Best Practices for Processing Raster Data in Soil Survey Applications

By NRCS Soil Scientists and GIS Specialists Tom D'Avello, Dwain Daniels, Adolfo Diaz, and Suzann Kienast-Brown.

Background

There is a wealth of raster data currently available for the development of covariates that can be used with countless soil survey-related activities. Because of the large spatial extents of many soil survey projects, planning is of the utmost importance in data development. No one wants to have to recreate a covariate that took 5 days of CPU time to complete because a problem was not discovered until the data was ready for analysis. Following a few guidelines during the preliminary stages of data processing can help eliminate or minimize problems during the application stages.

Processing Checklist

A quick workflow for processing raster data includes:

1) Verifying the data source
2) Verifying the projection parameters
3) Verifying the horizontal **and** vertical units
4) Verifying the resolution
5) Verifying the extent in terms of rows and columns
6) Verify raster statistics
7) Setting a snap raster
8) Buffering the project area
9) Using a watershed for hydrologically based derivatives
10) Processing data
11) Storing output in a common folder

Data source

Do you have complete coverage of raster data from the same source for the extent of your project area? A project area commonly will have raster data coverage that is delivered with the appearance of being a uniform, consistent product but actually has more than one source, and different sources utilize different data capture sensors/techniques, quality parameters, processing techniques, etc. Figure 1 shows an area where elevation data can be extracted as a single 3m resolution raster. The join line between two different sources of LiDAR data extends from the upper left to the lower right corner, and the difference in bare earth surface characterization and significant artifact occurrence is very clear. Just as users of "seamless" SSURGO data with a project area that extends across original soil survey area boundaries need to be aware of potential differences in the soil properties/interpretations and their impact on analyses, users of raster datasets for developing covariates need to be aware of differences and consider them in planning.

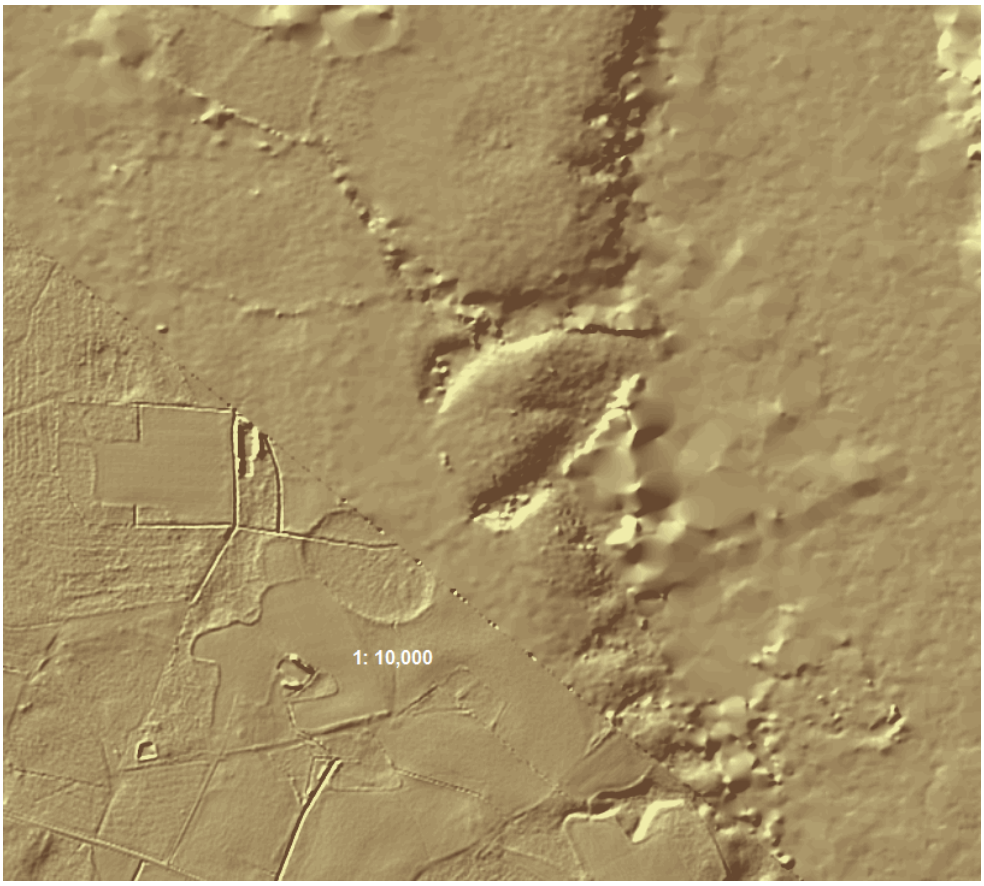


Figure 1. —Hillshade derivative example of a 3m resolution DEM comprised of LiDAR data from two different sources.

Projection parameters

ArcGIS provides dynamic projection capabilities; however, the best practice is to use one common projection for all of the raster data used in the GIS analyses. For DEMs that require re-projection, refer to the Job Aid that provides details related to that operation. If the raster data will also be utilized in applications such as R or ArcSIE, all layers must share common projections, resolutions, and extents. Save data in a GDAL compatible file format for use with R (see list at: http://www.gdal.org/formats_list.html) and in a ESRI GRID format for use with ArcSIE.

Units

Matching the horizontal and vertical units results in assumption-free terrain derivatives. Data provided in a geographic coordinate system, such as decimal degrees, must be converted to a projected coordinate system. Many users have been frustrated trying to interpret a slope-gradient layer generated from mismatched horizontal and vertical units. There are cases, primarily in engineering, in which users maintain DEMs with vertical units that differ from horizontal units, but there is no great need for this in soil survey applications. The primary reason for matching the horizontal and vertical units is to reduce file size. For example, a file with vertical units in meters may be stored as a floating-point, 64 or 32 bit file with a file size of 400 MB. Converting the vertical units to integer feet or centimeters would result in a file size of 200 MB. It is important to note that adoption of this space-saving option will limit the choices in developing slope gradient and curvature layers in GIS packages that accommodate the z scaling

parameter (ArcGIS, QGIS). In these packages, you must remember to use the proper z scaling factor. In addition, the 3 x 3 neighborhood utilized by ArcGIS imposes additional limitations (see the May 2016 NCSS Newsletter, page 16). Finally, this space-saving option will help avoid problems when developing compound terrain derivatives, such as wetness index or stream power index, that assume common horizontal and vertical units.

Resolution

Given the wide range of available data and sources, any given set of assembled data will have an array of resolutions. Developing an inventory that lists the data layers and their resolutions helps in keeping track of everything and documenting the development of metadata. A common cell resolution should be used for all analyses. There are no absolute rules for determining a working resolution. The best practice is to:

1) Determine the phenomena(on) of interest for the project
2) Determine the largest resolution that is required to appropriately map the phenomena(on)
3) Select the resolution from your data inventory that satisfies check #2
4) Resample all data to be used in analyses to the resolution selected in check #3

Two common methods available in ArcGIS for resampling are Resample and Aggregate. Resample is more versatile, allowing for increasing or decreasing to any resolution, while Aggregate only provides for decreasing resolution by whole factors. A Job Aid is available that discusses the use of the Resample tool. It is important that you select the bilinear or cubic convolution resampling technique when resampling continuous data types with the Resample tool. The only time you should select the nearest neighbor technique is when resampling categorical data types.

Common reference

Your objective is to develop all of your raster data with common extents, common number of columns and rows, and common cell alignment. There should be a 1:1 relationship between all cells among the data layers to be used in analyses. Figure 2 shows how multiple resolutions and misalignment cause problems.

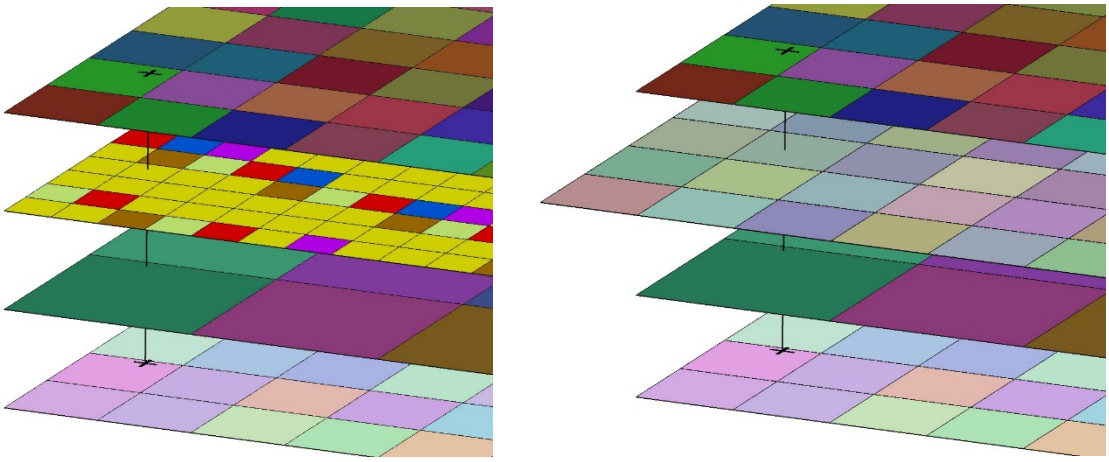


Figure 2.—These images illustrate how the value in one cell of a layer does not apply to the corresponding cells in other layers with different resolutions.

The requirement is that one cell matches all other cells, as shown in Figure 3.

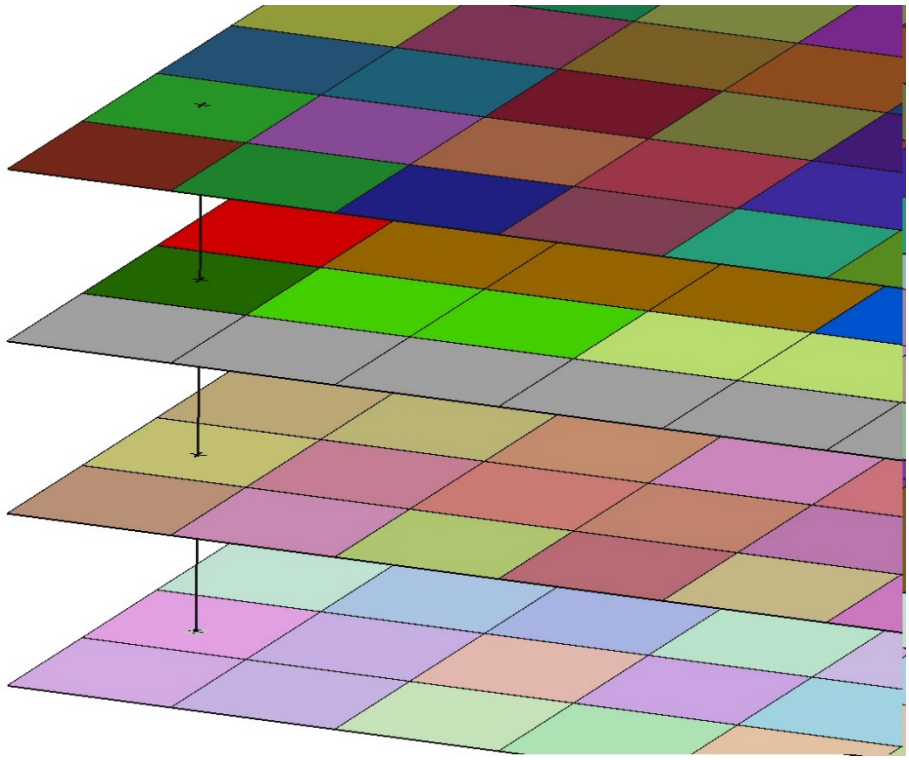
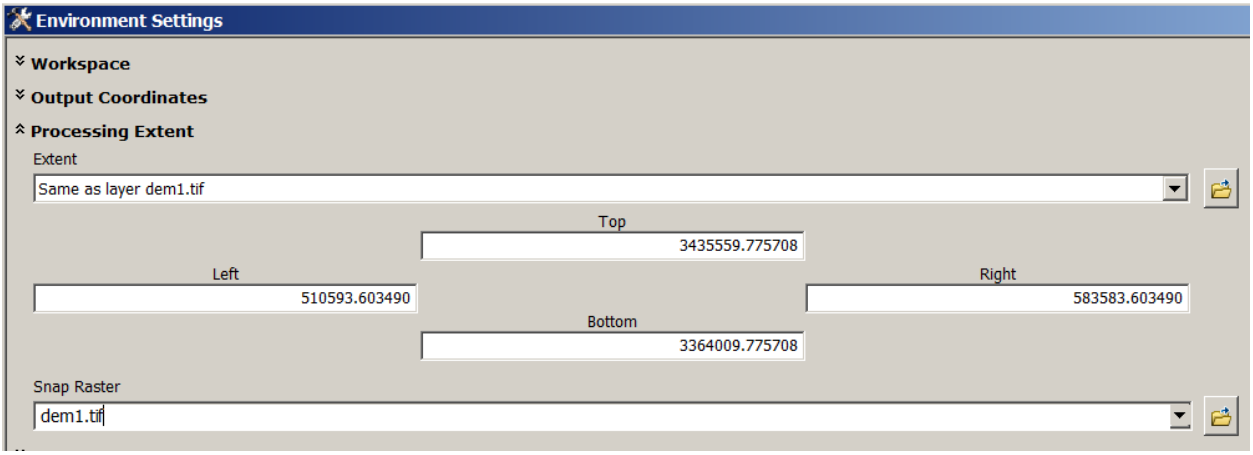
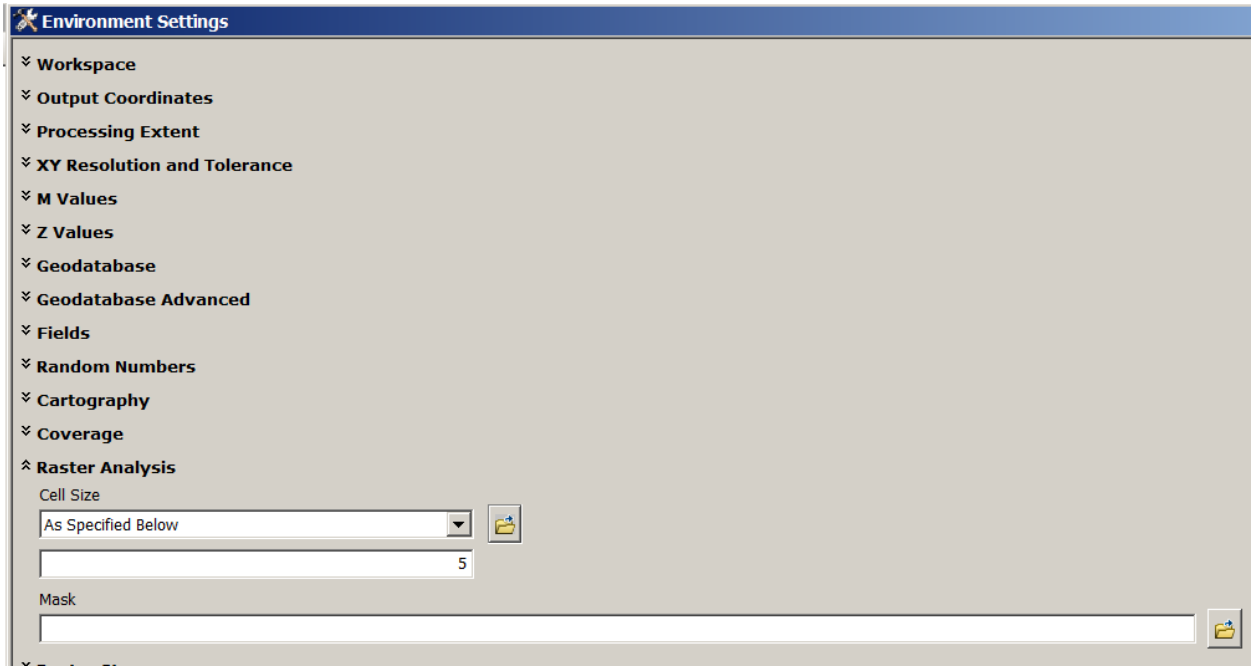


Figure 3. —This image illustrates raster layers that all share a common extent and resolution.

Use the Geoprocessing Environment in ArcGIS to specify the processing extent and snap raster in the Processing Extent drop-down menu. As a general rule, the snap raster should be set to the origin layer from which derivatives are being developed to ensure cell alignment.



Specify the cell size in the Raster Analysis drop-down menu:

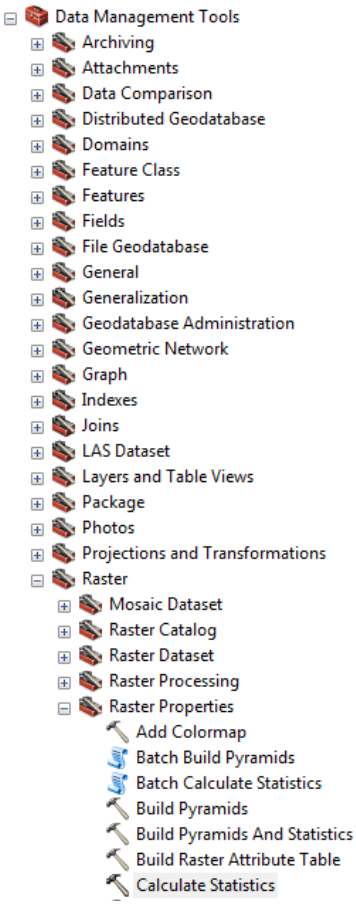


Raster statistics

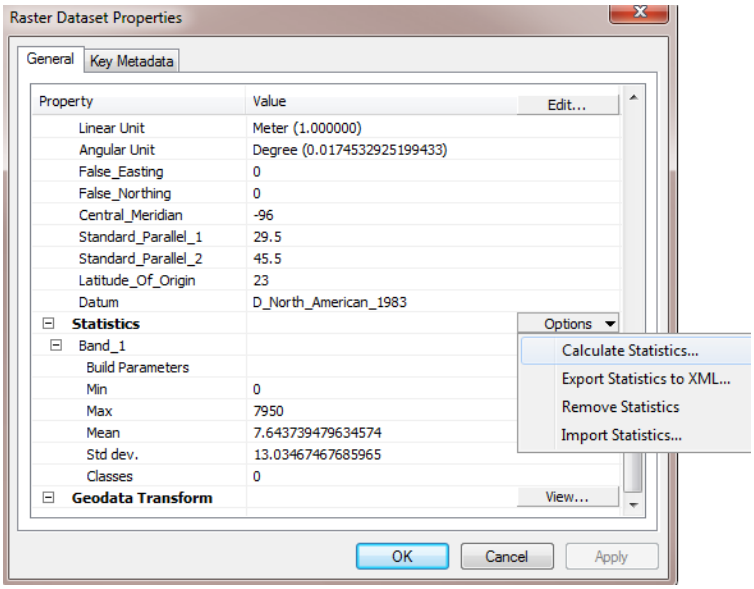
Calculating raster statistics allows ArcMap to properly stretch and symbolize raster data for display and can also serve as a quick quality assurance review by computing the value range of your data. The statistics that are calculated include the minimum and maximum pixel values, the mean and standard deviation of the calculated pixel values, and, if the dataset is thematic (such as a land cover dataset), the number of classes. If your dataset is continuous, such as a DEM, there will be no classes. Below is an example showing the statistics for a DEM whose minimum value may be questionable.

Statistics	Options
Band_1	
Build Parameters	
Min	-35.49980926513672
Max	4017.89794921875
Mean	679.1031711019898
Std dev.	533.9726007184406
Classes	0

There are multiple ways to calculate statistics for a raster dataset. The Calculate Statistics tool in the ArcGIS Data Management toolbox is the most direct way.



Another convenient way is through the Raster Dataset Properties interface, which is accessed by right-clicking on the raster dataset within ArcCatalog. Although you can access the raster properties in ArcMap, you cannot calculate them.



Buffering the project area

Raster processes use a neighborhood to determine the value of the middle cell. The edge of a raster layer will have many No Data (null) values, producing an "edge effect" along the border (see Figure 4).

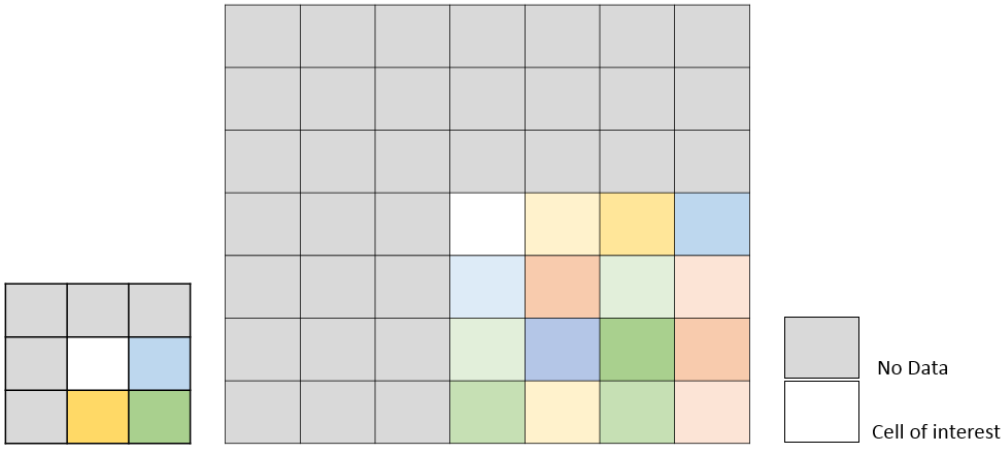
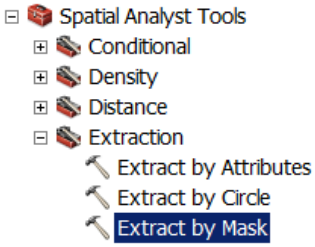


Figure 4. —The typical 3 x 3 neighborhood is on the left, and the larger 7 x 7 neighborhood is on the right.

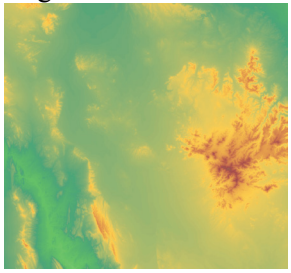
In Figure 4, the value for the center cell will be determined from 3 cells in the 3 x 3 neighborhood, rather than the typical 8, and from 15 cells in the 7 x 7 neighborhood, rather than 48. It is important to note that algorithms based on focal statistics have an optional parameter to "Ignore No Data in calculations." If that option is chosen, No Data input cells will remain as No Data in the output.

Given the variability that can occur at the edge of a raster dataset, you should make sure that your input raster layers extend well beyond the project area. There are no hard rules, but a buffer of 1,000 to 2,000 meters is reasonable. This buffer generally is created from a polygon file defining the project area, but a raster file may also be used. The Extract by MASK tool in the ArcGIS Spatial Analyst toolbox is the most direct way to subset data from a larger extent:



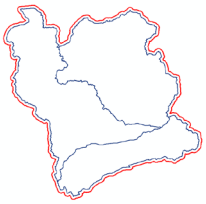
A typical workflow includes:

1) Using a seamless dataset of desired projection, resolution, and units

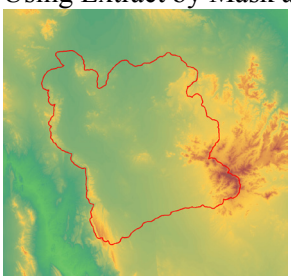


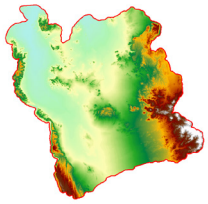
2) Ensuring the dataset defining the project area matches the projection parameters of layer #1

3) Selecting buffer layer #2



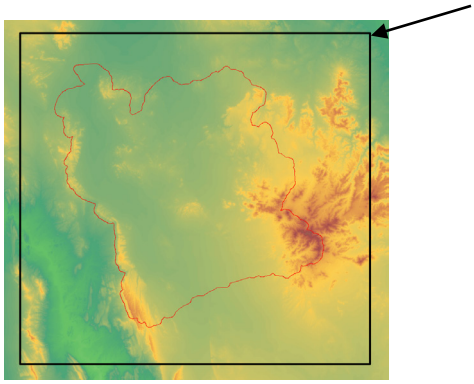
4) Using Extract by Mask and the Buffer layer to subset data





5) Setting the geoprocessing environment, extent, snap raster, and resolution to output layer from step #4
6) Using the buffer layer as a "clip" file for extractions on all other layers required for the project
7) Developing covariates from layers

You may alternatively use a bounding rectangle as a "clip" layer to extract subsets.



If you are processing smaller areas to facilitate computing but your eventual goal is one layer of large extent, you can "clip" the larger buffered layers using a buffer that slightly exceeds the original project area boundary. Adjacent layers can then be merged using Mosaic or Mosaic to New Raster to reform intact project areas with minimal match problems.

Using complete watersheds for hydrologically based covariates

Hydrologically based derivatives that have proven useful for soil survey applications include:

Flow accumulation
Slope length
Stream link
Stream order
Stream power index
Upslope contributing area
Watershed
Wetness Index

Brief descriptions of these covariates are available on the NRCS's Soil Geography website. The cell values for hydrologically based derivatives need to correspond among watersheds for analysis purposes. Using the entire extent of a watershed is the only sure way to satisfy this requirement. The 10- or 12-digit HUC boundaries serve as the most convenient reference. Follow the steps in the previous section and buffer the HUC 1,000 to 2,000 meters or use a bounding rectangle with a border well beyond the watershed edge, then "clip" input data to be processed (using the preceding process).

If the eventual goal is one layer of large extent, this procedure may be used to produce a mosaic dataset with minimal edge-matching problems.

For example, in Figure 5, the 1000 meter buffer file would be used to create a DEM subset. All hydrologically based covariates would be developed using this DEM. This process would be repeated for all adjacent watersheds using corresponding 1000 meter buffer files. The output covariate layers could then be subset using the 30 meter buffer file. The covariates from the 30 meter buffer would be used in the mosaic operation. The assumption is that adjacent layers will share relatively comparable data values within close proximity to their shared watershed divides. As a result, there is a smoother match when multiple layers are mosaicked into larger extents (see Figure 6).

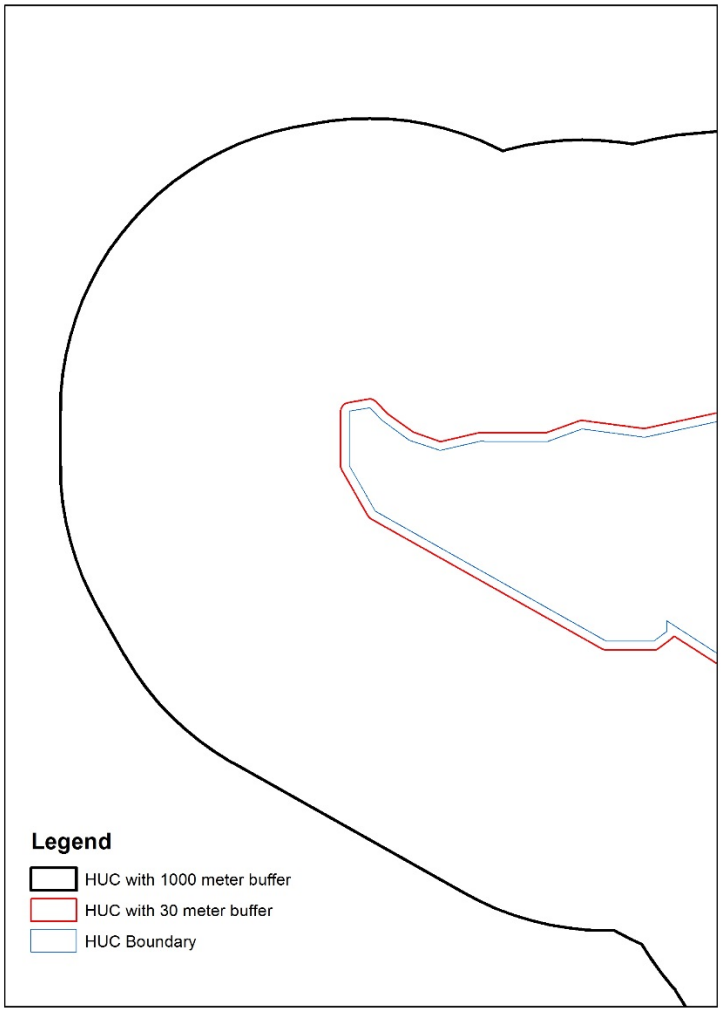


Figure 5. —Example of buffered watersheds from HUC layer.

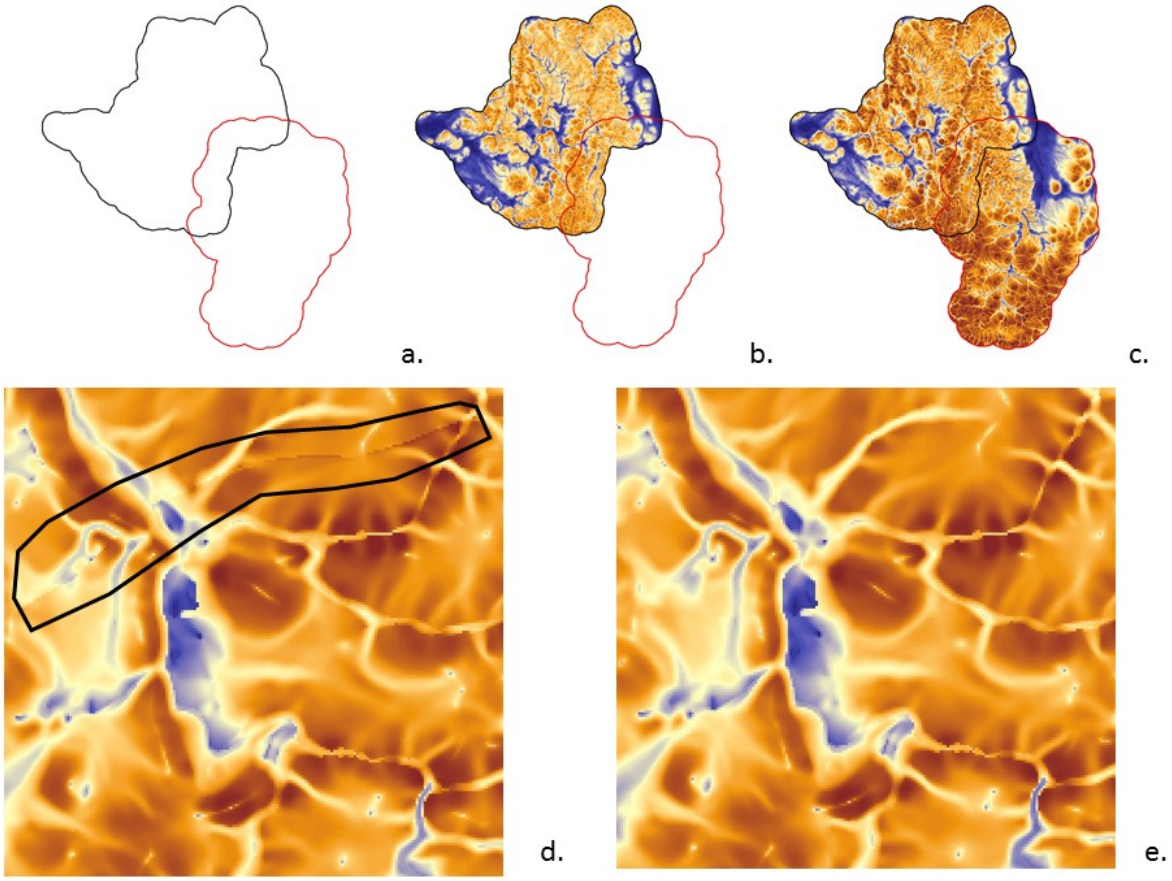


Figure 6.—Example of processing hydrologically based data. Image **a** shows two adjacent watersheds that have been buffered 1000 meters; image **b,** the wetness index layer for one of the watersheds; image **c,** the mosaicked wetness index; image **d**, the visibly poor match within the black polygon resulting from the mosaic of the 1000 meter buffered files; image **e,** the same area from a mosaic developed using the 30 meter buffered extracts of the original 1000 meter buffered wetness index layer with no visible match problem.

Storing output in a common folder

Place all covariates to be used for analyses in a common folder. This helps in data management, such as maintaining back-ups. More importantly, applications using raster data like R require all files to be stored in one common folder.

Develop a naming convention for files for use throughout the office. A Job Aid is available that provides useful guidance for data management. In general, GDAL compatible file formats, such as ERDAS Imagine or GeoTIFF, are preferred. If ArcSIE is being used for modeling purposes, use ESRI GRID format. It is also useful to store these files in a common folder.