*Journal:* Methods in Eco Evo [6000-7000 words including captions and references!] (Alt: Can J Fish Aquat Sci , Fisheries, Fisheries Research, Trans Am Fish Soc, Freshwater Bio, Ecological Applications, other Eco Modeling type journal)

**A geostatistical state-space model of abundance for stream networks**

Daniel J. Hocking[[1]](#footnote-1), *US Geological Survey, Leetown Science Center, S.O. Conte Anadromous Fish Research Laboratory, One Migratory Way, Turners Falls, MA 01376.*

James T. Thorson*, Fisheries Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA, 98112.*

Kyle O’Neil, *US Geological Survey, Leetown Science Center, S.O. Conte Anadromous Fish Research Laboratory, One Migratory Way, Turners Falls, MA 01376.*

Benjamin H. Letcher, *US Geological Survey, Leetown Science Center, S.O. Conte Anadromous Fish Research Laboratory, One Migratory Way, Turners Falls, MA 01376.*

**Running Title:** Spatio-temporal model for abundance in streams

**Word Count:**

Disclaimer: This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the US Geological Survey (USGS), it does not represent any official finding or policy.

**Abstract**

Fish and wildlife populations will have dynamics that are correlated in space and time due to correlations in environmental drivers as well as synchrony induced by individual dispersal. Many statistical analyses of populations ignore potential autocorrelations and assume that survey methods (distance and time between samples) eliminate these correlations, allowing samples to be treated independently. If these assumptions are incorrect, results and therefore inference may be biased and uncertainty under-estimated. Even if these assumptions are met, researchers may be missing important information about the spatial and temporal structure of populations through the design of their surveys to meet the assumptions of particular statistical methods. We developed a novel statistical method to account for spatiotemporal correlations within dendritic stream networks, while accounting for imperfect detection in the surveys. Through simulations, we found this model outperformed standard statistical methods when data were spatially correlated based on stream distance, and performed similarly when data were not correlated. We found that increasing the number of years surveyed substantially improved the model accuracy when estimating spatial and temporal correlation coefficients. Increased the number of survey sites within the network improved the performance of the non-spatial model but only marginally improved the abundance estimates in the spatiotemporal model. We then applied this model to Brook Trout data from the West Susquehanna Watershed in Pennsylvania collected over 30(???) years. We found … yoy abundance driven by xxxx with high spatial and temporal variability and moderate spatiotemporal correlation (???) … adult abundance less strongly affected by climatic conditions and less variable than YOY but with higher spatiotemporal correlation (???). [add 1-2 good conclusion sentences]

**Keywords:** Spatially-explicit, spatio-temporal, dendritic network, Gaussian random fields, Brook Trout, detection probability

**Introduction**

Ecologists are concerned with understanding the abundance and distribution of organisms in space and time, as well as the biological processes and interactions that cause these patterns. Surveys are frequently employed to estimate spatio-temporal variation in abundance, with the goal of inferring biological process. However, most statistical methods used in ecology have not explicitly accounted for spatial correlation in the data beyond including covariates that are themselves spatially autocorrelated (e.g. ANOVA, GLM, linear and generalized linear mixed models). Therefore to use these regression methods, researchers must design their studies to ensure that sample points are spaced such that statistical residuals are not correlated. It is difficult to know *a priori* how close is too close. Any residual autocorrelation violates regression model assumptions and leads to biased results and potentially incorrect inference regarding population distributions and environmental relationships. Additionally, information about the spatial and temporal patterns provides potentially interesting ecological insights that would not be gained if the data were collected in a way to avoid autocorrelation (give example). For these reasons, a large field of spatial statistics has been developed and applied to ecological problems (refs: Ross et al 2012, Conn et al 2015, Gelfand example, Cressie example).

Streams are likely to have significant correlation in time and space because of the hydrologic connections creating movements and gradients of chemical and physical properties. For example, (find good example). Similarly, organisms living in streams are likely to respond to these underlying conditions and their movements are often restricted to the dendritic network creating spatial correlation in the abundance and distribution of stream organisms. Spatial models that use Euclidean distance are likely to perform poorly in stream networks because streams in close overland proximity can be completely unconnected or have large hydrologic distances (ref: Ver Hoef et al. 2006).

A variety of statistical models have been developed to account for spatial correlations in dendritic networks. These include, but are not limited to, deriving valid covariance relationships for linear models (ref: Peterson et al 2007) and linear mixed models with moving averages that account for hydrologic distance and flow (Ver Hoef et al. 2006). Some models also include “tail-up”, “tail-down” or “two-tail” correlations to account for directional autocorrelation (ref: Ver Hoef and Peterson 2010, Peterson and Ver Hoef 2010). Additionally, block Kriging has been used for spatial interpolation (ref: Ver Hoef et al 2006).

While these models provide improved inference for many types of data, there are limitations with the current approaches. Current models account for spatial correlations but do not allow for changing spatial correlations over time as with spatiotemporal models (ref: Peterson et al. 2013 Ecology Letters). A second limitation is the inability to distinguish between process and observation error to account for imperfect detection (ref: Peterson et al. 2013 Ecology Letters). When performing count surveys of organisms, the probability of detecting each individual in the population is often less than one (imperfect). This results in a problem of inference regarding the populations and environmental effects on the population, particularly when the probability of detection is variable in time and space. To address this issue, a variety of hierarchical models have been developed separating information regarding abundance and detection (refs: royle, kery, dail and madsen). However, these models generally to not account for spatial correlation among sites (although exceptions exist; Royle and Wikle 2005). Those that do generally use random group effects assuming clustered sites to be more similar to each other than to other clusters (refs: Hocking et al., DeWeber). This coarse grouping does not allow for autocorrelation as a function of distance. For example, if sampling is done in a series of transects, all sites within a transect are treated the same (refs: e.g. Peterman et al., Hocking et al. ) even though it is likely that adjacent sites are more correlated than distant sites at the opposite ends of the transect. A final limitation of current spatial stream models is the computational challenges with analyzing large networks due to estimating large covariance structures (ref: Peterson et al. 2013 Ecology Letters).

We describe a novel and generalizable hierarchical model that includes spatiotemporal autocorrelation while accounting for imperfect detection. It also addresses unexplained random variation in abundance (log-normal overdispersion) not explained by deterministic covariates of abundance. We assessed the spatial component of this model with simulations varying the two parameters of the Ornstein-Uhlenbeck (OU) process used to define the spatial relationships in the network. We also performed a simulation study to evaluate the effects of spatial and temporal replication on model performance. We then applied this model to Brook Trout (*Salvelinus fontinalis*) data from the West Susquehanna watershed within Pennsylvania, USA. These data were collected by the Pennsylvania Boat and Fish Commission and are similar to stream fish surveys conducted by state and federal agencies and other researchers throughout the United States. Brook Trout were of particular interest as the only native trout in the eastern U.S. and are threatened by climate and land-use change, overfishing, and exotic species (refs).

**Materials and Methods**

**Overview**

In the following, we assume that data arise from a sampling design where *N* sites are visited in each of *T* years (we use vector-matrix notation throughout). These *N* sites are embedded within a stream network where there is only one unique path from each site to every other site (i.e., the stream network is acyclic), and each sample is conducted by eliminating the possibility of movement out of the sampled area (i.e., by placing nets above and below a selected stream segment) and then repeatedly counting and removing all individuals that are observed. We use the term “triple-pass depletion sampling” for this design, given that there are three removal samples conducted in each sampling occasion. Each removal sample has a lower expected count that the previous (because previous sampling has removed individuals), so this triple-pass design allows the detection probability to be estimated from the slope of this decline among passes.

We then modeled density at time and site (within units numbers per meter, i.e., where distances are measured along a one-dimensional stream reach) as a log-linked linear regression model with components representing the effect of measured habitat variables, as well as otherwise unexplained spatial, spatio-temporal, and independent variation. [Add ecological or sampling justification for each term in Eq. 1, and use opportunity to introduce notation in text prior to equation]. We therefore specify:

(1)

where is a row-vector of measured variables affecting abundance (which includes an intercept term) and is the estimated impact of these variables on log-abundance, is spatial variation, is temporal variation, is spatio-temporal variation, is overdispersion that is independent among sites and years (i.e., the nugget in a geostatistical model), and is the offset for length of stream sampled (length of survey / 100 m) so all abundances are relative to fish per 100 m of stream length.

Finally, we modeled counts assuming that each individual is equally likely to be captured in a given pass. This assumption results in a Poisson distribution for the first pass:

(2a)

where is the probability that each individual present at site and time will be captured. The second and third passes are then dependent upon not being captured in the earlier passes:

(2b)

and

(2c)

This detection probability is estimated simultaneously with parameters representing spatial and spatio-temporal variation in density . Refer to table 1 for summary descriptions of all model parameters.

*Spatio-temporal correlations on a stream network*

Working within a dendritic stream network, Euclidean distances are unlikely to represent the spatial similarity of population dynamics. Therefore, we instead approximate the similarity between two sites (i.e., correlations in spatial variation and spatio-temporal variation ) by the minimum distance between sites along the stream network (hydrologic distance). To do so, we augment the set of sampled sites (termed “sampling nodes”) with a set of “branching sites” (termed “branching nodes”) where two streams join, and note the direction of stream flow at each node. We then identify the “root” of the network as the node that is downstream of all other nodes in the network. We then move upstream from this “root” node, and identify the nearest node along the network (or nearest nodes if the root is a branching node). In this case, we label the root node as the “parent” and the nearest node (or nodes) as “children”. Then starting from these children, we again move upstream to the nearest node or nodes, and again record the parent-child relation between these nodes. This process is continued until we have reached the headwaters (or the highest sampling nodes) in each stream in the network. This description of the network has the important characteristics that each node is the “child” in one and only one “parent-child” relationship.

We also assume that changes in variables along the network are “memory-less”, i.e., the value of a variable defined at a set of points along a stream segment follows a first-order Markov process where sites that are not connected by a child-parent relationship are statistically independent (conditional on the value of at all other sites). This property arises from the assumption that the value of varies while moving along a stream network following a first-order stochastic differential equation.

Given these two properties (that the stream network is acyclic, and that spatial variation is a first-order Markov process), we can calculate the probability distribution for each site as a function of its value at the parent node for that site , and the distance between and its parent . This allows us to factor the joint probability of a spatial variable into a series of easy-to-calculate conditional probabilities:

(3)

We further assume that variation in arises from a mean-reverting Weiner process with movement along the network. This results in an Ornstein-Uhlenbeck process for the right-hand side of Eq. 3, as we now describe in detail.

*Ornstein-Uhlenbeck process for spatial variation*

We used the Ornstein-Uhlenbeck process to represent the spatial relationships along the network. The OU process is Gaussian, Markovian, and mean-reverting so that a child node will be correlated with its parent node as a function of distance following,

(4)

The variance () follows the OU process as

(5)

where is the exponential rate of decay in correlation between child and parent nodes with distance and larger values represent less correlation, and is the variation in the OU process. The expected correlation between points in the network is then represented by where

(6)

Eq. 5-6 are specified such that the pointwise variance of (i.e., the variance was drawn again from the same stochastic process) is .

*First-order autocorrelation for temporal variation*

We include a temporal term for each year to represent years that are better or worse on average across all sites. We model vector (representing in all years) using first-order autocorrelation

(7)

where is the variance in this temporal term, and is the correlation matrix for a first-order autocorrelation process:

(8)

where is the correlation between years and , separated by years, and is an estimated parameter representing the correlation in for two adjacent years.

*Ornstein-Uhlenbeck process for spatio-temporal variation*

We similarly used the OU process to represent the spatio-temporal relationships along the network. We use the vector to represent the spatio-temporal term for all years , and it varies along the network as an OU process:

(9)

where is again the spatio-temporal variance:

(10)

is the correlation due to spatial similarity:

(11)

and is the correlation due to temporal similarity, which we assume follows first-order autocorrelation (eq. Eq. 8, but replacing with , where is an estimated parameter representing the temporal correlation between two adjacent years in spatio-temporal variation ). We assumed that the decorrelation distance was identical for spatial and spatio-temporal variation (i.e., ), but the variance in the OU process was independent for the spatial and spatiotemporal components. This assumption could be relaxed in the future but may require large networks with a large amount of spatially and temporally replicated data to fit, potentially beyond what is available for most studies.

*Independent process for overdispersion*

[Please add definition of and sigmaIID]

**Spatial Simulations**

We conducted simulations to evaluate model performance. The first set of simulations was designed to test the ability to estimate spatial correlations and how well the model estimated abundance with varying levels of spatial autocorrelation compared with a non-spatial model. To accomplish this, we simulated data with all combinations of in {0.1, 0.5, 1, 2, 3} and in {0.1, 0.25, 0.5, 0.75, 1.0}. These values of represent a large range in correlations such that when = 0.1 then (90.5% correlation) whereas when then (5.0% correlation) at points 1 km apart. We also included a single covariate on abundance that differed by location but was not spatially autocorrelated We ran 200 simulations for each combination of and and fit each simulated dataset with the spatial model described (single year with no temporal or spatiotemporal variation) and with a non-spatial model. Both models accounted for imperfect detection, observed following a 3-pass depletion survey and were identical except for the inclusion of the spatial component of the model. The probability of detecting an individual that remained in the stream on any given pass was 50%. We ran the simulation using the White River watershed in Vermont with 359 nodes. Distances between child and parent nodes ranged from 0.17 km to 5.13 km with a mean of 1.13 km. The R code for simulating the data can be found at <https://github.com/djhocking/Trout_GRF/blob/master/Code/Spatial_Simulations.R>

**Spatiotemporal Power Analysis**

We also wanted to understand the effect of spatial and temporal replication on model performance. We simulated 200 independent data sets for the White River in Vermont over 20 years for each of the 359 nodes. For each simulation, we randomly sampled the data to represent surveying various numbers of sites and years (all combinations of 4, 8, 10, 15, and 20 years with 25, 50, 100, and 359 sites). For each survey combination and simulation, we fit the spatiotemporal model including spatial, temporal, and spatiotemporal dynamics (matching the data generating model) and a temporal model with no spatial or spatiotemporal dynamics. For each simulation, we used , , , , , , detection probability , site-level covariate on abundance , and mean abundance = 10.

**Brook Trout Case Study**

As the only trout native to eastern U.S. streams and rivers, Brook Trout are a species of social and economic importance in the region. State and federal agencies as well as organizations such as Trout Unlimited and the Eastern Brook Trout Joint Venture (EBTJV) have particular interest in supporting viable populations of Brook Trout. As such, there have been numerous recent modeling efforts to estimate occupancy, abundance, and population dynamics in response to landscape conditions, climate change, and management actions (refs: Wagner, Letcher, Kanno). However, beyond using random regional, watershed, or sub-basin effects, these models generally do not account for spatial correlations.

We identified the West Susquehanna watershed for our case study because it was a moderately-large network with a high density of good quality stream fish data over a long time period. The electrofishing data were collected by the state of Pennsylvania Boat and Fish Commission using standard methods common across agencies and researchers throughout the eastern U.S. We did not use the West Susquehanna watershed in our simulations because it is much larger than the White River network, with many more confluences, which would greatly slow the data simulation. Doing hundreds of simulations on such a large network under many conditions would have been prohibitively slow.

The West Susquehanna watershed contained 11,220 nodes, comprised of 349 survey sites and 10,871 stream reaches. Sites were survey in a total of 34 from 1981 and 2014. There were a total of 683 site visits with a mean of 2.0 and a range of 1 to 21 visits per site. The total drainage area of the watershed was 18,068 km2 and the smallest stream had a cumulative drainage area of 0.4 km2. The median drainage area was 4.4 km2. The mean distance between nodes in the network was 1.373 km and ranged from 0.001 to 11.613 km with a median of 1.114 km.

The watershed was primarily forest (mean percent forest cover = 79%) but with a range from 0 to 100% within individual stream catchments. We used percent forest cover as a fixed effect covariate in our model along with surficial coarseness, mean air temperatures from the summer (previous year), fall (previous year), winter, and spring prior to summer fish surveys, and mean daily precipitation for the same seasons. Daily temperature and precipitation data were obtained from daymet (refs) and spatially aggregated to the catchment scale. The surficial coarseness is the percentage of the catchment area covered by a parent soil material with texture described as sandy, gravelly, or a combination of the two. These classifications were obtained from the USDA National Resources Conservation Sciences Soil Survey Geographic Database (SSURGO; ref). Forest cover data was obtained from the 2011 National Land Cover Database (NLCD; ref). All basin characteristics were calculated as spatial sums (precipitation) or means within each zonal catchment layer as delineated based on the truncated NHDHRDV2 flowlines. All details and ArcPython scripts can be found at <http://conte-ecology.github.io/shedsGisData/>. The covariate summary statistics for the West Susquehanna watershed are presented in table ##.

We used the National Hydrography Dataset high resolution flowlines (ref) truncated to >0.75 km2 drainage area for spatial consistency and exclusion of highly ephemeral streams (ref: kyle github pages – he should really publish it as a data paper at the least). Any survey locations or other points of interest were then snapped to the flowlines. All survey points and confluences, including the base of the network and the terminal headwaters, were considered network nodes. Except for the base node, the distance from each child node was calculated to its downstream parent node to define the network relationships and distances. All hydrography processing was done using ArcPython (ref). The full description of the process, scripts, and links to the hydrography data is archived at (ref: gihub repo link). The hydrography for the region from Maine to Virginia, USA can be downloaded by hydrologic unit code 2 at <http://ecosheds.org/assets/nhdhrd/v2/>

*Model Selection*

For young of the year (YOY) and adult Brook Trout independently, we compared eight models with different combinations of spatial, temporal, and spatiotemporal correlations (Table 3). All other components of the model including fixed effect covariates were identical in all models. Meteorological conditions during the previous summer were used in the adult models but were excluded in the YOY models because spawning does not occur until the fall. We used Akaike’s Information Criterion (AIC) to select the best model balancing model fit and model complexity (refs).

**Results**

**Spatial Simulations**

We found that the spatial model did a good job estimating the spatial correlation decay rate () when there was strong spatial correlation but tended to slightly underestimate the decay rate when the correlation was low ( large; Figure 2a). Similarly, the variability in the estimation of was larger as the value of increased (Figure 2a). The spatial model estimated the mean abundance across the watershed much better than the non-spatial model (Figure 2b) and the mean uncertainty of the overall abundance was relatively similar among models, but the variation in this uncertainty was much larger for the non-spatial model compared with the spatial model when the spatial correlation was large (Figure 2c). The root mean squared error (RMSE) was far larger for the non-spatial model compared with the spatial model across all values of (Figure 2d), indicating that, although the average abundance across the watershed was estimated well, the individual location abundance estimates were much more accurate for the spatial model. This difference in uncertainty was largest with high levels of spatial correlation. The fixed effect coefficient for the single covariate () was estimated well across all values of , but the variation in this estimate was slightly smaller for the spatial model, especially at higher levels of spatial correlation (Figure 2e).

The range of also significantly influenced the parameter estimates and the differences between spatial and non-spatial models. At all levels, was recovered well using the spatial model with very slight underestimation on average (Figure 3a). The spatial and non-spatial models performed similarly in the estimation of mean abundance across the watershed when the true value of was small but the spatial model was more accurate and more precise compared with the non-spatial model as the level of increased (Figure 3b). The uncertainty in mean network abundance went up for the non-spatial model as increased but was constant for the spatial model across levels of (Figure 3c). The RMSE was again much smaller for the spatial model compared with the non-spatial model as increased. The variability in the RMSE also increased greatly for the non-spatial model as increased (Figure 3d). The fixed effect coefficient was estimated well for both models but the uncertainty increased in the non-spatial model as increased (Figure 3e).

**Spatiotemporal Power Analysis**

We found the mean network abundance was estimated fairly well for both the spatial and non-spatial models, but the spatial model tended to underestimate abundance slightly when few years were surveyed (Figure 4). However, the RMSE was lower for the spatial model compared with the non-spatial model and the difference increased with the number of years surveyed. The value of was recovered well regardless of the number of years sampled, but the variability in the accuracy of estimating decreased (improved) with the number of years sampled (Figure 4) and even more dramatically with the number of sites sampled (Figure 5). The variance in the spatial process ( was underestimated and the spatiotemporal variance () overestimated when sites were sampled for 10 or fewer years (Figure 4). Similarly, for both the spatial and non-spatial model it took 10-15 years to accurately recover the temporal autocorrelation, although the variability in the temporal process was recovered with approximately 8 years of sampling in the non-spatial model and possibly as little as 4 years with the spatial model (Figure 5). The value of the fixed effect covariate, , was estimated well for both models regardless of the number of years sites were sampled but the variation in the estimation was consistently lower for the spatiotemporal model (Figure 4).

The number of sites sampled similarly influenced the estimation of the spatial and spatiotemporal variance terms, with an increasing number of sites improving the spatial variance estimation and to a lesser extent, the spatiotemporal variance (Figure 5). The RMSE of the spatial model improved with an increasing number of sites sampled but it did not significantly improve for the non-spatial model (Figure 5). The fixed effect coefficient was recovered well for both models and the precision improved with the number of sites sampled.

**Brook Trout Case Study**

The top YOY model included temporal and spatiotemporal components. The null model was the worst and any model with a spatial or spatiotemporal component performed better than the temporal-only model (Table ##). For adult Brook Trout, the spatiotemporal model and the temporal plus spatiotemporal model were the top two models with a delta AIC of only 0.3 (Table ##). We chose to draw inference from the temporal plus spatiotemporal model for the easiest direct comparison with the YOY. As with the YOY models, the null and temporal-only models performed the worst, although the two models that included both spatial and spatiotemporal components did not converge and were excluded from model selection (Table ##).

Coefficient Tables

The estimated values of theta\_st for both YOY and adults were at the lower end of what we tested with simulations. Similarly, the values of sigma\_st were high for YOY (0.65) and adults (0.59). The combination of the two parameters indicate extremely high spatiotemporal autocorrelation, which is revealed by the estimate of rho\_st of 0.98 and 0.97 for YOY and adults, respectively (Table ##). Adults also exhibited strong temporal autocorrelation (rhot = 0.59) with low variability (sigmat = 0.16), whereas YOY exhibited no temporal autocorrelation (rhot = -0.05) but high stochastic temporal variability (sigmat = 0.76). Forest cover, the previous year’s mean summer temperature, spring temperature, and to a lesser extent the previous fall’s mean temperature were all significant predictors of adult abundance. For YOY, only forest cover and mean spring temperature had significant effects on abundance. Precipitation in any season did not have a significant effect on abundance for YOY or adults (Table ##).

Other Results?

Semivariograms/Torgegrams

* Of modeled abundance
* Of residuals

Figure: Panel of maps of abundance over time to show dynamics in space (YOY and Adults)

Figure: Line plot of abundance over time to show dynamics over time (YOY and Adults)

**Discussion**

Temporal interpretation – adds absolute number of fish each year to each site rather than relative change

The fact that the non-spatial model was just as good as the spatial model at recovering the mean network abundance but had higher (worse) RMSE indicates that the non-spatial model does okay on average but is not as good at estimating abundance at individual locations. This pattern of over homogenization of abundance estimates from the non-spatial model can be observed in figure 6. Many management actions occur at small scales and therefore understanding local population dynamics is important for prioritizing local actions and understanding the effects of those actions, particularly in an adaptive management framework. Such a situation could occur for decisions that are repeated and adjusted based on population responses such as stocking programs or setting stream-level fishing regulations (e.g. barbless hooks, catch-and-release, take limits). Even for one-time decisions such as in-stream habitat modification, and dam or culvert removal at a local site, it is important to have good estimates of local, rather than just watershed, abundance because the local change in abundance can help prioritize the location of the next project.

Reasons why precip not significant

**Acknowledgements**

Pennsylvania Boat and Fish Commission. E. Childress for discussions related to the use of data and inference from the models.

**Data Accessibility**

To enable readers to locate archived data, authors should list the database and the respective accession numbers or DOIs for all data from the manuscript that has been made publicly available in this section. E.g:  
- Species descriptions: uploaded as online supporting information- Phylogenetic data: TreeBASE Study accession no. Sxxxx- R scripts: uploaded as online supporting information- Sample locations, IMa2 input files and microsatellite data: DRYAD entry doi: xx.xxxx/dryad.xxxx

**References**

**Tables**

**tables and figures**

Supporting Information and Appendices

1. Corresponding Author:Frostburg State University, Department of Biology, 201 Compton Science Center, 101 Braddock Road, Frostburg, MD 21532. Email - [*djhocking@frostburg.edu*](mailto:djhocking@frostburg.edu)*.. Phone: 301-687-4343.* [↑](#footnote-ref-1)