Which modelling choices have been tested where marine survey data are sparse? An evidence map of 95 studies and a graded practitioner guide

Eben Afrifa-Yamoah[1,2]*, Alexandre C. Siqueira[2], 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

Aim. Marine conservation decisions increasingly rest on species distribution models fitted to sparse, clustered and irregularly sampled survey records, while guidance on fitting such models comes largely from data rich settings. We asked which modelling choices have actually been tested under the conditions marine practitioners face, what the tested evidence supports, and what remains untested.

Location. Global; marine and estuarine systems.

Methods. We screened 330 records and retained 95 peer reviewed marine applications that treated spatial autocorrelation explicitly. Each study was coded for the contrasts it tested (spatial treatment, algorithm family, sampling coverage, validation design), its study system, the sample size range at which performance was reported, and the metric family used. Twelve studies reporting direct comparisons, validation optimism or sample size experiments were extracted in detail. Values were never pooled across studies or metrics.

Results. The evidence map locates a specific gap. No marine study in the corpus crosses spatial treatment against algorithm family on the same data; the three that do are terrestrial or simulated. One marine study reports performance below 100 observations. Validation optimism has been quantified within study in marine data once, at 4,713 records. Within these limits the corpus supports three directional findings: adding spatial structure improves prediction more consistently than changing algorithm family; spatial coverage of sampling constrains reliability more than record count; and random cross validation is defensible for prediction within sampled conditions but optimistic beyond them. Method specific minimum data requirements, environment specific adjustments, and a constraint that two stage assessment workflows place on model structure are tabulated, each graded by the evidence supporting it.

Main conclusions. The principles marine practitioners are told to follow are established in spatial statistics but have not been tested at the sample sizes and sampling geometries marine conservation operates in. We provide the guidance that is defensible now, mark each element by its evidence grade, and specify the benchmark on harmonised trawl surveys that would close the gap.

Keywords. marine conservation; data limitation; evidence map; 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 three parts. First, which contrasts between modelling choices have been tested on marine data, at what sample sizes, and with what validation design. Second, what the tested evidence supports in direction, and where the marine magnitude is known. Third, what a practitioner should do now, with each element of the guidance marked by the evidence behind it, and what experiment would convert the untested elements into tested ones.

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. We have found that 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, has never been tested in the setting where it is most often invoked.

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.

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. The coding sheet is Appendix S3.

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-Arroita 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 in the paper appears in Table 5 with its source and its scope of generalisation.

2.4 Evidence grades

Each element of the practitioner guidance in Section 3.4 carries one of five grades, assigned from the evidence map. Grade A: a within study experiment on marine data at sample sizes below 200 observations. Grade B: a within study experiment on marine data at larger sample sizes. Grade C: a within study experiment on non marine or simulated data. Grade D: methods guidance or statistical theory without a within study experiment. Grade E: author judgement informed by the corpus, with no direct empirical test. The grades describe the evidence, not the confidence we hold; several grade D and E 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, which is the question the map is built to answer.

3 Results

3.1 The evidence map

Table 1 gives the map. Three features of it organise everything that follows.

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 has 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.

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 4, 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.

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 for the marine data limited regime: C for the crossed contrast, B for the within family marine contrast.

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 exists at the relevant sample size. Evidence grade: A, from one study.

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: B for the direction in marine data; D for the target specific mapping.

3.3 What the corpus does not support

Three statements that appear in the applied literature, are not supported by this corpus. No magnitude can be attached to the spatial structure advantage in marine data, because no marine study measures it against an algorithm contrast. 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.

3.4 A graded practitioner guide

The guidance below is what the corpus and the wider methods literature can defend now. Each element carries its evidence grade from Section 2.4. 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 2 gives indicative lower bounds by method, with the environment specific scaling of Appendix S5 merged in. Tree based presence absence classification is workable from roughly 50 to 100 well dispersed observations where habitat associations are strong (Luan et al., 2020; Mi et al., 2017; grade A and C). Boosted Regression Trees need roughly 100 as a lower bound and 200 for stable interaction estimation (Elith et al., 2008; grade D). 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 D). 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 C). 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 D). 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 E.

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). 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), D 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 2. Grade D. 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. 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 D.

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; spatial blocking at a block size exceeding the residual autocorrelation range is required for spatial marginal prediction; 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. Grade B for the direction, D for the mapping.

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 D.

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 D.

Method choice by objective. Table 4 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). Grade E for the pairings; D for the comparative rule.

4 Discussion

4.1 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 3.4 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.

4.2 What in this guidance is about small samples

The sample size question is a question about an interaction. Flexible algorithms earn their advantage by estimating interactions and unspecified functional forms; the variance of those estimates grows as sample size falls, and at some point the advantage inverts. Spatial random fields likewise require enough spatially distributed observations to estimate a range and a variance; below that, the field is either not identifiable or is shrunk to nothing, and the spatial method degrades to its non spatial counterpart. Whether the ordering spatial structure before algorithm choice survives at 30, 50 or 100 marine records is therefore an open question, and it is possible that it inverts at the low end, with a penalised spatial smooth outperforming a random field that cannot be estimated. Nothing in the corpus tests this. The guidance in Section 3.4 is what follows from established theory and from the two small sample experiments that exist; it is not a set of tested findings about the small sample regime, and we have graded it so that the reader is not led to think otherwise.

4.3 The experiment that would close the gap

The map specifies the missing experiment precisely, and the harmonised trawl survey resources now available make it feasible (Maureaud et al., 2024; Ward et al., 2025). Draw structured subsamples from a harmonised survey at sample sizes spanning 30 to 500 records, crossed with clustered and dispersed spatial coverage, and fit spatial and non spatial variants of at least two algorithm families to each subsample. Evaluate each fit against the held out remainder of the survey under both conditional and marginal targets, so that validation optimism is measured rather than inferred. Report the spatial structure effect, the algorithm effect and their interaction with sample size within one design. Such a study would establish, for the first time in marine data, whether the ordering this literature assumes holds at the sample sizes it is applied to, and it would attach the magnitudes that the present synthesis cannot. We have designed this experiment on the NWFSC West Coast Groundfish Bottom Trawl Survey as harmonised in surveyjoin, with five sample sizes, two coverage regimes, five models and two inference targets, and it is in progress; the present paper establishes why it is needed.

4.4 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 S6). 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.

4.5 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. They have not been tested at the sample sizes or under the survey geometries in which marine conservation operates, and the most quoted of them, the ordering of spatial structure over algorithm choice, has never been measured in a marine system. This paper maps that gap, gives the guidance that is defensible in the meantime with each element graded by its evidence, identifies the four respects in which marine sampling structure enters the modelling decision, and specifies the benchmark that would replace graded guidance with tested findings.

Tables

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

Table 2. 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.

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); A and C
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); D
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); D
Spatial GAM	Comparable to BRT; stable at small n with penalisation	As BRT	As BRT	Presence absence only	Wood (2017); D
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); C
Any method below about 30 observations	Exploratory only; presence only hypothesis generation; expert derived assessment	Same	Same	Same	Guillera-Arroita et al. (2015); D

Table 3. Matching validation design to inference target. Grade B for the direction in marine data (Becker et al., 2020); grade D for the target specific mapping (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

Table 4. 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 2 (grade E). 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

Table 5. 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

Figures

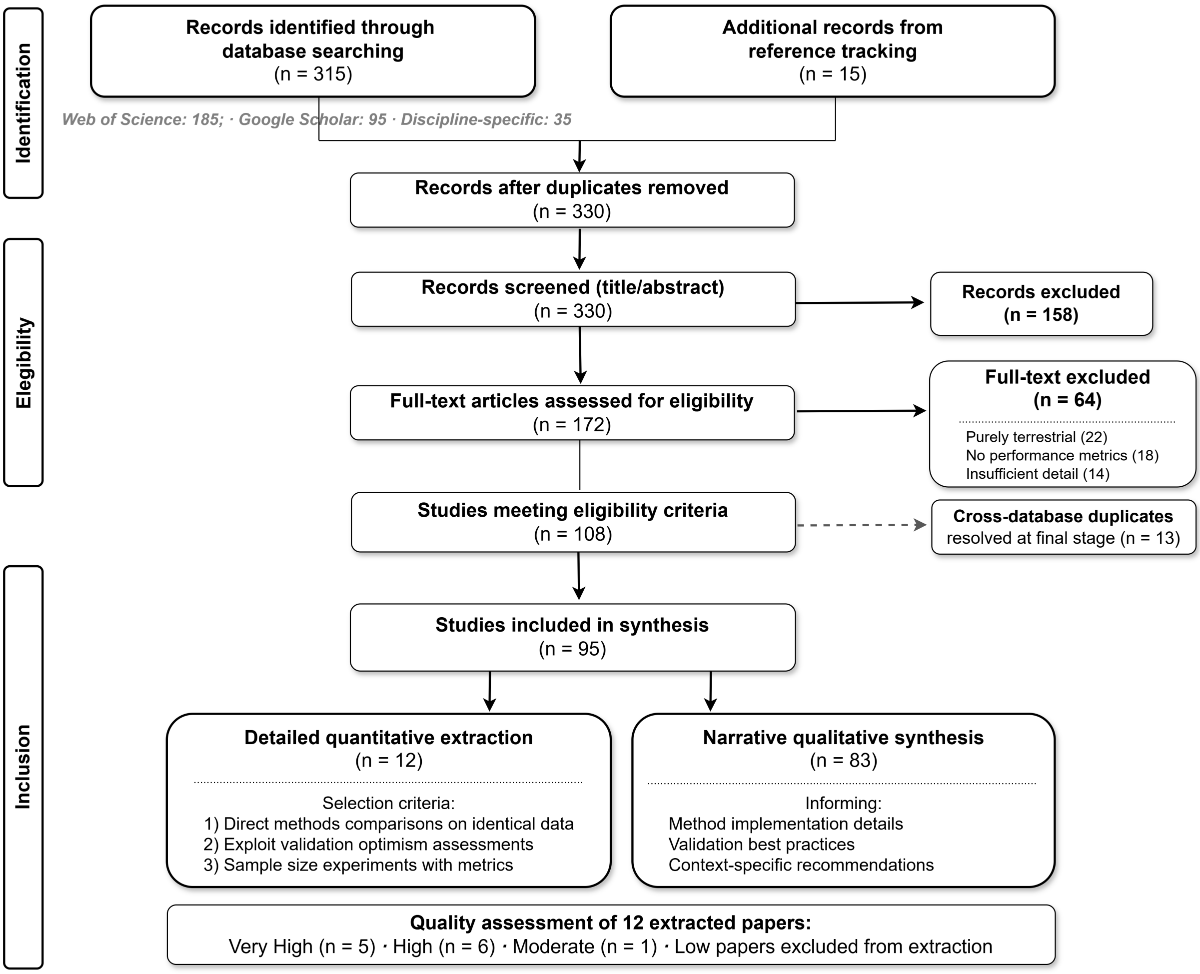

Figure 1. Identification, screening and inclusion of studies of spatial modelling applied to marine species distributions. Screening criteria are in Appendix S1. Peer reviewed sources only; stock assessment working group reports and other grey literature were not searched (Section 4.4).

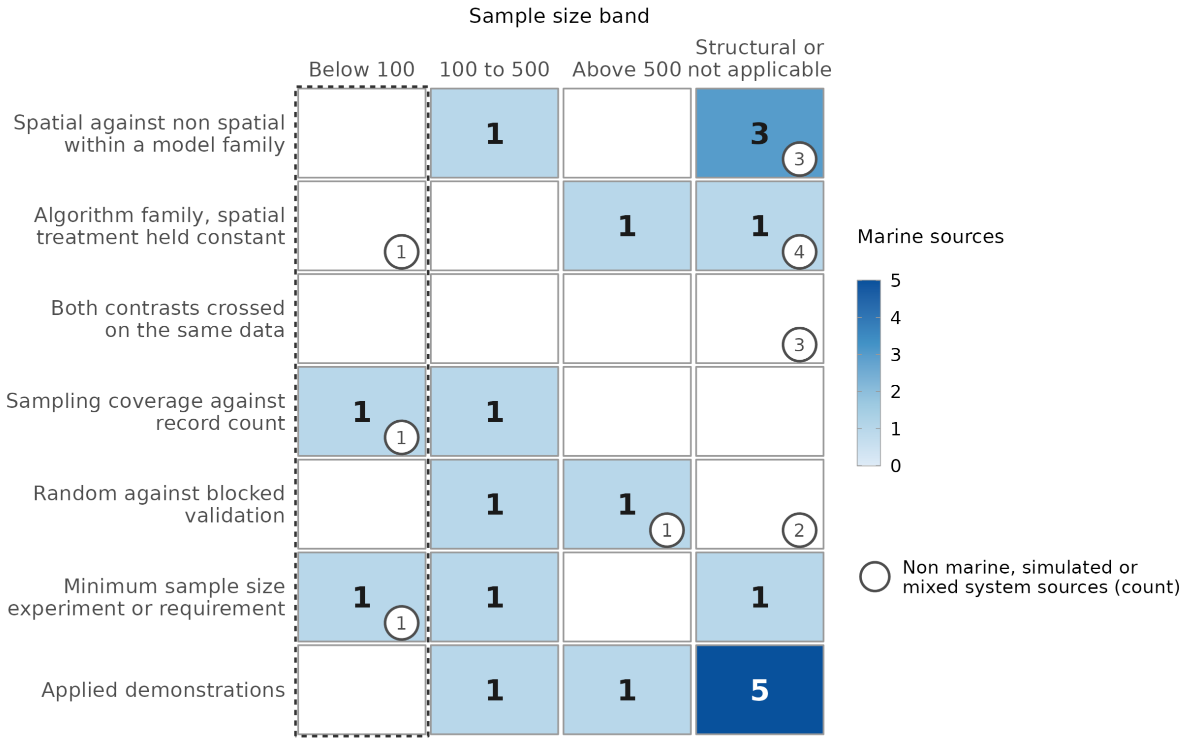

Figure 2. The evidence map as a matrix. Rows are the contrasts of Table 1; columns are sample size bands (below 100, 100 to 500, above 500, structural or not applicable). A study contributes to every band its reported sample size range spans, and to every contrast it tests, so a study can occupy more 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.

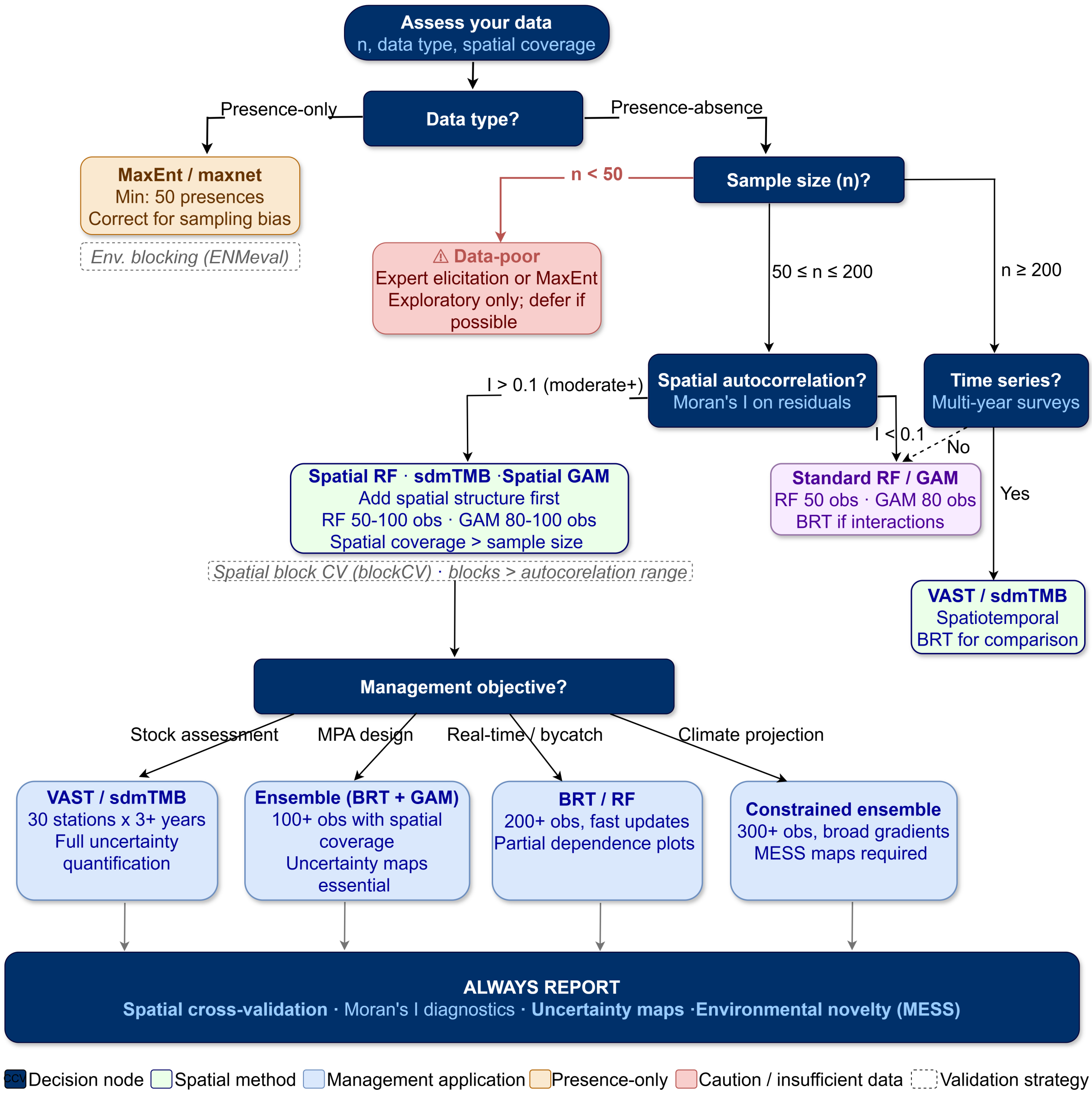

Figure 3. Decision framework for method and validation selection in data limited marine species distribution modelling. The sequence runs from data assessment through inference goal to method choice, with validation design fixed before model fitting. Each recommendation box carries the evidence grade of the corresponding element in Section 3.4.

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. Metrics, and the comparison protocol
- Appendix S4. Survey design: coverage, effort and data quality; environment specific scaling; two stage assessment workflows
- Appendix S5. Computational cost and software
- Appendix S6. Reporting protocol

Data and code availability

The detailed search protocol and evidence map coding are deposited with this preprint as a supplementary file and at https://github.com/EAfrifa-Yamoah/Spatial-structure-before-algorithmic-complexity-Supplement. No new empirical data were generated. The benchmark described in Section 4.3 uses the surveyjoin database (Ward et al., 2025; https://github.com/DFO-NOAA-Pacific/surveyjoin); its scripts and result tables will be released with the version of this work that reports it.

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, and support drafting the R code for the benchmark described in Section 4.3. All literature searching, screening, coding and evidence grading were performed by the authors; every citation and every numerical value was verified by the lead author against the primary source; and all code was executed and checked by the lead author. 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-Arroita, 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 Kakaï, 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), 260–267.

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

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

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