2010

A mixed-model moving-average approach to geostatistical modeling in stream networks

Erin E. Peterson

CSIRO Division of Mathematics, Informatics, and Statistics, 120 Meiers Road, Indooroopilly, Queensland 4068 Australia

Jay M. Ver Hoef

National Marine Mammal Lab, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service

Ecology, 91(3), 2010, pp. 644–651

Reports

A mixed-model moving-average approach to geostatistical modeling in stream networks

Erin E. Peterson[1,3] and Jay M. Ver Hoef[2]

[1]*CSIRO Division of Mathematics, Informatics, and Statistics, 120 Meiers Road, Indooroopilly, Queensland 4068 Australia*
[2]*National Marine Mammal Lab, Alaska Fisheries Science Center, NOAA National Marine Fisheries Service, 7600 Sand Point Way NE, Seattle, Washington 98115-6349 USA*

Abstract. Spatial autocorrelation is an intrinsic characteristic in freshwater stream environments where nested watersheds and flow connectivity may produce patterns that are not captured by Euclidean distance. Yet, many common autocovariance functions used in geostatistical models are statistically invalid when Euclidean distance is replaced with hydrologic distance. We use simple worked examples to illustrate a recently developed moving-average approach used to construct two types of valid autocovariance models that are based on hydrologic distances. These models were designed to represent the spatial configuration, longitudinal connectivity, discharge, and flow direction in a stream network. They also exhibit a different covariance structure than Euclidean models and represent a true difference in the way that spatial relationships are represented. Nevertheless, the multi-scale complexities of stream environments may not be fully captured using a model based on one covariance structure. We advocate using a variance component approach, which allows a mixture of autocovariance models (Euclidean and stream models) to be incorporated into a single geostatistical model. As an example, we fit and compare "mixed models," based on multiple covariance structures, for a biological indicator. The mixed model proves to be a flexible approach because many sources of information can be incorporated into a single model.

Key words: *geostatistics; hydrologic distance; moving average; scale; spatial autocorrelation; streams.*

Introduction

Tobler's first law of geography states that "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). In the field of geostatistics, this phenomenon is referred to as spatial autocorrelation or spatial autocovariance, which quantitatively represents the degree of statistical dependency between random variables using spatial relationships (Cressie 1993). Geostatistical models are somewhat similar to the conventional linear statistical model; they have a deterministic mean function, but the assumption of independence is relaxed and spatial autocorrelation is permitted in the random errors. For example, in a universal kriging model the deterministic mean is assumed to vary spatially and is modeled as a linear function of known explanatory variables (in contrast to ordinary kriging where the mean is unknown, but constant). Local deviations from the mean are then modeled using the spatial autocorrelation between nearby sites. Thus, geostatistical models are typically able to model additional variability in the response variable and make more accurate predictions when the data are spatially autocorrelated. For detailed information about geostatistical methods, please see Cressie (1993) or Chiles and Delfiner (1999).

Manuscript received 12 September 2008; revised 21 August 2009; accepted 15 October 2009. Corresponding Editor: J. A. Jones.
[3] E-mail: erin.peterson@csiro.au

Spatial autocorrelation is particularly relevant in freshwater stream environments where nested watersheds and flow connectivity may produce patterns that are not captured by Euclidean distance (Dent and Grimm 1999, Torgersen and Close 2004, Ganio et al. 2005, Monestiez et al. 2005, Peterson et al. 2006, Ver Hoef et al. 2006; Ver Hoef and Peterson, *in press*). Thus, aquatic ecologists may have been hesitant to use traditional geostatistical methods, which depend on Euclidean distance, because they did not make sense from an ecological standpoint. Covariance matrices based on Euclidean distance do not represent the spatial configuration, longitudinal connectivity, discharge, or flow direction in a stream network. In addition to being ecologically deficient, many common autocovariance functions are not generally valid when Euclidean distance is simply replaced with a hydrologic distance measure (Ver Hoef et al. 2006). A generally valid autocovariance function is guaranteed to produce a covariance matrix that is symmetric and positive-definite, with all nonnegative diagonal elements, regardless of the configuration of the stream segments or sample sites. If these conditions are not met, it may result in negative prediction variances, which violates the assumptions of geostatistical modeling. These issues

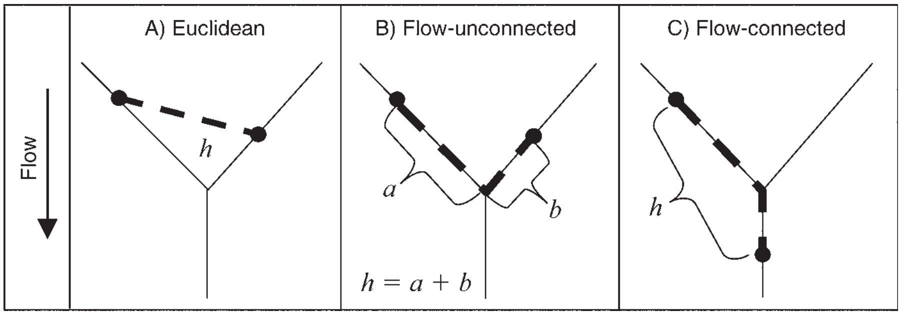


FIG. 1. The distance, h or $h = a + b$, between locations (black circles) is represented by a dashed line. (A) Euclidean distance is used to represent Euclidean relationships. Hydrologic distance can be used to represent both (B) flow-unconnected and (C) flow-connected relationships in a stream network. Water must flow from one site to another in order for the pair to be considered flow connected (C). In contrast, flow-unconnected sites do not share flow but do reside on the same stream network (B).

made it necessary to develop new geostatistical methodologies for stream networks, which permit valid covariances to be generated based on a variety of hydrologic relationships (Cressie et al. 2006, Ver Hoef et al. 2006; Ver Hoef and Peterson, *in press*). Our goal is to introduce these new autocovariance models to ecologists.

There are currently two types of distance measures that can be used to produce valid covariance matrices for geostatistical modeling in stream networks (when used with the appropriate autocovariance function): Euclidean distance and hydrologic distance. Euclidean distance is the straight-line distance between two locations and all locations within a study area have the potential to be spatially correlated when it is used (Fig. 1A). A hydrologic distance is simply the distance between two locations when measurement is restricted to the stream network. In contrast to Euclidean distance, locations within a study area do not automatically have the potential to be spatially correlated when a hydrologic distance is used. Instead, rules based on network connectivity and flow direction can be used to prevent spatial autocorrelation between locations and thus represent different hydrologic relationships. For example, a "flow-connected" relationship requires that water flow from one location to another for two sites to be correlated (Fig. 1C). When this condition is not met, sites have a "flow-unconnected" relationship (Fig. 1B) and these two sites can be made spatially independent. Likewise, whole networks (i.e., stream segments that share a common stream outlet anywhere downstream) can be made dependent.

Freshwater stream environments are typically considered open systems (Townsend 1996) with complex processes and interactions occurring between and within multiple aquatic and terrestrial scales. As a result, multiple patterns of spatial autocorrelation may be present in freshwater ecosystems (Peterson et al. 2006; Ver Hoef and Peterson, *in press*). The strength of each pattern may also vary at different spatial scales since the influence of environmental characteristics has been shown to vary with scale (Sandin and Johnson 2004). Consequently, we believe that a geostatistical model based on a mixture of covariances (i.e., multiple spatial relationships) may better fit the data than a model based on a single covariance structure.

Geostatistical models for stream network data are relatively new and may be unfamiliar to aquatic scientists. Here we review the current state of geostatistical modeling techniques for stream networks. A model-based biological indicator collected by the Ecosystem Health Monitoring Program (Bunn et al., *in press*) is used as a case study to demonstrate the approach. Simple worked examples and more details are given in the on-line appendices.

Stream network models

The stream network models of Ver Hoef et al. (2006) and Cressie et al. (2006) are based on moving-average (MA) constructions. MA constructions are flexible and can be used to create a large number of autocovariance functions (Barry and Ver Hoef 1996). They are developed by creating random variables as the integration of a MA function over a white noise random process. For our purposes, the key point is that spatial autocorrelation occurs when there is overlap between the MA function of one random variable and that of another, which we explain in greater detail for two classes of models.

Tail-up models

Models that are based on hydrologic distance and only allow autocorrelation for flow-connected relationships are referred to as "tail-up" (TU) models (Ver Hoef et al. 2006; Ver Hoef and Peterson, *in press*) because the tail of the MA function points in the upstream direction. There are a large number of TU MA functions that one could use such as the exponential, spherical, linear-with-sill, or mariah models (Ver Hoef et al. 2006), and each has a unique shape, which determines a unique autocorrelation function. Autocorrelation occurs when MA functions overlap among sites, with greater autocorrelation resulting from greater overlap. For example, in Fig. 2A, B, the shape of the TU MA

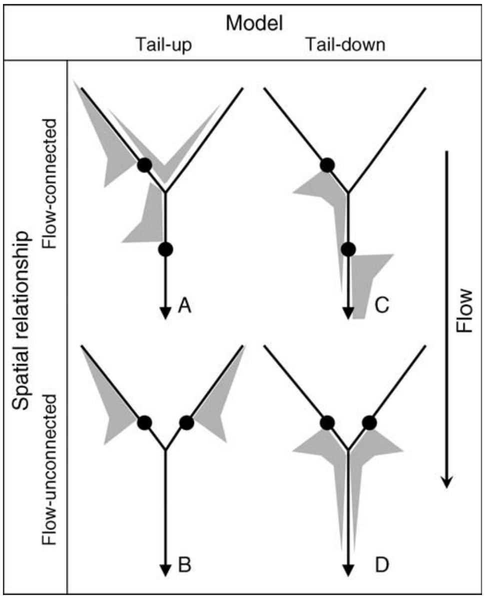


FIG. 2. The moving-average functions for the (A, B) tail-up and (C, D) tail-down models in both (A, C) flow-connected and (B, D) flow-unconnected cases. The moving-average functions are shown in gray with the width representing the strength of the influence for each potential neighboring location. Spatial autocorrelation occurs between locations when the moving-average functions overlap (A, C, and D).

function (shown in gray) dictates the relative influence of upstream values when the random variable for a specific location is generated. Values at short upstream hydrologic distances have a larger influence because there is a greater amount of overlap in the MA functions and this influence tends to decrease with distance upstream.

To better understand how the TU MA function is applied to the unique conditions of a stream network, imagine applying the function to a small network, moving upstream segment by segment. When two locations are flow-unconnected the tails of the MA functions do not overlap (Fig. 2B) and the locations do not have the potential to be spatially correlated. When two locations are flow-connected, the MA functions may overlap and one location has the potential to be spatially correlated with another location (Fig. 2A). When the MA function reaches a confluence in the network, segment weights (called segment PIs in Appendix A) are used to proportionally (i.e., they must sum to 1) allocate, or split, the function between upstream segments (Fig. 2A). The MA 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. Instead, segment weights can be used to ensure that locations residing on segments that have the strongest influence on conditions at a downstream location are given greater weight in the model (Ver Hoef et al. 2006). Thus, autocorrelation in the TU models depends on flow-connected hydrologic distance. However, it also depends on the number of confluences found in the path between flow-connected sites and the weightings assigned to each of the tributaries. This feature in particular makes stream network models different than classical geostatistical models based on Euclidean distance. It also makes TU models particularly useful for modeling organisms or materials that move passively downstream, such as waterborne chemicals.

The segment weights can be based on any ecologically relevant feature, such as discharge, which is thought to represent relative influence in a stream network. However, discharge data are rarely available for every stream segment throughout a region. As an alternative, watershed area is sometimes used as a surrogate for discharge (Peterson et al. 2007). 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 or area 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 (see Appendix A for details on how to construct valid segment weights). The segment weights may also be based on measures that represent the sum of the upstream measures (i.e., a segment does not contribute anything to itself), such as Shreve's stream order (Shreve 1966), as used by Cressie et al. (2006). Using segment weights that are normalized so that they sum to one ensures that all random variables have a constant variance (Ver Hoef et al. 2006) if desired, which is typical in geostatistics. As an aside, the MA construction could also allow for non-stationary variances, but those models will not be explored here.

The construction of a TU covariance matrix is based on the hydrologic distance *and* a spatial weights matrix (developed from the segment weights, as illustrated in Appendix A) between flow-connected locations. Furthermore, all flow-unconnected locations are uncorrelated. Additional details and a simple worked example are provided in Appendix A to more clearly illustrate the construction of a TU covariance matrix.

Tail-down models

Tail-down (TD) models allow autocorrelation between *both* flow-connected and flow-unconnected pairs of sites in a stream network (Ver Hoef and Peterson, *in press*). The MA function for a TD model is defined so that it is only non-zero downstream of a location. In other words, the tail of the MA function points in the downstream direction (Fig. 2C, D). Spatial autocorrelation is modeled somewhat differently in a flow-connected vs. flow-unconnected situation due to the way the overlap occurs in the MA functions (Fig. 2C, D). Notice also that the input data requirements are unique for each case. The total hydrologic distance, h (Fig. 1C), is used for flow-connected pairs, but the hydrologic distances, a and b, from each site to a common confluence are used for flow-unconnected pairs (Fig. 1B). As before, more overlap in the MA function implies more autocorrelation. Also, segment weights are not used to model flow-connected relationships since the MA function points downstream (Fig. 2C) and there is no need to split the function to maintain constant variances. Additional details and a simple worked example showing the TD construction of a covariance matrix are provided in Appendix A.

Although the TD model allows 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*) in the TD models. In fact, none of the current models are able to generate a TD model with significantly stronger spatial autocorrelation between flow-connected pairs (Ver Hoef and Peterson, *in press*) than flow-unconnected pairs for an equal hydrologic distance. These restrictions on spatial autocorrelation in the TD model may make sense for fish populations that have the tendency to invade upstream reaches, such as nonnative brook trout (*Salvelinus fontinalis*; Peterson and Fausch 2003). Yet, there are other situations where it would be useful to generate a model with stronger spatial autocorrelation between flow-connected pairs than flow-unconnected pairs. Peterson and Fausch (2003) also studied the movement characteristics of native cutthroat trout (*Oncorhynchus clarki*) and found that they moved downstream much more often than upstream. Here, a model with stronger spatial autocorrelation between flow-connected locations, which also allows for some spatial autocorrelation between flow-unconnected locations, might best fit the data. Yet, neither the TU or TD model can be used to represent this particular covariance structure. Therefore, another modeling solution is required. Here we turn to the mixed model, which is a very general approach, to address these issues.

Mixed models

The mixed model is closely related to the basic linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where the matrix $\mathbf{X}$ contains measured explanatory variables and the parameter vector $\boldsymbol{\beta}$ establishes the relationship of the explanatory variables to the response variable, contained in the vector $\mathbf{y}$. The random errors are contained in the vector $\boldsymbol{\varepsilon}$, and the general formulation is $\mathrm{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is a matrix. The mixed-model is simply a variance component approach, which allows the error term to be expanded into several random effects ($\mathbf{z}.$):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\sigma}_{\mathrm{EUC}}\mathbf{z}_{\mathrm{EUC}} + \boldsymbol{\sigma}_{\mathrm{TD}}\mathbf{z}_{\mathrm{TD}} + \boldsymbol{\sigma}_{\mathrm{TU}}\mathbf{z}_{\mathrm{TU}} + \boldsymbol{\sigma}_{\mathrm{NUG}}\mathbf{z}_{\mathrm{NUG}} \tag{2}$$

where $\mathrm{cor}(\mathbf{z}_{\mathrm{EUC}}) = \mathbf{R}_{\mathrm{EUC}}$, $\mathrm{cor}(\mathbf{z}_{\mathrm{TD}}) = \mathbf{R}_{\mathrm{TD}}$, $\mathrm{cor}(\mathbf{z}_{\mathrm{TU}}) = \mathbf{R}_{\mathrm{TU}}$ are matrices of autocorrelation values for the Euclidean (EUC), TD, and TU models, $\mathrm{cor}(\mathbf{z}_{\mathrm{NUG}}) = \mathbf{I}$ where NUG is the nugget effect, $\mathbf{I}$ is the identity matrix, and σ^2_{EUC}, σ^2_{TD}, σ^2_{TU}, and σ^2_{NUG} are the respective variance components. The mixed-model construction implies that covariance matrices based on different types of models, such as the EUC, TU, and TD are combined to form a valid covariance mixture:

$$\Sigma = \sigma^2_{\mathrm{EUC}}\mathbf{R}_{\mathrm{EUC}} + \sigma^2_{\mathrm{TD}}\mathbf{R}_{\mathrm{TD}} + \sigma^2_{\mathrm{TU}}\mathbf{C}_{\mathrm{TU}} \odot \mathbf{W}_{\mathrm{TU}} + \sigma^2_{\mathrm{NUG}}\mathbf{I} \tag{3}$$

where we have further decomposed $\mathbf{R}_{\mathrm{TU}} = \mathbf{C}_{\mathrm{TU}} \odot \mathbf{W}_{\mathrm{TU}}$ into the Hadamard product of the flow-connected autocorrelations $\mathbf{C}_{\mathrm{TU}}$ (unweighted) and the spatial weights matrix $\mathbf{W}_{\mathrm{TU}}$.

The variance component model is attractive for several reasons. First, it solves the problem mentioned in the previous section; namely that the combination of the TU covariance matrix and TD covariance matrix allows for the possibility of more autocorrelation among flow-connected pairs of sites, with somewhat less autocorrelation among flow-unconnected pairs of sites. Secondly, the multiple range parameters can capture patterns at multiple scales. Generally, large scale patterns are the most obvious and explanatory variables are incorporated to help explain them. Spatial patterns of intermediate scale, which have not been measured with explanatory variables, are captured by the range parameters for the EUC, TU, and TD models and the relative strength of each model is given by its variance component. The spatial weights are used to capture the influence of branching in the network, flow direction, and discharge. Finally, some spatial variation occurs at a scale finer than the closest measurements; these are modeled as independent error, which is represented by the nugget effect. We now turn to an example to make these concepts clearer.

Example

The Ecosystem Health Monitoring Program (EHMP) has been collecting indicators of biotic structure and ecosystem function throughout South East Queensland (SEQ), Australia (Fig. 3A) since 2002 (Bunn et al., *in press*). The program aims are to evaluate the condition and trend in ecological health of freshwater environments and to guide investments in catchment protection and rehabilitation. Metrics based on freshwater fish assemblages are commonly used as indicators of ecological health because they are thought to provide a holistic approach to assessment across broad spatial and temporal scales (Harris 1995). In this example, we used a model-based biological indicator, the proportion of native fish species expected (PONSE), which is simply the ratio of observed to expected native freshwater fish species richness (Kennard et al. 2006). The expected species composition data were generated using a referential model and represent the native species that are expected to be present in a physically similar, but undisturbed stream (Kennard et al. 2006). The observed species composition data were collected at 86 survey sites (Fig. 3B) in the spring of 2005. Hereafter, we will refer to these as the "observed sites." PONSE scores at the observed sites ranged from 0.09 to 1, with a mean of 0.76 and a median of 0.83. We also nonrandomly selected 137 "prediction sites" where PONSE was not generated. Additional information about the study area, the sampling methods, and the predictive model used to generate the PONSE scores is provided in Appendix B.

We generated the spatial data necessary for geostatistical modeling in a geographic information system (GIS). These included seven explanatory variables representing watershed-scale land use, land cover, and topographic characteristics for each observed and prediction site, as well as the hydrologic distances and spatial weights, which were based on watershed area. The EHMP classified streams based on elevation, mean annual rainfall, stream order, and stream gradient to create four EHMP regions (Bunn et al., *in press*), which we also included as a site-scale explanatory variable. Additional information about the GIS methodology and explanatory variables can be found in Appendix B.

We used a two-step model selection procedure to compare models. First we fixed the covariance structure and focused on selecting the explanatory variables using the Akaike information criterion (AIC; Akaike 1974). During the second phase of model selection we focused on selecting the most appropriate covariance structure. We fixed the selected explanatory variables, and then compared every linear combination of TU, TD, and EUC covariance structures, where four different autocovariance functions were tested for each model type. This resulted in a total of 124 models, each with a different covariance structure. In addition, we fit a classical linear model assuming independence to compare to models that use spatial autocorrelation. Once the final model was identified, universal kriging (Cressie 1993) was used to make predictions at the 137 prediction sites. Please see Appendix B for details on the model selection procedure.

Our results show that a model based on a mixture of covariances produced more precise PONSE predictions than models based on a single covariance structure (Table 1). When more than one covariance structure was incorporated, mixture models that included the TD model appeared to outperform other mixture types. In addition, all of the geostatistical models outperformed the classical linear model. The lowest RMSPE value was produced by the exponential TU/linear-with-sill TD mixture model. Hereafter this will be referred to as the "final model." The final model only contained one explanatory variable, mean slope in the watershed, which was positively correlated with PONSE. This statistical relationship may represent a physical relationship between PONSE and an anthropogenic disturbance gradient such as land use, water quality, channel or riparian condition, or in-stream habitat (Kennard et al. 2006), which is correlated to slope. For example, watersheds with steeper slopes might be expected to have less cleared or cropped land and, as a result higher PONSE scores. More details on the fitted explanatory variables and diagnostics are given in Appendix B.

We examined the percent of the variance explained by each of the covariance components to provide more information about the covariance structure of the models. In the final model, the TU, TD, and NUG components explained 22.43%, 64.32%, and 13.25% of the variance, respectively. Although the full covariance mixture (TU/TD/EUC) was not the best model in this example based on the RMSPE (Table 1), the loss in predictive ability was only 0.34% when it was used instead of the final model. One of the advantages of a mixed model is its flexibility; a covariance mixture can be used to represent the whole range of covariance structures used in the mixture (i.e., single EUC, TU, or TD or any combination of the three). Therefore, we recommend fitting a full covariance mixture; this allows the data to determine which variance components have the strongest influence, rather than making an implicit assumption about the spatial structure in the data by using a single covariance structure.

The predictions and prediction standard errors produced by the final model (Fig. 3C–F) exhibit some of the common characteristics of kriging predictions. Predictions and their standard errors vary depending on the estimated regression coefficients and distances to observed data sites. If the explanatory variables at the prediction site are not well represented in the observed data set a large standard error will be assigned to the prediction. The predictions change gradually along stream segments (Fig. 3C–F) and the prediction standard errors tend to be smaller near observed data and increase as a function of distance (Fig. 3C).

The predictions and prediction standard errors shown in Fig. 3 also demonstrate some characteristics that are unique to geostatistical models for stream networks. For example, the TU model allows discontinuities in predictions at confluences, which enables flow-unconnected tributaries to receive markedly disparate PONSE predictions (Fig. 3D). The effect of the spatial weights on prediction uncertainty is apparent if upstream segments have not been sampled (Fig. 3E). In this situation, uncertainty is relatively high upstream of a confluence because various combinations of the two PONSE scores could be contributing to the downstream observed score. Two sites located on the main stem are strongly correlated (based on a combination of the tail-up and tail-down models) and uncertainty in the

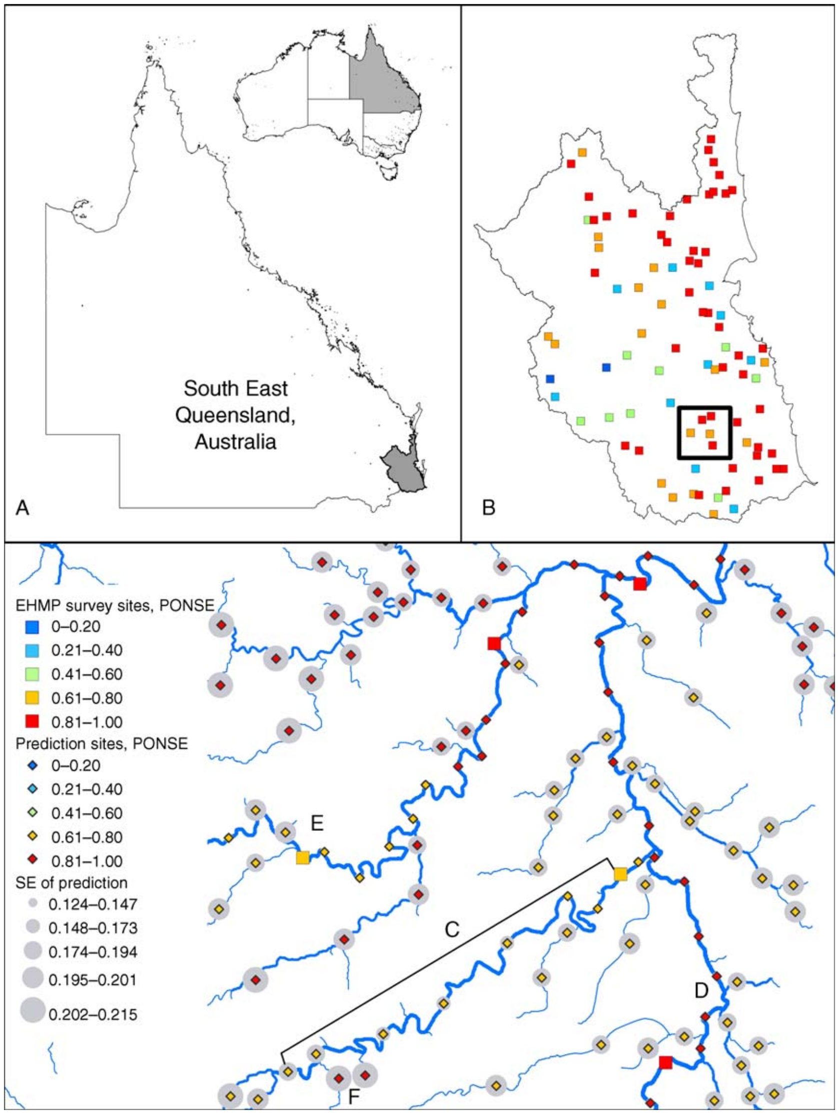


FIG. 3. (A, B) The Ecosystem Health Monitoring Program collects a suite of ecological indicators at survey sites located throughout South East Queensland, Australia. Predictions and prediction variances for the proportion of native fish species expected (PONSE) were produced using the exponential tail-up/linear-with-sill tail-down model (C–F). Observed sites are represented by larger squares, and prediction sites are represented by smaller diamonds. Prediction variances are shown in gray and are sized in proportion to their value (0.124–0.215). Thus, predictions with a large shaded gray area have less precision. Blue line segments symbolize stream segments, and the width of the line is proportional to the watershed area.

TABLE 1. A comparison of mixture models.

Mixture	Model 1	Model 2	Model 3	RMSPE
Nonspatial				0.2573
TU	linear-sill			0.2428
TD	linear-sill			0.2112
EUC	spherical			0.2283
TU/TD	exponential	linear-sill		0.2088
TU/EUC	linear-sill	Gaussian		0.2190
TD/EUC	linear-sill	spherical		0.2103
TU/TD/EUC	mariah	linear-sill	exponential	0.2094

Notes: Models shown represent the best model fit for each mixture type based on the root mean square prediction error (RMSPE). Models are tail-up (TU), tail-down (TD), and Euclidean (EUC).

downstream predictions is low (Fig. 3E). However, when the spatial weights are based on watershed area, and one upstream segment dominates a side branch, uncertainty in the upstream direction may also be low (Fig. 3C). In contrast, prediction uncertainty in small upstream tributaries is relatively large (based on the tail-up model) since the spatial correlation between the observed and prediction site is weak (Fig. 3F).

Discussion and Conclusions

Spatial autocorrelation is clearly a natural phenomenon given the open nature of stream ecosystems (Townsend 1996) and the complexity of process interactions occurring within and between the stream and the terrestrial environment. Observable patterns of spatial autocorrelation are likely caused by multiple spatially dependent factors (Wiens 2002). As a result, conceptualization and modeling in stream ecosystems requires tools that are able to account for dynamic multi-scale patterns ranging from the reach to the network scale (Townsend 1996). Yet many models do not account for these natural interdependencies. Ignoring spatial autocorrelation makes it possible to use traditional statistical methods, which rest on the assumption of independent random errors. Although convenient, our opinion is that it is better to develop new statistical methods that represent the unique ecological conditions found in the environment.

Traditional geostatistical methods account for spatial correlation in the error term, but they may not fully capture spatial autocorrelation structures in stream environments. Euclidean covariance functions, such as the spherical, cubic, exponential, and Gaussian, are all strikingly similar (Chiles and Delfiner 1999). As a result, two sites that are spatially correlated using one function are likely to be spatially correlated using another function. In contrast, the stream models have a markedly different autocovariance structure than the Euclidean models and represent a true difference in the way that spatial relationships are represented. The splitting of the covariance function along a branching network is also unique to stream models. Previously, autocorrelation has been restricted to a linear feature or to two-dimensional space, which cannot be used to capture hydrologic patterns of spatial autocorrelation in a branching network. The stream models given here provide an innovative statistical alternative because they were specifically developed to represent the spatial configuration, longitudinal connectivity, discharge, and flow direction in a stream network.

The mixed model is an extremely flexible approach because many sources of information can be incorporated into a single model. The explanatory variables are used to account for influential factors that can be measured. However, it is common for other influential variables to be left unmeasured due to a lack of resources or an incomplete understanding about the stream process. In a mixed model, autocorrelated errors can be modeled at multiple scales using a variety of distance measures. This produces a rich and complex covariance structure, that when combined with the explanatory variables, can be used to account for the effects of both measured and unmeasured variables at multiple scales. Given the multi-scale complexities of stream ecosystems, we expect these models to better represent the spatial complexity and interdependencies in a streams data sets.

The flexibility of the mixture model makes the method suitable for modeling a variety of variables collected within or near a stream network. Although we used a Gaussian response in the example, the autocovariance functions described here can also be used to produce covariance matrices for kriging Poisson or binomial variables, such as fish counts or the presence or absence of species. It may also be possible to model variables that are present in riparian areas rather than streams, but are expected to exhibit both Euclidean and hydrologic patterns of spatial autocorrelation. This might include riparian plant species that employ waterborne dispersal strategies or animal species that migrate along stream corridors. Finally, we chose to use watershed area to calculate the spatial weights because it made sense given the response variable, but any measurement, such as water velocity, stream width, or discharge, could be used as long as it meets the statistical requirements set out in Appendix A. In principle, multiple weighting schemes could be compared as part of the model selection criteria.

The mixed model is also useful from a management perspective since predictions with estimates of uncertainty may be generated throughout a stream network (Fig. 3). This ability provides a way to move from disjunct stream management, which is traditionally based on site or reach-scale samples, to a more continuous approach that yields location specific predictions and accounts for the network as a whole (Fausch et al. 2002). For example, the predicted PONSE values in two small tributaries were relatively high compared to those found in the main stem (Fig. 3F), which may indicate that those locations have the potential to act as natural refugia for native fish residing in marginally suitable habitats. The potential importance of these tributaries could go unnoticed without the ability to evaluate the network as a whole. This quality is particularly useful because it helps to ensure that management actions are targeted or scaled appropriately (Lake et al. 2007). Including estimates of uncertainty also enables users to gauge the reliability of the predictions and to target future sampling efforts in areas with large amounts of uncertainty or a greater potential for ecological impairment.

Geostatistical modeling in stream networks has the potential to be a powerful statistical tool for freshwater stream research and management. It can be used to capture and quantify spatial patterns at multiple scales, which may provide additional information about ecosystem structure and function (Levin 1992); a key step in developing new ecological theories. Our goal has been to make this methodology accessible to ecologists so that the models can be implemented, modified, and improved to derive additional information from streams data sets. We believe that when geostatistical models for stream networks are used in conjunction with sound ecological knowledge the result will be a more ecologically representative model that may be used to broaden our understanding of stream ecosystems.

Acknowledgments

We thank the SEQ Healthy Waterways Partnership for sharing the data set and Beth Gardner, Mark Kennard, Petra Kuhnert, and an anonymous reviewer for their constructive comments on the manuscript. This project received financial support from the National Marine Fisheries Service of NOAA, the CSIRO Division of Mathematical and Information Sciences, and the Australian Water for a Healthy Country Flagship.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–722.

Barry, R. P., and J. M. Ver Hoef. 1996. Blackbox kriging: spatial prediction without specifying variogram models. Journal of Agricultural, Biological, and Environmental Statistics 1:297–322.

Bunn, S., E. Abal, M. Smith, S. Choy, C. Fellows, B. Harch, M. Kennard, and F. Sheldon. *In press*. Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation. Freshwater Biology.

Chiles, J., and P. Delfiner. 1999. Geostatistics: modeling spatial uncertainty. John Wiley and Sons, New York, New York, USA.

Cressie, N. 1993. Statistics for spatial data. Revised edition. 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.

Dent, C. L., and N. B. Grimm. 1999. Spatial heterogeneity of stream water nutrient concentrations over successional time. Ecology 80:2283–2298.

Fausch, K. D., C. E. Torgersen, C. V. Baxter, and H. W. Li. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream reaches. BioScience 52: 483–498.

Ganio, L. M., C. E. Torgersen, and R. E. Gresswell. 2005. A geostatistical approach for describing spatial pattern in stream networks. Frontiers in Ecology and the Environment 3:138–144.

Harris, J. H. 1995. The use of fish in ecological assessments. Australian Journal of Ecology 20:65–80.

Kennard, M. J., B. D. Harch, B. J. Pusey, and A. H. Arthington. 2006. Accurately defining the reference condition for summary biotic metrics: a comparison of four approaches. Hydrobiologia 572:151–70.

Lake, P. S., N. Bond, and P. Reich. 2007. Linking ecological theory with stream restoration. Freshwater Biology 52:597–615.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology 73:1943–1967.

Monestiez, P., J.-S. Bailly, P. Lagacherie, and M. Voltz. 2005. Geostatistical modelling of spatial processes on directed trees: application to fluvisol extent. Geoderma 128:179–191.

Peterson, D. P., and K. D. Fausch. 2003. Upstream movement by nonnative brook trout (*Salvenlinus fontinalis*) promotes invasion of native cutthroat trout (*Oncorhynchus clarki*) habitat. Canadian Journal of Fisheries and Aquatic Sciences 60:1502–1516.

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.

Peterson, E. E., D. M. Theobald, and J. M. Ver Hoef. 2007. Geostatistical modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. Freshwater Biology 52:267–279.

Sandin, L., and R. K. Johnson. 2004. Local, landscape and regional factors structuring benthic macroinvertebrate assemblages in Swedish streams. Landscape Ecology 19:501–514.

Shreve, R. L. 1966. Statistical law of stream numbers. Journal of Geology 74:17–37.

Tobler, W. R. 1970. Computer movie simulating urban growth in Detroit region. Economic Geography 46:234–240.

Torgersen, C. E., and D. A. Close. 2004. Influence of habitat heterogeneity on the distribution of larval Pacific lamprey (*Lampetra tridentata*) at two spatial scales. Freshwater Biology 49:614–630.

Townsend, C. R. 1996. Concepts in river ecology: pattern and process in the catchment hierarchy. Archiv für Hydrobiologie 113(Large Rivers Supplement):3–21.

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.

Wiens, J. A. 2002. Riverine landscape: taking landscape ecology into the water. Freshwater Biology 47:501–515.

APPENDIX A

Formulas and examples for the construction of tail-up and tail-down covariance matrices (*Ecological Archives* E091-048-A1).

APPENDIX B

Additional details about the collection and analysis of the EHMP PONSE data set (*Ecological Archives* E091-048-A2).

Ecological Archives E091-048-A1

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

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

where $C_1(h \mid \boldsymbol{\theta})$ 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, $C_1(h \mid \boldsymbol{\theta})$ could be

$$C_1(h \mid \boldsymbol{\theta}) = \sigma_{TU}^2 \left(1 - \frac{h}{\alpha}\right) I\left(\frac{h}{\alpha} \le 1\right) \tag{A.2}$$

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

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

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

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

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

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

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

Note that creating a covariance matrix, $\mathbf{C}_1$, based on $C_1(h \mid \boldsymbol{\theta})$ alone, is not guaranteed to be a valid covariance matrix until it has been weighted appropriately using the spatial weights matrix, **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, $\boldsymbol{\Sigma}_{TU} = \mathbf{W} \odot \mathbf{C}_1$, 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 $\mathbf{\Sigma}_{TU}$ for the example data provided in Figs. A1 and A2.

$$\Sigma_{TU} = \begin{pmatrix} 4.00 & 1.33 & 0.80 & 0 \\ 1.33 & 4.00 & 1.87 & 0.27 \\ 0.80 & 1.87 & 4.00 & 2.40 \\ 0 & 0.27 & 2.40 & 4.00 \end{pmatrix} \circ \begin{pmatrix} 1 & 0 & 0.77 & 0.71 \\ 0 & 1 & 0.64 & 0.59 \\ 0.77 & 0.64 & 1 & 0.92 \\ 0.71 & 0.59 & 0.92 & 1 \end{pmatrix} = \begin{pmatrix} 4.00 & 0 & 0.62 & 0 \\ 0 & 4.00 & 1.20 & 0.16 \\ 0.62 & 1.20 & 4.00 & 2.21 \\ 0 & 0.16 & 2.21 & 4.00 \end{pmatrix} \tag{A.6}$$

Note that matrix **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,

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

the tail-down spherical model,

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

and the tail-down mariah model.

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

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)

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

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 (σ^2_{TD} = 4 and α = 15) we obtain the tail-down covariance matrix for the example data provided in Fig. A1

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

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.

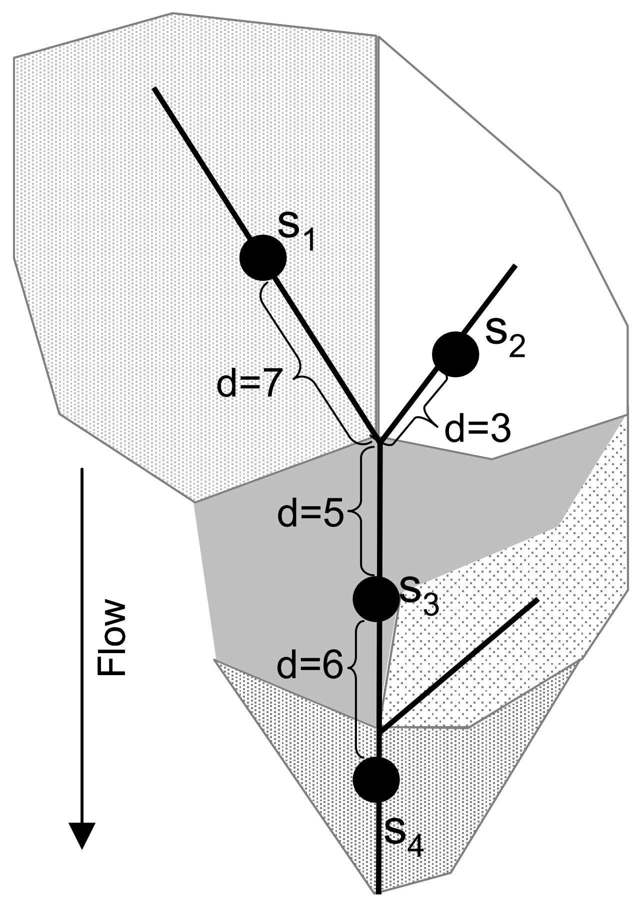
Handcock, M. S., and M. L. Stein. 1993. A Bayesian analysis of kriging. Technometrics 35:403–410.

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.



Downstream Hydrologic Distance

FROM Site (a)

TO Site (b)	s_1	s_2	s_3	s_4
s_1	0	3	0	0
s_2	7	0	0	0
s_3	12	8	0	0
s_4	18	14	6	0

(A)

$a+b=h$

Total Hydrologic Distance

FROM Site (h)

TO Site (h)	s_1	s_2	s_3	s_4
s_1	0	10	12	18
s_2	10	0	8	14
s_3	12	8	0	6
s_4	18	14	6	0

(B)

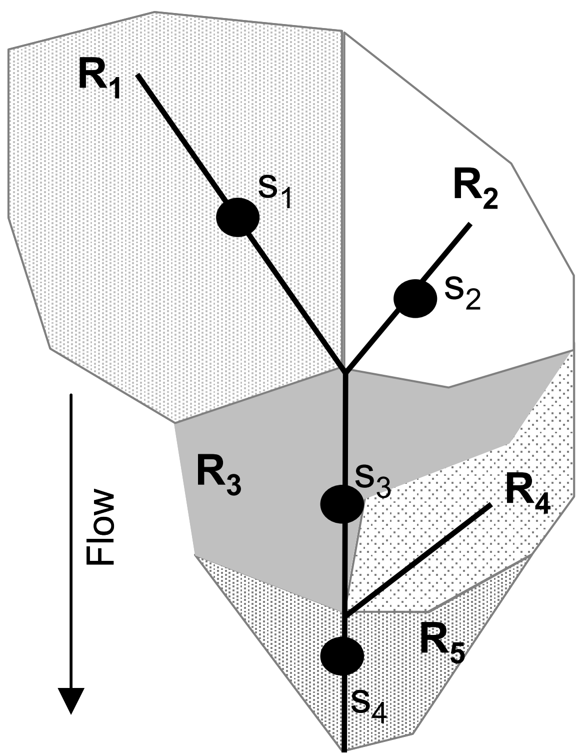
FIG. A1. The downstream hydrologic distance for each pair of sites, a and b, is needed to fit the tail-down model (A). The tail-up model requires the total hydrologic distance, h, between each pair (B). The total hydrologic distance between sites (B) can be derived from the downstream distance matrix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

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

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

(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)

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

FROM Site

TO Site	s_1	s_2	s_3	s_4
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.

Ecological Archives E091-048-A2

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 B: Additional details for the analysis of the PONSE data set.

South East Queensland (SEQ) is located on the eastern coast of Australia and is approximately 22,999 square km in size. It is a subtropical region with mean annual maximum temperatures ranging between 21 and 29 °C (EHMP, 2001) and total annual rainfall ranges between 900 and 1800 mm (Pusey et al. 1993). Elevation ranges between 0 meters in coastal areas to 1360 meters in the west along the Great Dividing Range. Regional discharge is seasonally variable and is related to elevation, slope, and rainfall (EHMP, 2001). The predominant land uses in the region are natural bushland (37%) and grazing (35%).

The Ecosystem Health Monitoring Program (EHMP) has been collecting indicators of biotic structure and ecosystem function throughout SEQ since 2002 (Bunn et al. *in press*). The program aims are to evaluate the condition and trend in ecological health of freshwater environments and to guide investments in watershed protection and rehabilitation. Metrics based on fish assemblage structure and function are commonly used as indicators of ecological health because they are thought to provide a holistic approach to assessment across broad spatial and temporal scales (Harris 1995). The EHMP uses ecosystem health indicators based on freshwater fish species richness (Kennard et al. 2006*a*), fish assemblage composition (Kennard et al. 2006*b*) and the relative abundance of alien species (Kennard et al. 2005). In this example, we used a model-based fish indicator, the proportion of native fish species expected (PONSE), which is the ratio of observed to expected native fish species richness (Bunn et al. *in press*, Kennard et al. 2006*a*). Here, species richness is simply the number of native fish species observed or predicted at a location. Native species richness is a commonly used summary indicator of ecosystem health that may reflect a variety of disturbances functioning at a range of spatial and temporal scales. Environmental degradation is expected to alter naturally diverse fish communities to simple assemblages dominated by only a few tolerant species. Therefore, species richness is expected to decline with increasing environmental stress. For example, Kennard and others (2006*a*) showed that native species richness was lower than expected at test sites in SEQ affected by intensive watershed land use (clearing, grazing, cropping and urbanization), degraded water quality (high diel temperature range, low pH, high conductivity, high turbidity) and riparian and aquatic habitat degradation.

The observed species richness values used to calculate the PONSE scores were based on fish assemblage data collected at 86 survey sites during the spring of 2005; hereafter referred to as the “observed sites”. Fish assemblages were sampled using a combination of backpack electrofishing and seine-netting where possible (Kennard et al. 2006*c*). Electrofishing was conducted using a Smith-Root model 12B Backpack Electroshocker, while seine-netting was conducted using a 10m long (1.5m drop) pocket seine of 10mm (stretched) mesh. An attempt was made to intensively sample all habitat unit types (pool, riffle, and run) at each site (Kennard et al. 2006*c*). When only one habitat unit type was present, at least two units were fished. The average length of fished stream was 75m and data from the entire length were combined to obtain an observed species richness value.

The expected species richness values used to calculate each PONSE score were generated using a referential model and represent the number of native species that are expected to be present in a physically similar, but undisturbed stream (Kennard et al. 2006*a*). A regression tree was used to partition a large set of minimally-disturbed reference sites into homogenous groups based on environmental characteristics including elevation, distance to the stream outlet, distance from the stream source, and the mean wetted stream width (Kennard et al. 2006*a*). The mean species richness for each reference site group was then calculated. The 86 observed sites were assigned to reference site groups based on the same environmental characteristics mentioned previously. Finally, each observed site was assigned an expected species richness value, which was equal to the mean species richness for the reference site group. For additional details about the calculation of the expected species richness values, please see Kennard and others (2006*a*) and EHMP (2001).

We generated the spatial data necessary for geostatistical modeling in a geographic information system (GIS) using ArcGIS version 9.2 software (ESRI 2006). The Functional Linkage of Waterbasins and Streams (FLoWS) toolset (Theobald et al. 2006) was used to construct a landscape network, which is a spatial data structure designed to store topological relationships between nodes, directed edges, and polygons (Theobald et al. 2005). Here, we used the landscape network to represent stream segments (directed edges) and confluences (nodes). The location of additional features, such as survey sites, can also be associated with the data structure. We incorporated the 86 observed sites and 137 non-randomly selected “prediction sites” (where PONSE scores were not measured) into the landscape network. Scripts written in Python version 2.4.1 (Van Rossum and Drake 2003) were used to generate the hydrologic distances and spatial weights between all observed and prediction sites based on the topological relationships stored in the landscape network. The spatial weights were based on watershed area, which was used as a surrogate for discharge. This seemed like a viable alternative since mean annual discharge has been shown to be correlated with watershed area in other regions (Vogel et al. 1999).

Watershed-scale explanatory variables for each observed and prediction site were calculated using a combination of FLoWS tools (Theobald et al. 2006) and customized scripts written in Python version 2.4.1 (Van Rossum and Drake 2003). The streams data were provided by the Moreton Bay Waterways and Catchments Partnership (2005); here we considered a single GIS polyline feature as a stream segment. A digital elevation model (Queensland Natural Resources and Water 2000) and the Queensland Land Use Mapping Program data set (Bureau of Rural Sciences 2002) were used to generate seven watershed explanatory variables (Table B1). The EHMP classified streams based on elevation, mean annual rainfall, stream order, and stream gradient (EHMP 2001) to create four EHMP regions: Tannin-stained, Coastal, Lowland, and Upland, which we also included as a site-scale explanatory variable (Table B1). We checked the explanatory variables for collinearity using the variance inflation factor statistic (Helsel and Hirsch 1992), which indicated that it was unnecessary to remove explanatory variables. Note that all of the statistical analyses described here were performed in R version 2.6.1 (R Development Core Team 2008).

We used a two-step model selection procedure to compare models. First we fixed the covariance structure and focused on the explanatory variables. All of the mixed models were fit using the exponential tail-up (TU), linear-with-sill tail-down (TD), and Gaussian Euclidean (EUC) as component models (formulas for component models are given in Appendix A). The choice of autocovariance function for the tail-up and Euclidean components was somewhat arbitrary because only one type of spatial relationship is represented by each of the functions (flow-connected or Euclidean). Differences in the strength of spatial autocorrelation between locations when different autocovariance functions are fit to the data result from the shape of the autocovariance function and are likely to be minimal. In the case of the tail-down model the difference in performance between autocovariance functions is attributed to both the shape of the function *and* the way in which flow-connected and flow-unconnected relationships are represented. More specifically, the relative strength of spatial autocorrelation for each relationship differs depending on the autocovariance function that is used. For example, consider a case where there are two pairs of locations. One pair is flow-unconnected and the other is flow-connected and the distance between pairs is equal ($a+b=h$). If the tail-down exponential model is fit to the data the strength of spatial autocorrelation between the two pairs is equal. However, if the linear-with-sill function is fit to the data the strength of spatial autocorrelation between the flow-unconnected pair may be up to 1.5 times that of the flow-connected pair. We chose the linear-with-sill tail-down autocovariance function because we wanted to model maximum autocorrelation among flow-unconnected sites relative to flow-connected sites. All tail-up models have zero autocorrelation among flow-unconnected sites, so any function allows us to model minimum autocorrelation among flow-unconnected sites relative to flow-connected sites. The combination of linear-with-sill tail-down plus any tail-up model in a mixed model allows the most flexible modeling of autocorrelation. More details may be found in Ver Hoef and Peterson (*in press*).

We estimated all of the parameters in the first phase of model selection using maximum likelihood. Outliers can be overly influential when likelihoods are calculated (Martin 1980) and when the covariance function is fit to the data (Cressie 1993). Therefore, a backwards stepwise model selection strategy was implemented so that the model diagnostics could be examined for every model. Nevertheless, no outliers were identified or removed during the analysis. We selected explanatory variables based on the smallest Akaike Information Criterion (AIC) (Akaike 1974) for the fitted models to determine which had the most support in the data.

During the second phase of model selection we focused on selecting the most appropriate covariance structure. We used the selected explanatory variable set as described above to compare every linear combination of TU, TD, and EUC covariance structure, where four different autocovariance functions were tested for each model type. For the TU and TD models, these included the spherical, exponential, mariah, and linear-with-sill functions (formulas are given in Appendix A). The EUC model included the spherical, exponential, Gaussian, and Cauchy functions (Chiles and Delfiner 1999). This resulted in a total of 124 models, each with a different covariance structure. In addition, we fit a classical linear model assuming independence to compare to models that use spatial autocorrelation. Maximum likelihood may produce biased estimates of the covariance parameters and so we used restricted maximum likelihood (REML) for parameter estimation (Cressie 1993). Note that REML was not used for parameter estimation in the first phase of model selection because a side effect of REML is that information criteria, such as AIC, can only be used if the explanatory variable set remains fixed (Verbeke and Molenberghs 2000). Once the fitted covariance matrix had been generated using REML, it was then used to estimate the fixed effects. This is referred to as "empirical" best linear unbiased estimation and is often used in software such as SAS (Littell et al. 1996).

Leave-one-out cross-validation predictions were generated at the observed sites for each of the 124 models using universal kriging (Cressie 1993) and then used to calculate the root-mean-squared-prediction error (RMSPE),

$$\text{RMSPE} = \sqrt{\sum_{i=1}^{n} (\hat{z}_i - z_i)^2 / n} \tag{B.1}$$

where $\hat{z}_i$ is the prediction of the *i*th datum after removing it from the observed data set. The RMSPE can be thought of as proportional to the prediction interval and its interpretation is fairly straight-forward. For example, if the RMPSE for the best model is one third that of a competing model the gain in using the best model is 66%. The RMSPE was used to compare the models based on different covariance structures. Other model selection criteria, such as information theoretic measures (Burnham and Anderson 1998), could also have been used. The method that we chose met our purpose, which was to identify the covariance mixture that provided the best predictions and it is not our intention to debate the merits of various approaches here. Once the model with the smallest RMSPE was identified, universal kriging was used to make predictions at the 137 prediction sites.

An empirical semivariogram of the model residuals, divided into flow-connected and flow-unconnected relationships (Fig. B1), was generated using the classical estimator as given in Cressie (1993). We also used the covariance parameter estimates to calculate the percent of the variance explained by each of the covariance components (%VC)

$$\%VC_{TU} = \frac{\sigma_{TU}^2}{\sigma_{TD}^2 + \sigma_{TU}^2 + \sigma_{NUG}^2} * 100, \qquad \text{(B.2)}$$

where σ_{TU}^2 and σ_{TD}^2 are the partial sill parameters for the TU and TD models and σ_{NUG}^2 is the nugget effect. Eq. B.2 demonstrates how to calculate the %VC for the TU model, which in this example is part of a two-component mixture model. However, the %VC can be calculated for any number of model components and the nugget effect. Since it is a percentage, the %VC for all of the components should always sum to 100.

The lowest RMSPE value was produced by the exponential TU/linear-with-sill TD mixture model, which we will hereafter refer to as the final model. The fixed effects for the final model are provided in Table B2. Note that the parameter estimates in Table B2 were generated using REML during the second phase of model selection, but the specific fixed effects were chosen in the first phase of model selection. The final model only contained one explanatory variable, mean slope in the watershed, which was positively correlated with PONSE. This statistical relationship may represent a physical relationship between PONSE and an anthropogenic disturbance gradient related to land use, water quality, channel or riparian condition, or in-stream habitat (Kennard et al. 2006*a*), rather than slope itself. For example, watersheds with steeper slopes might be expected to have less cleared or cropped land and, as a result higher PONSE scores. In addition, steep slopes are more likely to be found in headwater streams, which may be inaccessible to alien species if they are upstream of barriers to fish movement (Kennard et al. 2005). However, we did not have access to extensive information about anthropogenic disturbances at EHMP sites and were unable to specifically account for these effects in our model.

An empirical semivariogram of the final model residuals, divided into flow-unconnected and flow-connected relationships is shown in Fig. B1. The partial sill and range parameters for the TU component were 0.0162 and 1160.56 km, respectively. In contrast, the TD model had a larger partial sill (0.0464) and a smaller range parameter (110.67 km) than the TU model. The nugget effect was 0.0096. The TU range parameter was substantially larger than the largest flow-connected separation distance (262.19 km). The magnitude of the TU range parameter simply indicates that flow-connected sites are spatially correlated regardless of their location in the stream network. Covariance parameters are sometimes visually estimated from the semivariogram, but it would not be appropriate in this case since there is no weighting for the flow-connected locations. The weights affect the autocorrelation between sites and so there is no guarantee that sites close in space have a strong influence on one another if there is a confluence between them. Calculating the percent of the covariance explained by each of the components provides additional information about the covariance structure of the models. In the final model the TU, TD, and NUG components explained 22.43%, 64.32%, and 13.25% of the variance, respectively.

Our results show that a model based on a mixture of covariances produced the most precise PONSE predictions (Fig. B2). Models based on a covariance mixture tended to produce smaller RMSPE values than models based on a single covariance type. When more than one covariance type was incorporated, mixture models that included the TD model outperformed other mixture types. There appeared to be more within model type variation in the TD model compared to the TU and EUC model types (Fig. B2). As we discussed previously, this reflects differences in model fit related to both the shape of the autocovariance function and the way that spatial autocorrelation is represented between flow-connected and flow-unconnected pairs. All of the spatial models out performed the classical linear model, with the final model demonstrating an 18.86% gain based on the RMSPE.

Given that the final model only demonstrated a 1.17% gain in predictive ability on the TD model, some might question the usefulness of using a mixture model. Nevertheless, it is unclear which covariance type or mixture will provide the best model fit in advance. A geostatistical methodology is used to model the spatial structure in the residual error and so the observable patterns may not reflect the pattern or scale that would be expected. For example, the unexplained variability in fish distribution may be related to a strongly influential explanatory variable, such as elevation, that was not proposed during model selection. In that case, Euclidean patterns of spatial autocorrelation in fish distribution (resulting from elevation) might be observed at broad spatial scales even though fish movement is generally restricted to the stream network and some species may not migrate large distances. This is a simple example of how environmental factors can produce patterns of spatial autocorrelation; in a freshwater environment, extremely complex process interactions occurring within and between the stream and the terrestrial environment would be expected. Therefore, we recommend using a full covariance mixture (TU/TD/EUC) rather than attempting to make an a priori assumption about the covariance structure of the data. Although the full covariance mixture was not the best model in this example, the loss in predictive ability was only 0.34% when it was used. These results demonstrate an important feature of the covariance mixture approach; namely, that it is flexible enough to represent the entire range of covariance structures (i.e., single EUC, TU, or TD or any combination of the three). This flexibility also makes the approach suitable for a wide variety of data sets collected within a stream network.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6):716–722.

BRS (Bureau of Rural Sciences). 2002. Land use Mapping at Catchment Scale: Principles, Procedures and Definitions, Edition 2. Bureau of Rural Sciences, Department of Agriculture, Fisheries and Forestry – Australia. Kingston, ACT, Australia. p 46.

Bunn, S., E. Abal, M. Smith, S. Choy, C. Fellows, B. Harch, M. Kennard, and F. Sheldon. *In Press*. Integration of science and monitoring of river ecosystem health to guide investments in catchment protection and rehabilitation. Freshwater Biology.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Chiles, J., and P. Delfiner. 1999. Geostatistics: Modeling spatial uncertainty. John Wiley and Sons, New York, New York, USA.

Cressie, N. 1993. Statistics for Spatial Data, Revised Edition. Page 900 p. John Wiley and Sons, New York, New York, USA.

EHMP (Ecosystem Health Monitoring Program). 2001. Design and Implementation of Baseline Monitoring (DIBM3): Developing an ecosystem health monitoring program for rivers and streams in southeast Queensland, Report to the Southeast Queensland Regional Water Quality Management Strategy. Brisbane, Australia.

ESRI (Environmental Systems Research Institute). 2006. ArcGIS: Release 9.2 [software]. Redlands, California: Environmental Systems Research Institute.

Harris, J. H. 1995. The use of fish in ecological assessments. Australian Journal of Ecology 20:65–80.

Helsel, D. R., and R. M. Hirsch. 1992. Statistical Methods in Water Resources. Elsevier Science Publishing, New York, New York, USA.

Kennard, M. J., A. H. Arthington, B. J. Pusey, and B. D. Harch. 2005. Are alien fish a reliable indicator of river health? Freshwater Biology 50:174–193.

Kennard, M. J., B. D. Harch, B. J. Pusey, and A. H. Arthington. 2006*a*. Accurately defining the reference condition for summary biotic metrics: a comparison of four approaches. Hydrobiologia 572:151–170.

Kennard, M. J., B. J. Pusey, A. H. Arthington, B. D. Harch, and S. J. Mackay. 2006*b*. Development and application of a predictive model for freshwater fish assemblage composition to evaluate river health in eastern Australia. Hydrobiologia 572:33–57.

Kennard, M. J., B. J. Pusey, B. H. Harch, E. Dore, and A. H. Arthington. 2006*c*. Estimating local stream fish assemblage attributes: sampling effort and efficiency at two spatial scales. Marine and Freshwater Research 57:635–653.

Littell, R. C., R. C. Milliken, W. W. Stroup, and R. Wolfinger. 1996. SAS System for Mixed Models, Cary, North Carolina: SAS publishing.

Martin, R. D. 1980. Robust estimation of autoregressive models. Pages 228–254 *in* D. R. Brillinger and G. C. Tiao, editors. Directions in times series: Proceedings of the IMS special topics meeting on time series analysis. Institute of Mathematical Statistical, Haywood, California, USA.

Moreton Bay Waterways and Catchments Partnership. 2005. South East Queensland Streams and Catchments Version 2. Moreton Bay Waterways and Catchments Partnership, Brisbane, QLD, Australia.

Pusey, B. J., A. H. Arthington, and M. G. Read. 1993. Spatial and temporal variation in fish assemblage structure in the Mary River, South-Eastern Queensland – the influence of habitat structure. Environmental Biology of Fishes 37:355–380.

QNRW (Queensland Natural Resources and Water). 2000. Southeast Queensland 25 meter Digital Elevation Model. Queensland Natural Resources and Water, Indooroopilly, QLD, Australia.

R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Theobald, D. M., J. Norman, E. Peterson, and S. Ferraz. 2005. Functional Linkage of Watersheds and Streams (FLoWs): Network-based ArcGIS tools to analyze freshwater ecosystems. ESRI User Conference 2005. Environmental Systems Research Institute, Inc., San Diego, California, USA.

Theobald, D. M., J. B. Norman, E. E. Peterson, S. Ferraz, A. Wade, and M. R. Sherburne. 2006. Functional Linkage of Waterbasins and Streams (FLoWS) v1 User's Guide: ArcGIS tools for Network-based analysis of freshwater ecosystems. Natural Resource Ecology Lab, Colorado State University, Fort Collins, CO. Available at http://www.nrel.colostate.edu/projects/starmap/flows_index.htm

Van Rossum, G., and F. L. Drake, Jr. 2005. *The Python Language Reference Manual*. Network Theory, Ltd., Bristol, UK.

Verbeke, G., and G. Molenberghs. 2000. Linear mixed models for longitudinal data. Springer, New York, New York, USA.

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.

Vogel, R.M., I. Wilson, and C. Daly. 1999. Regional Regression Models of Annual Streamflow for the United States. Journal of Irrigation and Drainage Engineering May/June:148–157.

TABLE B1. Watershed and site-scale explanatory variables considered in the model selection procedure.

Explanatory Variable	Scale	Units	Source
Mean slope	Watershed	%	QNRW 2000
% Conservation	Watershed	%	BRS 2002
% Urban	Watershed	%	BRS 2002
% Mining	Watershed	%	BRS 2002
% Water	Watershed	%	BRS 2002
% Timber production	Watershed	%	BRS 2002
% Agriculture	Watershed	%	BRS 2002
EHMP region	Site	NA	EHMP 2001

TABLE B2. Fixed-effects estimates for the final mixture model (exponential TU/linear-with-sill TD). Parameters were estimated using restricted maximum likelihood (REML).

Effect	**Estimate**	**Std.Error**	**df**	**t-value**	**p-value**
Intercept	0.6669	0.0792	84	8.4209	0
Mean slope	0.0094	0.0061	84	1.5391	0.1275

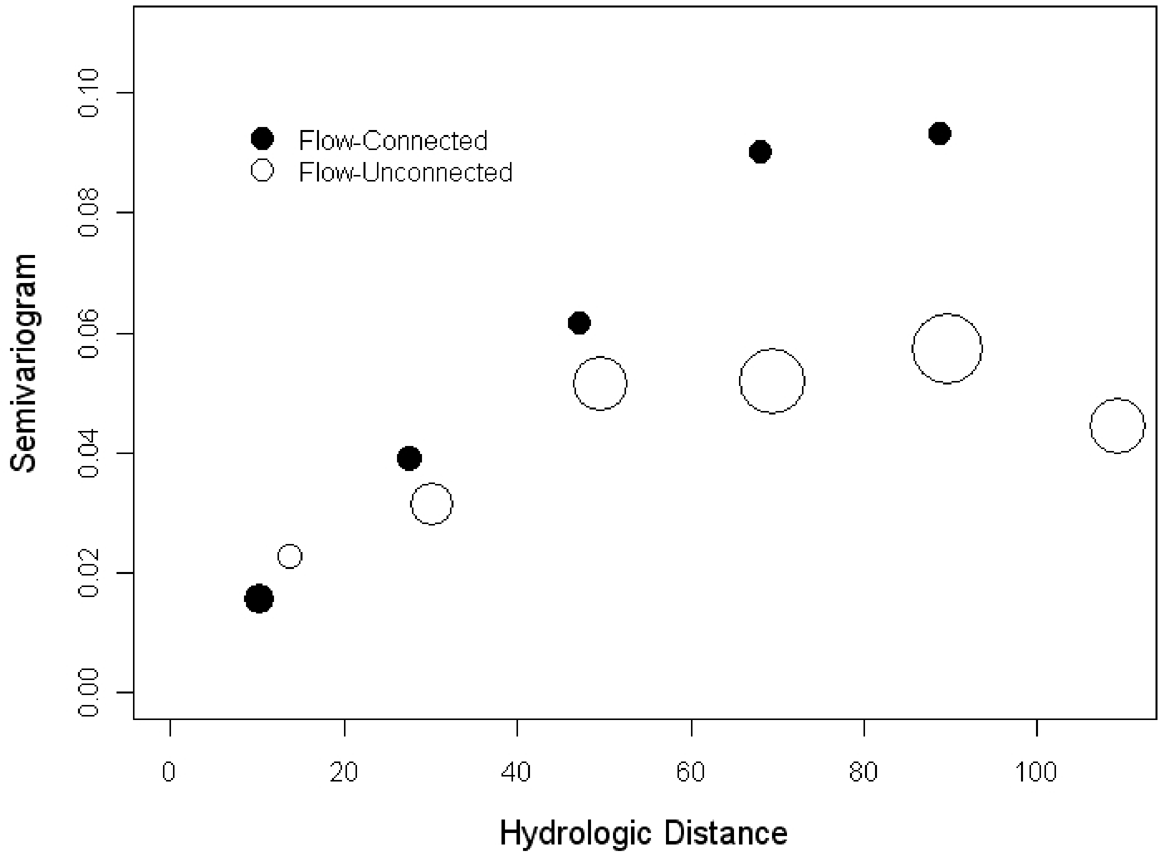


FIG. B1. Semivariogram of the final proportion of native species expected (PONSE) model (exponential TU/linear-with-sill TD) residuals, which includes the hydrologic distance (km) between pairs of sites based on flow-connected (black circles) and flow-unconnected (white circles) relationships. Only lags with > 10 pairs are shown and the size of the circles is proportional to the number of data pairs that are averaged for each value.

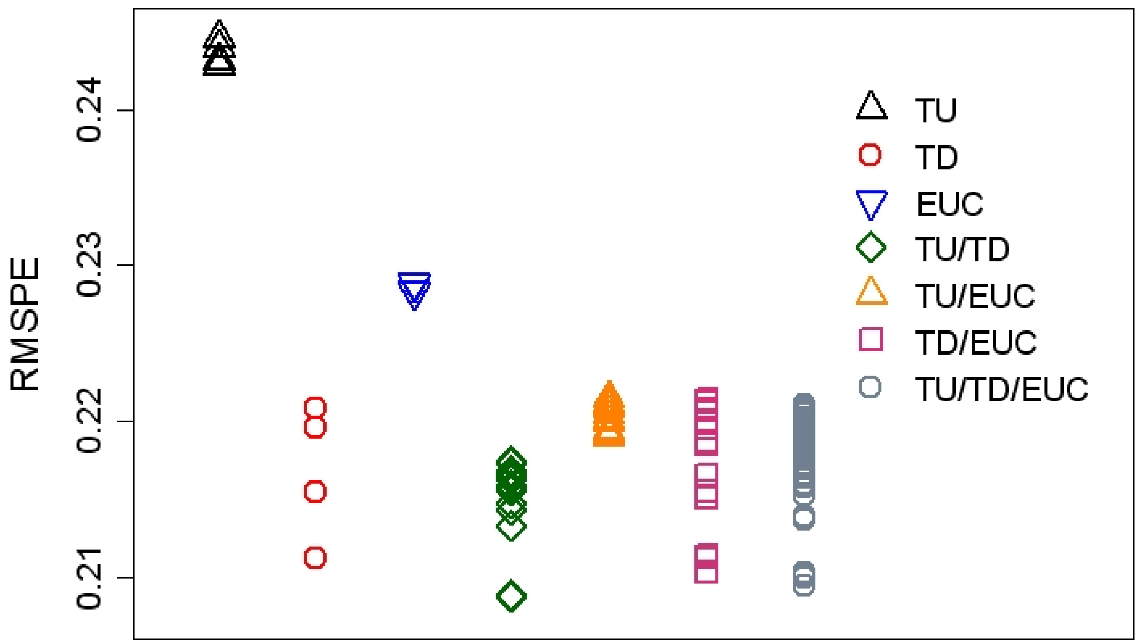


FIG. B2. Root mean square prediction error (RMSPE) values, by mixture type, which were used to select the covariance mixture during the second phase of model selection.