Data Development Workflow

Executive Summary

This document describes steps we took to obtain and prepare the key source datasets used for the DMT Habitat Prioritization Project. It includes brief descriptions of the data, steps taken to procure the data, and how the data are stored and accessed for subsequent analysis.

I. The National Hydrographic Dataset (v2) Data (NHD+)

Overview

The NHD+ is a suite of geospatial products derived from static snapshots of the National Hydrography Dataset (NHD) stream network (1:100,000-scale), Watershed Boundary Dataset (WBD) hydrologic units (12-digit), and National Elevation Dataset (NED) topography (30m) through a sophisticated data integration process. The NHD, WBD and NED are all maintained through USGS National Geospatial Program (NGP) led stewardship programs involving states and federal agencies. The derived components of NHDPlus are updated using the latest versions of the NHD, WBD and NED on a periodic basis, as determined by programmatic priorities and resources. For example, the production of NHDPlus Version 2 was largely driven by USGS Water Program requirements for improved regional water quality modeling capabilities. USGS Water and the EPA Office of Water worked together to update the NHD in preparation for NHDPlus Version 2 and, also, shared the cost for producing it.   
(Source: <http://www.horizon-systems.com/nhdplus/NHDPlusV2_home.php>)

The *DMT Habitat Prioritization Project* relies on the NHD+ (version 2) to provide both the spatial units of habitat prioritization (**NHD+ catchments**, eventually up-scaled to the **12-digit hydrologic units**) as well as numerous in-stream and up-slope attributes associated with these catchments. The project also uses a number of raster datasets included with NHD+ to derive additional instream and upslope catchment attributes. Table 1 lists all the datasets obtained from NHD+ used in habitat prioritization.

This section outlines the procedures used to obtain and prepare NHD+ data for use in developing the habitat models used to determine uplift potential.

1. Obtaining the data

Horizon Systems (<http://www.horizon-systems.com/NHDPlus/index.php>) hosts all NHD+ data and related documentation on its servers and allows data to be downloaded free of charge. Most data are bulk-downloaded for regional sections corresponding to major drainage areas; see <http://www.horizon-systems.com/NHDPlus/NHDPlusV2_data.php>. North Carolina touches three of these regions: The South Atlantic North (03N), the Tennessee (06), and the Ohio (05).

1. Uncompressing the files and the NHD+ directory format

We downloaded all the raster, vector, and tabular datasets listed in Table 1 for each of the three regions intersecting North Carolina. Some datasets for regions 03N and 05 were additionally parsed into sub-regional divisions (e.g. 03a and 03b). All subdivisions for each region were downloaded.

The datasets, which are obtained as compressed zip files, were decompressed into the native NHD+ directory format (Figure 1). At the end of this step, all required vector, raster, and tabular NHD+ data could be accessed on a local desktop. Uploading and merging the NHD+ data to the Nicholas School’s GIS Server

1. Importing data to the Nicholas School GIS Server

To facilitate sharing data across multiple machines (and avoid duplication among team workers), we uploaded the NHD+ datasets to the Nicholas School’s GIS Server (“NS-GIS.WIN.DUKE.EDU”) - a Windows 2008 Server running MS SQL Server 2008 (R2) integrated with ESRI’s ArcGIS Server (v 10.2). To do this, we created an ArcGIS Enterprise GeoDatabase on the server (“NHDPlusV2”) using Geodatabase administrative tools in ArcGIS desktop.

All but a few (exceptions listed below) of the NHD+ datasets for a single region (region 03N) were imported directly into the server geodatabase using ArcMap’s *Feature Class to Geodatabase*, *Raster to Geodatabase*, and *Table to Geodatabase* tools. Once region 03N datasets were imported into the server geodatabase, corresponding datasets from the regions 05 and 06 were merged directly into the existing server datasets using the ArcGIS *Mosaic* and *Append* tools for raster and vector/tabular datasets, respectively.

**Projections**

All spatial datasets maintained their original coordinate systems. Raster datasets are referenced to the Albers Equal Area (NAD 83) projection. Vector datasets are referenced to the NAD 1983 geographic coordinate system.

**Metadata**

FGDC Metadata from the original datasets downloaded from the Horizon Systems servers were copied over to the server datasets using ESRI’s Import Metadata tool.

**Permissions/Access**

Special case #1: NHDFlowlines

The NHDFlowlines feature class was handled slightly differently to allow a network dataset to be constructed from it. As ArcMap network datasets can only be created within feature datasets, we first manually created a feature dataset (“NHDFlowlines”) in the server geodatabase, using the coordinate system and XY tolerance information of the NHDFlowlines feature class as a template. We then imported the NHD Flowline dataset for NHD region 03N into this feature dataset and then appended the Flowline datasets from regions 05 and 06 to it (using ArcCatalog’s Load Dataset… function)

Special case #2: Mean annual and **monthly** precipitation, temperature, and runoff tables

The NHD+ provides the tabular data for precipitation, temperature, and runoff as both mean annual averages and as individual monthly averages. Prior to uploading these datasets into the server geodatabase, we merged the annual and monthly means into a single table for each parameter. We did this in two steps, with a Python script written to execute each one.

The first Python script (“[NHD\_MergeVPUAttributes.py](https://github.com/Duke-NSOE/EEP_DataDevelopment/blob/master/Scripts/NHD_MergeVPUAttributes.py)”) merged the individual parameter tables for the different NHD+ regions into a single table covering all regions. For example, the mean monthly runoff records for the month of January (ROMM01001.txt) for catchments in regions 05 and 06 were each appended to those records in region 03N. The appended tables were stored in a local file geodatabase as inputs for the second step. A list of the tables resulting from this step is shown in in Table 2.

The second Python script (“[NHD\_JoinVPUAttributes.py](https://github.com/Duke-NSOE/EEP_DataDevelopment/blob/master/Scripts/NHD_JoinVPUAttributes.py)”) joined the 12 monthly mean values to the annual mean tables for cumulative total precipitation, cumulative total temperature, incremental precipitation, incremental temperature, and runoff, respectively. Records were joined using the *FeatureID* attribute, and the value fields were renamed to include the month of record (e.g. “PRECIPVC\_01”). The result here is a single table for each parameter (precipitation, temperature, and runoff) that include both annual and monthly mean values.

We wrote an additional Python script to calculate the minimum and of the monthly mean values for each catchment. This script is titles “<NHD_CalculateVPUSummaries.py>”.

Special case #3: Incremental and Cumulative NLCD tables

The NHD+ also provides incremental and upstream cumulative area of NLCD land cover classes for each catchment via its VPU Attribute Extension (<http://www.horizon-systems.com/NHDPlus/V2NLCD2011.php>). As in the case with the precipitation, temperature, and runoff tables, we merged NLCD records for each region into a single table of all regions. We used the same Python script as above (“[NHD\_MergeVPUAttributes.py](NHD_CalculateVPUSummaries.py)”) to automate this and then uploaded the resulting tables into the ArcGIS Server geodatabase.

Additionally, however, we also combined several land cover attributes to produce more generalized land cover classes for analysis. Table 3 lists the regroupings of the original NLCD land cover classes. The Python script “NHD\_CalculateNLCDSummaries.py” created these merged NLCD class tables for incremental and cumulative land cover data, respectively.

1. Creating the NHD Flowlines network dataset

To enable network analyses on stream courses (e.g. finding upstream or downstream distances to dam locations), we constructed a network dataset from the NHD+ flowlines. We constructed the network dataset (“NHDFlowlines\_ND”) in the ArcGIS 10.1 format, using the Z-coordinate values from the NHDFlowlines feature class to define elevations and maintaining the default length cost attribute, defined in meters. We also added two other network attributes: one that can restrict flow in the downstream direction and another that can restrict flow in the upstream direction. These are called FlowDirection and Upstream, respectively.

1. Setting permissions/access to the NHD datasets

We set up an administrative “database owner” (DBO) account to upload datasets to the ArcGIS geodatabase. We created a separate read-only account (“EEPreader”) for consuming the data. The datasets are accessed by creating an ArcGIS Database Connection to NS-GIS2.WIN.DUKE.EDU, supplying EEPreader and the password, and setting the database to NHD. *For security purposes, however, the NS-GIS2 server can only be accessed within the Duke University firewall.* (We are looking into using ArcGIS Server as a public access point for these datasets but have not yet found a workable mechanism for this.)

The complete list of NHD+ data imported into the ArcGIS geodatabase is listed in Figure 2.

II. The 2011 National Land Cover Dataset (NLCD)

Overview

The Multi-Resolution Land Characteristic Consortium (MRLC) created and hosts the 2011 National Land Cover Database (NLCD 2011) on its servers. The NLCD comprises wall-to-wall land use/land cover, tree canopy cover, and percent developed impervious surface at a 30 m cell resolution for the United States. Additional information on the NLCD 2011 is found on the MLRC website: <http://www.mrlc.gov/nlcd2011.php>.

We use these NLCD to derive a number of environment layers for habitat evaluation: flowline land cover, percent stream shading, riparian zone composition, catchment percent canopy cover, and catchment percent impervious.

1. Obtaining the data

We obtained raster datasets for land cover, impervious surface, and tree canopy cover (analytical edition) from the MLRC download site: <http://www.mrlc.gov/nlcd11_data.php>. The respective datasets include the entire contiguous US and arrive as zipped Imagine IGE format files. The editions of the NLCD data used are listed in Table X.

1. Uncompressing and uploading the datasets to the Nicholas School GIS Server

After downloading to a local drive, we unzipped each of the three NLCD data files and imported them to an ArcGIS enterprise geodatabase created on the Nicholas School server using the *Raster to Geodatabase (multiple)* tool.

Projections

The three NLCD datasets maintained their original coordinate system: the Albers Conical Equal Area (NAD 83) projection.

Metadata

FGDC Metadata from the original datasets downloaded from the MRLC servers were copied over to the Nicholas School geodatabase server datasets using ESRI’s Import Metadata tool.

1. Setting permissions/access to the NLCD datasets

We set up an administrative “database owner” (DBO) account to upload datasets to the ArcGIS geodatabase. We created a separate read-only account (“EEPreader”) for consuming the data. The datasets are accessed by creating an ArcGIS Database Connection to NS-GIS2.WIN.DUKE.EDU, supplying EEPreader and the password, and setting the database to NLCD. *For security purposes, however, the NS-GIS2 server can only be accessed within the Duke University firewall.*

The complete list of NLCD data imported into the ArcGIS geodatabase is listed in Figure 3.

III. ESRI Landscape Layers

Overview

Environmental Systems Research Institute (ESRI) provides access to a growing number of nationwide raster and vector datasets pertaining to the physical structure of the land – what ESRI terms Landscape Layers (Source: <http://blogs.esri.com/esri/arcgis/2013/09/11/welcome-to-esri-landscape-layers/>). We calculate catchment summary statistics on a number of these layers to evaluate their relative importance in predicting habitat.

1. Accessing the ESRI Landscape Layers

An ESRI ArcGIS Online Organization Account is required to access these datasets, and unlike the NHD+ and NLCD datasets, these Landscape Layers cannot be downloaded in bulk; instead, the data are accessed directly through links from ArcGIS Desktop to ESRI’s ArcGIS Online server. Once this connection is made, however, spatial subsets of the data can be used in geoprocessing or downloaded locally. These subsets are limited to 24,000 x 24,000 pixels for raster data and up to 1000 features for vector data. The raster access limit is usually not an issue at the HUC 6 level.

The landscape layers reside on 5 different ArcGIS servers. We have created connection files to these servers, each stored within the *Data/ESRILayers* folder in the EEP workspace. Our analysis of these layers did not require creating local copies of data subsections. Instead, we performed analyses (zonal statistics) directly on geospatial layers created in Python scripts that access the copies of the data hosted on ESRI cloud. The Python script “[EEP\_ExtractESRILandscapeData.py](https://github.com/Duke-NSOE/EEP/blob/master/Scripts/EEP_ExtractESRILandscapeData.py)” includes code on how the ArcMap *MakeImageServerLayer* process can connect to an ESRI Landscape Layer (via the ArcGIS Server connections mentioned above) and how subsequent geoprocessing tools can work on these layers.

Projection

ESRI Landscape layers are referenced to the North America Albers Equal Area Conic projection.

IV. US Army Corps of Engineers Stream Temperature Regimes

Nancy Daly of DMT provided us with a dataset on in-stream thermal habitat for all North Carolina streams. We uploaded the original file, in ESRI shapefile format, to a server geodatabase (“NC”) on the Nicholas School server as “ThermalStreams” and granted read-only access to the “EEPreader” account. The dataset can be accessed via the NC.sde link found in the Data folder of the EEP workspace. (Access is available only to machines within the Nicholas School firewall.)

Projection

The NC DOT Roads Layer is referenced to the North Carolina State Plane (feet) coordinate system.

V. North Carolina Department of Transportation Roads Layer

The NC Department of Transportation (NC DOT) provides a statewide dataset of primary and secondary road arcs as part of its Linear Referencing System. We downloaded this dataset (in Shapefile format) from the NC DOT GIS resources page (<https://connect.ncdot.gov/resources/gis/pages/gis-data-layers.aspx>) to a local drive. The dataset was then imported to the “NC” server geodatabase on Nicholas School GIS server as NCDOT\_Roads. The dataset can be accessed via the NC.sde link found in the Data folder of the EEP workspace. (Access is available only to machines within the Nicholas School firewall.)

Projection

The NC DOT Roads Layer is referenced to the North Carolina State Plane (feet) coordinate system.

VI. National Inventory of Dams (NID) Reservoir locations

We downloaded federal, state, local, utility, private, and unlisted reservoir data from the US Army Corps of Engineers’ (USACE) National Inventory of DAMs (<http://nid.usace.army.mil/cm_apex/f?p=838:12>). From the national datasets, we extracted features from NC, VA, TN, and SC. We cleaned the data by removing features with duplicate NIDID (often levees around the dams or saddle dikes). We also removed features missing latitude/longitude coordinates, as these tend to be ash bonds, fossil fuel complex, etc. Finally, we clipped the dataset to our study area (all HUC 6s intersecting North Carolina). Table 5 lists the number of reservoirs at the various stages of cleaning.

The cleaned reservoirs feature class was uploaded to the “NC” server geodatabase on the Nicholas School GIS server as “NDI\_Reservoirs”. The dataset can be accessed via the NC.sde link found in the Data folder of the EEP workspace. (Access is available only to machines within the Nicholas School firewall.)

Projection

The NID reservoir layer is referenced to the NAD 1983 Geographic Coordinate System.

VII. North Carolina Aquatic Species Occurrence Data

Mark Endries of the US Fish and Wildlife Service compiled a comprehensive data of aquatic species observation records in North Carolina (<http://www.fws.gov/asheville/htmls/Maxent/Maxent.html>). He provided us with these locations as multi-point features in a shapefile. We uploaded this shapefile to the “NC” server geodatabase on Nicholas School GIS server as “AquaticSpeciesOccurrences”. The dataset can be accessed via the NC.sde link found in the Data folder of the EEP workspace. (Access is available only to machines within the Nicholas School firewall.)

Projection

The Aquatic Species Occurrence Layer is referenced to the North Carolina State Plane (meter) coordinate system.

**Table 1**. NHD+ (v2) Datasets obtained from Horizon Systems for NC Regions (03N, 05, & 06).

|  |  |  |
| --- | --- | --- |
| Format | Name | Description |
| Raster | elev\_cm | Elevation, in cm |
| Raster | cat | NHD+ catchments |
| Raster | fac | Flow accumulation |
| Raster | fdr | Flow direction |
| Raster | fdrnull | Flow direction; streams set to null values |
| Raster | shdrelief | Shaded relief |
| Vector | NHDflowline | NHD+ flowline features |
| Vector | BoundaryUnit | NHD+ regional boundary units |
| Vector | CatchmentFeatures | NHD+ catchment features |
| Vector | WBD\_Subwatershed | National watershed boundary database (HUC12 features) |
| Vector | WBDHU2 | 2-digit hydrologic unit polygons |
| Vector | WBDHU4 | 4-digit hydrologic unit polygons |
| Vector | WBDHU6 | 6-digit hydrologic unit polygons |
| Vector | WBDHU8 | 8-digit hydrologic unit polygons |
| Table (dbf) | CumulativeArea | Cumulative area upstream of an NHDFlowline feature |
| Table (dbf) | ElevSlope | Elevation and slope derived for NHDFlowline features |
| Table (dbf) | PlusFlowlineVAA | NHD+ “Value Added Attributes” for NHDFlowline features |
| Table (txt) | CumTotNLCD2011 | Area of each NLCD class found upstream of a given catchment |
| Table (txt) | IncrTotNLCD2011 | Area of each NLCD class found within a given catchment |
| Table (txt) | CumTotPrecipMA | Mean annual rainfall across the area upstream of a given catchment |
| Table (txt) | CumTotTempMA | Mean annual temperature across the area upstream of a catchment |
| Table (txt) | IncrPrecipMA | Mean annual rainfall over the of a given NHD+ catchment |
| Table (txt) | IncrPrecip*mm* (x12) | Mean monthly rainfall over the area of a given NHD+ catchment |
| Table (txt) | IncrTempMA | Mean annual temperature over the of a given NHD+ catchment |
| Table (txt) | IncrTemp*mm* (x12) | Mean monthly temperature over the of a given NHD+ catchment |
| Table (dbf) | EROM\_MA0001 | Extended unit runoff method (EROM) annual flow estimates |
| Table (dbf) | EROM\_*mm*0001 | Extended unit runoff method (EROM) monthly flow estimates |
| Table (txt) | ROMA0001 | Mean annual runoff recorded within a given NHD+ catchment |
| Table (txt) | ROMA*mm*0001 (x12) | Mean monthly runoff recorded within a given NHD+ catchment |

**Table 2.** List of appended tables of annual and monthly means for precipitation, temperature, and runoff.

|  |  |
| --- | --- |
| IncrPrecipMA | Mean annual precipitation |
| IncrPrecip*mm* (x12) | Mean monthly precipitation (mm = month) |
| IncrTempMA | Mean annual temperature |
| IncrTemp*mm* (x12) | Mean monthly temperature (mm = month) |
| ROMA0001 | Mean annual runoff |
| ROMA010001 (x12) | Mean monthly runoff (mm = month) |

**Table 3.** Reclassification of NLCD classes into more generalized land cover classes

|  |  |
| --- | --- |
| Original class | Generalized class |
| 11 – Open water | 1 – Open Water |
| 12 – Perennial Ice/Snow | *- Omitted -* |
| 21 – Developed, Open Space | 2 – Developed |
| 22 – Developed, Low intensity |  |
| 23 – Developed, Medium Intensity |  |
| 24 – Developed, High intensity |  |
| 31 – Barren | 3 – Barren |
| 41 – Deciduous forest | 4 – Forest |
| 42 – Evergreen forest |  |
| 43 – Mixed forest |  |
| 51 – Dwarf shrub | *- Omitted -* |
| 52 – Scrub shrub | 50 – Shrub |
| 71 – Grassland/herbaceous | 71 – Grassland/herbaceous |
| 72 – Sedge/herbaceous | *- Omitted -* |
| 73 – Lichens | *- Omitted -* |
| 74 – Moss | *- Omitted -* |
| 81 – Pasture/hay | 8 – Pasture/hay |
| 82 – Cultivated crops | 8 – Cultivated crops |
| 90 – Woody wetlands | 9 – Woody wetlands |
| 91 – Emergent herbaceous wetlands | 9 – Emergent herbaceous wetlands |

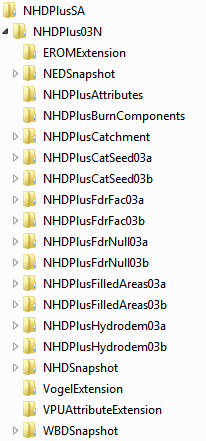
**Table 4.** Editions of NLCD (2011) layers downloaded and stored on the NS-GIS2 server

|  |  |
| --- | --- |
| Land Cover | 2014\_10\_10 |
| Percent Developed Imperviousness | 2014\_03\_31 |
| Tree Canopy Analytical | 2014\_03\_31 |

**Table 5.** Number of National Inventory of Dams reservoir locations by state.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Type** | **Phase** | **NC** | **SC** | **TN** | **VA** | **Total** |
| Federal | Filter by state | 63 | 41 | 96 | 68 | 268 |
|  | Remove Duplicates | 56 | 40 | 54 | 64 | 214 |
|  | Clip to study area | 56 | 40 | 24 | 19 | **139** |
| State | Filter by state | 61 | 78 | 155 | 74 | 368 |
|  | Remove Duplicates | 61 | 77 | 148 | 74 | 360 |
|  | Clip to study area | 61 | 77 | 5 | 18 | **161** |
| Local | Filter by state | 286 | 182 | 177 | 271 | 916 |
|  | Remove Duplicates | 286 | 179 | 174 | 265 | 904 |
|  | Clip to study area | 286 | 179 | 15 | 73 | **553** |
| Utility | Filter by state | 101 | 46 | 9 | 52 | 208 |
|  | Remove Duplicates | 86 | 25 | 9 | 50 | 170 |
|  | Clip to study area | 86 | 24 | 2 | 17 | **129** |
| Private | Filter by state | 2610 | 2089 | 793 | 1184 | 6676 |
|  | Remove Duplicates | 2603 | 2062 | 789 | 1166 | 6621 |
|  | Clip to study area | 2603 | 2058 | 181 | 302 | **5144** |
| Not Listed | Filter by state | 141 | 0 | 0 | 413 | 554 |
|  | Remove Duplicates | 141 | 0 | 0 | 402 | 543 |
|  | Clip to study area | 141 | 0 | 0 | 46 | **187** |

**Figure 1.** Directory format created when NHD+ downloaded zip files are decompressed.



**Figure 2. D**atasets included in the NHD server geodatabase located on NS-GIS2.WIN.DUKE.EDU

