

SSN2: The next generation of spatial stream network modeling in R

Michael Dumelle¹, Erin E. Peterson², Jay M. Ver Hoef³, Alan Pearce⁴, and Daniel J. Isaak⁵

¹ Office of Research and Development, United States Environmental Protection Agency (USEPA) ² EP Consulting and Queensland University of Technology, Centre for Data Science Brisbane Australia 4000 ³ NMFS Alaska Fisheries Science Center, United States National Oceanic and Atmospheric Administration (NOAA) ⁴ NIASRA, School of Mathematics and Applied Statistics, University of Wollongong ⁵ Rocky Mountain Research Station, United States Forest Service (USFS)

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [↗](#)

Submitted: 18 January 2024

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

The SSN2 R package provides tools for spatial statistical modeling, parameter estimation, and prediction on stream (river) networks. SSN2 is the successor to the SSN R package (Jay M. Ver Hoef, Peterson, Clifford, & Shah, 2014), which was archived alongside broader changes in the R-spatial ecosystem (Nowosad, 2023) that included 1) the retirement of rgdal (R. Bivand, Keitt, & Rowlingson, 2021), rgeos (R. Bivand & Rundel, 2020), and maptools (R. Bivand & Lewin-Koh, 2021) and 2) the lack of active development of sp (R. S. Bivand, Pebesma, & Gomez-Rubio, 2013). SSN2 maintains compatibility with the input data file structures used by SSN but leverages modern R-spatial tools like sf (E. Pebesma, 2018) and provides many useful features that were not available in SSN, including new modeling and helper functions, updated fitting algorithms, and simplified syntax consistent with other R generic functions.

Statement of Need

Streams provide vital aquatic services that sustain wildlife, provide drinking and irrigation water, and support recreational and cultural activities. Data are often collected at various locations on a stream network and used to characterize spatial patterns in stream phenomena. For example, a manager may need to know how the amount of a hazardous chemical changes throughout a stream network to inform mitigation efforts. Comprehensive formulations of spatial stream network (SSN) models are provided by Jay M. Ver Hoef & Peterson (2010), Peterson & Ver Hoef (2010), and Jay M. Ver Hoef et al. (2014).

SSN models use a spatial statistical modeling framework (Cressie, 1993) to describe unique and complex dependencies on a stream network resulting from a branching network structure, directional water flow, and differences in flow volume. SSN models relate a continuous or discrete response variable to one or more explanatory variables, a spatially independent error term (i.e., nugget), and up to three spatially dependent error terms: tail-up errors, tail-down errors, and Euclidean errors. Tail-up errors restrict spatial dependence to flow-connected sites (i.e., water flows from an upstream to a downstream site) and incorporate spatial weights through an additive function to describe the branching network between sites. Tail-down errors describe spatial dependence between both flow-connected and flow-unconnected (i.e., sites that share a common downstream junction but not flow) sites, but spatial weights are not required. Euclidean errors describe spatial dependence between sites based on straight-line distance and are governed by factors not confined to the stream network like regional geology. The length-scales of spatial dependence in the tail-up, tail-down, and Euclidean errors are controlled

by separate range parameters. In this paper, we show how to use the SSN2 R package to fit and inspect SSN models and make predictions at unobserved locations on a stream network.

Package Overview

Before fitting SSN models using SSN2, stream network and observation data sets must be pre-processed either by using the STARS toolset for ArcGIS Desktop versions 9.3x - 10.8x (Peterson & Ver Hoef, 2014) or by using the openSTARS R package (Kattwinkel, Szöcs, Peterson, & Schäfer, 2020), which leverages open-source GRASS GIS. Pre-processing using STARS or openSTARS ends with the creation of a .ssn folder, which is non-proprietary and contains all the spatial, topological, and attribute information needed to fit models to data on a stream network using SSN2. Relevant files residing in the .ssn folder are read into R (using `ssn_import()`) and placed into a special list called an SSN object. The SSN object contains geometry and topological information about the stream reaches and sites, as well as observed data and data for prediction at unsampled sites.

We first load SSN2 into our current R session:

```
library(SSN2)
```

The SSN2 packages comes with an example .ssn folder called MiddleFork04.ssn that represents water temperatures recorded from a stream network in the Middle Fork of the Salmon River in Idaho, USA during 2004. We may store the file path to this example data:

```
path <- system.file("lsndata/MiddleFork04.ssn", package = "SSN2")
```

Several functions in SSN2 for reading and writing data (which we use shortly) directly manipulate the .ssn folder. If it is not desirable to directly manipulate the MiddleFork04.ssn data installed alongside SSN2, MiddleFork04.ssn may be copied it into a temporary directory and the relevant path to this alternative location can be stored:

```
copy_lsn_to_temp()  
path <- paste0(tempdir(), "/MiddleFork04.ssn")
```

After specifying path (using `system.file()` or `copy_lsn_to_temp()`), we import the stream reaches, observed sites, and prediction sites (`pred1km`):

```
mf04p <- ssn_import(path, predpts = "pred1km")
```

We visualize the stream network, observed sites, and prediction sites (Figure 1) using `ggplot2` (Wickham, 2016):

```
library(ggplot2)  
ggplot() +  
  geom_sf(data = mf04p$edges) +  
  geom_sf(data = mf04p$preds$pred1km, pch = 17, color = "blue") +  
  geom_sf(data = mf04p$obs, color = "brown", size = 2) +  
  theme_bw()
```

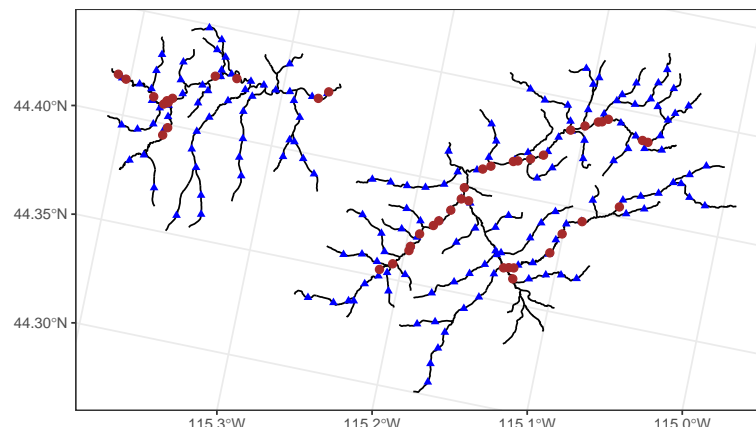


Figure 1: Middle Fork 2004 stream networks. Observed sites are represented by brown, closed circles. Prediction sites are represented by blue, closed triangles.

We supplement the `.ssn` object with hydrologic distance matrices that preserve directionality, which are required for statistical modeling:

```
ssn_create_distmat(mf04p, predpts = "pred1km", overwrite = TRUE)
```

Next, summer mean stream temperature (`Summer_mn`) is modeled as a function of elevation (`ELEV_DEM`) and watershed-averaged precipitation (`AREAWTMAP`) with exponential, spherical, and Gaussian structures for the tail-up, tail-down, and Euclidean errors, respectively. We fit and summarize this model:

```
ssn_mod <- ssn_lm(
  formula = Summer_mn ~ ELEV_DEM + AREAWTMAP,
  ssn.object = mf04p,
  tailup_type = "exponential",
  taildown_type = "spherical",
  euclid_type = "gaussian",
  additive = "afvArea"
)
```

A summary of the fitted model looks similar to a summary returned by `lm()` but also returns spatial dependence parameter estimates:

```
summary(ssn_mod)
```

```
##
## Call:
## ssn_lm(formula = Summer_mn ~ ELEV_DEM + AREAWTMAP, ssn.object = mf04p,
##       tailup_type = "exponential", taildown_type = "spherical",
##       euclid_type = "gaussian", additive = "afvArea")
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6393 -2.0646 -0.5952  0.2143  0.7497
##
## Coefficients (fixed):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  76.195041   7.871574   9.680  < 2e-16 ***
```

```

88 ## ELEV_DEM      -0.026905    0.003646   -7.379  1.6e-13 ***
89 ## AREAWTMAP     -0.009099    0.004461   -2.040   0.0414 *
90 ## ---
91 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
92 ##
93 ## Pseudo R-squared: 0.6124
94 ##
95 ## Coefficients (covariance):
96 ##           Effect      Parameter      Estimate
97 ##   tailup exponential de (parsill)  3.800e+00
98 ##   tailup exponential           range  4.194e+06
99 ##   taildown spherical de (parsill)  4.480e-01
100 ##   taildown spherical           range  1.647e+05
101 ##   euclid gaussian   de (parsill)  1.509e-02
102 ##   euclid gaussian           range  4.496e+03
103 ##           nugget           nugget  2.087e-02

104 SSN2 leverages the tidy(), glance(), and augment() functions (Robinson, Hayes, & Couch,
105 2021) to inspect the fitted model and provide diagnostics:

tidy(ssn_mod, conf.int = TRUE)

106 ## # A tibble: 3 x 7
107 ##   term      estimate std.error statistic  p.value conf.low conf.high
108 ##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
109 ## 1 (Intercept) 76.2      7.87      9.68 0      60.8     91.6
110 ## 2 AREAWTMAP   -0.00910  0.00446   -2.04 4.14e- 2  -0.0178 -0.000356
111 ## 3 ELEV_DEM    -0.0269   0.00365   -7.38 1.60e-13 -0.0341 -0.0198

glance(ssn_mod)

112 ## # A tibble: 1 x 9
113 ##       n      p  npar value    AIC  AICc logLik deviance pseudo.r.squared
114 ##   <int> <dbl> <int> <dbl> <dbl> <dbl> <dbl>    <dbl>          <dbl>
115 ## 1     45     3     7  59.3  73.3  76.3 -29.6     41.9            0.612

aug_mod <- augment(ssn_mod)
subset(aug_mod, select = c(Summer_mn, .fitted, .resid, .hat, .cooksd))

116 ## Simple feature collection with 45 features and 5 fields
117 ## Geometry type: POINT
118 ## Dimension: XY
119 ## Bounding box: xmin: -1530805 ymin: 2527111 xmax: -1503079 ymax: 2537823
120 ## Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
121 ## # A tibble: 45 x 6
122 ##   Summer_mn .fitted .resid .hat .cooksd geometry
123 ##   <dbl> <dbl> <dbl> <dbl> <dbl> <POINT [m]>
124 ## 1  11.4  14.4 -3.07 0.0915 0.0962 (-1512690 2531883)
125 ## 2  10.7  12.9 -2.20 0.114 0.00471 (-1512852 2531295)
126 ## 3  10.4  12.7 -2.25 0.0372 0.00724 (-1513400 2530706)
127 ## 4  10.1  12.3 -2.18 0.0251 0.00153 (-1514027 2530147)
128 ## 5  10.1  12.3 -2.13 0.0374 0.000583 (-1514309 2529902)
129 ## 6   9.81  12.0 -2.16 0.0602 0.0150 (-1515032 2529461)
130 ## 7   9.76  11.6 -1.85 0.0736 0.00739 (-1515513 2528810)
131 ## 8   9.77  11.6 -1.84 0.0648 0.00687 (-1515588 2528592)
132 ## 9   9.53  11.4 -1.87 0.112 0.00152 (-1516440 2527899)
133 ## 10  12.6  14.9 -2.28 0.0498 0.00964 (-1512464 2531195)
134 ## # i 35 more rows

```

Specific helper functions (e.g., `coef()`, `AIC()`, `residuals()`) can be used to obtain the same quantities returned by `tidy()`, `glance()`, and `augment()`:

```
coef(ssn_mod)

## (Intercept)      ELEV_DEM      AREAWTMAP
## 76.19504087 -0.02690478 -0.00909941

AIC(ssn_mod)

## [1] 73.2623

head(residuals(ssn_mod))

##           1           2           3           4           5           6
## -3.066413 -2.204147 -2.252004 -2.175337 -2.131527 -2.162417

Prediction at the unobserved sites is performed using augment (or predict()):

aug_pred <- augment(ssn_mod, newdata = "pred1km", interval = "prediction")
subset(aug_pred, select = c(.fitted, .lower, .upper))

## Simple feature collection with 175 features and 3 fields
## Geometry type: POINT
## Dimension:      XY
## Bounding box:  xmin: -1530631 ymin: 2521707 xmax: -1500020 ymax: 2540253
## Projected CRS: USA_Contiguous_Albers_Equal_Area_Conic_USGS_version
## # A tibble: 175 x 4
##   .fitted .lower .upper      geometry
##   <dbl>   <dbl>   <dbl>   <POINT [m]>
## 1    14.6    14.3    15.0 (-1520657 2536657)
## 2    15.0    14.7    15.4 (-1519866 2536812)
## 3    14.8    14.3    15.3 (-1521823 2536911)
## 4    15.0    14.5    15.5 (-1523183 2537256)
## 5    15.2    14.7    15.6 (-1523860 2537452)
## 6    15.1    14.8    15.5 (-1525443 2537698)
## 7    15.1    14.7    15.5 (-1526397 2537254)
## 8    15.0    14.6    15.4 (-1527436 2536803)
## 9    14.9    14.6    15.3 (-1529043 2536449)
## 10   14.9    14.5    15.2 (-1529689 2537313)
## # i 165 more rows
```

Here, `.fitted` are the predictions, `.lower` are the lower bounds of 95% prediction intervals, and `.upper` are the upper bounds of 95% prediction intervals.

Generalized spatial linear models for binary, count, proportion, and skewed data are available via the `ssn_glm()` function. Simulating data on a stream network is performed via `ssn_simulate()`.

Discussion

SSN models are valuable tools for statistical analysis of data collected on stream networks and help improve inference about vital stream ecosystems. These models have been employed (using SSN) to better understand and manage water quality (McManus et al., 2020; Scown, McManus, Carson Jr, & Nietch, 2017), ecosystem metabolism (Rodríguez-Castillo, Estévez, González-Ferreras, & Barquín, 2019), and climate change impacts on freshwater ecosystems (Isaak, Wenger, et al., 2017; Ruesch et al., 2012), as well as generate aquatic population estimates (Isaak, Ver Hoef, Peterson, Horan, & Nagel, 2017), inform conservation planning (Rodríguez-González et al., 2019; Sharma, Dubey, Johnson, Rawal, & Sivakumar, 2021), and assess restoration activities (Fuller, Leinenbach, Detenbeck, Labiosa, & Isaak, 2022), among

other applications. The breadth and applicability of SSN models are further enhanced by data aggregation tools like the National Hydrography Dataset (McKay et al., 2012), National Stream Internet Project (Nagel, Peterson, Isaak, Ver Hoef, & Horan, 2015) and StreamCat (Hill, Weber, Leibowitz, Olsen, & Thornbrugh, 2016).

There are several spatial modeling packages in R, including geoR (Ribeiro Jr et al., 2022), gstat (E. J. Pebesma, 2004), FRK (Sainsbury-Dale, Zammit-Mangion, & Cressie, 2022), fields (Nychka, Furrer, Paige, & Sain, 2021), R-INLA (Lindgren & Rue, 2015), and spmodel (Dumelle, Higham, & Ver Hoef, 2023), among others. However, these packages fail to account for the intricacies of stream networks. rtop (Skoien et al., 2014) allows for spatial prediction on stream networks but fails to provide options for model fitting and diagnostics. Thus, SSN2 is the most complete tool available in R for working with SSN models. To learn more about SSN2, visit our CRAN webpage at <https://CRAN.R-project.org/package=SSN2>.

Acknowledgements

The views expressed in this manuscript are those of the authors and do not necessarily represent the views or policies of USEPA, NOAA, or USFS. Any mention of trade names, products, or services does not imply an endorsement by the U.S. government, USEPA, NOAA, or USFS. USEPA, NOAA, or USFS do not endorse any commercial products, services or enterprises.

References

- Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, second edition*. Springer, NY. Retrieved from <https://asdar-book.org/>
- Bivand, R., Keitt, T., & Rowlingson, B. (2021). *rgdal: Bindings for the 'geospatial' data abstraction library*. Retrieved from <https://CRAN.R-project.org/package=rgdal>
- Bivand, R., & Lewin-Koh, N. (2021). *maptools: Tools for handling spatial objects*. Retrieved from <https://CRAN.R-project.org/package=maptools>
- Bivand, R., & Rundel, C. (2020). *rgeos: Interface to geometry engine - open source ('GEOS')*. Retrieved from <https://CRAN.R-project.org/package=rgeos>
- Cressie, N. (1993). *Statistics for spatial data (revised edition)*. Wiley: Hoboken, NJ.
- Dumelle, M., Higham, M., & Ver Hoef, J. M. (2023). spmodel: Spatial statistical modeling and prediction in R. *PLOS ONE*, 18(3), 1–32. doi:10.1371/journal.pone.0282524
- Fuller, M. R., Leinenbach, P., Detenbeck, N. E., Labiosa, R., & Isaak, D. J. (2022). Riparian vegetation shade restoration and loss effects on recent and future stream temperatures. *Restoration Ecology*, 30(7), e13626.
- Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R., & Thornbrugh, D. J. (2016). The stream-catchment (StreamCat) dataset: A database of watershed metrics for the conterminous united states. *JAWRA Journal of the American Water Resources Association*, 52(1), 120–128.
- Isaak, D. J., Ver Hoef, J. M., Peterson, E. E., Horan, D. L., & Nagel, D. E. (2017). Scalable population estimates using spatial-stream-network (SSN) models, fish density surveys, and national geospatial database frameworks for streams. *Canadian Journal of Fisheries and Aquatic Sciences*, 74(2), 147–156.
- Isaak, D. J., Wenger, S. J., Peterson, E. E., Ver Hoef, J. M., Nagel, D. E., Luce, C. H., Hostetler, S. W., et al. (2017). The NorWeST summer stream temperature model and scenarios for the western US: A crowd-sourced database and new geospatial tools foster a

- 219 user community and predict broad climate warming of rivers and streams. *Water Resources*
220 *Research*, 53(11), 9181–9205.
- 221 Kattwinkel, M., Szöcs, E., Peterson, E., & Schäfer, R. B. (2020). Preparing GIS data for
222 analysis of stream monitoring data: The r package openSTARS. *Plos one*, 15(9), e0239237.
- 223 Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical*
224 *Software*, 63(19).
- 225 McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., & Reah, A. (2012). *NHD-*
226 *Plus version 2: User guide*. Retrieved from [http://www.horizon-systems.com/NHDPlus/](http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php)
227 [NHDPlusV2_home.php](http://www.horizon-systems.com/NHDPlus/NHDPlusV2_home.php)
- 228 McManus, M. G., D'Amico, E., Smith, E. M., Polinsky, R., Ackerman, J., & Tyler, K.
229 (2020). Variation in stream network relationships and geospatial predictions of watershed
230 conductivity. *Freshwater Science*, 39(4), 704–721.
- 231 Nagel, D., Peterson, E., Isaak, D., Ver Hoef, J., & Horan, D. (2015). National stream internet
232 protocol and user guide. *US Forest Service, Rocky Mountain Research Station Air, Water,*
233 *and Aquatic Environments Program*.
- 234 Nowosad, J. (2023, June 4). Upcoming Changes to Popular R Packages for Spatial Data:
235 What You Need to Do. Retrieved from <https://geocompx.org/post/2023/rgdal-retirement>
- 236 Nychka, D., Furrer, R., Paige, J., & Sain, S. (2021). Fields: Tools for spatial data. Boulder,
237 CO, USA: University Corporation for Atmospheric Research. Retrieved from [https://github.](https://github.com/dnychka/fieldsRPackage)
238 [com/dnychka/fieldsRPackage](https://github.com/dnychka/fieldsRPackage)
- 239 Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data.
240 *The R Journal*, 10(1), 439–446. doi:[10.32614/RJ-2018-009](https://doi.org/10.32614/RJ-2018-009)
- 241 Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers &*
242 *Geosciences*, 30, 683–691.
- 243 Peterson, E. E., & Ver Hoef, J. M. (2010). A mixed-model moving-average approach to
244 geostatistical modeling in stream networks. *Ecology*, 91(3), 644–651.
- 245 Peterson, E. E., & Ver Hoef, J. M. (2014). STARS: An ArcGIS toolset used to calculate the
246 spatial information needed to fit spatial statistical models to stream network data. *Journal*
247 *of Statistical Software*, 56, 1–17.
- 248 Ribeiro Jr, P. J., Diggle, P., Christensen, O., Schlather, M., Bivand, R., & Ripley, B.
249 (2022). *geoR: Analysis of geostatistical data*. Retrieved from [https://CRAN.R-project.org/](https://CRAN.R-project.org/package=geoR)
250 [package=geoR](https://CRAN.R-project.org/package=geoR)
- 251 Robinson, D., Hayes, A., & Couch, S. (2021). *Broom: Convert statistical objects into tidy*
252 *tibbles*. Retrieved from <https://CRAN.R-project.org/package=broom>
- 253 Rodríguez-Castillo, T., Estévez, E., González-Ferreras, A. M., & Barquín, J. (2019). Estimating
254 ecosystem metabolism to entire river networks. *Ecosystems*, 22, 892–911.
- 255 Rodríguez-González, P. M., García, C., Albuquerque, A., Monteiro-Henriques, T., Faria, C.,
256 Guimarães, J. B., Mendonça, D., et al. (2019). A spatial stream-network approach assists
257 in managing the remnant genetic diversity of riparian forests. *Scientific Reports*, 9(1),
258 6741.
- 259 Ruesch, A. S., Torgersen, C. E., Lawler, J. J., Olden, J. D., Peterson, E. E., Volk, C. J., &
260 Lawrence, D. J. (2012). Projected climate-induced habitat loss for salmonids in the john
261 day river network, oregon, USA. *Conservation Biology*, 26(5), 873–882.
- 262 Sainsbury-Dale, M., Zammit-Mangion, A., & Cressie, N. (2022). Modelling big, heterogeneous,
263 non-gaussian spatial and spatio-temporal data using FRK. *arXiv:2110.02507*.

- 264 Scown, M. W., McManus, M. G., Carson Jr, J. H., & Nietch, C. T. (2017). Improving predictive
265 models of in-stream phosphorus concentration based on nationally-available spatial data
266 coverages. *Journal of the American Water Resources Association*, 53(4), 944–960.
- 267 Sharma, A., Dubey, V. K., Johnson, J. A., Rawal, Y. K., & Sivakumar, K. (2021). Dendritic
268 prioritization through spatial stream network modeling informs targeted management of
269 himalayan riverscapes under brown trout invasion. *Journal of Applied Ecology*, 58(11),
270 2415–2426.
- 271 Skoien, J. O., G. Blöschl, G. Laaha, E. Pebesma, J. Parajka, & A. Viglione. (2014). rtop: An
272 R package for interpolation of data with a variable spatial support, with an example from
273 river networks. *Computers & Geosciences*.
- 274 Ver Hoef, Jay M., & Peterson, E. E. (2010). A moving average approach for spatial statistical
275 models of stream networks. *Journal of the American Statistical Association*, 105(489),
276 6–18.
- 277 Ver Hoef, Jay M., Peterson, E. E., Clifford, D., & Shah, R. (2014). SSN: An R package for
278 spatial statistical modeling on stream networks. *Journal of Statistical Software*, 56, 1–45.
- 279 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
280 Retrieved from <https://ggplot2.tidyverse.org>

DRAFT