Accessing the PPS Production Archive using HTTPS and the *arthurhouhttps* Server

By Chris Cohoon and Owen Kelley for PPS, 05 June 2020
This document can be downloaded from the PPS website: https://pps.gsfc.nasa.gov

1. Introduction

At the end of 2020, PPS anticipates that it will replace the current FTP access to its Production data archive with FTPS and HTTPS access. In choosing between FTPS and HTTPS, select HTTPS in situations where firewall restrictions prevent FTPS access.

This document describes the two varieties of HTTPS access, both of which are provided by PPS's *arthurhouhttps* server. One option is to access *arthurhouhttps* with scripting tools like *curl* or *wget* and to request plain-text listings of directories in the archive. This option is best if one plans on parsing the responses in a script. Alternatively, one can access *arthurhouhttps* using a web browser and request HTML-formatted responses that contain clickable hyperlinks. This option is best if one plans on interactively exploring the archive's directory tree.

To obtain a plain-text directory listing, include "text/" following the server name, and to obtain an HTML-formatted directory listing, omit this "text/" string. For example, the top level of the PPS Production data archive is accessed at these URLs for plain-text or HTML responses, respectively:

https://arthurhouhttps.pps.eosdis.nasa.gov/text/
https://arthurhouhttps.pps.eosdis.nasa.gov/

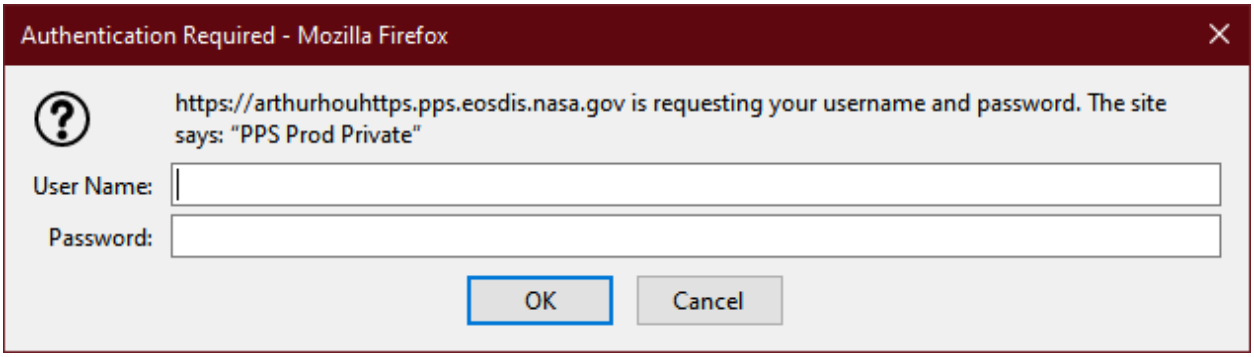
When accessing a directory, include a trailing forward slash ("/"). When accessing a data file, omit the trailing forward slash. If a trailing "/" is placed by mistake after a data file name, the server will return a "404 NOT FOUND" response.

2. User Registration

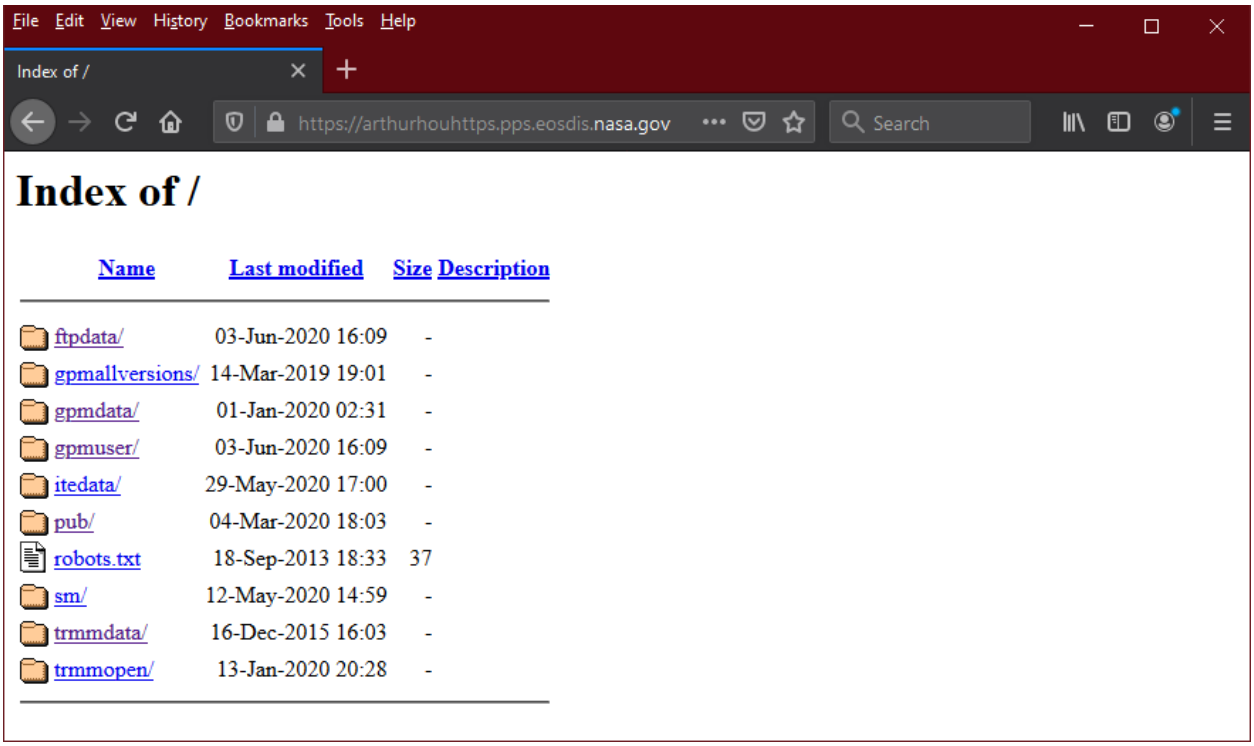
Before accessing the PPS archive, register your email address with PPS by visiting the following URL: *https://pps.gsfc.nasa.gov/register.HTML*

3. Using a Web Browser (HTML response)

To access the *arthurhouhttps* server go to this URL: *https://arthurhouhttps.pps.eosdis.nasa.gov/* . Before this page will display, the browser will prompt for a username and password, most likely in a pop-up window. The details may vary by browser, but regardless, type in your PPS-registered email address in both the username and password fields. (See the previous section of this document for registration instructions.) The username/password pop-up window will only appear the first time that the HTTPS server is accessed during a particular browser session.



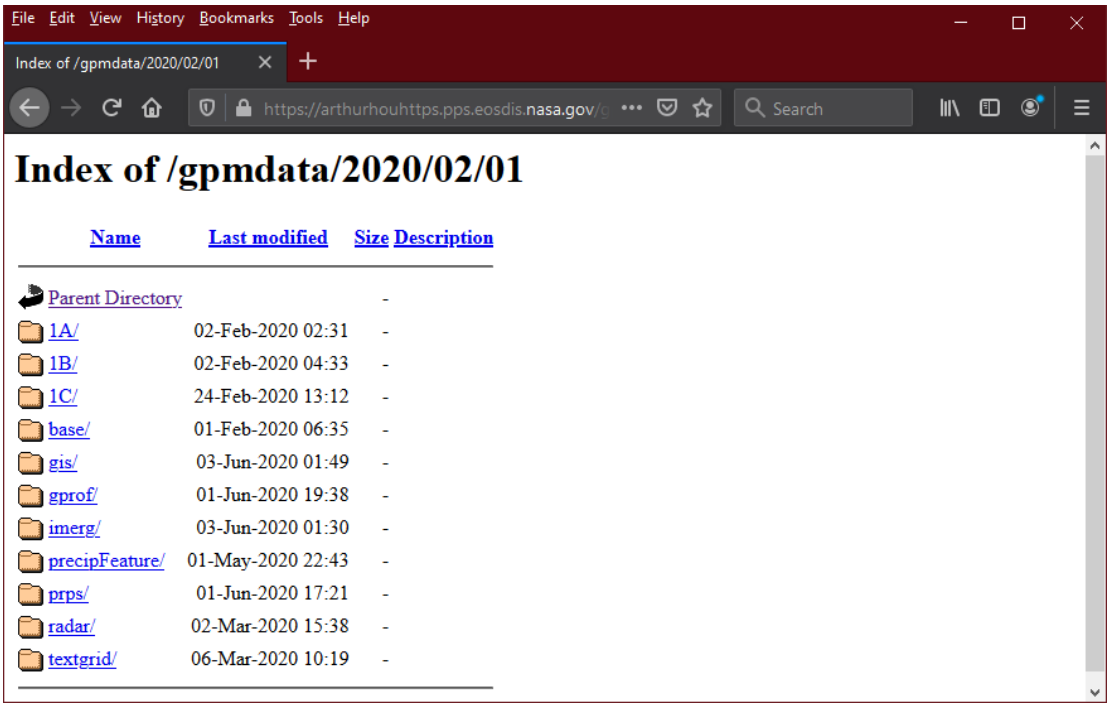
After filling in the username and password fields and clicking the OK button, your browser will display the top-level directory of the PPS Production archive, as shown in the screen capture below.



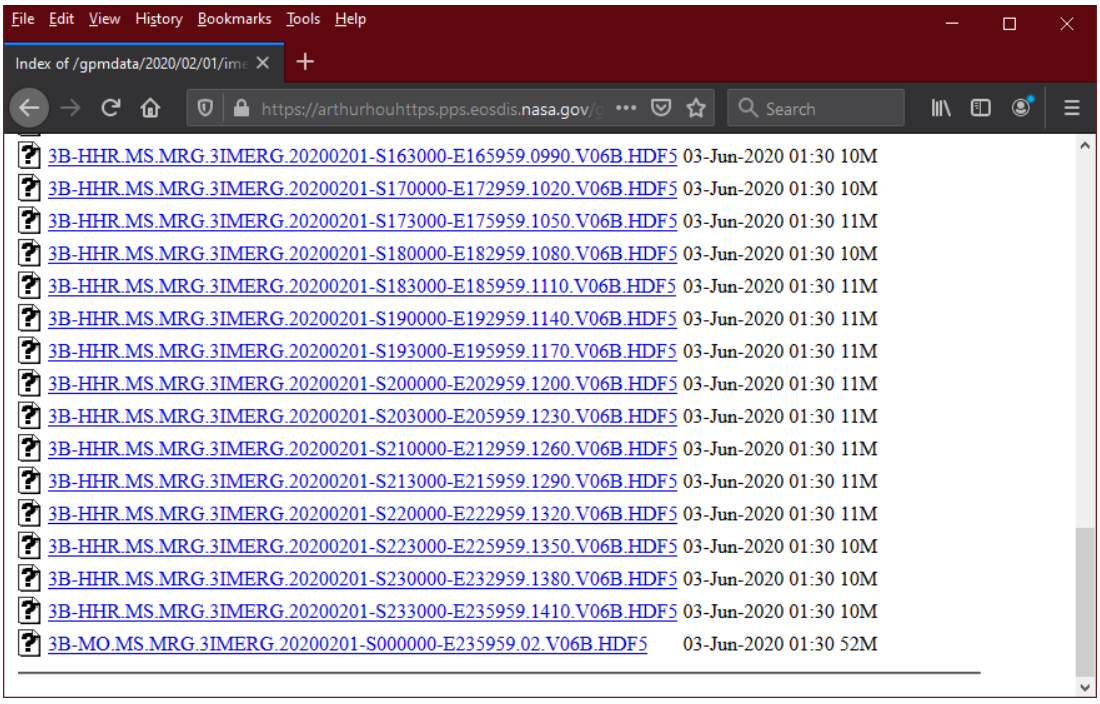
For many researchers, the files of interest will be within the gpmdata directory, which one can enter by clicking on the "gpmdata" link in your browser. The gpmdata directory contains the most recent, officially released version of all GPM data products. Another frequently visited directory is the ftpdata directory, which contains custom-subset files generated by the PPS STORM data-ordering system. To access the contents of ftpdata, use a URL given in an email received from STORM following the completion of an order.

The screen capture below shows what the browser would look like if one clicks on gpmdata and then enters the directory for data products generated from observations made on 1 February 2020. In other words, enter gpmdata/2020/02/01 by successively clicking on gpmdata, the year, the month, and the day of month. The data products for that day are in subdirectories based on the category of data product. For example, single-satellite passive-microwave estimates of precipitation rate are found in the

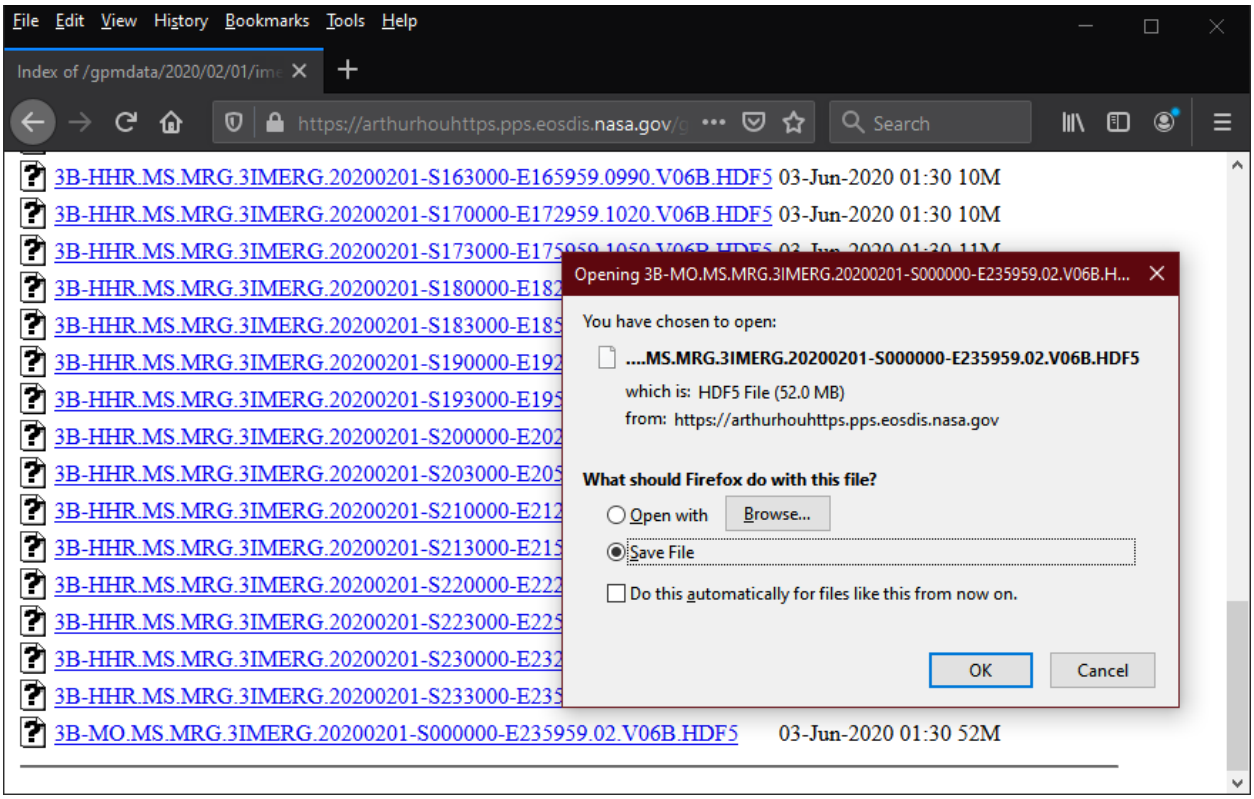
gprof and prps directories, depending on whether they are generated with the GPROF or PRPS science algorithms, respectively. The most commonly downloaded GPM product is the multi-satellite global-gridded precipitation-rate estimates generated by the IMERG algorithm. These files are located in the imerg directory.



Clicking on the imerg directory will give a listing of the IMERG products available for this day (1 February 2020 in this example), as shown in the screen capture below.



Left click on a filename to download that file. The majority of researchers will want to download GPM HDF5 files to their computer rather than immediately open files in a display application. A variety of languages and applications exist to enable researchers to examine HDF5 files including the C, Python, Matlab, and IDL languages. PPS provides a point-and-click desktop application for displaying GPM HDF5 files on a map of the Earth. This application is called THOR (Tool for High-resolution Observation Review) and it can be downloaded from the PPS Homepage: https://pps.gsfc.nasa.gov/ . THOR runs on Linux, Mac OS X, and Microsoft Windows systems.



4. Using Scripts (Text Response)

The *arthurhouhttps* server can also respond with text responses. This is useful when writing scripts or accessing data from the command line. If one is using *curl* or *wget* with HTTPS, the examples below assume that one has set up a .netrc file that lists the PPS-registered email address as both the username and password.

4a. Python Script

Below is a Python script that uses curl to download IMERG files for a user input date. To call this script the user would provide a date with the following format: YYYY-MM-DD. Note that a lot of error handling has been omitted from this script to make it briefer for including in this documentation.

In this program there are two functions that make calls to curl: `get_file_list` and `get_file`. `get_file_list` uses the given date to query *arthurhouhttps* for the directory listing. If there are imerg files for the given date a list of those files will be returned. The file list is looped over to send each filename to `get_file`, which call curl to download the file.

Users wishing to retrieve different types of files should modify the `get_file_list` for the specific desired file types.

```
#!/local/anaconda3/bin/Python3
import sys
import subprocess
import os

server = 'https://arthurhouhttps.pps.eosdis.nasa.gov/text'

def usage():
    print()
    print('Download imerg files for the given date')
    print()
    print('Usage: getImerg DATE')
    print('   DATE - Format is YYY-MM-DD')
    print()

def main(argv):
    # make sure the user provided a date
    if len(argv) != 2:
        usage()
        sys.exit(1)

    # make sure user gave a valid date
    year, month, day = argv[1].split('-')

    # loop through the file list and get each file
    file_list = get_file_list(year, month, day)
    for filename in file_list:
        get_file(filename)

def get_file_list(year, month, day):
    ''' Get the file listing for the given year/month/day
        using curl.
        Return list of files (could be empty).
    '''

    url = server + '/gpmdata/' + \
          '/'.join([year, month, day]) + \
          '/imerg/'
    cmd = 'curl -n ' + url
    args = cmd.split()

    process = subprocess.Popen(args,
                               stdout=subprocess.PIPE,
                               stderr=subprocess.PIPE)
    stdout = process.communicate()[0].decode()

    if stdout[0] == '<':
        print ('No imerg files for the given date')
        return []

    file_list = stdout.split()

    return file_list

def get_file(filename):
    ''' Get the given file from arthurhouhttps using curl. '''
```

```
        url = server + filename
        cmd = 'curl -n ' + url + ' -o ' + \
              os.path.basename(filename)
        args = cmd.split()

        process = subprocess.Popen(args,
                                   stdout=subprocess.PIPE,
                                   stderr=subprocess.PIPE)
        process.wait() # wait so this program doesn't end
                       # before getting all files

    if __name__ == '__main__':
        main(sys.argv)
```

The above code was copy-pasted into a file named getImerg.py. After execution, a directory listing shows that the files were successfully downloaded.

```
[ccohoon@gpmdev tmp]$ vi getImerg.py
[ccohoon@gpmdev tmp]$ ./getImerg.py 2020-02-01
[ccohoon@gpmdev tmp]$ ls
3B-HHR.MS.MRG.3IMERG.20200201-S000000-E002959.0000.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S123000-E125959.0750.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S003000-E005959.0030.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S130000-E132959.0780.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S010000-E012959.0060.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S133000-E135959.0810.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S013000-E015959.0090.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S140000-E142959.0840.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S020000-E022959.0120.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S143000-E145959.0870.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S023000-E025959.0150.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S150000-E152959.0900.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S030000-E032959.0180.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S153000-E155959.0930.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S033000-E035959.0210.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S160000-E162959.0960.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S040000-E042959.0240.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S163000-E165959.0990.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S043000-E045959.0270.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S170000-E172959.1020.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S050000-E052959.0300.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S173000-E175959.1050.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S053000-E055959.0330.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S180000-E182959.1080.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S060000-E062959.0360.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S183000-E185959.1110.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S063000-E065959.0390.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S190000-E192959.1140.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S070000-E072959.0420.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S193000-E195959.1170.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S073000-E075959.0450.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S200000-E202959.1200.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S080000-E082959.0480.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S203000-E205959.1230.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S083000-E085959.0510.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S210000-E212959.1260.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S090000-E092959.0540.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S213000-E215959.1290.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S093000-E095959.0570.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S220000-E222959.1320.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S100000-E102959.0600.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S223000-E225959.1350.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S103000-E105959.0630.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S230000-E232959.1380.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S110000-E112959.0660.V06B.HDF5  3B-HHR.MS.MRG.3IMERG.20200201-S233000-E235959.1410.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S113000-E115959.0690.V06B.HDF5  3B-MO.MS.MRG.3IMERG.20200201-S000000-E235959.02.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200201-S120000-E122959.0720.V06B.HDF5  getImerg.py*
[ccohoon@gpmdev tmp]$
```

4b. Bash shell script

Below is a Bash script that performs the same functionality as the above Python script: it downloads all available IMERG files for the user supplied filename pattern. Because *curl* appears to work under both Centos Linux and Mac OS X, it is used in this shell script rather than *wget*.

```
#!/bin/sh

#-------------------------------------------------------------------------
# Filename: https.sh
# Date: 4 June 2020
# Purpose: A Linux BASH shell script to download files matching
#   a filename pattern from the PPS HTTPS server for the
#   Production data archive, arthurhouhttps.
```

```
# Usage: Make this file executable using "chmod u+x https.sh". Execute
#   with "./https.sh".  The curl command appears to work under
#   both Mac OS X and Centos Linux, while wget only works under
#   Centos Linux.
#------------------------------------------------------------------------

# -- define the file pattern of interest and specify the PPS
#    registered email address that serves as both username and password
URLprefix="https://arthurhouhttps.pps.eosdis.nasa.gov/text/"
filePattern="${URLprefix}gpmdata/2020/01/01/imerg/*20200101-S00*"
# Replace the following email with your own email that you registered
# with PPS at https://registration.pps.eosdis.nasa.gov/registration/
email="owen.kelley@nasa.gov"
echo "$0: searching for filePattern $filePattern"

# -- get a list of files matching this pattern using wget or curl
##fileList=`wget -O - -q --user="$email" --password="$email" "$filePattern"`
fileList=`curl -s -u "$email:$email" "$filePattern"`

# -- return error if no files found
if [ "$fileList" == "" ] ; then
  echo "$0: error: no files found in archive"
  exit 99
else
  numFile=`echo $fileList | wc -w`
  count="0"
  echo "$0: $numFile files match filePattern"
fi

# -- loop over the files found and download, one at a time
for file in $fileList ; do
  fileNoPath=`basename $file`
  count="$((count +1))"
  countPattern="$count of $numFile"
  if [ -f "$fileNoPath" ] ; then
    echo "$0: file $countPattern already exists, skipping $fileNoPath"
  else
##  wget -q -N --user="$email" --password="$email" "${URLprefix}${file}"
    curl -sO -u "$email:$email" "${URLprefix}${file}"
    if [ -f "$fileNoPath" ] ; then
      echo "$0: file $countPattern downloaded $fileNoPath"
    else
      echo "$0: error: failed to download file $countPattern $fileNoPath"
    fi
  fi
done

# -- end of script -
```

The above code was copy-pasted into a file named https.sh. After execution, a directory listing shows that the files were successfully downloaded.

```
[ccohoon@gpmdev tmp2]$ vi https.sh
[ccohoon@gpmdev tmp2]$ chmod u+x https.sh
[ccohoon@gpmdev tmp2]$ ./https.sh
./https.sh: searching for filePattern https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2020/01/01/imerg/*20200101-S00*
./https.sh: 3 files match filePattern
./https.sh: file 1 of 3 downloaded 3B-HHR.MS.MRG.3IMERG.20200101-S000000-E002959.0000.V06B.HDF5
./https.sh: file 2 of 3 downloaded 3B-HHR.MS.MRG.3IMERG.20200101-S003000-E005959.0030.V06B.HDF5
./https.sh: file 3 of 3 downloaded 3B-MO.MS.MRG.3IMERG.20200101-S000000-E235959.01.V06B.HDF5
[ccohoon@gpmdev tmp2]$ ls
3B-HHR.MS.MRG.3IMERG.20200101-S000000-E002959.0000.V06B.HDF5  3B-MO.MS.MRG.3IMERG.20200101-S000000-E235959.01.V06B.HDF5
3B-HHR.MS.MRG.3IMERG.20200101-S003000-E005959.0030.V06B.HDF5  https.sh*
[ccohoon@gpmdev tmp2]$
```

5. Linux Command-line Retrieval (Text Response)

The following *curl* command will list files based on observations collected on 5 April 2010 and returning the response formatted as pure text:

curl -n https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2010/04/05/

The *wget* command is similar for this same directory:

wget -qO- https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2010/04/05/

The pure text that one obtains from either *curl* or *wget* is the following:

/gpmdata/2010/04/05/1A/
/gpmdata/2010/04/05/1B/
/gpmdata/2010/04/05/1C/
/gpmdata/2010/04/05/base/
/gpmdata/2010/04/05/gis/
/gpmdata/2010/04/05/gprof/
/gpmdata/2010/04/05/imerg/
/gpmdata/2010/04/05/precipFeature/
/gpmdata/2010/04/05/radar/
/gpmdata/2010/04/05/textgrid

When using *curl* or *wget*, one can include wildcards in the URL that is placed on the command line. The returned text will be the archive contents that match the wildcard expression. If nothing matches, the server will return a 404 NOT FOUND response. Multiple wildcards can be used in a single request. For example, the following command would list all the products from April 5, 2015 that fit the wildcard "*GMI*S08*":

curl -n https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2015/04/05/*/*GMI*S08*

In this example, the matching files are the following:

/gpmdata/2015/04/05/1A/1A.GPM.GMI.COUNT2016.20150405-S082303-E095535.006250.V05A.HDF5
/gpmdata/2015/04/05/1B/1B.GPM.GMI.TB2016.20150405-S082303-E095535.006250.V05A.HDF5
/gpmdata/2015/04/05/1C/1C-R.GPM.GMI.XCAL2016-C.20150405-S082303-E095535.006250.V05A.HDF5
/gpmdata/2015/04/05/1C/1C.GPM.GMI.XCAL2016-C.20150405-S082303-E095535.006250.V05A.HDF5
/gpmdata/2015/04/05/gprof/2A-CLIM.GPM.GMI.GPROF2017v1.20150405-S082303-E095535.006250.V05A.HDF5

/gpmdata/2015/04/05/gprof/2A.GPM.GMI.GPROF2017v1.20150405-S082303-E095535.006250.V05A.HDF5
/gpmdata/2015/04/05/precipFeature/1Z.GPM.DPRGMI.PF.20150405-S082303-E095535.006250.V06A.HDF5.tar.gz
/gpmdata/2015/04/05/precipFeature/2Z.GPM.DPRGMI.PF.20150405-S082303-E095535.006250.V06A.HDF5.tar.gz
/gpmdata/2015/04/05/radar/2B.GPM.DPRGMI.2HCSHv4-1.20150405-S082303-E095535.006250.V06A.HDF5
/gpmdata/2015/04/05/radar/2B.GPM.DPRGMI.CORRA2018.20150405-S082303-E095535.006250.V06A.HDF5
/gpmdata/2015/04/05/radar/3B-ORBIT.GPM.DPRGMI.3GCSHv6-0.20150405-S082303-E095535.006250.V06A.HDF5

When retrieving a file, do not put a trailing "/" at the end of the request. If a trailing "/" is placed after the file name the server will return a 404 NOT FOUND response. The following is an example *curl* and *wget* command for retrieving a data file:

curl -n \
https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2015/04/05/1C/1C.GPM.GMI.XCAL2016-C.20150405-S082303-E095535.006250.V05A.HDF5 --output <FILE>

wget \
https://arthurhouhttps.pps.eosdis.nasa.gov/text/gpmdata/2015/04/05/1C/1C.GPM.GMI.XCAL2016-C.20150405-S082303-E095535.006250.V05A.HDF5

Please send any questions about accessing the PPS data archive to the PPS Helpdesk, helpdesk@mail.pps.eosdis.nasa.gov.