*Journal:* Can J Fish Aquat Sci (Alt: Fisheries, Fisheries Research, Trans Am Fish Soc, Freshwater Bio, Ecological Applications, Methods in Eco Evo, other Eco Modeling type journal)

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

Daniel J. Hocking, *US Geological Survey, Leetown Science Center, S.O. Conte Anadromous Fish Research Laboratory, One Migratory Way, Turners Falls, MA 01376.* [*djhocking@frostburg.edu*](mailto:djhocking@frostburg.edu)

James T. Thorson

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.*

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**

Environmental conditions are often correlated in time and space, resulting in correlated fish and wildlife populations that depend on these conditions. Many statistical analyses of populations ignore potential autocorrelations and assume that survey methods (distance and time between samples) sufficiently reduce these correlations. 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 and performed similarly when data were not correlated. We found that increasing the number of years surveyed significantly improved the model accuracy and ability to recover 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]

**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 overland proximity can 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 (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. 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 is 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).

**Methods**

Abundance

- expected abundance at location i in year t

)

- overall mean abundance over space and time

– variance for the Poisson lognormal overdispersion

- fixed effect coefficients (could use capitals to mean vector coef and matrix data)

- spatial component

- temporal component (AR1)

spatiotemporal component

- offset for length of stream sampled (length of survey / 100 m) so all abundances are relative to fish per 100 m of stream length.

*Random Fields*

A random field is a stochastic process with spatial values that vary continuously over a set parameter space. Gaussian random fields follow a multivariate normal distribution with a spatially-structured covariance matrix. This multidimensional analog to the random variables frequently employed in hierarchical modeling of fish and wildlife populations (ref: royle, etc.) makes it a natural addition to these state-space models. However, dense covariance matrices make this type of geostatistical modeling computationally challenging (refs: Blangiardo book, Banerjee et al 2004). A recent solution to this is to represent the continuous Gaussian random field as a discretely indexed Gaussian Markov Random Field (GMRF) where the spatial dependency is defined based on the neighboring points and represented a stochastic partial differential equation (SPDE; ref: Lindgre et al 2011, Blangiardo book). In most two dimensional landscapes a Matern covariance function with continuous Euclidean distances between points (refs).

*Network Spatial Relationships*

Working within a dendritic stream network, Euclidean distances were unlikely to represent the spatial relationships of populations. Therefore, we had to define the network structure and hydrologic distances between locations. 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/>

*SPDE – Ornstein-Uhlenbeck Process*

We used the Ornstein-Uhlenbeck process to represent the spatial relationships with a SPDE 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,

The variance () follows the OU process as

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

Temporal

Spatiotemporal

We assumed that the spatial and spatiotemporal processes acted on the same spatial scales such that

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.

Detection Process

*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 ##.

*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. 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 modeling balancing model fit and model complexity (refs).

*Make into a table:*

1- Null -

2- Spatial -

3 - Temporal -

4 - s + t -

5 - spatiotemporal -

6 - s + st -

7 - t+ st -

8 - s + t + st -

**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 tended to 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 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 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).

2. 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.

3. West Susquehanna Case Study (Adults, YOY)

Model Selection

*YOY*

*M\_num Model AIC delta\_AIC*

*1 5 Temporal + ST 9592 0*

*2 6 S+T+ST 9596 4*

*3 4 Spatiotemporal 9663 71*

*4 8 Spatial + ST 9666 74*

*5 7 Spatial + Temporal 9739 147*

*6 3 Spatial 9801 209*

*7 2 Temporal 9925 333*

*8 1 Obs 10048 456*

*Adult*

*M\_num Model AIC delta\_AIC*

*1 4 Spatiotemporal 9408 0.0*

*2 5 Temporal + ST 9408 0.3*

*3 7 Spatial + Temporal 9583 174.9*

*4 3 Spatial 9588 180.2*

*5 2 Temporal 9783 375.3*

*6 1 Obs 9794 386.5*

Coefficient Tables

*YOY*

*Adult*

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. 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.

**Acknowledgments**

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

**Literature Cited**

**Tables**

Table ##. Covariate Summary for W. Susquehanna by Stream Reach

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Mean | Min | Max |
| Percent forest cover (%) | 79.15 | 0 | 100 |
| Percent surficial coarseness (AB?) | 6.62 | 0 | 100 |
| Previous summer mean temperature (C) | 17.74 | 15.21 | 21.66 |
| Previous fall mean temperature (C) | 3.49 | -0.09 | 7.3 |
| Winter mean temperature (C) | -1.77 | -7.99 | 2.87 |
| Spring mean temperature (C) | 14.63 | 10.31 | 17.31 |
| Previous summer mean precipitation (mm) | 3.78 | 1.59 | 8.92 |
| Previous fall mean precipitation (mm) | 2.99 | 1.29 | 5.01 |
| Winter mean precipitation (mm) | 2.58 | 1.1 | 4.73 |
| Spring mean precipitation (mm) | 2.91 | 1.42 | 6.9 |

**Figures**

Figure 2. (low res)

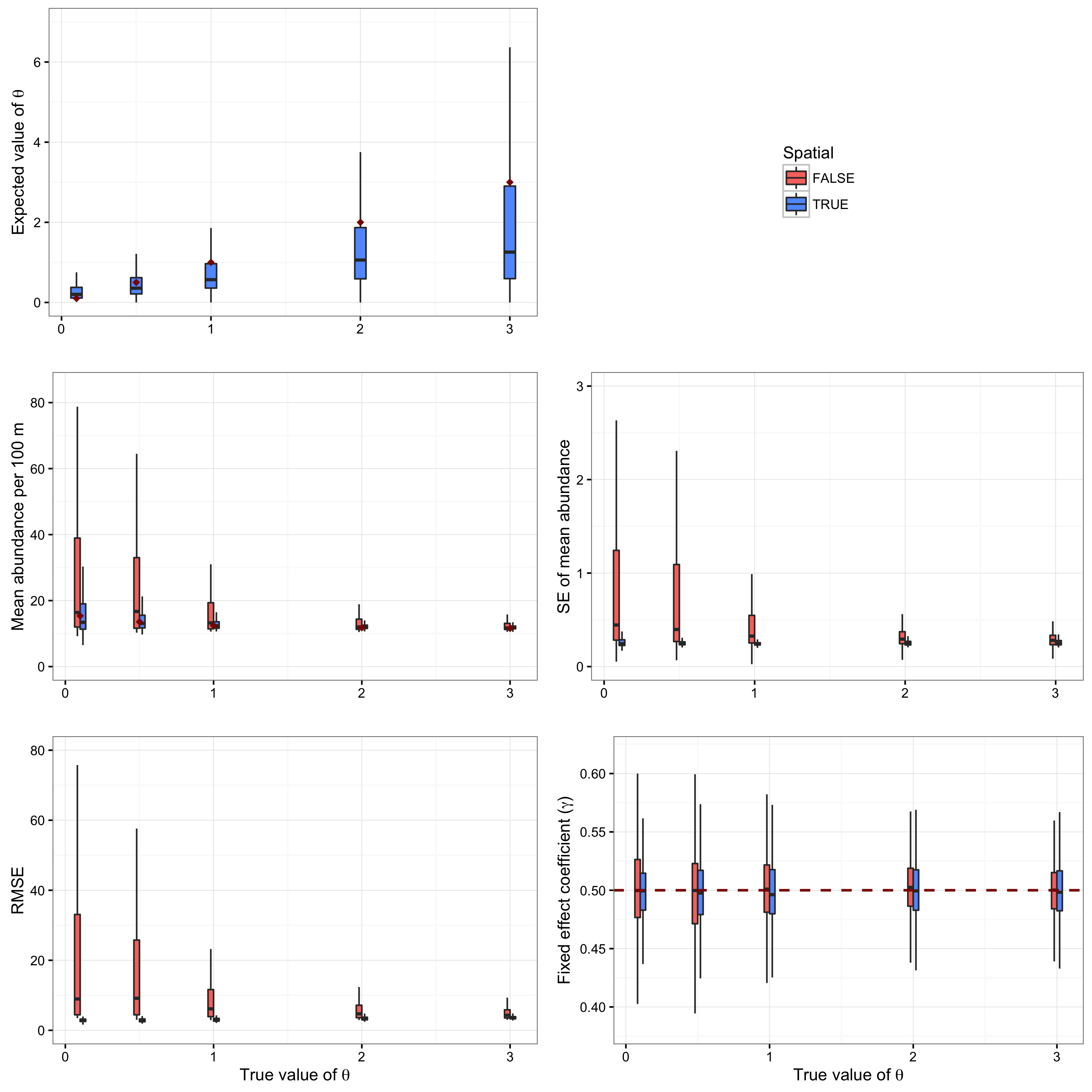


Figure 3. (low res)

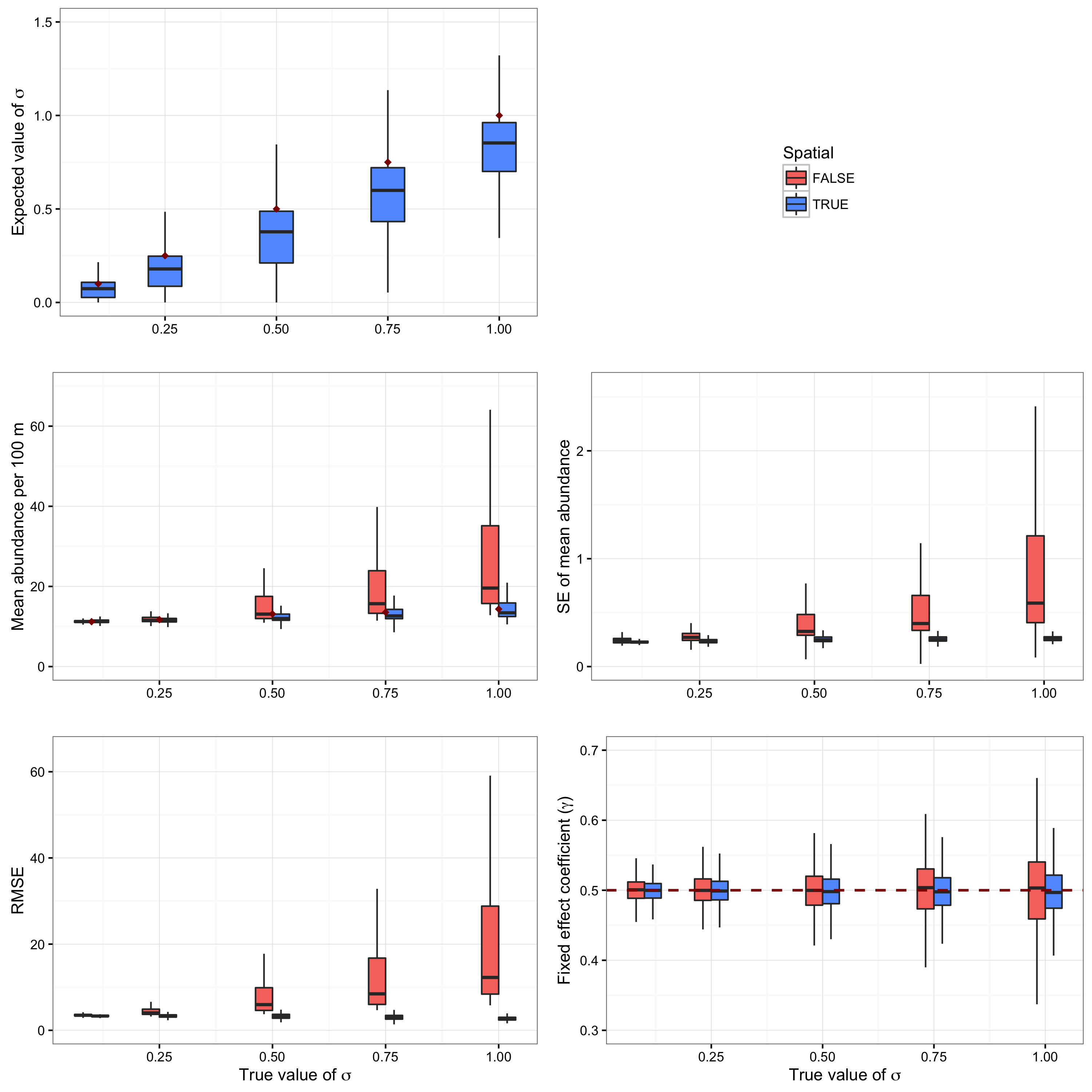


Figure: Effect of years surveyed

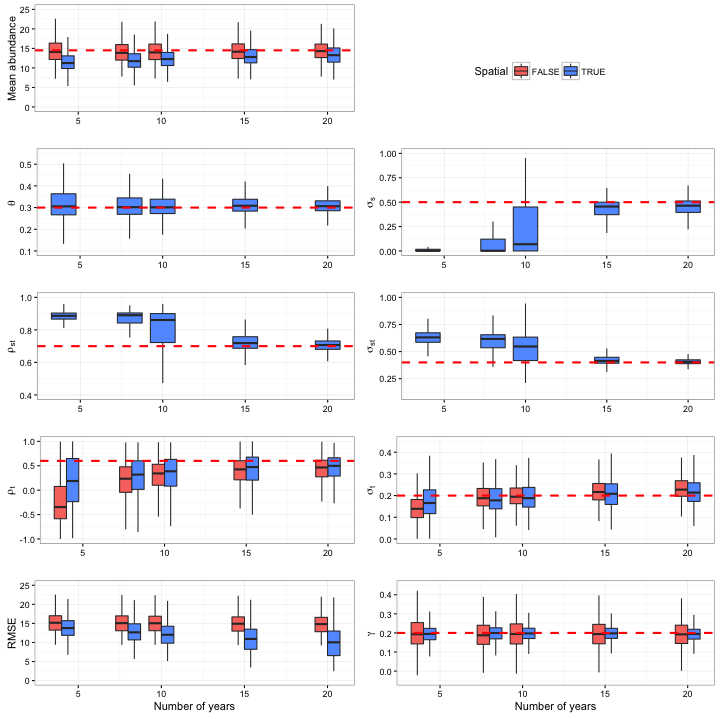


Figure: Effect of the number of sites surveyed (across all levels of years surveyed)

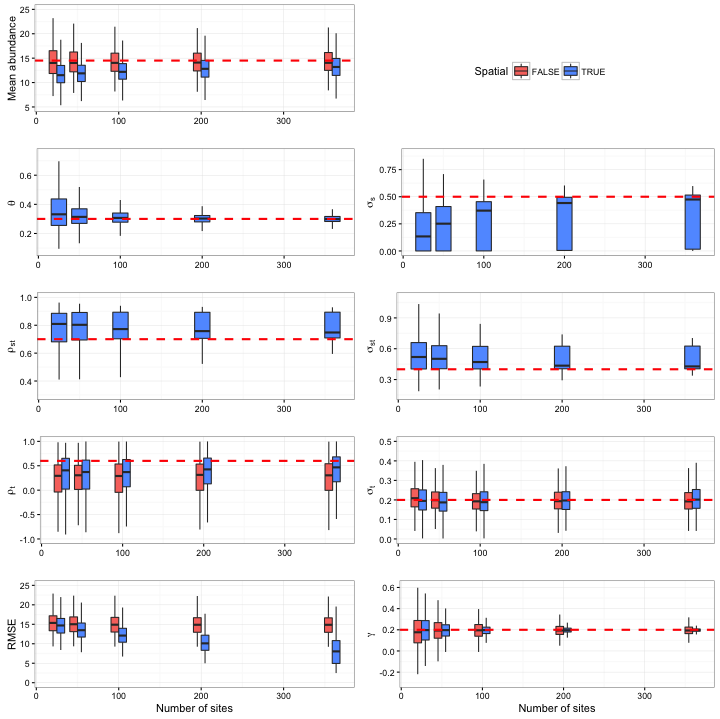


Figure XX (low res). Example of a spatiotemporal simulation of the abundance along a stream network over time. The top row shows the true (simulated) abundances and the middle row depicts the mean expected abundance based on the matching spatiotemporal model. The bottom row shows the mean expected abundance for a model with temporal autocorrelation but no spatial and spatiotemporal correlations. The simulation used values of theta, sigma, theta\_st, sigma\_st, AR1, sampled at N\_sites = 50 in N\_years = 8.

