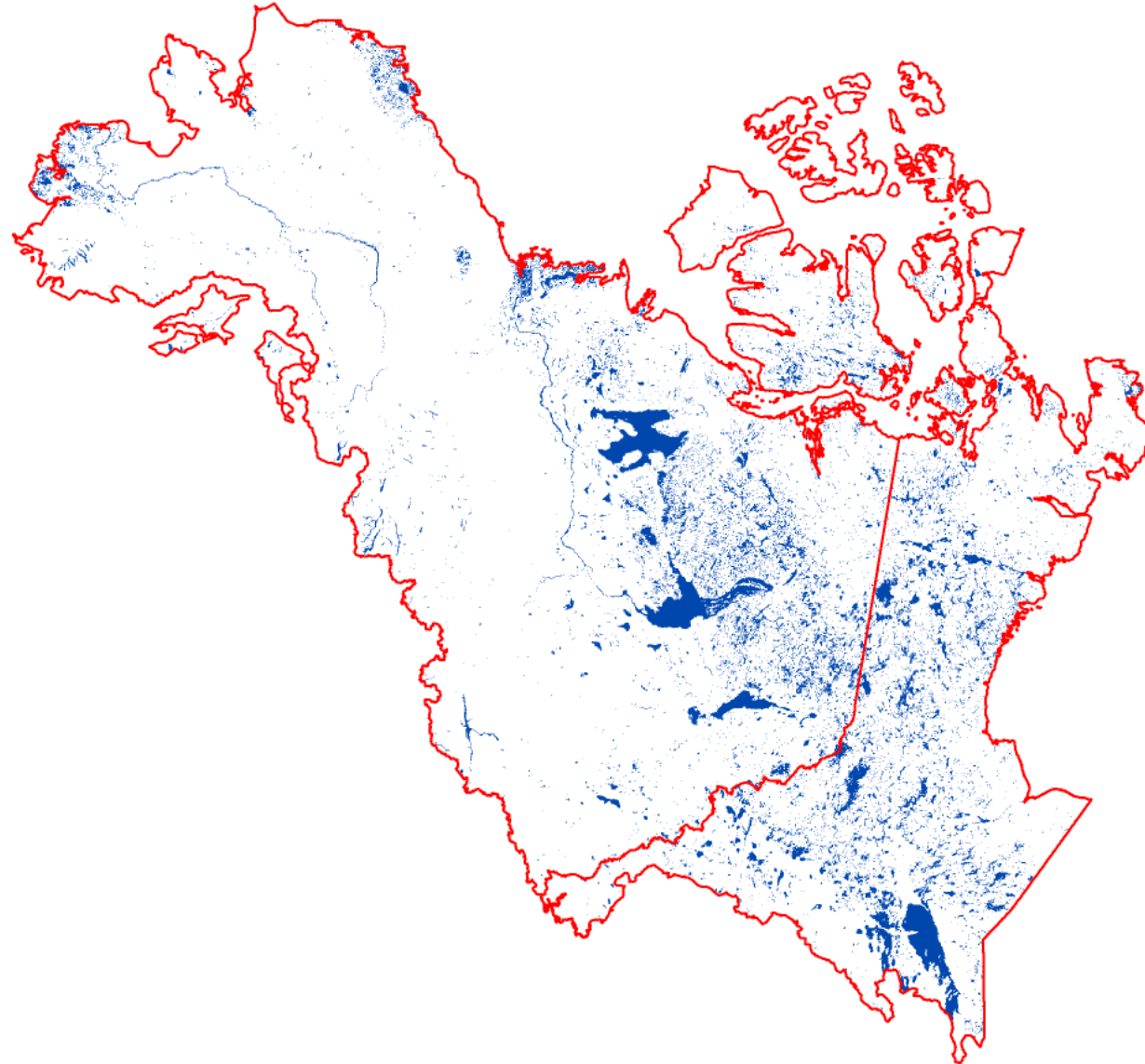
# Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign

- Original processing plan involved a couple of workstations and rotating data through an 8TB RAID
- Anticipated processing time 9 – 12 months
- Only final outputs would be kept online
- No time available for reprocessing
- Enter the Science Cloud at NCCS and GSFC High Performance Computing



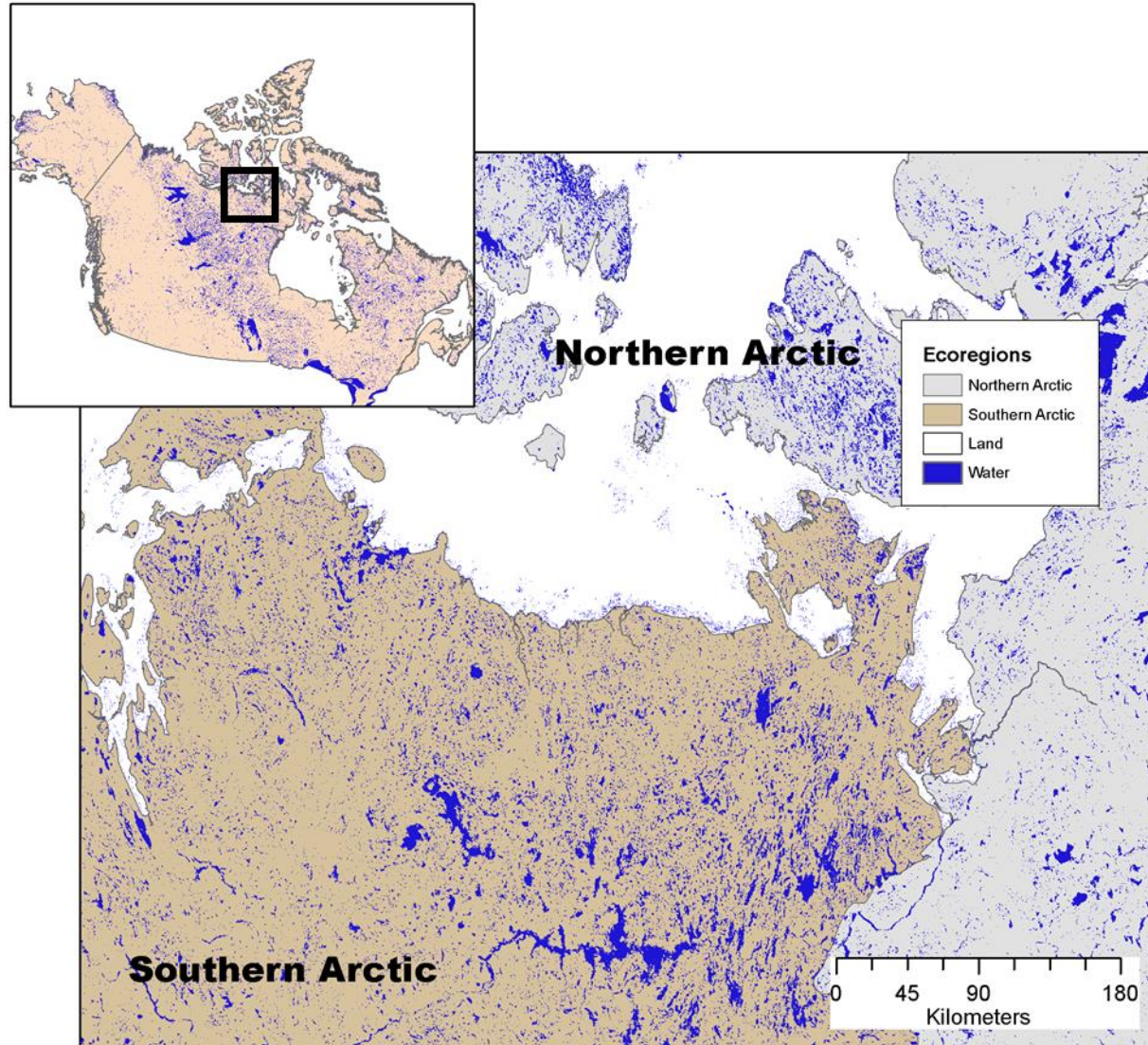
# Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign

- Final result is a time series of three maps 10 years apart that can be used to show not only the location of water at a given time period but also the change in surface water extent through time.



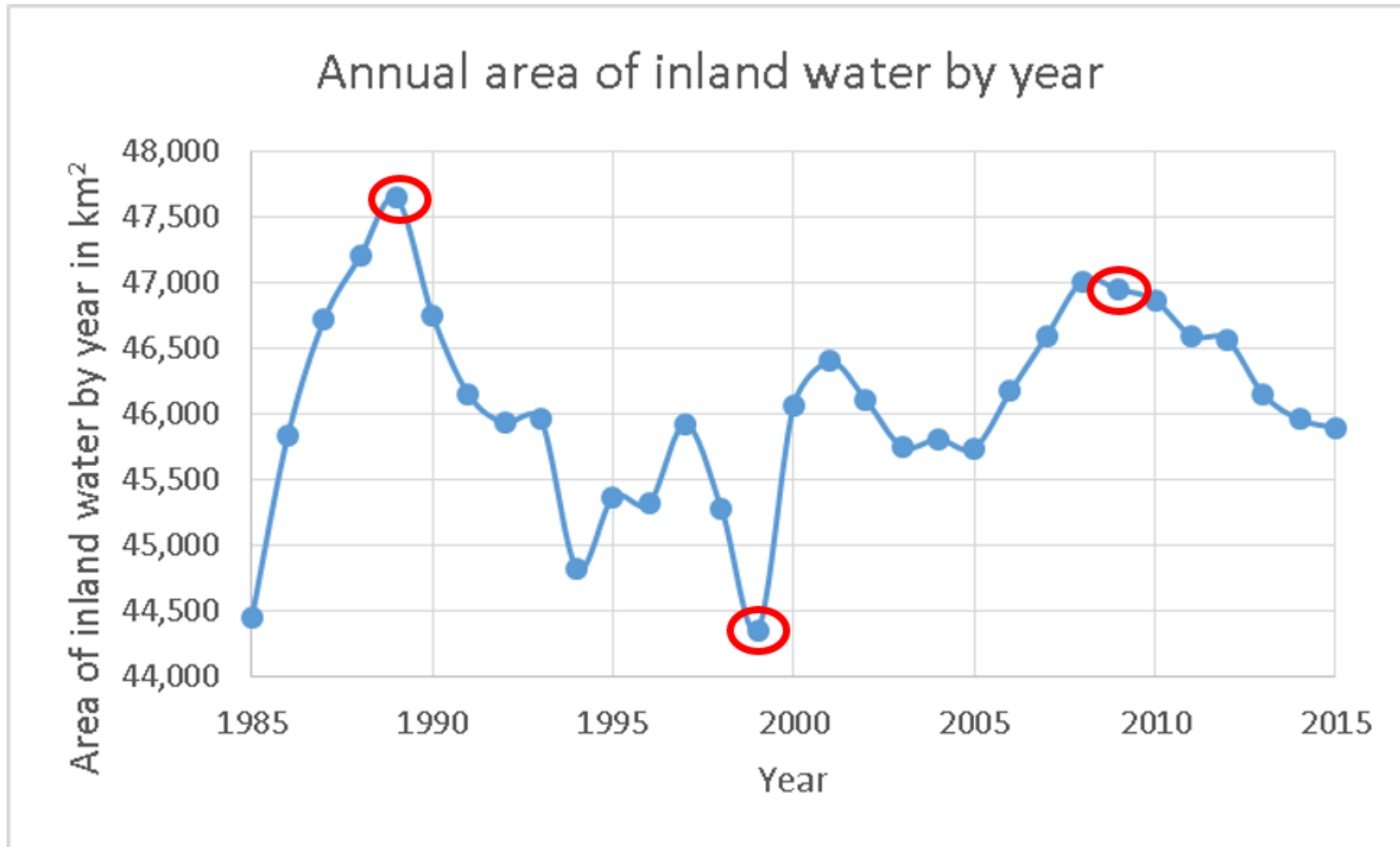


# Regional annual water



- Northern Nunavut province in Canada
- Includes Queen Maud Gulf Bird Sanctuary
- Limited impact from anthropogenic pressures

# Regional annual water



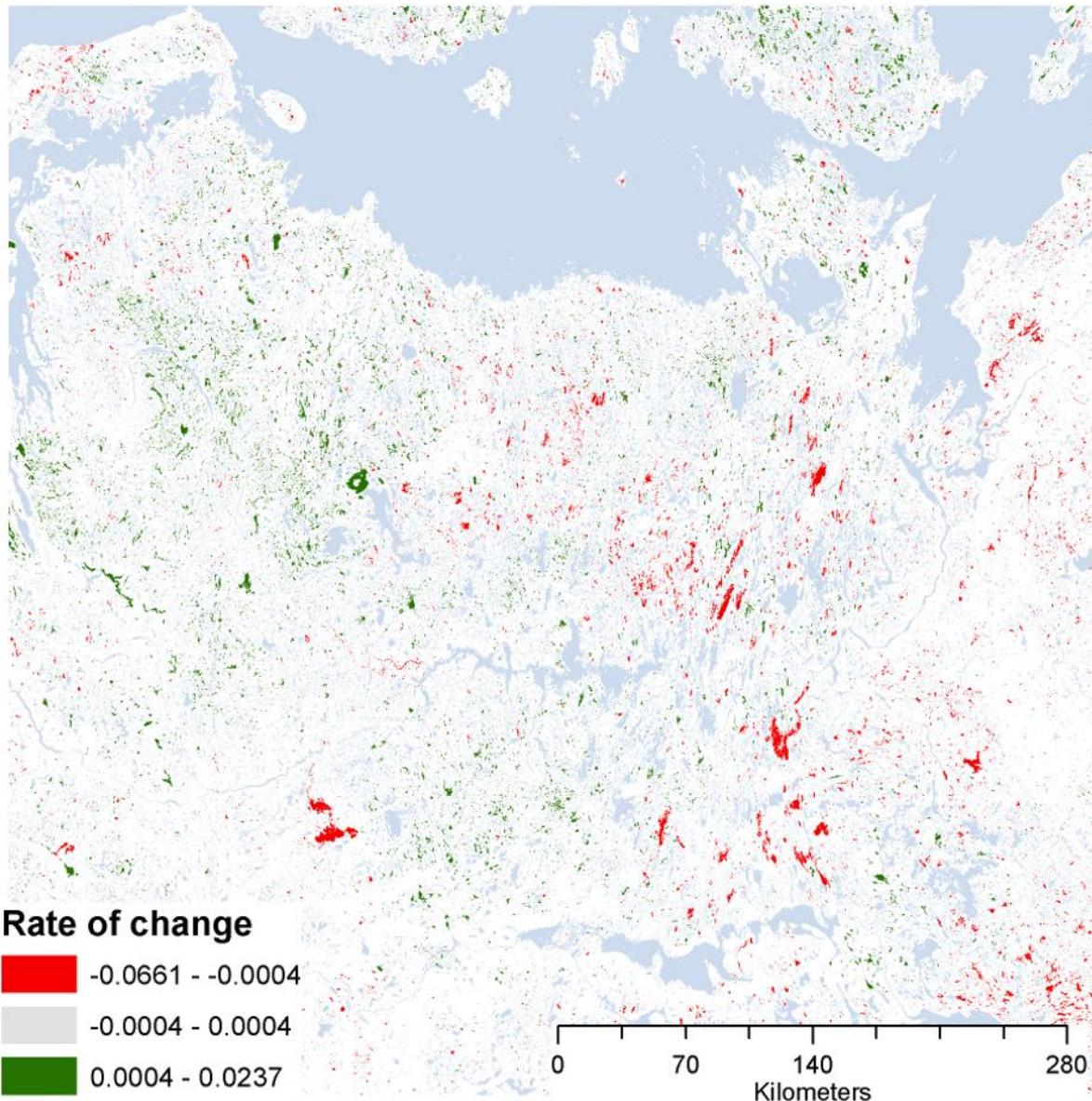
# Regional annual water

size in ha	<0.1	0.1 to 1	1 to 10	10 to 100	100 to 1,000	1,000 to 10,000	10,000 to 100,000	>100,000
total	251,884	202,412	167,450	48,495	4,836	257	29	9
change p0.05	52,475	55,081	45,724	13,330	1,369	62	4	2
fraction(total)	21%	27%	27%	27%	28%	24%	14%	22%
decreasing	31,810	21,438	17,216	4,960	527	32	3	2
increasing	20,665	33,643	28,508	8,370	842	30	1	0
d fraction	61%	39%	38%	37%	38%	52%	75%	100%
l fraction	39%	61%	62%	63%	62%	48%	25%	0%

- Over 60% of water bodies are < 1 ha
- Smallest and largest water bodies decreasing in size
- Middle sized water bodies are increasing



# Regional annual water

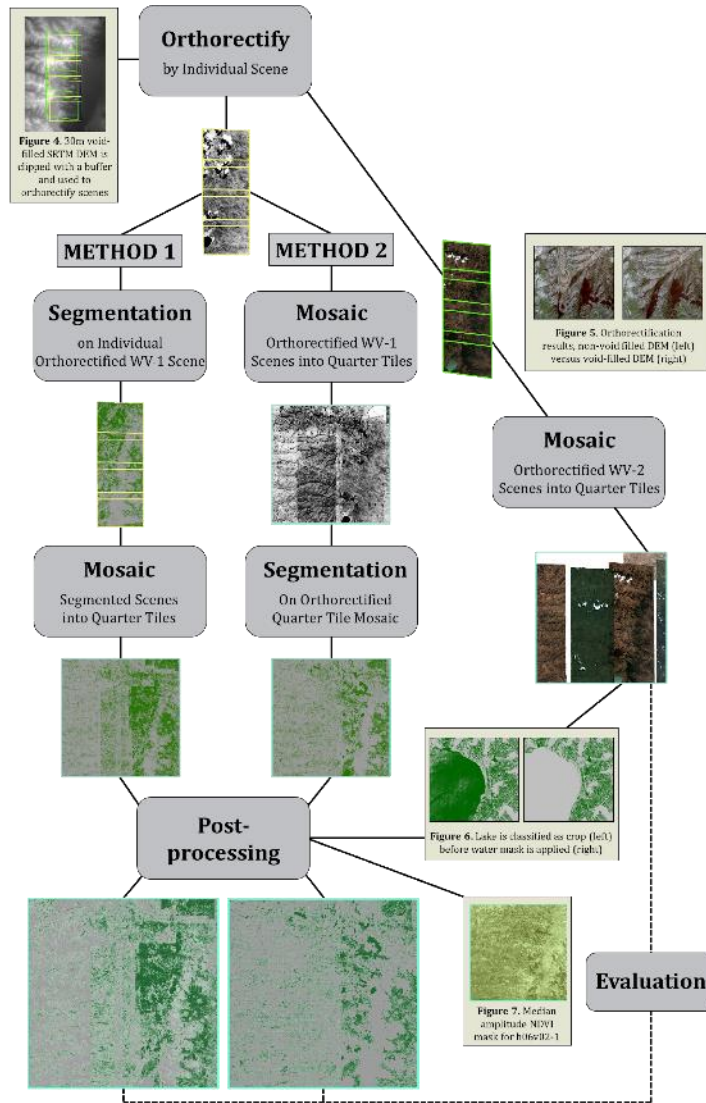


- Over 675,000 water bodies in the study area
- Linear regression (OLS) performed on area per water body per year
- Over 168,000 water bodies exhibit change with  $p < 0.05$



# Crop Mapping Tigray, Ethiopia

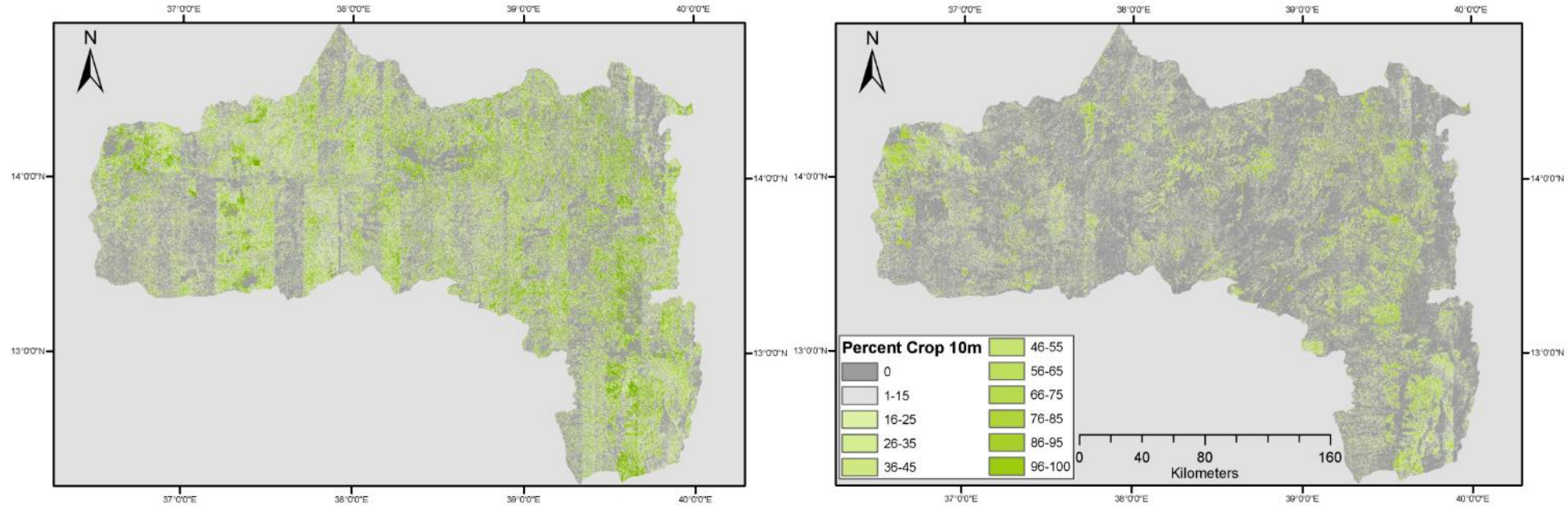
## Tigray crop mapping



- Initial plan was to use very high resolution data to train a classification algorithm
- Given data and processing capability in the cloud we transitioned to using VHR to generate the entire map
- Methodology began in linear fashion treating scenes individually
- Experience showed that generating mosaics prior to classification improved results
- Along the way a new approach to evaluation was created

# Crop Mapping Tigray, Ethiopia

## Method Evolution



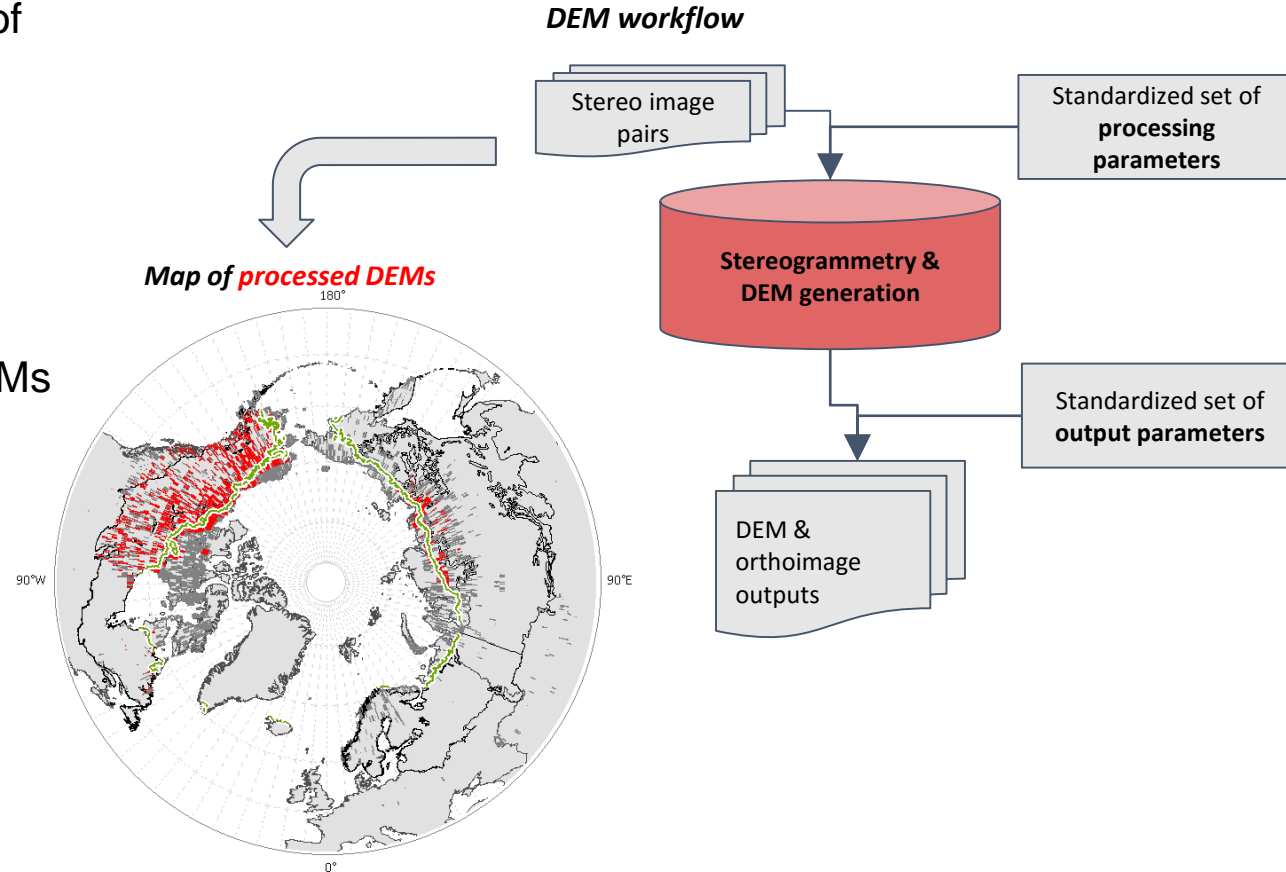
- Image on left shows original method using VHR data that resulted in distinct boundaries between scenes
- Image on right shows the improved method with far fewer errors and hard boundaries between scenes
- Working in the cloud enabled rapid reprocessing and assembly of thousands of scenes in just a few days
  - There are ~85 billion pixels in the mosaicked maps
  - Nearly 500 billion pixels were processed to get enough information to generate these mosaics
  - End to end processing can be completed in just a few days

# DEM Workflow: standardize & optimize in HEC

Paul Montesano and Chris Neigh

**Standardize** the processing of the image pairs returned from data discovery:

- with tested parameters
- incorporate lessons learned from **3000+** DEMs processed using **6000+** image strip pairs



# DEM Workflow: produce science-ready data

Reduce high start-up costs associated with working with massive datasets on HEC platforms

Extend science opportunities to other PIs

Our automated DEM workflow will streamline the (1) ingest & (2) processing of stereopairs, and (3) output of **science-ready DEMs and orthoimages.**

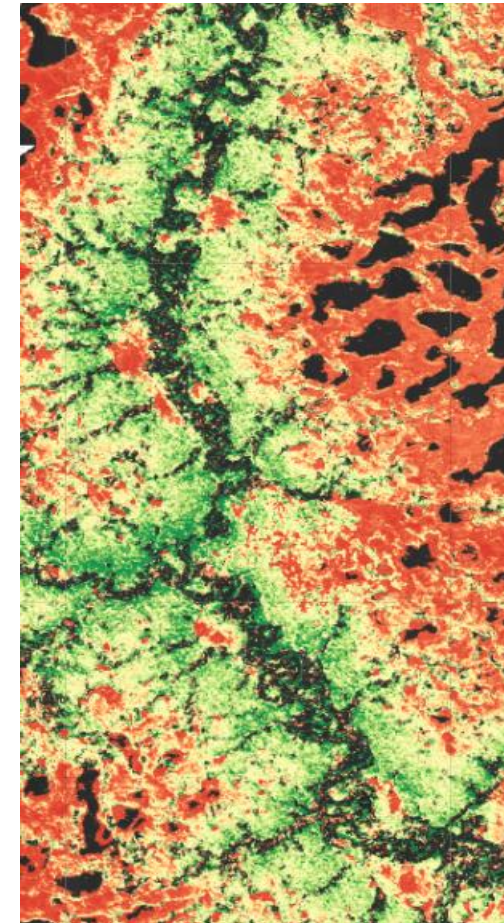
Vegetation Height (m)

10



0

1 km



*Example of a science product derived from an analysis of DSMs processed with a preliminary workflow. Vegetation height in an open-canopy boreal forest of western Siberian.*

*Montesano, Neigh et al. RSE 2017*



# Automated protocols for generating very high-resolution commercial validation products with NASA HEC resources

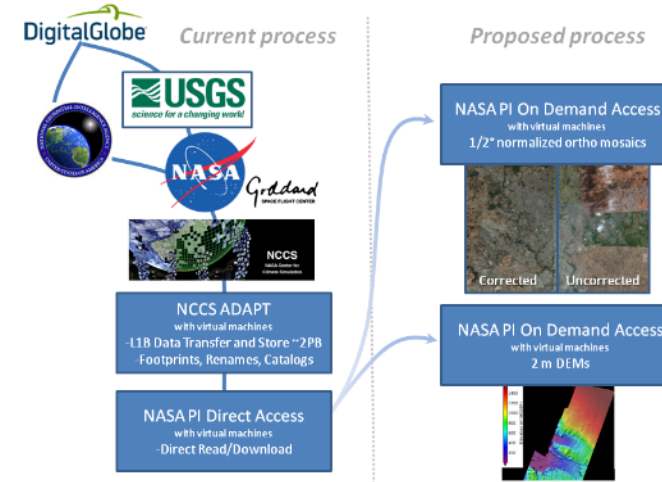
PI: Chris Neigh, NASA GSFC

## Goals and Objectives

Enhance scientific utility of sub-meter DigitalGlobe data by:

- 1) Improving VHR data discovery: using databases and mosaic datasets within NASA-GSFC's ADAPT global archive of DG VHR imagery;
- 2) Producing on demand VHR 1/2° degree mosaics: automating estimates of surface reflectance, ortho-rectifying and normalized 1 m mosaics for pan and 2 m for multi-spectral; and
- 3) Producing on demand 2 m posting DEMs: leveraging HEC processing and open source NASA-ARC ASP software.

## Architecture Overview



## Approach

Develop a HEC API to:

1. identify NASA-GSFC archived VHR DG data and Ortho-rectify, atmospherically correct, identify clouds/shadows, mosaic, and convert to GeoTiff in a standard GIS ready projection;
2. identify NASA-GSFC archived VHR DG stereo pair data and produce orthos and DEMs.

## Co-Is/Collaborators

Mr. Mark Carroll, Dr. Paul Montesano, Dr. Compton Tucker, Dr. Alexei Lyapustin, Dr. Daniel Slayback, Dr. David Shean, Dr. Oleg Alexandrov, Mr. Mathew Macander, Dr. Daniel Duffy, Dr. Jorge Pinzon, Dr. Gerald Frost and Dr. Scott Goetz

## Key Milestones

Automated database, beta	07/2018	TRL <sub>in</sub> = 2
Surface reflectance WV-2, beta	10/2018	
1/2° Mosaics and DEMs, beta	1/2019	
System Interface, API, beta	05/2019	
Optimization of performance	07/2019	
Client libraries and API tools completed	10/2019	TRL <sub>out</sub> = 5

↓

# Conclusions

- Six years ago all of my work was accomplished on local workstations
- Since then I have transitioned nearly all of my workflows into the cloud to take advantage of distributed and parallel processing
- This has freed up time to do analysis and enabled me to ask bigger questions
- Future plans all focus on use of cloud technologies to facilitate the processing of large datasets to answer science questions