A Moving Average Approach for Spatial Statistical Models of Stream Networks

Author(s): Jay M. Ver Hoef and Erin E. Peterson

Source: *Journal of the American Statistical Association*, March 2010, Vol. 105, No. 489 (March 2010), pp. 6-18

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/29747004

A Moving Average Approach for Spatial Statistical Models of Stream Networks

Jay M. VER HOEF and Erin E. PETERSON

In this article we use moving averages to develop new classes of models in a flexible modeling framework for stream networks. Streams and rivers are among our most important resources, yet models with autocorrelated errors for spatially continuous stream networks have been described only recently. We develop models based on stream distance rather than on Euclidean distance. Spatial autocovariance models developed for Euclidean distance may not be valid when using stream distance. We begin by describing a stream topology. We then use moving averages to build several classes of valid models for streams. Various models are derived depending on whether the moving average has a "tail-up" stream, a "tail-down" stream, or a "two-tail" construction. These models also can account for the volume and direction of flowing water. The data for this article come from the Ecosystem Health Monitoring Program in Southeast Queensland, Australia, an important national program aimed at monitoring water quality. We model two water chemistry variables, pH and conductivity, for sample sizes close to 100. We estimate fixed effects and make spatial predictions. One interesting aspect of stream networks is the possible dichotomy of autocorrelation between flow-connected and flow-unconnected locations. For this reason, it is important to have a flexible modeling framework, which we achieve on the example data using a variance component approach.

KEY WORDS: Euclidean distance; Geostatistics; Kernel convolution; Spatial autocorrelation; Spatial linear model.

1. INTRODUCTION

Some new models for stream networks, based on moving average constructions, have been described by Ver Hoef, Peterson, and Theobald (2006) and Cressie et al. (2006). These authors have developed models for stream chemistry data. One property of these models is statistical independence of random variables located on stream segments that do not share flowing water. Although these are reasonable models for the passive movement of particles in streams, they are not adequate for variables such as fish or insects that may move upstream. In this case it is desirable to allow spatial autocorrelation among random variables on stream segments that do not share flow. The goals of the present work were to (a) develop these new classes of models, (b) provide a unified view of moving average models for stream networks, (c) provide a flexible framework for modeling spatially continuous data from stream networks, and (d) apply these models to some data sets of national importance for monitoring water quality in Australia.

Streams and rivers are important; clean water is used by humans for drinking and recreation, and it provides critical habitat for certain plants and animals. A considerable amount of time and money has been spent on sampling and monitoring streams and rivers (e.g., Torgersen, Gresswell, and Bateman 2004; Yuan 2004). As with most environmental and ecological data, a sample must be taken from a possibly infinite population of values on a stream network; for example, sample units could be counts of fish from a small stream segment or water quality samples collected from points along a stream network in a large region. Freshwater ecologists have begun to explore stream network data using stream distance measures (Dent and Grimm 1999; Gardner, Sullivan, and Lembo 2003; Legleiter et al. 2003; Torgersen, Gresswell, and Bateman 2004; Ganio, Torgersen, and Gresswell 2005). Stream distance is defined as the shortest distance between two locations, where distance is computed only along the stream network. These methods are mostly descriptive and do not include valid models of spatial autocorrelation based on stream distance. By "valid," we mean an autocovariance model that yields positive definite covariance matrixes for any parameter values (within their range) and any number and configuration of locations. Ver Hoef, Peterson, and Theobald (2006) presented some traditional time series and geostatistical models that are not valid for stream networks when stream distance is substituted for Euclidean distance.

As mentioned earlier, new models based on stream distance and water flow are needed for stream networks. Ver Hoef, Peterson, and Theobald (2006) and Cressie et al. (2006) used spatial moving averages to develop models that are valid when using stream distance. Moving average models are being increasingly used to construct valid autocovariances when confronted with new problems, such as flexible, nonparametric models (Barry and Ver Hoef 1996), multivariate models (Ver Hoef and Barry 1998; Ver Hoef, Cressie, and Barry 2004), and nonstationary models (Higdon 1998; Higdon, Swall, and Kern 1999; Fuentes 2002). In this article we extend the use of moving average constructions to develop new classes of valid models for streams, and also provide a unified approach to all models developed to date.

After developing the models, we apply them to data from the Ecosystem Health Monitoring Program (EHMP; 2006) in Southeast Queensland (SEQ), Australia. This important data set is used to provide a regional assessment of the status and trend in ambient ecosystem health for 18 major watersheds, 18 estuaries, and Moreton Bay. One of the world's most comprehensive aquatic monitoring programs, the EHMP is funded and managed by the SEQ Healthy Waterways Partnership, which includes local, state, and national government agencies; catchment authorities; and industry, university, and community

Jay M. Ver Hoef is Statistician, National Marine Mammal Lab, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service, Seattle, WA 98115 (E-mail: *jay.verhoef@noaa.gov*). Erin E. Peterson is Research Scientist, CSIRO Division of Mathematical and Information Sciences, Indooroopilly, Queensland, Australia. This project was supported by the National Marine Fisheries Service as well as the CSIRO Division of Mathematical and Information Sciences. The authors thank Southeast Queensland's Healthy Waterways Partnership for collecting and sharing the data.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08248

groups. The program serves as a model for new integrated monitoring and reporting frameworks currently under development throughout the world.

The rest of the article is organized as follows. Section 2 presents a unified view of moving average constructions for stream networks. It begins with a stream topology and a review of the tail-up models used by Ver Hoef, Peterson, and Theobald (2006) and Cressie et al. (2006). It then introduces the tail-down and mixed models. Section 3 gives two examples from the EHMP data set from Queensland, Australia. Section 4 concludes with a discussion. The Appendix discusses two-tail models, which are not used for modeling because of computational difficulties.

2. MOVING AVERAGE MODELS

Yaglom (1987) showed that a large class of autocovariances can be developed by creating random variables as the integration of a moving average function over a white-noise random process,

$$Z(s|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x - s|\boldsymbol{\theta})\, dW(x), \tag{1}$$

where x and s are locations on the real line and $g(x|\boldsymbol{\theta})$ is the moving average function defined on $\mathcal{R}^1$, which is square-integrable. Recall that when $W(x)$ is Brownian motion, $E[Z(s|\boldsymbol{\theta})^2] = \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta})^2\, dx$. The moving average construction (1) allows the expression of a valid autocovariance between $Z(s)$ and $Z(s+h)$ as

$$C(h|\boldsymbol{\theta}) = \int_{-\infty}^{\infty} g(x|\boldsymbol{\theta}) g(x - h|\boldsymbol{\theta})\, dx. \tag{2}$$

Before making use of this result, we give some topology and notation that we use for stream networks.

2.1 Stream Topology and Notation

As described by Ver Hoef, Peterson, and Theobald (2006), here we define stream segments as lines between junctions in a stream network. Note that a segment does not represent the portion of stream located between two sampling locations, but rather is based on the branching characteristics of the stream. We consider a location downstream to be a lower real number than a location farther upstream. Stream networks are dendritic, and so the whole network will have a single most-downstream point (sometimes called the outlet), which we set to 0. This is the point from which all distances are computed. Any location on a stream network can be connected by a continuous line to the lowest point in that network; thus the distance from the lowest point is simply the length of that line. We define this as "distance upstream." To develop some of the moving average models, it will be convenient to let all terminal upstream segments go to ∞, and we extend the stream network downstream of the outlet as a single line to $-\infty$.

A stream network will contain a finite number of stream segments, which we index arbitrarily with $i = 1, 2, \ldots$. In a branching stream network, many locations will have the same distance from the outlet (our 0 point). To uniquely define individual locations and keep track of distance upstream, we identify each location according to its distance as x_i, where i indicates that the location is on the ith stream segment. Figure 1, shows three segments, with a location on each segment. Thus r_1 is the distance upstream, and it is located on the first segment. The location of the smallest upstream distance on the ith segment is denoted by l_i [Figure 1(b)], and the largest upstream distance on the ith segment is denoted by u_i. The arbitrary labeling of segments in Figure 1 is apparent from the subscripts on r_1, s_2, and t_3.

Let the whole set of stream segment indexes be denoted by A. The index set of stream segments upstream of x_i, including the ith segment, is denoted by $U_i \subseteq A$, and that excluding the ith segment is denoted by $U_i^* \subseteq A$. The index set of stream segments downstream of x_i, including the ith segment, is denoted by $D_i \subseteq A$, and that excluding the ith segment is denoted by $D_i^* \subseteq A$. Using these definitions, we can say that two locations r_i and s_j on a stream network are "flow-connected" if $U_i \cap U_j \neq \emptyset$ and are "flow-unconnected" if $U_i \cap U_j = \emptyset$. As another example, we can identify the set of stream segments between two flow-connected locations $r_i \leq s_j$, exclusive of the ith and jth segments, as $D_j^* \setminus D_i$. We also need notation for the upstream and downstream domains with respect to points in addition to segments. In that case we use $\vee_s$ to denote the domain upstream of a point s, including all branchings, and $\wedge_s$ to denote the domain downstream of s that follows flow only (i.e., it does not go downstream and then back upstream).

In Figure 1 the distance from each location to u_1 is labeled as a, b, or c. In general, for two flow-unconnected locations, we use a to indicate the distance from one location to the nearest junction downstream of which it shares flow with the other location. Likewise, we use b to indicate the distance of the other location to the same junction. We use c as the distance between the stream junction and the downstream location. In Figure 1, the distance between two flow-connected locations is $c + a$, but in general we denote this simply by h.

2.2 Tail-Up Models

The moving average construction in (1) and (2) is well known for the continuous real line from $-\infty$ to ∞, such as for time series models. Ver Hoef, Peterson, and Theobald (2006) and Cressie et al. (2006) used moving averages for a stream network to develop models as shown in Figures 1(a) and (b). We call these the "tail-up" models, because they are unilateral in the upstream direction. (Moving average function values are positive only upstream from a location.) In Figure 1(a), the moving average function goes upstream from r_1. When it reaches a fork at u_1, the function continues up each branch but is weighted. The weights can be proportional to flow volume or another meaningful metric. In Section 2.2.1 we show how to maintain stationary variances when weighting, but nonstationary models also could be developed. We do the integral in (1) piecewise, summing up all segments that contain the moving average function $g(x|\boldsymbol{\theta})$. Because of the "upstream" construction, we need only integrate the segments that are flow-connected and in U_j, where $r_i < s_j$, to compute the covariance between r_i and s_j.

In Figure 1(a), when the moving average function of r_1 overlaps with the moving average function of s_2, there will be autocorrelation between them. In Figures 1(a) and (b), r_1 is connected by flow to both s_2 and t_3, but s_2 and t_3 are not connected by flow. The limits of integration for the flow-connected

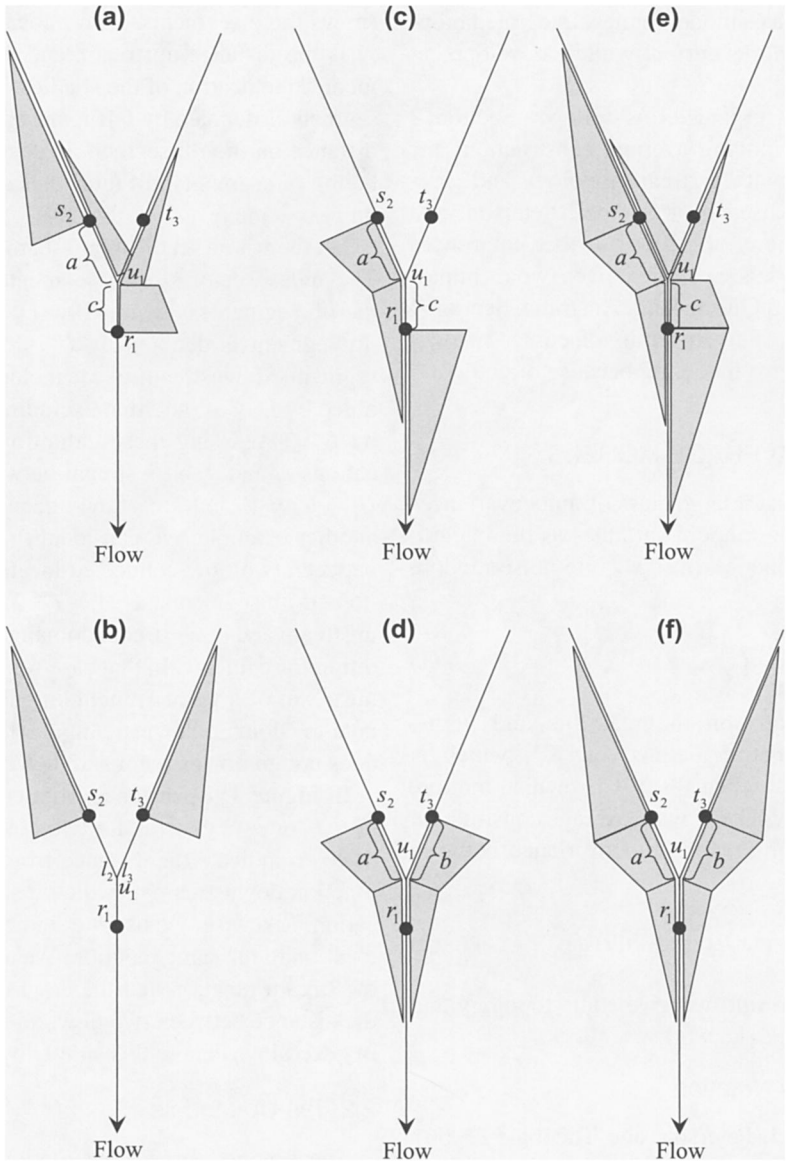

Figure 1. Three locations on a stream network, r_1, s_2, and t_3. The location of the farthest upstream distance on segment 1 is u_1, and the locations of the farthest downstream distances on segments 2 and 3 are l_2 and l_3, respectively. Effectively, $u_1 = l_2 = l_3$, but it is convenient to use the distinct notation. Two moving average functions are shown for tail-up flow-connected locations (a), tail-up flow-unconnected locations (b), tail-down flow-connected locations (c), tail-down flow-unconnected locations (d), two-tail flow-connected locations (e), and two-tail flow-unconnected locations (f).

case for the tail-up models are shown in Figure 2(a). For flow-connected locations, the unweighted covariance between two such locations is

$$C_t(h|\boldsymbol{\theta}) = \int_h^\infty g(x|\boldsymbol{\theta})g(x-h|\boldsymbol{\theta})\,dx, \tag{3}$$

where h is the stream distance between locations r_i and s_j [e.g., $h = a + c$ in Figure 1(a)]. Then the moving average construction, as described by Ver Hoef, Peterson, and Theobald (2006), is

$$C_u(r_i, s_j|\boldsymbol{\theta}) = \begin{cases} \pi_{i,j} C_t(h|\boldsymbol{\theta}) & \text{if } r_i < s_j \text{ are flow-connected} \\ 0 & \text{if } r_i \text{ and } s_j \text{ are flow-unconnected,} \end{cases} \tag{4}$$

where $\pi_{i,j}$ are weights that we describe in the next section. Note that there is no overlap in the moving average functions when they are not flow-connected [Figure 1(b)], and thus the covariance (2) is 0.

Table 1 gives some moving average functions that can be used in the construction (3), and, when used in (4), provide various tail-up models for stream networks. After integration in (3), we reparameterize to put the models in forms typically seen in the spatial statistical literature; for example, the "partial sill" is the variance, which is the covariance function when the distance is 0. This partial sill, denoted by θ_v, is usually some function of both θ_1 and θ_r in Table 1. Using Table 1 in (3), we obtain the following models:

- Tail-up linear with sill model

$$C_t(h|\boldsymbol{\theta}) = \theta_v\left(1 - \frac{h}{\theta_r}\right) I\left(\frac{h}{\theta_r} \le 1\right),$$

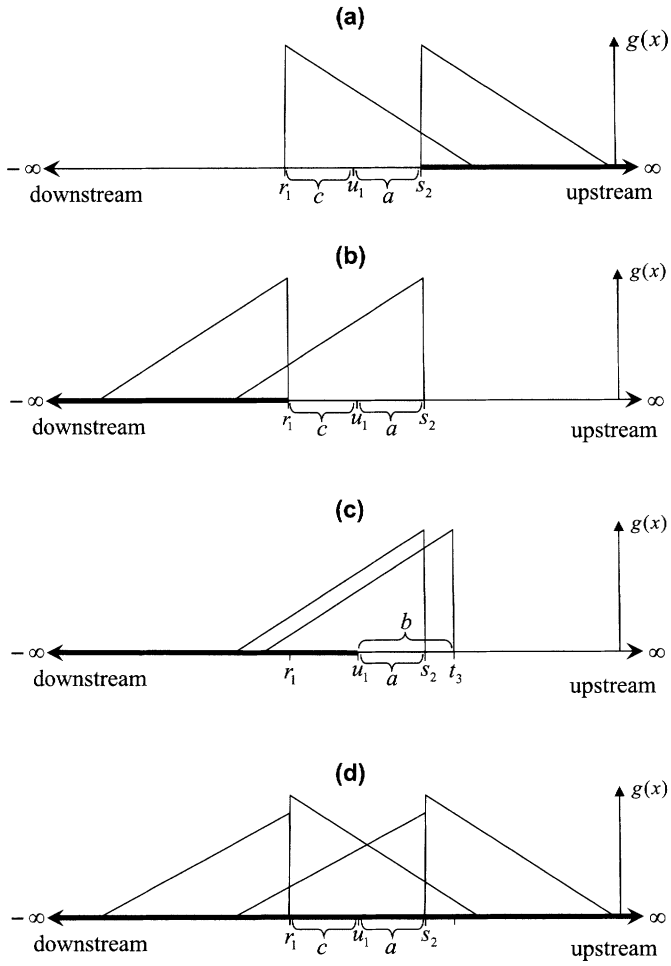

Figure 2. Moving average functions and limits of integration (solid black line) for tail-up flow-connected (a), tail-down flow-connected (b), tail-down flow-unconnected (c), and two-tail flow-connected (d) locations.

- Tail-up spherical model

$$C_t(h|\boldsymbol{\theta}) = \theta_v\left(1 - \frac{3}{2}\frac{h}{\theta_r} + \frac{1}{2}\frac{h^3}{\theta_r^3}\right)I\left(\frac{h}{\theta_r} \le 1\right),$$

- Tail-up exponential model

$$C_t(h|\boldsymbol{\theta}) = \theta_v \exp(-h/\theta_r),$$

- Tail-up Mariah model

$$C_t(h|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(\dfrac{\log(h/\theta_r + 1)}{h/\theta_r}\right) & \text{if } h > 0 \\ \theta_v & \text{if } h = 0. \end{cases}$$

Table 1. Moving average functions

Name	Moving average function
Linear with sill	$g(x\|\tilde{\boldsymbol{\theta}}) = \theta_1 I(0 \le x/\theta_r \le 1)$
Spherical	$g(x\|\tilde{\boldsymbol{\theta}}) = \theta_1(1 - x/\theta_r)I(0 \le x/\theta_r \le 1)$
Exponential	$g(x\|\tilde{\boldsymbol{\theta}}) = \theta_1 e^{-x/\theta_r} I(0 \le x)$
Mariah	$g(x\|\tilde{\boldsymbol{\theta}}) = \theta_1 \frac{1}{1+x/\theta_r} I(0 \le x)$

Here $\theta_v > 0$ is an overall variance parameter (also known as the partial sill), and $\theta_r > 0$ is the range parameter. All of these models were first described by Ver Hoef, Peterson, and Theobald (2006). The tail-up models are especially appropriate when we want to enforce zero autocorrelation when sites are flow-unconnected, which could occur when a variable is dominated by flow (e.g., when a pollutant enters a stream and can flow only downstream, causing measurements to be autocorrelated only when flow-connected).

2.2.1 Weighting in the Tail-Up Models. A unique feature of stream network models is the splitting of $g(x|\boldsymbol{\theta})$ as it goes upstream [Figure 1(a)]. This is achieved by assigning a weighting attribute to each stream segment. To account for the splitting, Cressie et al. (2006) modified (1) to construct a spatial process on a stream network as

$$Z(s_i|\boldsymbol{\theta}) = \int_{\vee s_i} g(x - s_i|\boldsymbol{\theta})\sqrt{\frac{\Omega(x)}{\Omega(s_i)}}\, dW(x),$$

where $\Omega(x)$ is an additive function to ensure stationarity in variance; that is, $\Omega(x)$ is constant within a stream segment, but then $\Omega(x)$ is the sum of each segment's value when two segments join at a junction. This definition leads to (4), where $\pi_{i,j} = \sqrt{\frac{\Omega(s_j)}{\Omega(r_i)}}$. Ver Hoef, Peterson, and Theobald (2006) constructed a spatial process on a stream network as

$$Z(s_i|\boldsymbol{\theta}) = \int_{s_i}^{u_i} g(x_i - s_i|\boldsymbol{\theta})\, dW(x_i) + \sum_{j \in U_i^*}\left(\prod_{k \in B_{i,j}} \sqrt{\omega_k}\right)\int_{l_j}^{u_j} g(x_j - s_i|\boldsymbol{\theta})\, dW(x_j),$$

where $B_{i,j} = D_j \setminus D_i$ is the set of segments between the ith and jth (inclusive of the jth but exclusive of the ith), and at each fork upstream of the ith stream segment, such that the upstream segments are denoted j and k, we require that $0 \le \omega_j, \omega_k \le 1$, and $\omega_j + \omega_k = 1$, which also ensures stationary variances among variables. This definition leads to (4), where $\pi_{i,j} = \prod_{k \in B_{i,j}} \sqrt{\omega_k}$. Now, what is the relationship between the specification of Cressie et al. (2006) and that of Ver Hoef, Peterson, and Theobald (2006)? To make this relationship clear, consider the example shown in Figure 3. In their application, Cressie et al. (2006) specified that all terminal segments, q_1 to q_6, equal 1. The additivity condition implies $q_7 = 2$, and so

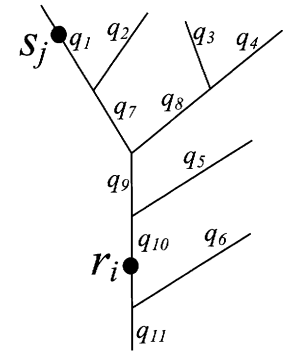

Figure 3. Example stream network illustrating weighting for the tail-up models.

on, with $q_{10} = 5$; these segment weights are known as Shreve's stream order (Shreve 1967). The additive function $\Omega(x_i)$ is equal to q_i; thus for this example, $\Omega(s_j) = 1$ and $\Omega(r_i) = 5$. Ver Hoef, Peterson, and Theobald (2006) used arbitrary values for all q_i; $i = 1, 2, \ldots, 11$, based on basin area, but then created weights that summed to 1 by taking $\omega_1 = q_1/(q_1 + q_2)$, $\omega_2 = q_2/(q_1 + q_2)$, $\omega_7 = q_7/(q_7 + q_8)$, and so on. The weights of Ver Hoef, Peterson, and Theobald (2006) would seem to be more general, being defined arbitrarily; however, as we show next, these weights can be turned into an additive function.

For the example shown in Figure 3, set $\Omega(x_{11}) = \omega_{11} = c$ for all x_{11} in stream segment 11; that is, the outlet segment is the only ω value that is arbitrary. Next, let $\Omega(x_{10}) = \omega_{11}\omega_{10}$ for all x_{10} in stream segment 10 and $\Omega(x_6) = \omega_{11}\omega_6$ for all x_6 in stream segment 6. Note that even though we have multiplied, $\Omega(x_i)$ is an additive function so far (as described earlier), because $\omega_6 + \omega_{10} = 1$. We continue these multiplications as we proceed upstream [e.g., $\Omega(x_9) = \omega_{11}\omega_{10}\omega_9$], maintaining the additive property. In general, we have $\Omega(x) = \prod_{k \in D_x} \omega_k$, where the set D_x comprises all stream segments downstream of x (inclusive of the segment containing x). Then, under this construction,

$$\sqrt{\frac{\Omega(s_j)}{\Omega(r_i)}} = \sqrt{\frac{\prod_{m \in D_i} \omega_m \prod_{k \in B_{i,j}} \omega_k}{\prod_{m \in D_i} \omega_m}} = \sqrt{\prod_{k \in B_{i,j}} \omega_k}$$

(see also Money, Carter, and Serre 2009); thus the weighting scheme of Cressie et al. (2006) and Ver Hoef, Peterson, and Theobald (2006) are equivalent.

Besides this connection, this result has an extremely important computational consequence. Using the apparently more general formulation of Ver Hoef, Peterson, and Theobald (2006), computing $\prod_{k \in B_{i,j}} \sqrt{\omega_k}$ for all *pairs* of locations might seem necessary. This would involve storage in an n^2 matrix (where n is the number of flow-connected pairs) plus the intensive GIS operations involved for each pair; indeed, Peterson, Theobald, and Ver Hoef (2007) showed how to do this. The additive function construction allows us to move up the stream network just once in a GIS and store the additive function value with each location, which involves only one intensive GIS operation and the storage of only the n $\Omega(s)$ values. The importance of the unconstrained constant c for the outlet segment is that it eliminates computational underflow when multiplying many ωs, some of which may be near 0, for a large stream network.

2.3 Tail-Down Models

We now turn our attention to the construction of the tail-down models, which have not been developed previously. We need to consider the two situations in which locations are connected by flow [Figure 1(c)] and are not connected by flow [Figure 1(d)]. The limits of integration for the flow-connected case for the tail-down models are shown in Figure 2(b) and those for the flow-unconnected case for the tail-down model are shown in Figure 2(c). We define the moving average function such that it is nonzero only downstream from a location. Note that we define all unilateral moving average functions with nonzero values on positive support only (as shown in Table 1). Using minus signs turns these into tail-down functions in the models that follow. Then, from (1) and Figure 2(b), we have for s_2 upstream of r_1, that is, $h = s_2 - r_1 > 0$,

$$C_c(h|\boldsymbol{\theta}) = \int_{-\infty}^{-h} g(-x|\boldsymbol{\theta}) g(-x - h|\boldsymbol{\theta})\, dx \tag{5}$$

for the flow-connected sites, where $g(-x|\boldsymbol{\theta})$ is a unilateral tail-down function with nonzero values only on the negative side of 0. From (1) and Figure 2(c), we have, for $b \geq a$,

$$C_n(a, b|\boldsymbol{\theta}) = \int_{-\infty}^{-b} g(-x|\boldsymbol{\theta}) g(-x - (b - a)|\boldsymbol{\theta})\, dx \tag{6}$$

for the flow-unconnected sites, where h, a, and b are as described in Section 2.1.

From the constructions in (5) and (6), using the moving average functions in Table 1, we can develop tail-down models for stream networks. On reparameterization, we obtain the following models:

- Tail-down linear with sill model, $b \geq a \geq 0$,

$$C_d(a, b, h|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(1 - \dfrac{h}{\theta_r}\right) I\left(\dfrac{h}{\theta_r} \leq 1\right) & \text{if flow-connected} \\ \theta_v\left(1 - \dfrac{b}{\theta_r}\right) I\left(\dfrac{b}{\theta_r} \leq 1\right) & \text{if flow-unconnected,} \end{cases}$$

- Tail-down spherical model, $b \geq a \geq 0$,

$$C_d(a, b, h|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(1 - \dfrac{3}{2}\dfrac{h}{\theta_r} + \dfrac{1}{2}\dfrac{h^3}{\theta_r^3}\right) I\left(\dfrac{h}{\theta_r} \leq 1\right) \\ \quad \text{if flow-connected} \\ \theta_v\left(1 - \dfrac{3}{2}\dfrac{a}{\theta_r} + \dfrac{1}{2}\dfrac{b}{\theta_r}\right)\left(1 - \dfrac{b}{\theta_r}\right)^2 I\left(\dfrac{b}{\theta_r} \leq 1\right) \\ \quad \text{if flow-unconnected,} \end{cases}$$

- Tail-down exponential model

$$C_d(a, b, h|\boldsymbol{\theta}) = \begin{cases} \theta_v \exp(-h/\theta_r) & \text{if flow-connected} \\ \theta_v \exp(-(a + b)/\theta_r) & \text{if flow-unconnected,} \end{cases}$$

- Tail-down Mariah model

$$C_d(a, b, h|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(\dfrac{\log(h/\theta_r + 1)}{h/\theta_r}\right) \\ \quad \text{if flow-connected, } h > 0 \\ \theta_v \quad \text{if flow-connected, } h = 0 \\ \theta_v\left(\dfrac{\log(a/\theta_r + 1) - \log(b/\theta_r + 1)}{(a - b)/\theta_r}\right) \\ \quad \text{if flow-unconnected, } a \neq b \\ \theta_v\left(\dfrac{1}{a/\theta_r + 1}\right) \\ \quad \text{if flow-unconnected, } a = b, \end{cases}$$

where $\theta_v > 0$ and $\theta_r > 0$. Note that the covariance between flow-connected sites is the same for the tail-down and tail-up models, apart from the weights in the tail-up models, as is apparent from Figures 2(a) and (b). Tail-down models are useful

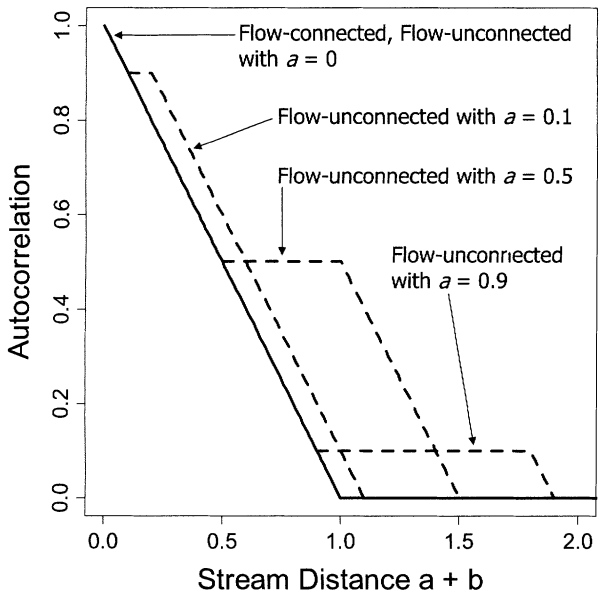

Figure 4. Tail-down linear with sill autocorrelation function (the sill parameter $\theta_v = 1$), where the range parameter is $\theta_r = 1$.

for modeling variables, such as fish or aquatic insects, that can move both upstream and downstream, creating autocorrelation among the flow-connected and flow-unconnected locations.

The tail-down linear with sill model shown in Figure 4 demonstrates how network autocorrelation models differ greatly from two-dimensional spatial (i.e., geostatistical) models and time series models. The combination of branching and flow, allowing for orientation of the moving average function, creates a rich and complex set of autocorrelation functions. Using Figure 4 to summarize, the tail-up linear with sill model has zero autocorrelation for flow-unconnected sites and has the same model for flow-connected sites as shown in Figure 4, but weighted by the additive function described in Section 2.2.1. A dichotomy of models between flow-connected and flow-unconnected also occurs for tail-down models, which we focus on next. Note that in Figure 4, for the same stream distance, flow-unconnected sites have more autocorrelation than flow-connected sites. A natural question arises: What if we want more autocorrelation in flow-connected sites than in flow-unconnected sites in tail-down models? We explore the relationship of autocorrelation between flow-connected and flow-unconnected sites in the next two sections.

2.4 Ratio of Two Autocovariance Functions

Unlike the tail-up models, tail-down models can have nonzero autocovariance between flow-connected and flow-unconnected locations, which can be a distinct function of stream distance. Here we investigate the interplay of autocovariances between flow-connected and flow-unconnected locations through their ratio. Consider a fixed distance, h, for two flow-connected sites and the distance, $a+b = h$, for two flow-unconnected locations. From the definitions of tail-up and tail-down covariance functions (4), (5), and (6), let

$$r_{uc}(a, b, \boldsymbol{\theta}) = \begin{cases} \dfrac{0}{\pi_{i,j} C_t(a+b|\theta)} & \text{for tail-up models} \\ \dfrac{C_n(a, b|\theta)}{C_c(a+b|\theta)} & \text{for tail-down models,} \end{cases} \tag{7}$$

which is the ratio of flow-unconnected to flow-connected autocovariances for a fixed distance and a set of covariance parameters $\boldsymbol{\theta}$. For example, in Figure 4, $r_{uc}(a, b) > 1$ because the flow-unconnected autocovariance is greater than the flow-connected autocovariance for the same stream distance. Note that the exponential model is a "symmetric" tail-down model, because $r_{uc}(a, b) = 1$ (where $h = a + b$). It appears that moving average functions with heavier shoulders than the exponential moving average function (like linear with sill and spherical) have relatively more autocorrelation among flow-unconnected sites than among flow-connected sites for an equal stream distance (see, e.g., Figure 4), making $r_{uc}(a, b) > 1$. Moving average functions with heavier tails (Mariah) have slightly less autocorrelation among flow-unconnected sites than among flow-connected sites (data not shown), making $r_{uc}(a, b) < 1$. This also makes sense when considering Figures 1(c) and (d).

To examine this further, consider the powered exponential moving average function,

$$g(x|\eta, \alpha) = \exp\left(-x^{\exp(\eta)}/10^{\alpha}\right) I(0 \le x). \tag{8}$$

This moving average function, with $\alpha = 0$, for various values of η, is shown in Figure 5(a). Note that the powered exponential model tends toward a linear with sill model (i.e., a rectangular moving average function) as $\eta \to \infty$. This is the exponential model at $\eta = 0$, and (8) has heavy-tail characteristics like the Mariah model as η becomes negative. Analytical solutions to the tail-down autocovariance functions (5) and (6) using the powered exponential moving average function (8) are not available for all values of η; however, we can use numeric integrals.

The autocovariance divided by the variance is known as the "autocorrelation." For (8), this can be expressed for flow-connected sites as

$$\rho_c(h, \eta, \alpha) = \frac{C_c(h|\eta, \alpha)}{C_c(0|\eta, \alpha)}. \tag{9}$$

Both ratios (7) and (9) are now viewed as functions of η and α. Figure 5(b) uses numeric integrals to compute (8) in (7) and (9) for a range of η and α values,with the distance between locations held constant at $a = b = 50$. This choice is completely arbitrary; any change in a fixed distance can be accommodated by a change in α. Thus we compute $r_{uc}(50, 50, \eta, \alpha)$ and $\rho_c(100, \eta, \alpha)$. Our goal is to investigate the ratio of autocovariance between flow-connected and flow-unconnected sites (7) as a function of flow-connected autocorrelation (9) (an expression of the range of the moving average function).

Figure 5(b) plots $r_{uc}(50, 50, \eta, \alpha)$ versus $\rho_c(100, \eta, \alpha)$ for all combinations of η and α ranging from -1.5 to 2.5 in increments of 0.1. Some combinations create numerical integrals that are beyond the precision of our computer; because these were essentially infinite values, they were eliminated. The y-axis in Figure 5(b) was cropped at 3, because the pattern is apparent. Several interesting features are shown in Figure 5(b). Note the effect of η and α in the $r_{uc}(50, 50, \eta, \alpha) \times \rho_c(100, \eta, \alpha)$ space. For example, if $\eta = 2$ is held constant [Figure 5(b)], then $r_{uc}(50, 50, \eta, \alpha)$ goes to ∞ for small values of α. As α increases, $r_{uc}(50, 50, \eta, \alpha) > 1$, but it approaches 1 as α [and thus $\rho_c(100, \eta, \alpha)$] increase. When $\eta = 0$ is held constant (the exponential model), Figure 5(b) shows that $r_{uc}(50, 50, \eta, \alpha) = 1$ for all values of α. When $\eta = -1$ is held constant, Figure 5(b)

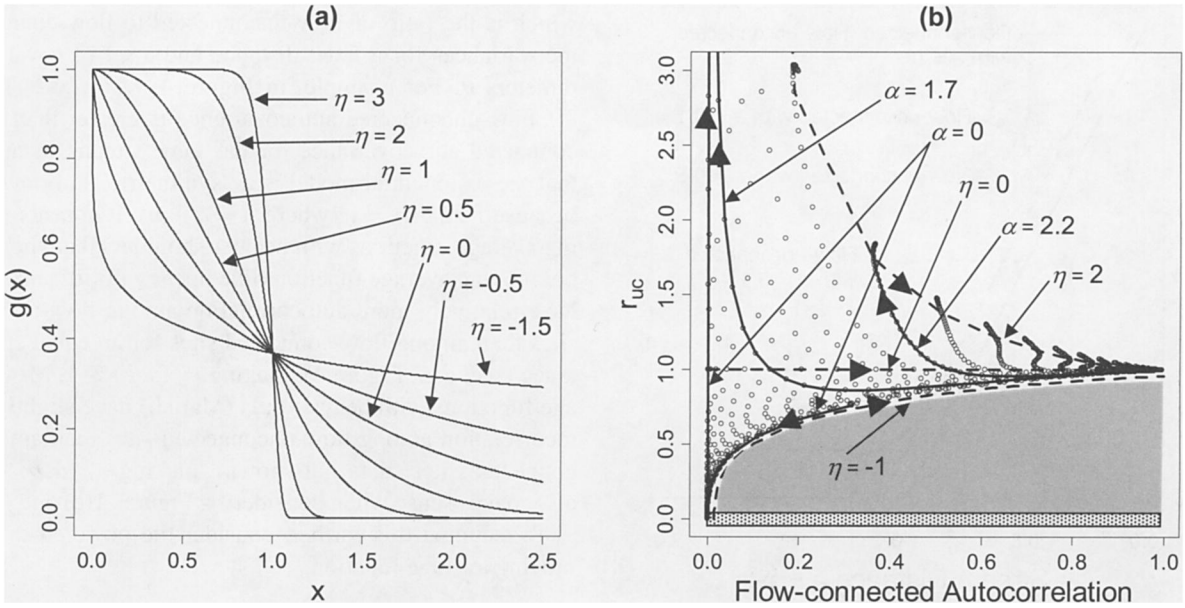

Figure 5. (a) The powered exponential moving average function $g(x) = \exp(-x^{\exp(\eta)}/10^{\alpha})$, where $\alpha = 0$, for different values of η. (b) r_{uc} versus flow-connected autocorrelation for tail-down models using the powered exponential moving average function where distance is fixed. The open circles show all combinations for which both α and η range from -1.5 to 2.5, in increments of 0.1. The solid lines are level curves, where α is held constant. The arrows show the path through r_{uc} × autocorrelation for increasing values of η. The dashed lines are level curves, where η is held constant. The arrows show the path through r_{uc} × autocorrelation for increasing values of α. The stippled bar at the bottom is the r_{uc} for tail-up models. The gray-shaded area cannot be attained with either the tail-down or tail-up models.

shows that $r_{uc}(50, 50, \eta, \alpha)$ goes to 0 for small values of α and $r_{uc}(50, 50, \eta, \alpha) < 1$, but approaches 1 as α increases.

Holding α constant also produces some interesting effects in Figure 5(b). If $\alpha = 2.2$, then $r_{uc}(50, 50, \eta, \alpha)$ is slightly less than 1 for heavy-tailed models with negative η values, and it increases for heavy-shouldered models to close to 2 for the largest values of η as $\rho_c(100, \eta, \alpha)$ decreases to near 0.4. A similar pattern emerges for $\alpha = 1.7$, except that $r_{uc}(50, 50, \eta, \alpha)$ tends toward ∞ and $\rho_c(100, \eta, \alpha)$ tends toward 0 as η increases. A different pattern emerges for the solid line labeled with $\alpha = 0$ in Figure 5(b); initially, $r_{uc}(50, 50, \eta, \alpha)$ and $\rho_c(100, \eta, \alpha)$ decrease as η increases, but at very small $\rho_c(100, \eta, \alpha)$ values, $r_{uc}(50, 50, \eta, \alpha)$ suddenly increases to ∞.

In summary, $r_{uc}(a, b, \eta, \alpha)$ is not independent of $\rho_c(h, \eta, \alpha)$. It is clear that as autocorrelation ($\rho_c(h, \eta, \alpha)$) increases (i.e., the range of the moving average increases), the ratio of autocovariance between flow-connected and flow-unconnected locations goes to 1. Of course, $r_{uc}(a, b, \eta, \alpha) = 0$ for all tail-up models. This means that we cannot attain some values of $r_{uc}(a, b, \eta, \alpha)$ for a given amount of autocorrelation. The gray-shaded area in Figure 5(b) is of particular interest. If, for example, the autocorrelation between flow-connected locations is high (say, 0.7), then we cannot have half as much autocovariance between flow-unconnected locations as that among flow-connected locations. The powered exponential covers the shapes of the four models with analytical solutions developed earlier—linear with sill, spherical, exponential, and Mariah—for which we can develop curves like the dashed lines shown in Figure 5(b). These all fall within the bounds formed by the open circles in Figure 5(b), and the gray-shaded area remains unattainable. This presents a problem, because from an environmental standpoint, it is reasonable to want high autocorrelation in our models [see the x-axis in Figure 5(b)] but to have somewhat less autocovariance for flow-unconnected locations than for flow-connected locations [the gray area in Figure 5(b)]; indeed, the example data in Section 3 demonstrate this. We could tackle this problem with the two-tailed models [Figures 1(e) and (f)] developed in the Appendix, but these models also have severe computational issues. Instead, we turn to mixed models, as discussed next.

2.5 Variance Component Models

The linear model is one of the foundations of statistical applications, including regression and ANOVA,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the dimension of $\mathbf{Y}$ is an $n \times 1$ vector. The relationship between the response variable and covariates is modeled through the design matrix $\mathbf{X}$ and parameters $\boldsymbol{\beta}$. The classical assumption is that the random errors $\boldsymbol{\epsilon}$ are independent, and so $\text{var}(\boldsymbol{\epsilon})$ is $\sigma_n^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. In spatial statistics, the independence assumption is relaxed, and random errors are allowed to be correlated. The geostatistical linear model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z} + \boldsymbol{\epsilon},$$

where $\mathbf{z}$ contains spatially autocorrelated random variables with $\text{var}(\mathbf{z}) = \sigma_z^2\mathbf{R}$, where $\mathbf{R}$ is a correlation matrix. When used for spatial prediction, this model is referred to as "universal" kriging (Le and Zidek 2006, p. 107), with "ordinary" kriging being the special case where the design matrix $\mathbf{X}$ is a single column of 1's (Cressie 1993, p. 119). The general formulation of the covariance matrix $\text{var}(\mathbf{Y}) = \boldsymbol{\Sigma}$ has too many parameters to estimate. Based on such assumptions as ergodicity and stationarity (Cressie 1993, p. 57), Euclidean distance has been used to reduce the number of parameters, and numerous models have

been proposed (e.g., Chiles and Delfiner 1999, p. 80). In the geostatistical linear model, $\text{var}(\mathbf{Y}) = \boldsymbol{\Sigma} = \sigma_z^2\mathbf{R} + \sigma_n^2\mathbf{I}$. In the geostatistical literature, σ_z^2 is called the "partial sill" and σ_n^2 is called the "nugget effect," but $\boldsymbol{\Sigma}$ can be viewed as a variance components model often seen in mixed models.

We can extend the variance component idea to include a mixture of tail-up and tail-down covariance models, along with a nugget effect. Like anisotropy models in geostatistics, this requires five parameters. A general purpose covariance model is

$$\boldsymbol{\Sigma} = \sigma_u^2\mathbf{R}_u + \sigma_d^2\mathbf{R}_d + \sigma_n^2\mathbf{I}, \tag{10}$$

where $\mathbf{R}_u$ is a matrix of autocorrelation values from the tail-up models, $\mathbf{R}_d$ is a matrix of autocorrelation values from the tail-down models, $\mathbf{I}$ is the identity matrix, and σ_u^2, σ_d^2, and σ_n^2 are the variance components. In fact, we could add a component using Euclidean distance with a traditional geostatistical covariance model as well. For example, suppose that autocorrelation among values was caused by an unmeasured covariate related to underlying bedrock characteristics. In such a case, Euclidean distance might be a better distance metric than stream distance. The variance component approach allows a mixture of tail-up and tail-down models that can yield values falling in the gray area of the $r_{uc} \times$ autocorrelation space in Figure 5(b). This is accomplished by creating a proportional mixture between the stippled bar at the bottom and the area encompassed by the open circles. The variance component approach is useful because it allows the development of models containing strong autocorrelation among flow-connected locations, while still maintaining some degree of autocorrelation among flow-unconnected locations. In the next section, we apply the variance component models to several important data sets from the EHMP in Australia.

3. EXAMPLE

The data for this example were provided by the Ecosystem Health Monitoring Program (EHMP) in Southeast Queensland (SEQ), Australia [Figure 6(a)]. We used pH ([H+]), collected in the spring of 2005, and conductivity (μS/cm), collected in the fall of 2006. Conductivity is a measure of the ability of a solution to carry an electrical charge based on ion concentration and temperature. The EHMP has 128 survey sites located throughout SEQ; however, there were missing data values for each variable. The full data set included 117 observations for pH [Figure 6(b)] and 125 observations for conductivity.

We generated the model covariates, hydrologic distances, and spatial weights in a GIS using the Functional Linkage of Waterbasins and Streams (FLoWS) toolset (Theobald et al. 2005, 2006). A suite of watershed covariates were calculated for each stream segment using a 25-m digital elevation model,

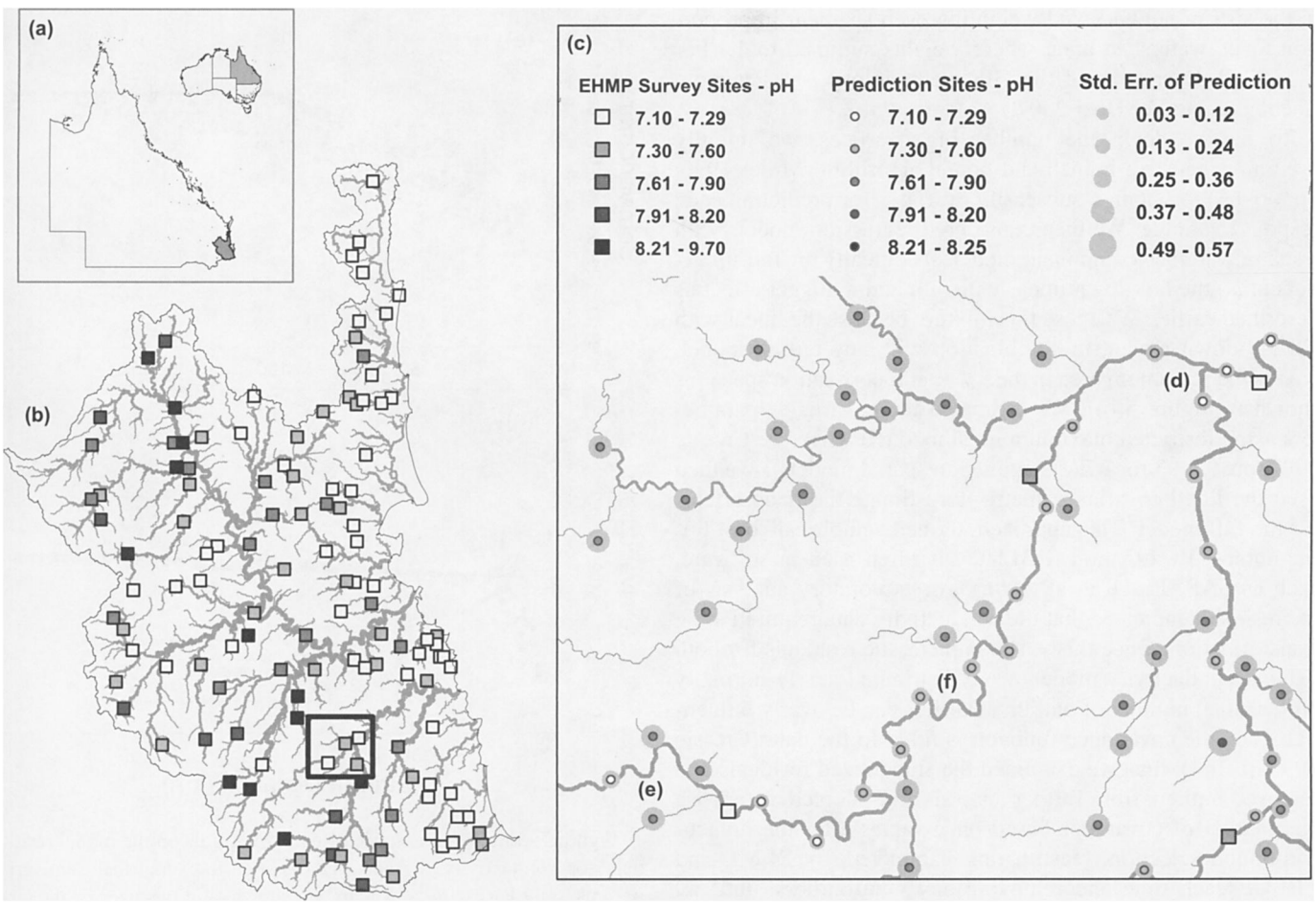

Figure 6. EHMP survey sites. (a) Study area in SEQ. (b) Network of all samples for pH. Line thickness is proportional to watershed area. (c) Close-up of the outlined box, including prediction sites and their standard errors.

the Queensland Geological Mapping data set, and the Queensland Land Use Mapping Program data set. The watershed covariates included % conservation area (land used primarily for conservation purposes, based on the maintenance of the natural ecosystem), % mining (mines, quarries, or tailings), % water (water features), % intensive animal production (intensive forms of animal production such as feedlots), watershed area (an area of land that drains downhill to a common stream outlet), mean slope (mean rise/run∗100), distance upstream from the stream outlet, % stratigraphic rock units (rocks formed in layered succession, including sedimentary, volcanic, and metamorphic rock types), % intrusive rock units (rocks formed by the solidification of cooled magma below the earth's surface), and % compound rock unit (two or more rock types). In addition, we used the elevation at the survey site and four categorical variables for stream type developed by the EHMP: tannin-stained, coastal, upland, and lowland regions. Several stream networks located in SEQ drain directly into the ocean (Figure 6). It is possible to fit spatial models to the data in the individual networks, which could result in unique fixed-effects and covariance parameter estimates for each network. But subdividing the data also would reduce the number of data observations used to fit the models, which in turn would decrease the reliability of the parameter estimates. Thus we chose to fit a single model to all of the observations, but to treat the networks as independent when developing the spatial covariance matrixes. We calculated the spatial weights by locating every confluence in the stream network and weighting each segment in proportion to its watershed area, where weights summed to 1. (For a detailed description of the GIS methodology, see Peterson, Theobald, and Ver Hoef 2007.)

To narrow down the number of covariates, we initially used an exhaustive branch-and-bound algorithm (Miller 1990, pp. 60–63) to obtain a subset of covariates for predicting each response variable. We then considered regression models with a spatial variance component model, specifically the tail-up exponential, the tail-down linear with sill, and a nugget effect as described earlier. We chose this mixture because the linear with sill tail-down model, in combination with any tail-up model, covers the maximum area in the $r_{uc} \times$ autocorrelation space [as shown in Figure 5(b)]. We estimated the covariance parameters using restricted maximum likelihood (REML) (see Cressie 1993, pp. 92–93 for REML applied to spatial models). We then used the fitted covariance matrix to estimate the fixed effects for the full model. This approach, termed "empirical" best linear unbiased estimation (EBLUE), is often used in software, such as SAS (Littell et al. 1996). An exploratory analysis of the residuals indicated that the conductivity data required a log transformation. Once this was complete, the residuals for both pH and conductivity models were distributed nearly normally with a small number of outliers. Outliers can be overly influential when the covariance function is fitted to the data (Cressie 1993, p. 144); thus we examined the studentized residuals and removed outliers from further analysis. We adopted a stepwise elimination of covariates based on p-values after the branch-and-bound selection, reestimating parameters by REML and EBLUE each time and removing any new outliers, until all covariates had a p-value < 0.15. The final data sets included 116 observations for pH [Figure 6(b)] and 122 observations for conductivity.

3.1 Fitted Covariance Model to Conductivity

REML estimates for the final model of conductivity were $\hat{\sigma}_u^2 = 0.00022$ and $\hat{\alpha}_u = 26{,}484.84$ for the tail-up exponential component, $\hat{\sigma}_d^2 = 0.00031$ and $\hat{\alpha}_d = 92.18$ for the tail-down linear with sill component, and $\hat{\sigma}_n^2 = 0.00006$ for the nugget effect. Empirical semivariograms of the residuals were generated using the classical estimator given by Cressie (1993, p. 75), using stream distance in place of Euclidean distance. Pairs of points that were flow-connected were separated from those that were flow-unconnected [Figure 7(a)]. These must be interpreted with some caution. Note that because there is no weighting for flow volume in the empirical semivariogram, using this semivariogram to estimate the covariance parameters would be inappropriate. The absence of spatial weights also hinders visual interpretation of the semivariogram, because sites that are close in space are not guaranteed to be highly correlated; however, this does provide a good visual diagnostic and is instructive. Figure 7(a) also shows the fitted model for each pair of points, again classified as flow-connected and flow-unconnected. The fitted semivariogram model is obtained by taking the estimated sill minus the autocovariance between sites r_i and s_j, $\gamma(r_i, s_j|\hat{\sigma}_n^2, \hat{\sigma}_u^2, \hat{\alpha}_u, \hat{\sigma}_d^2, \hat{\alpha}_d) = \hat{\sigma}_n^2 + \hat{\sigma}_u^2 + \hat{\sigma}_d^2 - C_u(r_i, s_j|\hat{\sigma}_u^2, \hat{\alpha}_u) - C_d(r_i, s_j|\hat{\sigma}_d^2, \hat{\alpha}_d)$.

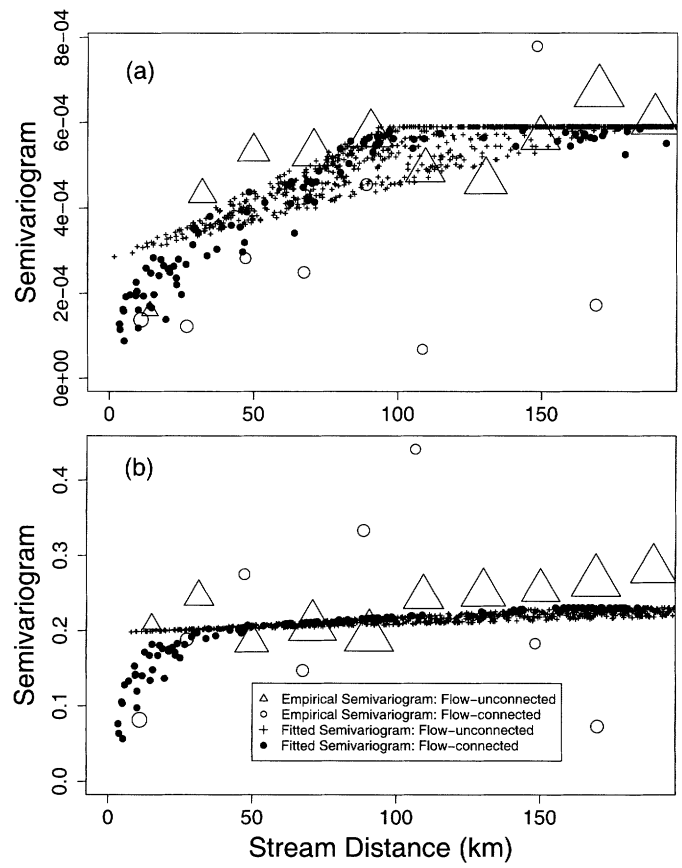

Figure 7. Fitted and empirical semivariograms on the model residuals for conductivity (a) and pH (b). For the empirical semivariograms, only lags with >15 pairs are shown, and the sizes of the circles/triangles are proportional to the number of data pairs averaged for each value. Note that the fitted models are not simple functions of distance.

Note how fitted models for stream networks differ from fitted models in geostatistics or time series, which provide smooth variogram or autocovariance curves as a function of distance. Consider, for example, the flow-connected points in Figure 7(a). For a given distance, the tail-down part of the model will yield a single semivariogram value; however, for the tail-up part of the model, there can be varying numbers of branches between two points, and these can have varying weights [the $\pi_{i,j}$ in (4)]. Thus there is no single fit for a given distance, and so we present the fitted values for all pairs of observed points as a cloud, to emphasize this fact (These could be binned and averaged, just like we do for the empirical semivariogram.) Next, consider the flow-unconnected points in Figure 7(a). From the tail-up part of the model, the covariance among all points is 0; however, from the tail-down part of the model, recall Figure 4. If a is small, then we obtain higher autocovariance, but the range of autocovariance is only slightly greater than $\hat{\alpha}_d = 92.18$; however, if a is just less than $\hat{\alpha}_d$, then we obtain lower autocovariance, but the range of autocovariance can be up to nearly $2\hat{\alpha}_d = 184.36$. These relationships are clearly evident in Figure 7(a). Once again, for flow-unconnected points, autocovariance is not a simple function of distance; thus we present the fitted values for all pairs of observed points.

Two other features of Figure 7(a) merit comment. First, note that $\hat{\alpha}_u = 26{,}484.84$, which means that $C_t(h|\boldsymbol{\theta}) \approx 1$ in (3) for all practical distances, and thus any decrease in the fitted autocovariance with distance is due almost completely to the weights $\pi_{i,j}$ in (4). Second, note that we have relatively strong autocorrelation among flow-connected sites, with somewhat weaker (but still substantial) autocorrelation among flow-unconnected sites. This demonstrates why we spent considerable effort (as described in Sec. 2.4) exploring the model relationships between flow-connected sites and flow-unconnected sites. Neither a pure tail-up model nor a pure tail-down model would fit these data well, and the model wants to be in the gray-shaded area in Figure 5(b), which is why we adopted the variance component model presented in Section 2.5. The covariance model fitted to conductivity data provides an instructive example of the behavior of the variance component models. We next turn to the pH data as an example of the full range of inference, from estimation of fixed effects to prediction of unsampled sites.

3.2 pH Example

REML estimates for the final model of pH were $\hat{\sigma}_u^2 = 0.1969$ and $\hat{\alpha}_u = 55.41$ for the tail-up exponential component, $\hat{\sigma}_d^2 = 0.0333$ and $\hat{\alpha}_d = 154.22$ for the tail-down linear with sill component, and $\hat{\sigma}_n^2 = 0.0002$ for the nugget effect. The tail-up model accounted for much of the variance explained in the pH model $[\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_d^2 + \hat{\sigma}_n^2) = 0.8546]$. The empirical semivariogram for pH [Figure 7(b)] corroborates these results, showing only slight evidence of spatial autocorrelation between flow-unconnected sites. In rivers, pH is a chemical attribute that moves passively downstream through the network. The large variance component for the tail-up model may reflect how pH is strongly influenced by water flowing downstream to a site and less so by the water in other portions of the stream network.

The fitted fixed effects are shown in Table 2. Because of the overparameterized model with streams classed as coastal, lowland, upland, or tannin-stained, the class "Coastal" was set to 0. Table 2 shows that pH decreased with increasing conservation, with less significant effects due to mining and water. The pH was highest in lowland streams, followed by (in decreasing order) upland, coastal, and tannin-stained streams. The importance of a spatial linear model for stream networks, such as that shown in Table 2, is that it allows scientists and managers to establish relationships between variables and gain insight into causes of pH variation throughout the stream network. But not all important factors can be known or measured, and considerable spatial variation in the residuals can remain, as demonstrated by Figure 7(b). This spatial variation, along with covariates, allows for predictions at unsampled locations throughout the stream network, which is another important goal for scientists and managers.

Figure 6(b) shows the locations and values of pH throughout the EHMP monitoring network. Predictions are possible for any location in the network. Figure 6(c) shows a close-up of the bold rectangle in Figure 6(b) that contains four observed values, with predictions made at many locations throughout the small area. One unique property of stream networks is due to branching, which has a major impact on predictions. Figure 6(d) shows a branch above an observed value. Notice how the standard errors increase when going up branches. With no other data nearby upstream, an observed value is obviously influenced by flow from the two upper branches, which increases

Table 2. Fixed-effects estimates for pH

Effect	Estimate[a]	SE[b]	df[c]	t-value	Probability t[d]
Intercept	7.275	0.121	109	60.076	<0.001
% conservation	−0.493	0.189	109	−2.602	0.011
% mining	−8.696	4.670	109	−1.862	0.065
% water	8.945	6.414	109	1.395	0.166
Stream class coastal	0	*NA*	*NA*	*NA*	*NA*
Stream class lowland	0.534	0.128	109	4.184	<0.001
Stream class upland	0.269	0.169	109	1.589	0.115
Stream class tannin-stained	−0.696	0.194	109	−3.583	0.001

[a]Estimated value.

[b]Estimated standard error of the estimate.

[c]Degrees of freedom.

[d]Probability of observed t-value under the null hypothesis.

uncertainty. For example, pH will be determined from the averaging of waters from the two upstream branches. This average could be due to two similar values, or to two widely differing values that yield the same average. Thus branching increases prediction uncertainty, unless there are other data on the upstream branches. Flow volume also has a significant effect. In Figure 6(c), the width of the stream segment is proportional to flow. Clearly, small terminal branches with low flow have very high uncertainty, and predictions tend to center on the expectation of the covariate model. Figure 6(e) shows that the dominance of flow volume allows relatively precise prediction upstream on the main channel, but the small side channels have much higher prediction variances, and predictions tend to drift to the expectation of the covariate model. Whereas branching and flow volume require unique interpretations of predictions for stream networks compared with classical geostatistics, other ideas remain intact; for example, the main stream segment [Figure 6(f)] shows the typical pattern in which prediction standard errors are smaller near the observed locations (both upstream and downstream), with the prediction standard errors increasing with distance from these two locations.

In summary, predictions are affected by the mean model and autocorrelation with nearby sites. The distinguishing features of stream networks are branching and flow volume, which alter autocorrelation relationships so that they are not simple functions of distance, which in turn affects predictions. It also has implications for sampling design on stream networks. This topic is too broad to investigate in the present work, but certainly merits further investigation. We do not expect the typical rules from geostatistics to always apply.

4. DISCUSSION

We have developed spatial autocovariance models for spatially continuous data on stream networks, using a constructive process based on stream distance and moving average functions. We presented four basic tail-up and tail-down models, but many other models remain to be developed. We found that one of the most important and fascinating aspects of stream networks is how autocovariance varies among locations that are flow-connected versus those that are flow-unconnected. Figure 5(b) points up a glaring deficiency in these models. For the one-tailed models, the tail-up models do not allow autocorrelation between flow-unconnected sites (the stippled bar at the bottom of the figure), and the tail-down models generally have more autocorrelation between flow-unconnected sites than between flow-connected sites. With either set of models, it is not possible to achieve substantial autocorrelation between flow-connected sites while at the same time obtaining a small, but still significant amount of autocovariance among flow-unconnected sites [the gray-shaded area in Figure 5(b)]. A two-tail approach could solve this problem, but this entails some computational difficulties.

A simpler alternative is the variance component approach. As shown in Figure 5(b), the highest r_{uc} values are obtained when $\eta \to \infty$; thus a good all-purpose model that covers the maximum extent in the $r_{uc} \times$ autocorrelation space is the linear with sill tail-down model in combination with any tail-up model and a nugget effect. We suggest this to be a useful and practical initial covariance model for an analysis of stream networks. If either the tail-up or tail-down parts of the model are not important (as demonstrated by the size of the variance components σ_i, where $i = u, d, n$), then these can be removed. We have provided several examples to illustrate this approach. Note that in both fitted models, there was higher autocorrelation among flow-connected sites, but some autocorrelation among flow-unconnected sites (Figure 7), justifying the importance of the variance component approach.

The ability to build valid models on stream networks presents many new research opportunities. As mentioned earlier, more models may be found by using moving averages and other methods. Stream network covariance matrixes can replace the more common ones based on Euclidean distance for Bayesian models (Le and Zidek 1992; Handcock and Stein 1993) and model-based geostatistics (Diggle, Tawn, and Moyeed 1998), just to name a few. Space-time models are an area of active research that combines spatial statistics and time series (Le and Zidek 2006), which also have important applications for stream networks. A moving average approach, as shown here, could be extended into three or even more dimensions for space-time models.

APPENDIX: TWO–TAIL MODELS

An obvious extension to the tail-up and tail-down models is a two-tail model. The moving average functions are depicted in Figures 1(e) and (f). Note that these figures are presented as if the tail-up part of the function were the same as the tail-down part of the function. In fact, allowing the halves to differ adds flexibility, but at the cost of a few extra parameters. Because integrals are done piecewise, stream segment by stream segment, there is no advantage to having one continuous moving average function in terms of solving the integrals. As we noted in Section 2, a one-sided moving average function is characterized by two parameters, and thus a flexible two-tail model generally will have four parameters. If we were to add a parameter for a nugget effect (see Sec. 2.5 on variance component models), then there could be up to five covariance parameters. This is very similar to geostatistical models with geometric anisotropy, which generally have five parameters: the nugget, partial sill, range, rotation, and an axis ratio parameter (see, e.g., Schabenberger and Gotway 2005, p. 151).

When two locations are flow-unconnected, the integrals for two-tail models will be equal to those for tail-down models, as is evident in Figures 1(d) and (f). For flow-connected locations [Figure 1(e)], the integrals can be broken into three parts; upstream of the upstream location will be equal to a tail-up model [Figure 1(a)], downstream of the downstream location will be equal to a tail-down model [Figure 1(c)], but the part in between is new. We did not fit two-tail models because of computational issues, which we now explain using Figure A.1 as an example. In Figure A.1, we need to further break up the integral into all segments between branches, which we arbitrarily label stream segments 1, 2, and 3 moving upstream. The tail-up function gets split at each junction with weight $\sqrt{\omega_i}$. Thus the integral between r_1 and s_3 is $\int_{r_1}^{u_1} g_d(s_3 - x|\boldsymbol{\theta}_d) g_u(x - r_1|\boldsymbol{\theta}_u)\, dx + \sqrt{\omega_2} \int_{l_2}^{u_2} g_d(s_3 - x|\boldsymbol{\theta}_d) g_u(x - r_1|\boldsymbol{\theta}_u)\, dx + \sqrt{\omega_2\omega_3} \int_{l_3}^{s_3} g_d(s_3 - x|\boldsymbol{\theta}_d) g_u(x - r_1|\boldsymbol{\theta}_u)\, dx$ where $g_d(x|\boldsymbol{\theta}_d)$ is the tail-down moving average function, controlled by parameters $\boldsymbol{\theta}_d$ and where $g_u(x|\boldsymbol{\theta}_u)$ is the tail-up moving average function, controlled by parameters $\boldsymbol{\theta}_u$. Recall that both $g_d(x|\boldsymbol{\theta}_d)$ and $g_u(x|\boldsymbol{\theta}_u)$ are unilateral with positive support, hence $g_d(s_3 - x|\boldsymbol{\theta}_d)$ takes on nonzero values downstream of s_3.

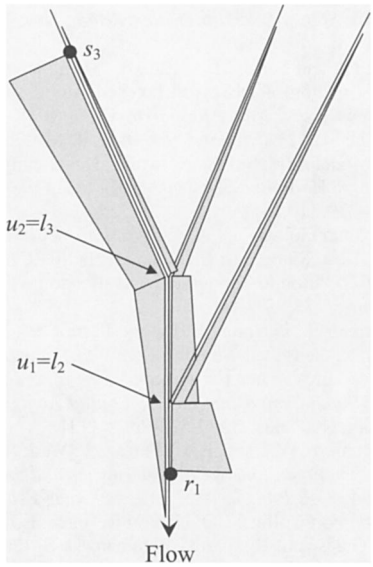

Figure A.1. An example of integrating the tail-up part times the tail-down part in a two-tail model.

In general, then, the integral between r_i and s_j, $r_i < s_j$, where the tail-up part is multiplied by the tail-down part, is

$$\begin{aligned} &C_b(r_i, s_j, u_i, \ldots, l_j, \ldots|\boldsymbol{\theta}) \\ &= \int_{r_i}^{u_i} g_u(x - r_i|\boldsymbol{\theta}_u) g_d(s_j - x|\boldsymbol{\theta}_d)\, dx \\ &\quad + \sum_{m \in D_j^* \setminus D_i} \left(\prod_{k \in D_m \setminus D_i} \sqrt{\omega_k} \right) \int_{l_m}^{u_m} g_u(x - r_i|\boldsymbol{\theta}_u) g_d(s_j - x|\boldsymbol{\theta}_d)\, dx \\ &\quad + \left(\prod_{k \in D_j \setminus D_i} \sqrt{\omega_k} \right) \int_{l_j}^{s_j} g_u(x - r_i|\boldsymbol{\theta}_u) g_d(s_j - x|\boldsymbol{\theta}_d)\, dx, \end{aligned} \tag{A.1}$$

where $\boldsymbol{\theta}_d$ are the parameters from the tail-down part of the moving average function and $\boldsymbol{\theta}_u$ are the parameters from the tail-up part of the moving average function. The two-tailed covariance between r_i and s_j, $r_i < s_j$, is

$$C_2(r_i, s_j|\boldsymbol{\theta}) = \begin{cases} C_n(a, b|\boldsymbol{\theta}_d) & \text{if flow-unconnected} \\ C_c(h|\boldsymbol{\theta}_d) + C_b(r_i, s_j, u_i, \ldots, l_j, \ldots|\boldsymbol{\theta}) & \\ \quad + \pi_{i,j} C_t(h|\boldsymbol{\theta}_u) & \text{if flow-connected,} \end{cases} \tag{A.2}$$

where $\boldsymbol{\theta}$ contains both sets of parameters $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_d$, $C_n(a, b|\boldsymbol{\theta}_d)$ is from (6), $C_c(h|\boldsymbol{\theta}_d)$ is from (5), and $\pi_{i,j} C_t(h|\boldsymbol{\theta}_u)$ is from (4). The main problem with this model is related to the limits of integration in (A.1). The pure tail-up and tail-down models allow us to work simply with two distances: the stream distances for flow-connected sites and the distances to a common junction for flow-unconnected sites. The integrals in (A.1) require us to store a ragged array of all segment distances and ω values between all pair of locations, which would be stored in a fashion similar as in Table A.1. For example, $r_i(1)$ is r distance upstream on the ith stream segment for location indexed by 1, and $B_{i,j}(1, 2)$ is the set of stream segments between locations indexed by 1 and 2. In the foregoing table, rows for pairs of locations that are not flow-connected can be eliminated. Storing information and computing a covariance matrix are much more difficult for two-tailed models than for one-tailed models. This is a consideration when fitting models, which often are iterative and thus require many evaluations of the covariances for the differing parameter values. For this reason, we believe that the variance component models presented earlier are viable alternatives that require much less computation.

Nevertheless, for completeness, we explore a few properties of the two-tail models and develop one model here. Note that the autocovariance ratio between the flow-unconnected and flow-connected sites for a two-tailed model is

$$r_{uc}(r_i, s_j, a, b|\boldsymbol{\theta}) = \frac{C_n(a, b|\boldsymbol{\theta}_d)}{C_c(a + b|\boldsymbol{\theta}_d) + C_b(\cdot|\boldsymbol{\theta}_d) + \pi_{i,j} C_t(h|\boldsymbol{\theta}_u)}. \tag{A.3}$$

Unlike (7), (A.3) depends on the exact locations r_i and s_j and the branching configuration between them. Because the positive values of $C_b(\cdot|\boldsymbol{\theta}_d)$ and $\pi_{i,j} C_t(h|\boldsymbol{\theta}_u)$ in the denominator can help fill in the gray-shaded area in Figure 5(b), this would be a desirable model if computational issues could be solved.

An example of a two-tail model is

$$g_u(x|\sigma_u, \alpha_u) = \sigma_u \exp(-x/\alpha_u) I(0 \le x \le \infty) \tag{A.4}$$

for the tail-up part and

$$g_d(x|\sigma_d, \alpha_d) = \sigma_d \exp(-x/\alpha_d) I(0 \le x \le \infty) \tag{A.5}$$

for the tail-down part. Then, from (A.1), again with $r_i < s_j$,

$$\begin{aligned} &C_b(r_i, s_j, u_i, \ldots, l_j, \ldots|\boldsymbol{\theta}) \\ &= \frac{\sigma_d \sigma_u \alpha_d \alpha_u e^{r_i/\alpha_u - s_j/\alpha_d}}{\alpha_u - \alpha_d} \Bigg(g(u_i, r_i|\alpha_u, \alpha_d) \\ &\quad + \sum_{k \in D_j^* \setminus D_i} \frac{\Omega_k}{\Omega_i} g(u_k, l_k|\alpha_u, \alpha_d) + \frac{\Omega_j}{\Omega_i} g(s_j, l_j|\alpha_u, \alpha_d) \Bigg), \end{aligned} \tag{A.6}$$

where

$$g(x, y|\alpha_u, \alpha_d) = e^{x(1/\alpha_d - 1/\alpha_u)} - e^{y(1/\alpha_d - 1/\alpha_u)}$$

and Ω_k is the additive function value for the kth stream segment, as described in Section 2.2.1. The covariance function for the two-tailed model is now obtained using (A.6) and the flow-connected parts of the tail-down exponential and tail-up exponential in (A.2). This general approach could be used to develop other models as well.

[Received May 2008. Revised January 2009.]

REFERENCES

Barry, R. P., and Ver Hoef, J. M. (1996), "Blackbox Kriging: Spatial Prediction Without Specifying Variogram Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 297–322. [6]

Chiles, J.-P., and Delfiner, P. (1999), *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley. [13]

Table A.1.

$r_i(1)$	$s_j(2)$	$u_i(1)$	$\{l_k, u_k, \omega_k; k \in B_{i,j}(1, 2)\}$	$\cdots$	$l_j(2)$
$r_i(1)$	$s_j(3)$	$u_i(1)$	$\{l_k, u_k, \omega_k; k \in B_{i,j}(1, 3)\}$	$\cdots$	$l_j(3)$
$\vdots$					
$r_i(N-1)$	$s_j(N)$	$u_i(N-1)$	$\{l_k, u_k, \omega_k; k \in B_{i,j}(N-1, N)\}$	$\cdots$	$l_j(N)$

NOTE: A unique index for each location indicated as the index in parentheses.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006), "Spatial Prediction on a River Network," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 127–150. [6,7,9,10]

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, New York: Wiley. [12,14]

Dent, C. I., and Grimm, N. B. (1999), "Spatial Heterogeneity of Stream Water Nutrient Concentrations Over Successional Time," *Ecology*, 80, 2283–2298. [6]

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics (Disc: P326–350)," *Journal of the Royal Statistical Society, Ser. C*, 47, 299–326. [16]

Ecosystem Health Monitoring Program (EHMP) (2006), "Ecosystem Health Monitoring Program 2004–2005 Annual Technical Report," technical report, Moreton Bay Waterways and Catchments Partnership, Brisbane, Australia. [6]

Fuentes, M. (2002), "Spectral Methods for Nonstationary Spatial Processes," *Biometrika*, 89, 197–210. [6]

Ganio, L. M., Torgersen, C. E., and Gresswell, R. E. (2005), "A Geostatistical Approach for Describing Spatial Pattern in Stream Networks," *Frontiers in Ecology and the Environment*, 3, 138–144. [6]

Gardner, B., Sullivan, P. J., and Lembo, A. J., Jr. (2003), "Predicting Stream Temperatures: Geostatistical Model Comparison Using Alternative Distance Metrics," *Canadian Journal of Fisheries and Aquatic Sciences*, 60, 344–351. [6]

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410. [16]

Higdon, D. (1998), "A Process-Convolution Approach to Modelling Temperatures in the North Atlantic Ocean (Disc: P191–192)," *Environmental and Ecological Statistics*, 5, 173–190. [6]

Higdon, D., Swall, J., and Kern, J. (1999), "Non-Stationary Spatial Modeling," in *Bayesian Statistics 6—Proceedings of the Sixth Valencia International Meeting*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. Smith, New York: Clarendon Press [Oxford University Press], pp. 761–768. [6]

Le, N. D., and Zidek, J. V. (1992), "Interpolation With Uncertain Spatial Covariances: A Bayesian Alternative to Kriging," *Journal of Multivariate Analysis*, 43, 351–374. [16]

——— (2006), *Statistical Analysis of Environmental Space-Time Processes*, New York: Springer. [12,16]

Legleiter, C. J., Lawrence, R. L., Fonstad, M. A., Marcus, W. A., and Aspinall, R. (2003), "Fluvial Response a Decade After Wildfire in the Northern Yellowstone Ecosystem: A Spatially Explicit Analysis," *Geomorphology*, 54, 119–136. [6]

Littell, R. C., Milliken, R. C., Stroup, W. W., and Wolfinger, R. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Publishing. [14]

Miller, A. J. (1990), *Subset Selection in Regression*, New York: Chapman & Hall. [14]

Money, E., Carter, G., and Serre, M. L. (2009), "Using River Distances in the Space/Time Estimation of Dissolved Oxygen Along Two Impaired River Networks in New Jersey," *Water Research*, 43 (7), 1948–1958. [10]

Peterson, E. E., Theobald, D. M., and Ver Hoef, J. M. (2007), "Support for Geostatistical Modeling on Stream Networks: Developing Valid Covariance Matrices Based on Hydrologic Distance and Stream Flow," *Freshwater Biology*, 52, 267–279. [10,14]

Schabenberger, O., and Gotway, C. A. (2005), *Statistical Methods for Spatial Data Analysis*, Boca Raton, FL: Chapman & Hall/CRC. [16]

Shreve, R. L. (1967), "Infinite Topographically Random Channel Networks," *Journal of Geology*, 75, 178–186. [10]

Theobald, D., Norman, J., Peterson, E. E., and Ferraz, S. (2005), "Functional Linkage of Watersheds and Streams (FLoWs): Network-Based ArcGIS Tools to Analyze Freshwater Ecosystems," in *Proceedings of the Second Annual International Symposium on GIS Spatial Analyses in Fishery and Aquatic Sciences*, Redlands, CA: ESRI Press. [13]

Theobald, D., Norman, J., Peterson, E. E., Ferraz, S., Wade, A., and Sherburne, M. R. (2006), *Functional Linkage of Waterbasins and Streams (FLoWs) v1 User's Guide: ArcGIS Tools to Analyze Freshwater Ecosystems*, Colorado State University, Fort Collins, CO: Natural Resource Ecology Lab. [13]

Torgersen, C. E., Gresswell, R. E., and Bateman, D. S. (2004), "Pattern Detection in Stream Networks: Quantifying Spatial Variability in Fish Distribution," in *Proceedings of the Second Annual International Symposium on GIS Spatial Analyses in Fishery and Aquatic Sciences*, eds. T. Nishida, P. J. Kailola, and C. E. Hollingworth, Saitama, Japan: Fishery GIS Research Group, pp. 405–420. [6]

Ver Hoef, J. M., and Barry, R. P. (1998), "Constructing and Fitting Models for Cokriging and Multivariable Spatial Prediction," *Journal of Statistical Planning and Inference*, 69, 275–294. [6]

Ver Hoef, J. M., Cressie, N., and Barry, R. P. (2004), "Flexible Spatial Models for Kriging and Cokriging Using Moving Averages and the Fast Fourier Transform (fft)," *Journal of Computational and Graphical Statistics*, 13, 265–282. [6]

Ver Hoef, J. M., Peterson, E. E., and Theobald, D. (2006), "Spatial Statistical Models That Use Flow and Stream Distance," *Environmental and Ecological Statistics*, 13, 449–464. [6-10]

Yaglom, A. M. (1987), *Correlation Theory of Stationary and Related Random Functions*, Vol. I, New York: Springer-Verlag. [7]

Yuan, L. L. (2004), "Using Spatial Interpolation to Estimate Stressor Levels in Unsampled Streams," *Environmental Monitoring and Assessment*, 94, 23–38. [6]

Comment: Statistical Dependence in Stream Networks

Noel CRESSIE and David O'DONNELL

This note is based on an invited discussion of the article, "A Moving Average Approach for Spatial Statistical Models on Stream Networks" by Jay M. Ver Hoef and Erin E. Peterson. Ver Hoef and Peterson (hereafter VHP) have extended the idea of flow-related statistical dependence in streams to one where dependence may not respect flow, such as might happen when modeling data on fish in connected streams. We congratulate VHP for their innovative paper on using moving average models in stream networks.

1. STREAM NETWORKS AS GRAPHS

A stream network could be viewed as a directed graph with nodes defined by the stream confluences, edges defined by the stream segments between confluences, and the directions of the edges defined by the direction of flow. Assuming the network under study has no side channels or a complicated delta, then it has a tree-like appearance. In fact, if the directions defined by flow are *reversed*, then the directed graph is a *rooted tree* (e.g., Lauritzen 1996, p. 6), where the "root node" is the outlet and the "branches" are smaller and smaller streams. In this discussion, we shall draw analogies between statistical models built on graphs and the sort of models VHP are building. For example, the analogous models to VHP's *tail-down models* are those based on rooted trees. It is important to be aware that the

Noel Cressie is Director of the Program in Spatial Statistics and Environmental Statistics, Professor of Statistics, and Distinguished Professor of Mathematical and Physical Sciences, The Ohio State University (OSU), Columbus, OH 43210-1247 (E-mail: *ncressie@stat.osu.edu*). David O'Donnell is Ph.D. Student in the Department of Statistics, the University of Glasgow, Glasgow G12 8QW, Scotland. Cressie's research was supported by The Office of Naval Research under grant N00014-08-1-0464, and the discussion was prepared while O'Donnell was visiting the Department of Statistics at OSU under a Jim Gatheral Scholarship awarded through the University of Glasgow.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap09530