Empowering Data Management, Diagnosis, and Visualization of Cloud-Resolving Models (CRM) by Cloud Library upon Spark and Hadoop

AIST-14-0022

Wei-Kuo TAO (PI, GSFC)
Xian-He Sun (Co-PI, IIT)
Shujia Zhou (Co-PI, Northrop Grumman Information Technology)
Toshihisa Matsui (Co-PI, ESSIC UMD)
Xiaowen Li (Co-I, GESTAR MSU)
Dan Duffy (Collaborator, GSFC)

John Thompson, Garrsion Vaughan, Michael Bowen, Carrie Spear (System support, GSFC)

Earth Science Technology Office

Goals

- Make Cloud Resolving Model output more usable by science community
  - Accelerate visualization of output (over 20 Terabytes).
  - Inter-compare large volumes of output from high-resolution or long-time simulations.
  - Diagnose key processes for cloud-precipitation.

- Demonstrate the value to distribute, visualize, analyze and inter-compare Cloud Resolving Model output and data with GCE and NU-WRF

GCE: Goddard Cumulus Ensemble model (1982 -)

NU-WRF: NASA Unified Weather Research Forecast (2010 -)

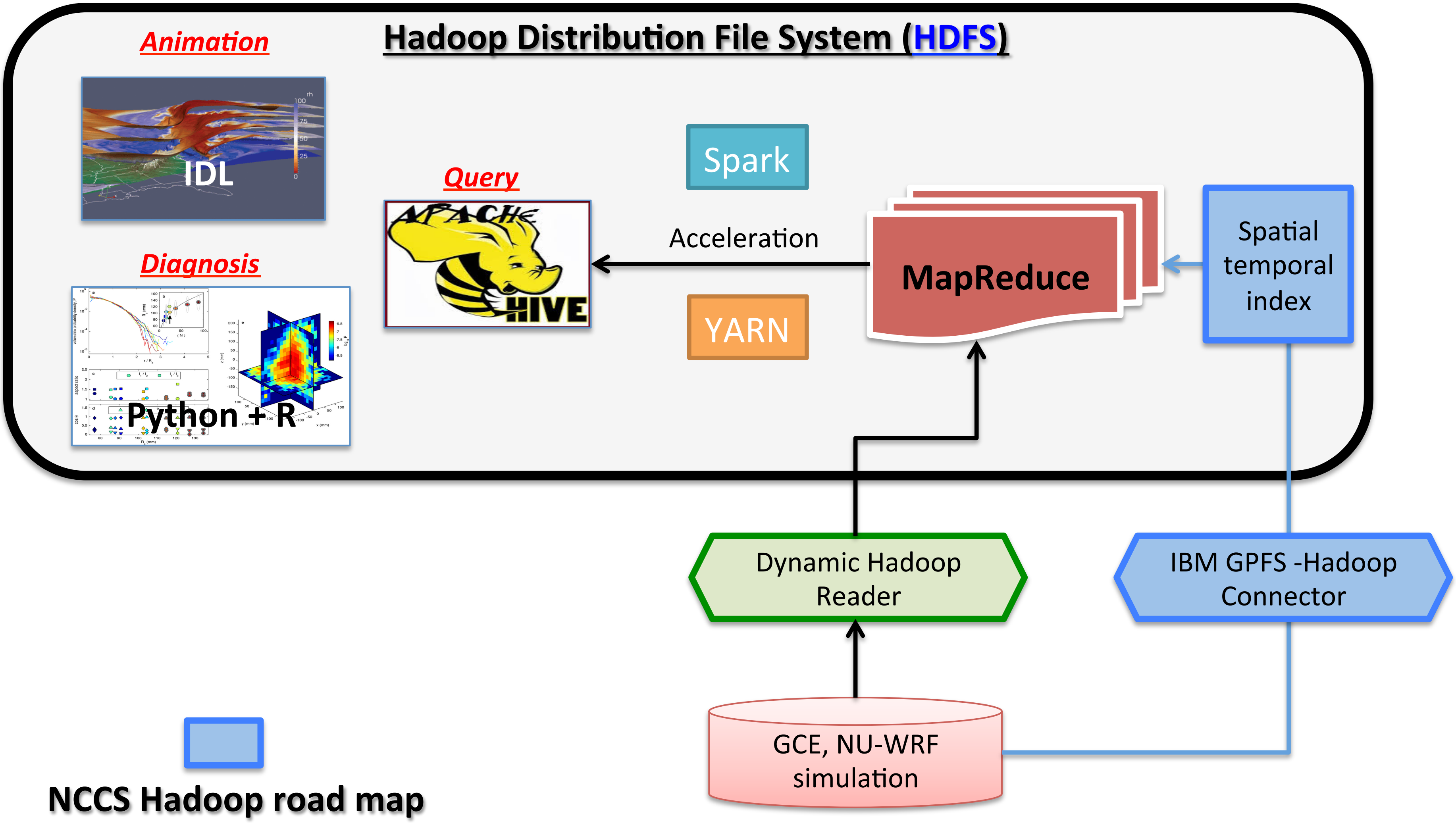
Approach

- Develop Super Cloud Library (SCL) supporting Cloud Resolving Model using Spark and Hadoop.
  - *Create cloud data files*
  - *Develop data model and Hadoop format transformer*
  - *Develop dynamic transfer tool to Hadoop*
  - Develop subset and visualization APIs (Application Programming Interfaces)
  - Develop Web User Interface
- Conduct **Demo** of GCE and NU-WRF diagnosis on NCCS

- ***Hadoop/Spark Ecosystem is powerful, mature, broadly deployed***
- ***Data lake uses HIVE tables to store and manipulate big data***

ESTO
Earth Science Technology Office

Super Cloud Library (SCL)



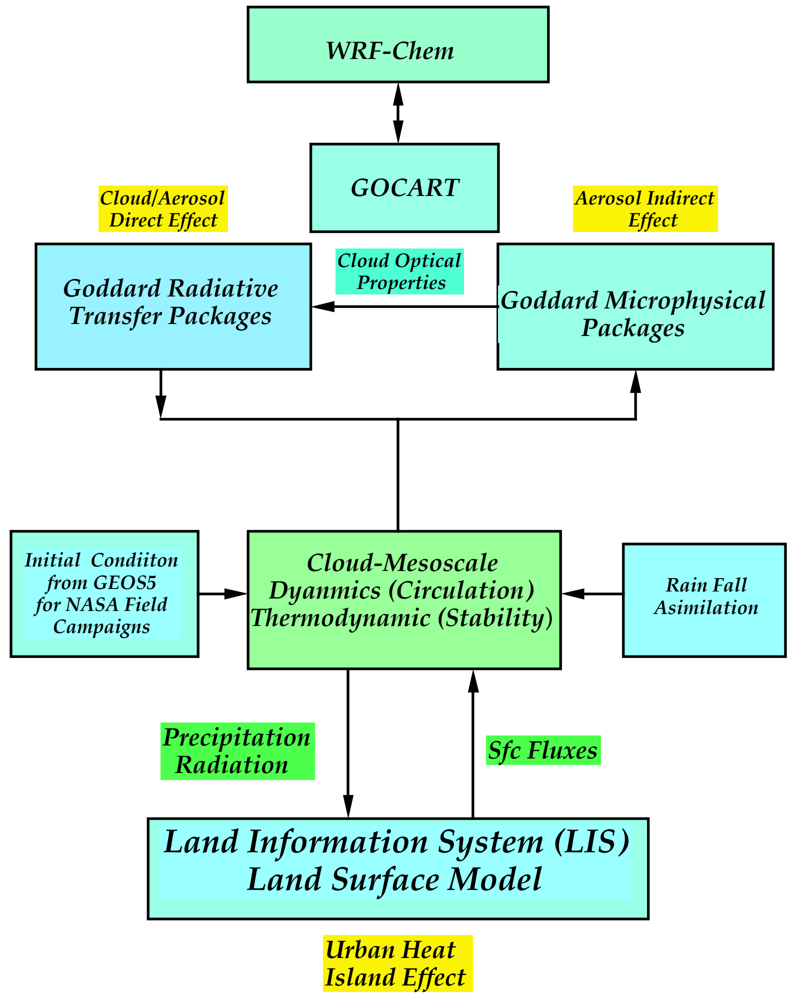
Model and Data

NASA-Unified WRF

Peters-Lidard *et al.* (2015)

Tao *et al.* (2011, 2013, 2016), Wu *et al.* (2016), Matsui *et al.* (2014)

- Real time forecast to support GPM ground validation field campaigns
- Precipitation processes study to support satellite missions (GPM and CaPPM)
- Cloud-aerosol-precipitation interaction (direct and indirect)
- Regional climate (downscaling from climate models)
- Dust effect
- Land surface effect on precipitation
- Hydrological study (flood, draught)



Blue box:
Goddard Physical Packages

Supported by NASA MAP, PMM, IDS

NU-WRF Real Cases

Long-Term Case

- **Grid (9km)**: ***600x400x50***
- **Date**: West African Monsoon (June-July-August in 2006)
- **Output frequency**: 3hr
- **Data Sizes**: 0.34TB
- **Sensitivity Runs**: Fixed-SST simulation to see the impact of dynamic SST on West African Monsoon.

Semi-Giga Cases

- **Grid (2km)**: ***2500x2500x50***
- **Date**: Tornado Outbreak (6days), Tropical Storm Bill (6days)
- **Output frequency**: 1hr
- **Data Sizes**: 145 files x 2, 1.73TB x 2
- **Sensitivity Runs (2014 Tornado Outbreak):** Used 2011/2014 microphysics radiation to see the impact of physics update.
- **Sensitivity Runs (Tropical Storm Bill)**: Time-lag simulations, which initialized 1day later to see the impact of forecast leading time.

West African Monsoon Case

(Long-Term Simulations)

Domain and Duration

- Grid Spaces: 9km domain for analysis (LDA/EDA) (600x400x50).
- Duration: 3month Wet (JJA) in 2006: SOP for intercomparison Aug 9 ~ 17 2006 (within AMMA SOP) and Peak Monsoon:

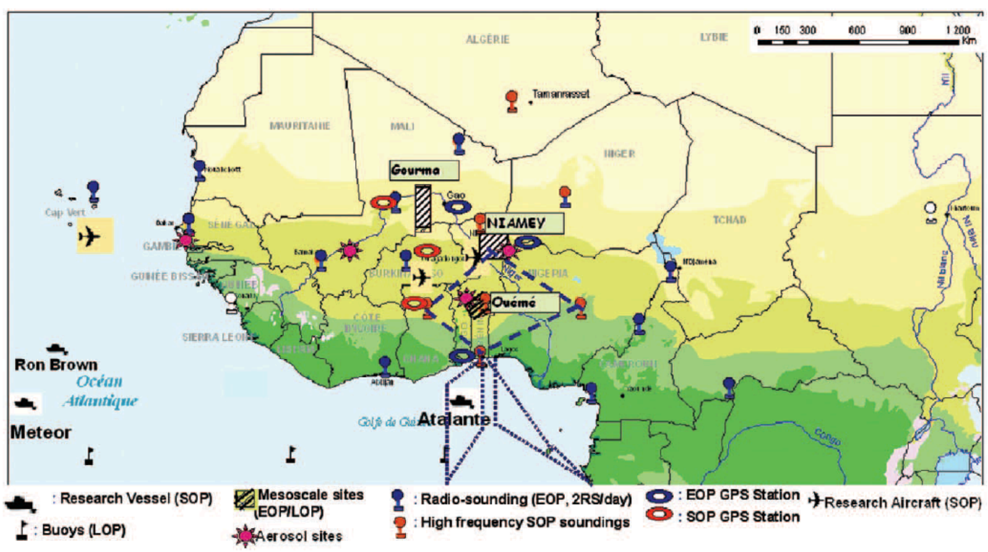


FIG. 3. Field implementation of AMMA observations based on nested networks. Circles indicated the atmospheric sounding network activated during the SOP.

Goal and Data

Objective

- Seasonal simulation: Long-term free running simulations to investigate seasonal variability of energy and water budget through SCL.

SOP III is designed for model intercomparison between free-running and analysis through NU-WRF data assimilations.

Goddard Cumulus Ensemble (GCE) Model
(1982 – , over 150 papers published)

Water and energy cycles in the tropical climate system,

Redistribution of ozone and trace,

Deep convection related to global warming,

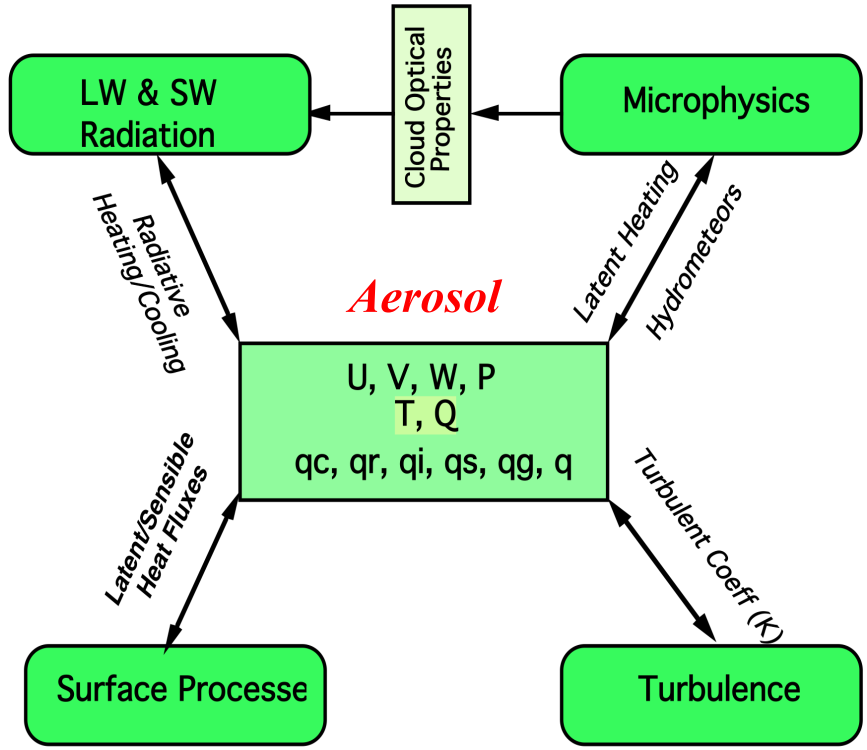
Precipitation processes - precipitation efficiency,

The aerosol impact on precipitation and rainfall,

Surface process on precipitation and rainfall,

Representation of cloud microphysical processes and their interaction with radiative forcing

TRMM and GPM rainfall and latent heating retrieve algorithm performance



GCE Model Description: Tao and Simpson (1993), Tao *et al.* (2003), Tao (2003), **Tao *et al.* (2014)**
CRM review paper: Tao and Moncrieff (1999 – Geophy Rev)
Aerosol review paper: Tao *et al.* (2012 – Geophy Rev)

Supported by NASA PMM and MAP

GCE Real Cases

Long-Term Case

- **Grid (1km)**: 256x256x106
- **Case & Date**: DYNAMO (Nov. 1~Dec. 10, 2011)
- **Output frequency**: 1hr
- **Data Sizes: 0.82** TB
- **Sensitivity Runs**: Used Goddard and Morrison microphysics to see the impact of different microphysics scheme.

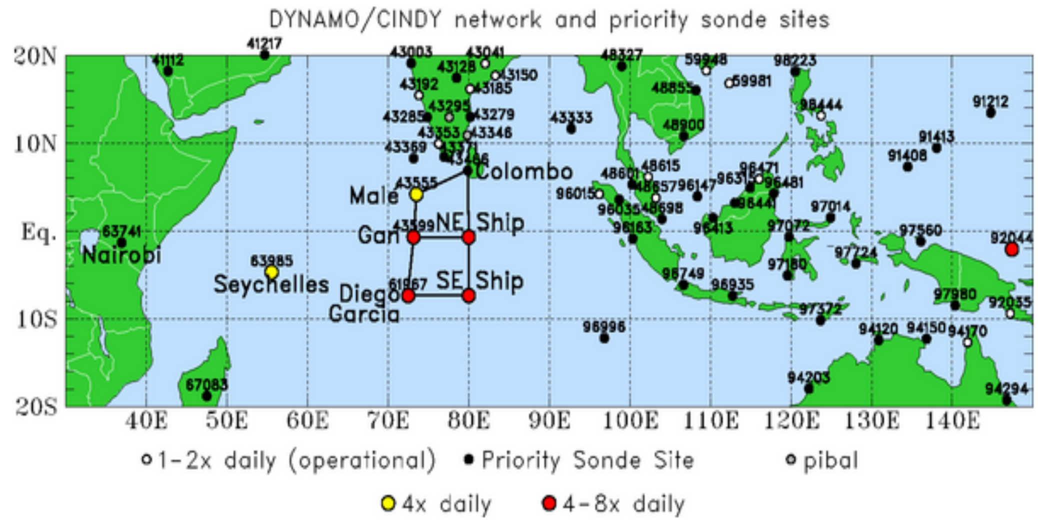
Giga Case

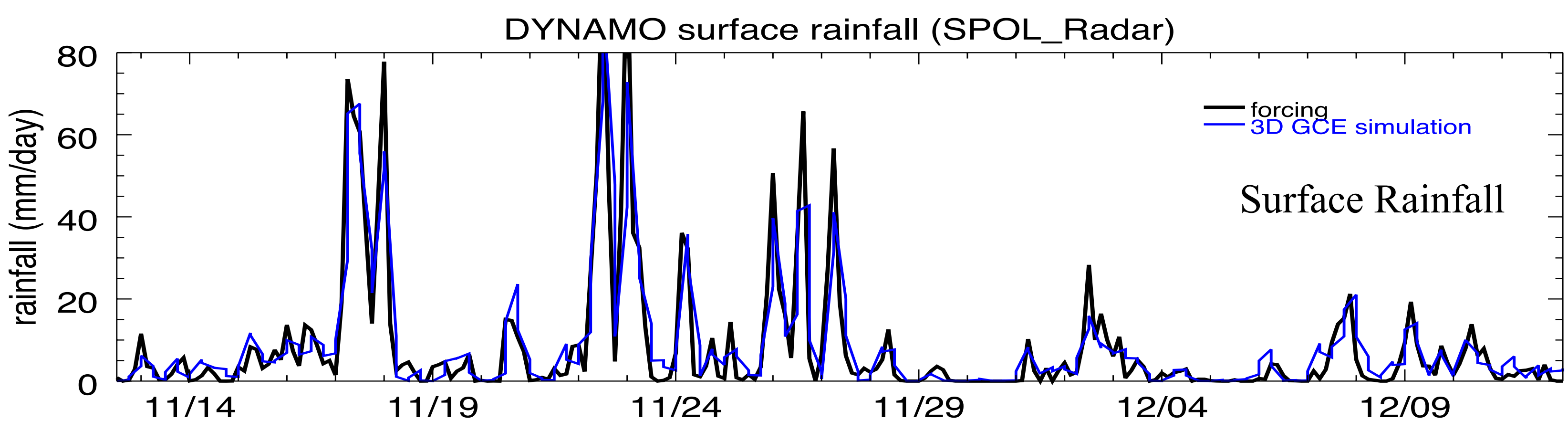
- **Grid (250m)**: 4096x4096x106 ≈ 1.8 giga points
- **Case & Date**: DYNAMO (Nov. 23~Nov. 29, 2011)
- **Output frequency**: 1hr
- **Data Sizes: 15**TB
- **Sensitivity Runs**: None

GCE DYNAMO Simulation

DYNAMO: DYNAmics of the MjO

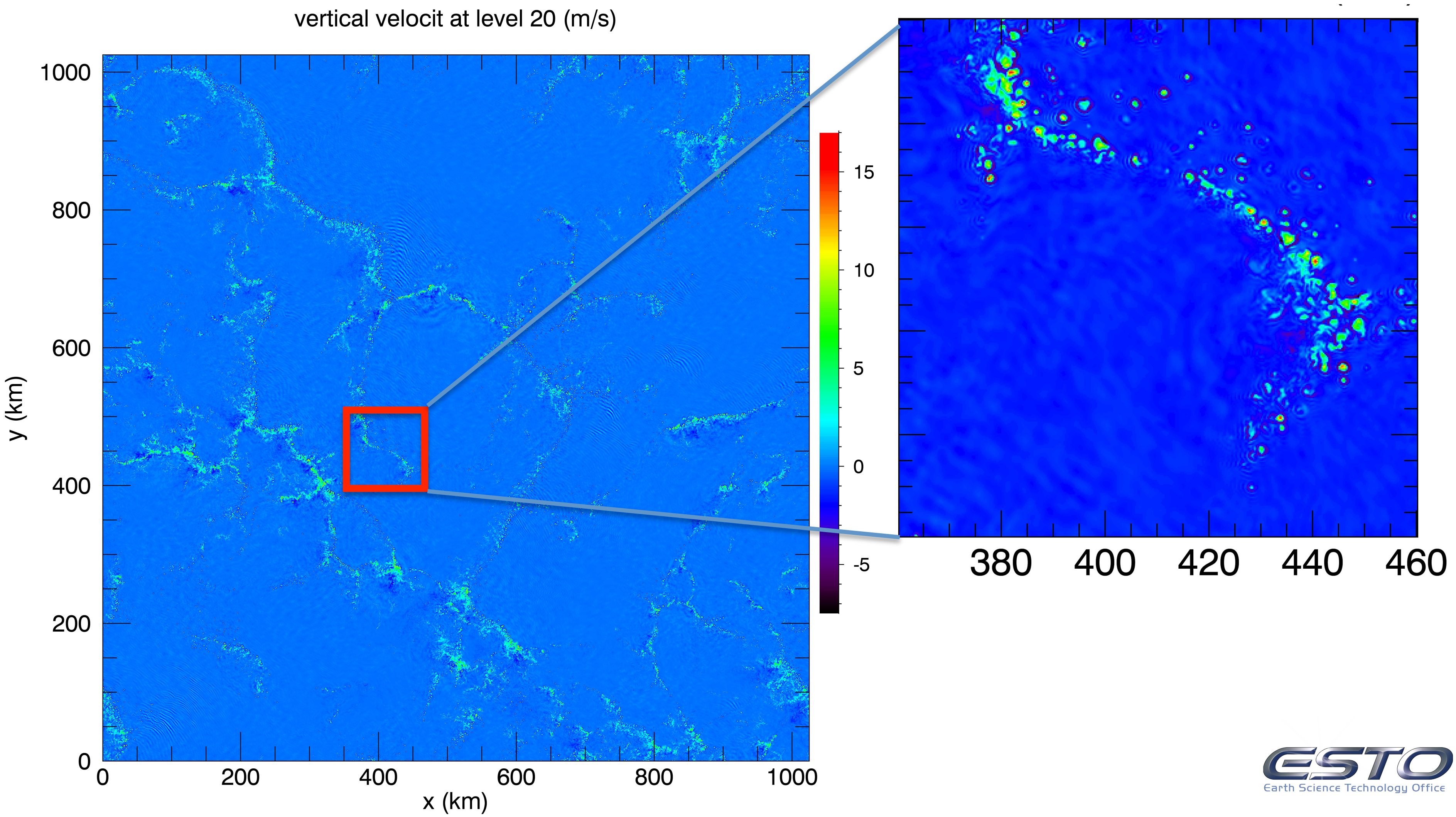
MJO: Madden-Julian Oscillation





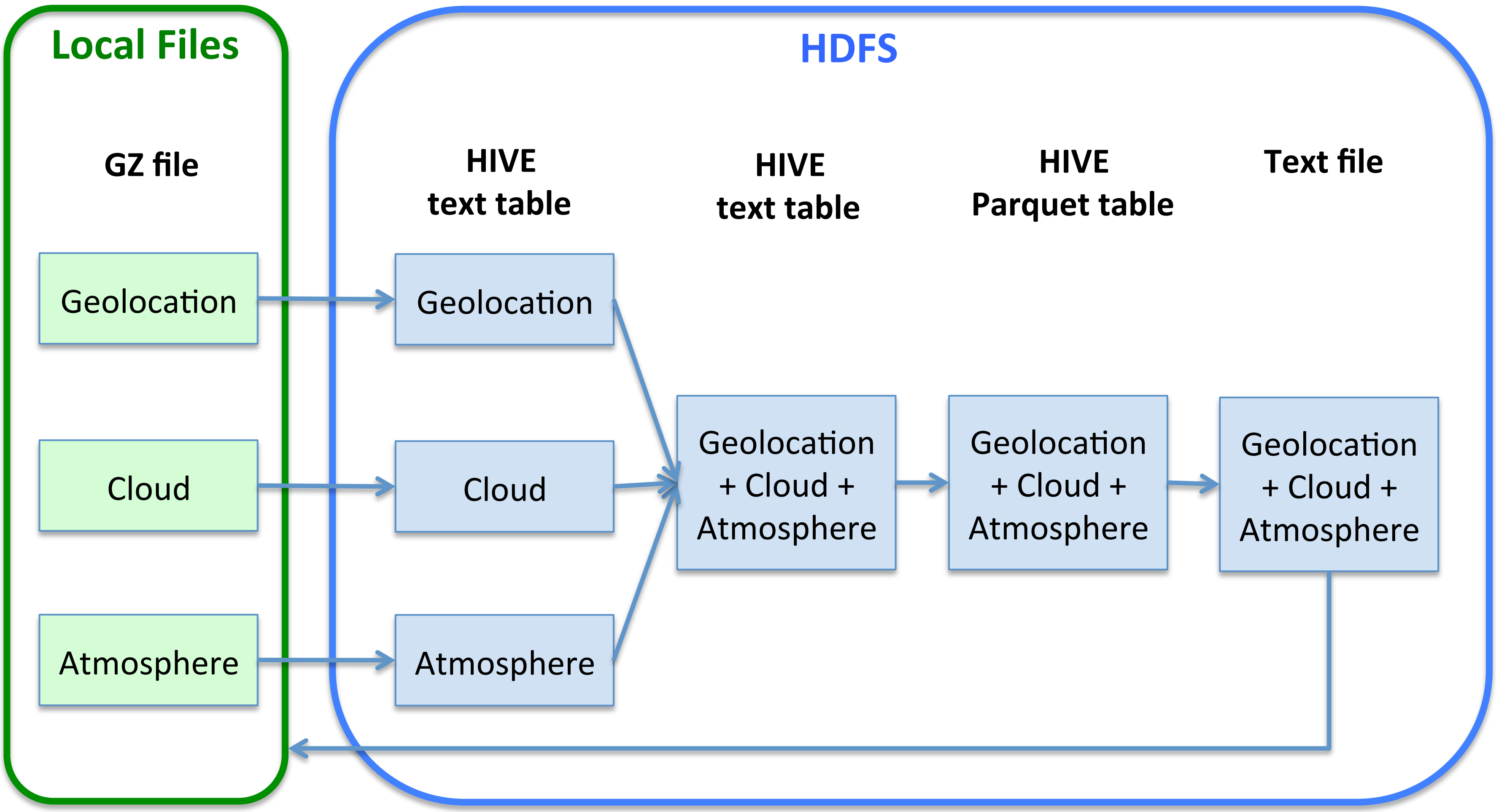
GCE Giga Case (Δ=0.25km)
DYNAMO simulated vertical velocity details

4096x4096x106 grid points



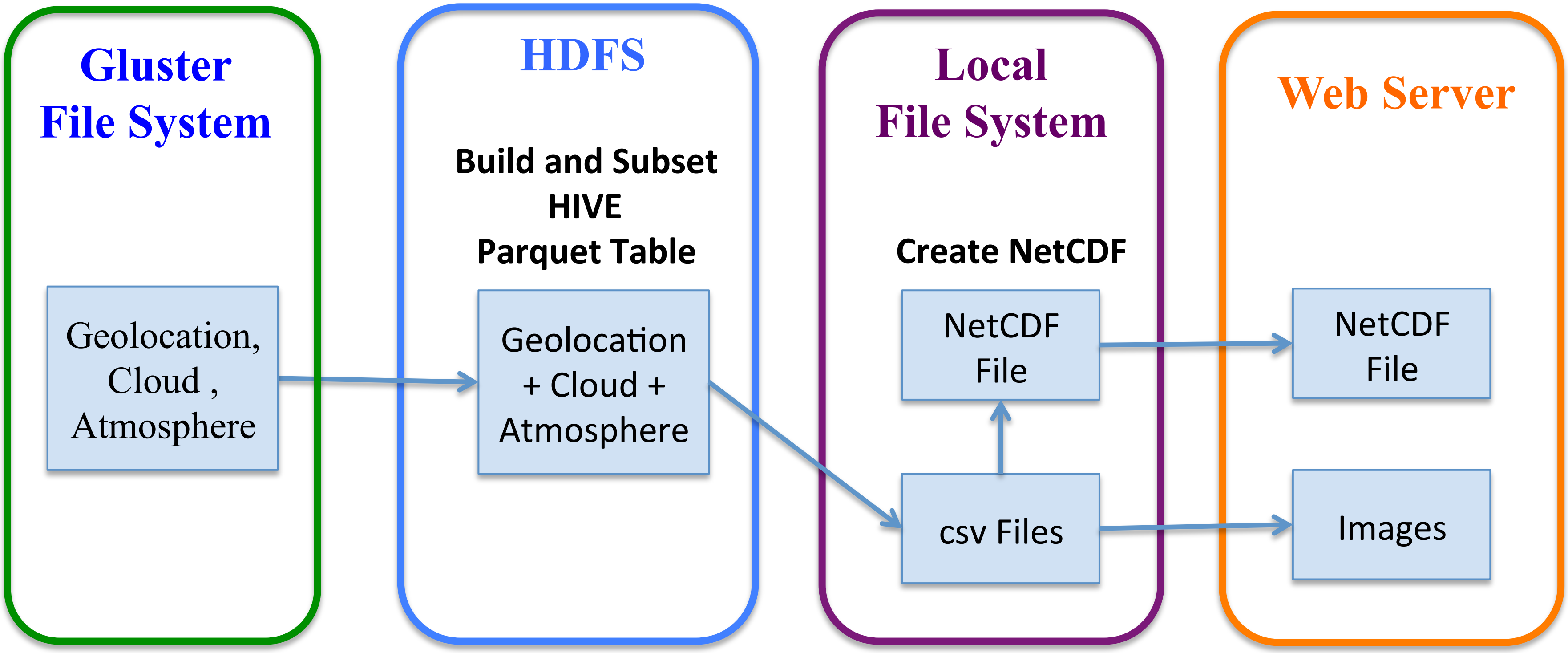
Subset, NetCDF Creation and Distribution

Adaptively Subsetting: Data Flow



Using multiple tables saves storage as well as be flexible. Combining tables is compute-intensive.

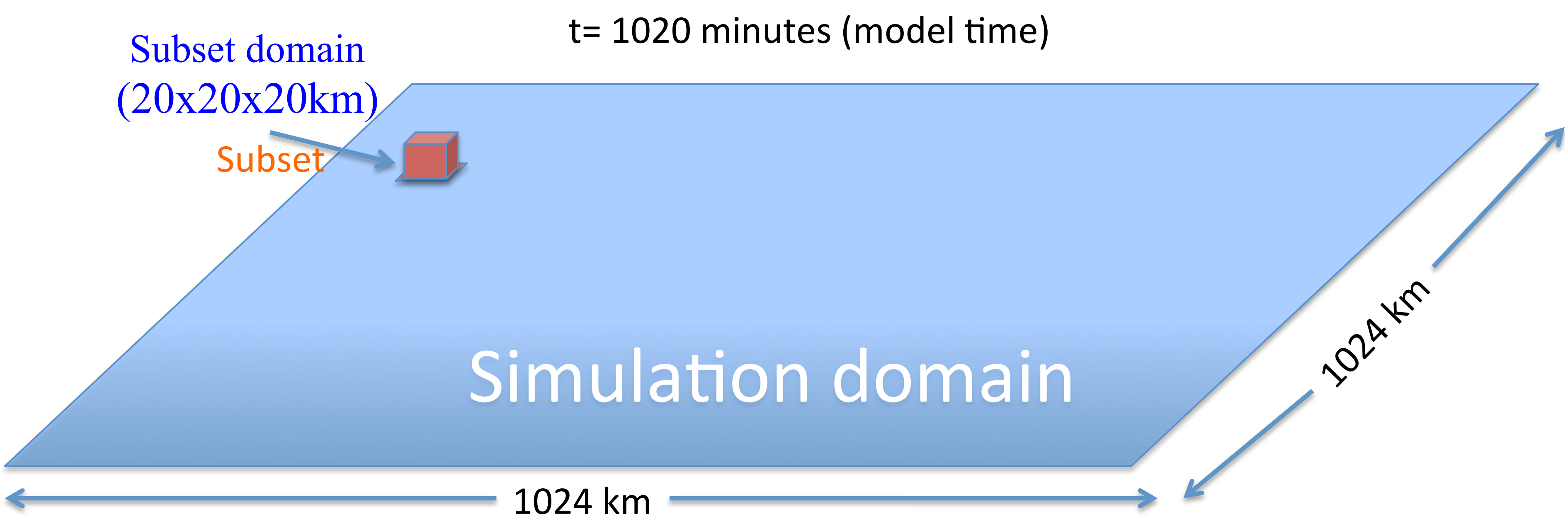
End-to-End Data Analysis and Distribution: Data Flow



A user can use Hive tables, subset, perform sophisticated analysis with Python and IDL, create a NetCDF file and download it.

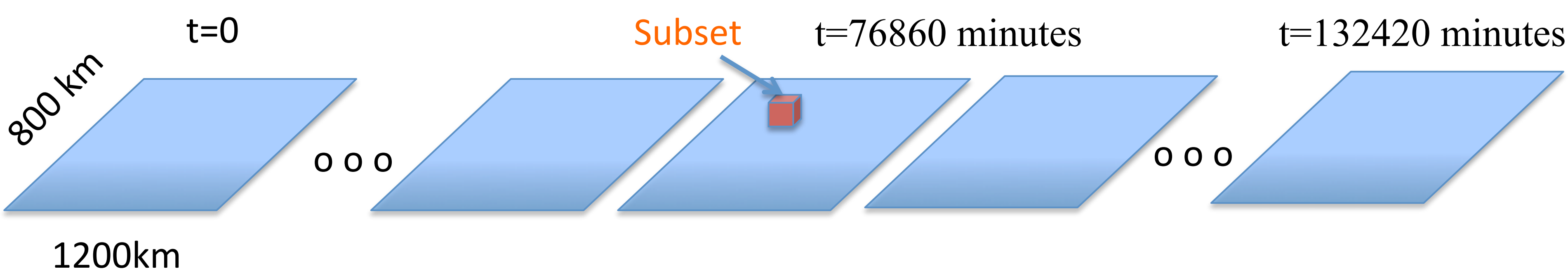
Create HIVE Tables and Adaptively Subsetting (High-Resolution Simulation)

- **GCE**
  - Data
    - Gridn: 4096*4096*106
    - Time: 6 days.
    - Size: ~1.0TB for a 3-hour Hive table.
  - Subset 3D and 2D tables for the maximum w (upward wind velocity) in a particular time



Create HIVE Tables and Adaptively Subsetting (Long-Term Simulation)

- **NU WRF**
  - Data
    - Grid: 600*400*50
    - Time: June, July, August (92 days)
    - Size: ~5TB for a 92-day HIVE table.
  - Build tables and subset at the production Hadoop cluster.

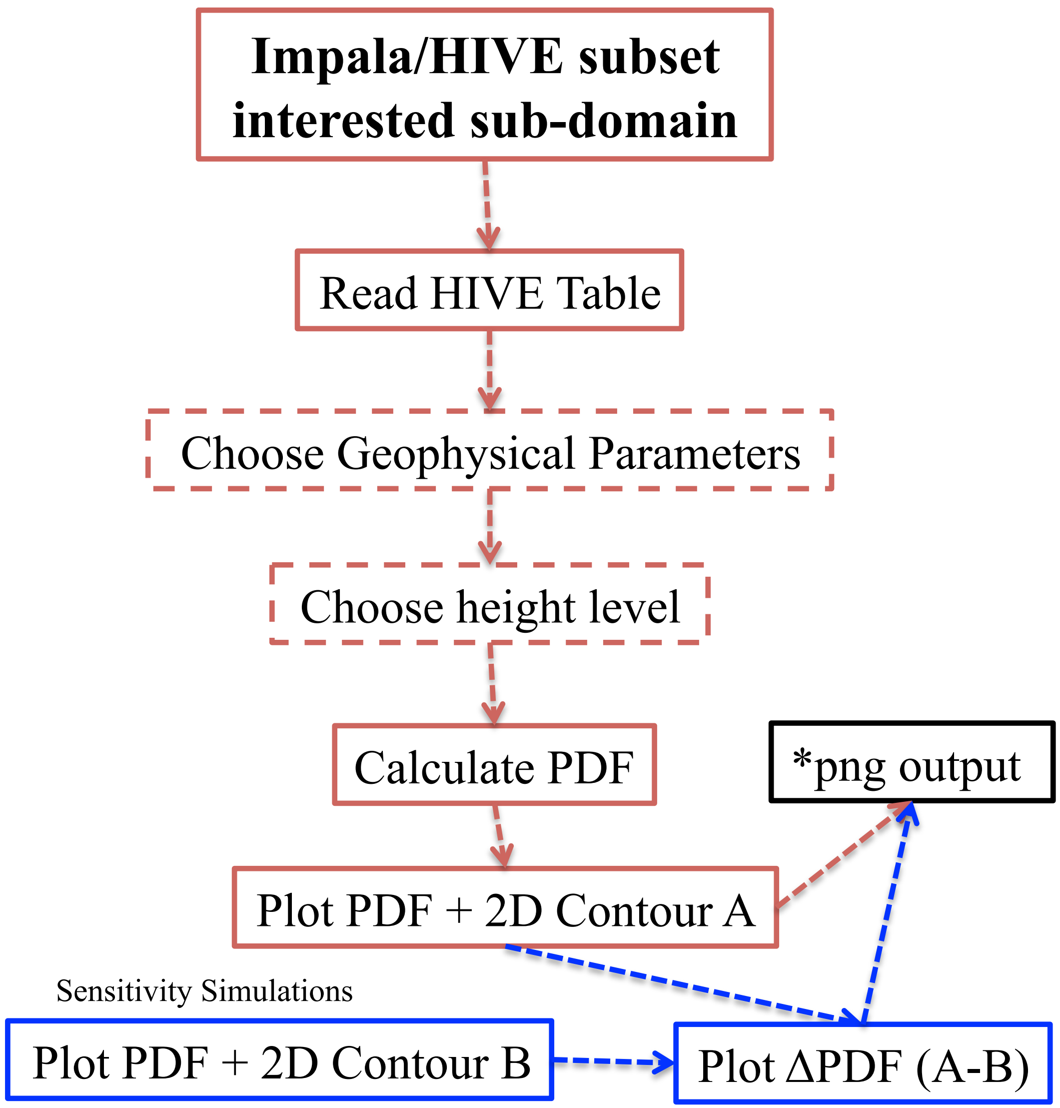


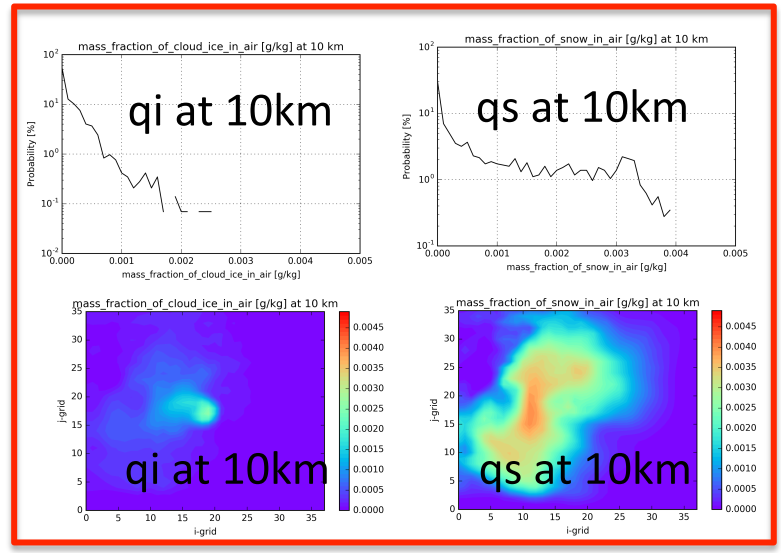
Diagnosis and Visualization via Python, IDL, R, Hadoop

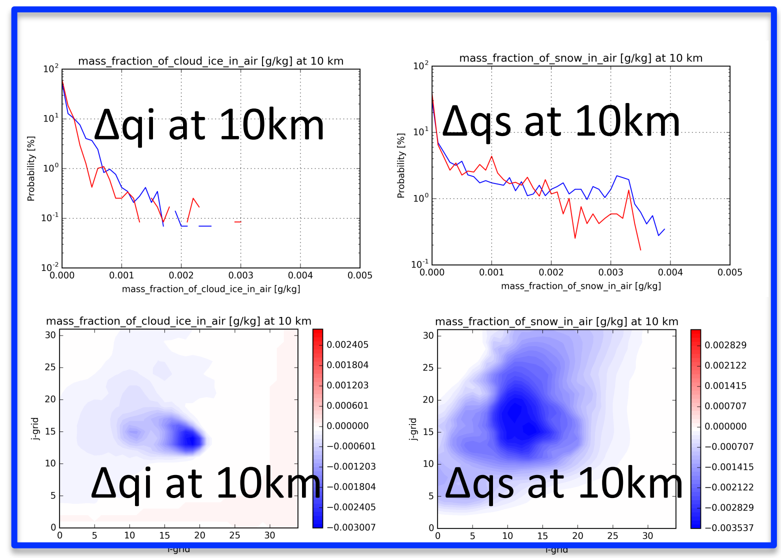
Earth Science Technology Office

Python CRM Diagnostic Module

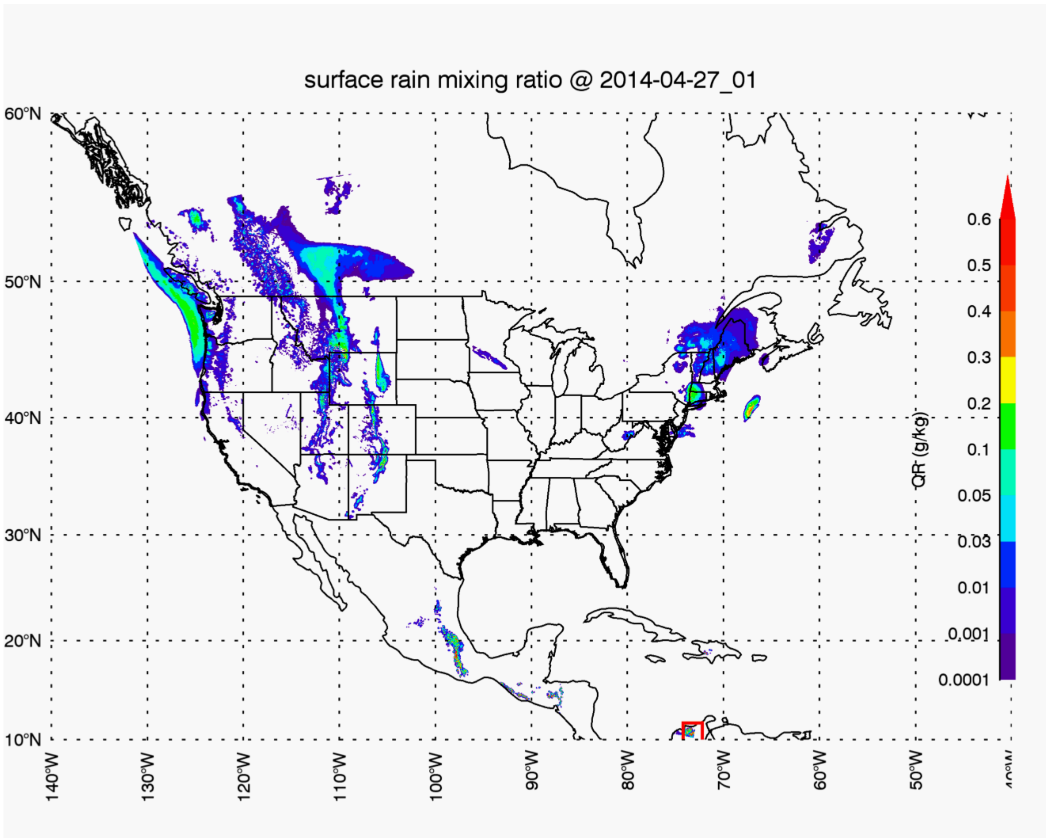
-Probability Density Function (PDF)-







WRF Semi-giga Scale (*2500x2500*) Simulation
Impala Subset + IDL



Surface Rain Mixing Ratio

+

Maximum Rain Mixing Ratio in Red Box

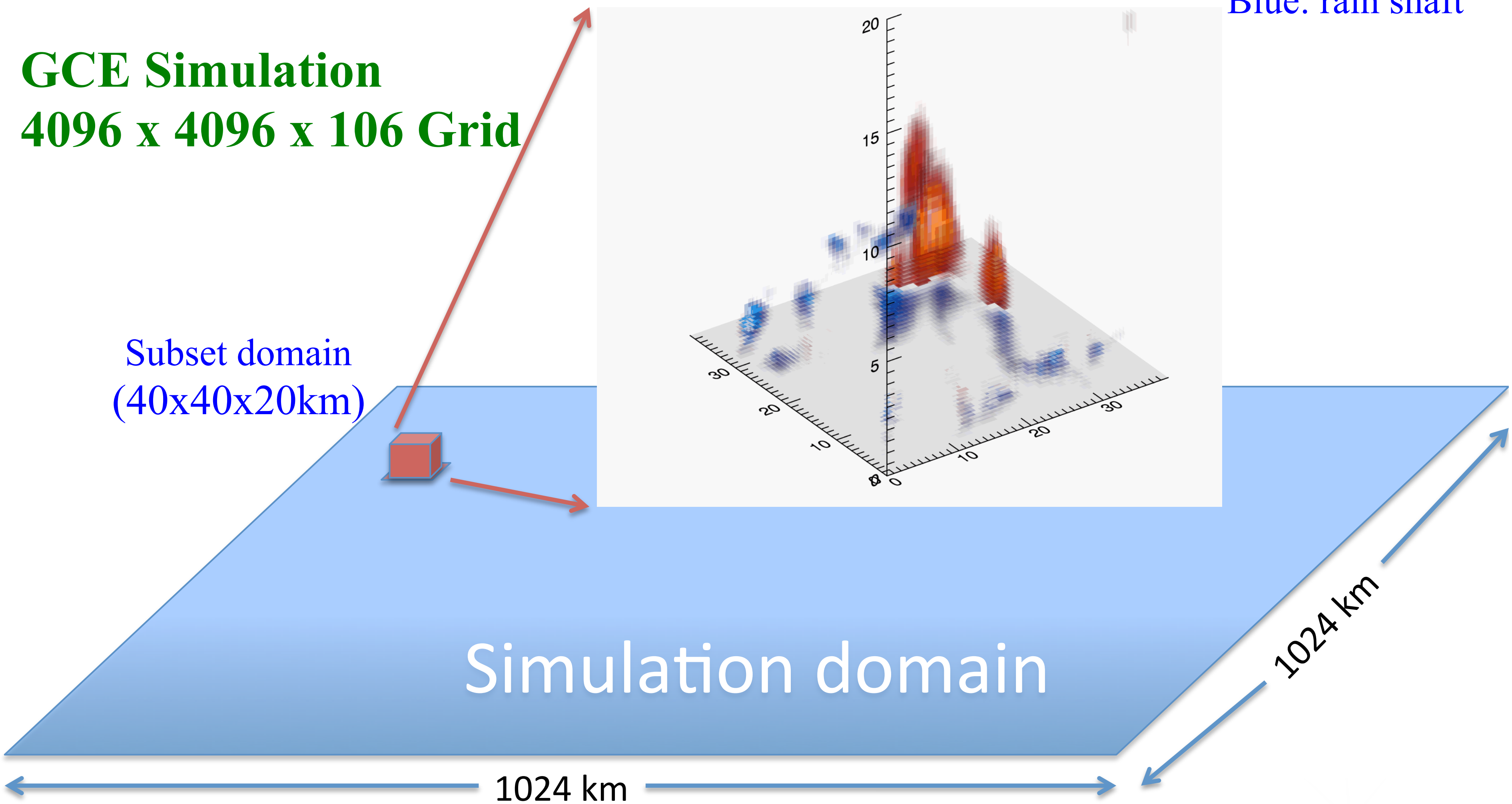
Earth Science Technology Office

Impala Subset + IDL

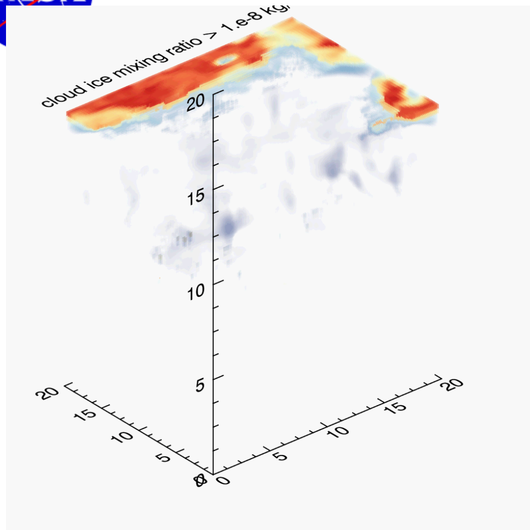
Can Overlay Multiple Variables to Study Their Relations

Red: updraft core
Blue: rain shaft

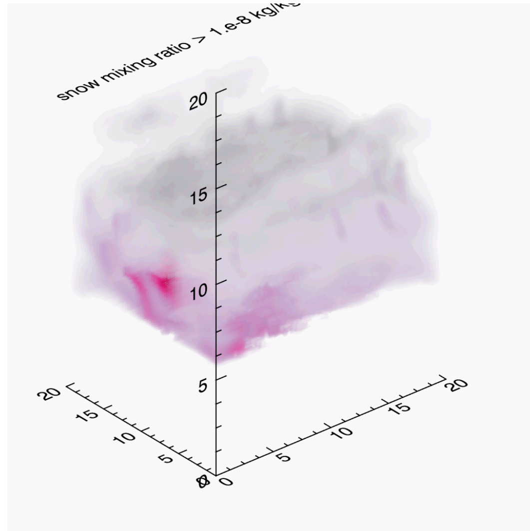
GCE Simulation
4096 x 4096 x 106 Grid



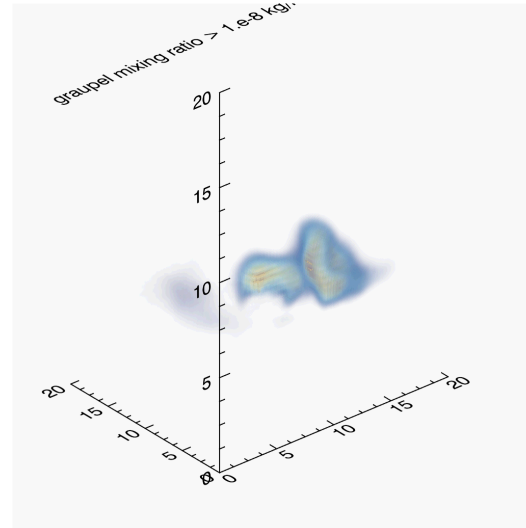
Pristine Ice



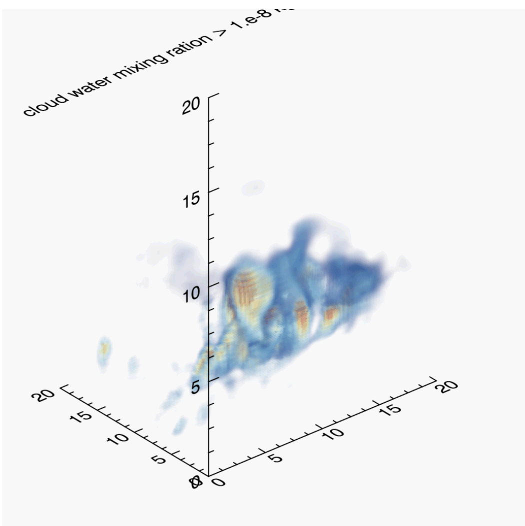
Snow



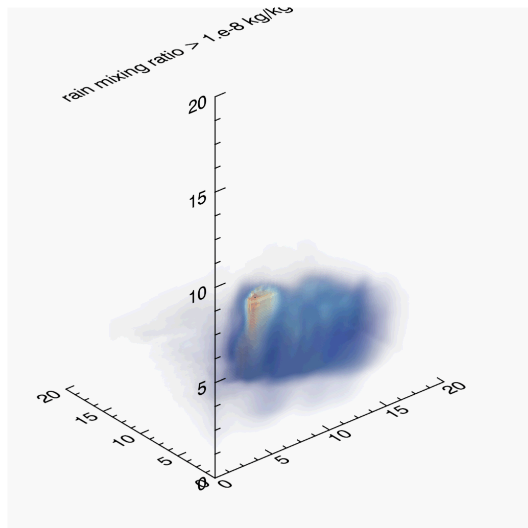
Graupel



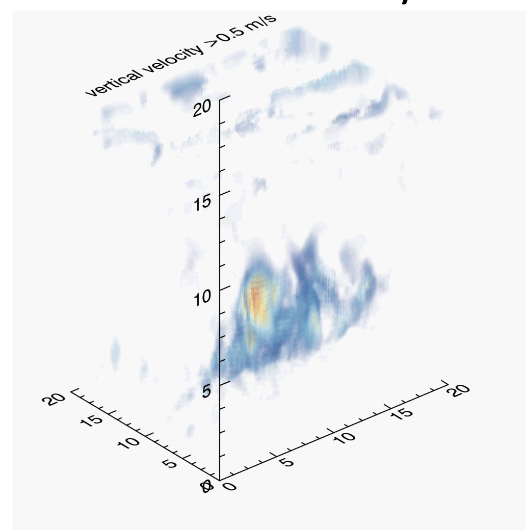
Cloud Water



Rain



Vertical Velocity

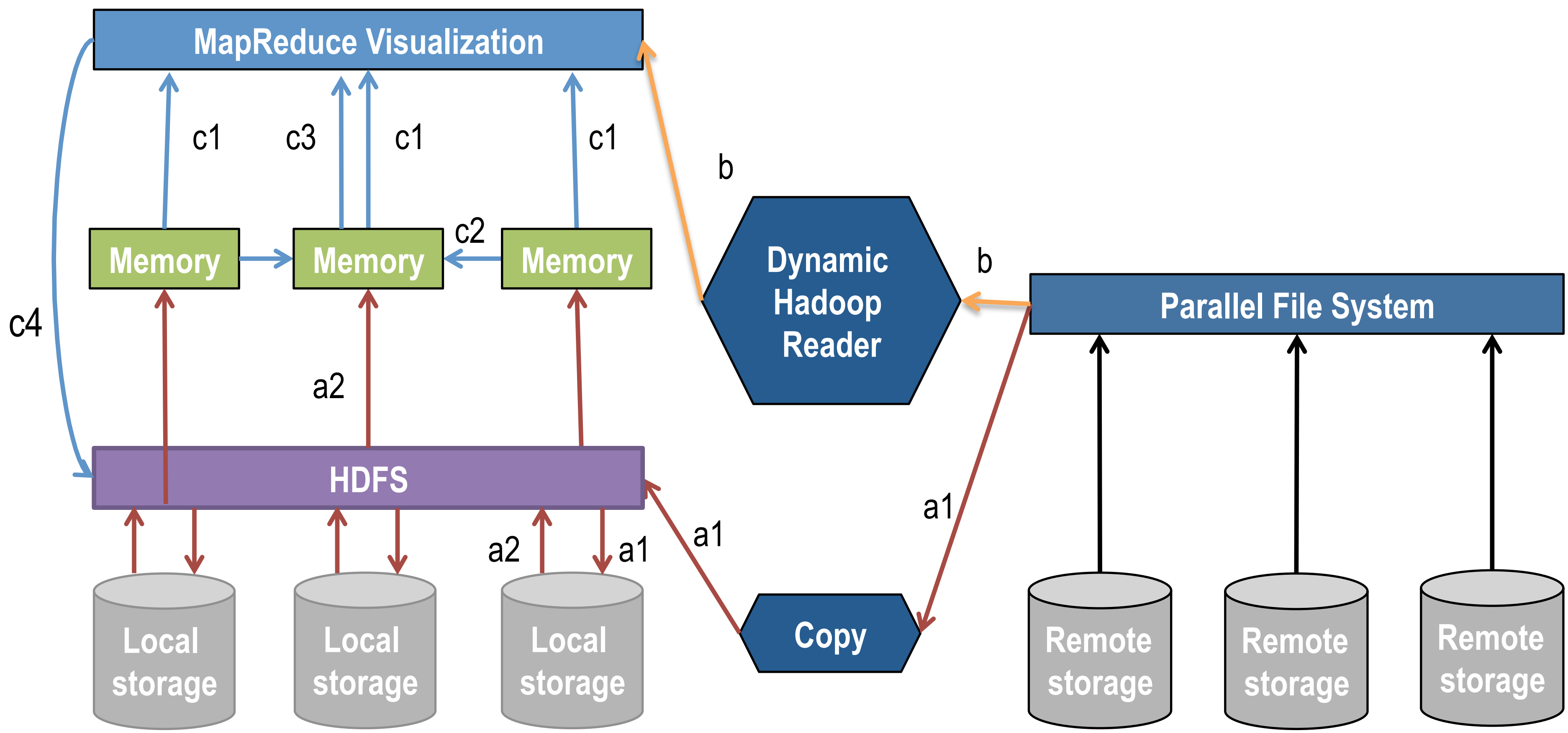


Giga GCE: 4096 x 4096 x 106 → 79 x 79 x 106

Earth Science Technology Office

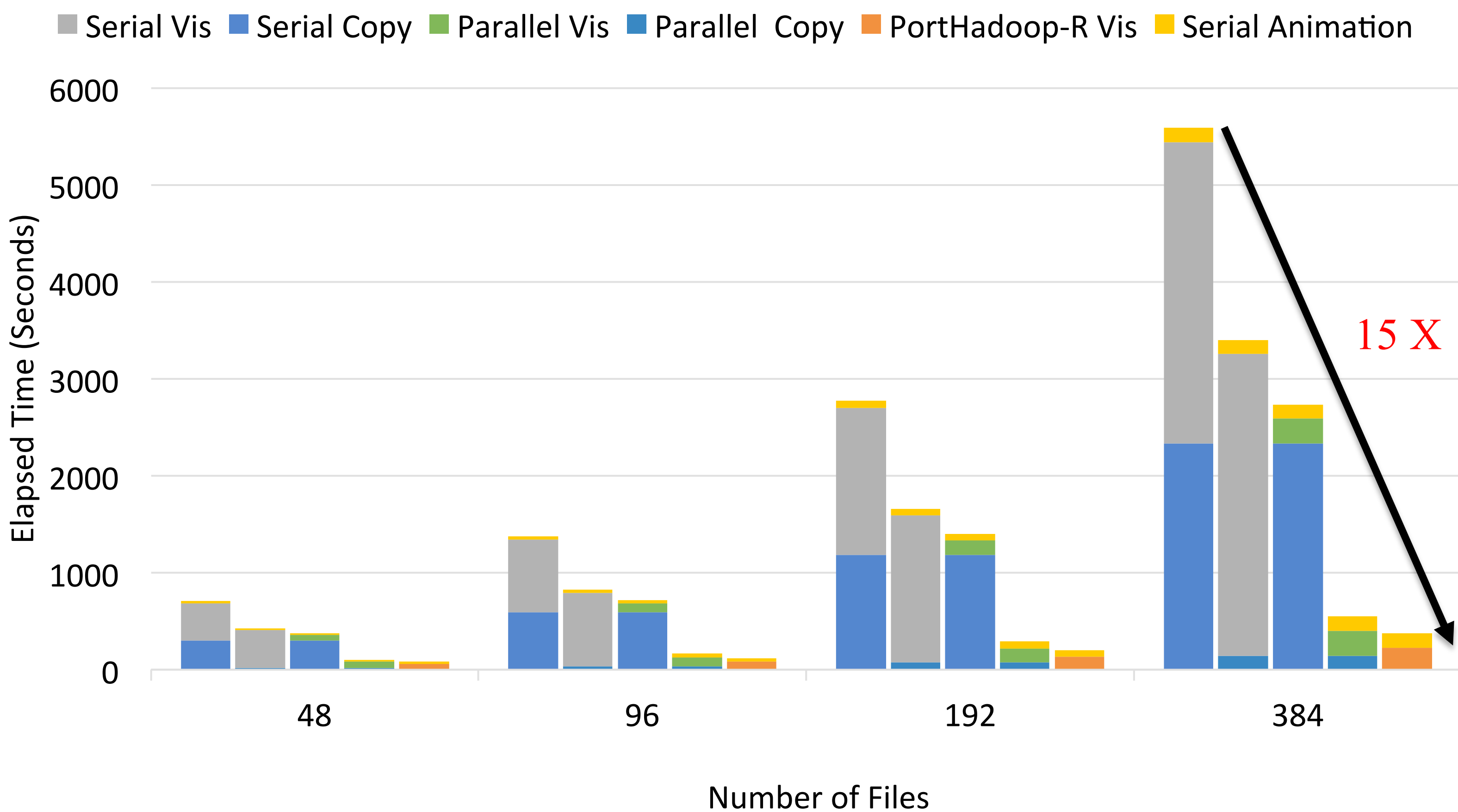
Cross-Platform Hadoop (PortHadoop), GPFS Interface, R Interface R Hadoop, Spark visualization

Dynamic Visualization with R and PortHadoop: System



PortHadoop reads data from PFS to memory of Hadoop cluster directly and concurrently

R-PortHadoop VS Conventional Methods: 15 times faster



Serial Copy: one copy command issued on one Hadoop node
Parallel Copy: 4 copy commands issued on each Hadoop node (32 in total)

Serial Vis: read all data into one fat node, visualize and animate
Parallel Vis: visualize all files on all Hadoop nodes and animate on the master node

PortHadoop-R Vis: visualize all files into images
Serial Animation: merge all images into one animation

Web Interface

Earth Science Technology Office

Illustration of Dynamic Web Interface: Setting

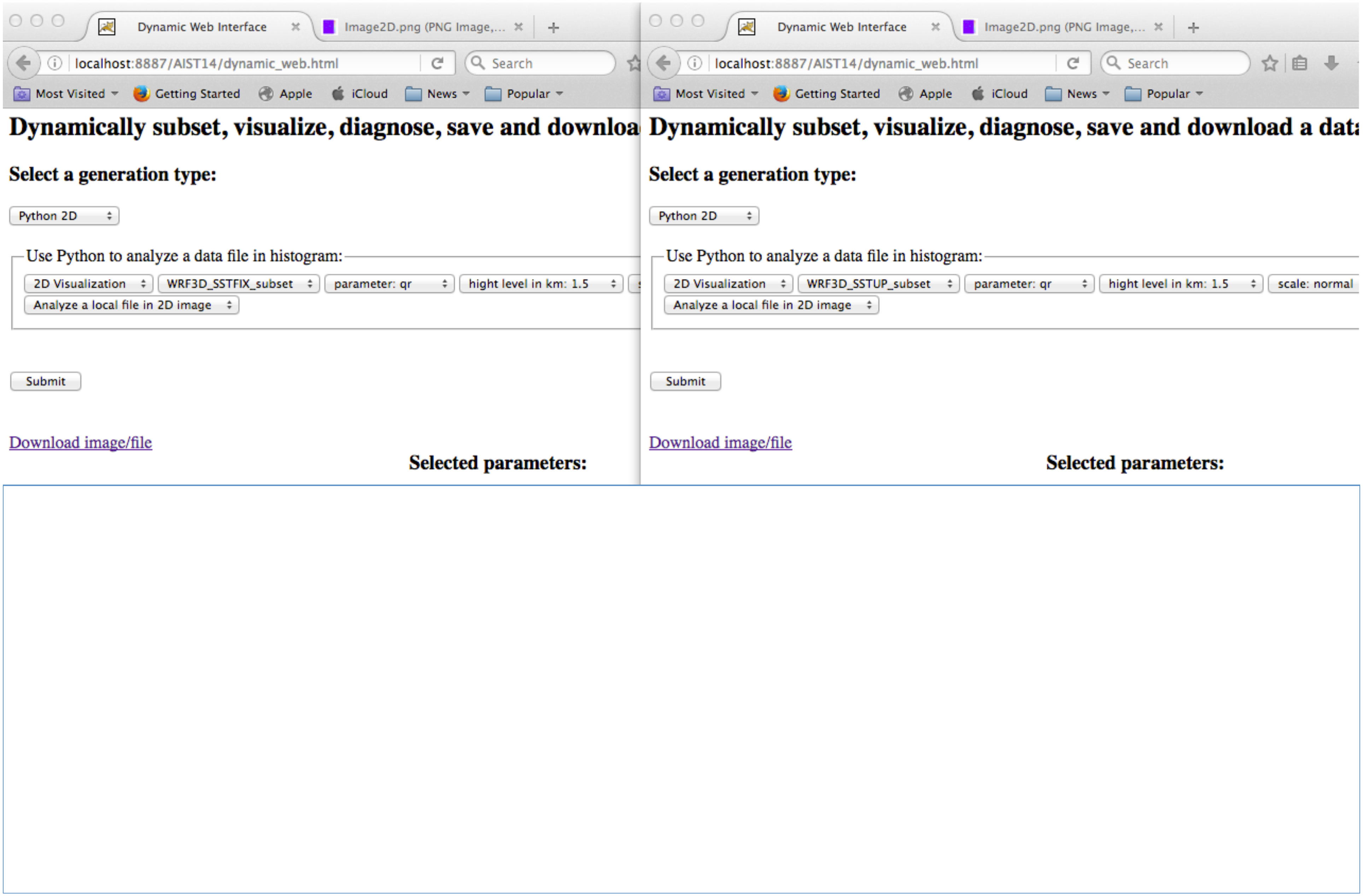
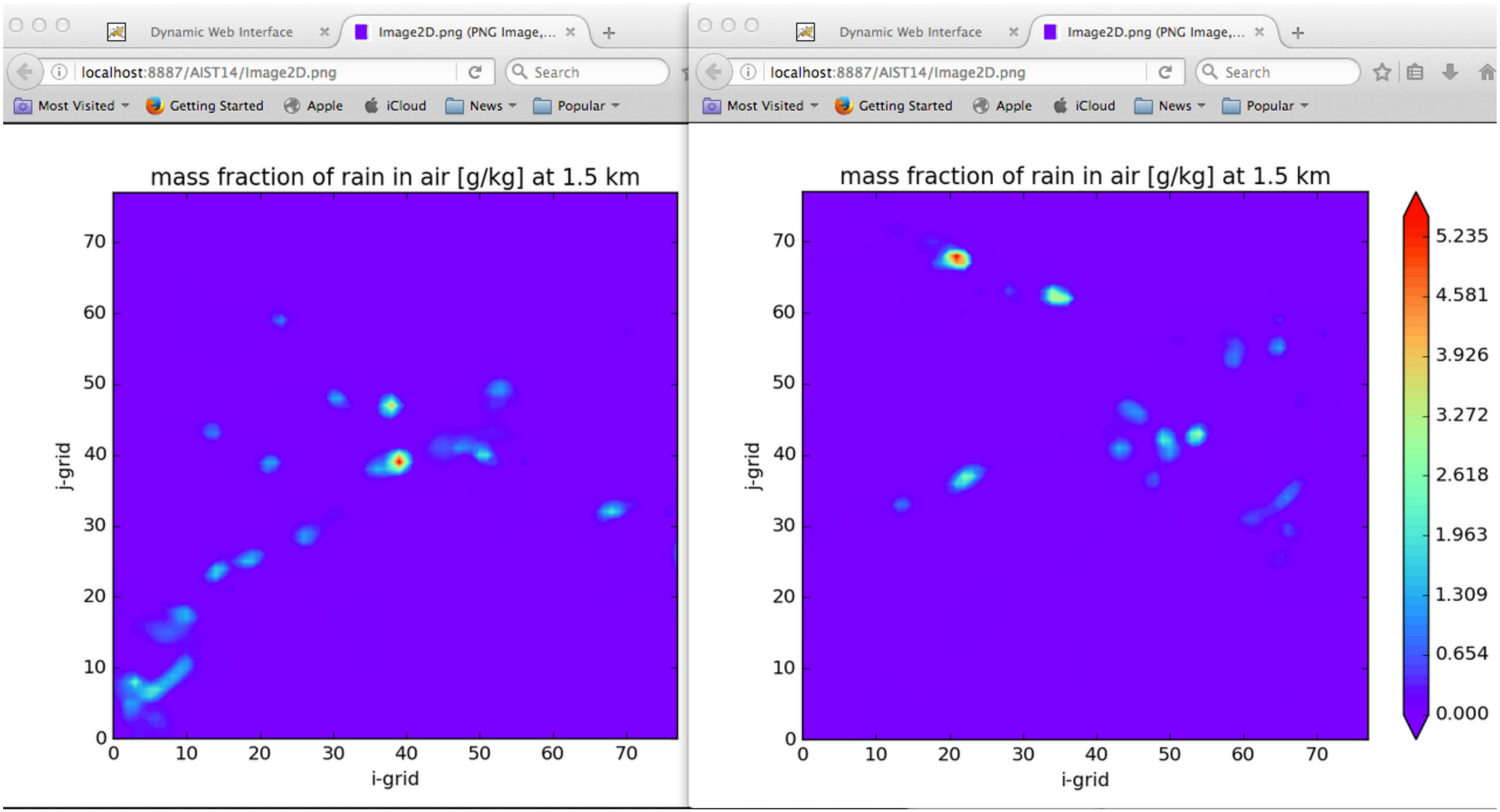


Illustration of Dynamic Web Interface: Diagnosis



Summary

- GCE giga-scale (4096x4096x106), NU-WRF semi-giga-scale (2500x2500x50), 3-month NU WRF and other simulations were performed and analyzed with IDL, Python, R, Hadoop and Spark
- Web interface allows for adaptively subsetting a Terabyte-size data with around a critical point such as maximum vertical wind speed, performing diagnosis and visualization, and converting the subset data to a downloadable NetCDF file
- HIVE Schema was developed to store and process big data efficiently and flexibly
  - Data Lake approach
  - Facilitate inter-comparison of models
- Impala and SparkSQL codes were developed to process big data efficiently
- PortHadoop was developed to support cross-platform data analysis