

Erin E. Peterson and Jay M. Ver Hoef. 2010. A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology* 91:644–651.

Appendix A: Calculating covariance matrices using the tail-up and tail-down models.

Computing Hydrologic Distances

A distance matrix that contains the hydrologic distance between any two sites in a study area is needed to fit a geostatistical model using the tail-up and tail-down autocovariance functions. However, the hydrologic distance information needed to model the covariance between flow-connected and flow-unconnected locations differs in most cases. The total hydrologic distance is a directionless measure; it represents the hydrologic distance between two sites, ignoring flow direction. The hydrologic distance from each site to a common confluence is generally used when creating models for flow-unconnected pairs (Fig. A1.A), which we term “downstream hydrologic distance”. In contrast, the total hydrologic distance is used for modeling flow-connected pairs (Fig. A1.B), which we term “total hydrologic distance”.

A downstream hydrologic distance matrix provides enough information to meet the data requirements for both the tail-up and tail-down models. When two locations are flow-connected, the downstream hydrologic distance from the upstream location to the downstream location is greater than zero, but it is zero in the other direction. Using the data contained in Fig. A1 as an example, the

downstream hydrologic distance from s_2 to $s_3 = 3 + 5 = 8$ (Fig. A1.A). In contrast, the downstream hydrologic distance from s_3 to $s_2 = 0 + 0 = 0$. When two locations are flow-unconnected the downstream hydrologic distance will be greater than zero in both directions. For example, the downstream hydrologic distance from s_1 to $s_2 = 7$, while the downstream hydrologic distance from s_2 to $s_1 = 3$ (Fig. A1.A). Notice that a site's downstream hydrologic distance to itself is equal to zero. When two sites do not reside on the same stream network (do not share a common stream outlet anywhere downstream) there is no hydrologic path between the pair and as such, a hydrologic distance cannot be calculated. To address this issue, an extremely large downstream hydrologic distance is recorded in both directions, essentially equal to infinity. This ensures that the downstream hydrologic distance is greater than the range parameter and no spatial correlation will be permitted between sites. The format of the downstream hydrologic distance matrix is efficient because only one distance matrix is needed to fit both the tail-up and tail-down models. For example, a matrix containing the total hydrologic distance between sites is easily calculated by adding the downstream distance matrix to its transpose (Fig. A1.B).

Tail-up models

Tail-up models represent flow-connected relationships and so water must flow from one location to another for two sites to be spatially correlated. The total hydrologic distance (Fig. A1.B) does not represent flow direction; instead, the spatial weights are used to restrict the symmetric correlation to include only flow-connected sites. A variety of weighting schemes may be used in the tail-up

moving-average function, but there are rules about the way that weights may be constructed in order to maintain stationary variances.

The construction of the spatial weights will be explained in greater detail in subsequent sections.

A valid tail-up autocovariance between flow-connected locations on the stream network can be constructed as

$$C_{TU}(s_i, s_j | \boldsymbol{\theta}) = \begin{cases} 0 & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected} \\ \prod_{k \in B_{s_i, s_j}} \sqrt{w_k} C_1(h | \boldsymbol{\theta}) & \text{if } s_i \text{ and } s_j \text{ are flow-connected} \end{cases} \quad (\text{A.1})$$

where $C_1(h | \boldsymbol{\theta})$ is the unweighted autocovariance function constructed using a tail-up moving-average function which depends on

parameters $\boldsymbol{\theta}$, h is the total hydrologic distance, and $\prod_{k \in B_{s_i, s_j}} \sqrt{w_k}$ represents the spatial weights; we describe four such models derived by

Ver Hoef et al. (2006) next.

Although Eq. A.1 may appear unfamiliar, the practical construction is relatively straight-forward. For example, $C_1(h | \boldsymbol{\theta})$ could be

$$C_1(h | \boldsymbol{\theta}) = \sigma_{TU}^2 \left(1 - \frac{h}{\alpha}\right) I\left(\frac{h}{\alpha} \leq 1\right) \quad (\text{A.2})$$

where h is the total hydrologic distance, $\boldsymbol{\theta}$ is the parameter vector containing $\sigma_{TU}^2 > 0$ (the partial sill or variance component in the mixed model) and $\alpha > 0$ (the spatial range parameter), and $I(\bullet)$ is the indicator function. Eq. A.2 combined with Eq. A.1 is the tail-up linear-with-sill model. When $C_1(h | \boldsymbol{\theta})$ is

$$C_1(h | \boldsymbol{\theta}) = \sigma_{TU}^2 \exp(-h / \alpha) \quad (\text{A.3})$$

and combined with Eq. A.1, it is the tail-up exponential model. When $C_1(h | \boldsymbol{\theta})$ is

$$C_1(h | \boldsymbol{\theta}) = \sigma_{TU}^2 \left(1 - \frac{3h}{2\alpha} + \frac{1}{2} \frac{h^3}{\alpha^3} \right) I\left(\frac{h}{\alpha} \leq 1\right) \quad (\text{A.4})$$

and combined with Eq. A.1, it is the tail-up spherical model. Finally, when $C_1(h | \boldsymbol{\theta})$ is

$$C_1(h|\boldsymbol{\theta}) = \begin{cases} \sigma_{TU}^2 \frac{\log(h/\alpha + 1)}{h/\alpha} & \text{if } h > 0 \\ \sigma_{TU}^2 & \text{if } h = 0 \end{cases} \quad (\text{A.5})$$

and combined with Eq. A.1 it is the tail-up mariah model.

Note that creating a covariance matrix, \mathbf{C}_1 , based on $C_1(h|\boldsymbol{\theta})$ alone, is not guaranteed to be a valid covariance matrix until it has been weighted appropriately using the spatial weights matrix, \mathbf{W} , where the i,j^{th} element is $\mathbf{W}[i,j] = \prod_{k \in B_{s_i, s_j}} \sqrt{w_k}$ (Eq. A.1). The Hadamard (element-wise) product is applied to the two matrices, $\boldsymbol{\Sigma}_{TU} = \mathbf{W} \square \mathbf{C}_1$, and the product is a covariance matrix, $\boldsymbol{\Sigma}_{TU}$, that meets the statistical assumptions necessary for geostatistical modeling (Cressie et al. 2006; Ver Hoef et al. 2006).

Weighting for the Tail-up Model

The tail-up moving-average function is named as such because the tail of the function points in the upstream direction. Since stream networks are dendritic, a weighting scheme must be used to proportionally allocate (i.e., they must sum to 1), or split, the tail-up moving-average function between upstream segments. The function could simply be split evenly between the two (or more) upstream branches, but this would not accurately represent differences in influence related to factors such as discharge or watershed area. Instead, segment weights are used to ensure that locations residing on segments that contribute the strongest influence (i.e.,

discharge, area, or stream order) to a downstream location are allocated a stronger influence, or weighting, in the model (Cressie et al. 2006, Ver Hoef et al. 2006). It is intuitive to think about a site's influence on downstream conditions in terms of discharge or watershed area; stream segments that contribute the most discharge to a downstream location are likely to have a strong influence on the conditions found there. However, spatial weights can be based on any measure as long as some simple rules are followed during their construction.

In a previous publication, we described a process that can be used to calculate the spatial weights in a geographic information system (GIS) (Peterson et al. 2007). However, that process can be computationally intensive because the topological relationships of the stream segments are used to identify the features that make up the hydrologic path between every pair of sites (both observed and predicted). Consequently, it becomes challenging to calculate the spatial weights as the number of segments in the stream network or the number of observed or predicted sites increases. Here we present an alternative method for calculating the spatial weights, which is based on an additive function. It is less computationally intensive and produces spatial weights that are identical to those produced using the methods described in Peterson et al. (2007).

Calculating the spatial weights is a three step process: 1) calculating the segment proportional influence (PI), 2) calculating the additive function value for each stream segment, and 3) calculating the spatial weights (Peterson et al. 2007). The segment PI is defined as the relative influence that a stream segment has on the segment(s) directly downstream (Fig. A2.A). In this example, the segment PI is based on watershed area, but other measures could also be used. To begin, watershed area is calculated for the downstream node of each stream segment in the network. The cumulative watershed area at each confluence is calculated by summing

the watershed area for the stream segments that flow into it. The segment PI for each of the segments that flow into the confluence is then equal to the proportion of the cumulative watershed area that it contributes (Fig. A2.A). The segment PIs directly upstream from a confluence always sum to 1 because they are proportions. This is particularly important because it ensures that stationarity in the variances is maintained (Ver Hoef et al. 2006).

The second step is to calculate the additive function value (AFV) for each stream segment (Fig. A2.A). First, the stream segment directly upstream from the stream outlet is identified and assigned a segment AFV equal to 1. The outlet segment is the most downstream segment in the network and represents features that drain to the ocean or out of a predefined study area. Working upstream from the outlet segment by segment, the product of the segment PIs is taken and assigned to the individual segments; this value represents the segment AFV, which is constant within a segment. Although the segment AFV is a product, it is also considered an additive function because the segment PIs always sum to 1 at the stream confluences (Fig. A2.A). Using the data in Fig. A2.A as an example, the segment AFV of $R_5 = 1$ because it is the most downstream segment in the network. It follows that the segment AFV of $R_3 = 1 * 0.85 = 0.85$ and the segment AFV of $R_2 = 1 * 0.85 * 0.41 = 0.35$. This process continues until each segment in the stream network is assigned a stream AFV. If there is more than one stream network in the study area (multiple stream outlets) then the segment AFV is calculated for each network separately.

The third and final step is to assign the site AFV values and to calculate the spatial weights. The site AFV value (Ω_i) is derived simply; it is equal to the segment AFV value on which it resides (Fig. A2.B). This also holds true when multiple sites reside on a single stream segment. The spatial weight for a pair of flow-connected sites is $\sqrt{\Omega_i / \Omega_j}$, where Ω_i is the upstream site AFV and Ω_j is

the downstream site AFV (Fig. A2.C). If two sites are not flow-connected their spatial weight is equal to 0. A sites spatial weight on itself, or any other site located on the same segment, is equal to 1.

The spatial weights are stored as a symmetric matrix, meaning that there is a symmetric correlation between upstream and downstream locations (Fig. A2.C). For example, the spatial weight from s_2 to s_4 is equal to the spatial weight from s_4 to s_2 . This may at first seem confusing since flow-connected relationships are meant to represent flow direction in a stream network. Nevertheless, there is a symmetric correlation between flow-connected sites. Even though a downstream site does not directly influence an upstream site, it does provide information about the conditions found upstream. In addition, attempting to enforce an asymmetric correlation would violate one of the assumptions of geostatistical modeling; namely, that the covariance matrix is symmetric positive-definite (Chiles and Delfiner 1999). Instead, flow direction is preserved by restricting the symmetric correlation to include only flow-connected sites (Fig. A2.C) and the strength of the spatial autocorrelation is represented using the spatial weights and the hydrologic distances. Since the spatial weights are based on the product of proportions, the influence decreases as the distance between two locations increases for flow-connected locations. However, note that in Fig. A2, s_2 is physically closer to s_4 than s_1 is to s_4 , yet the weight between s_1 and s_4 is higher in A2.C, so weights carry information other than simple distance.

Constructing a Valid Tail-up Covariance Matrix

The example data provided in Figs. A1 and A2 can be used to further illustrate the construction of a valid covariance matrix using a tail-up model. Rather than estimating covariance parameters, we set them to $\sigma_{TU}^2 = 4$ and $\alpha = 15$. Of course, for a real data set, the autocovariance parameters are not specified; they are left as free parameters and estimated from the data. This can be accomplished using a variety of methods, such as weighted least squares (Cressie et al. 2006), maximum likelihood (Peterson et al. 2006), restricted maximum likelihood (REML) (Ver Hoef et al. 2006), or Markov Chain Monte Carlo (MCMC) in a Bayesian framework (Handcock and Stein 1993). We obtain matrix \mathbf{C}_1 by applying the linear-with-sill autocovariance function (Eqs. A.1 and A.2) to the hydrologic distances given in Fig. A1.B. Then we take the Hadamard product of \mathbf{C}_1 and \mathbf{W} to obtain Σ_{TU} for the example data provided in Figs. A1 and A2.

$$\Sigma_{TU} = \begin{pmatrix} 4.00 & 1.33 & 0.80 & 0 \\ 1.33 & 4.00 & 1.87 & 0.27 \\ 0.80 & 1.87 & 4.00 & 2.40 \\ 0 & 0.27 & 2.40 & 4.00 \end{pmatrix} \square \begin{pmatrix} 1 & 0 & 0.77 & 0.71 \\ 0 & 1 & 0.64 & 0.59 \\ 0.77 & 0.64 & 1 & 0.92 \\ 0.71 & 0.59 & 0.92 & 1 \end{pmatrix} =$$

$$\begin{pmatrix} 4.00 & 0 & 0.62 & 0 \\ 0 & 4.00 & 1.20 & 0.16 \\ 0.62 & 1.20 & 4.00 & 2.21 \\ 0 & 0.16 & 2.21 & 4.00 \end{pmatrix} \quad (\text{A.6})$$

Note that matrix \mathbf{W} is the spatial weights matrix (Fig. A2.C). Also, notice that sites s_1 and s_2 are uncorrelated because they are flow-unconnected, while s_1 and s_4 are uncorrelated because they are beyond the range of the autocovariance function.

Tail-down Models

A tail-down model allows spatial correlation between *both* flow-connected and flow-unconnected pairs of sites in a stream network. However, the autocovariance functions used to model these two relationships differ (Ver Hoef and Peterson in press). Thus, the spatial data inputs needed to construct a tail-down model are different than the tail-up model. The total hydrologic distance, h (Fig. A1.B), is used for flow-connected pairs, in the same way that it was used in the tail-up model (Eq. A.1). However, for flow-unconnected pairs the downstream hydrologic distance is used (Fig. A1.A). As we mentioned previously, only one distance matrix is necessary to calculate these two distances since $a+b=h$, (Figs. A1.A and A1.B). Also, the total hydrologic distance between flow-connected sites is not weighted in a tail-down model. This is because the tails of the moving-average function point downstream. It is generally not necessary to split the function since a dendritic network tends to converge to a single outlet. In the case of a braided channel we do not split the function in the downstream direction and thus data must be restricted to a single channel; however the moving-average theory could accommodate splitting downstream as well.

Examples of the tail-down models include the tail-down linear-with-sill model,

$$C_{TD}(s_i, s_j | \boldsymbol{\theta}) = \begin{cases} \sigma_{TD}^2 \left(1 - \frac{\max(a, b)}{\alpha}\right) I\left(\frac{\max(a, b)}{\alpha} \leq 1\right) & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected,} \\ \sigma_{TD}^2 \left(1 - \frac{h}{\alpha}\right) I\left(\frac{h}{\alpha} \leq 1\right) & \text{if } s_i \text{ and } s_j \text{ are flow-connected;} \end{cases} \quad (\text{A.7})$$

the tail-down spherical model,

$$C_{TD}(s_i, s_j | \boldsymbol{\theta}) = \begin{cases} \sigma_{TD}^2 \left(1 - \frac{3 \min(a, b)}{2 \alpha} + \frac{1 \max(a, b)}{2 \alpha}\right) \times \\ \left(1 - \frac{\max(a, b)}{\alpha}\right)^2 I\left(\frac{\max(a, b)}{\alpha} \leq 1\right) & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected,} \\ \sigma_{TD}^2 \left(1 - \frac{3 h}{2 \alpha} + \frac{1 h^3}{2 \alpha^3}\right) I\left(\frac{h}{\alpha} \leq 1\right) & \text{if } s_i \text{ and } s_j \text{ are flow-connected;} \end{cases} \quad (\text{A.8})$$

and the tail-down mariah model.

$$C_{TD}(s_i, s_j | \boldsymbol{\theta}) = \begin{cases} \sigma_{TD}^2 \left(\frac{\log(a/\alpha + 1) - \log(b/\alpha + 1)}{(a-b)/\alpha} \right) & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected,} \\ & a \neq b, \\ \sigma_{TD}^2 \left(\frac{1}{a/\alpha + 1} \right) & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected, } a = b, \\ \sigma_{TD}^2 \left(\frac{\log(h/\alpha + 1)}{h/\alpha} \right) & \text{if } s_i \text{ and } s_j \text{ are flow-connected, } h > 0, \\ \sigma_{TD}^2 & \text{if } s_i \text{ and } s_j \text{ are flow-connected, } h = 0. \end{cases} \quad (\text{A.9})$$

Notice that some form of the downstream hydrologic distances (a and b) are used in Equations A.7, A.8, and A.9 rather than the total hydrologic distance ($a+b=h$). This modification is necessary because the covariance matrix is not guaranteed to be positive definite when Euclidean distance is simply replaced with total hydrologic distance ($a+b = h$). The only known exception to this rule is the tail-down exponential model (Ver Hoef et al. 2006)

$$C_{TD}(s_i, s_j | \boldsymbol{\theta}) = \begin{cases} \sigma_{TD}^2 \exp(-(a+b)/\alpha) & \text{if } s_i \text{ and } s_j \text{ are flow-unconnected,} \\ \sigma_{TD}^2 \exp(-h/\alpha) & \text{if } s_i \text{ and } s_j \text{ are flow-connected.} \end{cases} \quad (\text{A.10})$$

Here, the exponential model (Eq. A.10) is a pure hydrologic-distance model ($a+b=h$) since the autocovariance is calculated in the same way for both flow-connected and flow-unconnected pairs. Please see Ver Hoef et al. (2006) for an example that demonstrates

how invalid covariance matrices are generated when total hydrologic distance is used in autocovariance models that were developed for Euclidean distance.

Using the linear-with-sill autocovariance function (Eq. A.7), the example hydrologic distances (Fig. A1.A and Fig. A1.B), and set covariance parameters ($\sigma_{TD}^2 = 4$ and $\alpha = 15$) we obtain the tail-down covariance matrix for the example data provided in Fig. A1

$$\begin{pmatrix} 4.00 & 2.13 & 0.80 & 0 \\ 2.13 & 4.00 & 1.87 & 0.27 \\ 0.80 & 1.87 & 4.00 & 2.40 \\ 0 & 0.27 & 2.40 & 4.00 \end{pmatrix} \quad (\text{A.11})$$

Although the tail-down model can account for spatial autocorrelation between both flow-connected and flow-unconnected pairs, the relative strength of spatial autocorrelation for each type is restricted (Ver Hoef and Peterson in press). For example, consider the situation where there are two pairs of locations, one pair is flow-connected and the other flow-unconnected, and the distance between the two pairs is equal, $a + b = h$. In this case, the strength of spatial autocorrelation is generally equal or greater for flow-unconnected pairs (Ver Hoef and Peterson in press). This characteristic can be seen in the tail-down covariances that were derived from the example data in Fig. A1 (Eq. A.11). Notice that the covariance between flow-unconnected sites s_1 and s_2 was 2.13, which is stronger than a neighboring flow-connected pair, s_2 and s_3 (1.87), even though the flow-connected pair has a shorter hydrologic distance (Figs.

A1.A and A1.B). However, if we use a mixture model approach and add the tail-up covariance matrix to the tail-down covariance matrix, then the flow-connected pair (s_2 and s_3) may have greater autocovariance than the flow-unconnected pair (s_1 and s_2).

LITERATURE CITED

- Chiles, J., and P. Delfiner. 1999. Geostatistics: Modeling spatial uncertainty. John Wiley and Sons, New York, New York, USA.
- Cressie, N., J. Frey, B. Harch, and M. Smith. 2006. Spatial prediction on a river network. *Journal of Agricultural Biological and Environmental Statistics* 11:127–150.
- Handcock, M. S., and M. L. Stein. 1993. A Bayesian analysis of kriging. *Technometrics* 35:403–410.
- Peterson, E. E., D. M. Theobald, and J. M. Ver Hoef. 2007. Geostatistical modeling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology* 52:267–279.
- Peterson, E. E., A. A. Merton, D. M. Theobald, and N. S. Urquhart. 2006. Patterns of spatial autocorrelation in stream water chemistry. *Environmental Monitoring and Assessment* 121:569–594.
- Ver Hoef, J. M., and E. E. Peterson. *In press*. A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association*.
- Ver Hoef, J. M., E. E. Peterson, and D. M. Theobald. 2006. Some new spatial statistical models for stream networks. *Environmental and Ecological Statistics* 13:449–464.

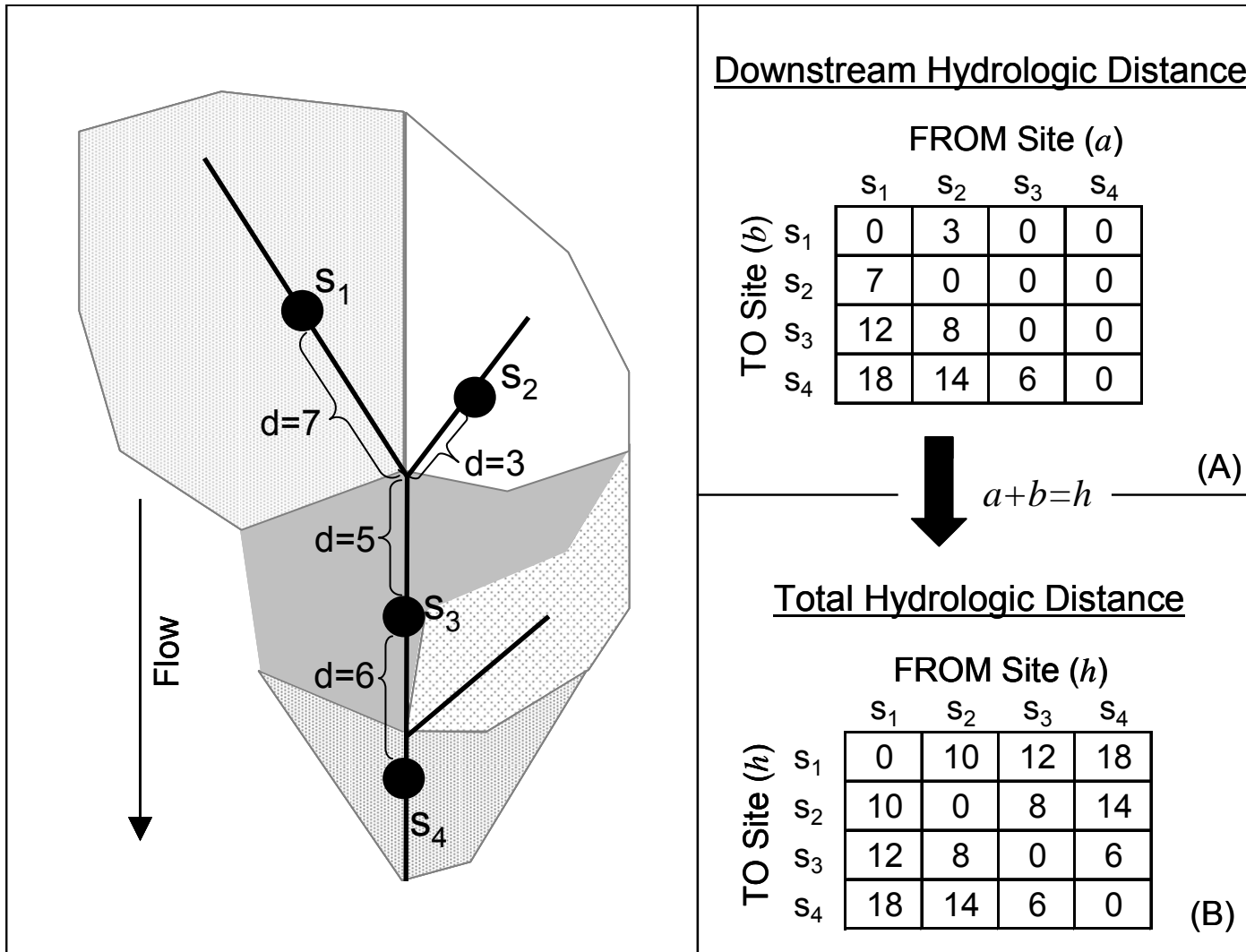
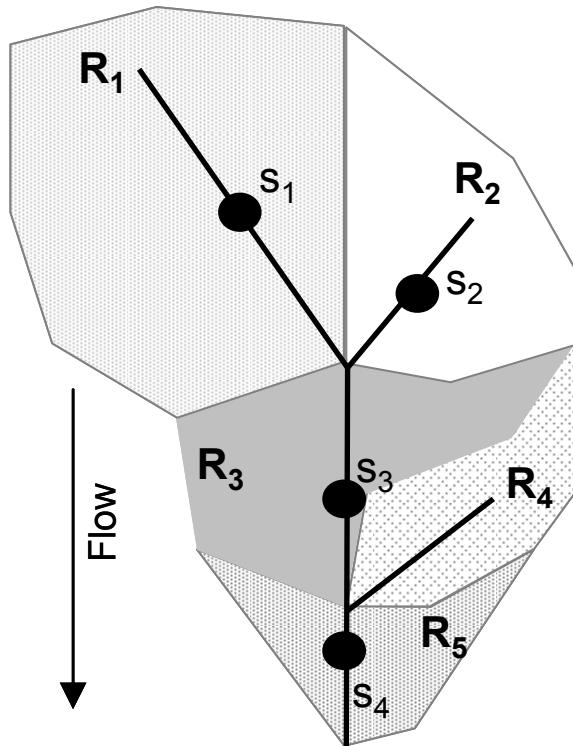


FIG. A1. The downstream hydrologic distance for each pair of sites, a and b , is needed to fit the tail-down model (A). The tail-up model requires the total hydrologic distance, h , between each pair (B). The total hydrologic distance between sites (B) can be derived from the downstream distance matrix (A) since $a + b = h$.

Segment ID	Watershed Area (q _i)	Segment PI (ω _k)	Segment AFV
R ₁	50	0.59	0.5
R ₂	35	0.41	0.35
R ₃	115	0.85	0.85
R ₄	20	0.15	0.15
R ₅	160	1	1

$$\omega(R_3) = \frac{q(R_3)}{q(R_3) + q(R_4)} = \frac{115}{115 + 20} = 0.85$$

$$\text{Segment AFV } R_1 = \omega(R_5) * \omega(R_3) * \omega(R_1) = 1 * 0.85 * 0.59 = 0.50 \quad (\text{A})$$



Site ID	Site AFV (Ω _i)
s ₁	0.5
s ₂	0.35
s ₃	0.85
s ₄	1

$$\Omega(s_4) = \text{Segment AFV } R_5 = 1$$

(B)

$$W[s_1, s_3] = \prod_{k \in B_{s_1, s_3}} \sqrt{w_k} = \sqrt{\frac{\Omega(s_1)}{\Omega(s_3)}} = \sqrt{\frac{0.5}{0.85}} = 0.77$$

		FROM Site			
		s ₁	s ₂	s ₃	s ₄
TO Site	s ₁	1	0	0.77	0.71
	s ₂	0	1	0.64	0.59
	s ₃	0.77	0.64	1	0.92
	s ₄	0.71	0.59	0.92	1

(C)

FIG. A2. The spatial weights matrix is constructed using the segment proportional influences (PI), which represent the proportion of the watershed area that each segment (R_1, R_2, R_3, R_4, R_5) contributes to a confluence (A). The segment PIs are used to calculate the segment additive function value (AFV). First, the stream segment directly upstream from the stream outlet is identified (R_5) and assigned a segment AFV equal to one. Working upstream from the outlet segment by segment, the product of the segment PIs is taken and assigned to the individual segments (A). The site AFV is equal to the segment AFV on which it resides (B). The spatial weights represent a symmetric correlation between flow-connected sites. They are calculated by taking the square root of the upstream site AFV divided by the downstream site AFV (C). If two sites are not flow-connected their spatial weight is equal to zero and a sites spatial weight on itself, or any other site located on the same segment, is equal to one.