The NAS Data Portal enables you to share large, multi-gigabyte datasets with either the public or, for restricted datasets, with specified collaborators. After you submit a dataset, it will be reviewed by the Principal Investigator (PI) for your project and the NAS Data Publication and Discovery team; if approved, it will be made available for download. Once your dataset is published, you can modify it at any time. Please note that you as the data owner will responsible for maintaining backup copies of the dataset. While NAS will keep the data as current and consistent as possible, there might be issues with other uncertainties beyond our control. We recommend you copy critical files to the tape archives.

To learn how to submit a dataset for sharing, update your submitted datasets, and download datasets shared by other NAS users, see:

- Submitting Datasets to the NAS Data Portal
- Submitting Datasets to the NAS Data Portal's THREDDS server
- Updating Datasets in the NAS Data Portal
- Downloading Datasets from the NAS Data Portal

Dataset Types

You can share raw or processed datasets, either with the public or with specified collaborators. Each dataset type is associated with specific security restrictions.

Raw or Processed

When you submit a dataset, you must specify whether it contains *raw* data or *processed* data:

- Raw datasets contain data unaltered from their original form.
- Processed datasets contain data that have been modified, reduced, filtered, or interpreted in some way.

At any time, you can edit or append new data to raw or processed datasets (for example, new time series data). In both cases, the modified dataset will be made available for download once the change is approved.

Public or Restricted

You can choose to make your dataset *public,* which means it is open and available to anyone, or *restricted*, which means it is available only to certain collaborators that you specify (currently, only those with a NASA identity).

Both public and restricted datasets must be hosted on one of the Lustre /nobackup filesystems. Both types of datasets should be left in a hierarchical directory structure, but they must all reside under one top-level directory placed under a directory called either "public" (for public datasets) or "private" (for restricted datasets). For example:

- Public: /nobackupp8/*username*/public/*datasetname*
- Private: /nobackupp8/*username*/private/*datasetname*

To migrate your dataset files to a /nobackup filesystem, use the shiftc command, which automates the required file striping.

Notes:

- If it is not feasible to house all of the dataset files in this recommended directory structure (for example, if they are spread across multiple filesystems), you can indicate this in the form when you submit a new dataset or amend an existing one. The NAS team will contact you to discuss the dataset locations.
- If your dataset needs to be modified at regular intervals (more than a few times a year), please contact the Data Portal team to set up a system for frequent updates.

If you choose to make your dataset restricted, and you want to share it with non-NASA collaborators, those individuals will be provided guest accounts via the Launchpad authentication system used to login to the Data Portal. Once the guest accounts are set up, the collaborators can log in to access to the shared dataset files.

Security Requirements for Each Dataset Type

Datasets shared on the NAS Data Portal must not contain any Sensitive-But-Unclassified (SBU) data, or data that is restricted under the International Traffic in Arms Regulations (ITAR). Each dataset type carries its own set of security restrictions, summarized in this table:

NAS Data Portal Security Requirements

Dataset Type	Required Documents
Public (Raw Data)	None
Public (Processed Data)	NASA Form 1676B is needed only if NASA Scientific and Technical (STI) data is included.
Restricted (Raw or Processed Data)	NASA Form 1676B or Space Act Agreement (SAA) or Software Usage Agreement (SUA) is needed for non-NASA and/or non-U.S. citizen collaborators for STI data.

Public (Raw) Datasets

Before you share a public dataset containing raw data, you must ensure that it meets these security requirements:

- The dataset must be public ("non-excepted" per NASA Policy Directive [NPD] 2230.1).
- The dataset must not contain Sensitive-But-Unclassified (SBU), TS, ITAR/EAR, or Personally Identifiable Information (PII) data.
- There must be no restrictions on data release due to agreements between various project members/organizations.

Public (Processed) Datasets

For any public dataset with processed data that does not contain any NASA Scientific and Technical Information (STI), follow the guidelines listed above for public (raw) datasets.

For all public datasets with processed data that *does* contain STI, a NASA Scientific and Technical (STI) Document Availability Authorization—NASA Form 1676B—is required. This form is used to ensure that your dataset contains only public domain data, and that none of the data is subject to export control restrictions.

The completed form must be attached when you submit a public, processed dataset that contains STI data. You can access the form here: NASA Form 1676B (requires login to NASA's Access Launchpad).

For more details, refer to "excepted" data per NASA Policy Directive (NPD) 2230.1.

Restricted Datasets

A restricted dataset is shared only with specified collaborators (both NASA and non-NASA).

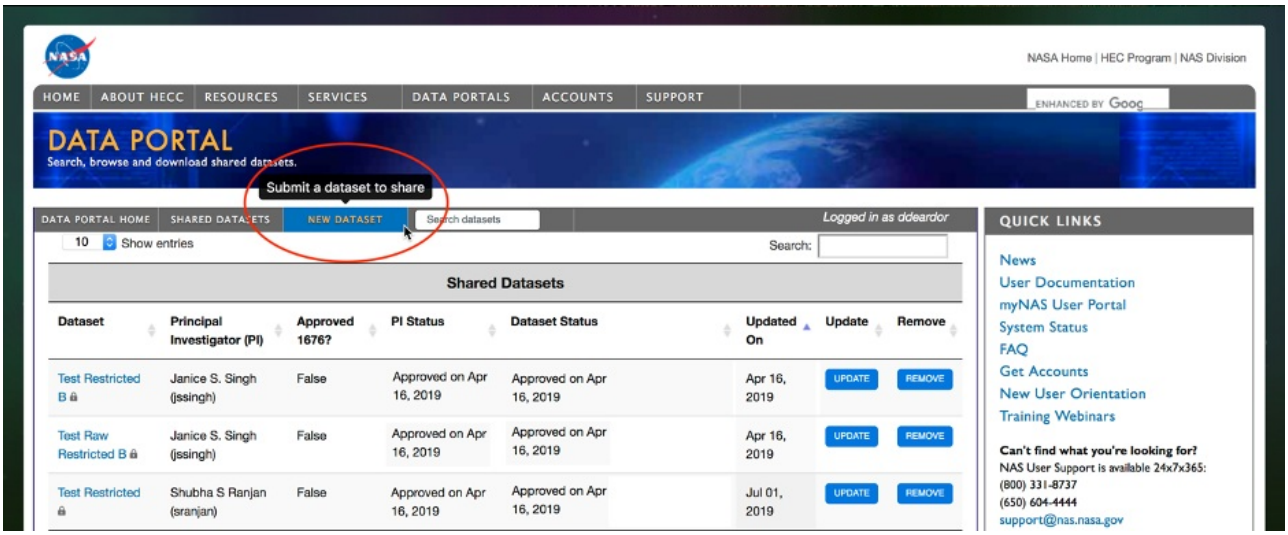
If your restricted dataset is to be shared with non-NASA collaborators, either a NASA Form 1676, or Space Act Agreement (SAA), or Software Usage Agreement (SUA) will be required depending on the individuals it's being shared with. Please contact the Data Portal team for further assistance.

Status API for the Data Portal

Each dataset on the Data Portal is accompanied with an endpoint that will return the status of the dataset when it is queried. The return value provided by this feature is a JSON message that will indicate the availability status of the dataset that is being queried, along with a general message regarding the dataset. Specific datasets will provide further granularity into specific subsets of the data. To retrieve the JSON, add status.php to the end of your Data Portal URL. For example:

https://data.nas.nasa.gov/*dataset*/status.php

On the NAS Data Portal submission site, you can submit a dataset to share with others and view a list of all of the datasets you have already submitted. Please note that currently, only NAS users can submit datasets. Additionally, you as the data owner will be responsible for maintaining backup copies of the dataset. While NAS will keep the data as current and consistent as possible, there might be issues with other uncertainties beyond our control. We recommend you copy critical files to the tape archives.



Note: The Data Portal is currently close to its storage capacity. Depending on the size of your dataset, the Data Portal team may not be able to deploy it as this time. We are, however, actively working to acquire more storage, and are currently building a wait queue for researchers who want to submit their datasets for sharing through the Data Portal. Please go ahead and submit your dataset to the Data Portal, and we will contact you when we are able to proceed with this deployment. Also, if you have funds available to support storage for your dataset, please contact the Big Data Team Lead, Shubha Ranjan, at shubha.ranjan@nasa.gov.

Submitting Your Dataset

To submit your dataset:

1. Review each type of dataset and its security requirements in NAS Data Portal: Overview.
2. Go to the NAS Data Portal dataset submission site.
3. Log in using your NAS username and password.
4. Select **New Dataset** in the menu.

The submission form will be displayed. This section describes each field of the form.

- For NASA-affiliated collaborators, you can provide just the email address.
- For non-NASA collaborators, you will need to provide an email address, first and last name, and the country of citizenship. You may also optionally provide a phone number. In addition, for non-NASA collaborators you will need to provide the ID number of the release document (e.g. SAA or SUA) for each specified collaborator, and the expiration date of that document; there is also a checkbox for you to acknowledge responsibility for obtaining the release documents. You may specify as many collaborators as you need. LaunchPad guest accounts will be set up for non-NASA collaborators, so that they can log into the Data Portal to access the restricted dataset.

Principal Investigator
: Select the Principal Investigator (PI) for this project from the dropdown list, which includes the PIs for all of the projects you are assigned to. Once you submit the dataset, the PI will be notified; if approved, it will go to the NAS team for final validation and deployment on the Data Portal.

Name
: Provide a name for your dataset. This name will be used to identify the dataset in email notifications.

Short Name
: If your dataset is public (not restricted), provide a short project name for the dataset. This name will be used as the directory name for the dataset in the Data Portal's filesystem.

Description
: Provide a description of your dataset, explaining what the data is, how it was generated, what it is used for, and so on.

Location
: Specify the full path to the dataset on the /nobackup filesystem (all shared datasets must reside on one of the /nobackup filesystems). Ensure that all of the files are located in one top-level directory under public (for public) or private (for restricted). For example:

  - /nobackupp8/*username*/public/*datasetname*/
  - /nobackupp8/*username*/private/*datasetname*

  Directory names in and under the top-level directory must not contain any of the following characters: #$%^&*()+=\[\]:?~'\";)

  The best way to copy your dataset files to a /nobackup filesystem is to use the shiftc command. This automates the required

file striping.

Note: If it is not feasible to house all of your public dataset files on one filesystem as described above, please list one location in this field, and use the **Notes** field below to specify the additional locations. The NAS team will contact you to discuss the locations.

Processed
: Check this box if the dataset contains processed data (where *processed* is defined as data that has been modified, reduced, filtered, or interpreted in some way).

Note: Public processed datasets that contain Scientific and Technical Information (STI) must have an approved NASA Form 1676B (per NASA Policy Directive (NPD) 2230.1), which must be uploaded when submitting this dataset (see the **Uploaded 1676B** field below). The form is available here: NASA Form 1676B (requires login to NASA's Access Launchpad)

Restricted
: Check this box if the dataset is to be restricted to specified collaborators. Checking the box reveals an additional field:

Restricted Users: Specify the people you would like to share your dataset with. You may specify as many collaborators as needed, as described below.

Note: If your restricted dataset is to be shared with non-NASA collaborators, either a NASA Form 1676, or a Space Act Agreement (SAA), or Software Usage Agreement (SUA) will be required depending on the individuals it's being shared with. If you need further assistance with setting up non-NASA collaborators, please contact the Data Portal team.

- For NASA-affiliated collaborators, you can provide just the email address.
- For non-NASA collaborators, you will need to provide an email address, first and last name, and the country of citizenship. You may also optionally provide a phone number. Additionally, for non-NASA collaborators, you will need to provide the ID number of the release document (e.g. SAA or SUA) for each specified collaborator, and the expiration date of that document. There is also a checkbox for you to acknowledge responsibility for obtaining the release documents. LaunchPad guest accounts will be set up for any non-NASA collaborators, so that they can log into the Data Portal to access the restricted dataset.

Notes
: If you would like to convey any special notes about this dataset, include them here. This can be used to describe additional locations of dataset files, for instance.

Associated Websites
: Provide one or more websites that pertain to this dataset, as needed.

Tags
: Provide a comma-separated list of keywords (tags) that will be associated with your dataset on the Data Portal. Other users can search for your dataset using these tags.

No 1676B Needed
: Form 1676B is required only for the following types of datasets:

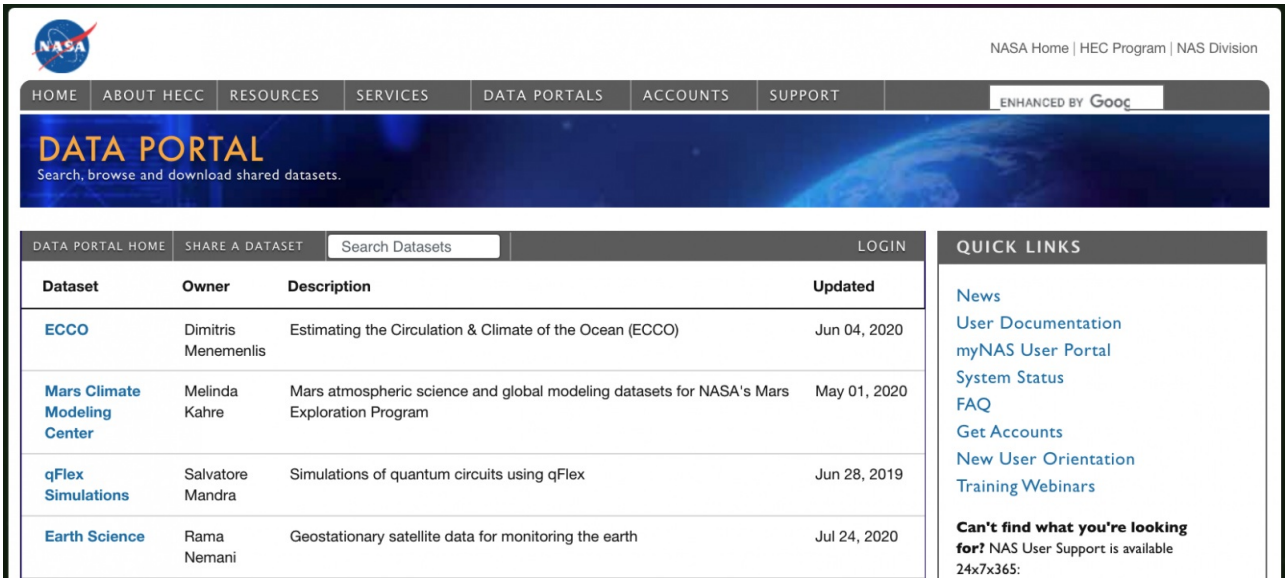
- Public, processed datasets that include STI information
- Restricted datasets where foreign nationals are included as collaborators (refer to "excepted" data per NPD 2230.1).

Check this box to acknowledge that the dataset is not subject to form 1676B requirements; otherwise, you must upload a completed form using the provided file browser described below.

Uploaded 1676B
: If your dataset requires Form 1676B, use this file browser to upload the completed form. (Note that neither this field nor the one above, **No 1676B Needed**, will be shown if the dataset is designated as a raw public dataset.)

Security Requirements
: All datasets need to meet the security requirements listed on the form. Please ensure your dataset meets these requirements and check this box. See NAS Data Portal: Overview for details about security requirements for each data type.

To download shared datasets submitted by other NAS users, visit the NAS Data Portal download site:



The homepage lists all of the datasets available for download. You can also search for a specific dataset by its name, owner, or tag.

- **Public datasets:** No login is required.
- **Restricted datasets:** To view and download restricted datasets for which you have been specified as a collaborator, use LaunchPad to authenticate. This will enable you to gain access to your datasets.
  Note: If you are a non-NASA collaborator, you will authenticate using the email address and password that was set up for your LaunchPad Guest Account.

For more information about dataset types and security requirements, see NAS Data Portal: Overview.

Downloading Files in Bulk

In order to download all the files in a directory, you can use the following command:

```
for i in `lynx -dump -listonly download_directory_URL | grep download_data.php | awk '{print $2}'`; \
do file=`basename $i`; wget -O $file $i ; done
```

where *download_directory_URL* is the URL of the NAS Data Portal directory containing the desired download files.
For example: https://data.nas.nasa.gov/ecco/data.php?dir=/eccodata/llc_270/iter42/input

Note: The command line above is too long to be formatted as one line, so it is broken with a backslash (\).

You can make this command into a shell script and pass the download URL to the script. For example:

Sample Script: get_download_data.sh

```
#!/bin/bash

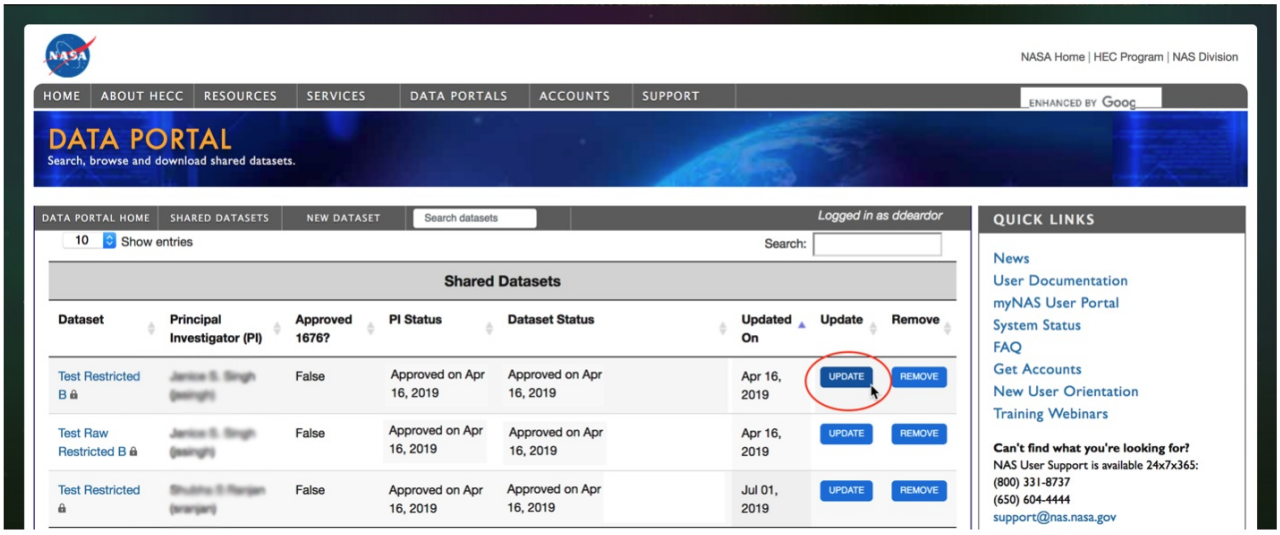
for i in `lynx -dump -listonly $1 | grep download_data.php | awk '{print $2}'`
do
file=`basename $i`
wget -O $file $i
done
```

Run the script as follows:

```
./get_download_data.sh download_directory_URL
```

Once you have submitted a dataset for sharing in the NAS Data Portal, you can continue to add, remove, or delete files and/or directories by submitting a dataset change request.

Your request will be reviewed by the Principal Investigator (PI) for your project and by the NAS Data Publication and Discovery team; if approved, the dataset will be updated.



To submit changes, log into the Data Portal. You will see a list of datasets that you have submitted for sharing, as shown above. Click **Update** for the dataset you would like to update, and a change request form will open. Each field of the change request form is described below:

Name
: You can edit the name of your dataset. This name will be used to identify the dataset in email notifications.

Location
: You can edit the primary location of your dataset files. All shared datasets must reside on one of the /nobackup filesystems). Ensure that all of the files are located in one top-level directory under public (for public) or private (for restricted). For example:

- /nobackupp8/*username*/public/*datasetname*/
- /nobackupp8/*username*/private/*datasetname*

Directory names in and under the top-level directory must not contain any of the following characters: #$%^&*()+=\[\]:?~'\";)

Note: If it is not feasible to house all of your public dataset files on one filesystem as described above, please list one location in this field, and use the **Notes** field below to specify the additional locations. The NAS team will contact you to discuss the locations.

Description
: You can edit your dataset description. Use this field to explain what the data is, how it was generated, what it is used for, and so on.

Notes
: You can add notes about your dataset. This field can be used to describe additional locations of dataset files, for instance.

Associated Websites
: You can edit the list of associated websites.

Delete Dataset
: If you no longer want to share the dataset, you can request that it be deleted from the Data Portal.

Message for Change
: Provide a short descriptive label for this requested change. The label will be used to identify this change request (for instance, in emails).

Description of Change
: Describe the dataset change you are submitting. Note which files have been added, removed or changed. This is needed to communicate to the NAS team exactly the which files to update.

Note: You can edit restricted users or tags without PI approval by clicking the **Edit Restricted Users** or **Edit Tags** buttons at the bottom of the page. Clicking either button will bring up a new form that you can use to submit these updates.

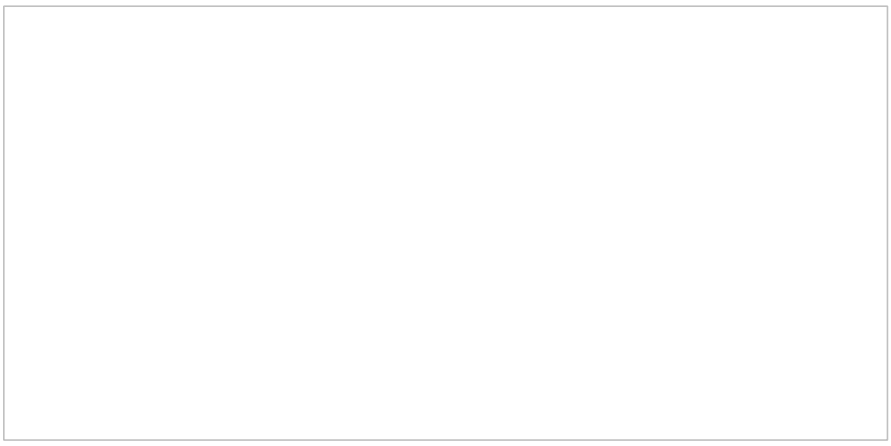
The Thematic Realtime Environmental Distributed Data Service (THREDDS) application was developed by the Unidata organization to provide access to metadata and scientific data. The NAS THREDDS server publishes a catalog of pre-generated datasets for some of the most commonly requested datasets on the NAS Data Portal, and allows users to extract specific data from a larger dataset file using a specific set of parameters.

Key features of THREDDS include:

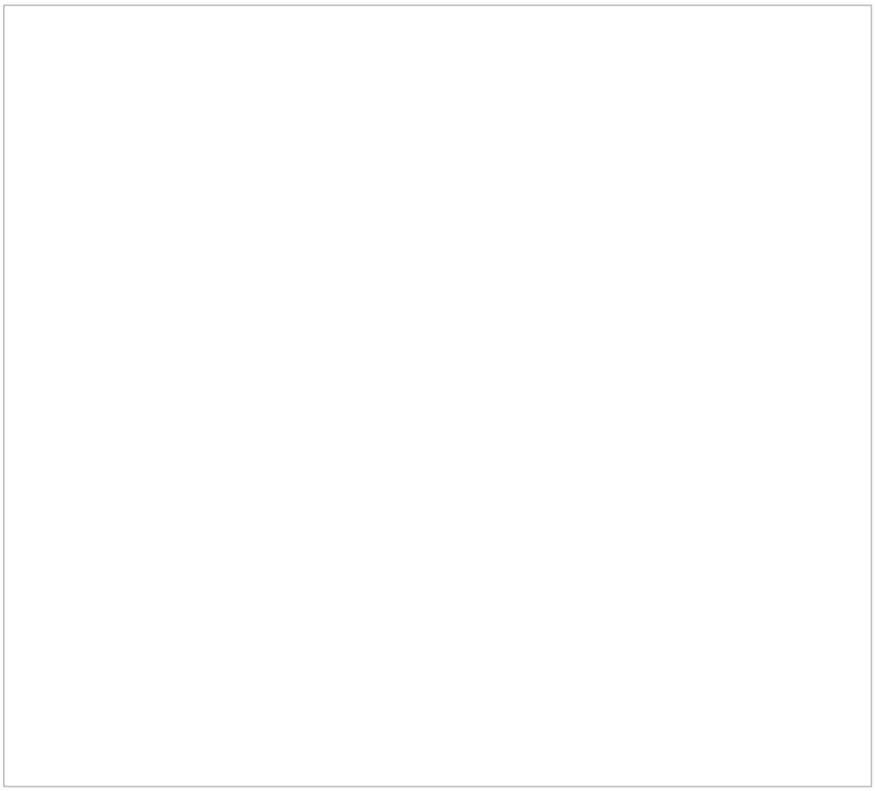
- The Netcdf-Java/CDM library reads NetCDF, OpenDAP, and HDF5 datasets, as well as other binary formats such as GRIB and NEXRAD, into a Common Data Model (CDM)—essentially, an (extended) netCDF view of the data. Datasets that can be read through the Netcdf-Java library are called CDM datasets.
- An integrated server provides OPeNDAP access to any CDM dataset (OPeNDAP is a widely used subsetting data access method that extends the HTTP protocol).
- The integrated NetCDF Subset Service allows subsetting certain CDM datasets in coordinate space, using a REST API. Gridded data subsets can be returned in CF-compliant netCDF-3 or netCDF-4 files. Point data subsets can be returned in CSV, XML, WaterML, or CF-DSG (discrete sampling geometry) v1.6 netCDF files.

Accessing the NAS THREDDS Data Server

The NAS THREDDS data server is available at https://data.nas.nasa.gov/thredds. The TDS homepage lists all available datasets in a series of catalog entries:



The entries link to lists of data files that you can view through the browser or download by using the various protocols available:



The server provides a set of standard data services that are always available (unless explicitly disabled) and can be referenced from any configuration catalog. These utilities work for most model data served through THREDDS:

http
: HTTP File Download Server.

netcdfSubsetGrid
: Interface for choosing and downloading subsetted **gridded** datasets as NetCDF files.

netcdfSubsetPointv
: Interface for choosing and downloading subsetted **point** datasets as NetCDF files.

NCML
: NetCDF Markup Language; provides an XML representation of NetCDF data.

opendap
: Delivers subsetted datasets to OPeNDAP clients.

wcs
: Delivers subsetted datasets to web coverage service (WCS) clients.

wms
: Delivers map images to web mapping service (WMS) clients.

Submitting Datasets to the NAS THREDDS Server

If you are the Principal Investigator (PI) of an existing datatset on the NAS Data Portal, and you want to add it to the THREDDS server, please submit a ticket describing your request to support@nas.nasa.gov.

If you are a PI of a prospective dataset that has not yet been added to the NAS Data Portal, please follow the instructions outlined in Submitting Datasets to the NAS Data Portal.

Accessing Datasets from the NAS THREDDS Server

Example MatLab Script

The script described in this section demonstrates how to access data on the NAS THREDDS Server using the NetCDF functions in Matlab.

First, specify a .ncml URL for the desired dataset:

```
url='http://data.nas.nasa.gov/thredds/path/to/dataset/file.ncml'
```

You can use the ncinfo() function to access the list of variables and metadata information contained in the dataset:

```
metadata = ncinfo(url)
disp({metadata.Variables.name});
```

You can use the ncreadatt() function to grab a specific attribute value:

```
source = ncreadatt(url,'/','source');
```

You can use the ncread() function to grab actual data for the variables that have been selected. The following code will retrieve time and temperature as example variables:

```
dtime = ncread(url, 'time');
dtime = dtime/(60*60*24)+datenum(1900,1,1); %Convert to Matlab time
temp = ncread(url, 'temp');
temp(temp==0)=NaN; %Remove bad points
```

Please note that the example above grabs data for the full range of time in a dataset. Additional parameters can be passed into the ncread() function to subset the the data being retrieved:

```
ind = find(dtime>datenum(2015,1,15) & dtime<datenum(2015,1,20));
dtime2 = ncread(url,'time',ind(1),length(ind),1);
dtime2 = dtime2/(60*60*24)+datenum(1900,1,1); %Convert to Matlab time
temperature2 = ncread(url,'temp',ind(1),length(ind),1);
temperature2(temperature2==0)=NaN; %Remove bad points
```

For more information, please refer to Matlab's documentation for ncinfo(), ncreadatt(), and ncread().

Example Python Script

The following script highlights how to access data on the NAS THREDDS Server using the Pydap Library.

First, import the open_url() function from the Pydap library:

```
from pydap.client import open_ur
```

Next, specify the the dataset that is desired and open the Pydap connection:

```
url = 'http://data.nas.nasa.gov/thredds/path/to/dataset';
dataset=open_url(url)
```

You can use the .keys() function to find out the variables that are available in the dataset:

```
print dataset.keys()
```

You can retrieve additional information about specific variables in the dataset:

```
dataset['example_variable'].dimensions
dataset['example_variable'].shape
dataset['example_variable'].type
dataset['example_variable'].attribute
```

Data can now be accessed from variables. Using [:] will grab all available data in the file. You might need to subset the query if the dataset is very large.

```
time = dataset['time'][:]
var1 = dataset['example_variable1'][:]
var2 = dataset['example_variable2'][:]
var3 = dataset['example_variable3'][:]
```

Convert the datasets time variable, if necessary:

```
from datetime import datetime
dtime = time/(24*3600) + datetime(1900,1,1).toordinal(
```

Now, you can construct a plot from the dataset:

```
#Import necessary libraries
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
# Close any existing plots
plt.close('all')
# Setup the plot
fig, ax = plt.subplots()
# Turn on the plot grid
plt.grid()
# Add the scatterplot with data
sc = plt.scatter(dtime,pressure, c=np.abs(temp), marker='o', edgecolor='none')
# Format the date axis
df = mdates.DateFormatter('%Y-%m-%d')
ax.xaxis.set_major_formatter(df)
fig.autofmt_xdate()
# Reverse the y-axis
ax.set_ylim(ax.get_ylim()[::-1])
# Plot labels
plt.ylabel(dataset['example_variable'].name+&amp;quot; (&amp;quot;+ dataset['example_variable'].units + &amp;quot;)&amp;quot;)
plt.title(dataset.attributes['NC_GLOBAL']['source'], fontsize=11)
# Add the colorbar
clabel = dataset['example_variable1'].name +&amp;quot; (&amp;quot;+ dataset['example_variable2'].units + &amp;quot;)&amp;quot;
fig.colorbar(sc, ax=ax, label=clabel)
# Display the plot (this is a necessary step in Python) and save the figure to a file.
plt.show()
fig.savefig('pydap_fig1.png')
```

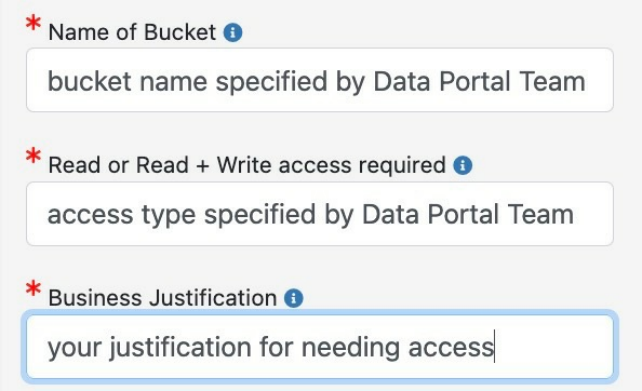
For more information, please refer to the Pydap documentation or contact NAS Support with any questions.

The Data Portal's new Object Store, based on S3 (Amazon Simple Store Service), is used to store Data Portal dataset files. As a dataset owner, you may need to access and maintain your dataset files directly on the Object Store within the NAS enclave. In that case, you need to submit a NAMS request to access your dataset bucket.

First, contact the Data Portal team at dataportal@nas.nasa.gov. They will provide the name of the bucket for the dataset you want to access and the type of access to request ("Read" vs. "Read+Write").

Next, make a NAMS request for the Object Store bucket access:

- Go to the NAMS website at https://nams.nasa.gov/nams/user/
- Fill out the NAMS request for "ARC NAS User Permission Request for Data Portal Object Store Prod" (id: 270962), as shown in the following fields of the form shown below.
  - For Name of Bucket, enter your bucket name provided by the Data Portal Team.
  - For the Access Type, enter the type specified by the Data Portal Team ("Read" or "Read + Write").
  - For the Business Justification, explain why you need access.



Once you submit the NAMS request, your sponsor (manager) will approve it. The Data Portal Approver will then approve and a ticket will automatically be created.

The Data Portal Team will create S3 credentials for you and send them to you with instructions.

After you access your dataset bucket, you can continue your work as normal.

JupyterHub is a powerful platform that provides a multi-user web environment for interactive computing using Jupyter Notebook and JupyterLab. It enables multiple users to have their own isolated and customizable Jupyter servers, all accessible through a standard web browser. This makes JupyterHub an ideal solution for:

- **Collaborative Work**: Teams can work on projects simultaneously in their own environments.
- **Education**: Instructors can provide individual computing resources for students.
- **Research**: Researchers can access and analyze data in dedicated environments.

Accessing JupyterHub

To access the NAS Systems Environment (NSE) JupyterHub platform, please submit a NAMS request. Once you have been granted access to JupyterHub, you can access it at: https://jupyterhub.nas.nasa.gov/

User Space

When you use the NAS Data Portal JupyterHub you will be given your own user space with about 49 GB of disk space. Note that this data gets deleted 24 hours after you end your session/shut down your environment.

Limitations

The Data Portal instance of JupyterHub has the following limitations due to the hardware setup:

- If you want to perform random reads on a file, we advise that you copy the dataset over to your local storage.
- All user space environments are wiped every 24 hours after shutdown. Please copy your data out of the environment after you are done working.

Getting Support

If your team requires more resources and you want to discuss possible funding upgrades, or if you require other assistance or encounter issues, please contact the NAS Control Room at support@nas.nasa.gov.

Available Resources

The NAS JupyterHub instance offers convenient access to several key resources:

- **Data Portal Public Data Files** : You can directly browse and utilize public datasets available through the data portal within your Jupyter environment.
- **Pre-built Containers**: A wide range of pre-configured software containers are available, including popular options for:
  - For machine learning and deep learning tasks.
  - PySci: For scientific computing with Python (likely including libraries such as NumPy, SciPy, and Matplotlib).

For additional information, see the following articles:

- Secure Setup for Using Jupyter Notebook on NAS Systems
- How to Set Up R Kernel in JupyterLab
- Machine Learning at NAS