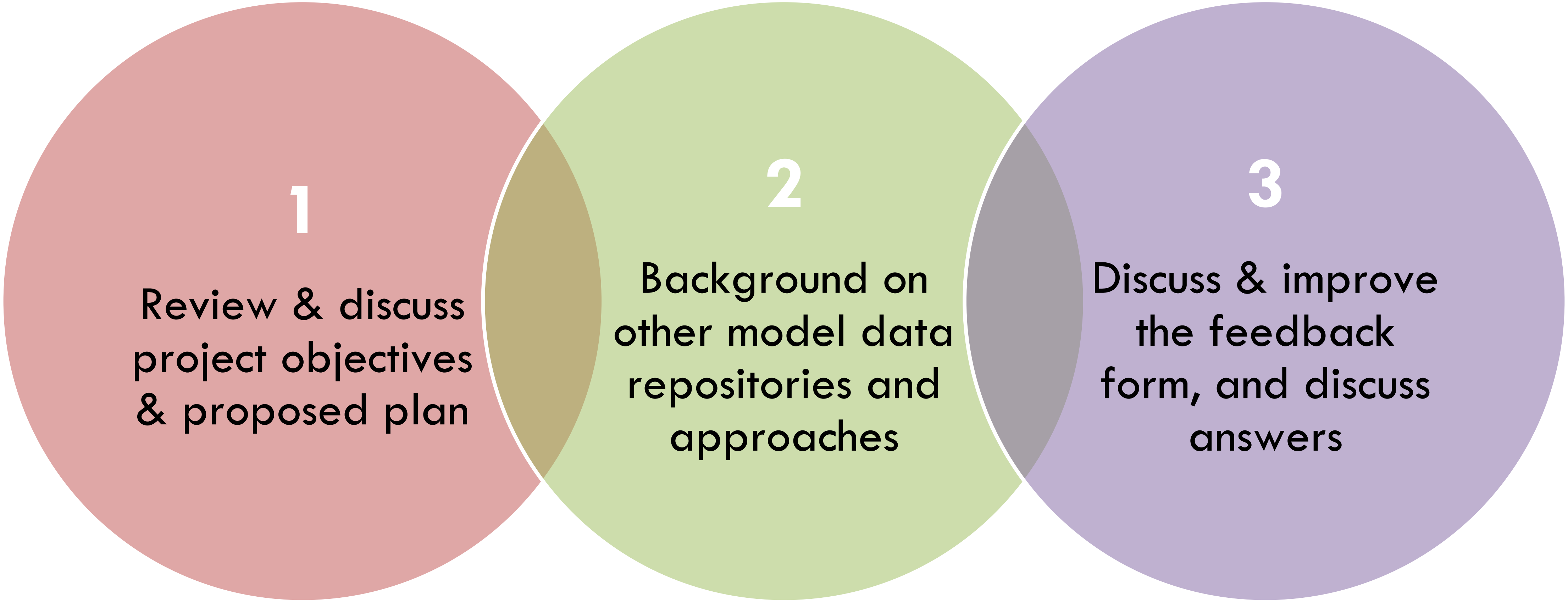
ESS-DIVE Webinar:
Model Data Archiving

Maegen Simmonds, William Riley,
Charuleka Varadharajan
November 22, 2019

Webinar Goals



1

Review & discuss project objectives & proposed plan

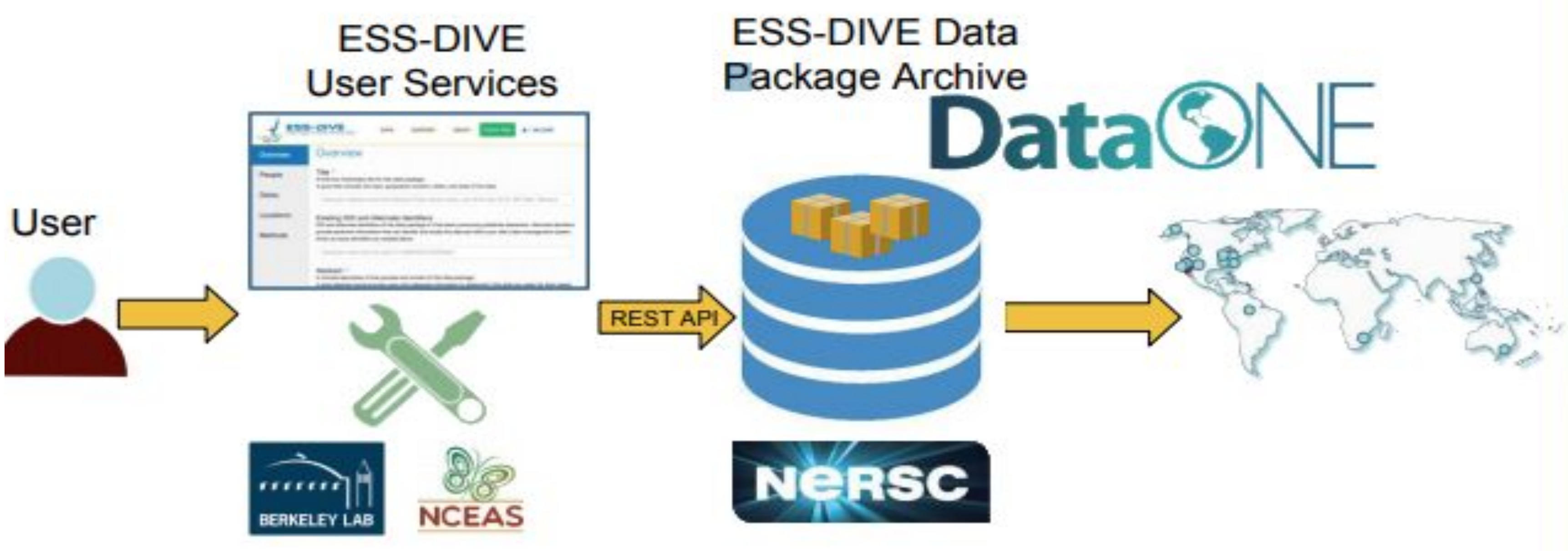
2

Background on other model data repositories and approaches

3

Discuss & improve the feedback form, and discuss answers

ESS-DIVE: Current Functionality for Model Data Archiving

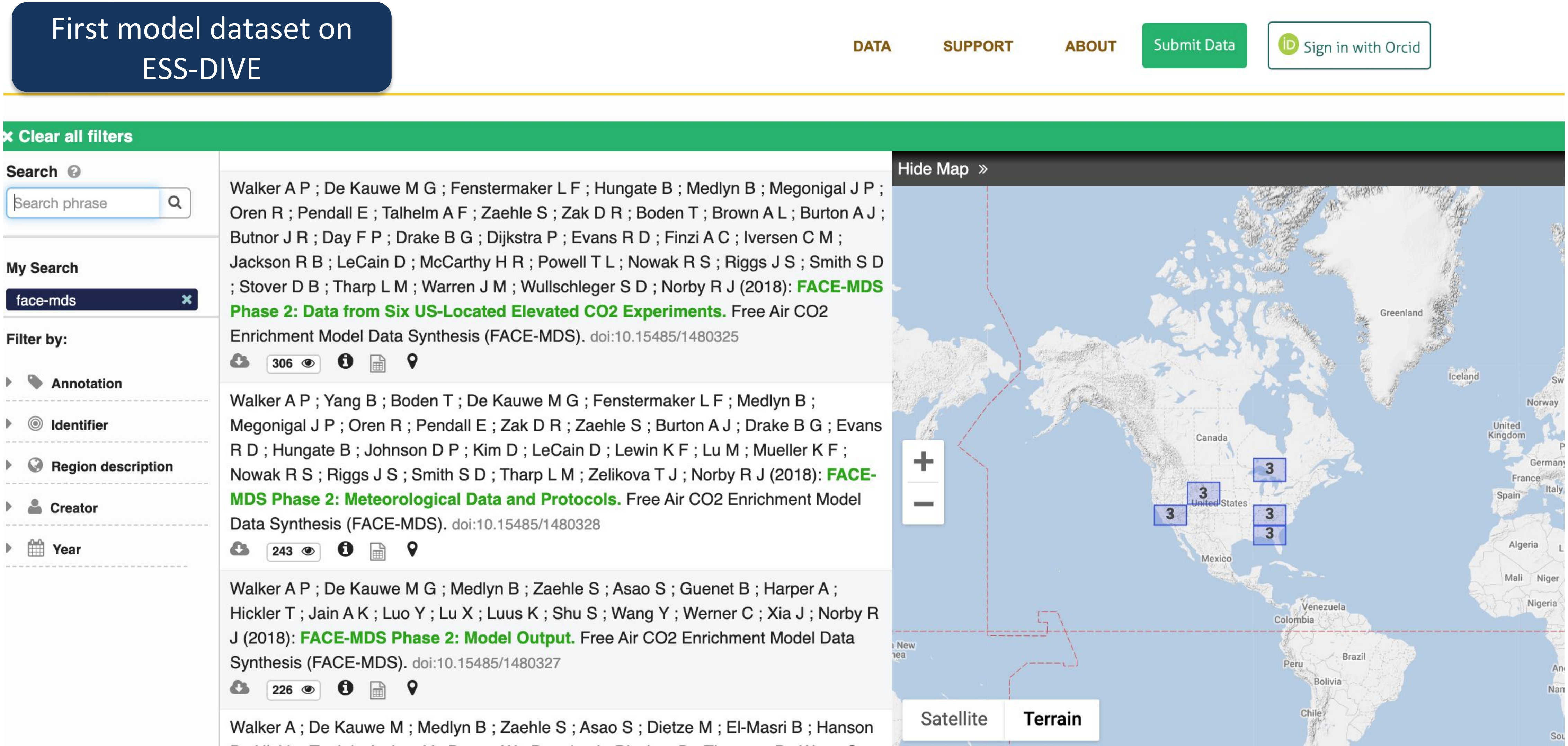


CREATE SINGLE OR MULTIPLE DATA PACKAGES

GOALS & PROPOSED PLAN

ESS-DIVE: Current Functionality for Model Data Archiving

First model dataset on ESS-DIVE



Walker A P ; De Kauwe M G ; Fenstermaker L F ; Hungate B ; Medlyn B ; Megonigal J P ; Oren R ; Pendall E ; Talhelm A F ; Zaehle S ; Zak D R ; Boden T ; Brown A L ; Burton A J ; Butnor J R ; Day F P ; Drake B G ; Dijkstra P ; Evans R D ; Finzi A C ; Iversen C M ; Jackson R B ; LeCain D ; McCarthy H R ; Powell T L ; Nowak R S ; Riggs J S ; Smith S D ; Stover D B ; Tharp L M ; Warren J M ; Wullschleger S D ; Norby R J (2018): **FACE-MDS Phase 2: Data from Six US-Located Elevated CO2 Experiments.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480325

306

Walker A P ; Yang B ; Boden T ; De Kauwe M G ; Fenstermaker L F ; Medlyn B ; Megonigal J P ; Oren R ; Pendall E ; Zak D R ; Zaehle S ; Burton A J ; Drake B G ; Evans R D ; Hungate B ; Johnson D P ; Kim D ; LeCain D ; Lewin K F ; Lu M ; Mueller K F ; Nowak R S ; Riggs J S ; Smith S D ; Tharp L M ; Zelikova T J ; Norby R J (2018): **FACE-MDS Phase 2: Meteorological Data and Protocols.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480328

243

Walker A P ; De Kauwe M G ; Medlyn B ; Zaehle S ; Asao S ; Guenet B ; Harper A ; Hickler T ; Jain A K ; Luo Y ; Lu X ; Luus K ; Shu S ; Wang Y ; Werner C ; Xia J ; Norby R J (2018): **FACE-MDS Phase 2: Model Output.** Free Air CO2 Enrichment Model Data Synthesis (FACE-MDS). doi:10.15485/1480327

226

Walker A ; De Kauwe M ; Medlyn B ; Zaehle S ; Asao S ; Dietze M ; El-Masri B ; Hanson

ESS-DIVE: Current Functionality for Model Data Archiving

Opportunities for improvement!

ESS-DIVE: Current Functionality for Model Data Archiving

Problem Statement

Model data storage is limited

- currently only archiving a limited set of small-sized model outputs
- infrastructure limitations on data size:
    - upload limits 2GB/file on portal and 1GB/file on API
    - architecture limits how much data ESS-DIVE can store and serve
- web interface not the best tool for uploading/downloading large datasets
- API helps but there are still physical limitations

No community consensus yet on what to archive, standards, storage space needed, etc.

Main objectives of this project

Assess:

- What model data should be archived, purposes of storage, storage capacity needed
- Best approach to store data

Deliverable:

- White paper describing data storage capabilities ESS modeling projects need based on community feedback and a few use-cases, and potential storage options

GOALS & PROPOSED PLAN

Timeline

Feb 2019

CI meeting on model data

Nov 2019

ESS-DIVE webinar kick-off of model data white paper project

Model Data Archiving Feedback Form

Dec-Apr 2020

Deeper dive- one-on-one conversations

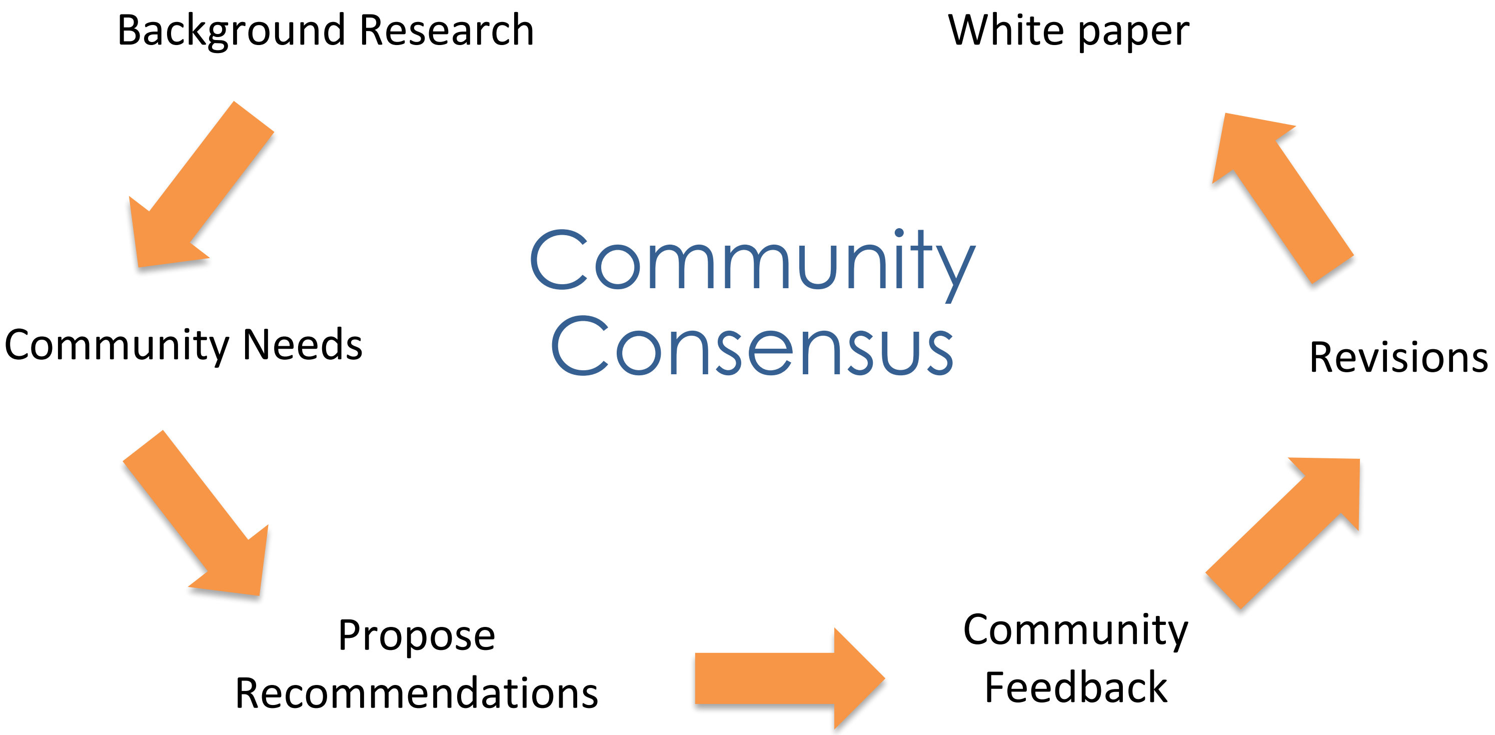
Feb/Mar 2020

ESS-DIVE Virtual Model Data Workshop

May 2020

Finalize white paper

Our Process



Background Research

Community Needs

Propose Recommendations

Community Feedback

Revisions

White paper

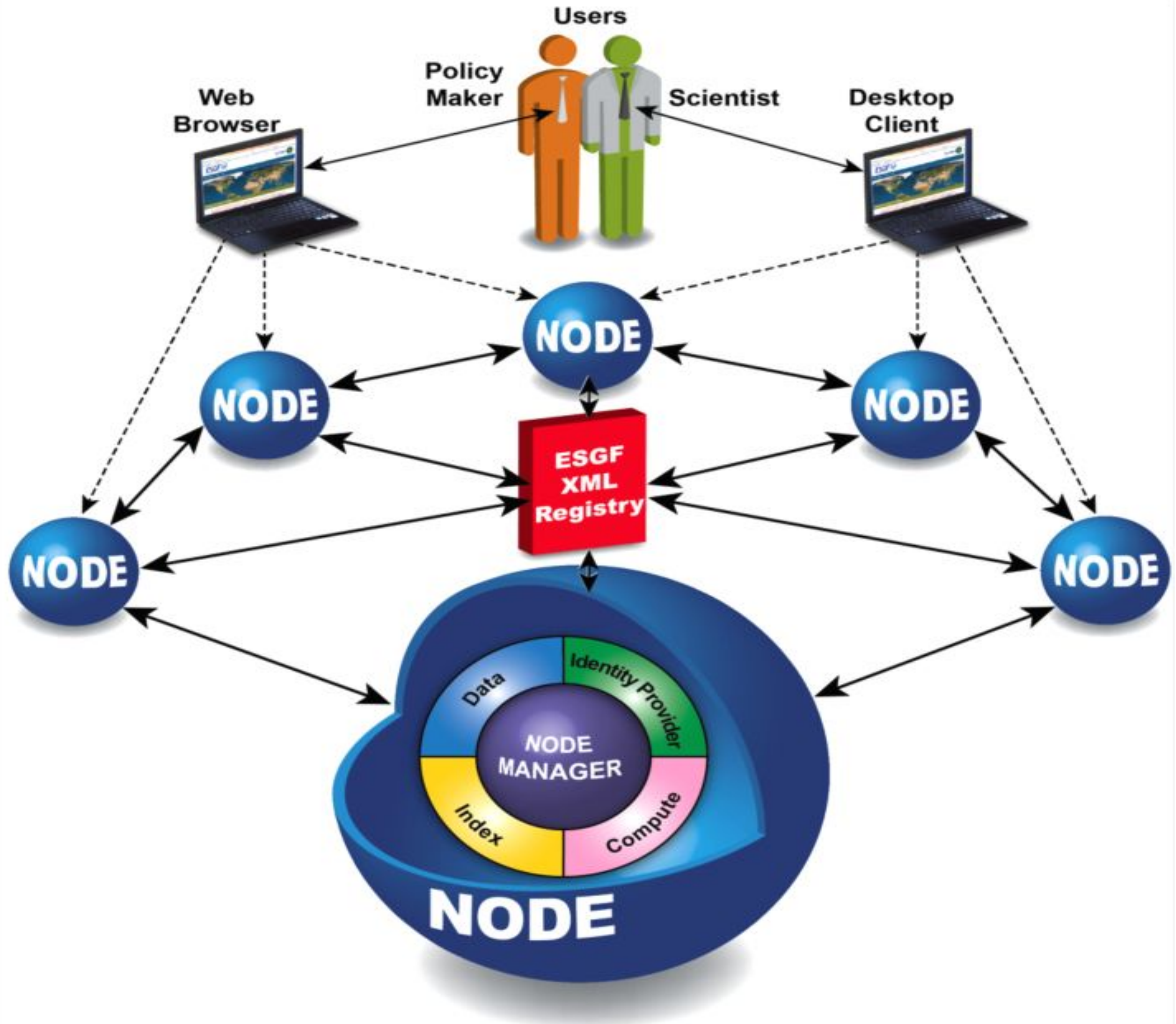
Community Consensus

Examples of cloud-based web storage for model & observational data

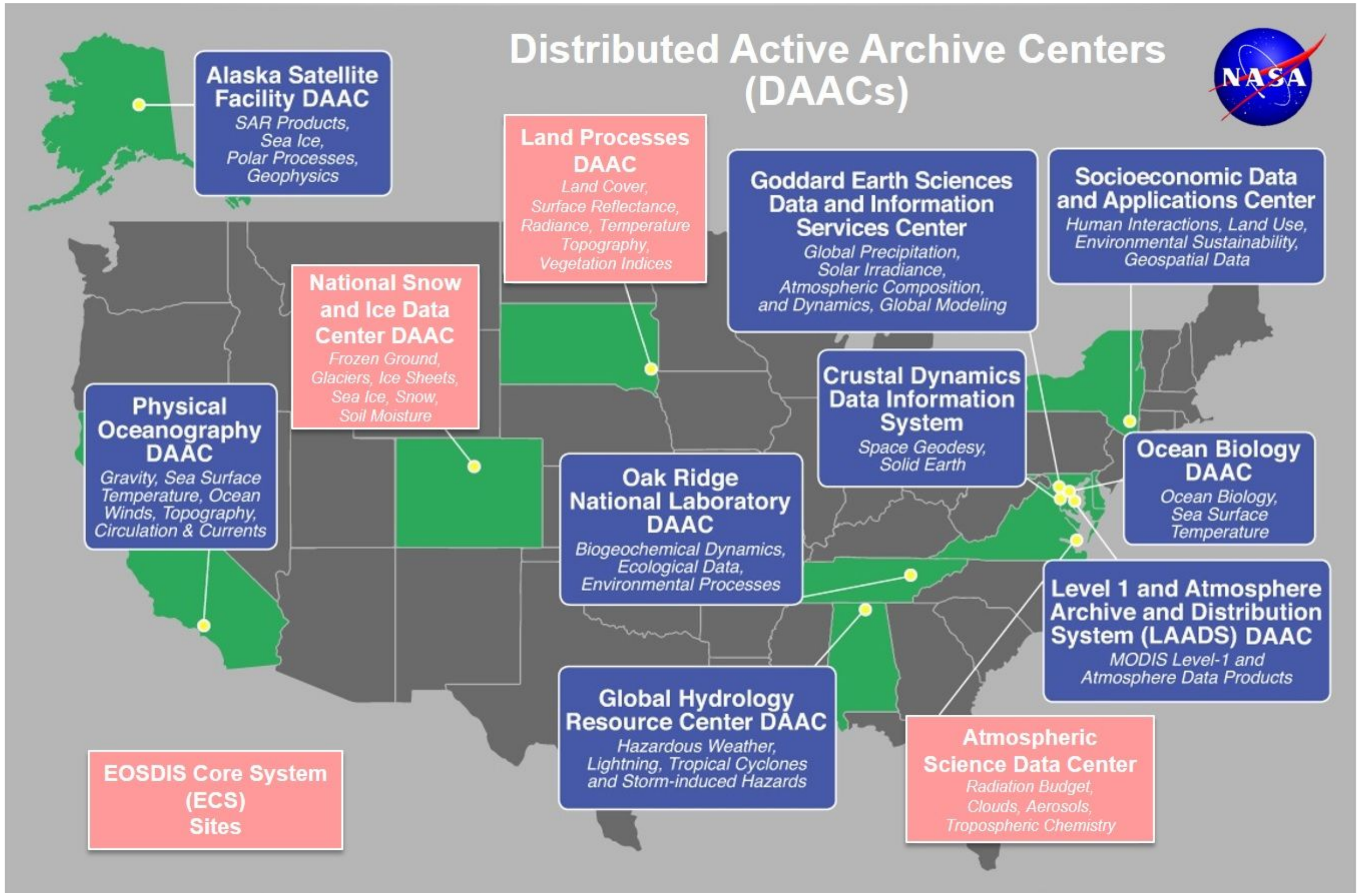
- Earth System Grid Federation (ESGF)
- NASA's Earth Observing System Data and Information System (EODIS)
- NCAR's Earth Observatory Laboratory (EOL) Data Archive
- NCAR's Research Data Archive (RDA)
- CYVERSE

ESGF

- Federated system for storing and serving data from multiple locations and sources
- Developed for sharing climate model (CMIP) data
- Currently stores some ESS modeling data
- NOT guaranteed to be a long-term host of data, to follow digital library standards, or to assign DOIs to data



EOSDIS


Distributed Active Archive Centers (DAACs)
NASA
Alaska Satellite Facility DAAC
SAR Products, Sea Ice, Polar Processes, Geophysics
Land Processes DAAC
Land Cover, Surface Reflectance, Radiance, Temperature Topography, Vegetation Indices
Goddard Earth Sciences Data and Information Services Center
Global Precipitation, Solar Irradiance, Atmospheric Composition, and Dynamics, Global Modeling
Socioeconomic Data and Applications Center
Human Interactions, Land Use, Environmental Sustainability, Geospatial Data
National Snow and Ice Data Center DAAC
Frozen Ground, Glaciers, Ice Sheets, Sea Ice, Snow, Soil Moisture
Physical Oceanography DAAC
Gravity, Sea Surface Temperature, Ocean Winds, Topography, Circulation & Currents
Crustal Dynamics Data Information System
Space Geodesy, Solid Earth
Ocean Biology DAAC
Ocean Biology, Sea Surface Temperature
Oak Ridge National Laboratory DAAC
Biogeochemical Dynamics, Ecological Data, Environmental Processes
Level 1 and Atmosphere Archive and Distribution System (LAADS) DAAC
MODIS Level-1 and Atmosphere Data Products
Global Hydrology Resource Center DAAC
Hazardous Weather, Lightning, Tropical Cyclones and Storm-induced Hazards
Atmospheric Science Data Center
Radiation Budget, Clouds, Aerosols, Tropospheric Chemistry
EOSDIS Core System (ECS) Sites

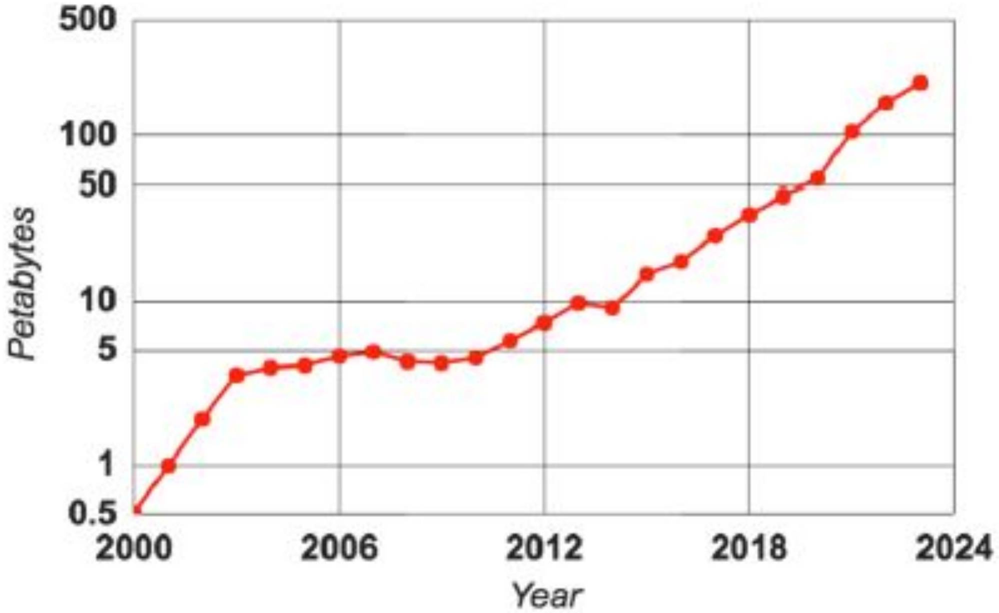
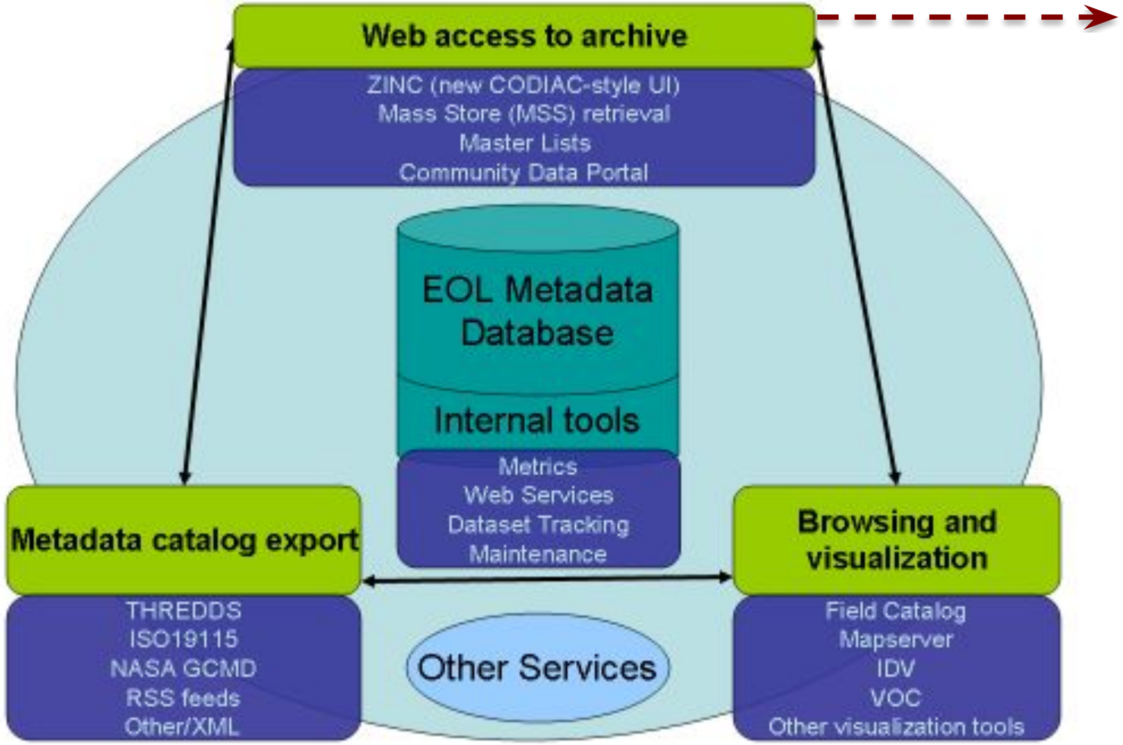


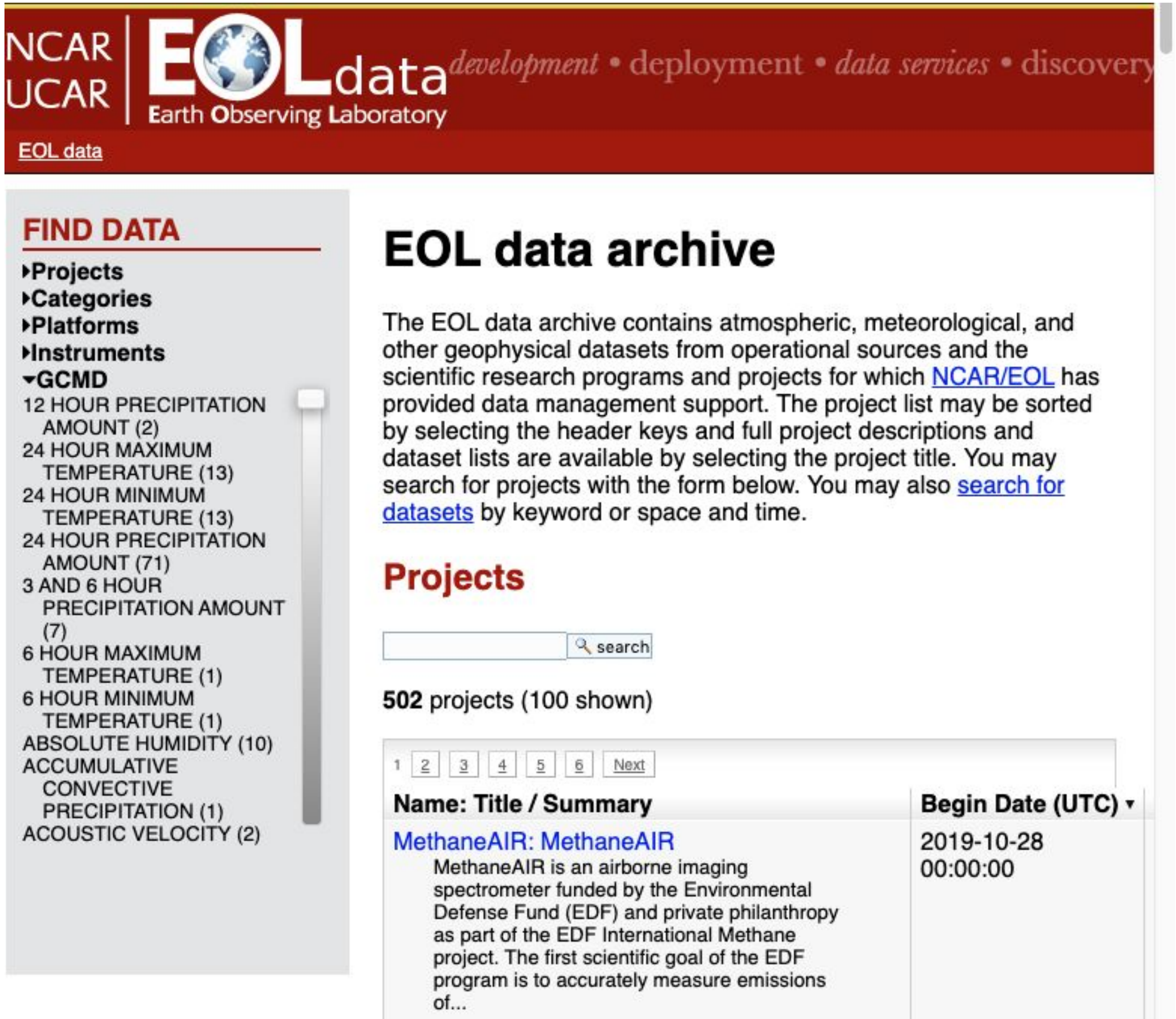
FIGURE 1. HISTORICAL AND PROJECTED CUMULATIVE ARCHIVE VOLUME IN EOSDIS. (YEARS RUN FROM OCTOBER TO SEPTEMBER.)

BACKGROUND

Earth Observatory Lab (EOL) Data Archive

EOL Metadata Database and Cyberinfrastructure (EMDAC)




NCAR
UCAR
EOLdata
Earth Observing Laboratory
development • deployment • data services • discovery
EOL data

FIND DATA

- Projects
- Categories
- Platforms
- Instruments
- GCMD

12 HOUR PRECIPITATION AMOUNT (2)
24 HOUR MAXIMUM TEMPERATURE (13)
24 HOUR MINIMUM TEMPERATURE (13)
24 HOUR PRECIPITATION AMOUNT (71)
3 AND 6 HOUR PRECIPITATION AMOUNT (7)
6 HOUR MAXIMUM TEMPERATURE (1)
6 HOUR MINIMUM TEMPERATURE (1)
ABSOLUTE HUMIDITY (10)
ACCUMULATIVE CONVECTIVE PRECIPITATION (1)
ACOUSTIC VELOCITY (2)

EOL data archive

The EOL data archive contains atmospheric, meteorological, and other geophysical datasets from operational sources and the scientific research programs and projects for which NCAR/EOL has provided data management support. The project list may be sorted by selecting the header keys and full project descriptions and dataset lists are available by selecting the project title. You may search for projects with the form below. You may also search for datasets by keyword or space and time.

Projects

search

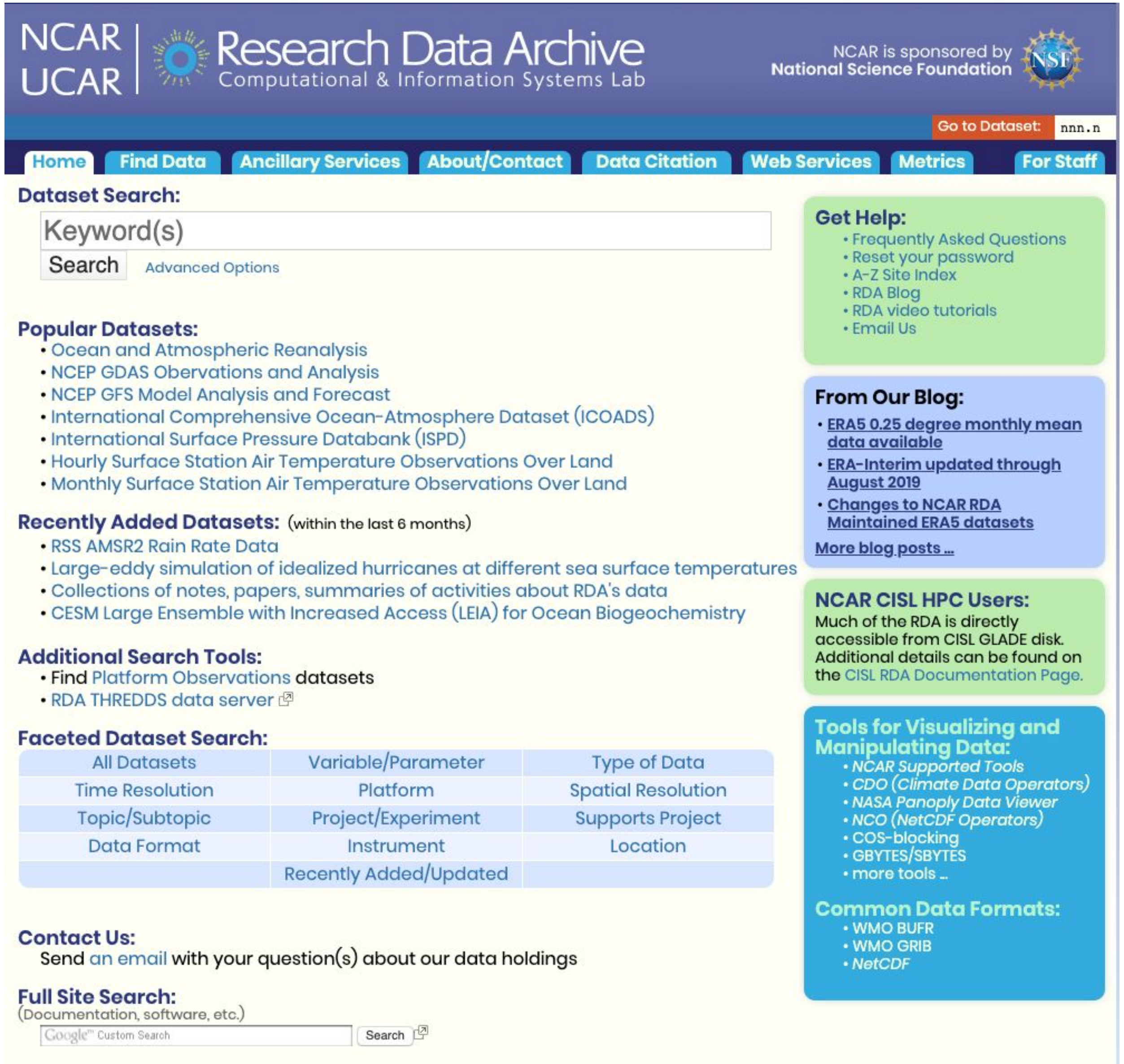
502 projects (100 shown)

1 2 3 4 5 6 Next

Name: Title / Summary	Begin Date (UTC) ▾
MethaneAIR: MethaneAIR MethaneAIR is an airborne imaging spectrometer funded by the Environmental Defense Fund (EDF) and private philanthropy as part of the EDF International Methane project. The first scientific goal of the EDF program is to accurately measure emissions of...	2019-10-28 00:00:00

BACKGROUND

Research Data Archive (RDA)



Dataset Search:

Popular Datasets:

- Ocean and Atmospheric Reanalysis
- NCEP GDAS Obervations and Analysis
- NCEP GFS Model Analysis and Forecast
- International Comprehensive Ocean-Atmosphere Dataset (ICOADS)
- International Surface Pressure Databank (ISPD)
- Hourly Surface Station Air Temperature Observations Over Land
- Monthly Surface Station Air Temperature Observations Over Land

Recently Added Datasets: (within the last 6 months)

- RSS AMSR2 Rain Rate Data
- Large-eddy simulation of idealized hurricanes at different sea surface temperatures
- Collections of notes, papers, summaries of activities about RDA's data
- CESM Large Ensemble with Increased Access (LEIA) for Ocean Biogeochemistry

Additional Search Tools:

- Find Platform Observations datasets
- RDA THREDDS data server

Faceted Dataset Search:

All Datasets	Variable/Parameter	Type of Data
Time Resolution	Platform	Spatial Resolution
Topic/Subtopic	Project/Experiment	Supports Project
Data Format	Instrument	Location
	Recently Added/Updated	

Contact Us:

Send an email with your question(s) about our data holdings

Full Site Search:

(Documentation, software, etc.)

Get Help:

- Frequently Asked Questions
- Reset your password
- A-Z Site Index
- RDA Blog
- RDA video tutorials
- Email Us

From Our Blog:

- ERA5 0.25 degree monthly mean data available
- ERA-Interim updated through August 2019
- Changes to NCAR RDA Maintained ERA5 datasets

More blog posts ...

NCAR CISL HPC Users:

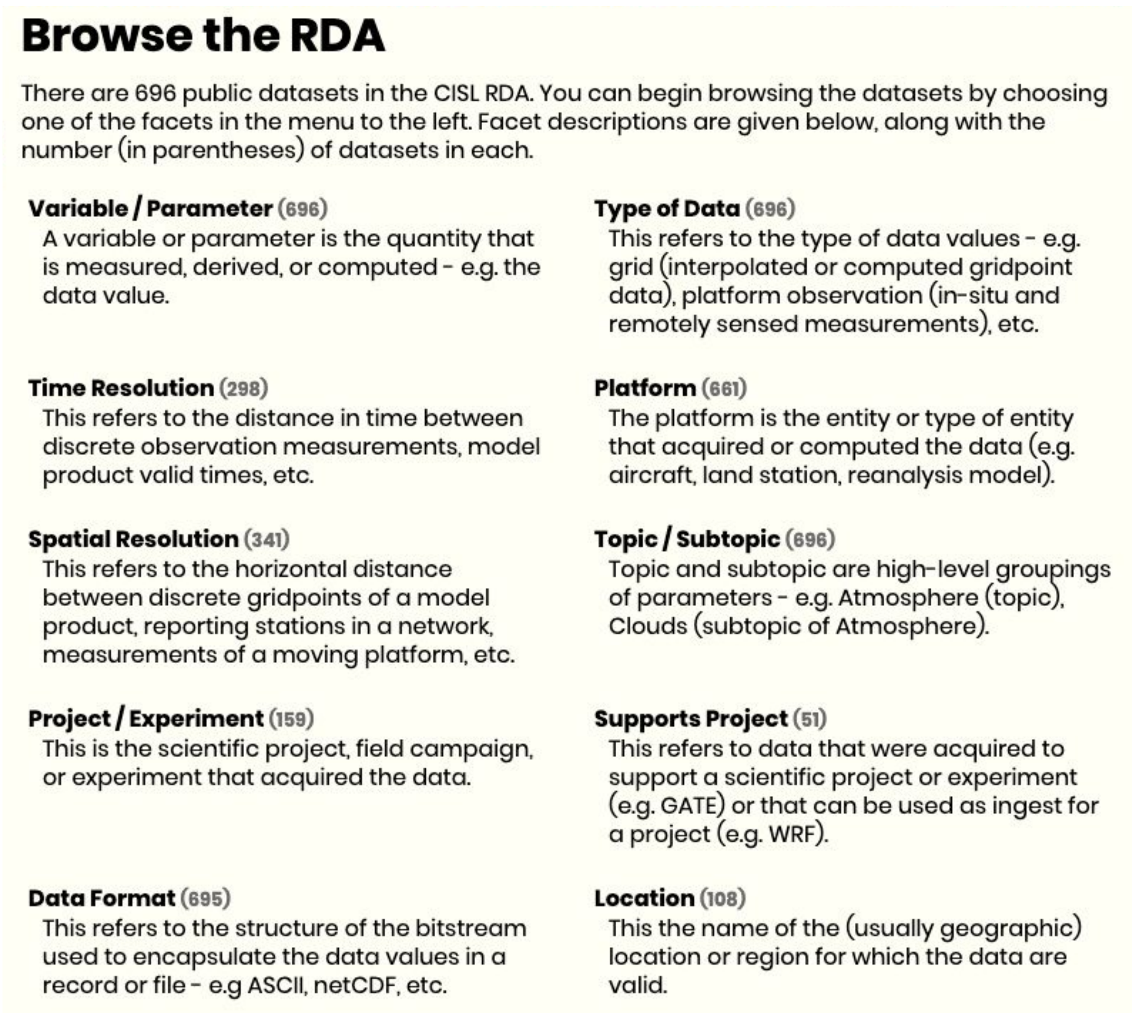
Much of the RDA is directly accessible from CISL GLADE disk. Additional details can be found on the CISL RDA Documentation Page.

Tools for Visualizing and Manipulating Data:

- *NCAR Supported Tools*
- *CDO (Climate Data Operators)*
- *NASA Panoply Data Viewer*
- *NCO (NetCDF Operators)*
- COS-blocking
- GBYTES/SBYTES
- more tools ...

Common Data Formats:

- WMO BUFR
- WMO GRIB
- *NetCDF*



Browse the RDA

There are 696 public datasets in the CISL RDA. You can begin browsing the datasets by choosing one of the facets in the menu to the left. Facet descriptions are given below, along with the number (in parentheses) of datasets in each.

Variable / Parameter (696)
A variable or parameter is the quantity that is measured, derived, or computed - e.g. the data value.

Time Resolution (298)
This refers to the distance in time between discrete observation measurements, model product valid times, etc.

Spatial Resolution (341)
This refers to the horizontal distance between discrete gridpoints of a model product, reporting stations in a network, measurements of a moving platform, etc.

Project / Experiment (159)
This is the scientific project, field campaign, or experiment that acquired the data.

Data Format (695)
This refers to the structure of the bitstream used to encapsulate the data values in a record or file - e.g ASCII, netCDF, etc.

Type of Data (696)
This refers to the type of data values - e.g. grid (interpolated or computed gridpoint data), platform observation (in-situ and remotely sensed measurements), etc.

Platform (661)
The platform is the entity or type of entity that acquired or computed the data (e.g. aircraft, land station, reanalysis model).

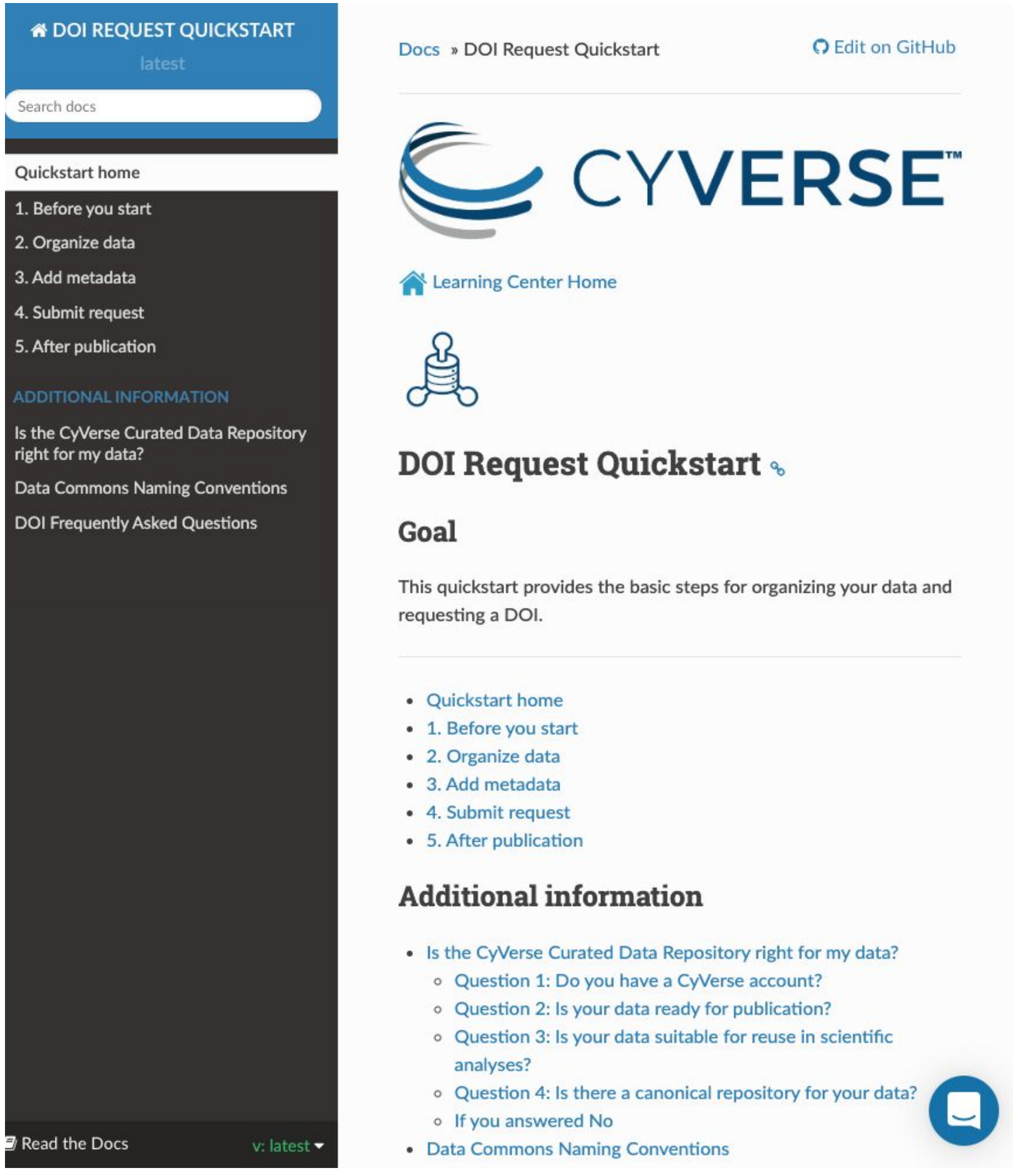
Topic / Subtopic (696)
Topic and subtopic are high-level groupings of parameters - e.g. Atmosphere (topic), Clouds (subtopic of Atmosphere).

Supports Project (51)
This refers to data that were acquired to support a scientific project or experiment (e.g. GATE) or that can be used as ingest for a project (e.g. WRF).

Location (108)
This the name of the (usually geographic) location or region for which the data are valid.

BACKGROUND

CYVERSE (originally iPlant)



DOI REQUEST QUICKSTART
latest
Search docs

Quickstart home
1. Before you start
2. Organize data
3. Add metadata
4. Submit request
5. After publication

ADDITIONAL INFORMATION
Is the CyVerse Curated Data Repository right for my data?
Data Commons Naming Conventions
DOI Frequently Asked Questions

Read the Docs v: latest

Docs » DOI Request Quickstart Edit on GitHub

CYVERSE™

Learning Center Home

DOI Request Quickstart

Goal

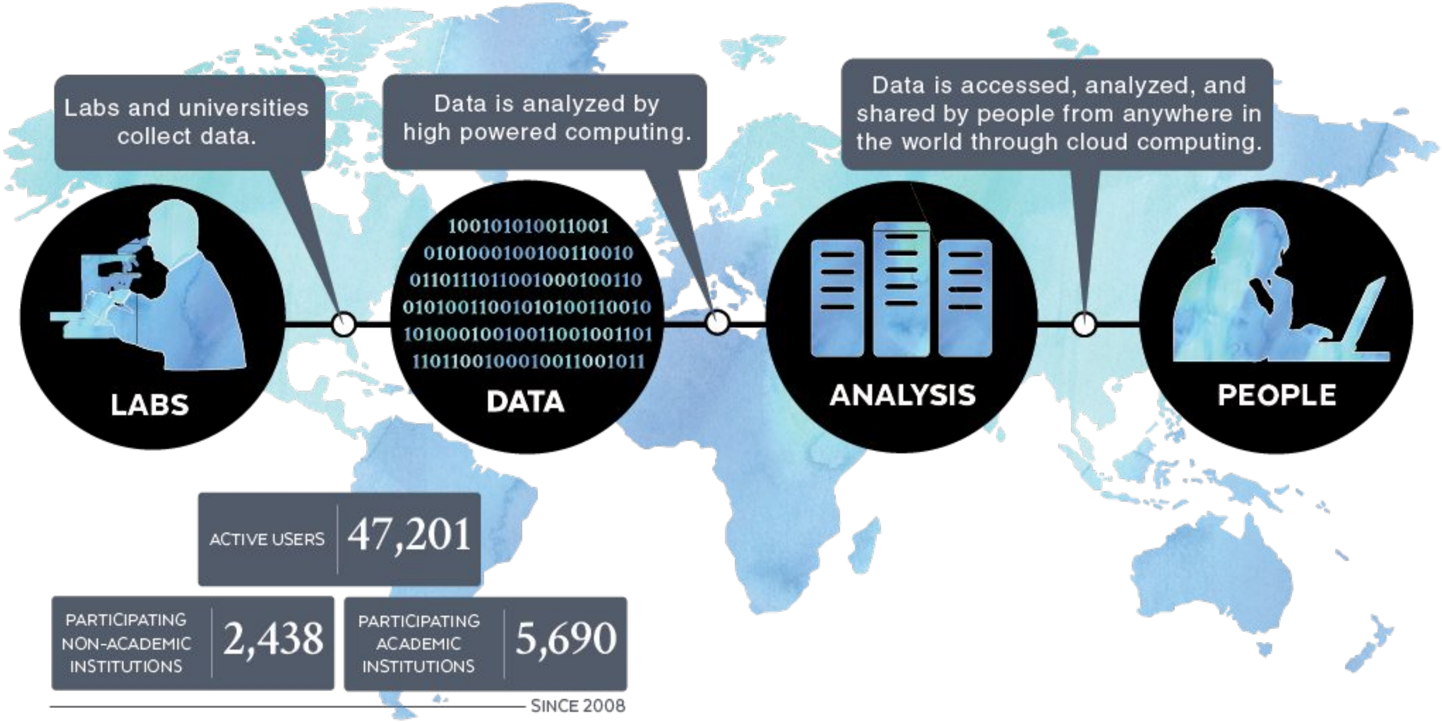
This quickstart provides the basic steps for organizing your data and requesting a DOI.

- Quickstart home
- 1. Before you start
- 2. Organize data
- 3. Add metadata
- 4. Submit request
- 5. After publication

Additional information

- Is the CyVerse Curated Data Repository right for my data?
  - Question 1: Do you have a CyVerse account?
  - Question 2: Is your data ready for publication?
  - Question 3: Is your data suitable for reuse in scientific analyses?
  - Question 4: Is there a canonical repository for your data?
  - If you answered No
- Data Commons Naming Conventions

- Data storage geared specifically towards data analysis
- Interactive, web-based analytical platform
- Cloud computing, analysis and storage
- Support services for scaling up computational algorithms & on how to use CI



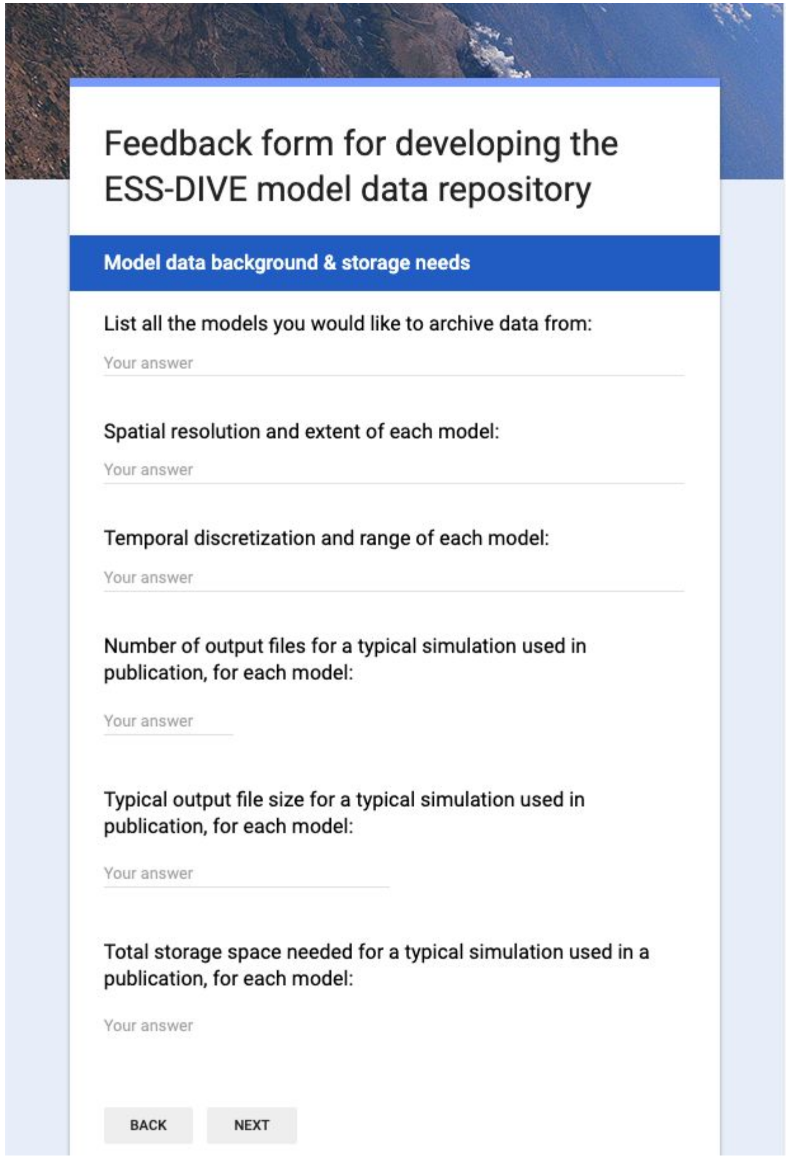
Labs and universities collect data.
Data is analyzed by high powered computing.
Data is accessed, analyzed, and shared by people from anywhere in the world through cloud computing.
LABS
100101010011001
0101000100100110010
0110111011001000100110
0101001100101010011001O
1010001001001100100110l
11011001000100110O1011
DATA
ANALYSIS
PEOPLE
ACTIVE USERS 47,201
PARTICIPATING NON-ACADEMIC INSTITUTIONS 2,438
PARTICIPATING ACADEMIC INSTITUTIONS 5,690
SINCE 2008

Any thoughts so far?

DISCUSSION & POLLING

(DRAFT) Feedback form for developing the ESS-DIVE model data repository

- Currently comprised of 20 questions to:
  - inventory models and assess their specific data storage needs
  - evaluate what's worth archiving and for how long
  - get recommendations for archiving protocols and storage options
- Let's review for completeness and start discussing answers!



Feedback form for developing the ESS-DIVE model data repository

Model data background & storage needs

List all the models you would like to archive data from:

Your answer

Spatial resolution and extent of each model:

Your answer

Temporal discretization and range of each model:

Your answer

Number of output files for a typical simulation used in publication, for each model:

Your answer

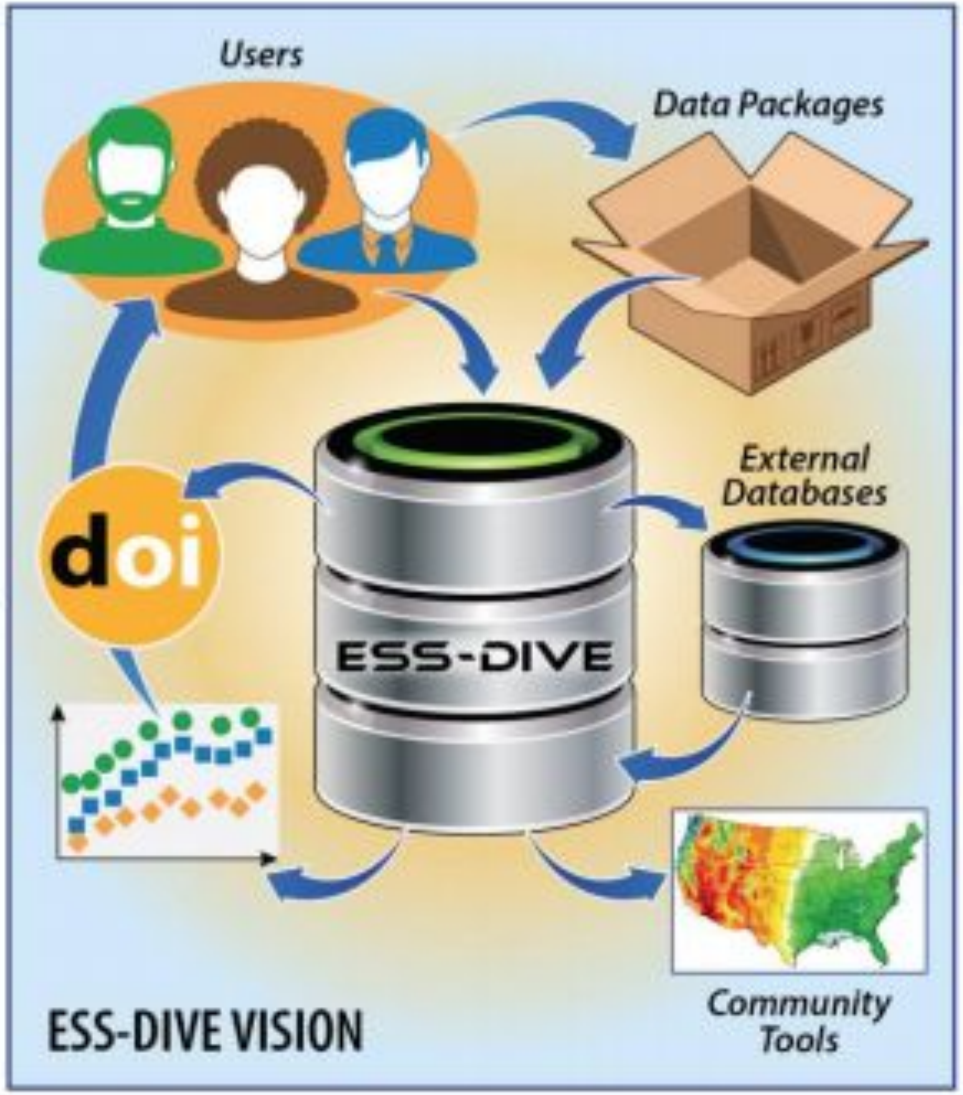
Typical output file size for a typical simulation used in publication, for each model:

Your answer

Total storage space needed for a typical simulation used in a publication, for each model:

Your answer

ESS-DIVE: Vision



Next steps

1 Synthesize our discussion today: compile preliminary poll of responses to feedback form, and email link for additional comments on it from the ESS community.

2 Revise feedback form and distribute to everyone in ESS community.

3 Connect for follow-up discussions.

4 Email any questions or more ideas to Maegen (mbsimmonds@lbl.gov) and Bill (wjriley@lbl.gov).