sdmTMB: an R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields

Sean C. Anderson[1]*, Eric J. Ward[2], Philina A. English[1], Lewis A. K. Barnett[3]
[1]Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, BC, Canada
[2]Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, USA
[3]Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA, USA

Running headline: sdmTMB: spatial and spatiotemporal GLMMs

Abstract

1. Geostatistical data—spatially referenced observations related to some continuous spatial phenomenon—are ubiquitous in ecology and can reveal ecological processes and inform management decisions. However, appropriate models to analyze these data, such as generalized linear mixed effects models (GLMMs) with Gaussian random fields, are often computationally intensive and challenging to implement, interpret, and evaluate.

2. Here, we introduce the R package sdmTMB, which implements predictive-process SPDE- (stochastic partial differential equation) based spatial and spatiotemporal models. Estimation is conducted via maximum marginal likelihood with Template Model Builder (TMB) but can be extended to penalized likelihood or Bayesian inference. We describe the statistical model, illustrate the package's use through two case studies, and compare the functionality, speed, and interface to related software.

3. We highlight advantages of using sdmTMB for this class of models: (1) sdmTMB provides a flexible interface familiar to users of `glm()`, lme4, glmmTMB, or mgcv; (2) estimation is often faster than alternatives; (3) sdmTMB provides simple out-of-sample cross validation; (4) non-stationary processes (time-varying and spatially varying coefficients) are easily constructed with a formula interface; and (5) sdmTMB includes features not available as a combination in related packages (e.g., delta/hurdle models, penalized smoothers and break-point effects, anisotropy, abundance index standardization).

4. We hope that sdmTMB's user-friendly interface will open this useful class of models to a wider audience within species distribution modelling and beyond.

Keywords: Gaussian Markov random fields (GMRF), generalized linear mixed effects models (GLMM), INLA, SPDE, species distribution modelling, R package, spatio-temporal or spatial-temporal, Template Model Builder

*Corresponding author: sean.anderson@dfo-mpo.gc.ca

Introduction

Ecological data are often collected in space or in space repeatedly over time. While such data are a rich source of information about ecological processes (Legendre & Fortin 1989; Rossi *et al.* 1992; Tilman *et al.* 1997), they are challenging to properly model—data closer in space and time are usually more similar to each other than data farther apart due to measured and unmeasured variables (Cressie 1993; Diggle & Ribeiro 2007; Cressie & Wikle 2011). While measured variables can be accounted for with predictors in a model (e.g., measuring and modelling temperature effects on species abundance), unmeasured variables (e.g., everything influencing species abundance but not explicitly modelled) can cause residual spatial correlation. Accounting for this residual correlation is important because doing so allows for valid statistical inference (Legendre & Fortin 1989; Dormann *et al.* 2007), can improve predictions (e.g., Shelton *et al.* 2014), and can be of ecological interest itself by, for example, identifying locations with similar population responses (e.g., Thorson 2019b; Barnett *et al.* 2021).

Geostatistical GLMMs (generalized linear mixed effects models) with spatially correlated random effects are a class of models appropriate for these data (Rue & Held 2005; Diggle & Ribeiro 2007; Cressie & Wikle 2011). Similarly to how random intercepts can account for correlation among groups, spatial or spatiotemporal random effects can account for unmeasured variables causing observations to be correlated in space or space and time. A common approach to modelling these spatial effects is with Gaussian random fields (GRFs), where the random effects describing the spatial patterning are assumed to be drawn from a multivariate normal (MVN) distribution, constrained by some covariance function such as the exponential or Matérn (Cressie 1993; Chilés & Delfiner 1999; Diggle & Ribeiro 2007).

Such models quickly become computationally limiting due to the need to invert large matrices to keep track of covariation among data (e.g., Rue & Held 2005; Latimer *et al.* 2009). Many solutions have been proposed, such as predictive processes (Banerjee *et al.* 2008; Latimer *et al.* 2009), the stochastic partial differential equation (SPDE) approximation to GRFs (Lindgren *et al.* 2011), and nearest-neighbour Gaussian processes (Datta *et al.* 2016; Finley *et al.* 2021). These approaches reduce the scale of the covariance estimation problem while providing a means to evaluate the data likelihood, thereby allowing fitting via Bayesian (Gelfand & Banerjee 2017) or maximum likelihood methods. This can greatly improve computational efficiency (e.g., Heaton *et al.* 2019). The SPDE approach is a solution popularized via the INLA R package (Rue *et al.* 2009; Lindgren *et al.* 2011; Lindgren & Rue 2015) and an implementation in TMB (Template Model Builder, Kristensen *et al.* 2016) that partially relies on INLA to create input matrices (e.g., Osgood-Zimmerman & Wakefield 2021). Details are beyond the scope of this paper and are not necessary to use the software discussed here, but the idea is that the solution to a specific SPDE is a GRF with a Matérn covariance function and this 'trick' enables one to efficiently fit approximations to GRFs to large spatial datasets (Lindgren *et al.* 2011).

Systems for specifying statistical models that can include the SPDE, such as INLA and TMB, are flexible and powerful but are challenging to use for many applied ecologists. For example, TMB requires the user to program in a C++ template and it can be slow to experiment with multiple models when writing bespoke model code. Packages such as lme4 (Bates *et al.* 2015) and glmmTMB (Brooks *et al.* 2017) let users quickly iterate and explore statistical models—focusing on evaluating fit and comparing models—but do not have built-in SPDE functionality. Packages such as VAST (Thorson 2019a) and inlabru (Bachl *et al.* 2019) are powerful user interfaces to fit spatial models that use the SPDE, but they either lack a modular interface

familiar to those who have used lme4 or glmmTMB, or lack some functionality. We provide a more detailed comparison of related software packages in Table 1 and the Discussion.

Here, we introduce the R package sdmTMB, which implements geostatistical spatial and spatiotemporal GLMMs using TMB for model fitting and INLA to set up SPDE matrices. Our aim is not to replace the above-mentioned statistical packages, but to provide a fast, flexible, and user-friendly interface that is familiar to users of lme4, glmmTMB, or mgcv (Wood 2017), for a specific class of spatial and spatiotemporal models. One common application is for species distribution models (SDMs), hence the package name. This paper describes the basic functionality of this R package and its underlying statistical model, illustrates its use through two case studies, and concludes with a comparison to related software.

Model description

sdmTMB fits GLMMs to spatial or spatiotemporal geostatistical data. Geostatistical data simply refers to data observed at specific spatial coordinates reflecting some underlying spatial process (Rossi *et al.* 1992; Diggle & Ribeiro 2007). These data can be collected across discrete points in time. Areal data (data aggregated to polygon or grid level) may be analyzed using other spatial models, including conditional (CAR) autoregressive models (e.g., Ver Hoef *et al.* 2018). sdmTMB can also fit models with areal data if each polygon has an associated centroid. A benefit of the geostatistical approach over CAR or similar models is that the parameters describing spatial covariance can be more easily interpreted (Wall 2004).

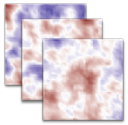
The process component of an sdmTMB model can be formed by any combination of main ("fixed") effects, spatial intercept random fields, spatiotemporal intercept random fields, IID (independent and identically distributed) random intercepts (but not currently random slopes), time-varying effects, and spatially varying effects (Fig. 1, Appendix 1). This process component is combined with an observation error family (e.g., Gaussian, Gamma, binomial, Tweedie) and link (e.g., identity, log, logit) as in any generalized linear model (GLM) (McCullagh & Nelder 1989). Some families can be combined into a two-part "delta" or "hurdle" model (Aitchison 1955) to model the zero vs. non-zero observations separately from the positive observations.

The GLMMs underpinning sdmTMB models are spatially explicit—they estimate interpretable parameters of a spatial covariance function: parameters defining the magnitude of spatial variation and the rate of correlation decay with distance. In contrast, semi- and non-parametric approaches do not estimate spatial covariance functions (e.g., randomForest (Liaw & Wiener 2002), MaxEnt (Phillips *et al.* 2006), and most smooths in mgcv (Wood 2017)). The random fields in sdmTMB are structured as MVN constrained by a Matérn covariance function (Matérn 1960). The Matérn can accommodate a range of shapes and can be both isotropic (covariance decays the same in all directions) or anisotropic (covariance in the latitudinal direction may differ from the longitudinal direction) (Haskard 2007). The Matérn standard deviations are estimated separately for the various fields and the range—the distance at which spatial correlation decays to ~ 0.13 (Lindgren & Rue 2015)—can be shared or estimated separately (`share_range` argument).

Model component	Illustration	Description	Example	Notation	Example code

Main effects		Linear, smoother, or breakpoint effects	Linear temperature, spline of depth, or breakpoint effect of oxygen on abundance	$\boldsymbol{X}^{\text{main}}_{\boldsymbol{s},t}\boldsymbol{\beta}$	`formula = y ~ x`
					`formula = y ~ s(x)`
					`formula = y ~ breakpt(x)`

Spatial random effects		All spatially correlated effects from variables that are constant in time but are omitted from the model (or a model without a time element)	Depth, latitude, or substrate effects if omitted from model	$\omega_{\boldsymbol{s}} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\omega})$	`spatial = 'on'` `spatial = 'off'`

Spatiotemporal random effects		All spatially correlated effects from variables that change through time but are omitted from the model	Temperature, oxygen, prey abundance effects if omitted from the model	$\epsilon_{\boldsymbol{s},t} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\epsilon})$	`spatiotemporal = 'iid'`
					`spatiotemporal = 'ar1'`
					`spatiotemporal = 'rw'`
					`spatiotemporal = 'off'`

IID random intercepts		Group-level effects that are constrained by normal distributions	Transect ID, vessel ID	$\alpha_g \sim \text{N}(0, \sigma^2_{\alpha})$	`formula = y ~ (1|g)`
Time-varying effects		Effects that vary though time	Relationship between depth and fish abundance changing through time	$\boldsymbol{X}^{\text{tvc}}_{\boldsymbol{s},t}\gamma_t$ $\gamma_t \sim \text{N}(\gamma_{t-1}, \sigma^2_{\gamma})$	`time_varying =` `  ~ 0 + depth`
Spatially varying effects		Effects ('slopes') that vary in space	(1) Local trends in abundance over time; (2) when a climate index is high, hotspots look one way, and vice versa	$\boldsymbol{X}^{\text{svc}}_{\boldsymbol{s},t}\zeta_{\boldsymbol{s}}$ $\zeta_{\boldsymbol{s}} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{\zeta})$	`spatial_varying =` `  ~ 0 + climate_index`

Observation error		Error from observing or sampling the process	Counting birds or fish in a survey; recording presence/ absence in a quadrat		`family = binomial(link = "logit")`
					`family = nbinom2(link = "log")`
					`family = tweedie(link = "log")`

Figure 1: Components of an sdmTMB model with illustrations, descriptions, examples, notation, and example code. An sdmTMB model can be built from any combination of the process components (first six rows) plus an observation component (last row). The examples are from an SDM context, but the model can be fit to any spatially referenced point data. Notation: We refer to design matrices as $\boldsymbol{X}$. The indexes $\boldsymbol{s}$, t, and g index spatial coordinates, time, and group, respectively. The σ and $\boldsymbol{\Sigma}$ symbols represent standard deviations and covariance matrices, respectively. All other symbols refer to the described model components (e.g., $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ refer to a vector of main effects and spatial random field deviations, respectively). See Appendix 1 for a full description of the model. Note that `s()` denotes a smoother as in mgcv (Wood 2017), `breakpt()` denotes a breakpoint 'hockey-stick' shape (e.g., Barrowman & Myers 2000), `(1|g)` denotes a random intercept by group `g`, and `~ 0` is used in an R formula to omit an intercept.

By default, if spatiotemporal fields are included, they are assumed IID; however, additional options allow them to be modelled as a random walk or first-order autoregressive, AR(1), process (Fig. 1; Appendix 1). Turning off both spatial and spatiotemporal effects allows comparison with a standard non-spatial GLM or GLMM. We include additional flexibility in specifying the linear fixed effect matrix: covariates can be modelled as penalized smooth functions (generalized additive models, GAMs) using the same `s()` syntax as in mgcv (Wood 2017) (thereby allowing automatic selection of smoother 'wiggliness'), or can be modelled with threshold shapes; for example, hockey-stick models, `breakpt()`, (Barrowman & Myers 2000) or logistic functions, `logistic()` (Appendix 1). The smoothers in sdmTMB can mimic most `mgcv::s()` smoothers including bivarate smoothers (`s(x, y)`), smoothers varying by continuous or categorical variables (`s(x1, by = x2)`), cyclical smoothers `s(x, bs = "cc")`, and smoothers with specified basis dimensions `s(x, k = 4)` (Wood 2017).

The sdmTMB model is fit by maximum marginal likelihood. Internally, a TMB (Kristensen *et al.* 2016) model template is used to calculate the marginal log likelihood and its gradient, and the negative log likelihood is minimized via the non-linear optimization routine `stats::nlminb()` in R (Gay 1990; R Core Team 2021). Random effects are estimated at values that maximize the log likelihood conditional on the estimated fixed effects and are integrated over via the Laplace approximation (Kristensen *et al.* 2016). After rapid model exploration with maximum likelihood, one can optionally pass an sdmTMB model to the R package tmbstan (Monnahan & Kristensen 2018) to estimate the joint posterior distribution for Bayesian inference.

sdmTMB models can include penalized likelihoods by assigning priors to model parameters (`?sdmTMBpriors`). These priors may be useful in cases where estimation is difficult because of identifiability issues or relatively flat likelihood surfaces, or to impart prior information or achieve regularization. Following other recent SPDE implementations in TMB (Osgood-Zimmerman & Wakefield 2021; Breivik *et al.* 2021), penalized complexity (PC) priors (Simpson *et al.* 2017; Fuglstad *et al.* 2019) (`?pc_matern`) can constrain the spatial range and variance parameters.

The results from an sdmTMB model can be used to generate various quantities of interest. For example, sdmTMB provides functionality for generating population-level summaries for each time slice including the center of gravity (`get_cog()`) (e.g., Thorson *et al.* 2016b) and total population index (`get_index()`) given a user-supplied grid.

Introductory example using fish survey data

An sdmTMB model requires a data frame that contains a response column, columns for any predictors, and columns for spatial coordinates. It usually makes sense to convert the spatial coordinates to an equidistant projection such as the Universal Transverse Mercator (UTM) to ensure that distance remains constant throughout the study region (e.g., using `sf::st_transform()`, Pebesma 2018, or `sdmTMB::add_utm_columns()`). Here, we illustrate a spatial model fit to Pacific cod (*Gadus macrocephalus*) trawl survey data from Queen Charlotte Sound, British Columbia, Canada. Our model contains a main effect of depth as a penalized smoother, a spatial random field, and Tweedie observation error. Our data frame `pcod` (built into the package) has a column `year` for the year of the survey, `density` for biomass density of Pacific cod in the area swept for a given survey tow, `depth` for depth in meters of that tow, and spatial coordinates `X` and `Y`, which are UTM coordinates in kilometres.

```
library(sdmTMB)
head(pcod)
```

```
#>    year density depth     X     Y
#>   <int>   <dbl> <dbl> <dbl> <dbl>
#> 1  2003   113.    201  446. 5793.
#> 2  2003    41.7   212  446. 5800.
#> 3  2003     0     220  449. 5802.
```

We start by creating a mesh object that contains matrices to apply the SPDE approach.

```
mesh <- make_mesh(pcod, xy_cols = c("X", "Y"), cutoff = 10)
```

Here, `cutoff` defines the minimum allowed distance between mesh vertices in the units of `X` and `Y`. Alternatively, we could have created any mesh via the INLA package and supplied it to `make_mesh()`. We can inspect our mesh object with the associated plotting method (`plot(mesh)`; see Fig. 2A).

We can then fit a model with spatial random fields (`spatial = "on"`) via the function `sdmTMB()`. We use a penalized smoother for depth as a main effect via `s()` from the mgcv package. We specify the family as Tweedie to account for positive continuous density values that also contain zeros. An alternative would be the `delta_gamma()` family to specify a delta/hurdle model (Aitchison 1955) or to model catch weight with an offset for log(area swept).

```
fit <- sdmTMB(
  density ~ s(depth),
  data = pcod,
  family = tweedie(link = "log"),
  mesh = mesh,
  spatial = "on"
)
```

We can get a summary of the fit with the `print()` or `summary()` methods or extract parameter estimates as a data frame with the `tidy()` method.

```
summary(fit)
```

```
#> Spatial model fit by ML ['sdmTMB']
#> Formula: density ~ s(depth)
#> Mesh: mesh
#> Data: pcod
#> Family: tweedie(link = 'log')
#>             coef.est coef.se
#> (Intercept)     2.37    0.21
#> sdepth          6.17   25.17
#>
#> Smooth terms:
#>            Std. Dev.
#> sds(depth)     13.93
#>
```

```
#> Dispersion parameter: 12.69
#> Tweedie p: 1.58
#> Matern range: 16.39
#> Spatial SD: 1.86
#> ML criterion at convergence: 6402.136
```

The output indicates our model was fit by maximum (marginal) likelihood (`ML`). We also see the formula, mesh, fitted data, and family. Next we see any estimated main effects including the linear component of the smoother (`sdepth`), the standard deviation on the smoother weights (`sds(depth)`), the Tweedie dispersion and power parameters, the Matérn range distance, the marginal spatial field standard deviation, and the negative log likelihood at convergence.

We can make predictions with the `predict()` method (`?predict.sdmTMB`) and optionally use the `newdata` argument to predict on a new data frame with any locations and values for the predictor columns. Here, we will predict on a 2x2 km grid (`qcs_grid`) that covers the entire region of interest so we can visualize the predictions spatially (or calculate a standardized population index with a spatiotemporal model). The grid contains spatial covariate columns and all predictors used in the model set at values for which we want to predict.

```
p <- predict(fit, newdata = qcs_grid)
```

The output of `predict()` is a data frame containing overall estimates in link space, estimates from the non-random-field components (intercept and depth), and estimates from the random field components. We show a basic plot of the estimated spatial random field, predictions across a depth gradient, and predictions in space in Fig. 2B–D.

We could extend this spatial model to be a spatiotemporal one simply by supplying the year column name to the `time` argument and specifying how we want the spatiotemporal random fields to be structured. Here we will keep the spatial field and structure the spatiotemporal random fields as AR(1).

```
fit_spatiotemporal <- sdmTMB(
  density ~ s(depth), family = tweedie(link = "log"), data = pcod, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "ar1"
)
```

We could generate an area-weighted population index (e.g., a relative or absolute index of abundance or biomass) that is independent of sampling locations by predicting from that fit on a grid covering the area of interest and summing the predicted biomass with the `get_index()` function (Fig. 2E, Appendix 2).

Example of a spatially varying coefficient with citizen science data

Snowy Owls (*Bubo scandiacus*) breed on the arctic tundra and are irruptive migrants, meaning that they appear across the mid-latitudes of North America in much greater numbers in some winters than others. The reasons for this interannual variation in the number of individuals migrating south are not well understood but seem to be related to high abundances of food during the breeding season and therefore sharp increases in breeding ground population densities (Robillard *et al.* 2016). The North Atlantic Oscillation Index (NAO) has been linked to productivity of both owls and their prey in Europe (Millon *et al.* 2014). Because both

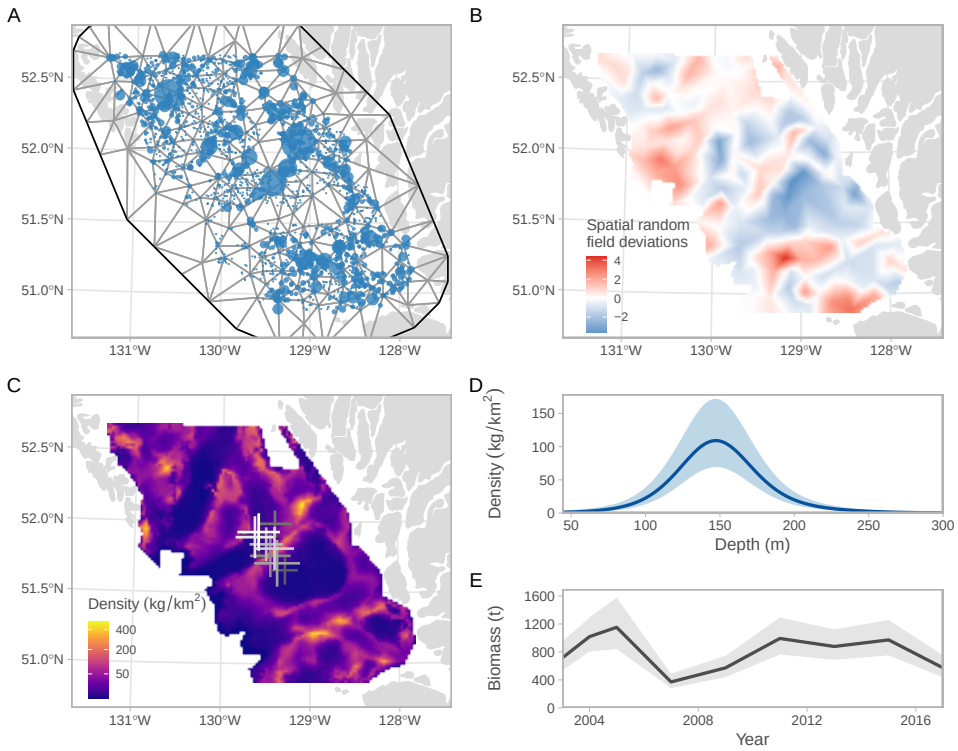


Figure 2: Output from the Pacific cod spatial model example in Queen Charlotte Sound, BC, Canada. (A) SPDE mesh (lines) combined with the trawl survey observations (points). Finer meshes will be slower to fit but generally increase the accuracy of the SPDE approximation. The circle area corresponds to biomass density of Pacific cod caught on individual trawls. (B) Spatial random field: these values are shown in link (log) space and represent spatially correlated deviations that are not accounted for by the depth effect. (C) Overall prediction: these estimates represent the combination of all fixed and random effects. Crosses illustrate center of gravity (`get_cog()`) from a spatiotemporal version; lighter crosses are more recent. (D) Conditional effect of depth modelled as a penalized smoother. These predictions are made omitting the spatial random field. (E) Standardized area-weighted population index derived from a spatiotemporal version (`get_index()`). The line represents the estimate and the ribbon indicates a 95% confidence interval (+/- 2 SEs).

productivity and the choice of wintering location could be influenced by climate, we tested for a spatially varying effect of annual mean NAO index on winter abundance across the southern boundary of their winter distribution. We fit counts observed in North America during annual Christmas Bird Counts (National Audubon Society 2021) using a negative binomial (NB2) distribution, random intercepts for year, spatial and spatiotemporal random fields, and a spatially varying coefficient associated with the NAO.

```
mesh <- make_mesh(snow, xy_cols = c("X", "Y"), cutoff = 1.5)
fit_owls <- sdmTMB(
  count ~ 1 + nao + (1 | year_factor),
  spatial_varying = ~ 0 + nao,
  family = nbinom2(link = "log"), data = snow, mesh = mesh,
  time = "year", spatial = "on", spatiotemporal = "iid"
)
```

We found a weak average positive effect of annual mean NAO on overall counts, but a southeast to northwest gradient in the intensity of the effect (Fig. 3, Appendix 3). This result is consistent with owls closest to the Atlantic coast and those migrating the furthest south being the most affected by NAO.

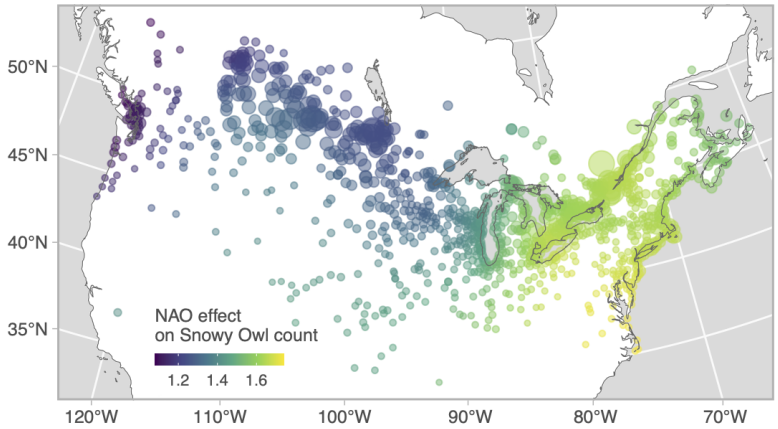


Figure 3: Spatially varying coefficient for effect of mean annual NAO (North Atlantic Oscillation) on counts of Snowy Owls observed on annual Christmas Bird Counts 1979–2020 in Canada and the US. Points represent all count locations and circle area is scaled to the mean number of owls observed per year (range: 0 to 8). The effect is multiplicative on owl count per NAO unit.

Model validation and selection

Validation and selection of state-space models is challenging, particularly when using the Laplace approximation (Thygesen *et al.* 2017). We provide several approaches to assist this process: (1) The Akaike Information Criterion (AIC, Akaike 1974) can be calculated with `AIC()`, although AIC has well-documented biases with mixed-effects models (Liang *et al.* 2008). (2) Alternatively, k-fold cross validation with `sdmTMB_cv()` can be used with user-specified or randomly chosen folds for model selection (e.g., via expected log predictive density [ELPD], Vehtari *et al.* 2017) or to evaluate goodness of fit according to user-calculated criteria (e.g., mean squared error, area under the curve). (3) An sdmTMB model can be passed to the tmbstan package (Monnahan & Kristensen 2018) to sample from the joint posterior with Stan (Carpenter *et al.*

2017), evaluate the accuracy of the Laplace approximation, or perform posterior predictive checks (see `?extract_mcmc`). (4) The `residuals()` method by default returns randomized quantile (Dunn & Smyth 1996) or probability integral transform (PIT) (Smith 1985) residuals. For state-space models, these residuals have known statistical issues with the Laplace approximation (Thygesen *et al.* 2017) but are quick to calculate. A version that uses Markov chain Monte Carlo (MCMC) to avoid this issue is recommended but slower (`?residuals.sdmTMB`). Simulation-based residuals (Hartig 2021) (`dharma_residuals()`) are also possible. (5) The `simulate.sdmTMB()` method can simulate from fitted models and the `sdmTMB_simulate()` function can simulate entirely new data to which models can be fit to ensure identifiability, evaluate bias and precision in parameter estimation, or evaluate the consequences of model misspecification.

Package comparisons

There are many R packages capable of fitting geostatistical spatial or spatiotemporal models (e.g., Heaton *et al.* 2019). sdmTMB, VAST, INLA/inlabru, and spaMM (Rousset & Ferdy 2014) are the most closely related, as they all provide a user interface to SPDE-based GRF models; we also include spBayes in our software comparison as it is a prominent package that can fit related predictive-process models without the SPDE (Table 1). sdmTMB, VAST, and mgcv can estimate anisotropic covariance whereas INLA/inlabru and spBayes are currently limited to isotropic covariance. sdmTMB and mgcv focus on univariate response data, whereas VAST, INLA/inlabru, spaMM, and spBayes extend to multivariate responses with various limitations. To our knowledge, VAST is the only package to implement spatial (Thorson *et al.* 2015) and spatial dynamic factor analysis (Thorson *et al.* 2016a) and spatial empirical orthogonal function (EOF) regression (Thorson *et al.* 2020). Of these packages, only sdmTMB and inlabru can currently fit threshold (e.g., hockey-stick) covariate relationships. spaMM is limited to a spatial random field and spBayes implements spatiotemporal fields, but only as a random walk. There is considerable variability in the available observation likelihoods across packages (Table 1). We provide comparisons of the syntax and the reproducibility of results from models fit using INLA or inlabru (Appendix 4) and VAST (Appendix 5).

We ran a simple speed comparison between sdmTMB, inlabru/INLA, and mgcv for fitting an SPDE spatial random field model to 1000 data points with Gaussian error across a range of mesh resolutions (Fig. 4, Appendix 6). In this test, sdmTMB was the fastest across all mesh resolutions; however, inlabru/INLA and spaMM were less affected by mesh resolution than sdmTMB. mgcv was most affected by mesh resolution. Our test was restricted to one core and default R algebra libraries; all packages could run faster with optimized libraries and parallel processing. Results with optimized math libraries on one core (openBLAS: Xianyi & Kroeker (2021) and PARDISO: Bollhöfer *et al.* (2019)) resulted in a ~10% speed increase for sdmTMB and inlabru and a ~7–9-fold speed increase for mgcv.

Discussion

How does one choose among the related packages mentioned in this paper to fit SPDE-based geostatistical GLMMs? Assuming a given package can fit the model of interest (Table 1), we suggest the primary differences are the user interface and speed. We think users familiar with `stats::glm()`, lme4, or glmmTMB will find sdmTMB most approachable. Users familiar with INLA will find inlabru approachable. Users familiar with mgcv can adapt mgcv to fit similar models with custom code (Miller *et al.* 2019) and INLA/inlabru and mgcv are also general purpose modelling packages. VAST is the sole option for fitting some multivariate models;

Table 1: Comparison of functionality between several R packages that can fit geostatistical GLMMs.

	sdmTMB	VAST	INLA/inlabru	mgcv	spBayes	spaMM
Time-varying coefficients	✓	–[1]	✓	✓	✓	–
Spatially varying coefficients (SVC)	✓	✓	✓	✓	✓	–
GAMs[2]	✓	–	✓	✓	–	–
Threshold covariates	✓	–	✓[3]	–	–	–
Offsets	✓	✓	✓	✓	✓	✓
Spatiotemporal fields	✓	✓	✓	✓	✓[4]	–
Spatial + spatiotemporal fields	✓	✓	✓	✓	–	–
Anisotropy	✓	✓	–	✓	–	–
Correlation barriers	✓	✓	✓	✓	–	–
Separate range parameters for fields	✓	–	✓	✓	–	–
Share range parameters across fields	✓	✓	✓	–	–	–
SPDE-based	✓	✓	✓	✓[5]	–[6]	✓
NB1 distribution	✓	–	✓	✓	–	✓
NB2 distribution	✓	✓[7]	✓	✓	–	✓
Zero-truncated distributions	✓	–	✓	–	–	✓
Zero-inflated distributions	✓	✓	✓	–	–	✓
Tweedie distribution	✓	✓	✓	✓[8]	–	–
Student-t distribution	✓	–	✓	✓	–	–
Censored Poisson distribution	✓	–	✓	–	–	–
log Gaussian Cox processes	–[9]	–[9]	✓	–[9]	–[9]	–[9]
Multivariate responses	–	✓	✓	–	✓	✓
Built-in delta/hurdle models	✓	✓	✓	–[10]	–	✓
Poisson-link delta model	✓	✓	✓	–	–	–
Likelihood weights	✓	–	✓	✓	✓	✓
Maximum/marginal likelihood	✓	✓	–	✓	–	–
Bayesian/optionally Bayesian	✓	✓	✓	✓	✓	–
Priors/penalties	✓	–	✓	–	✓	–
Matern PC priors	✓	–	✓	–	–	–
Spatial (or spatial dynamic) factor analysis	–	✓	–	–	–	–
Empirical Orthogonal Function (EOF) analysis	–	✓	–	–	–	–
Built-in area-weighted index standardization	✓	✓	–	–	–	–
Built-in cross-validation	✓	–	–	–	–	–

Note:
[1]Technically possible but non-trivial. [2]Penalized smoother GAMs that determine 'wiggliness'. [3]inlabru but not INLA. [4]Spatiotemporal fields as random walk only. [5]SPDE approach as in Miller et al. (2019). [6]Does have predictive process knots. [7]Zero-inflated NB2 only. [8]Tweedie power parameter fixed for `mgcv::gamm()`. [9]Possible as log-linked Poisson GLMM with aggregated data. [10]Hurdle models possible by fitting components separately.

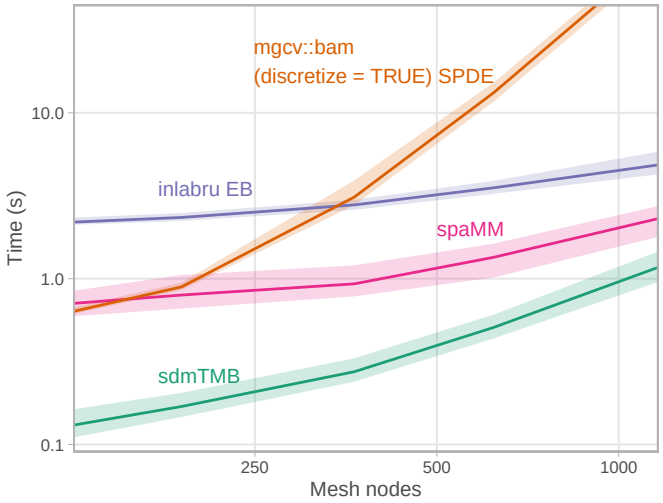


Figure 4: Comparison of time to fit an SPDE spatial random field model with 1000 observations, an intercept and one predictor, Gaussian error, and a sequence of SPDE resolutions. Lines represent means and ribbons 95% quantiles across 50 random iterations. Note the log x and y axes. VAST is similar to sdmTMB and so is not shown. inlabru used the empirical Bayes integration strategy and Gaussian approximation with `bru_max_iter = 1`, and the `like()` formulation. mgcv used `bam()`, `method = "fREML"`, and discretized covariates (Miller *et al.* 2019). Note that spaMM only fits spatial, not spatiotemporal, models. All platforms were restricted to one core and could be faster with parallel computation or optimized algebra libraries.

alternatively, because VAST focuses on multivariate delta models and fisheries applications (Appendix 5), users fitting "simple" univariate spatial/spatiotemporal GLMMs in non-fisheries contexts may find sdmTMB more straightforward to use. Users looking for calculation, with uncertainty, of derived variables such as area-weighted population indexes, may favour sdmTMB or VAST (although such quantities can be post hoc derived with other packages).

Speed-wise, sdmTMB (and by association VAST) were fastest up to at least 1000 mesh nodes at approximately 4–16 times faster than INLA/inlabru across the range of mesh complexities considered here. These speed increases can allow for more rapid and thorough model exploration and experimentation with a class of computationally intensive models. However, for users ultimately interested in Bayesian inference, the approximate Bayesian inference offered by INLA/inlabru is likely to be considerably faster than passing the same model from sdmTMB/VAST to tmbstan for full Bayesian inference. Furthermore, although both INLA/inlabru and sdmTMB can use parallel processing, the PARDISO library (Bollhöfer *et al.* 2019) within INLA/inlabru allows memory to be shared across cores.

Additional functionality in sdmTMB not already mentioned includes interpolating across missing time slices and forecasting, the barrier SPDE model of Bakka *et al.* (2019), time-varying spatiotemporal covariance parameters, and simulation from the parameter joint precision matrix. Future development may include additional zero-inflated models, improvements to Bayesian sampling efficiency (e.g., Monnahan *et al.* 2021), continuous time models (e.g., Blangiardo & Cameletti 2015), multivariate responses, and non-Gaussian random fields (e.g., Anderson & Ward 2019). The included TMB `.cpp` file also provides a high-quality and tested model template that can be modified to add additional features.

Spatially and spatiotemporally explicit data are increasingly collected in ecology and have the power to reveal new ecological processes (e.g., Dinnage *et al.* 2020; English *et al.* 2022) and improve ecological management (Sofaer *et al.* 2019). These data present statistical challenges to modelling them effectively and efficiently since appropriate models such as GLMMs with random fields are often computationally intensive and challenging to implement, interpret, and evaluate. We hope the development of user-friendly interfaces such as sdmTMB opens this useful class of models to a wider audience of users.

Acknowledgements

sdmTMB would not be possible without the TMB (Kristensen *et al.* 2016) and INLA (Rue *et al.* 2009; Lindgren *et al.* 2011; Lindgren & Rue 2015) R packages. sdmTMB is heavily inspired by and in some places code has been adapted from both the VAST (Thorson 2019a) and glmmTMB (Brooks *et al.* 2017) R packages. Penalized spline support was possible thanks to mgcv (Wood 2017). We thank the authors of all these packages. We thank S. Kotwicki, M. Lindmark, C.C. Monnahan, P.M. Regular, J.T. Thorson, and J. Watson for helpful comments that substantially improved the manuscript.

Conflict of Interest statement

The authors declare no conflict of interest.

Author Contributions

SA led development of the software and EW contributed key software code. All authors contributed software code, guided software development, and participated in testing. PE led the Snowy Owl case study. All authors contributed to appendix code. SA led the writing of the manuscript; all authors contributed critically to the drafts and gave final approval for publication.

Supporting Information

Appendix 1: sdmTMB statistical model description
Appendix 2: Example of index standardization of fishery-independent survey data with sdmTMB
Appendix 3: Example of a spatially varying effect of climate on citizen science count data with sdmTMB
Appendix 4: Species distribution model comparison between INLA and sdmTMB
Appendix 5: Comparison of index standardization of survey data with VAST and sdmTMB
Appendix 6: sdmTMB speed testing methods and model validation description

Data Availability

A development version of sdmTMB is available at https://github.com/pbs-assess/sdmTMB. sdmTMB will be available on CRAN before publication. Paper source code is available at https://github.com/seananderson/sdmTMB-paper and will be archived on Zenodo before publication.

References

Aitchison, J. (1955). On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin. *Journal of the American Statistical Association*, **50**, 901.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Anderson, S.C. & Ward, E.J. (2019). Black swans in space: Modelling spatiotemporal processes with extremes. *Ecology*, **100**, e02403.

Bachl, F.E., Lindgren, F., Borchers, D.L. & Illian, J.B. (2019). inlabru: An R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, **10**, 760–766.

Bakka, H., Vanhatalo, J., Illian, J., Simpson, D. & Rue, H. (2019). Non-stationary Gaussian models with physical barriers. *arXiv:1608.03787 [stat]*. Retrieved from https://arxiv.org/abs/1608.03787

Banerjee, S., Gelfand, A.E., Finley, A.O. & Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **70**, 825–848.

Barnett, L.A.K., Ward, E.J. & Anderson, S.C. (2021). Improving estimates of species distribution change by incorporating local trends. *Ecography*, **44**, 427–439.

Barrowman, N. & Myers, R.A. (2000). Still more spawner-recruitment curves: The hockey stick and its generalizations. *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 665–676.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Blangiardo, M. & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with R-INLA*. John Wiley & Sons.

Bollhöfer, M., Eftekhari, A., Scheidegger, S. & Schenk, O. (2019). Large-scale sparse inverse covariance matrix estimation. *SIAM Journal on Scientific Computing*, **41**, A380–A401.

Breivik, O.N., Aanes, F., Søvik, G., Aglen, A., Mehl, S. & Johnsen, E. (2021). Predicting abundance indices in areas without coverage with a latent spatio-temporal Gaussian model (S. Kotwicki, Ed.). *ICES Journal of Marine Science*, fsab073.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M. & Bolker, B.M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **9**, 378–400.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Chilés, J.-P. & Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. John Wiley & Sons.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, N.A.C. & Wikle, C.K. (2011). *Statistics for spatio-temporal data*. Wiley, Hoboken, N.J.

Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Am. Stat. Assoc.*, **111**, 800–812.

Diggle, P.J. & Ribeiro, P.J. (2007). *Model-based geostatistics*. Springer.

Dinnage, R., Skeels, A. & Cardillo, M. (2020). Spatiophylogenetic modelling of extinction risk reveals evolutionary distinctiveness and brief flowering period as threats in a hotspot plant genus. *Proceedings of the Royal Society B: Biological Sciences*, **287**, 20192817.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kühn, I., Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B.,

Schurr, F.M. & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography*, **30**, 609–628.

Dunn, P.K. & Smyth, G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244.

English, P.A., Ward, E.J., Rooper, C.N., Forrest, R.E., Rogers, L.A., Hunter, K.L., Edwards, A.M., Connors, B.M. & Anderson, S.C. (2022). Contrasting climate velocity impacts in warm and cool locations show that effects of marine warming are worse in already warmer temperate waters. *Fish and Fisheries*, **23**, 239–255.

Finley, A.O., Datta, A. & Banerjee, S. (2021). spNNGP R package for Nearest Neighbor Gaussian Process models. *ArXiv200109111 Stat*. Retrieved from https://arxiv.org/abs/2001.09111

Fuglstad, G.-A., Simpson, D., Lindgren, F. & Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, **114**, 445–452.

Gay, D.M. (1990). Usage summary for selected optimization routines. *Computing Science Technical Report*, **153**, 1–21.

Gelfand, A.E. & Banerjee, S. (2017). Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and Its Application*, **4**, 245–266.

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. Retrieved from https://CRAN.R-project.org/package=DHARMa

Haskard, K.A. (2007). *An anisotropic Matern spatial covariance model: REML estimation and properties*. PhD thesis, The University of Adelaide.

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D.W., Sun, F. & Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol.*, **24**, 398–425.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B.M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, **70**, 1–21.

Latimer, A.M., Banerjee, S., Sang Jr, H., Mosher, E.S. & Silander Jr, J.A. (2009). Hierarchical models facilitate spatial analysis of large data sets: A case study on invasive plant species in the northeastern United States. *Ecol. Lett.*, **12**, 144–154.

Legendre, P. & Fortin, M.J. (1989). Spatial pattern and ecological analysis. *Vegetatio*, **80**, 107–138.

Liang, H., Wu, H. & Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773–778.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, **2**, 18–22.

Lindgren, F. & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, **63**, 1–25.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **73**, 423–498.

Matérn, B. (1960). *Spatial Variation*. Springer, New York.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models*. Chapman & Hall, New York.

Miller, D.L., Glennie, R. & Seaton, A.E. (2019). Understanding the Stochastic Partial Differential Equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*.

Millon, A., Petty, S.J., Little, B., Gimenez, O., Cornulier, T. & Lambin, X. (2014). Dampening prey cycle overrides the impact of climate change on predator population dynamics: A long-term demographic study on tawny owls. *Global Change Biology*, **20**, 1770–1781. Retrieved November 25, 2021,

Monnahan, C.C. & Kristensen, K. (2018). No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PLOS ONE*, **13**.

Monnahan, C.C., Thorson, J.T., Kotwicki, S., Lauffenburger, N., Ianelli, J.N. & Punt, A.E. (2021). Incorporating vertical distribution in index standardization accounts for spatiotemporal availability to acoustic and bottom trawl gear for semi-pelagic species. *ICES Journal of Marine Science*, fsab085.

National Audubon Society. (2021). The Christmas bird count historical results. Retrieved from www.christmasbirdcount.org

Osgood-Zimmerman, A. & Wakefield, J. (2021). A statistical introduction to Template Model Builder: A flexible tool for spatial modeling. *ArXiv210309929 Stat*. Retrieved from https://arxiv.org/abs/2103.09929

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, **10**, 439–446. Retrieved from https://doi.org/10.32614/RJ-2018-009

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Robillard, A., Therrien, J.F., Gauthier, G., Clark, K.M. & Bêty, J. (2016). Pulsed resources at tundra breeding sites affect winter irruptions at temperate latitudes of a top predator, the snowy owl. *Oecologia*, **181**, 423–433. Retrieved November 3, 2021, from http://link.springer.com/10.1007/s00442-016-3588-3

Rossi, R.E., Mulla, D.J., Journel, A.G. & Franz, E.H. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecological Monographs*, **62**, 277–314.

Rousset, F. & Ferdy, J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, **37**, 781–790.

Rue, H. & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. CRC press.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 319–392.

Shelton, A.O., Thorson, J.T., Ward, E.J. & Feist, B.E. (2014). Spatial semiparametric models improve estimates of species abundance and distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, **71**, 1655–1666.

Simpson, D., Rue, H., Riebler, A., Martins, T.G. & Sørbye, S.H. (2017). Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, **32**, 1–28.

Smith, J.Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting*, **4**, 283–291.

Sofaer, H.R., Jarnevich, C.S., Pearse, I.S., Smyth, R.L., Auer, S., Cook, G.L., Edwards Jr, T.C., Guala, G.F., Howard, T.G., Morisette, J.T. & others. (2019). Development and delivery of species distribution models to inform decision-making. *BioScience*, **69**, 544–557.

Thorson, J.T. (2019a). Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research*, **210**, 143–161.

Thorson, J.T. (2019b). Measuring the impact of oceanographic indices on species distribution shifts: The spatially varying effect of cold-pool extent in the eastern Bering Sea. *Limnology and Oceanography*, **64**, 2632–2645.

Thorson, J.T., Cheng, W., Hermann, A.J., Ianelli, J.N., Litzow, M.A., O'Leary, C.A. & Thompson, G.G. (2020). Empirical orthogonal function regression: Linking population biology to spatial varying environmental conditions using climate projections. *Global Change Biology*, **26**, 4638–4649.

Thorson, J.T., Ianelli, J.N., Larsen, E.A., Ries, L., Scheuerell, M.D., Szuwalski, C. & Zipkin, E.F. (2016a). Joint dynamic species distribution models: A tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, **25**, 1144–1158. Retrieved December 8, 2021,

Thorson, J.T., Pinsky, M.L. & Ward, E.J. (2016b). Model-based inference for estimating shifts in species distribution, area occupied and centre of gravity. *Methods in Ecology and Evolution*, **7**, 990–1002.

Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, **6**, 627–637.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K. & Nielsen, A. (2017). Validation of ecological state space models using the Laplace approximation. *Environmental and Ecological Statistics*, **24**, 317–339.

Tilman, D., Kareiva, P.M. & others. (1997). *Spatial ecology: The role of space in population dynamics and interspecific interactions*. Princeton University Press.

Vehtari, A., Gelman, A. & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–1432.

Ver Hoef, J.M., Peterson, E.E., Hooten, M.B., Hanks, E.M. & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, **88**, 36–59.

Wall, M.M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121**, 311–324.

Wood, S.N. (2017). *Generalized additive models: An introduction with R*, 2nd edn. Chapman and Hall/CRC.

Xianyi, Z. & Kroeker, M. (2021). OpenBLAS. *http://www.openblas.net*.