



# Federated Observational and Simulation Data in the NASA Center for Climate Simulation

## Data Management System Project

Glenn Tamkin, Scott Sinno, Roger Gill, David Fladung,  
Dave Ripley, Savannah Strong & John Schnase

*NASA Center for Climate Simulation (NCCS)*

*NASA Goddard Space Flight Center*

# NCCS Mission

- *Traditional*

- Enable scientists to increase their understanding of the Earth and the universe by providing state-of-the-art high performance computing, storage, network, and application solutions
- Provide large-scale compute engines, analytics, data sharing, and high-end computing services

- *Future*

- Develop a data services capability to better support the climate research communities and prepare the way for technology advances





# New Challenges

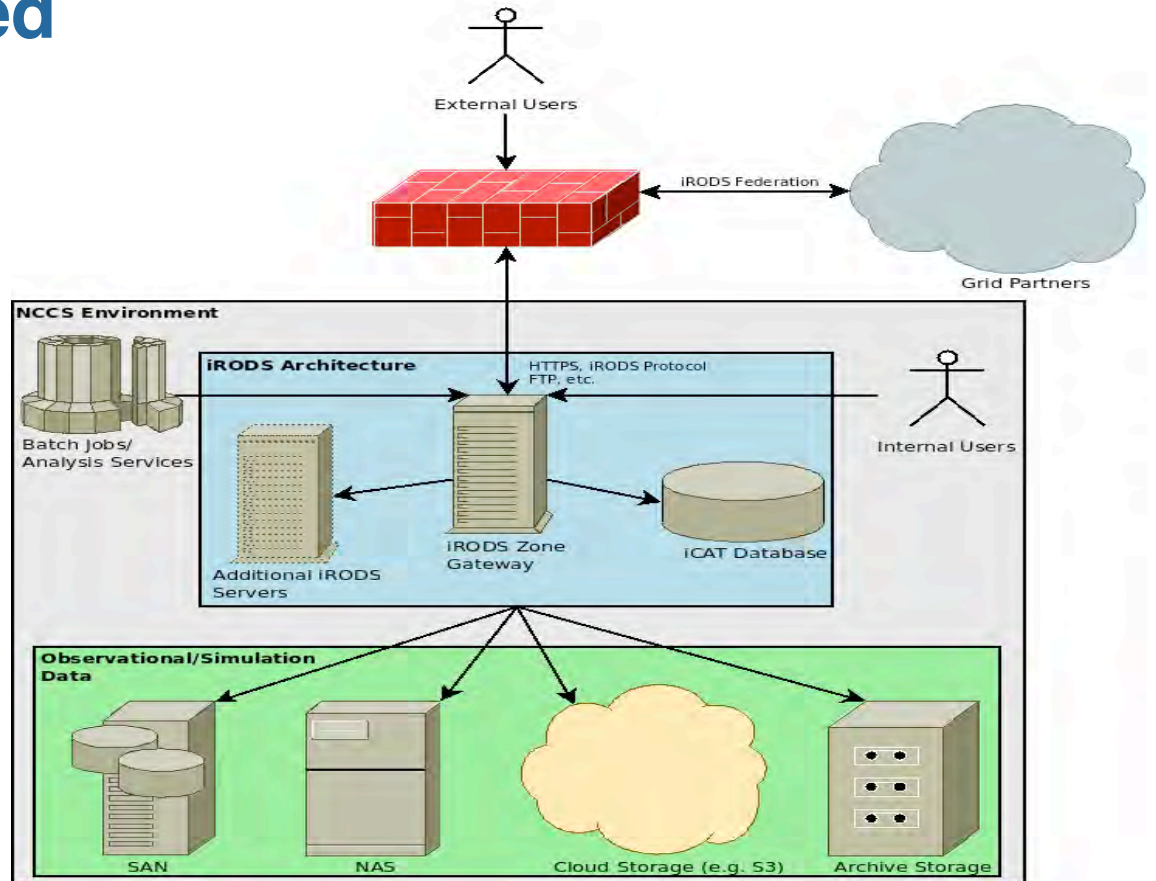
- *Finding* observational and model data for use in climate and weather studies
- *Accessing* the geographically distributed data
- *Managing* the massive digital holdings, which are measured in petabytes and hundreds of millions of files
- *Maintaining* the data, which must often be preserved for decades
- *Supporting* data sharing, data publication, and data stewardship





# NCCS Test Bed

- iRODS abstracts physical location of data
- iRODS assists with archive management







# Preliminary Tests – Ingest/Registration

- iRODS rules and microservices allow data to be stored in configurable collections based on data policies
- Replication to backup storage resources also supported

Name	Resource	Size	Date Modified
MERRA100.prod.assim.instM_3d_	demoResc	125.88 MB	July 14, 2010, 11:02 am
MERRA100.prod.assim.instM_3d_	demoResc	126.24 MB	July 14, 2010, 11:07 am
MERRA100.prod.assim.instM_3d_	demoResc	127.29 MB	July 14, 2010, 11:11 am
MERRA100.prod.assim.instM_3d_	demoResc	128.23 MB	July 14, 2010, 11:15 am
MERRA100.prod.assim.instM_3d_	demoResc	126.73 MB	July 14, 2010, 11:19 am
MERRA100.prod.assim.instM_3d_	demoResc	126.46 MB	July 14, 2010, 11:24 am
MERRA100.prod.assim.instM_3d_	demoResc	124.78 MB	July 14, 2010, 11:29 am
MERRA100.prod.assim.instM_3d_	demoResc	125.34 MB	July 14, 2010, 11:33 am
MERRA100.prod.assim.instM_3d_	demoResc	126.65 MB	July 14, 2010, 11:37 am
MERRA100.prod.assim.instM_3d_	demoResc	128.45 MB	July 14, 2010, 11:41 am
MERRA100.prod.assim.instM_3d_	demoResc	127.45 MB	July 14, 2010, 11:46 am
MERRA100.prod.assim.instM_3d_	demoResc	125.85 MB	July 14, 2010, 11:50 am



# Preliminary Tests - Search

- iRODS rules and microservices can be used to assign metadata
- iRODS provides advanced search capabilities over the metadata

The screenshot displays the iRODS web interface. An 'Advanced Search' dialog box is open, showing search criteria for a file. The search results table below the dialog lists several files with their sizes and modification dates. The file 'HDFEOS\_V2.14' is highlighted, and its metadata is displayed in a table below the search results.

Attributes	Value
Name	Name of Partial Name, case sensitive
Modified Within	Any Time
Owner	Owner of the file
Resource	Resource of the file
Only	<input type="checkbox"/> Under Current Collection
Current Collection	/merra_Zone/home/public/merra/1979

Size	Date Modified
25.69 MB	July 14, 2010, 11:02 am
26.24 MB	July 14, 2010, 11:07 am
27.29 MB	July 14, 2010, 11:11 am
28.23 MB	July 14, 2010, 11:15 am
28.73 MB	July 14, 2010, 11:19 am
26.46 MB	July 14, 2010, 11:24 am
24.78 MB	July 14, 2010, 11:29 am
25.34 MB	July 14, 2010, 11:33 am
26.65 MB	July 14, 2010, 11:37 am
28.45 MB	July 14, 2010, 11:41 am
27.45 MB	July 14, 2010, 11:46 am
25.85 MB	July 14, 2010, 11:50 am

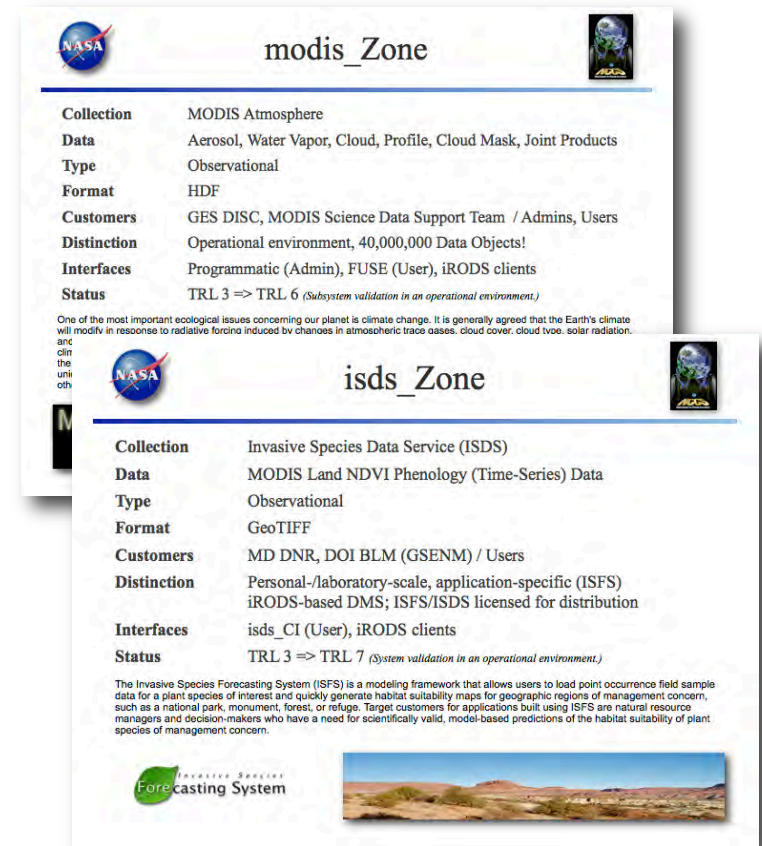
Metadata	Value
variables	like Surface Geopotential
checksum	= 67abb8ca2e184ad2f9e92
Name	Op Value
Name	Op Value
Name	Op Value

hdfeosversion	HDFEOS_V2.14
dimensions	TIME:EOSGRID = 1, YDim:EOSGRID = 144, XDim:EOSGRID = 288, Height:EOSGRID = 42
conventions	CF-1.0
contact	http://gmao.gsfc.nasa.gov/
comment	GEOS-5.2.0
checksum	91ddec7eee867abb8ca2e184ad2f9e92

# Preliminary Tests – Observational Data

- Developed an iRODS data grid that publishes Moderate Resolution Imaging Spectroradiometer (MODIS) observational data
  - 54 million registered files, 630 TB of data, and over 300 million defined metadata values
- Developed an iRODS data grid that focuses on a small-scale, multi-product, application-specific data service
  - The Invasive Species Data Service (ISDS) manages a collection of MODIS data products for ecological forecasting applications



The image shows two screenshots of iRODS metadata pages. The top screenshot is for 'modis\_Zone' and the bottom is for 'isds\_Zone'. Both pages include a NASA logo, a title, and a table of metadata. The 'modis\_Zone' page also includes a small image of a satellite and a paragraph of text about climate change. The 'isds\_Zone' page includes a small image of a globe and a paragraph of text about the Invasive Species Forecasting System (ISFS).

Collection	MODIS Atmosphere
Data	Aerosol, Water Vapor, Cloud, Profile, Cloud Mask, Joint Products
Type	Observational
Format	HDF
Customers	GES DISC, MODIS Science Data Support Team / Admins, Users
Distinction	Operational environment, 40,000,000 Data Objects!
Interfaces	Programmatic (Admin), FUSE (User), iRODS clients
Status	TRL 3 => TRL 6 <i>(Subsystem validation in an operational environment.)</i>

One of the most important ecological issues concerning our planet is climate change. It is generally agreed that the Earth's climate will modify in response to radiative forcing induced by changes in atmospheric trace gases, cloud cover, cloud type, solar radiation, and other factors.


Collection	Invasive Species Data Service (ISDS)
Data	MODIS Land NDVI Phenology (Time-Series) Data
Type	Observational
Format	GeoTIFF
Customers	MD DNR, DOI BLM (GSENM) / Users
Distinction	Personal-/laboratory-scale, application-specific (ISFS) iRODS-based DMS; ISFS/ISDS licensed for distribution
Interfaces	isds_CI (User), iRODS clients
Status	TRL 3 => TRL 7 <i>(System validation in an operational environment.)</i>

The Invasive Species Forecasting System (ISFS) is a modeling framework that allows users to load point occurrence field sample data for a plant species of interest and quickly generate habitat suitability maps for geographic regions of management concern, such as a national park, monument, forest, or refuge. Target customers for applications built using ISFS are natural resource managers and decision-makers who have a need for scientifically valid, model-based predictions of the habitat suitability of plant species of management concern.




# Preliminary Tests – Simulation Data

- Developed an iRODS data grid that manages Modern Era Retrospective-Analysis for Research and Applications (MERRA) simulation data
  - 360 files, 47 GB of data, and 4000 metadata values
- Developed an iRODS data grid that publishes Year of Tropical Convection (YOTC) data sets
  - 134,000 files, 12 TB of data, and 400,000 metadata values




## merra\_Zone




---

<b>Collection</b>	Modern Era Retrospective-Analysis for Research and Applications (MERRA)
<b>Data</b>	Monthly products from the past 15 years
<b>Type</b>	Observational/Simulation
<b>Format</b>	NETCDF
<b>Customers</b>	NCCS, GES DISC, ESG, Nebula / Admins, Managers
<b>Distinction</b>	merra_Zone @ NCCS + merra_Zone @ Nebula
<b>Interfaces</b>	merra_CI (Admin), iRODS clients
<b>Status</b>	TRL 3 => TRL 5 <i>(System validation in a relevant test environment.)</i>

Retrospective-analyses (or reanalyses) have been a critical tool in studying weather and climate variability for the last 15 years.




## yotc\_Zone



---

<b>Collection</b>	Year of Tropical Convection (YOTC)
<b>Data</b>	Satellite, in-situ and simulation/prediction model data sets
<b>Type</b>	Observational/Simulation
<b>Format</b>	NETCDF
<b>Customers</b>	NCCS / Admins, Users
<b>Distinction</b>	Operational environment, iRODS-mediated archive management
<b>Interfaces</b>	yotc_CI (Admin), FUSE (User), iRODS clients
<b>Status</b>	TRL 3 => TRL 7 <i>(System validation in an operational environment.)</i>

The realistic representation of tropical convection in our global atmospheric models is a long-standing grand challenge for numerical weather forecasts and global climate predictions. To address the challenge of tropical convection, collaborative organizations from around the world have proposed a year of coordinated observing, modeling and forecasting of organized tropical convection and its influences on predictability. This effort is intended to exploit the vast amounts of existing and emerging observations, the expanding computational resources and the development of new, high-resolution modeling frameworks, with the objective of advancing the characterization, diagnosis, modeling, parameterization and prediction of multi-scale convective/dynamic interactions, including the two-way interaction between tropical and extra-tropical weather/climate.





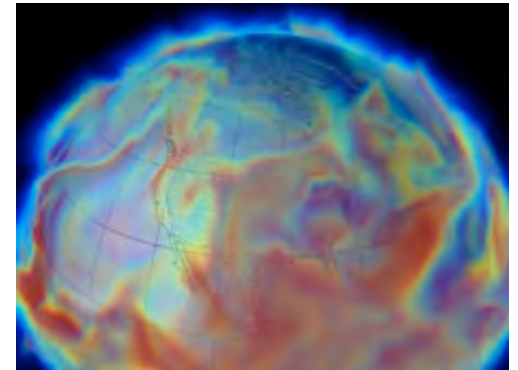


# Preliminary Tests – Federation

- Tested and evaluated iRODS data federation
  - Federated the YOTC and MODIS grids to simulate the union of observational and simulation data
- Explored the integrated management of observational and simulation data
  - Implemented an interface that enables comingling of remote and local observational and simulation data for advanced scientific study

# Preliminary Results

- iRODS is a promising technology for exposing services for data management, publication, and analysis
- The iRODS catalog (ICAT) demonstrated adequate scaling for data registration
  - Optimization desired for searching huge datasets
- Good collaboration with the iRODS development team
- NCCS has made the decision to operationalize iRODS





# New Goals

- *IPCC / AR5*
  - Provide the data management services and analytical tools necessary to support the publication requirements of the Intergovernmental Panel on Climate Change (IPCC).
- *Observation/Simulation Data Integration*
  - Bring the climate modeling and observational communities together to work toward the goal of integrating model outputs and observational data
- *Next Generation HEC Requirements for Modeling and Assimilation*
  - Contribute emerging technologies to address computing requirements for Earth system modeling that will increase significantly in the coming years

# Questions

