*Journal of Statistical Software*

*January 2014, Volume 56, Issue 2.* *http://www.jstatsoft.org/*

# STARS: An ArcGIS Toolset Used to Calculate the Spatial Information Needed to Fit Spatial Statistical Models to Stream Network Data

**Erin E. Peterson**
CSIRO

**Jay M. Ver Hoef**
NOAA

**Abstract**

This paper describes the **STARS** ArcGIS geoprocessing toolset, which is used to calculate the spatial information needed to fit spatial statistical models to stream network data using the **SSN** package. The **STARS** toolset is designed for use with a landscape network (`LSN`), which is a topological data model produced by the **FLoWS** ArcGIS geoprocessing toolset. An overview of the **FLoWS** `LSN` structure and a few particularly useful tools is also provided so that users will have a clear understanding of the underlying data structure that the **STARS** toolset depends on. This document may be used as an introduction to new users. The methods used to calculate the spatial information and format the final `.ssn` object are also explicitly described so that users may create their own `.ssn` object using other data models and software.

*Keywords*: GIS, spatial statistical modeling, streams, **STARS**, **FLoWS**, **SSN**.

## 1. Introduction

Spatial autocorrelation is an intrinsic characteristic in freshwater stream environments where nested watersheds (i.e., the entire land area that drains to a single location within the stream), dendritic network structure, and the directional flow of water may produce spatial relationships that are not explained by Euclidean distance. Yet, many common autocovariance functions used in spatial models are statistically invalid when Euclidean distance is replaced with hydrologic, or within-network, distance (Ver Hoef, Peterson, and Theobald 2006). This issue made it necessary to develop new spatial statistical methodologies for stream networks, which permit valid covariances to be generated based on a variety of hydrologic, or within-network, relationships (Ver Hoef and Peterson 2010).

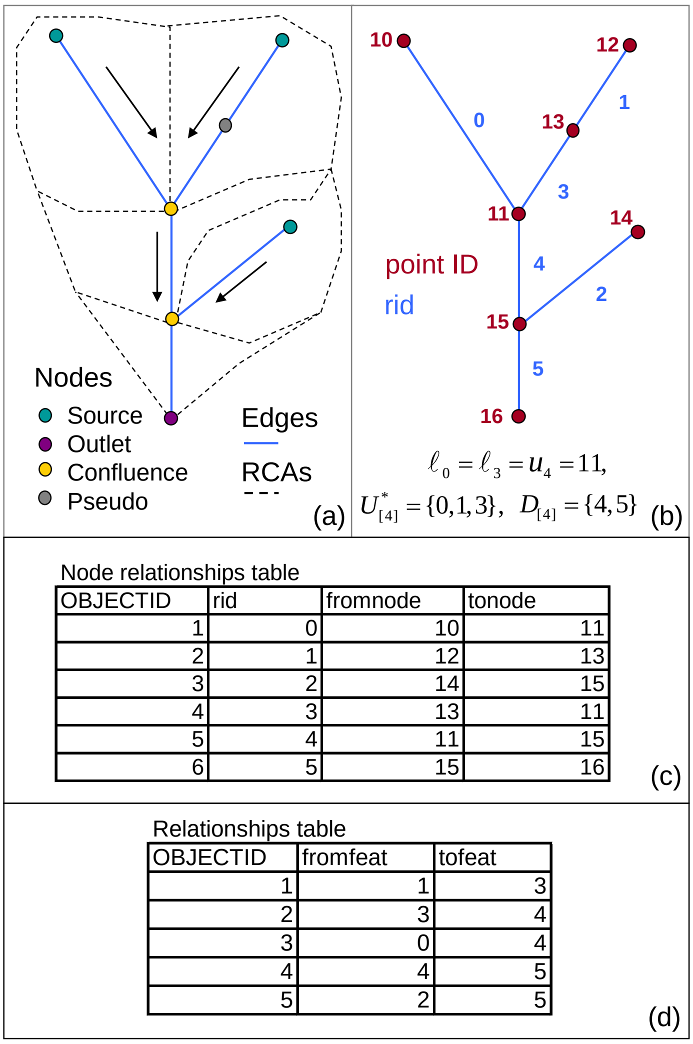
Fitting spatial statistical models to stream network data is challenging because it requires multidisciplinary skills in aquatic ecology, geographic information science, and spatial statistics. In addition, specialized geographic information system (GIS) tools are needed to generate the spatial information needed to fit spatial models to stream network data. Two ArcGIS version 9.3 (ESRI 2009) geoprocessing toolboxes have been provided to help users generate these spatial data: the functional linkage of waterbasins and streams (**FLoWS**) toolbox (Theobald, Norman, Peterson, Ferraz, Wade, and Sherburne 2006) and the spatial tools for the analysis of river systems (**STARS**) toolbox. The **FLoWS** toolbox is a set of graph theoretic-based analysis tools that functionally link aquatic and terrestrial components of the landscape based on hydrologic processes (Theobald *et al.* 2006). These tools provide an efficient framework for navigating throughout the network, which makes it possible to calculate a variety of attributes related to network distance, flow direction, and terrestrial contributing areas.

The **STARS** toolbox was developed to take advantage of the **FLoWS**-based landscape network `LSN` and provides tools to generate and format the feature geometry, attribute data, and topological relationships of GIS datasets so that they may be used to fit spatial statistical models to streams data. The **STARS** tools create a new directory to store this information, which we refer to as a `.ssn` object. Once the data has been reformatted and exported as a `.ssn` object, it can be imported and analyzed in the R environment for statistical computing and graphics (R Core Team 2013) using the **SSN** package (Ver Hoef, Peterson, Clifford, and Shah 2014). This document provides an overview of the `LSN`, an introduction to the **STARS** toolset, and the methods used to create the `.ssn` object.

## 2. Mathematical framework and notation

We represent each stream segment as a line (i.e., edge) bounded by nodes, with multiple stream segments forming a dendritic stream network (Figure 1a). Water flows downstream and so the lines have direction. All of the directed lines within a network drain to a single most-downstream location, referred to as the stream outlet (Figure 1a). Any location on a stream network can be connected by a continuous line to the stream outlet, and hence distance from the outlet is simply the length of that line. We define this as “upstream distance”.

There are a finite number of stream segments within a network and we index them arbitrarily using a reach identifier (rid), $j = 1, 2, \dots, n$ (Figure 1b). Many locations will have the same upstream distance in a dendritic stream network. Therefore, we denote each location as $x^j$, which is the upstream distance on the $j$th stream segment, in order to uniquely define each location and keep track of upstream distance. It is also convenient to arbitrarily assign indices to points in the network, which we denote as $x_i^j$ for the $i$th point, dropping the superscript when knowledge of stream segment is unnecessary. The most downstream location on the $j$th segment is denoted as $\ell^j$, and the most upstream location as $u^j$. As an example, $\ell^0 = \ell^3 = u^4 = x_{11}$ (Figure 1b). The whole set of stream segment indices is denoted as $I$. The index set of stream segments upstream of segment $j$, including $j$ is denoted $U_j$, while the index set of stream segments upstream of segment $j$, excluding $j$ is denoted as $U_j^*$. Likewise, the index set of stream segments downstream of segment $j$, including $j$ is denoted as $D_j$, while the index set of stream segments downstream of segment $j$, excluding $j$ is denoted as $D_j^*$. As an illustration, $U_4^* = \{0, 1, 3\}$ and $D_4 = \{4, 5\}$ in Figure 1b.



Node relationships table

| OBJECTID | rid | fromnode | tonode |
|---|---|---|---|
| 1 | 0 | 10 | 11 |
| 2 | 1 | 12 | 13 |
| 3 | 2 | 14 | 15 |
| 4 | 3 | 13 | 11 |
| 5 | 4 | 11 | 15 |
| 6 | 5 | 15 | 16 |

Relationships table

| OBJECTID | fromfeat | tofeat |
|---|---|---|
| 1 | 1 | 3 |
| 2 | 3 | 4 |
| 3 | 0 | 4 |
| 4 | 4 | 5 |
| 5 | 2 | 5 |

Figure 1: A stream network is made up of a group of directed edges that drain to a single stream outlet (a). Source and outlet nodes occur at the end points of the network, while confluence and pseudo nodes mark locations where edges intersect one another (a). Reach contributing areas (RCAs) form a tessellation of non-overlapping polygons (a). There is a one-to-one relationship between edges and RCAs, which represent the aerial extent that would theoretically contribute overland flow to a given edge. An identifier is allocated to each edge (rid) and node (pointID) in the landscape network (b). The geographic coincidence of nodes and edges is recorded in the node relationships table (c), while the spatial relationship of edges to one another is stored in the relationships table (d). Both the node relationships (c) and relationships (d) tables retain information concerning flow direction in the network.

## 3. FLoWS geoprocessing toolset

The ability to quantify and efficiently represent topological relationships allows users to investigate questions relating to connectivity, adjacency, proximity, and directional relationships in a network. However, some common GIS data formats, such as shapefiles and feature classes, are non-topological data structures; meaning that the topological relationships between features (e.g., edges) are not explicitly stored. Instead, topological relationships must be identified based on the spatial coincidence of features (e.g., edges intersect at confluences or pseudo nodes (Figure 1a), with information about the nodes, edges, and their topological relationships stored in attribute and relationship tables (Fischer 2004). A variety of topological data models have been developed for use in a GIS; for example, `Vector Networks` are used in **GRASS** (Neteler and Mitasova 2008), while the `Network Dataset` and `Geometric Network` are provided in ArcGIS (ESRI 2009). Often, platform-specific toolsets, such as the **v.net** modules for **GRASS** and the **Network Analyst** in ArcGIS, are provided to analyze network data structures. In addition, the ArcGIS **Arc Hydro** toolset (Maidment 2002) is based on the `Geometric Network` and provides an extensive set of tools used to derive new information about network relationships in streams.

The **STARS** toolset relies on a `LSN`, which is a topological data model generated using the **FLoWS** ArcGIS geoprocessing toolset (Theobald *et al.* 2006). The **FLoWS** toolset includes a suite of tools used to analyze relationships in the `LSN` data structure, which have been described elsewhere (Theobald, Norman, Peterson, and Ferraz 2005; Theobald *et al.* 2006). However, we provide an overview of the `LSN` structure and a few particularly useful tools in this section so that users will have a clear understanding of the underlying data structure that the **STARS** toolset depends on; allowing users to generate the same topological information using other data models and software.

A `LSN` is a directional graph used to represent spatial context and relationships with additional geographic information (Theobald *et al.* 2006). It is created using the **FLoWS** toolset and stored as an Environmental Systems Research Institute (ESRI) ArcGIS personal geodatabase (ESRI 2009). The personal geodatabase is stored as an **Access** database, which may contain point, polyline (a.k.a. line), and polygon feature classes. Each feature class is made up of similar spatial features and an **Access** table (i.e., attribute table), which contains attribute data for every spatial feature.

The nodes feature class represents topologic breaks in the stream network such as confluences, stream sources, or stream outlet points, while edges represent flow paths (i.e., line segments) from node to node (Figure 1a). The stream reach contributing area (RCA) corresponds to the aerial extent that would theoretically contribute overland flow to a given edge in the absence of other hydrologic processes such as infiltration or evaporation. RCAs have a one-to-one relationship with edges, and form a non-overlapping, contiguous tessellation of polygons (Figure 1a). The **FLoWS** toolset also allows RCAs to be incorporated into the `LSN`, but they do not need to be explicitly included for spatial statistical modeling. Instead, landscape characteristics such as land use, elevation, or road density, can be summarized as areas, means, or counts for each RCA, RCA$_j$, and then recorded in the edges attribute table.

The `LSN` is based on a `ForwardStar` data structure (Ahuja, Magnanti, and Orlin 1993), which provides a way to store, search, and calculate new spatial information based on the geographic coincidence of features. These spatial relationships are stored within three tables, which may be related to one another through the node and edge identifiers (`pointID` and `rid`, respectively). The `nodexy` table contains four attributes: the `OBJECTID` (internal GIS identifier), `pointID`, and the $x$-, $y$-coordinates for each node. The `noderelationships` table links the nodes to their associated edges based on the `rid`, $j$, and the `pointID` (Figures 1b,c), while the `relationships` table represents the downstream flow paths from edge to edge, which are based on the digitized direction of the line segments. Topological errors are common in GIS data and must be corrected before the `LSN` is constructed to ensure that the `noderelationships` (Figures 1b,c) and `relationships` (Figures 1b,d) tables accurately represent direction and connectivity within the network. Together, these tables provide a wealth of information about the topology of the network, including the spatial location of each node, the most upstream and downstream location on each edge (`fromfeat` $= u^j$ and `tofeat` $= \ell^j$, respectively), the direction of the edges, and the spatial relationship between edge features.

One challenge of working with GIS data is that survey sites collected within a stream usually do not intersect the edge, or segment, where they were collected. This is a common phenomenon that can result for a variety of reasons. Coordinates collected using global positioning system (GPS) units typically have uncertainty associated with them, which translates to locational uncertainty in the site data. Streams have area (i.e., both length and width), but are often represented by lines in a GIS. Consequently, survey sites located on the banks of a stream may not fall directly on a line segment. There are also various degrees of mapping errors and generalizations in digital streams datasets, such as the absence of small tributaries and the homogenization of form. The magnitude of these errors is dependent upon the spatial resolution of the dataset. In addition, the physical location of streams is temporally dynamic, with some segments meandering from their mapped position over time. Regardless of the error source, the survey sites must fall exactly on an edge so that they may be incorporated into the `LSN` framework.

The **FLoWS** tools are used to modify site locations and incorporate site data into the `LSN` (Theobald *et al.* 2006). This includes both observed site locations where measurements were collected, as well as locations where predictions will be generated. Each site, $s_i$, is associated with an edge using dynamic segmentation. The process involves intersecting the site with the closest edge segment, physically moving the site to the new location, $x_i^j$, and calculating the distance ratio, $r_i$, from the most-downstream location on the edge to the site location:

$$r_i = \frac{d(\ell^j, x_i^j)}{L_j}, \tag{1}$$

where $d(\ell^j, x_i^j)$ is the distance travelled along edge $j$ (i.e., hydrologic distance) between locations $\ell^j$ and $x_i^j$ and $L_j$ is the total length of the $j$th edge, $L_j = d(\ell^j, u^j)$.

When the tool has finished successfully, a new point feature class is created in the `LSN` containing the modified site data. The sites attribute table contains the original attribute data, as well as, a number of new fields; the most relevant of which are the `ratio` (field type = double), `rid` (field type = long), `NEAR_X` and `NEAR_Y` (field type = double) fields. The `rid` field indicates which edge the site has been moved to, while `ratio` provides the exact location along the edge. These two pieces of information allow attributes to be calculated for site-specific locations. In addition, the `NEAR_X` and `NEAR_Y` fields provide the $x$-, $y$-coordinates for the modified site locations, which are used to calculate the Euclidean distance matrix in the **SSN** package.

A `LSN` must contain six datasets before it can be used to calculate the spatial information needed to fit the spatial statistical models described by Ver Hoef and Peterson (2010): three feature classes representing the edges, nodes, and survey sites, as well as, three **Access** tables containing information about the spatial relationship between features (Figure 2). Additional feature classes representing prediction locations may also be included. This information is used to navigate upstream and downstream between features, allowing new attributes to be generated based on spatial context and the relationship between features. As such, the `LSN`

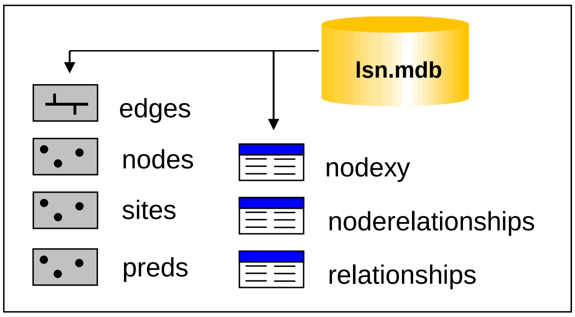


Figure 2: A landscape network (LSN) must contain six datasets before it can be used to calculate the data needed to fit the spatial statistical models: three feature classes: edges, nodes, and sites, as well as three **Access** tables: nodexy, noderelationships, and relationships. Additional feature classes representing prediction locations (preds) may also be included.

provides a powerful and flexible GIS-based hydrologic modeling framework for the **STARS** toolset to operate on. For a more detailed description of the LSN, please see Theobald *et al.* (2006). For step-by-step instructions on how to calculate an LSN appropriate for spatial statistical modeling in stream networks, please see Peterson (2011).

# 4. STARS geoprocessing toolset

The **STARS** geoprocessing toolset is used to generate and format the spatial data needed to fit spatial statistical models to streams data in the **SSN** package. These data include information about hydrologic distances (with flow-direction preserved), the spatial additive function used to calculate the spatial weights, and the covariates for all observed and prediction locations in the stream network. In addition, the topological structure of the network is exported to a format that can be efficiently stored, accessed, and analyzed in the **SSN** package. The **STARS** toolset was developed for ArcGIS version 9.3.1 (ESRI 2009), which runs on a Windows operating system (OS), with scripts written in Python version 2.5 (Hammond and Robinson 2000). A Python editor for Windows, **PythonWin** (Hammond and Robinson 2000), must also be installed to successfully run the toolset.The **STARS** toolset contains eight tools that are used to transform an LSN into a .ssn object (Figure 3), which is the data structure the **SSN** package has been designed to utilize. The tools are grouped into three general categories: Pre-processing, Calculate, and Export.

## 4.1. Pre-processing

### *Identify complex confluences*

It is *extremely* critical that the edges are topologically correct to ensure that hydrologic distances and spatial relationships are calculated properly using the **STARS** toolset. The process of topologically correcting the data can easily be the most time-consuming aspect of the modeling process, especially for large datasets. There are a number of useful tools provided in the **FLoWS** toolset to help identify and remove errors (Theobald *et al.* 2006) and so the majority of the pre-processing occurs before the final LSN is generated. However,

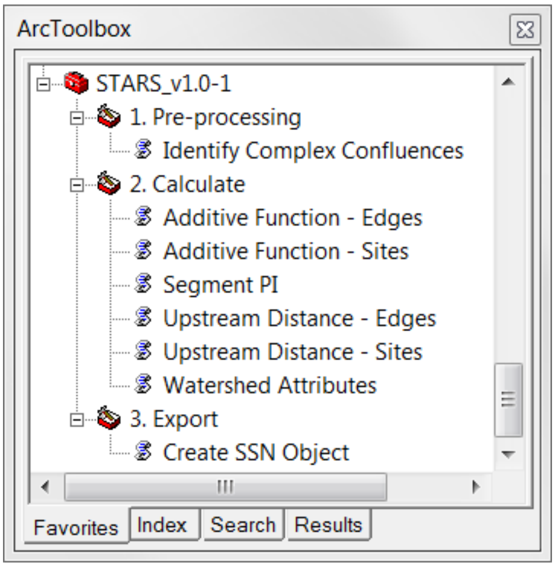


Figure 3: The spatial tools for the analysis of river systems (**STARS**) toolset contains eight tools that are used to transform a landscape network into a `.ssn` object. The tools are grouped into three general categories: `Pre-processing`, `Calculate`, and `Export`.

`LSN`s used to create `.ssn` objects have two unique topological restrictions, which are not considered a true topological error in a GIS or within a `LSN`. First, converging stream nodes are not allowed. These nodes occur at the downstream node of two edges that converge (Figures 4a,b), but do not flow into another downstream edge. This commonly occurs at the boundaries of the streams dataset (Figure 4a) or may be the result of topological errors within the stream network (Figure 4b). Converging nodes are identified using the **FLoWS** `Check Network Topology` tool and must be manually removed when the `LSN` is being generated. The second restriction is that only two edges may converge and flow into a single downstream edge at a confluence. The `Identify Complex Confluences` tool has been included in the **STARS** toolset to help users to identify `LSN` nodes that violate this condition (Figure 4c). The sole input to the tool is the `LSN` and a text file is produced that contains the `pointID` for every node with more than 2 upstream edges (Figure 5a). Once these errors have been identified, they must be manually edited and a new `LSN` generated before a `.ssn` object can be created. Note, only the topological relationships shown in Figure 1a are permitted in an `LSN` used to generate a spatial statistical model using the **SSN** package.

## 4.2. Calculate

Six tools are provided to calculate the spatial data necessary for spatial statistical modeling: `Watershed Attributes`, `Segment PI`, `Additive Function - Edges`, `Additive Function - Sites`, `Upstream Distance - Edges`, and `Upstream Distance - Sites`.

### *Watershed attributes*

The RCA is used to characterize landscape conditions found near each edge (Figure 1a), but in most cases it is only a subcomponent of the watershed; the exception being source edges

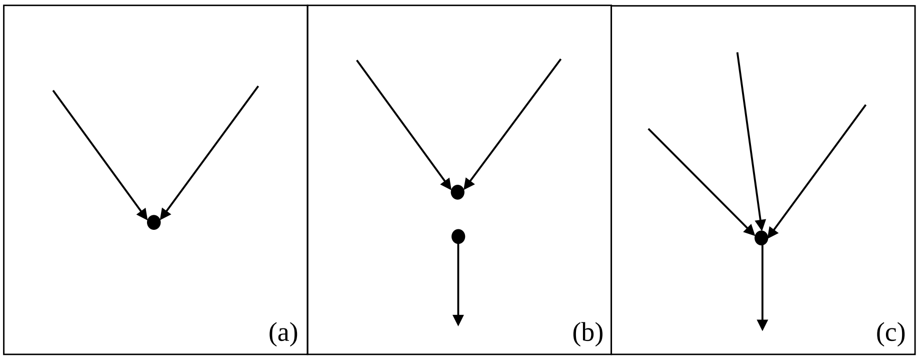


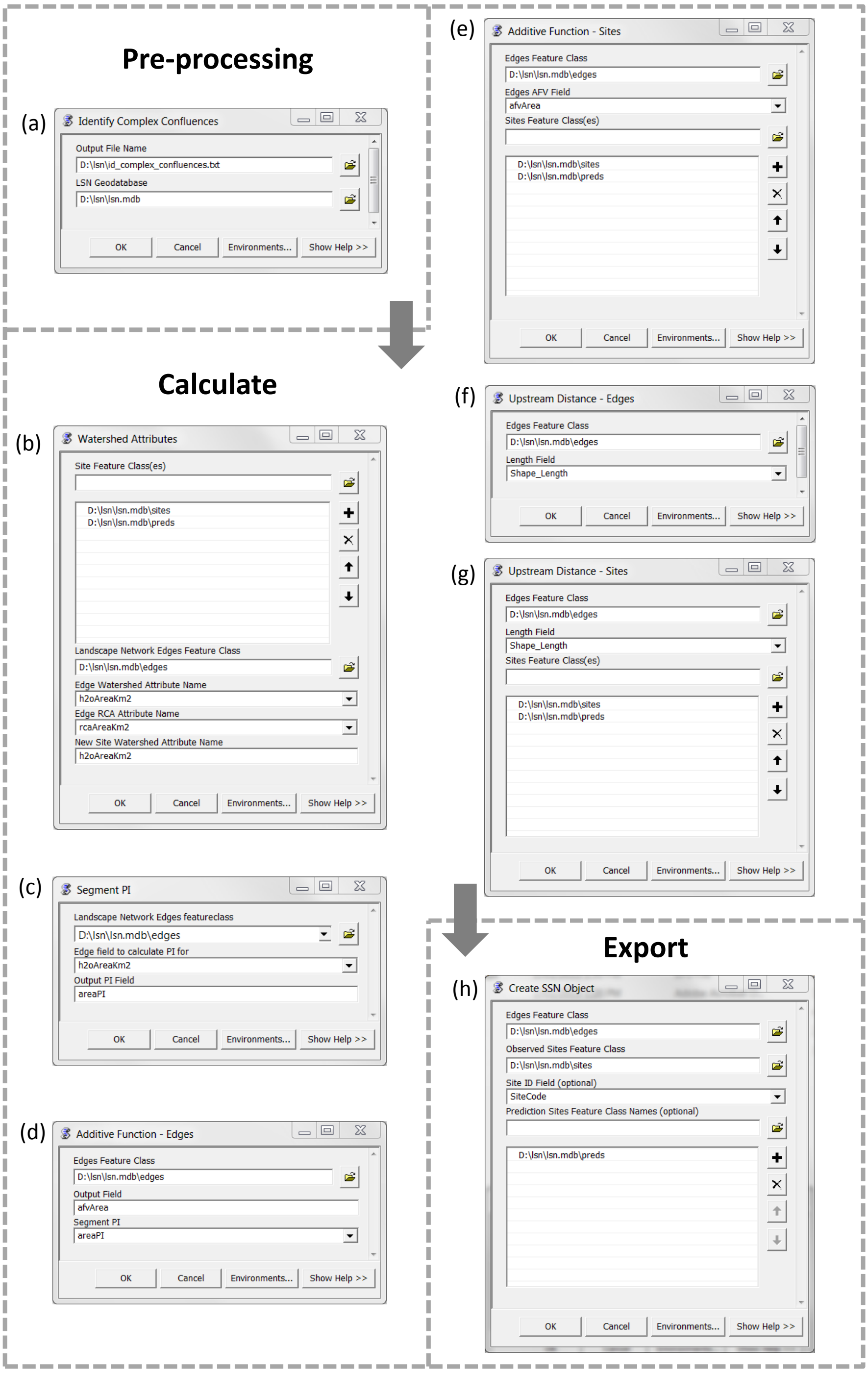
Figure 4: Landscape networks (`LSN`s) used to create `.ssn` objects have two unique topological restrictions; converging nodes (a, b) and complex confluences (c) are not allowed. Converging nodes occur at the downstream node of two edges that converge (a, b), but do not flow into another downstream edge. Complex nodes occur anytime more than two edges converge at, or flow into, a node (c).

where the watershed and the RCA are equivalent (Figure 1a). Instead, the watershed is commonly used as an analytical unit to summarize characteristics about the landscape that have the potential to affect in-stream conditions (Johnson and Host 2010). For example, common watershed attributes of interest include area, proportional land use or land cover type (e.g., urban area / watershed area), or the number of road crossings upstream. The **FLoWS** toolset provides the capability to sum edge attributes downstream (`Accumulate Values Downstream` tool). When an RCA attribute is summed, a watershed attribute is generated for the most downstream location on each edge. Yet, survey sites may fall anywhere along an edge. The **STARS** `Calculate Watershed Attributes` tool was developed to estimate site-specific watershed attributes:

$$W_i = (1 - r_i)\mathrm{RCA}_i + \sum_{k \in U_i^*} \mathrm{RCA}_k$$

where $W_i$ is the watershed attribute for each survey site $i$, $\mathrm{RCA}_i$ is an attribute summarized over the RCA where the $i$th site resides, and $r_i$ is as defined in Equation 1. Inputs to the tool include the sites feature class, the edges feature class, the watershed attribute name, and the RCA attribute name (Figure 5b). A new field is added to the sites attribute table that contains the $W_i$ for each site. Note, $W_i$ is simply an estimate of the true watershed attribute because $\mathrm{RCA}_i$ does not contain spatial information about attribute variability at the sub-RCA level.

Although it is relatively simplistic, the `Watershed Attributes` tool is useful for extracting watershed attributes from publically available, nationally-attributed stream datasets, such as the United States **NHDPlus** (Horizon Systems Corporation 2007) and the Australian Hydrological Geospatial Fabric (**Geofabric**) Surface Network dataset (Bureau of Meteorology 2012). In both cases, each edge is associated with an RCA; though they are referred to as catchments and subcatchments in the **NHDPlus** and **Geofabric**, respectively. Environmental attributes, such as land use, number of road crossings or dams, or climate statistics are provided at the RCA scale in separate relational databases (Stein, Hutchinson, and Stein 2012; United States Geological Survey 2010). These attributes may be joined to the edges and summed down-



(a) Identify Complex Confluences

Output File Name: D:\lsn\id_complex_confluences.txt

LSN Geodatabase: D:\lsn\lsn.mdb

(b) Watershed Attributes

Site Feature Class(es):
D:\lsn\lsn.mdb\sites
D:\lsn\lsn.mdb\preds

Landscape Network Edges Feature Class: D:\lsn\lsn.mdb\edges

Edge Watershed Attribute Name: h2oAreaKm2

Edge RCA Attribute Name: rcaAreaKm2

New Site Watershed Attribute Name: h2oAreaKm2

(c) Segment PI

Landscape Network Edges featureclass: D:\lsn\lsn.mdb\edges

Edge field to calculate PI for: h2oAreaKm2

Output PI Field: areaPI

(d) Additive Function - Edges

Edges Feature Class: D:\lsn\lsn.mdb\edges

Output Field: afvArea

Segment PI: areaPI

(e) Additive Function - Sites

Edges Feature Class: D:\lsn\lsn.mdb\edges

Edges AFV Field: afvArea

Sites Feature Class(es):
D:\lsn\lsn.mdb\sites
D:\lsn\lsn.mdb\preds

(f) Upstream Distance - Edges

Edges Feature Class: D:\lsn\lsn.mdb\edges

Length Field: Shape_Length

(g) Upstream Distance - Sites

Edges Feature Class: D:\lsn\lsn.mdb\edges

Length Field: Shape_Length

Sites Feature Class(es):
D:\lsn\lsn.mdb\sites
D:\lsn\lsn.mdb\preds

(h) Create SSN Object

Edges Feature Class: D:\lsn\lsn.mdb\edges

Observed Sites Feature Class: D:\lsn\lsn.mdb\sites

Site ID Field (optional): SiteCode

Prediction Sites Feature Class Names (optional):
D:\lsn\lsn.mdb\preds

Figure 5: Eight tools are included in the **STARS** geoprocessing toolset: the (a) `Identify Complex Confluences`, (b) `Watershed Attributes`, (c) `Segment Proportional Influence (PI)`, (d) `Additive Function - Edges`, (e) `Additive Function - Sites`, (f) `Upstream Distance - Edges`, (g) `Upstream Distance - Sites`, and (h) `Create SSN Object` tools.

stream using the **FLoWS** `Accumulate Values Downstream` tool before extracting estimates of watershed attributes using the `Watershed Attributes` tool. However, other more spatially-explicit methods may also be used to calculate site-specific watershed attributes. For example, the ArcGIS **Arc Hydro** toolset can be used to calculate spatially explicit watershed attributes when source data, such as land use or climate, are available. Raster-based approaches may also be implemented independently of a topological data model using the ArcGIS **Spatial Analyst** extension (ESRI 2009), **FRAGSTATS** (McGarigal, Cushman, and Ene 2012), or a suite of **GRASS** toolsets, such as **r.stream** (Jasiewicz and Metz 2011), **r.watershed**, **r.flow**, and **r.stats** (Neteler and Mitasova 2008), among others. In addition, a variety of deterministic models could also be used to derive watershed characteristics (e.g., Rathburn and Wohl 2001). Given the plethora of sophisticated and well-documented methods used to extract watershed attributes, automated approaches used to calculate spatially explicit watershed attributes have not been included in the **STARS** toolset.

### *Segment proportional influence*

Calculating the spatial weights needed to fit a spatial statistical model to streams data is a three step process: 1) calculating the segment proportional influence (PI), 2) calculating the additive function values, and 3) calculating the spatial weights (Peterson and Ver Hoef 2010). Steps 1 and 2 are performed using the **STARS** toolset, while step 3 is undertaken in R using the **SSN** package.

The PI for each edge, $\omega_j$, is defined as the relative influence that the $j$th edge, or segment, has on the edge directly downstream. In the following example, $\omega_j$ is based on watershed area, but other simple measures, such as Shreve's stream order (Shreve 1966), could also be used (e.g., Cressie, Frey, Harch, and Smith 2006). To begin, watershed area is calculated for the most downstream location of each edge in the network: $A^j = \sum_{k \in U_i} \mathrm{RCA}_k$, using the **FLoWS** `Accumulate Values Downstream` tool. When two edges, denoted $j$ and $j'$, join at a node, the PI, $\omega_j$, for the $j$th edge that flows into the node is then:

$$\omega_j = \frac{A^j}{A^j + A^{j'}}.$$

Note that, the $\omega_j$ values for edges directly upstream from a single confluence always sum to 1 because they are proportions. Inputs to the `Segment PI` tool include the `LSN` edges feature class, as well as, the field in the edges attribute table that will be used to calculate the segment PI. The tool outputs a new field to the `edges` attribute table (field type = double) that contains $\omega_j$ (Figure 5c).

### *Additive function – Edges and sites*

Two tools have been provided in the **STARS** toolset, which are used to calculate the additive function value (AFV) for every edge and site in the `LSN`: `Additive Function - Edges` and `Additive Function - Sites`. Separate tools are provided so that the AFV can be calculated for multiple sets of sites without having to recalculate the values for the edges.

The AFV for the $j$th edge, $\mathrm{AFV}_j$, is equal to the product of the segment PIs found in the path downstream from the $j$th edge to the stream outlet:

$$\mathrm{AFV}_j = \prod_{k \in D_j} \omega_k.$$

Given that $\omega_j$ is a proportion, the $\mathrm{AFV}_j$ will always range between 0 and 1, with the $\mathrm{AFV}_j$ for the most downstream edge in the network equal to 1. The AFV for the $i$th site, $\mathrm{AFV}_i$, is simply equal to the $\mathrm{AFV}_j$ of the edge it lies on. As a result, multiple sites located on a single edge will always have the same AFV value.

Inputs to the `Additive Function - Edges` tool include the `LSN` edges feature class and the segment PI value generated using the `Segment PI` tool. The `LSN` edges feature class is also an input to the `Additive Function - Sites` tool, as well as, the `LSN` sites feature class(es) that the AFV values will be assigned to. The `Additive Function - Edges` (Figure 5d) and `Additive Function - Sites` (Figure 5e) tools create a new field in both the edges and sites attribute tables representing the AFV (field type = double). For additional details, please see Peterson and Ver Hoef (2010) and Peterson (2011).

### *Upstream distance – Edges and sites*

There are two tools provided in the **STARS** toolset to calculate the upstream distance for each of the edges and sites: `Upstream Distance - Edges` and `Upstream Distance - Sites`. The `Upstream Distance - Edges` tool is used to calculate:

$$\mathrm{upDist}_j = \sum_{k \in D_j} L_k,$$

where $\mathrm{upDist}_j$ is the upstream distance from the stream outlet to the upper end of the $j$th edge ($u^j$) and $L_j$ is the total length of the $j$th edge. The `Upstream Distance - Sites` tool is used to compute the upstream distance from the outlet to the $i$th site, $\mathrm{upDist}_i$, i.e., the $x$ part of $x_i^j$, and is given by:

$$\mathrm{upDist}_i = r_i L_i + \sum_{k \in D_j^*} L_k,$$

where $L_i$ is the total length of the $j$th edge of $x_i^j$ and recall that $D_j^*$ is the set of all segments downstream of $x_i^j$, excluding the $j$th segment.

Inputs to the `Upstream Distance - Edges` (Figure 5f) and `Upstream Distance - Sites` (Figure 5g) tools include the `LSN` edges feature class and the field in the edges feature class that contains the length of each segment. The sites feature class(es) that the upstream distance values will be assigned to must also be specified in the `Upstream Distance - Sites` tool. The $\mathrm{upDist}_j$ and $\mathrm{upDist}_i$ values are recorded in the edges and sites attribute tables, respectively, within a new field named `upDist` (field type = double). The information stored in the two upstream distance attributes provides part of the spatial information needed to calculate pair-wise hydrologic distances between locations in the **SSN** package.

## 4.3. Export

### *Create SSN object*

The purpose of the `Create SSN Object` tool is to reformat the feature geometry, attribute data, and topological relationships of each spatial dataset contained in the `LSN` into a `.ssn` object so that it can be imported into R using functions provided in the **SSN** package. As we mentioned previously, a new directory is created to store this information, with the naming convention `lsn-name.ssn` (i.e., `lsn.ssn`); we refer to this as the `.ssn` object (Figure 6).

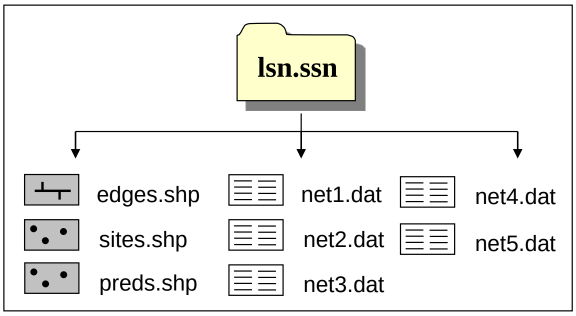


Figure 6: The `.ssn` object contains the spatial, attribute, and topological information of the landscape network (`LSN`). It will always contain two shapefiles: edges and sites, as well as multiple text files containing the edge binary identifiers for each stream network in the `LSN`. Multiple shapefiles representing the prediction locations may also be included.

Shapefiles can be easily imported into `R` using the **maptools** package (Bivand, Pebesma, and Gómez-Rubio 2008), with all of the associated shape geometry and attributes. Consequently, the spatial datasets (edges and sites) are converted from ESRI (2009) feature class to shapefile format and stored in the `.ssn` object directory. However, shapefiles in standard form cannot be used to represent the topological relationships of the `LSN`. Our solution was to generate network and binary identifiers (IDs), which provide an efficient way to assess connectivity and spatial relationships between features on the network.

The process of assigning binary IDs is relatively straightforward. First, the outlet edge (i.e., the most downstream edge in the network) is identified and assigned a binary ID equal to 1 (Figure 7). The information stored in the relationships table (Figure 1d) is used to identify edges that are directly upstream from the outlet edge. Binary IDs are assigned to the upstream edge(s) by arbitrarily appending a 0 or 1 to the downstream binary ID. For example, binary IDs 10 and 11 are directly upstream from binary ID 1 in Figure 7. This process of moving upstream and assigning binary IDs continues until every edge in the stream network has been assigned a binary ID.

The binary IDs are useful because they contain information about whether locations have a flow-connected or flow-unconnected relationship. Two locations are considered flow-connected when water flows from an upstream location to a downstream location. In contrast, two locations are flow-unconnected when they share a common outlet downstream, but water does not flow between them. In Figure 7, the binary ID for the edge where site B resides is completely nested within the binary ID for the edge where site A lies ("1" is nested within "1110"), which indicates that the two locations are flow-connected. In contrast, the binary IDs for the edges where sites C and D reside are not nested ("1110" is not nested within "1111") because the two locations are flow-unconnected. The binary IDs for flow-unconnected locations also contain information about the closest common downstream location. As an example, the most common downstream confluence between sites C and D is $u^{111}$ (Figure 7) and so "111" is where the two binary IDs diverge.

It is common for `LSN`s to contain multiple stream networks, with unique stream outlets. As an example, consider a streams dataset with two individual stream networks in the edges feature class (Figure 7). In this case, two edges from different networks will have the same binary ID. Thus, a network identifier (`netID`, field type = long) is also assigned to the edges,

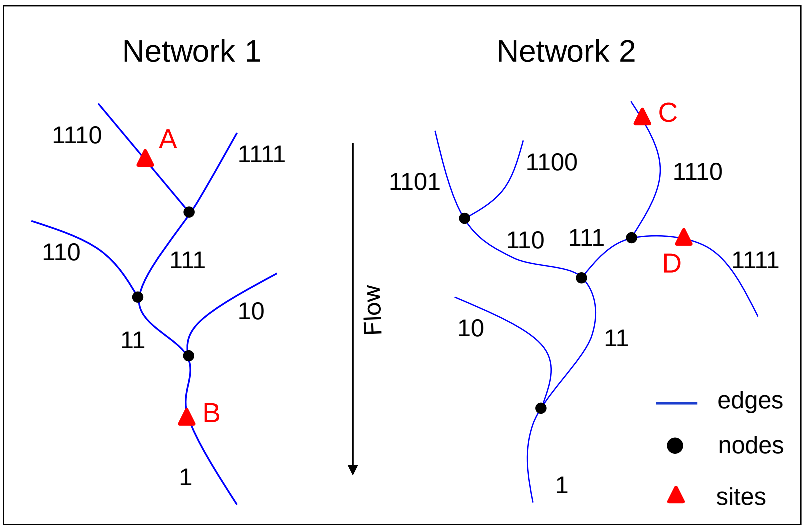
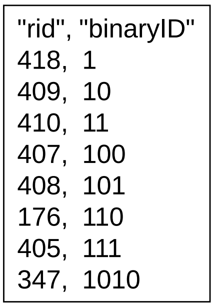


Figure 7: Binary identifiers (`binaryID`) are assigned to each edge in the landscape network.



```
"rid", "binaryID"
418,  1
409,  10
410,  11
407,  100
408,  101
176,  110
405,  111
347,  1010
```

Figure 8: Binary identifiers (`binaryID`) and their associated reach identifier (`rid`) are stored in a comma delimited text file format within the `.ssn` object.

sites, and prediction sites (if included) attribute tables to differentiate between locations. In addition, a location ID (`locID`, field type = long) and a unique point ID (`pid`, field type = long) are assigned to the sites and prediction sites attribute tables in order to distinguish between repeated measurements at a single location.

Inputs to the `Create SSN Object` tool include the `LSN` edges feature class and the `LSN` observed sites feature class (Figure 5h). Optional inputs include a site identifier (ID) field and a prediction sites feature class. As an output, the `Create SSN Object` tool stores the `rid` and `binaryID` for each edge in a comma delimited text file (Figure 8), with a separate file for each network. The naming convention for these files corresponds to the network ID (i.e., `net1.dat`, `net2.dat`, etc.). Once the binary and network IDs have been calculated, the edges, sites, and prediction sites feature classes are converted to shapefiles and exported to the `.ssn` object directory. When the `Create SSN Object` tool is complete, the `.ssn` object will have the file structure shown in Figure 6 and will contain the information necessary to fit a spatial statistical model using the **SSN** package.

## 5. Accessing the tools

The **FLoWS** and **STARS** custom ArcGIS toolsets, as well as the **SSN** package for the R environment for statistical computing and graphics are freely available through the "**SSN** and **STARS**" website hosted by the U.S. Forest Service, Boise Aquatic Science Lab at http://www.fs.fed.us/rm/boise/AWAE/projects/SpatialStreamNetworks.shtml. A tutorial for **STARS** (Peterson 2011) is also available from the website; it is written for users who have relatively little experience with GIS and provides detailed instructions about how to generate the spatial information needed to fit spatial statistical models for stream networks (Ver Hoef and Peterson 2010). An example dataset is available for download and is referred to throughout the tutorial.

## 6. Future developments

The **STARS** toolset contains all of the functionality needed to create the `.ssn` object, which can then be imported into the **SSN** package and used to fit a spatial statistical model to stream data. The `LSN` is currently stored as a personal geodatabase, which provides a self-contained data structure for multiple feature types (e.g., edges, nodes, survey sites). However, the personal geodatabase is not a recommended data format in ArcGIS versions released after 9.3. Therefore, the next major development will be to select a new data format for the `LSN`. The ArcGIS file geodatabase provides a straight-forward alternative; it is similar in structure to the personal geodatabase and an application programming interface (API) has been released for 32 and 64-bit Windows and Linux OS, which should provide better non-ArcObjects based access to the data. In addition, the **FLoWS** and **STARS** toolsets could be merged to provide a single custom toolset solely used for spatial statistical modeling on stream networks. A second alternative is to redesign the **STARS** tools to take advantage of the ESRI `Geometric Network` and the **Arc Hydro** toolset. **Arc Hydro** does not currently provide out-of-box tools to calculate and format the spatial information needed to create a `.ssn` object. However, a conversion to this data model would allow users to take advantage of the extensive set of stream-network analysis tools provided in **Arc Hydro**, such as more spatially explicit methods for calculating, rather than estimating, watershed attributes. A third alternative would be to recreate the **FLoWS** and **STARS** functionality in an open-source GIS software, such as **GRASS**, that can be directly linked with R using the **spgrass6** package (Bivand *et al.* 2008). Although the third alternative requires more labor to implement, it is attractive for a number of reasons. First, the ability to link an active GIS session directly with R reduces the chances of user error when files are transferred between software packages. Second, a direct linkage would allow some of the current functionality, such as calculating covariates for observed and prediction sites, to be more accessible in R. This would allow users who are unfamiliar with GIS software to calculate new covariates without having to move between software programs. Third, ArcGIS is proprietary software, while R is open-source with a vast community of contributors and users. Implementing a modeling framework based on open-source GIS and statistical software would not only be powerful, but also philosophically appropriate. In the meantime, we have tried to provide enough detail about the structure of the `.ssn` object so that users can recreate it using other, proprietary and non-proprietary GIS software packages and data models if necessary.

## Acknowledgments

Support for this work was provided by the United States (US) Forest Service, the US Geological Survey, and the Oregon State Office of the Bureau of Land Management. Some of this work was conducted as part of the working group entitled "Spatial Statistical Models for Stream Networks," supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant #EF-0553768), the University of California, Santa Barbara, and the State of California. This project also received financial support from the CSIRO Water for a Healthy Country Flagship, and the NOAA's National Marine Fisheries Service to the Alaska Fisheries Science Center. We thank Andy Last, Dona Horan, and Jenny Frieden for their assistance in testing the **STARS** tools and developing the example dataset included in http://www.fs.fed.us/rm/boise/AWAE/projects/SSN_STARS/software_data.html#doc. We also thank David Clifford for his constructive review of a previous version of this manuscript. Reference to trade names, and the findings and conclusions in the paper, are those of the authors, and do not imply endorsement by the National Marine Fisheries Service, NOAA.

## References

Ahuja RK, Magnanti TL, Orlin JB (1993). *Network Flows: Theory, Algorithms, and Applications.* Prentice-Hall.

Bivand RS, Pebesma EJ, Gómez-Rubio V (2008). *Applied Spatial Data Analysis with R.* Springer-Verlag, New York.

Bureau of Meteorology (2012). "Australian Hydrological Geospatial Fabric (**Geofabric**) Product Guide." URL http://www.bom.gov.au/water/geofabric/documentation.shtml.

Cressie N, Frey J, Harch B, Smith M (2006). "Spatial Prediction on a River Network." *Journal of Agricultural, Biological, and Environmental Statistics*, **11**(2), 127–150.

ESRI (2009). "ArcGIS Desktop: Release 9.3.1." *Technical report*, Environmental Systems Research Institute, Redlands, California.

Fischer MM (2004). "GIS and Network Analysis." In *Handbook of Transport Geography and Spatial Systems*, pp. 391–408. Elsevier, Amsterdam.

Hammond M, Robinson A (2000). *Python Programming on Win32: Help for Windows Programmers.* O'Reilly, Sebastopol.

Horizon Systems Corporation (2007). "National Hydrography Dataset Plus: Documentation." URL http://www.horizon-systems.com/NHDPlus/NHDPlusV2_documentation.php.

Jasiewicz J, Metz M (2011). "A New **GRASS** GIS Toolkit for Hortonian Analysis of Drainage Networks." *Computers and Geosciences*, **37**(8), 1525–1531.

Johnson LB, Host GE (2010). "Recent Developments in Landscape Approaches for the Study of Aquatic Ecosystems." *Journal of the North American Benthological Society*, **29**(1), 41–66.

Maidment D (ed.) (2002). ***Arc Hydro**: GIS for Water Resources.* ESRI Press.

McGarigal K, Cushman SA, Ene E (2012). ***FRAGSTATS** v4: Spatial Pattern Analysis Program for Categorical and Continuous Maps.* University of Massachusettes, Amherst, Massachusettes. URL http://www.umass.edu/landeco/research/fragstats/fragstats.html.

Neteler M, Mitasova H (2008). *Open Source GIS: A **GRASS** GIS Approach.* Number 773 in The International Series in Engineering and Computer Science, 3rd edition. Springer-Verlag.

Peterson EE (2011). "**STARS**: Spatial Tools for the Analysis of River Systems – A Tutorial." *Technical Report EP111313*, Commonwealth Scientific Industrial Research Organisation (CSIRO). URL http://www.fs.fed.us/rm/boise/AWAE/projects/SSN_STARS/software_data.html#doc.

Peterson EE, Ver Hoef JM (2010). "A Mixed-Model Moving-Average Approach to Geostatistical Modeling in Stream Networks." *Ecology*, **93**(3), 644–651.

Rathburn SL, Wohl EE (2001). "One-Dimensional Sediment Transport Modeling of Pool Recovery Along a Mountain Channel after a Reservoir Sediment Release." *Regulated Rivers: Research and Management*, **17**(3), 251–273.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Shreve RL (1966). "Statistical Law of Stream Numbers." *Journal of Geology*, **74**(1), 17–37.

Stein JL, Hutchinson MF, Stein JA (2012). "National Catchment and Stream Attributes Database Version 1.1.5." URL http://www.ga.gov.au/topographic-mapping/national-surface-water-information.html.

Theobald D, Norman J, Peterson EE, Ferraz S (2005). "Functional Linkage of Watersheds and Streams (**FLoWS**): Network-Based ArcGIS Tools to Analyze Freshwater Ecosystems." In *Proceedings of the 2005 ESRI International User Conference*. ESRI Press.

Theobald D, Norman J, Peterson EE, Ferraz S, Wade A, Sherburne MR (2006). *Functional Linkage of Waterbasins and Streams (**FLoWS**) v1 User's Guide: ArcGIS Tools to Analyze Freshwater Ecosystems.* Natural Resource Ecology Lab, Colorado State University, Fort Collins, Colorado.

United States Geological Survey (2010). "Attributes for **NHDPlus** Catchments (Version 1.1) for the Conterminous United States (DS-490)." USGS National Water-Quality Assessment (NAWQA) Program, URL http://www.bom.gov.au/water/geofabric/documentation.shtml.

Ver Hoef JM, Peterson EE (2010). "A Moving Average Approach for Spatial Statistical Models of Stream Networks." *Journal of the American Statistical Association*, **105**(489), 6–18.

Ver Hoef JM, Peterson EE, Clifford D, Shah R (2014). "**SSN**: An R Package for Spatial Statistical Modeling on Stream Networks." *Journal of Statistical Software*, **56**(3), 1–43. URL http://www.jstatsoft.org/v56/i03/.

Ver Hoef JM, Peterson EE, Theobald D (2006). “Spatial Statistical Models That Use Flow and Stream Distance.” *Environmental and Ecological Statistics*, **13**(1), 449–464.

## Affiliation:

Erin E. Peterson
Division of Computational Informatics
Commonwealth Scientific and Industrial Research Organisation (CSIRO)
PO Box 2583
Brisbane, QLD 4001, Australia
E-mail: Erin.Peterson@csiro.au
URL: http://www.csiro.au/org/CMIS

Jay M. Ver Hoef
NOAA National Marine Mammal Laboratory
NMFS Alaska Fisheries Science Center
International Arctic Research Center, Room 351
University of Alaska Fairbanks
Fairbanks, AK 99775-7345, United States of America
E-mail: jay.verhoef@noaa.gov
URL: http://sites.google.com/site/jayverhoef/

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/

Volume 56, Issue 2 *Submitted:* 2012-08-24
January 2014 *Accepted:* 2013-04-22