# Geoprocess DEMs by watershed in Linux

This section describes the steps to collect Hydrologic Unit Boundaries, 6th level (HUC6) and Hydrologic Unit Boundaries, 8th level (HUC8) data inside Linux system.

The data computation requires digital elevation model (DEM) file as the primary input, and utilizes SAGA-GIS (6.4.0, <https://sourceforge.net/projects/saga-gis/>) and GDAL (2.3.1, <https://www.gdal.org/>) to calculate and generate data. Due to the large number of watersheds by HUC6 (total of 336), and by HUC8 (total of 2121), we use two High Performance Computing (HPC) systems to conduct the computations, the Discovery (<https://hpc.nmsu.edu/>) from New Mexico State University and the RMACC Summit (<https://www.colorado.edu/rc/resources/summit>) from University of Colorado Boulder, in order to reduce the time required to finish computation of both HUC6 and HUC8 data.

To start the computation on either of the HPC systems, the general steps are as follow:

1. Install SAGA-GIS version 6.4.0, and GDAL version 2.3.1, and all the required dependent libraries and software.
2. Download *geoprocess\_DEP\_10\_v2.3.sh, Geoprocess\_single\_watershed\_DEP\_10\_v2.3.sh, geoprocess\_DEP\_30\_v2.1.sh,* and *Geoporcess\_single\_watershed\_DEP\_30\_v2.1.sh*, and as the filename suggested, both *geoprocess\_DEP\_10\_v2.3.sh* and *Geoprocess\_single\_watershed\_DEP\_10\_v2.3.sh* are used to process and generate HUC8 data, whereas *geoprocess\_DEP\_30\_v2.1.sh,* and *Geoporcess\_single\_watershed\_DEP\_30\_v2.1.sh* are for HUC6 data. In order for the scripts to work on different systems, it is required to modify the directory of the input files, and the destination of the output files for the respective scripts. More detail in the following section.
3. For HPC systems, job scheduler is typically deployed to manage system resource usage, which means submitting a job to start the computation with aforementioned scripts is required, in our case, the SLURM (<https://slurm.schedmd.com/documentation.html>) scheduler is deployed on both Discovery and RMACC Summit HPC systems, therefore, to start computation on such HPC system, use following command:  
   *sbatch geoprcess\_DEP\_10\_v2.3.sh List\_of\_watersheds.txt* ( for HUC8 )  
    or  
   *sbatch geoprocess\_DEP\_30\_v2.1.sh List\_of\_watersheds.txt* ( for HUC6)  
   where *List\_of\_watershed.txt* contains a list of either HUC6 or HUC8 format watershed numbers.  
     
   For computer systems without a job scheduler, simply use following command to start computation:  
   *./geoprcess\_DEP\_10\_v2.3.sh List\_of\_watersheds.txt* ( for HUC8 )  
    or  
   *./geoprocess\_DEP\_30\_v2.1.sh List\_of\_watersheds.txt* ( for HUC6)  
   where *List\_of\_watershed.txt* contains a list of either HUC6 or HUC8 format watershed numbers.

geoprocess\_DEP scripts

Both *geoprocess\_DEP\_10\_v2.3.sh* and *geoprocess\_DEP\_30\_v2.1.sh* scripts perform the same steps to set up the computation environments, excepts the differences in input file directories, size of buffers and execute the per-watershed computation script *Geoprocess\_single\_watershed\_DEP\_10\_v2.3.sh* and *Geoporcess\_single\_watershed\_DEP\_30\_v2.1.sh* respectively. The general steps are as follow:

1. SLURM submission settings:  
   This part sets up the hardware resource request for SLURM scheduler, changes should be made accordingly for different requirements, as well as different job scheduler.
2. Input file directories:  
   This part sets up the paths for input files, changes should be made accordingly based on the file location.
   * $INPUT\_DIR: The main directory for input files
   * $DEM: The location of DEM file
   * $indexA: The location of HUC8 file
   * $indexB: The location of HUC8 file
   * $bufferA: The size of primary buffer distance in number of pixels
   * $bufferB: The size of secondary buffer distance in number of pixels
3. Parse input and execute single-watershed script  
   The part parses the input text file which contains a list of HUC6/HUC8 format numbers, and loop through each HUC6/HUC8 number as one of the input parameters to execute *Geoprocess\_single\_watershed\_DEP\_10\_v2.3.sh* or *Geoporcess\_single\_watershed\_DEP\_30\_v2.1.sh*, in serial. The other parameter is the number of cores used during data computation, as SAGA-GIS and GDAL are capable of multiprocessing, and currently it’s setting to the max number of cores available on the computing unit.

Geoprocess\_single\_watershed scripts

Both *Geoprocess\_single\_watershed\_DEP\_10\_v2.3.sh* and *Geoporcess\_single\_watershed\_DEP\_30\_v2.1.sh* perform the same steps to generate data for each watershed of HUC6/HUC8 respectively, with exceptions on the output destinations and number derivatives generated. The general steps are as follow:

1. Loading SAGA-GIS and parsing input variables:
   * The Environment Modules (<http://modules.sourceforge.net/>) is typically used in HPC systems to manage software efficiently. “module load” is an Environment Modules command which alters the Linux environmental variables to add the path of executables and/or libraries so that the users can access such executables and/or libraries. In our case, the program “saga-6.4.0” is loaded, which itself also loads “gdal-2.3.1” and other dependent software and libraries.
   * For system without the Environment Modules, but already have both SAGA-GIS, GDAL and other dependent software and libraries installed in Linux default location, simple delete this line.
   * For system without the Environment Modules but has installed required software and libraries in other locations, please delete this line, and add the necessary path to the environmental variables such as PATH, LD\_LIBRARY\_PATH and others accordingly here.
2. Preprocessing:  
   This part sets the output directory and define bash script function *Compress()* to compress output data.
   * $working\_dir: The location of output
3. Calculate Derivatives:

This part defines bash script functions *Trim\_gdalwarp()* to calculate one derivative, and *Delete\_Temp\_Files()* to delete files generated during calculation of one derivative, and then perform multiple derivative calculations with previously defined bash script functions.

\*This work utilized the RMACC Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.