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

Daniel J. Hocking, *US Geological Survey, S.O. Conte Anadromous Fish Research Center, One Migratory Way, Turners Falls, MA 01376.* [*dhocking@usgs.gov*](mailto:dhocking@usgs.gov)

James Thorson

Ben Letcher, *US Geological Survey, S.O. Conte Anadromous Fish Research Center, One Migratory Way, Turners Falls, MA 01376*

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

Kasper Kristensen or others – for help with OU and TMB?

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**

**Introduction**

Ecologists are concerned with understanding the abundance and distribution of organisms in space and time and the conditions that influence these patterns. Surveys are frequently employed to study these patterns. However, most statistical methods used in ecology have not explicitly accounted for spatial correlation in the data beyond deterministic covariate parameters (e.g. ANOVA, GLM, linear and generalized linear mixed models). Therefore to use these analyses, researchers must design their studies to ensure that sample points are not overly close in time or space such that unaccounted for autocorrelation could occur. 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 proximity can be separated by ridges and be completely unconnected or have large hydrologic distances.

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 with little to no chance of double counting individuals. 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. 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.). 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 accounts for spatial, temporal, and spatiotemporal autocorrelation while accounting for imperfect detection. It also addresses unexplained random variation in abundance (log-normal overdispersion).

We include R and TMB code that allows for rapid, efficient estimation of these models.

We assessed the spatial component of this model with

1. Spatial Simulations

We also performed a small simulation study to evaluate the effects of spatial and temporal replication on model performance.

2. Spatiotemporal Power Analysis Simulations

We then applied this model to Brook Trout

3. West Susquehanna Case Study (Adults, YOY, Stock-Recruit, Recruit)

* commonly collected by state agencies throughout the US
* Declining and threatened by climate and landuse change, overfishing, exotic species
* Likely to be spatially correlated

**Methods**

*Describe model in general*

Describe how to structure a network (parent-child, distances) – include figure

Random Fields

SPDE – Ornstein-Uhlenbeck Process

Spatial

where

is the exponential rate of decay in correlation between child and parent with distance

is the distance between the child and parent nodes

is the variation in the process

The correlation between points in the network is then

Temporal

Spatiotemporal

Detection Process

*Testing and Use:*

1. 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 {0.5, 1, 2, 3, 4, 5} and in {0.2, 0.3, 0.4, 0.5}. These values of represent a large range in correlations such that when =0.5 then (60.7% correlation) whereas when then at points 1 km apart. We also included a single covariate on abundance that differed randomly by location but was not spatially autocorrelated We ran 200 simulations for each combination of and and estimated 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 R code for simulating the data can be found at <https://github.com/djhocking/Trout_GRF/blob/master/Code/Spatial_Simulations.R>

2. Spatiotemporal Power Analysis

I ran 200 simulations of the spatiotemporal model for the White River in Vermont with 359 nodes (~330 stream reaches) over 20 years. For each iteration I then sampled the data to simulate surveying various numbers of sites for differing numbers of years (all combinations of 4, 8, 10, 15, 20 years for 25, 50, 100, 359 sites). For each survey combination in each iteration I fit the model including either spatial, temporal, and spatiotemporal dynamics (attempting to match the data generating model) or just a temporal model with no spatial or spatiotemporal dynamics.

3. West Susquehanna Case Study (Adults, YOY, Stock-Recruit, Recruit)

We chose 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.

**Results**

1. 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 consistently underestimated 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 and non-spatial models both estimated the mean abundance across the watershed well (Figure 2b) and the mean uncertainty of the overall abundance was similar among models, but the variation in this uncertainty was 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 much 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. The fixed effect coefficient for the single covariate () was estimated well across all values of , but the variation in this estimate was much smaller for the spatial model, especially at higher levels of spatial correlation (Figure 2e).

The range of had less effect on the estimates and on 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 and in the confidence of those estimates. The uncertainty went up for both models as this random variability increased (Figure 3b,c). The RMSE was again much smaller for the spatial model compared with the non-spatial model. The variability in the RMSE increased greatly for the non-spatial model as increased but it only increased slightly for the spatial model (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).

Figure 2. (low res)

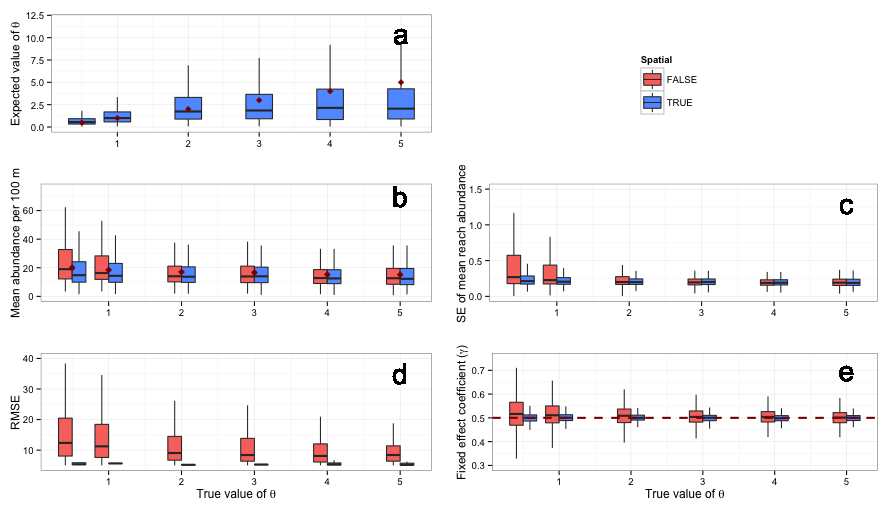
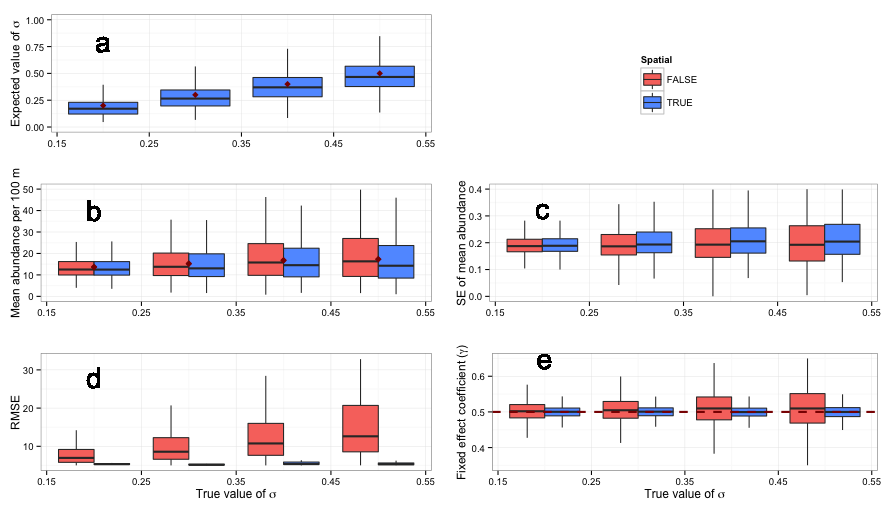
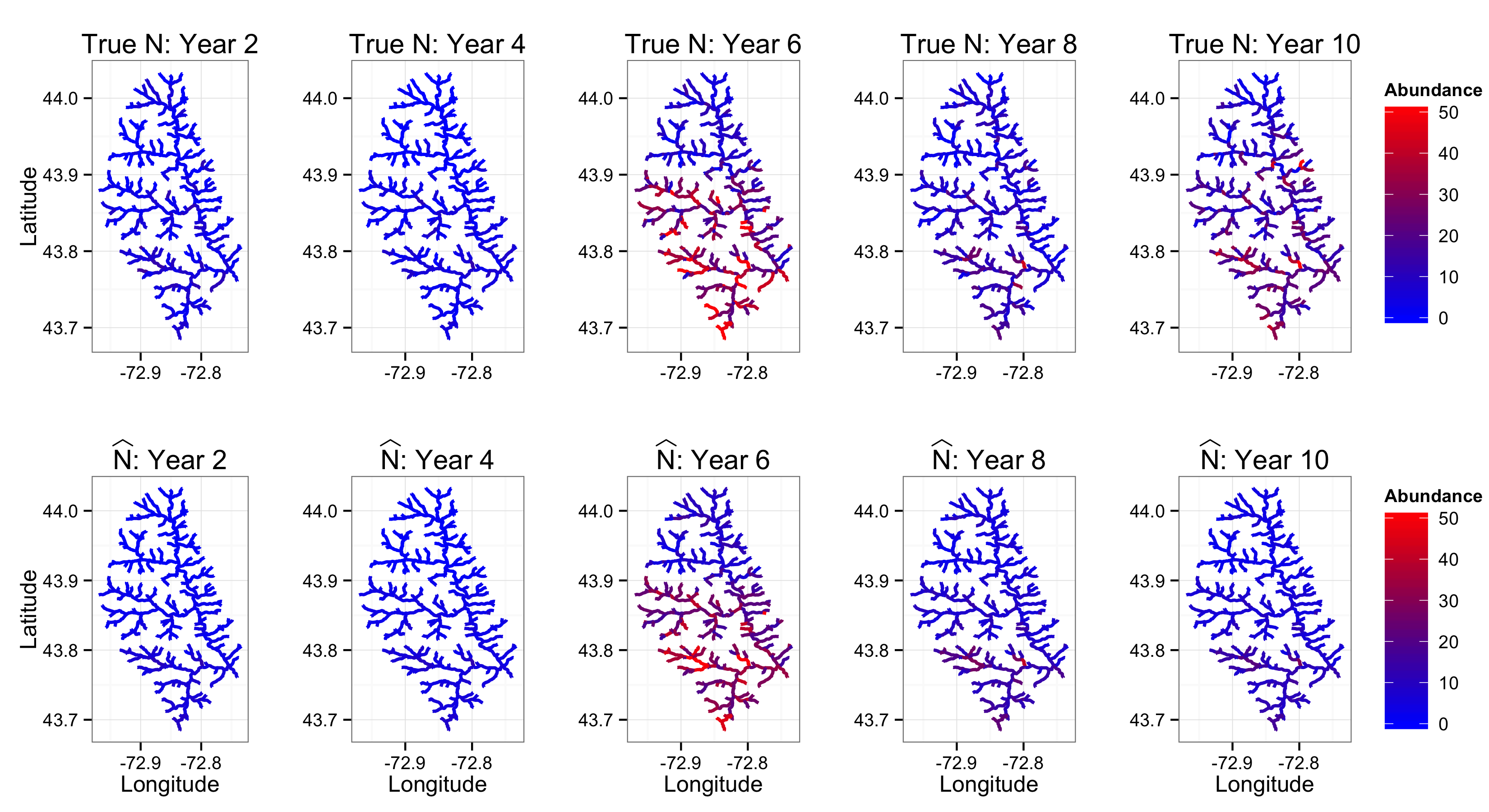


Figure 3. (low res)



2. Spatiotemporal Power Analysis

Figure XX (low res). Example of a spatiotemporal simulation of the abundance along a stream network over time. The top row shows the true (simulated) abundances and the bottom row depicts the mean expected abundance based on the model. The simulation used values of theta, sigma, theta\_st, sigma\_st, AR1, sampled at N\_sites = 50 in N\_years = 8.



3. West Susquehanna Case Study (Adults, YOY, Stock-Recruit, Recruit)

Semivariograms and Torgegrams

* Of modeled abundance
* Of residuals

**Discussion**

**Acknowledgments**

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

**Literature Cited**