Ecography

ECOG-04707
Brodie, S., Thorson, J. T., Carroll, G., Hazen, E. I., Bograd, S., Haltuch, M., Holsman, K., Kotwicki, S., Samhouri, J., Willis-Norton, E. and Selden, R. 2019. Trade-offs in covariate selection for species distribution models: a methodological comparison. – Ecography doi: 10.1111/ecog.04707

Supplementary material

Appendix 1

Material and methods

Simulated data

Spatially-explicit abundance of a hypothetical species was simulated over 20 years in a 20° x 20° gridded area with 1° x 1° grid cells, and used to fit three types of species distribution models. Data were simulated using two major steps. First, habitat suitability grid layers (1° resolution) were created for each year with four sources of variability: environmental, spatial, temporal, and spatiotemporal (Figure 1), where spatial, temporal, and spatiotemporal variability represent proxies for unmeasured ecological processes. Habitat suitability ranged from 0 to 1, with highest habitat suitability being 1. Second, these habitat grid layers were used to inform species occurrence and abundance at each grid cell for each year.

Habitat suitability grid layers integrated the four sources of variability using five covariates. The relationship between each covariate and suitability was pre-defined as a function (Table A1; Figure A1) using the *virtualspecies* package (Leroy et al., 2016) in the R statistical computing software (Table A1; Figure A1; R Core Team, 2018). Environmental habitat suitability included two covariates, temperature (°C) and topography (m). Temperature preference was parameterized using a normal distribution with a mean of 6 °C, while topographic preference was parameterized using a logistic distribution to approximate a linear relationship (Table A1; Figure A1). Temperature had an additional temporal trend (spatially constant), where temperature gradually increased by 2°C over the 20-year simulation period, with the species' preference remaining constant. Spatial habitat suitability included two covariates, latitude and longitude, with spatial preference parameterized as a normal distribution with a mean of 10° (Table A1; Figure A1). Spatiotemporal habitat suitability was considered as a latent factor, where

the spatiotemporal process plays a role in species distribution and abundance but the process is not directly observed or measured. This latent spatiotemporal had three primary modes of high, medium, and low extent, and was parameterized to alternate between three modes over the 20-year simulation period (Figure a1). Random noise was added to each grid cell of the spatially-varying habitat suitability layers (temperature, topography, latitude, longitude, spatiotemporal). Noise was randomly sampled from a uniform distribution, with the lower and upper limits of the uniform distribution specified for each habitat suitability layer (*jitter* R function; Table A1). Adding noise decreased intra-year correlations between temperature and latitude, and topography and longitude (Pearson's correlation < 0.6; Table A1). Each habitat suitability layer (n=3: environmental, spatial, spatiotemporal) was then combined as a weighted average (environment weighting was doubled), to create a final habitat suitability layer for each year.

Annual habitat suitability was then used to inform the probability of species occurrence and abundance. For each grid cell in the area, habitat suitability was converted to species occurrence (0 or 1) using a logistic function as part of the *convertToPA* function in the "virtualspecies" R package. The logistic function has a defined threshold of 0.5 (i.e. inflection point) and a defined alpha of -0.05 that controls the steepness of the curve (Table A1), and ultimately influences the probability of habitat suitability being converted to a presence (1) or absence (0). Species abundance in each grid cell was generated as a function of occurrence, where abundance was zero if the species was absent. If the species was present, then abundance was estimated from a log-normal distribution with log-mean two and log-standard deviation of 0.1 (Table A1; Figure A1). Abundance in each grid cell was then multiplied by the habitat suitability value in the same grid cell, to provide a habitat-informed abundance. Simulated data were generated for each grid cell (n=400) once per year, for 20 years, resulting in 8000 data points. 400 samples per year is within the range that is obtained by fishery resource surveys for many regions

worldwide, including the NOAA eastern Bering Sea shelf bottom trawl survey (~375 samples per year) and the West Coast shelf-slope survey (~650 samples per year).

Model parameterization

Below we briefly outline the specific parameterization details for each of the three modeling approaches. Where possible, we implemented default or commonly used model settings for simplicity (Table A2).

BRTs were implemented using the *dismo* R package (Elith et al., 2008). The occurrence component of the delta model was modeled using a Bernoulli distribution, and the abundance component modeled using a gaussian distribution, with a logged non-zero response variable. The BRTs had a tree complexity of 3, a bag fraction of 0.6, and a learning rate small enough to ensure >1000 trees during the fitting process (Elith et al., 2008). All covariates for configuration one (latitude, longitude, year), configuration two (temperature, topography, year), and configuration three (latitude, longitude, temperature, topography, year) were simply added as individual covariates, with tree complexity internally creating interactions (Table A2). BRT models were optimized using the 'gbm.simplify' function to determine if any covariates should be omitted to optimize model fit. Fitted data for each grid cell from both delta components were multiplied to provide BRT estimates of species abundance. Each BRT model was then fit 10 times and the standard deviation of predicted species abundance across the 10 models was determined to provide an estimate of model error (Hazen et al. 2018).

GAMs were implemented using the *mgcv* R package. The occurrence component of the delta model was modeled using a binomial distribution, and the abundance component modeled using a gaussian distribution, with a logged non-zero response variable. GAMs included spatial covariates as a tensor

gaussian process smooth between latitude and longitude, spatiotemporal covariates as a tensor gaussian process smooth between latitude and longitude by year, and temporal covariate year as a factor (Table A2). Environmental covariates (temperature and topography) were included as an *s*-type gaussian process smooth, with the number of knots not pre-specified. Fitted data from both delta components were multiplied to provide GAM estimates of species abundance.

VAST models were implemented using the *VAST* R package (https://github.org/james-thorson/VAST). VAST implements a hierarchical approach to spatiotemporal models and is widely used for determining spatially-explicit biomass estimates in fisheries research and management (Thorson and Barnett, 2017; Thorson, 2019; Xu et al., 2019). The occurrence component of the delta model was modeled using a binomial distribution, and the abundance component modeled using a lognormal distribution. Year was considered a fixed effect, while model residuals were attributed to spatial and spatiotemporal variation and treated as random effects described by a three-dimensional Gaussian process with 400 locations (one for each of the 400 simulated grid cells). A pre-specified bias correction method, an argument within the VAST package, was applied to correct for re-transformation bias (Thorson and Kristensen, 2016). Environmental covariates (temperature and topography) were included as quadratic forms in the model to allow for non-linear responses (Thorson et al., 2017).

Table A1 Parameters, units, equations, and range limits used to simulate species occurrence and abundance. Minimum (maximum) limits for the uniform distribution used to add random noise are also shown.

Name	Description	Units	Range	Parameter 1	Parameter 2	Distribution	Noise limit
Latitude	Spatial covariate	Degrees	1-20	$\mu = 10$	$\sigma = 20$	Normal	-1 (1)
Longitude	Spatial covariate	Degrees	1-20	$\mu = 10$	$\sigma = 20$	Normal	-1 (1)
Temperature	Environmental covariate that increases by 2°C over the 20-year study period	°C	min = 0.1053Y + 1.8974 max = 0.1053Y + 5.8947	$\mu = 6$	$\sigma = 12$	Normal	-3.5 (3.5)
Topography	Environmental covariate	m	-3000 - 0	$\alpha = 100$	$\beta = 100$	Logistic	-65 (65)
Spatiotemporal	Latent spatiotemporal covariate		0-1				-50 (50)
Occurrence	Species occurrence as a function of habitat suitability	0 or 1		$\alpha = -0.05$	$\beta = 0.5$	Logistic	
Abundance	Abundance value if species considered present in a grid cell	kg grid cell^{-1}	3-8	$\log(\mu) = 2$	$\log(\sigma) = 0.1$	log-normal	

Table A2 Summary of model configurations and parameterization in R syntax. Model configurations include continuous spatiotemporal covariates latitude (lat), longitude (lon), and year, and environmental covariates temperature (temp) and topography (topo). For boosted regression trees (BRT), gbm.x indicates the argument used to input names of predictor variables. For Generalized Additive Models (GAM), *te* and *s* indicate smooth terms, bs=gp implements a gaussian process smooth. For BRTs and GAMs, the model parameterization is the same for both the occurrence and abundance forms of the delta model. In the vector-autoregressive spatiotemporal model (VAST), FieldConfig is the argument used to turn on (1) or off (0) spatial and spatiotemporal variation in the occurrence and abundance components of the delta models. For VAST, 400 spatial knots were used for the simulated data and 100 knots for each of the species data.

Configuration	**BRT**	**GAM**	**VAST**
Configuration 1:			
spatiotemporal	gbm.x = lat, lon, factor(year)	*te*(lat, lon, bs=gp) + *te*(lat, lon, by = year, bs=gp) + factor(year)	FieldConfig=c(1,1,1,1)
Configuration 2:			
environmental	gbm.x = temp, topo, factor(year)	*s*(temp, bs=gp) + *s*(topo, bs=gp) + factor(year)	FieldConfig=c(0,0,0,0), temp and temp ^{2}, topo and topo^2
Configuration 3:			
spatiotemporal & environmental	gbm.x = lat, lon, temp, topo, factor(year)	*te*(lat, lon, bs=gp) + *te*(lat, lon, by = year, bs=gp) + *s*(temp, bs=gp) + *s*(topo, bs=gp) + factor(year)	FieldConfig=c(1,1,1,1), temp and temp^2, topo and topo^2

Table A3 Parameter estimates and significance for the three delta (occurrence and abundance) model types (Boosted Regression Trees BRT; Generalized Additive Model GAM, Vector autoregressive spatiotemporal model VAST). % denotes the percent relative influence of each covariate in a BRT; μ is the mean and gives the average results for all years (for simplicity); * indicates that not all years were significant; *edf* is the estimated degrees of freedom in GAM models, with *te* and *s* denoting the smoother type used; σ is the standard deviation of the given process.

		Occurrence Models		Abundance Models	
Model	**Parameter**	**Estimate**	**Significance**	**Estimate**	**Significance**
BRT	year		9%		8%
	latitude		27%		21%
	longitude		23%		17%
	temperature		21%		21%
	depth		21%		32%
GAM	year	μ = 0.6	p<0.05*	μ = 0.07	p<0.001*
	te(lat,lon)	*edf* = 20	p<0.0001	*edf* = 24	p<0.0001
	te(lat,lon):year	*edf* = 13	p<0.0001	*edf* = 23	p<0.0001
	s(temperature)	*edf* = 7	p<0.0001	*edf* = 11	p<0.0001
	s(depth)	*edf* = 5	p<0.0001	*edf* = 7	p<0.0001
VAST	year	μ = 0.9	p>0.05	μ = 6	p<0.0001
	temperature	-0.01	p=0.84	-0.001	p=0.52
	temperature^2	-0.02	p=0.5	-0.002	p=0.17
	depth	-0.01	p=0.94	-0.015	p<0.05
	depth^2	0.03	p=0.64	0.015	p<0.001
	Spatial variation	σ = 3.4	p<0.0001	σ = 0.5	p<0.0001
	Spatiotemporal variation	σ = -0.8	p<0.0001	σ = -0.06	p<0.0001

Table A4 Summary of catch-per-unit-effort data for the three case study species. Mean and range of abundance CPUE, range of latitude (°N) and longitude (°W) when species is present, and the percent of trawls with species occurring (n=12866 trawls in total) are shown.

	Mean	**Range**	**Latitude Range**	**Longitude Range**	**Trawls**
Arrowtooth flounder	0.09	0 - 3.9	54.7 – 61.7	178.2 – 158.3	43%
Pacific cod	0.007	0 - 0.34	54.7 – 62	178.2 – 158.3	57%
Walleye pollock	1.06	0 - 75	54.6 – 63	178.2 – 158.3	90%

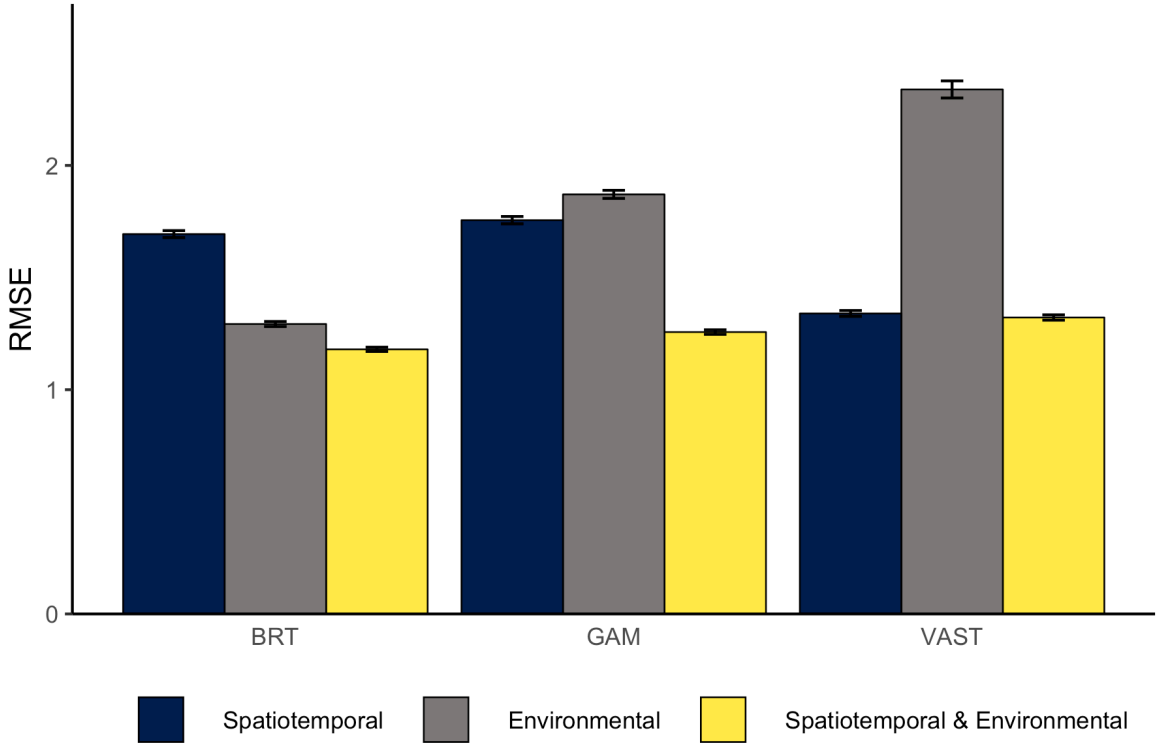

Figure A1 Mean of model root mean square error (RMSE) from ten simulation replicates (± standard error). Results are grouped by three model types, and colored by three covariate configurations.

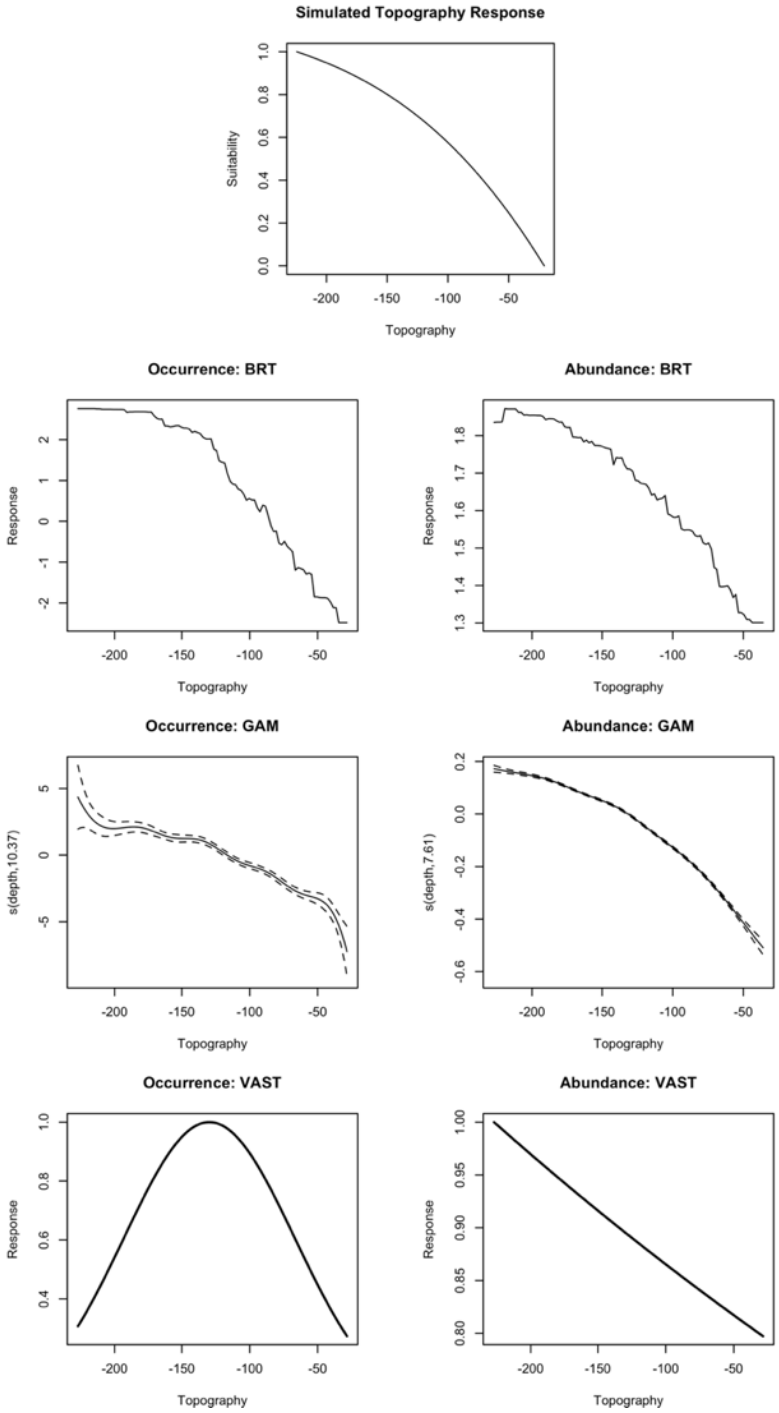

Figure A2 Topography preference curves for simulated data (top) and the three model types (Boosted Regression Trees BRT; Generalized Additive Model GAM; and Vector autoregressive spatiotemporal model VAST) using configuration three (spatiotemporal and environmental covariates). Models are delta models and so each occurrence and abundance component are presented.

Elith, J., Leathwick, J.R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4)**,** 802-813.

Hazen, E.L., Scales, K.L., Maxwell, S.M., Briscoe, D.K., Welch, H., Bograd, S.J., et al. (2018). A dynamic ocean management tool to reduce bycatch and support sustainable fisheries. *Science advances* 4(5)**,** eaar3001.

Leroy, B., Meynard, C.N., Bellard, C., and Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography* 39(6)**,** 599-607.

R Core Team (2018). "R: A language and environment for statistical computing. R Foundation for Statistical Computing". (Vienne, Austria).

Thorson, J.T. (2019). 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., and Barnett, L.A.K. (2017). Comparing estimates of abundance trends and distribution shifts using single-and multispecies models of fishes and biogenic habitat. *ICES Journal of Marine Science* 74(5)**,** 1311-1321.

Thorson, J.T., Ianelli, J.N., and Kotwicki, S. (2017). The relative influence of temperature and size-structure on fish distribution shifts: A case-study on Walleye pollock in the Bering Sea. *Fish and fisheries* 18(6)**,** 1073-1084.

Thorson, J.T., and Kristensen, K. (2016). Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fisheries research* 175**,** 66-74.

Xu, H., Lennert-Cody, C.E., Maunder, M.N., and Minte-Vera, C.V. (2019). Spatiotemporal dynamics of the dolphin-associated purse-seine fishery for yellowfin tuna (Thunnus albacares) in the eastern Pacific Ocean. *Fisheries Research* 213**,** 121-131.