The NASA Center for Climate Simulation DMS Project

Initial Deployment and Preliminary Experiences with an iRODS-Based Data Management Service

Glenn Tamkin

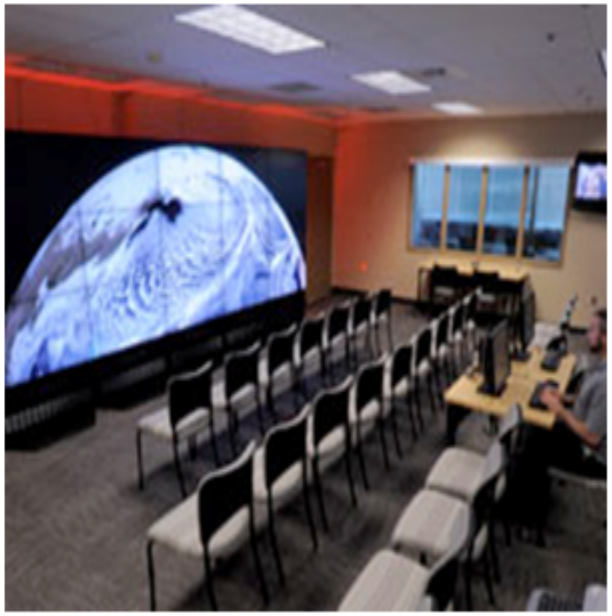
David Fladung, Scott Sinno, Roger Gill, Dave Ripley, Savannah Strong, & John Schnase

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 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

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

Technology - integrated Rule-Oriented Data System (iRODS)

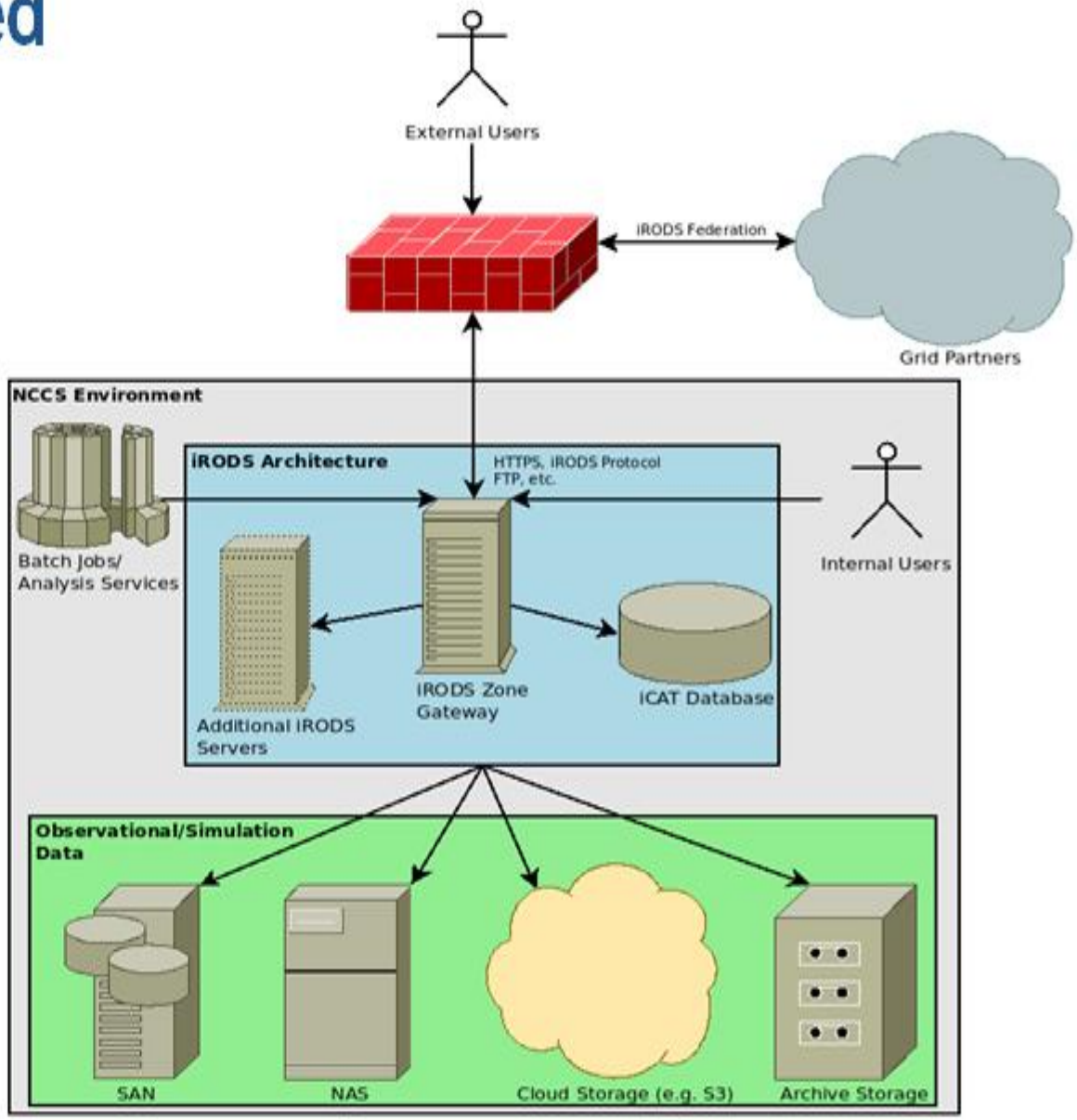
- Open source data grid software developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina
- Targets large repositories and digital preservation
- Supports the federation of independent, distributed collections
- Supports server-side workflows that are implemented by chaining execution rules together based on data policies
- Includes features such as domain-specific validation, automatic replication, and digital signature/checksum computation
- Validates assertions about data such as integrity and authenticity

Technology - integrated Rule-Oriented Data System (iRODS)

- Open source data grid software developed by the Data Intensive Cyber Environments (DICE) group, University of North Carolina
- Targets large repositories and digital preservation
- Supports the federation of independent, distributed collections
- Supports server-side workflows that are implemented by chaining execution rules together based on data policies
- Includes features such as domain-specific validation, automatic replication, and digital signature/checksum computation
- Validates assertions about data such as integrity and authenticity

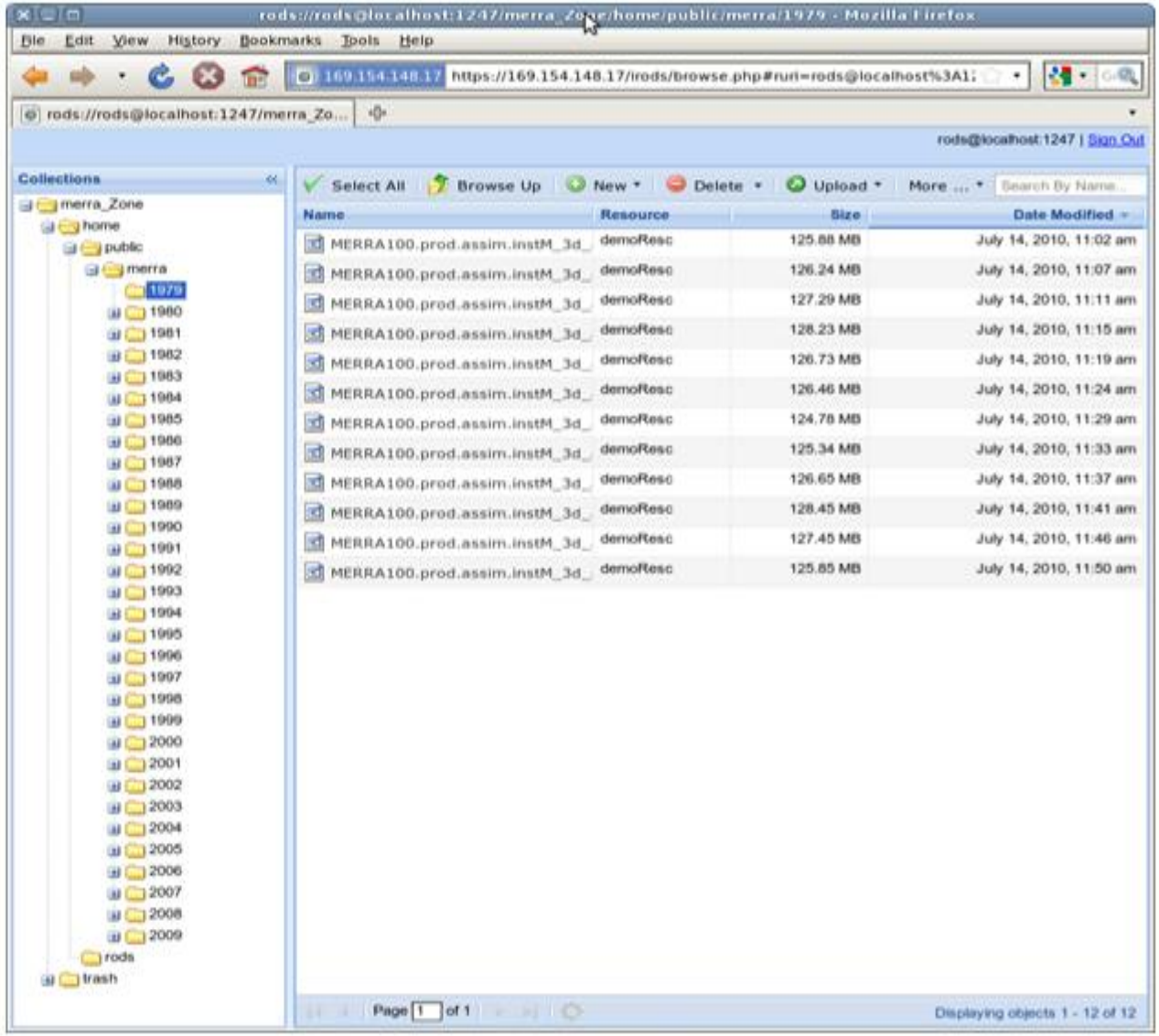
NCCS Test Bed

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



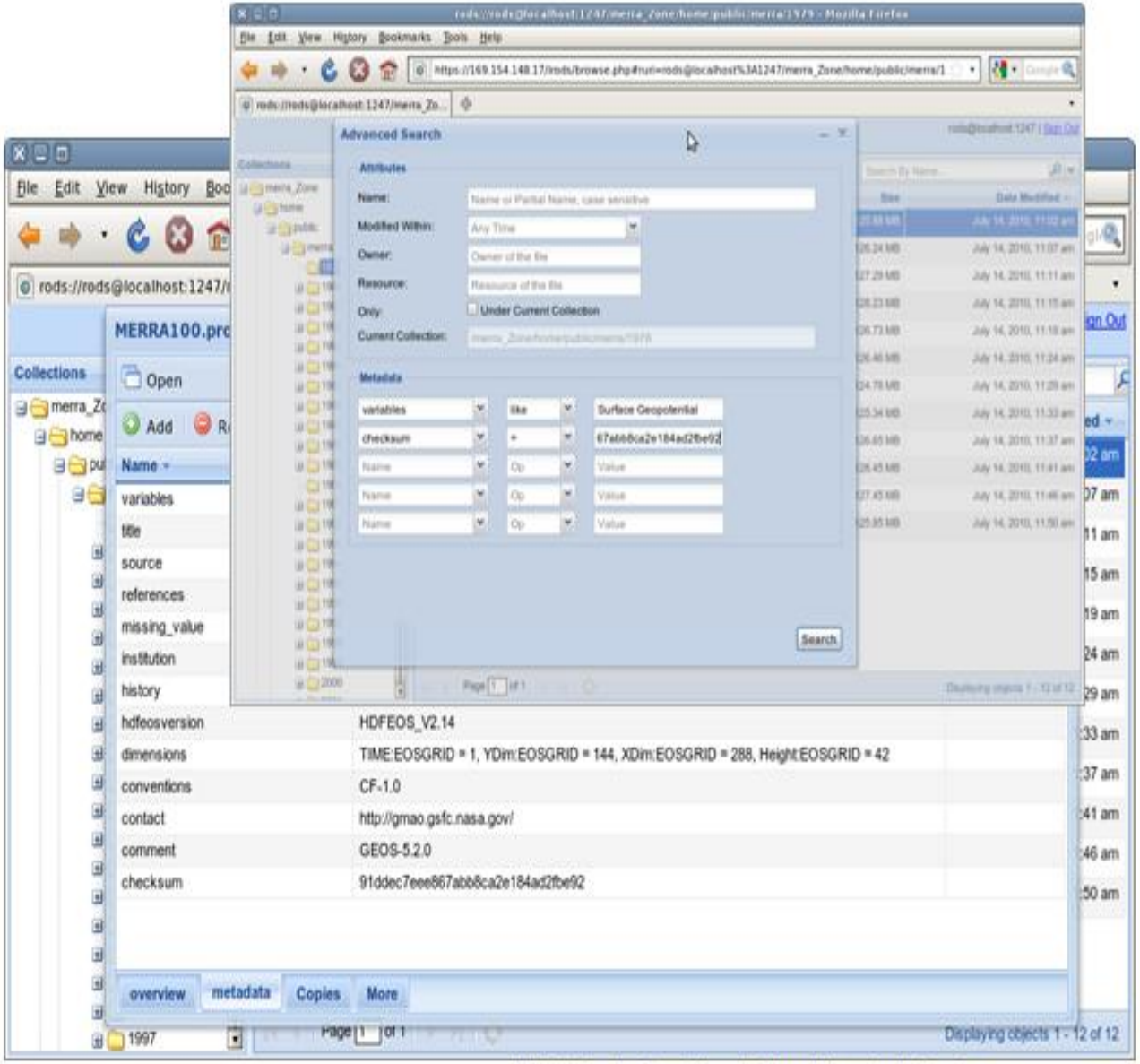
Preliminary Tests – Ingest/Registration

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



Preliminary Tests - Search

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



Preliminary Tests – Observational Data

- Developed an iRODS data grid that publishes Moderate Resolution Imaging Spectro-radiometer (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

modis_Zone

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 *(Subsystem validation in an operational environment.)*

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,

isds_Zone

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 *(System validation in an operational environment.)*

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.

Forecasting System

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

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

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

yotc_Zone

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

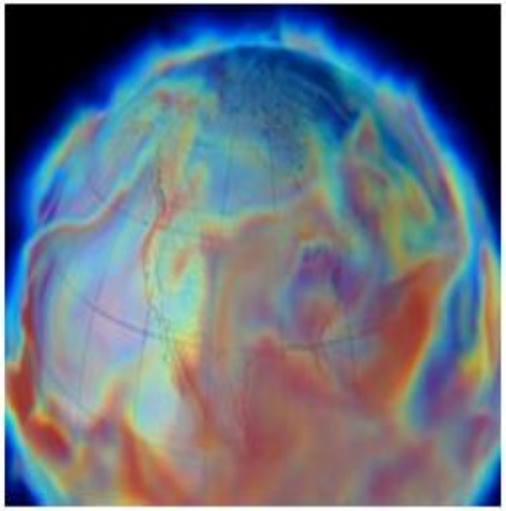
The realistic representation of tropical convection in our global atmospheric models is a long-standing grand challenge for numerical weather forecasts and global climate predictions.To address the challenge of tropical convection, collaborative organizations from around the world have proposed a year 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

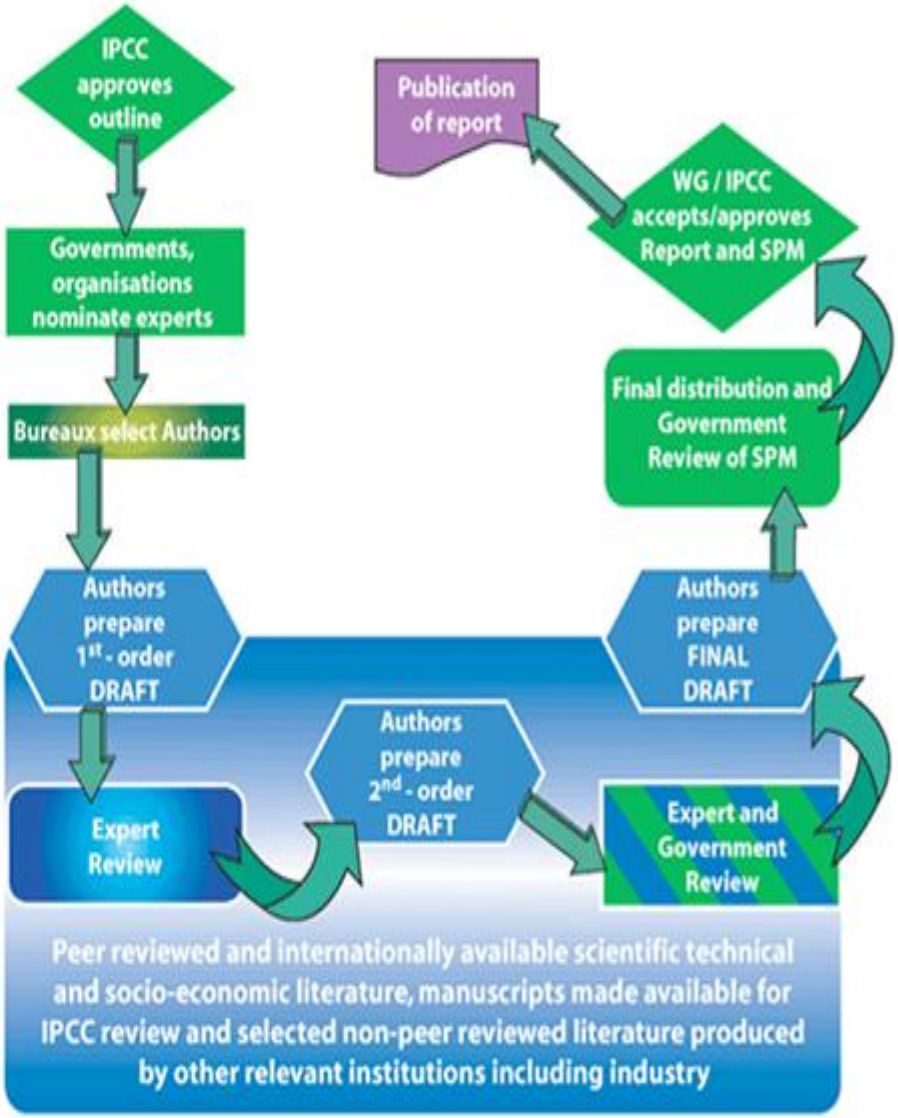
Preliminary Results

- 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



Next Steps – Data Services

- *NCCS Data Management System*
  - Use iRODS to enable all data- and analysis-related aspects of AR5 research
  - Define and implement the relevant OAIS-compliant (Open Archival Information System) data policies that will govern the AR5 archive through iRODS
  - Develop interfaces and archive administration and end-user data management tools



Next Steps – Data Services

- *Earth System Grid*
  - Use iRODS to address the publication requirements of IPCC/AR5, which requires integration of ESG
- *Data Integration*
  - Create the framework for integrating observational and simulation data
- *Cloud Provisioning*
  - Incorporate cloud computing to better serve the extended research community

Questions