

Which modelling choices are supported when marine survey data are limited? An evidence map of 95 studies and a benchmark on structured subsamples of a trawl survey

Authors. Eben Afrifa-Yamoah^{1,2*}, Alexandre C. Siqueira², Yaw Kwaafo Awuah-Mensah^{1,2},
Francky Fouedjio^{1,3}, Ute Mueller^{1,2}

¹ Mathematical Applications and Data Analytics, School of Science, Edith Cowan University,
Australia

² Centre for Marine Ecosystems Research, School of Science, Edith Cowan University, Australia

³ Kaplan Business School Pty Ltd, Data Science & Analytics, 10 William Street, Perth, WA
6000, Australia

* Corresponding author: e.afrifayamoah@ecu.edu.au; ORCID 0000-0003-1741-9249

Abstract

1. Marine management decisions increasingly rest on species distribution models fitted to sparse, clustered survey records, while the guidance for fitting them comes from data-rich settings. We asked which modelling choices have been tested under the conditions marine practitioners face, and whether the ordering that guidance assumes, spatial structure before algorithm choice, holds at the sample sizes at which it is applied.
2. We screened 330 records and retained 95 peer-reviewed marine applications that treated spatial autocorrelation explicitly, coding the contrasts tested, the sample sizes reported and the metric family. We then benchmarked, on the NWFSC West Coast Groundfish Bottom Trawl Survey, subsamples of three demersal species at 30 to 500 records crossed with dispersed and clustered coverage and with interpolation and extrapolation targets, fitting random forests and generalised additive models with and without spatial terms and a Gaussian Markov random field, evaluated against the held-out survey and by random and spatial block cross-validation.
3. No marine study crosses spatial treatment against algorithm family on the same data, and one reports performance below 100 records. In the benchmark, spatial structure helped or harmed according to how it was added: a coordinate smooth in a GAM reduced held-out AUC by 0.06 to 0.15 below 200 records and at every size under clustered sampling, a Matérn random field recovered most of that loss, and coordinates in a random forest helped only with dispersed coverage and an interpolation target. Dispersed coverage was worth as much as the gain from 30 to 500 records. Random cross-validation erred by up to 0.13

AUC in either direction according to coverage, and spatial blocking moved the estimate by less than 0.03. Likelihood models were overconfident at 30 to 50 records; a random forest with coordinates was nearly calibrated.

4. Spatial against non-spatial is not a contrast on which to base a recommendation; coverage is the first-order decision; internal validation does not reveal extrapolation loss when the sample is a band; and at the smallest sizes model choice affects the probability scale more than the ranking. The corpus and benchmark yield a practitioner guide graded by the evidence behind each element.

Keywords. marine conservation; sample size; spatial autocorrelation; spatial cross-validation; species distribution models; survey design

1 Introduction

Conservation decisions in marine systems are often made on thin evidence. Marine protected area boundaries, dynamic closures, climate refugia designations and spatial catch controls rest on predicted distributions, and those predictions are typically built from survey records that are sparse, clustered along shipping tracks or depth contours, irregular in time, and drawn from a three-dimensional environment sampled almost entirely at its surface. Sampling is expensive, weather-constrained and logistically bounded, so the density of observation that terrestrial biogeographers take for granted is rarely available (Robinson et al., 2017; Melo-Merino et al., 2020).

Species distribution models are fitted under these conditions anyway, and a large methodological literature offers guidance on how. The difficulty is where that guidance was generated. Comparative studies that establish the value of flexible algorithms typically use thousands of records (Elith et al., 2006), and the mechanism through which those algorithms earn their advantage, the estimation of interactions and unspecified functional forms, is the mechanism that fails first as sample size falls. Marine survey geometry adds a second layer: records lie along tracks rather than across grids, effort varies by station and year, detection varies by gear and depth, and the sampled surface is a thin slice of the habitat volume. Guidance established on dense, gridded terrestrial data is a hypothesis about marine survey data, not a finding.

We use spatial machine learning to denote approaches along the statistical learning spectrum (Hastie et al., 2009) that share an explicit representation of spatial dependence: algorithmic learners such as Random Forest and Boosted Regression Trees with spatial predictors (Breiman, 2001; Friedman, 2001; Hengl et al., 2018), Gaussian Markov random field methods such as VAST, sdmTMB and tinyVAST (Thorson, 2019b; Anderson et al., 2025; Thorson et al., 2025), and spatial generalised additive models (Wood, 2017; Miller et al., 2022). Practitioners choose among these as alternatives, and applied marine studies compare them. Deep learning is outside scope for the data volumes considered here (Kellenberger et al., 2026; Kapur et al., 2026).

Existing reviews cover marine species distribution modelling in general (Robinson et al., 2017; Pasanisi et al., 2024), machine learning in marine systems (Melo-Merino et al., 2020; Rubbens

et al., 2023) and methodological trends (Klaassen et al., 2025). None asks where the guidance practitioners rely on has been tested against the conditions those practitioners face. That is the question this paper takes up, in four parts: which contrasts between modelling choices have been tested on marine data, at what sample sizes, and with what validation design; what the tested evidence supports in direction, and where the marine magnitude is known; whether the ordering the guidance assumes holds when it is tested in marine data at the sample sizes it is applied to; and what a practitioner should do now, with each element of the guidance marked by the evidence behind it.

The distinction between the first two parts is the contribution. A review that asserts a principle and then acknowledges that its support is non-marine leaves the reader to guess how much weight the acknowledgement carries. An evidence map states the weight directly: which studies, which system, which sample sizes, which contrasts crossed. The map is more informative than the principles it qualifies, because it shows that the central claim of the applied literature, that spatial structure matters more than algorithm choice, had never been tested in the setting where it is most often invoked. Harmonised trawl survey resources (Maureaud et al., 2024; Ward et al., 2025) make the test feasible: sparse subsamples of known provenance can be drawn from a dense survey, competing models fitted to each, and every fit evaluated against the held-out remainder, so that true out-of-sample error is known and the optimism of any cross-validation scheme is measured rather than inferred. The second half of the paper reports that test, and the guidance in the final section is graded on both the corpus and the benchmark.

2 Methods

2.1 Search, screening and inclusion

We searched Web of Science, Google Scholar and discipline-specific journals for studies applying spatial modelling to marine species distributions, using Boolean combinations of method terms, validation terms, data limitation terms and marine ecosystem descriptors (Appendix S1). Searches returned 330 unique records; 172 passed title and abstract screening; 108 passed full-text assessment; 95 met all inclusion criteria after cross-database duplicates were resolved (Figure 1). Studies were included if they reported a peer-reviewed marine or estuarine application, addressed spatial autocorrelation explicitly in the model or its validation, and used an approach with an available implementation in R or Python. Purely terrestrial studies were excluded unless they supplied a methodological result with no marine equivalent; the eight retained under this exception are identified at every point of use and are listed with their reason for retention in Appendix S1.

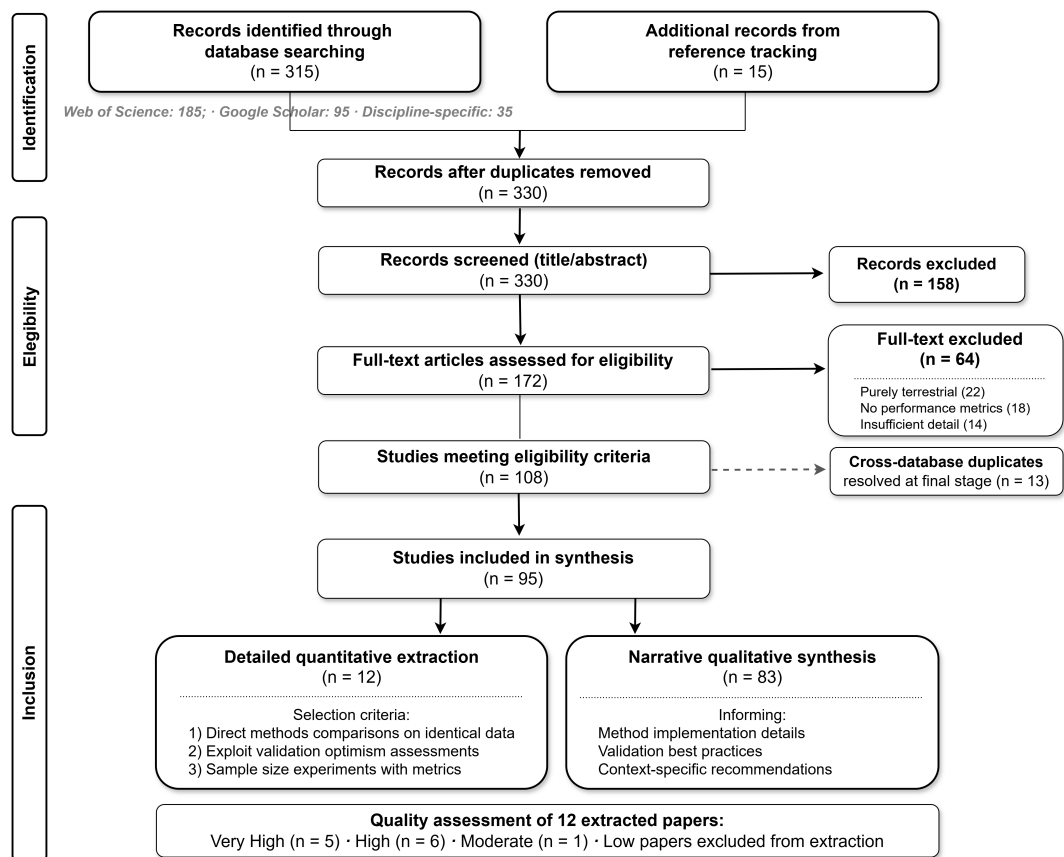


Figure 1. Identification, screening and inclusion of studies of spatial modelling applied to marine species distributions. Screening criteria are in Appendix S1. Peer-reviewed sources only; stock assessment working group reports and other grey literature were not searched (Section 4.4).

2.2 Evidence map coding

Each of the 95 studies was coded on five attributes. The contrast tested: whether the study compared spatial against non-spatial treatment within a model family, compared algorithm families holding spatial treatment constant, crossed both, varied sampling coverage or sample size experimentally, or compared validation designs. The study system: marine, estuarine, terrestrial, or simulated. The sample size range at which performance was reported, recorded as the minimum and maximum used in any performance estimate. The metric family: discrimination, error, correlation, index precision, or projection outcome. And whether the inference target of the validation, conditional or marginal in the sense of Section 3.2, was stated. A study contributing to more than one contrast was counted under each. Counts are lower bounds on corpus contribution, since a study may inform a finding without reporting the experiment that would test it.

Twelve studies were selected for detailed extraction on the basis that they reported a direct method comparison on identical data, an explicit assessment of validation optimism, a sample size experiment with performance metrics, or an authoritative performance benchmark for a specific method (Appendix S2). Selection preceded extraction and was not conditioned on the direction of the result. Quality was rated on sample adequacy, validation rigour, reporting clarity, statistical robustness and relevance to data-limited marine settings; ratings determined whether a study could anchor a numerical claim and were not used as weights.

2.3 Comparison protocol

Studies report performance in metrics that are not interconvertible. Discrimination measures are insensitive to calibration and depend on the extent from which absences are drawn (Lobo et al., 2008; Guillera-Aroita et al., 2015); error measures are scale-dependent; index precision is a property of a survey design as much as of a model. No transformation relates a difference in one to a difference in another. We therefore applied three rules. Within-study contrasts are reported with attribution to the source, its taxon, its system and the metric it used. No value is averaged, weighted or otherwise combined across studies. Where studies agree in the direction of a contrast but not in metric or magnitude, the direction is reported with the number of studies supporting it and no magnitude is attached. Every numerical value drawn from the corpus appears in a table of quantitative values in Section 3.2 with its source and its scope of generalisation.

2.4 Evidence grades

Each element of the practitioner guidance in Section 3.5 carries one of three grades. Grade A: a within-study experiment on marine data at sample sizes below 200 observations, which before this study meant one source (Luan et al., 2020) and now also means the benchmark of Section 3.4, marked this study wherever it is the basis. Grade B: a within-study experiment elsewhere, on marine data at larger sample sizes or on non-marine or simulated data. Grade C: methods guidance, statistical theory or author judgement informed by the corpus, without a

within-study experiment. The grades describe the evidence, not the confidence we hold; several grade C elements follow from mechanisms that are not in doubt. Their purpose is to let a reader see at once which parts of the guidance rest on marine experiments at the relevant sample size. Where the benchmark tested an element and the result qualified rather than confirmed it, the guidance says so.

2.5 Benchmark design

Survey data. We used the surveyjoin database (Ward et al., 2025), which harmonises bottom trawl surveys of the Northeast Pacific with standardised species codes, effort and haul-level covariates, restricted to the NWFSC West Coast Groundfish Bottom Trawl Survey (NWFSC.Combo) for 2003 to 2023 excluding 2020, a single annual survey with one design and one gear. Satisfactory hauls with position, effort and bottom temperature numbered 12,414 after 258 lacking bottom temperature were dropped. Hauls record catch by species, so absences are confirmed rather than inferred. Covariates were depth, bottom temperature and year, entered as smooth terms in the additive and random field models and as numeric predictors in the tree-based models; the set is deliberately small, since the study concerns the treatment of space and sample size rather than covariate selection.

Species. Three species spanning prevalence were selected by a rule fixed before any model was fitted: the highest prevalence candidate above 40 percent of hauls, the highest between 10 and 40 percent, and the highest between 2 and 10 percent, from six common West Coast groundfish. The rule selected Dover sole (*Microstomus pacificus*, prevalence 0.84), lingcod (*Ophiodon elongatus*, 0.33) and canary rockfish (*Sebastes pinniger*, 0.08). Responses are presence-absence by haul.

Subsampling design. The survey was split once for each species into a training pool and a held-out evaluation set, in two ways corresponding to the two inference targets of Table 3. For the conditional target the evaluation set was a random 30 percent of hauls across the whole domain, so that evaluation locations lie within the spatial and environmental envelope of the pool. For the marginal target the evaluation set was the northern 30 percent of hauls by projected northing, so that every training subsample lies south of the region to be predicted and evaluation requires projection along the survey’s dominant latitudinal gradient. From the training pool we drew subsamples at $n = 30, 50, 100, 200$ and 500 hauls, crossed with two coverage designs. Dispersed subsamples were stratified across a 25 km grid so that they span the pool. Clustered subsamples were drawn at random inside a latitudinal band one-tenth of the pool’s north-to-south extent, placed at random; a fixed band width holds coverage constant across sample sizes, whereas nearest neighbour clustering collapses to a single station at $n = 30$ and expands with n , confounding the two factors. The realised extent of each subsample as a fraction of the pool’s bounding box was recorded (0.85 to 0.99 for dispersed and 0.01 to 0.04 for clustered draws) and is the quantity that operationalises coverage. Figure 2 maps the partitions and example subsamples; Appendix S6 maps one example for every cell. Subsamples with fewer than five presences or five absences were redrawn up to 50 times. The minimum has a consequence the reader should hold in mind: at $n = 30$ it conditions every retained subsample of a species with 8 percent prevalence on a draw

with at least 17 percent prevalence, so small n results for the rarest species describe favourable draws rather than typical ones. Each cell was replicated 10 times with independent draws, giving 600 scheduled replicate runs and 15,000 primary fits; the number of replicates was set by the compute available and its consequence for precision is stated in Section 4.6.

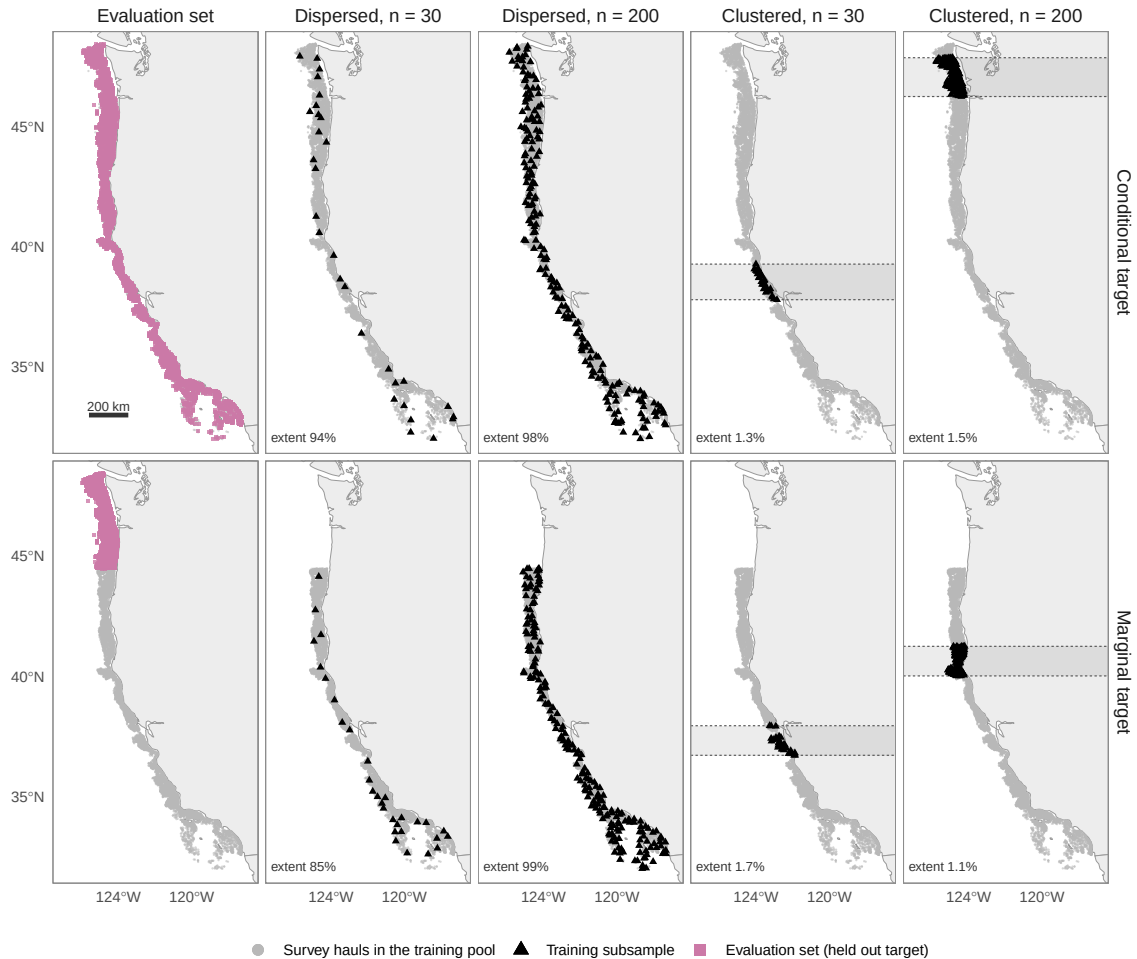


Figure 2. Sampling designs of the benchmark, drawn from the NWFSC West Coast Groundfish Bottom Trawl Survey (2003 to 2023, 12,414 hauls). Rows are the two inference targets. For the conditional target a random 30 percent of hauls is held out as the evaluation set and the remaining 70 percent form the training pool; for the marginal target the northern 30 percent of hauls by projected northing is held out, so every training subsample lies south of the evaluation region. Columns show the evaluation set and one example training subsample for each coverage design at $n = 30$ and $n = 200$. Dispersed subsamples are stratified over a 25 km grid so that they span the pool; clustered subsamples are drawn at random inside a latitudinal band one-tenth of the pool’s north-to-south extent, placed at random (band limits dashed and shaded). Extent is the subsample’s bounding box as a percentage of the pool’s. Grey points are pool hauls not drawn. Species: Dover sole; draws are examples and change with the replicate seed.

Models. Five models were fitted to every subsample, forming a two-by-two contrast of spatial treatment against algorithm family plus a geostatistical reference: a random forest on covariates only (ranger; Wright and Ziegler, 2017; 500 trees, minimum node size 5); a random forest with projected coordinates and distance to shore added as predictors, following the coordinate representation guidance for tree-based learners (Hengl et al., 2018); a binomial generalised additive model with thin-plate smooths of the covariates (mgcv; Wood, 2017; REML; basis dimension scaled to n); the same with a two-dimensional thin-plate smooth of the coordinates

whose basis dimension was capped relative to n so that it is estimable at $n = 30$; and a binomial Gaussian Markov random field model with covariate smooths and a Matérn spatial field on an SPDE mesh (sdmTMB; Anderson et al., 2025), with mesh cutoff set to one eighth of the subsample extent and a 5 km minimum so that mesh nodes remain fewer than records at every sample size. Settings were fixed in advance and held constant across cells, since tuning at $n = 30$ would introduce variance the design cannot separate. All models used identical covariates within a subsample.

Evaluation and validation. Each fit produced predictions on the held-out evaluation set for its target, giving true out-of-sample performance, measured by the area under the ROC curve, the Brier score and the calibration slope from a logistic recalibration on the evaluation set. Within each subsample we also ran the two cross-validation schemes a practitioner would: random five-fold, and spatial block five-fold (blockCV; Valavi et al., 2019) with block side set to the residual autocorrelation range of a covariate-only GAM fitted to the subsample, capped at the subsample extent divided by the ceiling of the square root of the number of folds, with the cap recorded. Optimism for a scheme is the cross-validated estimate minus the held-out value, positive when the scheme overstates accuracy. The block side estimate is unstable at these sample sizes (Appendix S6); under clustered coverage it was 3 to 25 km, smaller than the spacing between hauls, so the spatial partition is in effect a random one. This is reported as a property of the recommended procedure rather than corrected, because it is what a practitioner applying that procedure to a small clustered sample obtains; its consequences are examined in Section 3.4.3.

Analysis. The estimand for every contrast is the paired difference in a metric between two models fitted to the same subsample and evaluated on the same held-out set, or between the two coverage designs at the same species, replicate, sample size and target; pairing removes the between-subsample variance that dominates at small n . For each cell we report the mean paired difference, a t-based 95 percent interval over the pooled replicates of the three species, and the proportion of replicates in which the difference is positive, which is the robustness statistic a practitioner choosing between two methods on one dataset needs. Five contrasts answer the questions: the spatial structure effect within each family (random forest with coordinates minus random forest; GAM with coordinate smooth minus GAM; random field model minus GAM with coordinate smooth), the algorithm effect at each spatial treatment (random forest minus GAM, with and without spatial terms), the coverage effect (dispersed minus clustered, by model), and optimism by scheme. A linear mixed model with fixed effects for model, sample size, coverage and target and random intercepts for species and for replicate within species was fitted as a check; its constant residual variance is contradicted by the data (replicate standard deviation of a paired difference about 0.02 at $n = 500$ and 0.15 at $n = 30$), so its intervals are conservative and it is reported in Appendix S7 rather than as the primary inference. Failure modes were tabulated by cell: fits returning an error or no finite prediction, degenerate predictions (standard deviation of predicted probability below 0.001 across the evaluation set), and the package sanity flag for the random field model. Performance summaries exclude failed fits and are therefore conditional on obtaining a usable fit, with the failure rate reported alongside.

228 *Reproducibility.* All analyses were run in R 4.3.3 with sdmTMB 1.1.0, blockCV 4.0.0, ranger, mgcv
 229 and surveyjoin 0.0.5 on a 2 core Linux workstation (2 h 18 min wall time for the full design with two
 230 workers). Replicate seeds are indexed by design row, so any cell can be rerun or extended to more
 231 replicates without changing the draws reported here. The subsampling seeds, the cell-by-replicate
 232 result table, the scripts, the figure sources and the evidence map coding sheet are archived at
 233 <https://github.com/EAfrifa-Yamoah/Which-modelling-choices-are-supported>.

234 3 Results

235 3.1 The evidence map

236 Table 1 gives the map and Figure 3 draws it as a matrix. Three features of it organise everything
 237 that follows.

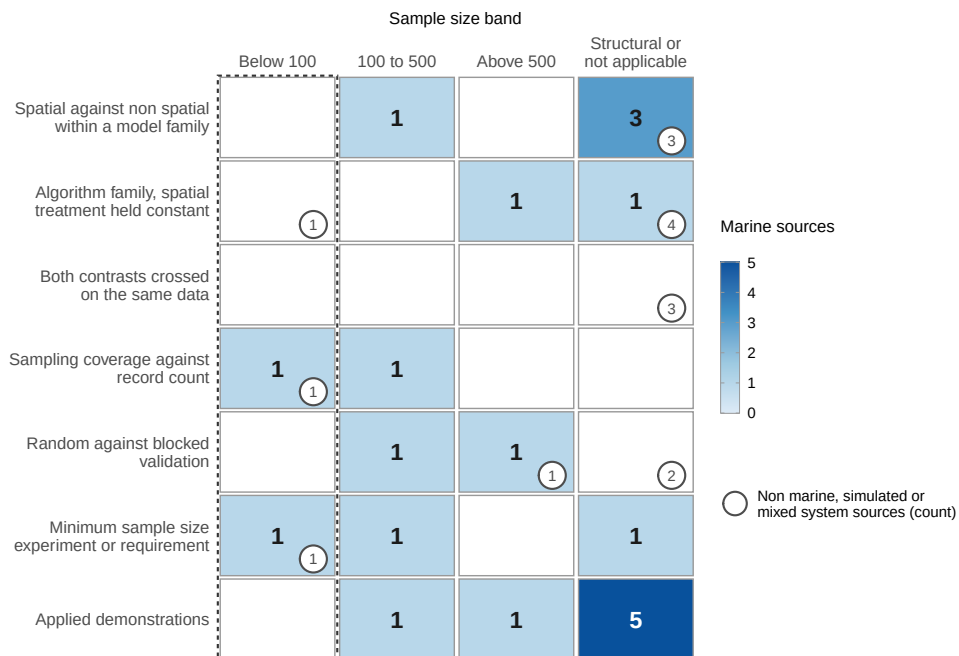


Figure 3. The evidence map as a matrix, generated from the deposited coding sheet. Rows are the contrasts of Table 1; columns are sample size bands (below 100, 100 to 500, above 500, structural or not applicable). A study contributes to every band its reported sample size range spans, and to every contrast it tests, so a study can occupy more than one cell. Cell fill and the central number give the count of marine sources on a single sequential colour-blind-safe ramp, with zero drawn as white; grey outlined markers give the count of non-marine, simulated and mixed system sources. The dashed frame marks the column below 100 records. The empty cells at the intersection of the first three rows, marine systems and fewer than 100 records are the figure's content.

Table 1. Evidence map. For each contrast between modelling choices, the number of corpus sources testing it, the number that are marine, the number reporting performance below 100 observations, the number that are both marine and below 100 observations, and the number that cross the contrast with a second contrast on the same data. Counts are sources cited in support of the finding in this paper and are lower bounds on corpus contribution. Non-marine and simulated sources are listed in Appendix S1.

Contrast	Sources	Marine	n < 100 reported	Marine and n < 100	Crossed with second contrast	Metric families	Note
Spatial against non-spatial treatment within a model family	6	3	0	0	3 (all non-marine)	R ² , error, variance explained, correlation	Marine sources vary spatial structure within one family only
Algorithm family, holding spatial treatment constant	5	1	1	0	3 (all non-marine)	AUC, rank correlation, R ² , error	Marine source is Becker et al. (2020) at n = 4,713
Both contrasts crossed on the same data	3	0	0	0	3	R ² , error	Hengl et al. (2018), Talebi et al. (2022), Mushagalusa et al. (2024)
Sampling coverage against record count	6	4	2	1	0	Performance stability, AUC	Marine experiment: Luan et al. (2020), 40 to 120 records
Random against blocked validation	4	3	0	0	1 (simulated)	AUC, conceptual	Marine quantification: Becker et al. (2020), one comparison
Minimum sample size experiment	3	1	2	1	0	AUC, performance stability, structural guidance	Thorson (2019b) supplies a structural requirement rather than an experiment
Applied demonstrations informing the decision framework	15	15	0	0	0	Mixed	Rich in single-system demonstrations; no controlled comparison

The contrast the applied literature invokes most often is the least tested. Six sources in the corpus bear on whether spatial treatment matters more than algorithm family; three of them cross the two contrasts on the same data, and all three are non-marine (Hengl et al., 2018, soil; Talebi et al., 2022, geoscience; Mushagalusa et al., 2024, simulated Gaussian process data). The three marine sources (Brodie et al., 2020; Grüss and Thorson, 2019; Grüss et al., 2021) vary spatial structure within a single model family and do not compare families. No marine study in the corpus fits spatial and non-spatial variants of two algorithm families on the same data. The ordering of the two effects had therefore never been measured in a marine system.

Performance below 100 observations has been reported in two studies, one marine. Luan et al. (2020) evaluated Random Forest across 40 to 120 demersal fish records under different sampling designs; Mi et al. (2017) evaluated Random Forest, MaxEnt and Boosted Regression Trees on 33 to 75 records of wetland cranes. Neither crosses spatial treatment against algorithm family, and Luan et al. (2020) does not vary the algorithm. The corpus contains no marine comparison of algorithm families at the sample sizes for which the guidance in Section 3.4 is intended. Every other marine method comparison in the corpus uses hundreds to thousands of records (Becker et al., 2020, $n = 4,713$; Thorson et al., 2015, 28 species across a full survey series; Zelli et al., 2025, $n = 1,247$).

Validation optimism, the inflation of apparent accuracy when random partitioning is applied to autocorrelated data, is quantified within-study in marine data by one source, at one sample size. Becker et al. (2020) report a difference of 0.11 in the area under the curve between k-fold and leave-one-year-out partitioning for cetacean models in the California Current, from 4,713 sightings. The reference demonstration is terrestrial (Ploton et al., 2020, 11.8 million forest biomass observations) and the controlled measurement is simulated (Mushagalusa et al., 2024). How optimism scales with sample size, coverage and the strength of residual autocorrelation in marine survey data is not reported anywhere in the corpus, and that scaling is the quantity a practitioner with 60 records needs. Section 3.4 supplies it for one survey.

Two further patterns are visible in the map. The inference target of the validation is stated in a minority of studies, so that a reported performance value often cannot be assigned to conditional or marginal prediction (Section 3.2). And the applied marine literature, which supplies fifteen of the sources for the decision framework in Section 3.5, is rich in demonstrations of what geostatistical and tree-based methods can do in specific systems and almost empty of controlled comparisons between them.

3.2 What the corpus supports in direction

Within the limits set by the map, three findings recur across the corpus. Each is reported as a direction with its supporting studies and the sample size at which the marine evidence, where it exists, was generated; Table 2 collects every quantitative value cited from the corpus, with its source and scope.

Spatial structure improves prediction more consistently than algorithm choice. Across the studies

that report either contrast, the difference between spatially aware and spatially naive versions of the same family is more consistent than the difference between families holding spatial treatment constant. The direction is uniform among the three crossed studies. The most cited of them is also the most often misread. Hengl et al. (2018) compared Random Forest with buffer distance predictors against ordinary kriging on benchmark data, and kriging retained an advantage, reaching an R^2 of approximately 0.60 against approximately 0.41 for the spatial Random Forest on the Meuse dataset. The contribution of that study is not that the algorithmic method wins. It is that supplying spatial features to a learner with no other access to spatial dependence moves its performance into the regime occupied by geostatistical methods, closing most of a gap that the standard formulation leaves open. Talebi et al. (2022) and Mushagalusa et al. (2024) report the same ordering on error metrics. Marine support is indirect: Brodie et al. (2020) show that covariate choices including the representation of spatial and dynamic habitat structure alter predictive performance more than the choice among comparable formulations, and Grüss and Thorson (2019) and Grüss et al. (2021) show that spatio-temporal random fields recover distributional structure that non-spatial formulations of the same data do not. Evidence grade from the corpus: B for the crossed contrast and for the within-family marine contrast; A after Section 3.4, where the crossed contrast is measured in marine data below 100 records.

Coverage constrains reliability more than record count. Luan et al. (2020) provide the only marine experiment, finding that spatially distributed samples outperform clustered samples of equal or larger size across 40 to 120 demersal fish records. Mi et al. (2017) reach a compatible conclusion from wetland cranes, obtaining discrimination between 0.82 and 0.89 from 33 to 75 records where habitat associations were strong and sampling spanned the environmental gradient. The mechanism is representation of environmental space: clustered samples leave regions of it unobserved, so prediction there is extrapolation, and adding clustered records does not repair that. Robinson et al. (2011) and Canonico et al. (2019) supply the marine survey design context. This is the one finding for which a marine experiment existed at the relevant sample size before the present study. Evidence grade: A, from one study in the corpus and from Section 3.4.

Validation must match the inference target. Random partitioning gives an approximately unbiased estimate of error for conditional prediction, meaning prediction at unsampled locations within the sampled spatial and environmental envelope, because test cases resemble the operational prediction cases. For marginal prediction, meaning projection into locations, periods or conditions not represented in training, random partitioning is optimistic, since autocorrelation places near neighbours of every training point in the test set (Roberts et al., 2017; Ploton et al., 2020; Valavi et al., 2019). Becker et al. (2020) quantify this in marine data once, as above. The corrective runs in both directions: spatial blocking is not universally required, and the reflex that it is always correct is as imprecise as the practice it displaced. Table 3 maps inference targets to partitioning schemes. Evidence grade from the corpus: B for the direction; C for the target-specific mapping. Section 3.4 measures the optimism under both targets below 500 records and raises the conditional and spatial marginal rows of Table 3 to grade A.

Table 2. Quantitative values in this paper, with source and scope. Every value is within-study or reported directly by the named source; none is pooled.

Claim	Value as reported	Metric	Source	Scope
Random partitioning inflates apparent accuracy relative to temporally blocked partitioning	Difference of 0.11	Area under the ROC curve	Becker et al. (2020)	Cetaceans, California Current, n = 4,713, one comparison; direction corroborated widely, magnitude specific to this study
Algorithmic and additive models differ modestly and metric dependently	0.85 against 0.83; density surface GAM better on rank correlation in the same study	AUC; rank correlation	Becker et al. (2020)	Same study; illustrates metric dependence of the apparent winner
Geostatistical models improve index precision over design-based estimators	About 20% reduction in standard errors	Standard error of abundance index	Thorson et al. (2015)	West Coast groundfish; magnitude varies with survey design
Forecast skill decays with horizon	r = 0.6 to 0.7 at one year; 0.3 to 0.4 at five years	Correlation	Thorson (2019a)	Eastern Bering Sea retrospective experiment
Environmental forcing explains a substantial share of distributional variability	About 40% attributed to cold pool extent; cross-validation correlation 0.72	Variance explained; correlation	Grüss et al. (2021)	Eastern Bering Sea pollock
Geostatistical benchmarks are not uniformly beaten by spatial tree ensembles	R ² about 0.60 kriging against about 0.41 spatial Random Forest	R ²	Hengl et al. (2018)	Meuse benchmark, terrestrial; cited for mechanism
Small-sample classification is workable	AUC 0.82 to 0.89 from 33 to 75 observations	Area under the ROC curve	Mi et al. (2017)	Cranes, wetland, strong habitat associations
Distributed sampling outperforms clustered sampling of equal or greater size	Reported across 40 to 120 observations	Performance stability	Luan et al. (2020)	Demersal fish; direction robust, magnitude study-specific
Climate projection reduces suitable habitat	About 23% retained under high emissions, from 1,247 observations	Projected habitat retention	Zelli et al. (2025)	Cold-water corals and sponges, New Zealand
Climate projection reduces suitable habitat	7.9% to 37.4% loss by 2100 across scenarios	Projected habitat loss	Mestre-Tomás et al. (2025)	Yellowfin tuna

Table 3. Matching validation design to inference target. Grade A for the conditional and spatial marginal rows (this study, Section 3.4.3; Becker et al., 2020 for the direction at large n); grade C for the temporal and environmental rows (Roberts et al., 2017; Valavi et al., 2019).

Inference target	Prediction task	Appropriate partitioning	Consequence of random partitioning
Conditional	Prediction at unsampled locations within the sampled spatial and environmental envelope	Random k-fold is defensible	Approximately unbiased for the relevant task
Marginal, spatial	Projection to areas outside the surveyed domain	Spatial block partitioning at a block size exceeding the residual autocorrelation range	Optimistic; near neighbours leak into the test set
Marginal, temporal	Forecasting to future years or unsurveyed periods	Leave-one-year-out or forward chaining	Optimistic; within-year structure leaks into the test set
Marginal, environmental	Projection under changed conditions, including climate scenarios	Partition by environmental strata; report extrapolation diagnostics alongside performance	Optimistic and potentially uninformative where novel conditions are unrepresented
Mixed or unspecified	Model to be reused for purposes not fixed at fitting time	Spatial blocking as default; report both estimates	Understates uncertainty for the harder task the model will face

3.3 What the corpus does not support

Three statements that appear in the applied literature, and that appeared in earlier versions of this synthesis, are not supported by the corpus and are withdrawn from it. No magnitude can be attached, from the corpus, to the spatial structure advantage in marine data, because no marine study measures it against an algorithm contrast; Section 3.4 attaches one for one survey, and it is not an advantage. No threshold in any metric can be given at which an algorithmic method should displace an additive one, because the thresholds reported are in non-comparable metrics and, at the sample sizes considered here, the difference they would need to exceed is not separable from between-fold variability. And no magnitude can be attached to the gain from adaptive monitoring, because the values circulating in secondary sources do not trace to primary studies in the corpus. The three withdrawn quantities are the three most quoted.

3.4 The benchmark

Values in this section are stated for this benchmark setting. Appendix S7 gives every cell.

3.4.1 Spatial structure against algorithm choice across sample sizes

Held-out discrimination was similar across the five models once the training subsample reached 200 records and the sample spanned the domain, and diverged sharply below that (Table 4; Figure S5). Averaged over coverage, target and species, mean held-out AUC over all usable fits at $n = 30$ was 0.77 for both random forests, 0.74 for the Gaussian Markov random field, 0.73 for the covariate-only GAM and 0.64 for the GAM with a coordinate smooth, with standard deviations across fits of 0.09 to 0.13; at $n = 500$ the range of means had closed to 0.79 to 0.81 and the spread to 0.06 to 0.12. The random forests lost about 0.03 AUC between 500 and 30 records, the covariate-only GAM 0.08, the random field model 0.06 and the coordinate smooth GAM 0.15. At 50 records the coordinate smooth GAM was the weakest model and the random field model and the random forest with coordinates the strongest in each of the three species (Table S11), so the small n ordering does not depend on prevalence.

Table 4. Held-out AUC on the matched target by model and training sample size in the benchmark: mean over all usable fits pooled across coverage, target and species, with the standard deviation across those fits in parentheses, and the number of usable fits. The standard deviation is the spread a practitioner faces across draws and cells, not the uncertainty of the mean (standard errors are 0.006 to 0.015).

Model	Statistic	$n = 30$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
Random forest	Mean (SD)	0.769 (0.098)	0.772 (0.094)	0.784 (0.085)	0.796 (0.074)	0.802 (0.063)
	Usable fits	113	120	120	120	120
	Usable fits	113	120	120	120	120
Random forest with coordinates	Mean (SD)	0.768 (0.117)	0.774 (0.111)	0.791 (0.106)	0.799 (0.099)	0.813 (0.071)
	Usable fits	113	120	120	120	120
	Usable fits	113	120	120	120	120
GAM	Mean (SD)	0.729 (0.129)	0.744 (0.139)	0.783 (0.123)	0.805 (0.102)	0.811 (0.093)
	Usable fits	113	120	120	120	120
	Usable fits	113	120	120	120	120
GAM with coordi- nate smooth	Mean (SD)	0.641 (0.107)	0.649 (0.134)	0.700 (0.131)	0.763 (0.143)	0.790 (0.123)
	Usable fits	93	99	109	109	108
	Usable fits	93	99	109	109	108
Gaussian Markov random field	Mean (SD)	0.743 (0.118)	0.739 (0.147)	0.755 (0.132)	0.795 (0.089)	0.803 (0.097)
	Usable fits	90	92	100	111	119
	Usable fits	90	92	100	111	119

The question the evidence map could not answer, whether spatial structure or algorithm family matters more and at what sample sizes, has a different answer for each of the three ways of adding spatial structure (Figure 4). Coordinates as predictors in a random forest helped only when the sample spanned the domain and the target was interpolation, by 0.02 to 0.03 AUC, with the difference positive in 66 percent of replicates at $n = 30$ and in 97 to 100 percent from $n = 100$. Under clustered coverage the same predictors were neutral or harmful (-0.04 at $n = 30$, interval -0.07 to -0.01), and for the marginal target they were within 0.025 of zero at every sample size. A coordinate smooth in a binomial GAM was harmful in every cell below 200 records, by 0.06 to 0.15 AUC with intervals excluding zero, and remained harmful at every sample size when the sample was a band (-0.10 at $n = 500$ under clustered conditional coverage, positive in 17 percent of replicates). The one cell in which it improved on the covariate-only

351 GAM was dispersed conditional coverage at $n = 500$, by 0.014. A Matérn random field in place
 352 of the coordinate smooth recovered most of that loss: the random field model exceeded the
 353 coordinate smooth GAM by 0.07 to 0.15 AUC at 30 to 50 records in every cell and by 0.10 to
 354 0.15 at every sample size under clustered conditional coverage, and was indistinguishable from it
 355 at 200 and 500 records under dispersed coverage (differences of -0.02 to $+0.004$). Against the
 356 covariate-only GAM, the random field model was roughly equal from 100 records upward and
 357 slightly better below, though that comparison is conditional on the fits that converged (Section
 358 3.4.4).

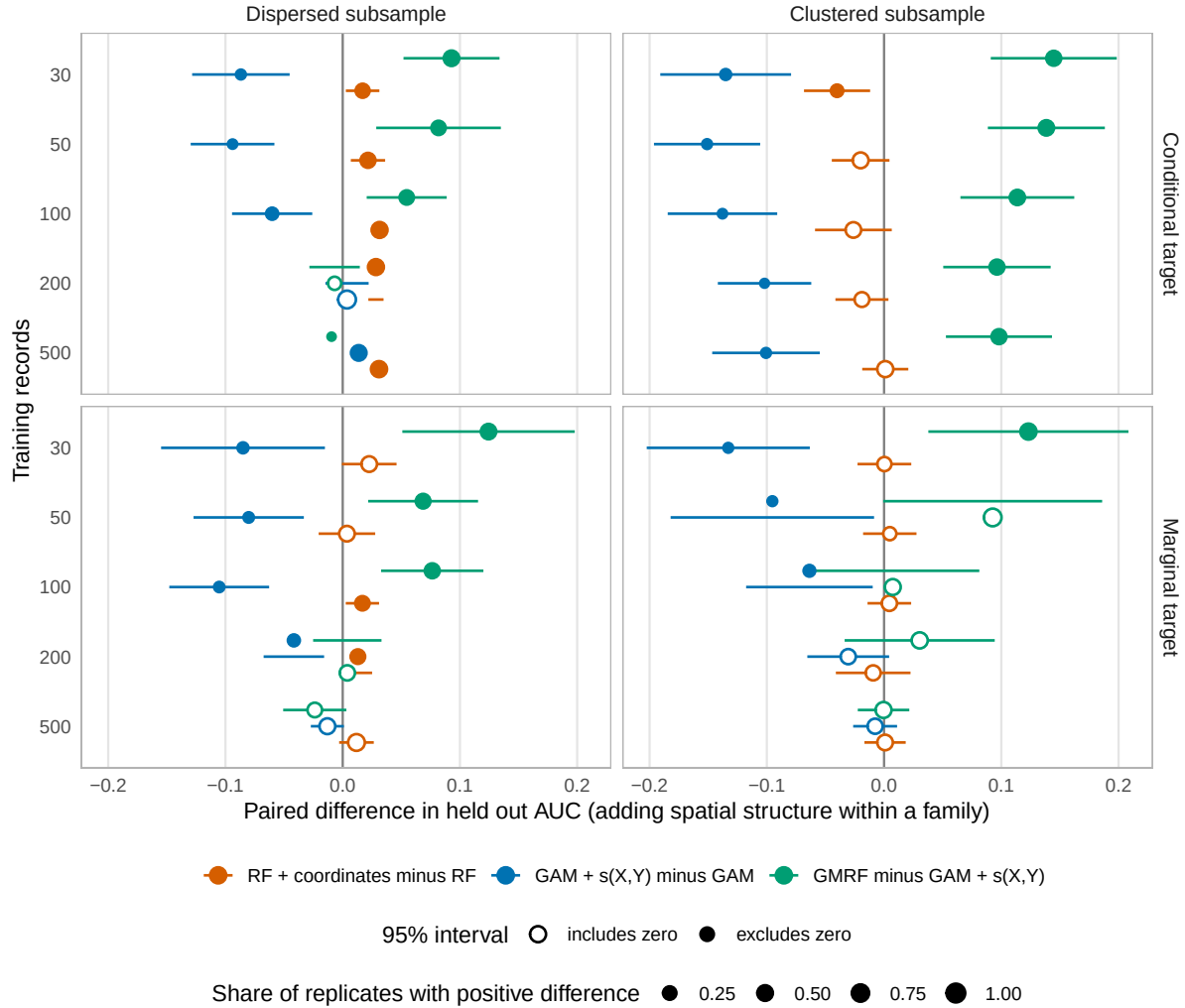


Figure 4. Effect of adding spatial structure within a model family, as the paired difference in held-out AUC between the spatial and non-spatial variant fitted to the same subsample. Rows are training sample sizes, whiskers the 95 percent interval over replicates (three species pooled), filled markers where the interval excludes zero, marker size the share of replicates in which the difference was positive. Panels are coverage design by inference target. Colours identify the spatial variant and are the model colours used in every benchmark figure.

359 The algorithm effect changed sign across the sample size range. Without spatial terms, the
 360 random forest exceeded the GAM at $n = 30$ in all four cells (by 0.03 to 0.05; positive in 54 to 70
 361 percent of replicates), the forest still led by 0.01 to 0.05 at 50 records, the two were within 0.025 of
 362 each other at 100, and the GAM was marginally better from 200 records under dispersed coverage

(−0.007 to −0.022; positive in 17 to 43 percent of replicates). With spatial terms the random forest with coordinates exceeded the GAM with a coordinate smooth by 0.11 to 0.16 AUC at 30 to 50 records and by 0.08 to 0.12 at 100, shrinking to 0.00 to 0.07 by 200 records (Figure S6). A comparison of algorithm families at a single sample size therefore carries no information about the ordering at another, which is the property that makes the single- n comparisons in the evidence map hard to transfer.

Calibration separated the models more than discrimination did at the smallest sample sizes (Figure 5). The median calibration slope on the held-out target at 30 to 50 records was 0.20 to 0.29 for the covariate-only GAM, 0.06 to 0.07 for the coordinate smooth GAM and 0.34 to 0.45 for the random field model, all far below 1 and indicating predicted probabilities much more extreme than the outcomes warrant. The random forest with coordinates had a median slope of 0.94 to 1.10 at every sample size and the plain random forest 0.65 to 0.83. The Brier score told the same story: 0.132 for the random forest against 0.184 for the GAM and 0.286 for the coordinate smooth GAM at $n = 30$, a larger gap than any AUC difference at that size. By 200 records the likelihood models had recovered to slopes of 0.46 to 0.77. At 30 to 50 records, in this setting, the choice of algorithm affected the probabilities far more than the ranking.

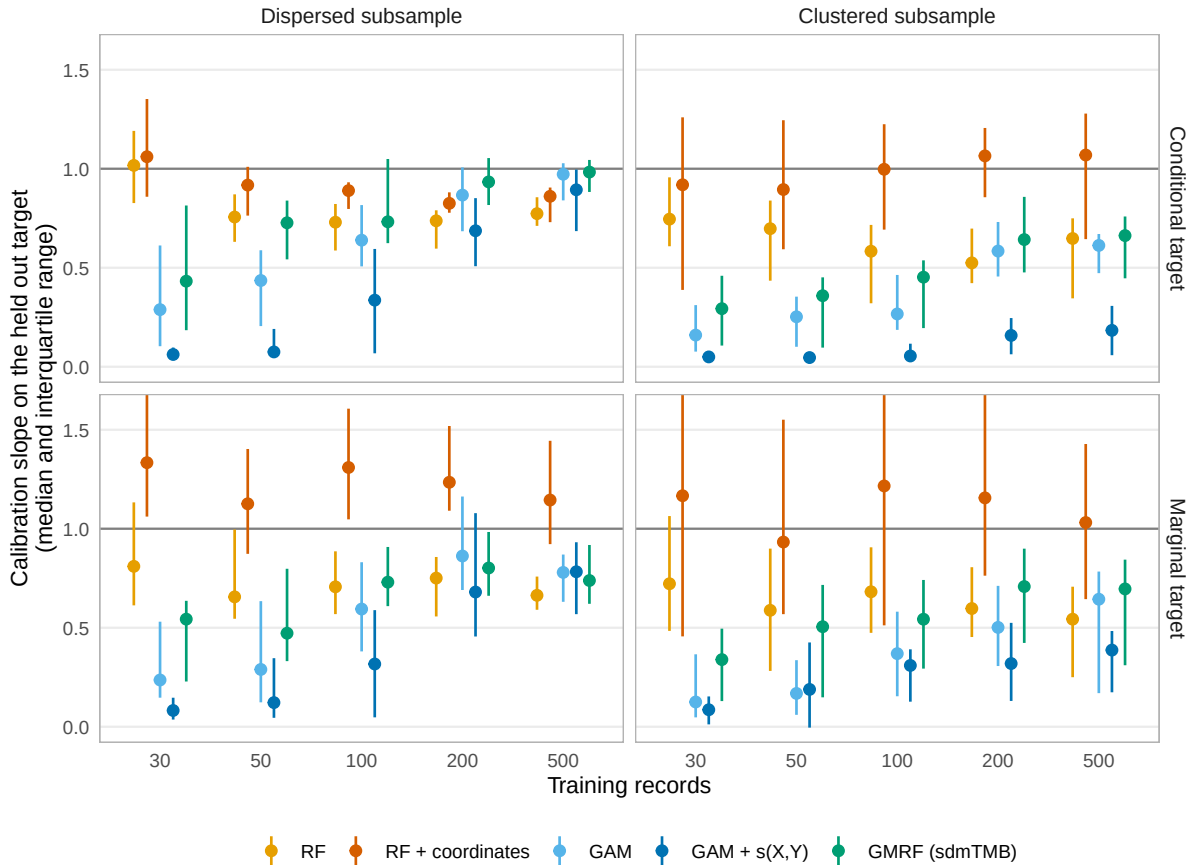


Figure 5. Calibration slope on the held-out target, median and interquartile range over replicates, by model and sample size; a slope of 1 (reference line) indicates calibrated probabilities and a slope below 1 indicates predictions more extreme than the observed frequencies.

3.4.2 Coverage against sample size For the conditional target, a dispersed subsample outperformed a clustered subsample of the same size for every model at every sample size (Figure 6, upper panel). The paired gain was 0.03 to 0.06 AUC for the random forest and the covariate-only GAM, 0.04 to 0.08 for the random field model, 0.08 to 0.10 for the random forest with coordinates and 0.08 to 0.16 for the coordinate smooth GAM, positive in 63 to 100 percent of replicates from 100 records upward. Coverage was worth as much as the whole gain from 30 to 500 records for the non-spatial models, and more than any algorithm contrast at sample sizes above 50. The models that use location most directly lost most from clustering, which is the expected consequence of asking a spatial term to extrapolate from a band one-tenth of the domain wide. For the marginal target the coverage effect was positive but smaller and less consistent (positive in 33 to 90 percent of replicates), because a northern holdout is extrapolation under both coverage designs and the band’s position relative to the target then matters as much as its width.

3.4.3 Validation optimism Neither cross-validation scheme estimated held-out performance well, and the direction of the error depended on sample size, coverage and target rather than on the scheme (Figure 7; Table 5). With a dispersed sample and an interpolation target, random five-fold cross-validation was pessimistic at small n: it underestimated held-out AUC by 0.11 at 30 records and by 0.035 at 50, reaching zero only at 500. Each fold model at $n = 30$ is trained on 24 records and scored on 6, and the sample is too small for the fold models to resemble the full model. Under extrapolation or clustered coverage the error ran the other way and grew with n. For the clustered marginal cell optimism rose from +0.03 at 30 records to +0.13 at 500; for dispersed marginal from 0 to +0.07; for clustered conditional from 0 to +0.06. More records inside a band make the internal estimate more precise about the band and no more informative about the outside.

Table 5. Optimism (cross-validation AUC minus held-out AUC), models pooled, for random and spatial block five-fold cross-validation by sample size, coverage and target. Entries are random; spatial block. Standard errors are 0.002 to 0.020 (Table S14).

n	Dispersed, conditional	Dispersed, marginal	Clustered, conditional	Clustered, marginal
30	−0.110; −0.084	−0.025; −0.036	−0.004; −0.006	+0.025; +0.026
50	−0.035; −0.028	+0.003; +0.003	+0.009; +0.021	+0.075; +0.081
100	−0.024; −0.033	+0.025; +0.011	+0.041; +0.047	+0.085; +0.067
200	−0.011; −0.014	+0.040; +0.045	+0.033; +0.031	+0.080; +0.079
500	−0.001; −0.009	+0.069; +0.066	+0.057; +0.063	+0.125; +0.123

Spatial block cross-validation did not remove the optimism. Across the 40 cells, models pooled, the block estimate differed from the random estimate by between −0.018 and +0.025 and by less than 0.01 in most, and in the cells where the optimism was largest the two were within 0.003 of each other. The fold maps in Appendix S6 show why. Under clustered coverage the residual autocorrelation range that sets the block side was 3 to 25 km, smaller than the spacing between hauls inside a band about 100 km wide, so the blocks partitioned the sample as a random assignment would. Under dispersed coverage the blocks were 40 to 150 km and the folds

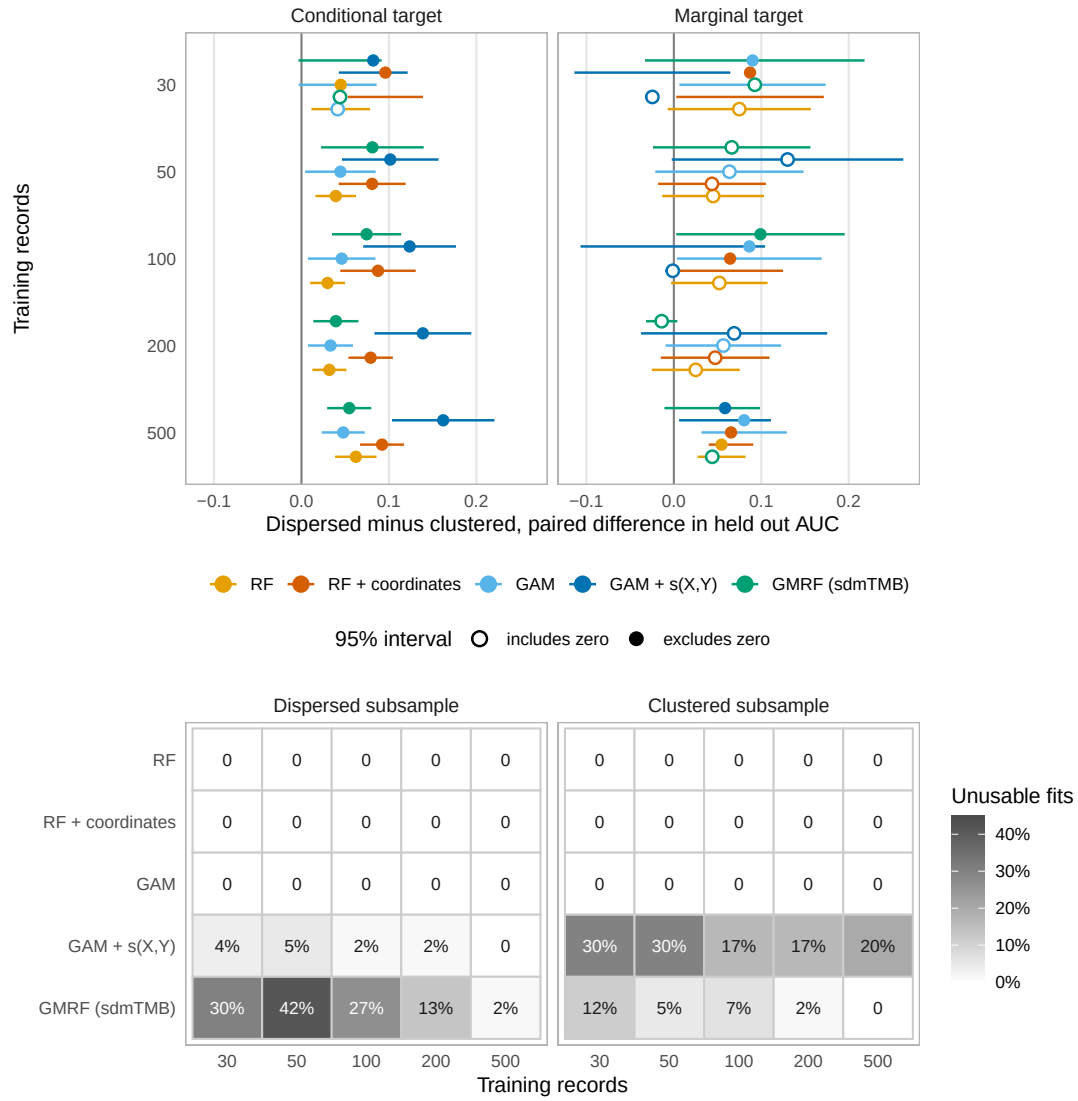


Figure 6. Upper panel: effect of coverage as the paired difference in held-out AUC between a dispersed and a clustered subsample of the same size, by model, sample size and target; conventions as in Figure 4. Lower panel: proportion of replicates in which a model returned no usable fit (error, no finite prediction or a degenerate prediction surface), by model, sample size and coverage.

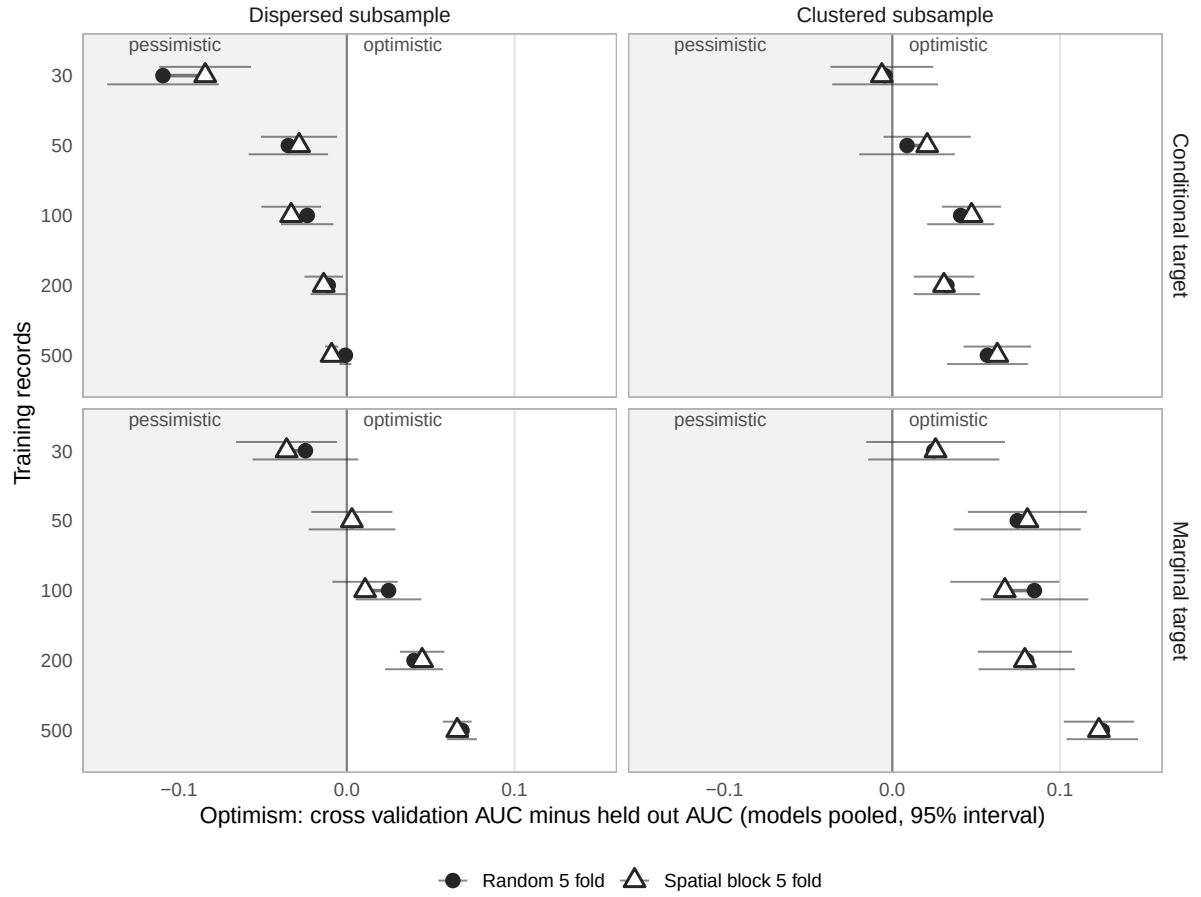


Figure 7. Optimism of internal validation, defined as the cross-validation estimate of AUC minus the AUC on the held-out target, models pooled. For each cell the random five-fold and spatial block five-fold estimates sit on one row joined by a segment, so the effect of blocking is the segment length; whiskers are 95 percent intervals. The shaded region marks pessimism (the internal estimate below the held-out value).
[File: Fig3_v2_optimism_dumbbell.png]

did occupy distinct stretches of coast, but the target there was interpolation, where blocking has nothing to correct. Block validation in this setting was a diagnostic of within-sample spatial dependence and not a substitute for coverage or for a held-out region on the side of the sample being predicted.

3.4.4 Failure modes The two models that add a spatial surface to a likelihood were also the two that failed (Figure 6, lower panel). The coordinate smooth GAM produced a degenerate surface, with predictions that did not vary across the evaluation set, in 17 percent of draws at 30 to 50 records and about 10 percent at 100 to 500, concentrated under clustered coverage (30 percent at 30 to 50 records against 4 to 5 percent dispersed). The random field model returned no usable prediction in 21 to 23 percent of draws at 30 to 50 records, 17 percent at 100, 8 percent at 200 and 1 percent at 500, and here the concentration was under dispersed coverage (30 to 42 percent at 30 to 50 records against 5 to 12 percent clustered), where the mesh spans the whole coast with a handful of records. The random forests and the covariate-only GAM never failed. Performance summaries above exclude failed fits, so they describe what a practitioner sees on obtaining a fit and flatter the two spatial smooth models at small n by the failure rate. The package’s stricter sanity flag for the random field model was failed by most usable fits at every sample size with no material difference in held-out AUC between flagged and unflagged fits (Table S15) and is treated as a diagnostic rather than an exclusion criterion. Seven of the 20 dispersed draws at $n = 30$ for the rarest species could not meet the presence minimum within 50 attempts and contributed no result.

3.5 A graded practitioner guide

The guidance below is what the corpus, the wider methods literature and the benchmark can defend now, and Figure 8 draws it as a decision framework. Each element carries its evidence grade from Section 2.4, and where the benchmark tested it the result is stated, including where the result qualified the guidance. Elements specific to marine survey data are marked as such, since the question of what in this guidance is marine rather than general is a fair one and the answer is not everything.

Minimum data by method. Table 6 gives indicative lower bounds by method, with the environment-specific scaling of Appendix S3 merged in. Tree-based presence-absence classification is workable from roughly 50 to 100 well-dispersed observations where habitat associations are strong, and in the benchmark a random forest held 0.77 AUC at 30 records against 0.80 at 500 while remaining the best calibrated family (Luan et al., 2020; Mi et al., 2017; this study; grade A and B). Boosted Regression Trees need roughly 100 as a lower bound and 200 for stable interaction estimation (Elith et al., 2008; grade C). Geostatistical spatio-temporal models have a structural requirement rather than a count: on the order of 30 stations across at least three years for basic configurations, with spatial structure approximated through 100 to 500 mesh nodes (Thorson, 2019b; grade C). In the benchmark a spatial only random field model failed to return a usable fit in about a fifth of draws at 30 to 50 records and in fewer than 2 percent at 500, and where it fitted it was the best

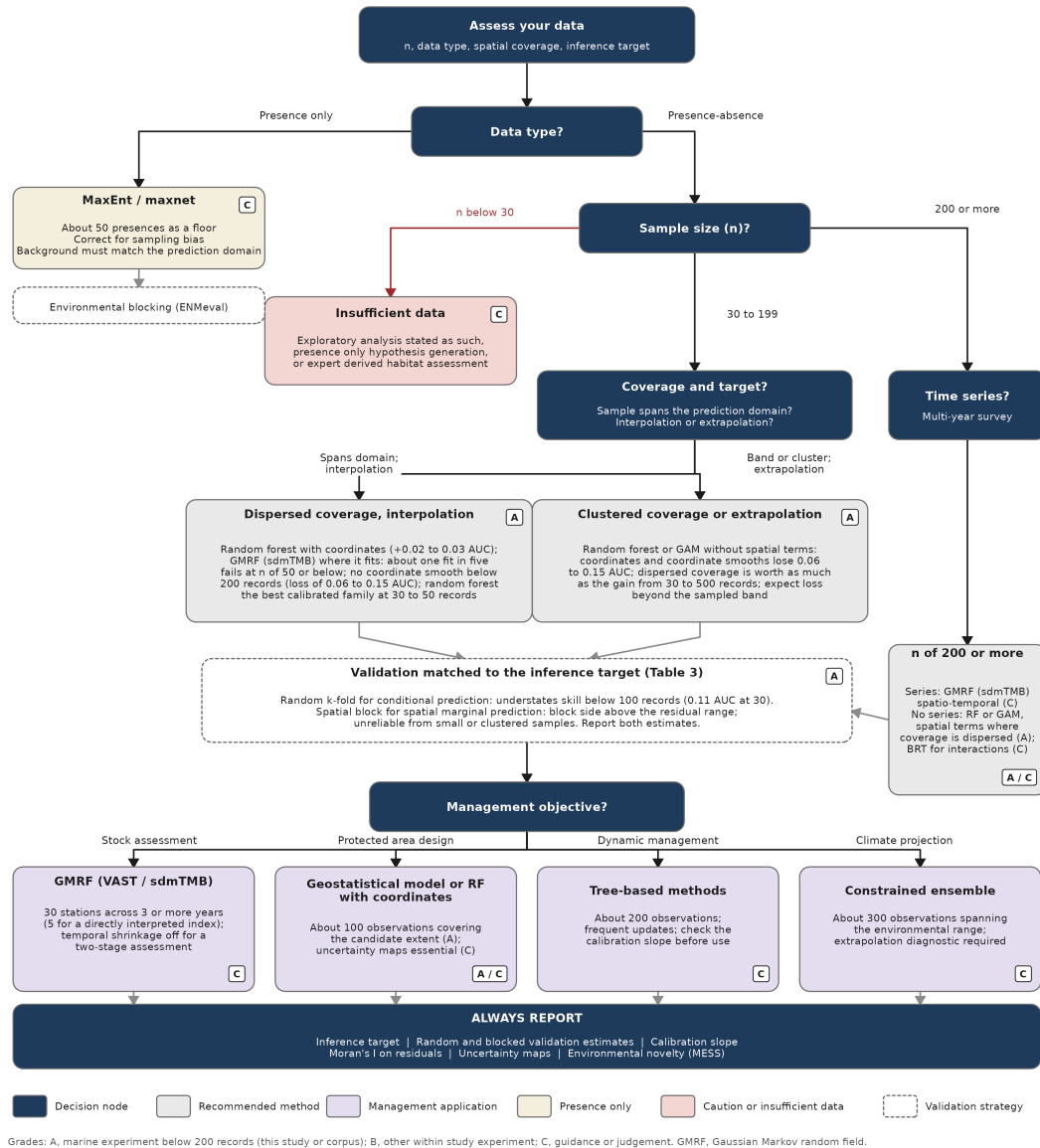


Figure 8. Decision framework for method and validation selection in data-limited marine species distribution modelling, drawn from the graded guide of Section 3.5. The sequence runs from data assessment through data type and sample size to the coverage and inference target of the sample, which the benchmark identifies as the decision that determines whether spatial terms help or harm (Section 3.4), then to the validation design matched to the target (Table 3), the management objective (Table 7) and the quantities to report. Each recommendation box carries its evidence grade (A, marine experiment below 200 records, this study or corpus; B, other within-study experiment; C, guidance or judgement). Numerical thresholds are the indicative minima of Tables 3, 6 and 7 and the benchmark effects of Section 3.4; GMRF, Gaussian Markov random field.

of the likelihood models below 100 records (grade A). A coordinate smooth in a GAM is not a substitute for the random field at these sizes: it reduced held-out AUC by 0.06 to 0.15 below 200 records in every cell of the benchmark (grade A). Joint species models need adequate replication per species and gain most where ten or more ecologically similar species co-occur (Norberg et al., 2019; grade B). Below roughly 30 observations the defensible options are exploratory analysis stated as such, presence-only hypothesis generation, or expert derived habitat assessment (Guillera-Arroita et al., 2015; grade C). The scaling factors by environment (pelagic roughly 1.5 times the demersal reference; coastal roughly 0.75; deep sea 0.6 to 1.0 with abundance modelling not recommended) are author estimates ordered by environmental variability, and are graded C.

Survey design before modelling (marine-specific). Four requirements bind a design in the order listed: coverage sufficient to represent the environmental space over which prediction is required; a sample size meeting the method-specific minimum; temporal replication where a spatio-temporal method is intended; and covariates at a resolution and temporal alignment matched to the response. Under a fixed budget, breadth of coverage should be preferred to intensity of local effort (Canonico et al., 2019); in the benchmark, dispersed coverage was worth as much held-out AUC as the whole gain from 30 to 500 records for the non-spatial models, and the models that use location most directly lost most from clustering (Section 3.4.2). Environmental stratification across temperature, depth and productivity gradients addresses the extrapolation problem most directly; geographic stratification is the practical approximation where the structuring gradients are not known in advance; temporal stratification is a prerequisite rather than an option for spatio-temporal methods. Grade A for the coverage priority (Luan et al., 2020; this study, Section 3.4.2), C for the ordering. The marine specificity lies in the geometry: survey records concentrate along tracks and contours, so coverage in environmental space is routinely poorer than the count suggests.

Sampling effort and detection (marine-specific). Gaussian Markov random field methods represent varying effort within the likelihood, so unequal sampling intensity across stations and years is accommodated in estimation. Tree-based learners treat records as exchangeable unless effort is supplied as a predictor, so an unmodelled effort gradient is absorbed into the fitted habitat relationship and reappears as apparent habitat preference. Imperfect and variable detection across gear and depth reduces effective sample size below the nominal count, which pushes working requirements above the values in Table 6. Grade C. This is a difference between method families that matters more under uneven marine survey effort than under gridded terrestrial sampling, and it is a reason to prefer the geostatistical families where effort varies and cannot be controlled.

Coordinate representation for tree-based methods (marine-specific). Tree-based learners split on one feature axis at a time, so rotating the coordinate system produces different trees and different predictions (Hastie et al., 2009). This is a property of how space is encoded rather than a defect of the method. The remedy is to supply several coordinate representations: latitude and longitude together with at least one ecologically meaningful axis such as distance from shore, depth contour distance or distance to a shelf break. The benchmark qualifies the advice: coordinates and distance to shore improved a random forest by 0.02 to 0.03 AUC only when the

Table 6. Indicative minimum data by method, with environment-specific scaling. Values are lower bounds under favourable conditions, namely strong habitat associations and coverage of the environmental gradient, and do not substitute for the coverage requirement. Scaling factors are author estimates ordered by environmental variability and are not empirical coefficients. Grades follow Section 2.4; grade A entries marked this study rest on Section 3.4.

Method	Indicative minimum (demersal reference)	Pelagic (about 1.5 times)	Coastal (about 0.75 times)	Deep sea (0.6 to 1.0 times)	Basis and grade
Random Forest, presence-absence	About 50 to 100 observations	About 75 to 150	About 40 to 75	30 to 50; presence-absence preferred	Luan et al. (2020), Mi et al. (2017); this study; A and B
Random Forest, abundance	Higher than classification; not quantified in corpus	Higher	Higher	Not recommended	Luan et al. (2020), qualitative; A
Boosted Regression Trees	About 100 lower bound; about 200 for stable interactions	About 150 to 300	About 75 to 150	Not recommended below 100	Elith et al. (2008); C
Geostatistical spatio-temporal (VAST, sdmTMB, tinyVAST)	About 30 stations across at least 3 years; 100 to 500 mesh nodes	About 45 stations	About 25 stations	Not recommended	Thorson (2019b); C; failure rates at 30 to 50 records from this study, A
Spatial GAM	Comparable to BRT; stable at small n with penalisation	As BRT	As BRT	Presence-absence only	Wood (2017); this study; C for the count, A for the harm of a coordinate smooth below 200 records
Joint species models	Adequate replication per species; benefit with 10 or more co-occurring species	As reference	As reference	Not recommended	Warton et al. (2015), Norberg et al. (2019); B
Any method below about 30 observations	Exploratory only; presence-only hypothesis generation; expert derived assessment	Same	Same	Same	Guillera-Arroita et al. (2015); C

sample spanned the domain and the target was interpolation, were neutral for extrapolation, and were harmful at 30 records when the sample was a band (Section 3.4.1), so they should be supplied when coverage is good and withheld, or the model recalibrated, when it is not. Marine habitat is organised along exactly such axes, which is why the choice of coordinate representation carries more weight in marine applications than the terrestrial literature on spatial Random Forest suggests. Grade C for the choice of axes; the benefit of coordinates in a random forest at all is grade A from Section 3.4.1, where it was confined to dispersed samples and interpolation targets.

Validation matched to inference target. Table 3 maps inference targets to partitioning schemes. State the target before any model is fitted. Random k-fold is defensible for conditional prediction, with the caution from the benchmark that below about 100 records it understates skill (by 0.11 AUC at 30 records), so a poor internal estimate at that size is not evidence against the model; spatial blocking at a block size exceeding the residual autocorrelation range is the recommended design for spatial marginal prediction, with the second caution that a block size estimated from a small or clustered sample is unreliable and, where it falls below the spacing of the records, delivers a random partition under a spatial name; leave-one-year-out or forward chaining for temporal marginal prediction; environmental strata for projection under changed conditions. Where objectives are mixed or the model will be reused, spatial blocking is the conservative default and reporting both estimates is better still, since the difference between them is itself informative about how far the model can be trusted beyond its training domain. Block size should be set from the residual variogram, and the nugget to sill ratio should be read alongside the range: below roughly 0.3 the range estimate is reliable; above roughly 0.8 nugget variance dominates and the range should be treated as a guide rather than a bound. Section 3.4.3 adds a caution that is grade A: a block size estimated from a small clustered sample was smaller than the spacing of the records, so blocked and random partitions gave the same estimate and neither revealed the extrapolation loss. Grade A for the direction and for the conditional and spatial marginal rows; C for the temporal and environmental rows.

Two-stage assessment workflows (marine-specific). Where a model produces annual abundance indices that are passed to a stock assessment, the assessment generally treats the index as a series of observations with independent year-specific errors. A model that imposes strong temporal autoregression on the abundance field produces indices correlated across years, which are not plausible draws from that error structure; the assessment then understates uncertainty and the correlation propagates into estimated trends as retrospective pattern. This is a reason for a specific structural choice in VAST and sdmTMB (Thorson, 2019b): rich spatial structure through random fields, without strong temporal autoregression on the field. Methods that impose AR(1) or random-walk structure on the abundance field yield smoother indices that look better in isolation and are less appropriate for this workflow. The constraint applies to two-stage workflows only and concerns temporal rather than spatial structure. It has no terrestrial analogue of comparable importance, since the two-stage workflow is a fisheries institution. Grade C.

Transfer and extrapolation. Transferability degrades with environmental novelty rather than geographic distance (Wenger and Olden, 2012; Yates et al., 2018). Report the proportion of the

prediction domain falling outside the training envelope in each covariate and in their multivariate combination, and present predictions there as extrapolations rather than mapping them with the visual authority of interpolations. Response curve behaviour outside the sampled range is a property of the method: tree-based learners extrapolate as flat surfaces, additive and random field methods according to the fitted basis or covariance, and neither is correct without checking against the extrapolation diagnostic. Grade C.

Method choice by objective. Table 7 pairs common marine management objectives with methods and indicative data. Continuous or zero-inflated responses with a requirement for calibrated uncertainty point to Gaussian Markov random field methods. Presence-absence with modest samples and strong habitat associations point to tree-based methods with spatial predictors, or spatial generalised additive models where interpretability is required for stakeholder communication. Community data with rare target species point to joint models with shrinkage on spatial latent variables, which outperform their non-spatial counterparts (Norberg et al., 2019). We give no numerical threshold at which an algorithmic method should displace an additive one, for the reason stated in Section 3.3; the defensible rule is comparative and local. Fit candidates under the same validation design matched to the target, compare against between-fold variability, and prefer the more interpretable or better calibrated model unless the algorithmic alternative exceeds that variability consistently. At the sample sizes considered here the comparison is usually inconclusive, and ensemble averaging weighted by cross-validated performance is a reasonable resolution (Araújo and New, 2007). The benchmark adds one element that the corpus could not: below about 100 records the models differed far more in calibration than in discrimination, so whichever model is chosen, its probabilities should be recalibrated before they are used as probabilities. Grade C for the pairings; A for the comparative rule, since Section 3.4.1 shows the algorithm contrast changing sign between 30 and 200 records, which is what makes a local comparison necessary.

4 Discussion

4.1 What the benchmark adds to the map

The evidence map identified an empty cell: no marine study had crossed spatial treatment against algorithm family on the same data below 100 records. In this benchmark setting that cell has a shape. Whether spatial structure helps depends on how it is added, and the three ways behave differently enough that spatial against non-spatial is not a useful contrast to carry into a recommendation. Coordinates in a tree-based learner are cheap and, with good coverage and an interpolation target, modestly and consistently helpful; a coordinate smooth in a GAM is harmful below about 200 records and harmful at any size when the sample is a band; a Matérn random field behaves like the coordinate smooth at large n and far better at small n , at the cost of fit failures in a fifth of small draws. The algorithm contrast that dominated the marine literature at large n , random forest against GAM, was worth 0.03 to 0.05 AUC at 30 records, changed sign by 200, and was in every cell smaller than the coverage contrast. The mechanism

Table 7. Objective to method pairings for common marine management applications, with indicative minimum data. The geostatistical requirement is reported by Thorson (2019b) and the joint model requirements follow Norberg et al. (2019); the remainder are author guidance consistent with Table 6 (grade C). Pairings are starting points; method selection should also account for computational resources, expertise, interpretability requirements and the resolution the decision needs.

Objective	Recommended methods	Indicative minimum data	Marine-specific consideration
Protected area habitat identification	Geostatistical model with uncertainty maps, or spatial Random Forest	About 100 observations covering the candidate extent	Coverage of the candidate extent matters more than count
Abundance index for direct interpretation	Geostatistical spatio-temporal model with a multi-year series	About 30 stations across at least 5 years	Effort represented in the likelihood
Abundance index feeding a two-stage assessment	Geostatistical model in default configuration, temporal shrinkage off	30 stations across at least 3 years (Thorson, 2019b)	Indices must be temporally independent in error structure (Section 3.4)
Climate projection	Constrained ensemble; mechanistic and correlative combined	About 300 observations spanning the environmental range	Marginal target; extrapolation diagnostic required
Biodiversity hotspot mapping	Multi-species geostatistical model, stacked single-species models, or joint model with spatial latent factors	About 50 to 100 per species, or 20 to 30 per species across 50 or more sites with 10 or more co-occurring species	Joint models gain where rare targets co-occur with better sampled congeners
Early detection of range expansion	Presence-only initially, moving to tree-based methods as monitoring accumulates	20 to 50 initially, expanding	Background extent must match the prediction domain
Dynamic management for mobile species	Tree-based methods for operational reasons; geostatistical alternatives untested operationally	About 200 observations; frequent updates	Fast prediction and no mesh construction favour tree-based methods; calibrated uncertainty is then absent
Community level impact assessment	Joint species models with environmental and spatial random effects	30 or more per species across 40 or more sites spanning the impact gradient	Variational inference where assessment timelines are tight

proposed in the Introduction, that flexible learners lose their advantage as n falls while spatial terms need spatially distributed observations to identify, is consistent with the pattern, with one refinement: the random forest’s advantage at small n came through calibration and the Brier score more than through ranking. The corpus finding of Section 3.2, that spatial structure improves prediction more consistently than algorithm choice, therefore survives in a qualified form: it holds for a random field and for coordinates in a forest under good coverage, and it fails for the way most applied marine studies add space to an additive model.

4.2 Coverage as the first-order decision

The coverage effect for the interpolation target was of the same size as the whole gain from 30 to 500 records for the non-spatial models. This is the quantitative form of the guidance the corpus supplied from single studies (Luan et al., 2020; Canonico et al., 2019): where a survey can be redesigned, spreading n stations across the domain buys more than adding stations to a cluster, and where it cannot, no choice of algorithm recovers the loss. The models that use location most directly, which are also the ones recommended most often for spatial data, were the ones that lost most from clustering. A practitioner with a fixed budget and a choice between a wider design and a denser one should, on this evidence, take the wider one for any target inside the domain; for a target outside it, width helps less and the position of the sample relative to the target matters as much.

4.3 Validation at small-sample sizes

The sample size question is a question about an interaction, and the benchmark shows the interaction operating on validation as well as on models. The literature’s expectation of optimism from random partitioning under spatial dependence (Roberts et al., 2017; Ploton et al., 2020) presupposes a sample large enough for the fold models to resemble the full model. At 30 records with good coverage the estimate was pessimistic by 0.11 AUC, and a practitioner reading it would reject a model that in fact performed at 0.81. The opposite error, optimism growing with n under extrapolation or clustering, is the one the literature describes, and blocking did not correct it here because a block side estimated from a band is smaller than the spacing of the hauls. Neither property is an argument against spatial cross-validation; both are arguments for choosing the validation design from the inference target rather than from the sample. A partition aligned with the direction of prediction, holding out the northern fold of the training sample when the target is northern, is the natural test and is proposed as the next extension of the design. The single marine quantification of optimism in the corpus, 0.11 AUC at 4,713 records (Becker et al., 2020), sits inside the range measured here at 500 records under extrapolation, which is some reassurance that the scaling is not peculiar to this survey.

4.4 Calibration

Discrimination is the metric the marine literature reports, and it was the metric on which the models differed least at small n . The calibration slopes of 0.06 to 0.45 for the likelihood models at 30 to 50 records mean that a predicted probability of 0.9 corresponded to an observed frequency far below it, which matters for any use of the model that treats probabilities as probabilities: expected habitat area, encounter rates, thresholds set on the probability scale. The random forest with coordinates was close to calibrated at every sample size. If a single recommendation for 30 to 50 records were to come out of this benchmark it would concern the probability scale, not the ranking, and it would be to recalibrate any likelihood model fitted at that size before its probabilities are used.

4.5 What in this guidance is marine

The principles that spatial dependence must be represented, that coverage matters, and that validation must match the target are general. What is marine is the set of conditions that determine how they bite. Marine survey records are collected along tracks and contours by gears with depth and habitat-dependent detection, at effort that varies by station and year, from a habitat volume sampled at one surface. Four consequences follow that have no terrestrial counterpart of comparable weight, and Section 3.5 states them: effort must be represented in the likelihood or as a predictor, which separates the method families; coordinate representation for tree-based learners should follow the shore, shelf and depth axes that organise marine habitat; environment-specific data requirements are ordered by oceanographic variability rather than accessibility; and the two-stage assessment workflow constrains the temporal structure a model may impose. These are not general spatial modelling principles applied to a marine example. They are marine sampling structure entering the modelling decision. The benchmark adds three features of the setting that are properties of trawl surveys and bear on transfer. The survey is a long thin strip along a coast, so a latitudinal band is the natural clustered design and a northern region the natural extrapolation target; a survey over a two-dimensional shelf or an archipelago would admit other geometries and the coverage effect would need to be measured again. Hauls are fixed-effort samples with confirmed absences, so prevalence is known and the presence minimum is a design choice rather than a data limitation; presence-only data would change the calibration results in particular. Depth structures the fauna strongly, so three covariates carry most of the signal and the spatial terms have less to add than they would where the environment is less well measured; that favours the covariate-only models and is the direction in which the benchmark results are most likely to be specific to this survey.

4.6 Limitations

Metric heterogeneity prevents quantitative synthesis of the existing literature whatever effort is applied to it, and is the reason the findings above are directional. The remedy is upstream,

in standardised multi-metric reporting by primary studies (Appendix S5). Publication practice biases the observable literature toward positive comparative results, so the corpus cannot support statements about how often a given approach fails. The corpus is English language and peer-reviewed; regional grey literature, including stock assessment working group reports, is under-represented, and marine practice in those documents may differ from what is published. The evidence map counts sources cited in support of each finding and is a lower bound on corpus contribution. The environment scaling factors and objective method pairings are author judgement and are graded as such. The benchmark has its own bounds. Ten replicates per cell is enough to fix the sign and approximate size of the large contrasts, which held across species and cells, but not to settle the weaker ones: the coordinate benefit in random forests under dispersed coverage and the sign change of the algorithm effect between 30 and 200 records have intervals near zero at this replication and should be read as suggestive. Three species from one survey, with three covariates and fixed default settings for every model, define the setting; tuning at each n would have introduced variance the design cannot separate, at the cost of favouring the models most sensitive to defaults. The presence minimum conditions the small n rare species cells on favourable draws. The block side rule was applied as recommended and is unstable at these sample sizes; a fixed minimum block is a different design and is proposed rather than run. Performance summaries are conditional on a usable fit. Response type is presence-absence only, with no abundance and no multi-species models, and the study makes no attempt at cross-metric mapping, which is the subject of a separate protocol. None of these limits the structural findings, which held across species and cells, but each bounds how far a number should travel.

4.7 Conclusions

The marine species distribution literature tells practitioners to add spatial structure before comparing algorithms, to spread survey effort rather than concentrate it, and to match validation to the inference target. Those instructions are well founded in spatial statistics, and this paper has mapped how far they had been tested in marine data and then tested them, in one survey, at the sample sizes at which they are applied. Two of the three survive intact: coverage is the first-order decision, and validation must be chosen from the target. The first survives only in a qualified form, because how spatial structure is added matters more than whether it is added, and the most common way of adding it to an additive model was harmful below 200 records. The benchmark also supplied two findings the corpus could not: internal validation understates skill at 30 records with good coverage and does not reveal extrapolation loss from a clustered sample under any partition, and at the smallest sample sizes model choice affects the probability scale far more than the ranking. The guidance in Section 3.5 states what follows, graded on the evidence behind each element, and the design and scripts are archived so that the grades can rise as other surveys are tested.

Supporting information

- Appendix S1. Search strategy, screening, inclusion criteria, and non-marine sources retained.
- Appendix S2. Extraction protocol, extracted studies and quality assessment.
- Appendix S3. Survey design: coverage, effort and data quality; environment-specific scaling.
- Appendix S4. Computational cost and software.
- Appendix S5. Reporting protocol.
- Appendix S6. Benchmark sampling and validation designs
- Appendix S7. Benchmark additional results

Data and code availability

Survey data for the benchmark are from the surveyjoin R package (Ward et al., 2025; <https://github.com/DF0-NOAA-Pacific/surveyjoin>). The subsampling seeds, the cell-by-replicate result table (results_all.rds), the benchmark scaffold, the analysis and figure scripts are archived at <https://github.com/EAfrifa-Yamoah/Which-modelling-choices-are-supported>.

Author contributions

EAY conceived the study, conducted the search, screening and coding, and drafted the manuscript. Co-authors provided resources and critically reviewed the manuscript state roles. All authors read and approved the manuscript.

Funding

EAY was supported by the ECU Early-Mid Career Researchers (EMCR) Grant Scheme 2026 (G1008306).

Conflicts of interest

The authors declare no conflicts of interest.

Statement on the use of AI assistance

Generative AI (Anthropic Claude) was used during preparation of this manuscript to assist with restructuring earlier drafts, editing prose, support drafting the R code for the benchmark. All literature searching, screening, coding and evidence grading were performed by the authors; every citation and every numerical value was verified by the authors against the primary source or

reproduced by the authors from the archived scripts; and all code was executed and checked by the authors. The authors take full responsibility for the content.

References

- Addison, P. F. E., Collins, D. J., Trebilco, R., et al. (2018). A new wave of marine evidence-based management: Emerging challenges and solutions to transform monitoring, evaluating, and reporting. *ICES Journal of Marine Science*, 75(3), 941–952.
- Anderson, S. C., Ward, E. J., English, P. A., Barnett, L. A. K., & Thorson, J. T. (2025). sdmTMB: An R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *Journal of Statistical Software*, 115(2), 1–46.
- Araújo, M. B., & New, M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1), 42–47.
- Becker, E. A., Carretta, J. V., Forney, K. A., et al. (2020). Performance evaluation of cetacean species distribution models developed using generalized additive models and boosted regression trees. *Ecology and Evolution*, 10(12), 5759–5784.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Brodie, S. J., Thorson, J. T., Carroll, G., et al. (2020). Trade-offs in covariate selection for species distribution models: A methodological comparison. *Ecography*, 43(1), 11–24.
- Canonico, G., Buttigieg, P. L., Montes, E., et al. (2019). Global observational needs and resources for marine biodiversity. *Frontiers in Marine Science*, 6.
- Elith, J., Graham, C. H., Anderson, R. P., et al. (2006). Novel methods improve prediction of species’ distributions from occurrence data. *Ecography*, 29(2), 129–151.
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Grüss, A., & Thorson, J. T. (2019). Developing spatio-temporal models using multiple data types for evaluating population trends and habitat usage. *ICES Journal of Marine Science*, 76(6), 1748–1761.
- Grüss, A., Thorson, J. T., Stawitz, C. C., et al. (2021). Synthesis of interannual variability in spatial demographic processes supports the strong influence of cold-pool extent on eastern Bering Sea walleye pollock (*Gadus chalcogrammus*). *Progress in Oceanography*, 194.
- Guillera-Aroita, G., Lahoz-Monfort, J. J., Elith, J., et al. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24(3), 276–292.

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, 6, e5518.
- Kapur, M. S., Adams, G., Lapeyrolerie, M., & Thorson, J. T. (2026). AI for fisheries science: Neural network tools for forecasting, spatial standardization, and policy optimization. *bioRxiv*.
- Kellenberger, B., Winner, K., & Jetz, W. (2026). The performance and potential of deep learning for predicting species distributions. *Global Ecology and Biogeography*, 35, e70184.
- Klaassen, M., Marques, T. A., Alves, F., & Fernandez, M. (2025). Trends in marine species distribution models: A review of methodological advances and future challenges. *Ecography*, e07702.
- Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2), 145–151.
- Luan, J., Zhang, C., Xu, B., Xue, Y., & Ren, Y. (2020). The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research*, 227.
- Maureaud, A. A., Palacios-Abrantes, J., Kitchel, Z., et al. (2024). FISHGLOB_data: An integrated dataset of fish biodiversity sampled with scientific bottom-trawl surveys. *Scientific Data*, 11.
- Melo-Merino, S. M., Reyes-Bonilla, H., & Lira-Noriega, A. (2020). Ecological niche models and species distribution models in marine environments: A literature review and spatial analysis of evidence. *Ecological Modelling*, 415.
- Mestre-Tomás, J., Fuster-Alonso, A., Bellido, J. M., & Coll, M. (2025). GLOSSA: A user-friendly R Shiny application for Bayesian machine learning analysis of marine species distribution. *Methods in Ecology and Evolution*, 16(1), 1–13.
- Mi, C., Huettmann, F., Guo, Y., Han, X., & Wen, L. (2017). Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ*, 5, e2849.
- Miller, D. L., Becker, E. A., Forney, K. A., et al. (2022). Estimating uncertainty in density surface models. *PeerJ*, 10, e13950.
- Mushagalusa, C. A., Fandohan, A. B., & Glèlè Kakai, R. (2024). Random forest and spatial cross-validation performance in predicting species abundance distributions. *Environmental Systems Research*, 13.
- Norberg, A., Abrego, N., Blanchet, F. G., et al. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3), e01370.
- Pasanisi, E., Pace, D. S., Orasi, A., Vitale, M., & Arcangeli, A. (2024). A global systematic

review of species distribution modelling approaches for cetaceans and sea turtles. *Ecological Informatics*, 82.

Ploton, P., Mortier, F., Réjou-Méchain, M., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1).

Roberts, D. R., Bahn, V., Ciuti, S., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929.

Robinson, L. M., Elith, J., Hobday, A. J., et al. (2011). Pushing the limits in marine species distribution modelling: Lessons from the land present challenges and opportunities. *Global Ecology and Biogeography*, 20(6), 789–802.

Robinson, N. M., Nelson, W. A., Costello, M. J., Sutherland, J. E., & Lundquist, C. J. (2017). A systematic review of marine-based species distribution models (SDMs) with recommendations for best practice. *Frontiers in Marine Science*, 4.

Rubbens, P., Brodie, S., Cordier, T., et al. (2023). Machine learning in marine ecology: An overview of techniques and applications. *ICES Journal of Marine Science*, 80(7), 1829–1853.

Talebi, H., Peeters, L. J. M., Otto, A., & Tolosana-Delgado, R. (2022). A truly spatial random forests algorithm for geoscience data analysis and modelling. *Mathematical Geosciences*, 54, 1–22.

Thorson, J. T. (2019a). Forecast skill for predicting distribution shifts: A retrospective experiment for marine fishes in the Eastern Bering Sea. *Fish and Fisheries*, 20(1), 159–173.

Thorson, J. T. (2019b). 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., Anderson, S. C., Goddard, P., & Rooper, C. N. (2025). tinyVAST: R package with an expressive interface to specify lagged and simultaneous effects in multivariate spatio-temporal models. *Global Ecology and Biogeography*, 34(7), e70035.

Thorson, J. T., Shelton, A. O., Ward, E. J., & Skaug, H. J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES Journal of Marine Science*, 72(5), 1297–1310.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2), 225–232.

Ward, E. J., English, P. A., Rooper, C. N., et al. (2025). surveyjoin: A standardized database of scientific trawl surveys in the Northeast Pacific Ocean. *PeerJ*, 13, e19964.

Warton, D. I., Blanchet, F. G., O’Hara, R. B., et al. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, 30(12), 766–779.

Wenger, S. J., & Olden, J. D. (2012). Assessing transferability of ecological models: An underappreciated aspect of statistical validation. *Methods in Ecology and Evolution*, 3(2),

810 260–267.

811 Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman
812 and Hall/CRC.

813 Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high
814 dimensional data in C++ and R. *Journal of Statistical Software*, 77(1), 1–17.

815 Yates, K. L., Bouchet, P. J., Caley, M. J., et al. (2018). Outstanding challenges in the transfer-
816 ability of ecological models. *Trends in Ecology & Evolution*, 33(10), 790–802.

817 Zelli, E., Ellis, J., Pilditch, C., Rowden, A. A., et al. (2025). Identifying climate refugia for
818 vulnerable marine ecosystem indicator taxa under future climate change scenarios. *Journal of*
819 *Environmental Management*, 373.