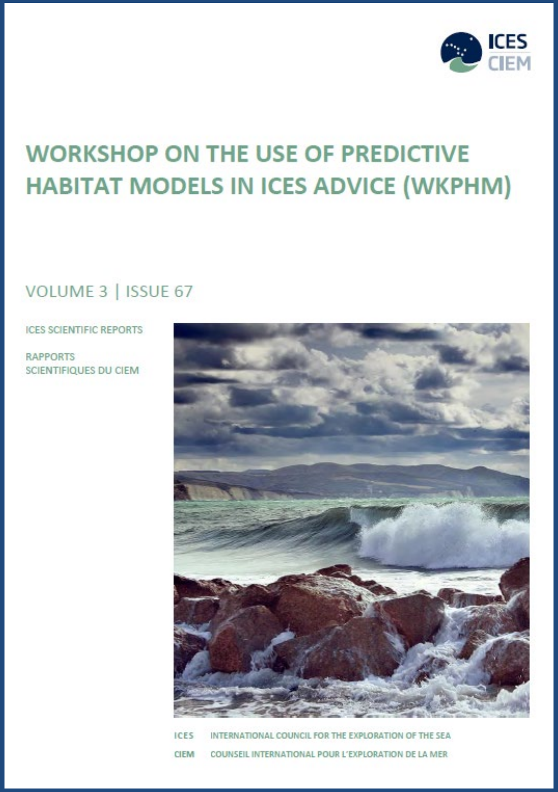
Using species distribution modeling to predict deep-sea coral and sponge communities, hotspots, diversity and indicators

CN Rooper (DFO)
ICES WKPHM
MF Sigler (NOAA)
P Thompson (DFO)
O Gemmell (SFU)

PICES Workshop W1
Sept. 24, 2022



ICES
CIEM

WORKSHOP ON THE USE OF PREDICTIVE HABITAT MODELS IN ICES ADVICE (WKPHM)

VOLUME 3 | ISSUE 67

ICES SCIENTIFIC REPORTS

RAPPORTS SCIENTIFIQUES DU CIEM

ICES INTERNATIONAL COUNCIL FOR THE EXPLORATION OF THE SEA
CIEM CONSEIL INTERNATIONAL POUR L'EXPLORATION DE LA MER

February 1-5, 2021
~30 participants

Objectives of the workshop:

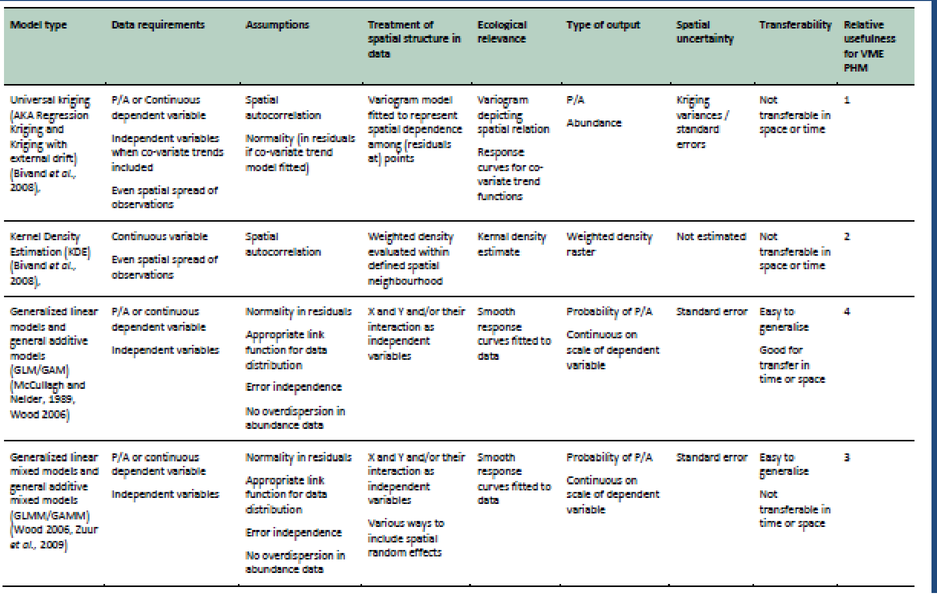
Identify the methods for modelling the distribution of VMEs that would be most appropriate for use within ICES advice

Detail ‘required’ and ‘desirable’ criteria in data, model techniques, display of results, validation and performance

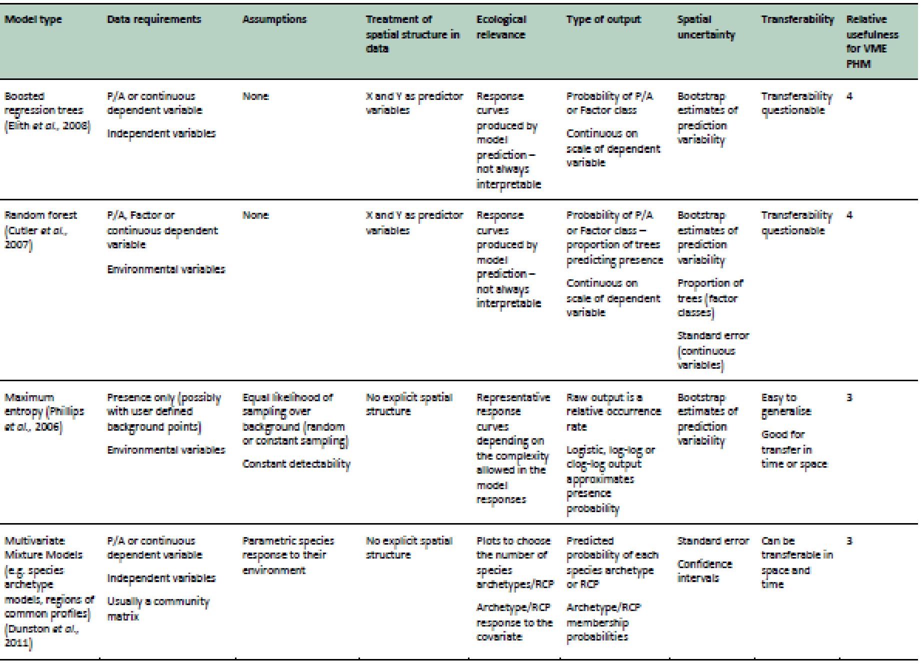
Develop clear standards for recording the caveats and assumptions inherent in the modelling method

Review and recommend a set of criteria, similar to the existing ICES benchmarking system for regional fish stock assessments, under which new and existing predictive habitat models can be used for ICES scientific advice related to the distribution of VMEs

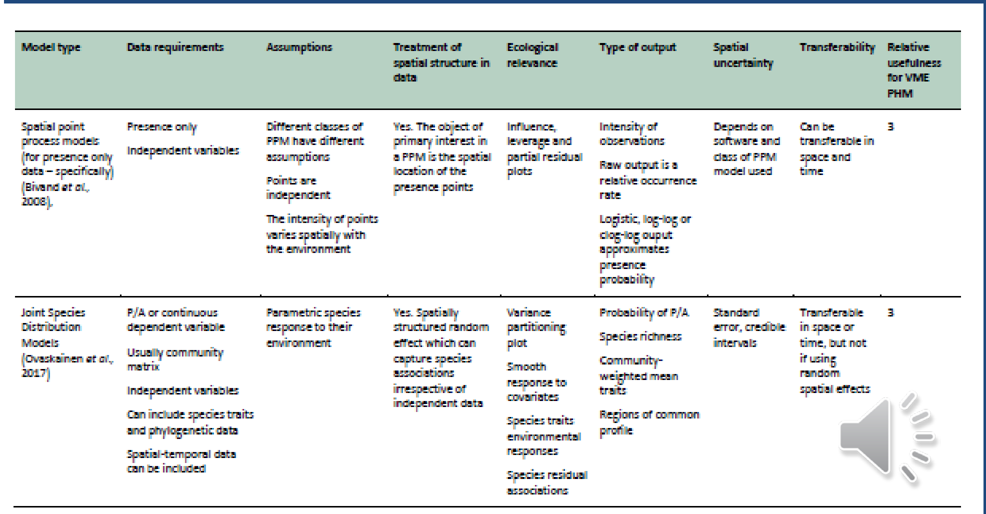
Model types and usefulness for VME



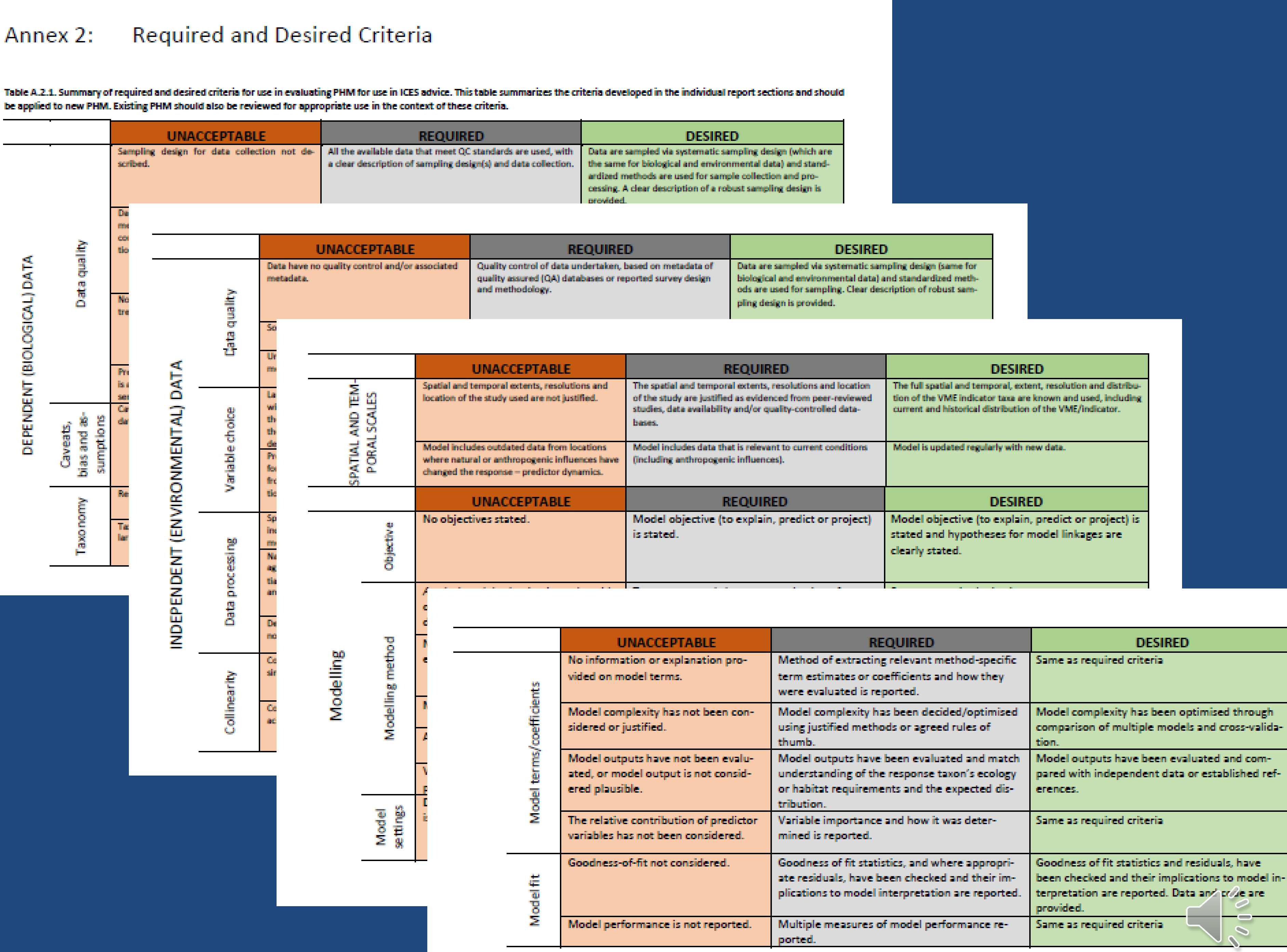
Model type	Data requirements	Assumptions	Treatment of spatial structure in data	Ecological relevance	Type of output	Spatial uncertainty	Transferability	Relative usefulness for VME PHM
Universal kriging (AKA Regression Kriging and Kriging with external drift) (Bivand *et al.*, 2008).	P/A or Continuous dependent variable Independent variables when co-variate trends included Even spatial spread of observations	Spatial autocorrelation Normality (in residuals if co-variate trend model fitted)	Variogram model fitted to represent spatial dependence among (residuals at) points	Variogram depicting spatial relation Response curves for co-variate trend functions	P/A Abundance	Kriging variances / standard errors	Not transferable in space or time	1
Kernel Density Estimation (KDE) (Bivand *et al.*, 2008).	Continuous variable Even spatial spread of observations	Spatial autocorrelation	Weighted density evaluated within defined spatial neighbourhood	Kernal density estimate	Weighted density raster	Not estimated	Not transferable in space or time	2
Generalized linear models and general additive models (GLM/GAM) (McCullagh and Nelder, 1989, Wood 2006)	P/A or continuous dependent variable Independent variables	Normality in residuals Appropriate link function for data distribution Error independence No overdispersion in abundance data	X and Y and/or their interaction as independent variables	Smooth response curves fitted to data	Probability of P/A Continuous on scale of dependent variable	Standard error	Easy to generalise Good for transfer in time or space	4
Generalized linear mixed models and general additive mixed models (GLMM/GAMM) (Wood 2006, Zuur *et al.*, 2009)	P/A or continuous dependent variable Independent variables	Normality in residuals Appropriate link function for data distribution Error independence No overdispersion in abundance data	X and Y and/or their interaction as independent variables Various ways to include spatial random effects	Smooth response curves fitted to data	Probability of P/A Continuous on scale of dependent variable	Standard error	Easy to generalise Not transferable in time or space	3



Model type	Data requirements	Assumptions	Treatment of spatial structure in data	Ecological relevance	Type of output	Spatial uncertainty	Transferability	Relative usefulness for VME PHM
Boosted regression trees (Elith *et al.*, 2008)	P/A or continuous dependent variable Independent variables	None	X and Y as predictor variables	Response curves produced by model prediction – not always interpretable	Probability of P/A or Factor class Continuous on scale of dependent variable	Bootstrap estimates of prediction variability	Transferability questionable	4
Random forest (Cutler *et al.*, 2007)	P/A, Factor or continuous dependent variable Environmental variables	None	X and Y as predictor variables	Response curves produced by model prediction – not always interpretable	Probability of P/A or Factor class – proportion of trees predicting presence Continuous on scale of dependent variable	Bootstrap estimates of prediction variability Proportion of trees (factor classes) Standard error (continuous variables)	Transferability questionable	4
Maximum entropy (Phillips *et al.*, 2006)	Presence only (possibly with user defined background points) Environmental variables	Equal likelihood of sampling over background (random or constant sampling) Constant detectability	No explicit spatial structure	Representative response curves depending on the complexity allowed in the model responses	Raw output is a relative occurrence rate Logistic, log-log or clog-log output approximates presence probability	Bootstrap estimates of prediction variability	Easy to generalise Good for transfer in time or space	3
Multivariate Mixture Models (e.g. species archetype models, regions of common profiles) (Dunstan *et al.*, 2011)	P/A or continuous dependent variable Independent variables Usually a community matrix	Parametric species response to their environment	No explicit spatial structure	Plots to choose the number of species archetypes/RCP Archetype/RCP response to the covariate	Predicted probability of each species archetype or RCP Archetype/RCP membership probabilities	Standard error Confidence intervals	Can be transferable in space and time	3



Model type	Data requirements	Assumptions	Treatment of spatial structure in data	Ecological relevance	Type of output	Spatial uncertainty	Transferability	Relative usefulness for VME PHM
Spatial point process models (for presence only data – specifically) (Bivand *et al.*, 2008).	Presence only Independent variables	Different classes of PPM have different assumptions Points are independent The intensity of points varies spatially with the environment	Yes. The object of primary interest in a PPM is the spatial location of the presence points	Influence, leverage and partial residual plots	Intensity of observations Raw output is a relative occurrence rate Logistic, log-log or clog-log ouput approximates presence probability	Depends on software and class of PPM model used	Can be transferable in space and time	3
Joint Species Distribution Models (Ovaskainen *et al.*, 2017)	P/A or continuous dependent variable Usually community matrix Independent variables Can include species traits and phylogenetic data Spatial-temporal data can be included	Parametric species response to their environment	Yes. Spatially structured random effect which can capture species associations irrespective of independent data	Variance partitioning plot Smooth response to covariates Species traits environmental responses Species residual associations	Probability of P/A Species richness Community-weighted mean traits Regions of common profile	Standard error, credible intervals	Transferable in space or time, but not if using random spatial effects	3



Annex 2: Required and Desired Criteria

Table A.2.1. Summary of required and desired criteria for use in evaluating PHM for use in ICES advice. This table summarizes the criteria developed in the individual report sections and should be applied to new PHM. Existing PHM should also be reviewed for appropriate use in the context of these criteria.

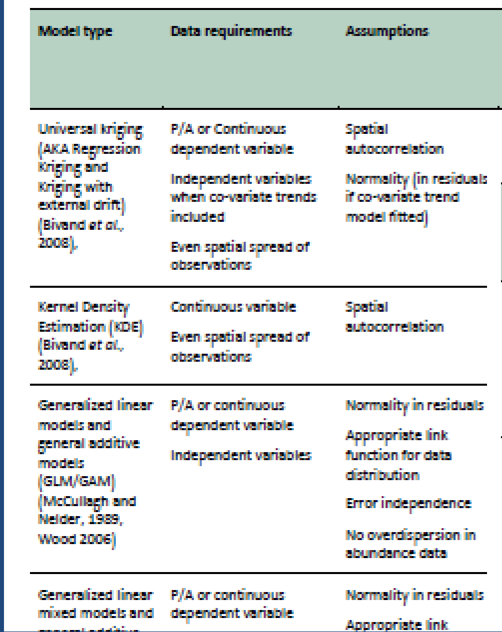
		UNACCEPTABLE	REQUIRED	DESIRED
DEPENDENT (BIOLOGICAL) DATA	Data quality	Sampling design for data collection not described.	All the available data that meet QC standards are used, with a clear description of sampling design(s) and data collection.	Data are sampled via systematic sampling design (which are the same for biological and environmental data) and standardized methods are used for sample collection and processing. A clear description of a robust sampling design is provided.
	Caveats, bias and assumptions			
	Taxonomy			

		UNACCEPTABLE	REQUIRED	DESIRED
INDEPENDENT (ENVIRONMENTAL) DATA	Data quality	Data have no quality control and/or associated metadata.	Quality control of data undertaken, based on metadata of quality assured (QA) databases or reported survey design and methodology.	Data are sampled via systematic sampling design (same for biological and environmental data) and standardized methods are used for sampling. Clear description of robust sampling design is provided.
	Variable choice			
	Data processing			
	Collinearity			

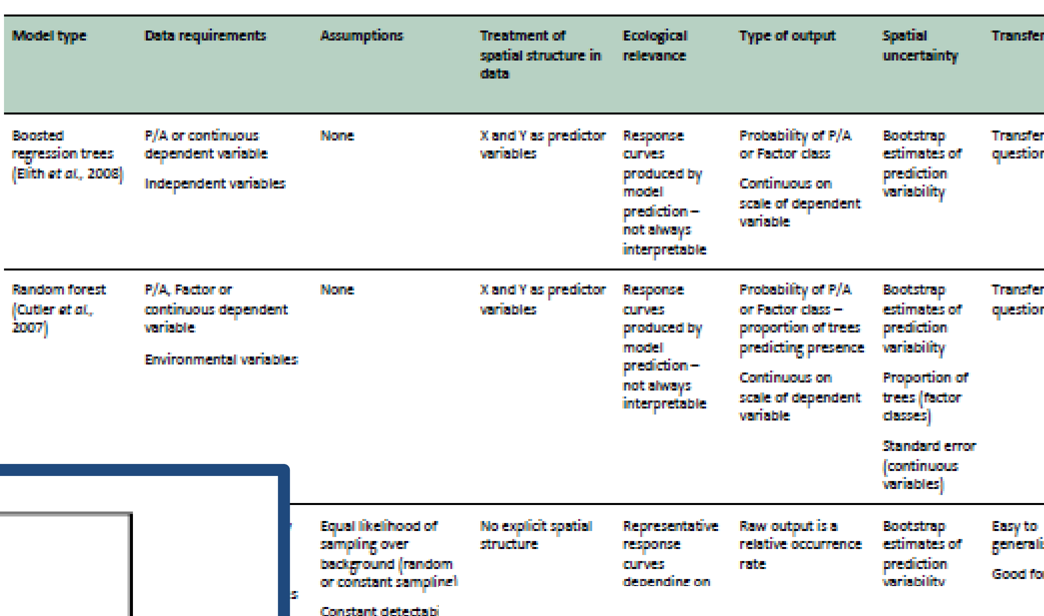
	UNACCEPTABLE	REQUIRED	DESIRED
SPATIAL AND TEMPORAL SCALES	Spatial and temporal extents, resolutions and location of the study used are not justified.	The spatial and temporal extents, resolutions and location of the study are justified as evidenced from peer-reviewed studies, data availability and/or quality-controlled databases.	The full spatial and temporal, extent, resolution and distribution of the VME indicator taxa are known and used, including current and historical distribution of the VME/indicator.
	Model includes outdated data from locations where natural or anthropogenic influences have changed the response – predictor dynamics.	Model includes data that is relevant to current conditions (including anthropogenic influences).	Model is updated regularly with new data.

		UNACCEPTABLE	REQUIRED	DESIRED
Modelling	Objective	No objectives stated.	Model objective (to explain, predict or project) is stated.	Model objective (to explain, predict or project) is stated and hypotheses for model linkages are clearly stated.
	Modelling method			
	Model settings			

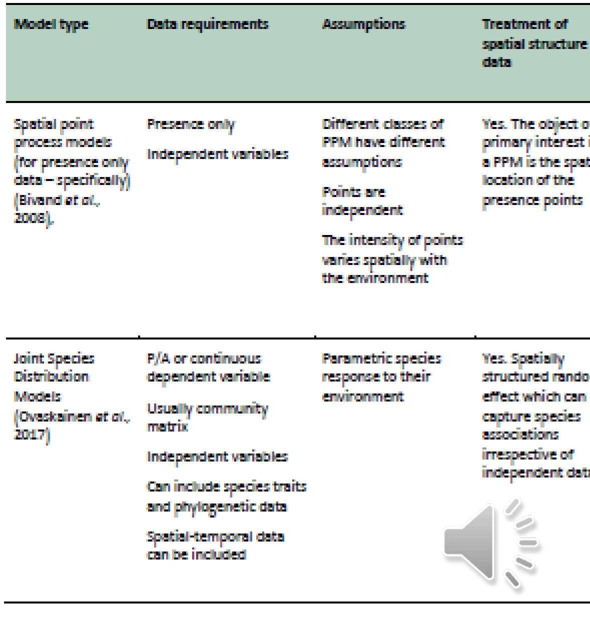
	UNACCEPTABLE	REQUIRED	DESIRED
Model terms/coefficients	No information or explanation provided on model terms.	Method of extracting relevant method-specific term estimates or coefficients and how they were evaluated is reported.	Same as required criteria
	Model complexity has not been considered or justified.	Model complexity has been decided/optimised using justified methods or agreed rules of thumb.	Model complexity has been optimised through comparison of multiple models and cross-validation.
	Model outputs have not been evaluated, or model output is not considered plausible.	Model outputs have been evaluated and match understanding of the response taxon's ecology or habitat requirements and the expected distribution.	Model outputs have been evaluated and compared with independent data or established references.
	The relative contribution of predictor variables has not been considered.	Variable importance and how it was determined is reported.	Same as required criteria
Model fit	Goodness-of-fit not considered.	Goodness of fit statistics, and where appropriate residuals, have been checked and their implications to model interpretation are reported.	Goodness of fit statistics and residuals, have been checked and their implications to model interpretation are reported. Data and [illegible] are provided.
	Model performance is not reported.	Multiple measures of model performance reported.	Same as required criteria



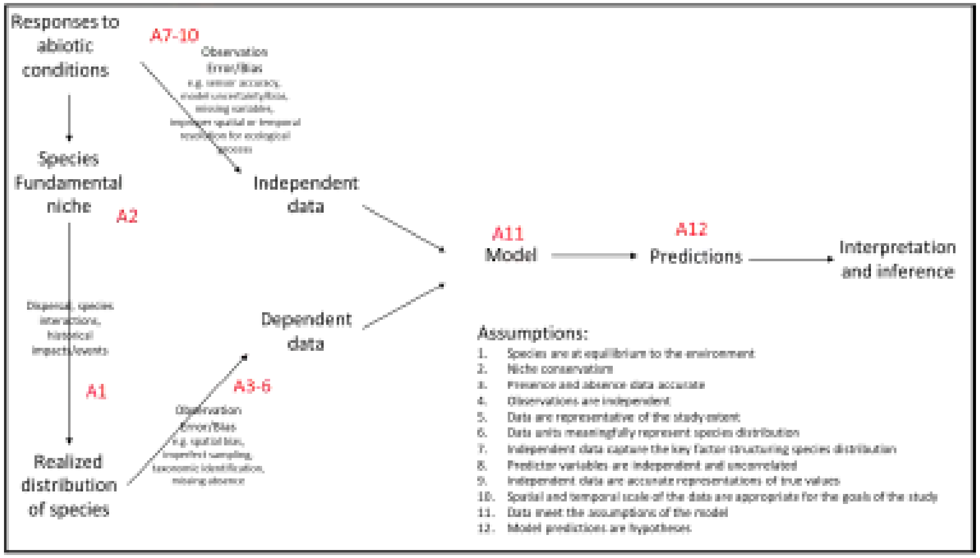
Model type	Data requirements	Assumptions	Treatment of spatial structure in data	Ecological relevance	Type of output	Spatial uncertainty	Transferability	Relative usefulness for VME PHM
Universal kriging (AKA Regression Kriging and Kriging with external drift) (Bivand *et al.*, 2008).	P/A or Continuous dependent variable Independent variables when co-variate trends included Even spatial spread of observations	Spatial autocorrelation Normality (in residuals if co-variate trend model fitted)	Variogram model fitted to represent	Variogram depicting	P/A Abundance	Kriging variances /	Not transferable in	1
Kernel Density Estimation (KDE) (Bivand *et al.*, 2008).	Continuous variable Even spatial spread of observations	Spatial autocorrelation						
Generalized linear models and general additive models (GLM/GAM) (McCullagh and Nelder, 1989, Wood 2006)	P/A or continuous dependent variable Independent variables	Normality in residuals Appropriate link function for data distribution Error independence No overdispersion in abundance data						
Generalized linear mixed models and	P/A or continuous dependent variable	Normality in residuals Appropriate link						



Model type	Data requirements	Assumptions	Treatment of spatial structure in data	Ecological relevance	Type of output	Spatial uncertainty	Transfer
Boosted regression trees (Elith *et al.*, 2008)	P/A or continuous dependent variable Independent variables	None	X and Y as predictor variables	Response curves produced by model prediction – not always interpretable	Probability of P/A or Factor class Continuous on scale of dependent variable	Bootstrap estimates of prediction variability	Transfer question
Random forest (Cutler *et al.*, 2007)	P/A, Factor or continuous dependent variable Environmental variables	None	X and Y as predictor variables	Response curves produced by model prediction – not always interpretable	Probability of P/A or Factor class – proportion of trees predicting presence Continuous on scale of dependent variable	Bootstrap estimates of prediction variability Proportion of trees (factor classes) Standard error (continuous variables)	Transfer question
	s	Equal likelihood of sampling over background (random or constant sampling) Constant detectabi	No explicit spatial structure	Representative response curves depending on	Raw output is a relative occurrence rate	Bootstrap estimates of prediction variability	Easy to generali Good fo
		Parametric species response to their environment					



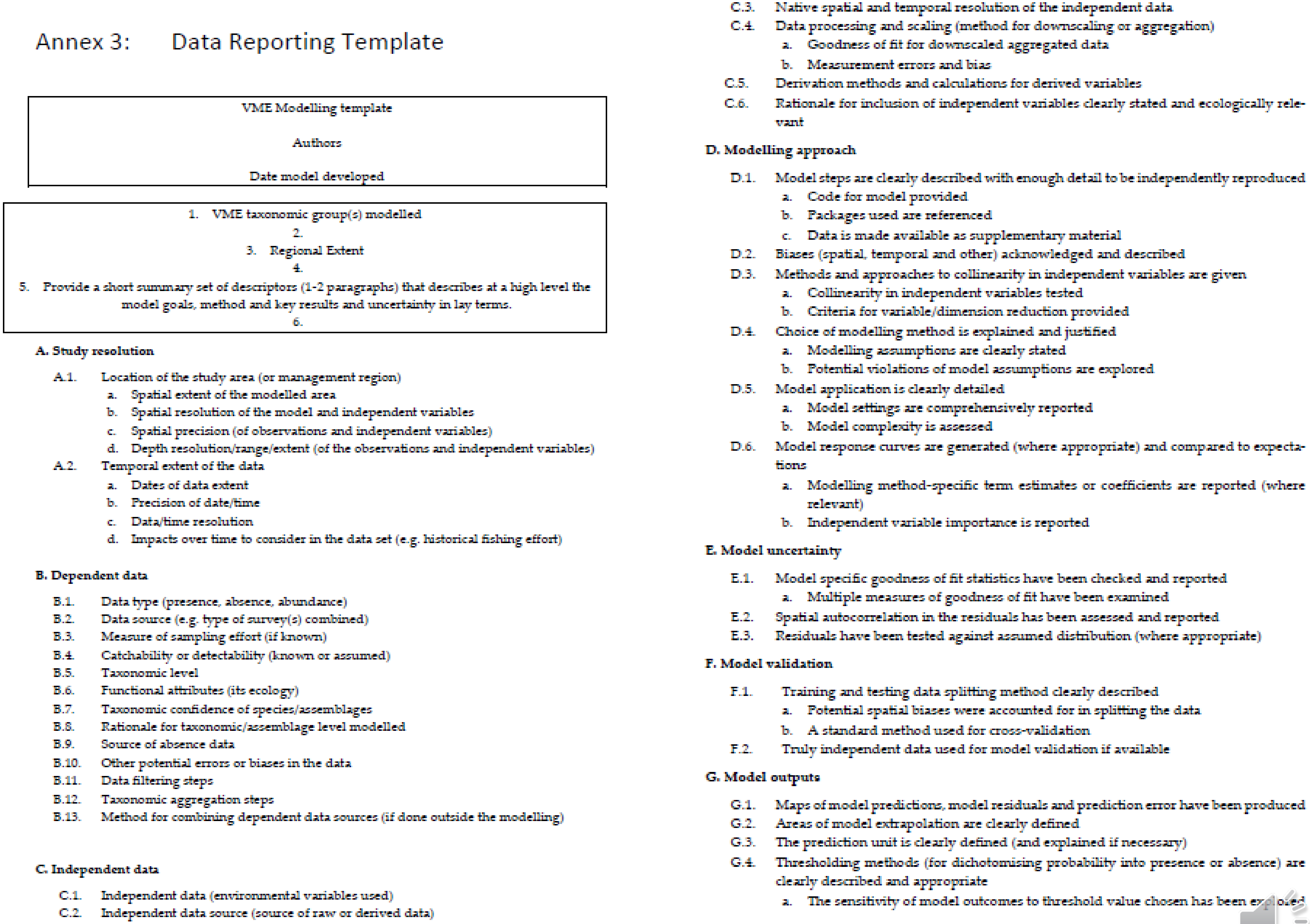
Model type	Data requirements	Assumptions	Treatment of spatial structure data
Spatial point process models (for presence only data – specifically) (Bivand *et al.*, 2008).	Presence only Independent variables	Different classes of PPM have different assumptions Points are independent The intensity of points varies spatially with the environment	Yes. The object of primary interest i a PPM is the spat location of the presence points
Joint Species Distribution Models (Ovaskainen *et al.*, 2017)	P/A or continuous dependent variable Usually community matrix Independent variables Can include species traits and phylogenetic data Spatial-temporal data can be included	Parametric species response to their environment	Yes. Spatially structured rando effect which can capture species associations irrespective of independent dat



Assumptions:

1. Species are at equilibrium to the environment
2. Niche conservation
3. Presence and absence data accurate
4. Observations are independent
5. Data are representative of the study extent
6. Data units meaningfully represent species distribution
7. Independent data capture the key factor structuring species distribution
8. Predictor variables are independent and uncorrelated
9. Independent data are accurate representations of true values
10. Spatial and temporal scale of the data are appropriate for the goals of the study
11. Data meet the assumptions of the model
12. Model predictions are hypotheses

Figure 2.2.1: Conceptual illustration of how the assumptions provided in Table 2.2.1 relate to the ecological processes that determine species distributions, the independent and dependent data, the model, the predictions, and interpretation and inference. The arrows illustrate the logical flow between these different components with indications of where each of the assumptions is relevant. The smaller text associated with the arrows provides examples of why the assumptions may not be met. The red text indicates the corresponding assumption number listed in the Assumptions list.



Annex 3: Data Reporting Template

VME Modelling template

Authors

Date model developed

1. VME taxonomic group(s) modelled
2.
3. Regional Extent
4.
5. Provide a short summary set of descriptors (1-2 paragraphs) that describes at a high level the model goals, method and key results and uncertainty in lay terms.
6.

A. Study resolution

A.1. Location of the study area (or management region)
  a. Spatial extent of the modelled area
  b. Spatial resolution of the model and independent variables
  c. Spatial precision (of observations and independent variables)
  d. Depth resolution/range/extent (of the observations and independent variables)
A.2. Temporal extent of the data
  a. Dates of data extent
  b. Precision of date/time
  c. Data/time resolution
  d. Impacts over time to consider in the data set (e.g. historical fishing effort)

B. Dependent data

B.1. Data type (presence, absence, abundance)
B.2. Data source (e.g. type of survey(s) combined)
B.3. Measure of sampling effort (if known)
B.4. Catchability or detectability (known or assumed)
B.5. Taxonomic level
B.6. Functional attributes (its ecology)
B.7. Taxonomic confidence of species/assemblages
B.8. Rationale for taxonomic/assemblage level modelled
B.9. Source of absence data
B.10. Other potential errors or biases in the data
B.11. Data filtering steps
B.12. Taxonomic aggregation steps
B.13. Method for combining dependent data sources (if done outside the modelling)

C. Independent data

C.1. Independent data (environmental variables used)
C.2. Independent data source (source of raw or derived data)
C.3. Native spatial and temporal resolution of the independent data
C.4. Data processing and scaling (method for downscaling or aggregation)
  a. Goodness of fit for downscaled aggregated data
  b. Measurement errors and bias
C.5. Derivation methods and calculations for derived variables
C.6. Rationale for inclusion of independent variables clearly stated and ecologically relevant

D. Modelling approach

D.1. Model steps are clearly described with enough detail to be independently reproduced
  a. Code for model provided
  b. Packages used are referenced
  c. Data is made available as supplementary material
D.2. Biases (spatial, temporal and other) acknowledged and described
D.3. Methods and approaches to collinearity in independent variables are given
  a. Collinearity in independent variables tested
  b. Criteria for variable/dimension reduction provided
D.4. Choice of modelling method is explained and justified
  a. Modelling assumptions are clearly stated
  b. Potential violations of model assumptions are explored
D.5. Model application is clearly detailed
  a. Model settings are comprehensively reported
  b. Model complexity is assessed
D.6. Model response curves are generated (where appropriate) and compared to expectations
  a. Modelling method-specific term estimates or coefficients are reported (where relevant)
  b. Independent variable importance is reported

E. Model uncertainty

E.1. Model specific goodness of fit statistics have been checked and reported
  a. Multiple measures of goodness of fit have been examined
E.2. Spatial autocorrelation in the residuals has been assessed and reported
E.3. Residuals have been tested against assumed distribution (where appropriate)

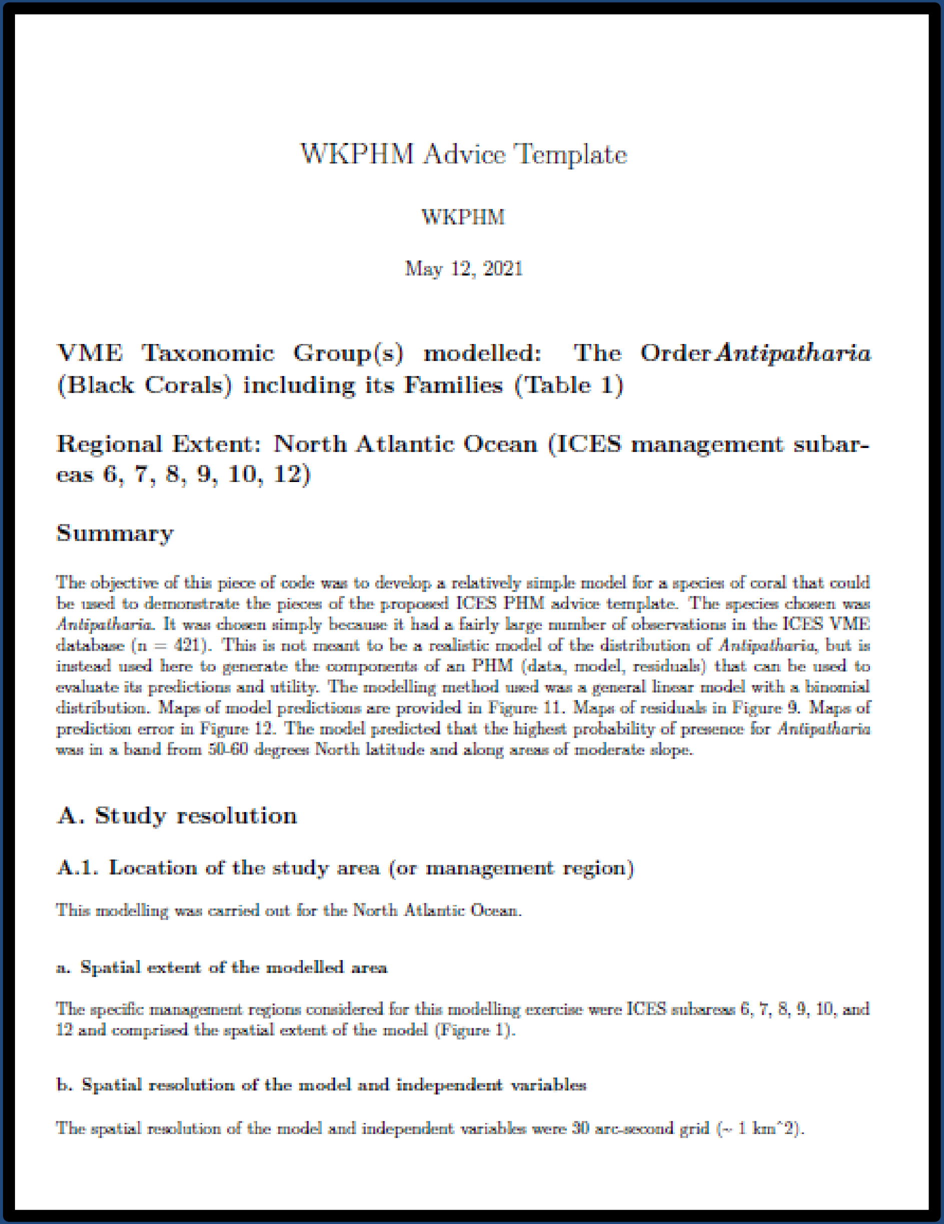
F. Model validation

F.1. Training and testing data splitting method clearly described
  a. Potential spatial biases were accounted for in splitting the data
  b. A standard method used for cross-validation
F.2. Truly independent data used for model validation if available

G. Model outputs

G.1. Maps of model predictions, model residuals and prediction error have been produced
G.2. Areas of model extrapolation are clearly defined
G.3. The prediction unit is clearly defined (and explained if necessary)
G.4. Thresholding methods (for dichotomising probability into presence or absence) are clearly described and appropriate
  a. The sensitivity of model outcomes to threshold value chosen has been explored

Example template and code



WKPHM Advice Template

WKPHM

May 12, 2021

VME Taxonomic Group(s) modelled: The Order *Antipatharia* (Black Corals) including its Families (Table 1)

Regional Extent: North Atlantic Ocean (ICES management subareas 6, 7, 8, 9, 10, 12)

Summary

The objective of this piece of code was to develop a relatively simple model for a species of coral that could be used to demonstrate the pieces of the proposed ICES PHM advice template. The species chosen was *Antipatharia*. It was chosen simply because it had a fairly large number of observations in the ICES VME database (n = 421). This is not meant to be a realistic model of the distribution of *Antipatharia*, but is instead used here to generate the components of an PHM (data, model, residuals) that can be used to evaluate its predictions and utility. The modelling method used was a general linear model with a binomial distribution. Maps of model predictions are provided in Figure 11. Maps of residuals in Figure 9. Maps of prediction error in Figure 12. The model predicted that the highest probability of presence for *Antipatharia* was in a band from 50-60 degrees North latitude and along areas of moderate slope.

A. Study resolution

A.1. Location of the study area (or management region)

This modelling was carried out for the North Atlantic Ocean.

a. Spatial extent of the modelled area

The specific management regions considered for this modelling exercise were ICES subareas 6, 7, 8, 9, 10, and 12 and comprised the spatial extent of the model (Figure 1).

b. Spatial resolution of the model and independent variables

The spatial resolution of the model and independent variables were 30 arc-second grid (~ 1 km^2).

https://github.com/ices-eg/WKPHM

Recommendations

- Transparency in data and methods
- Clearly state the objective of the PHM
- Include all available data that meets criteria and standards
- Collect independent data to validate model predictions
- Include existing and new models in developing ICES management advice
- Facilitate communication between science and management
- Develop a systematic approach to PHM in ICES

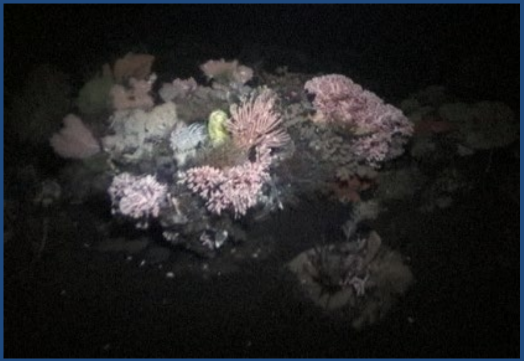
What to model (x & y variables)?

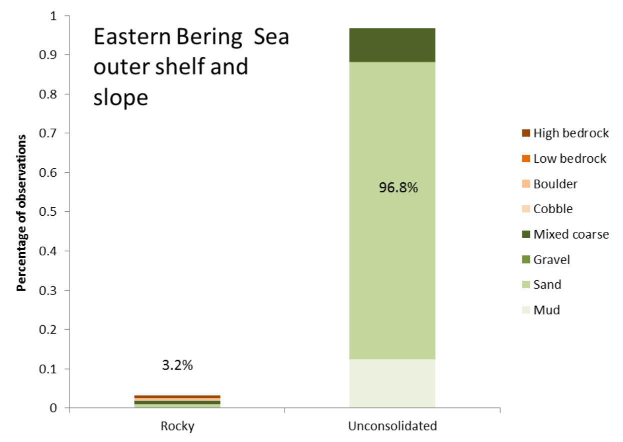
- Single taxa
- Multi-taxa
- Density hotspots
- Indicators
- Diversity

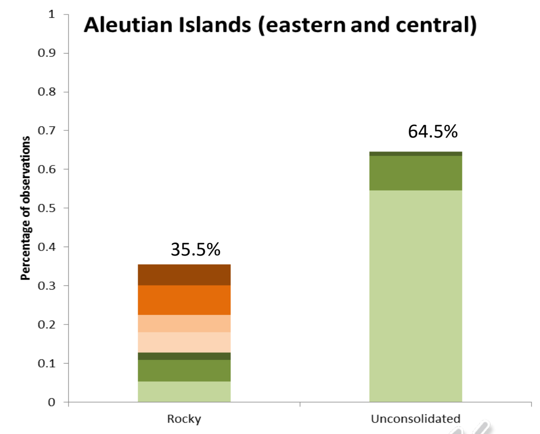
Presence/absence always better than presence only

Region	Transects with rocky habitat	Transects with coral
Gulf of Alaska	35%	30%
Aleutian Islands	63%	60%
Bowers Bank	42%	47%
Eastern Bering Sea	19%	13%

- Feasible mechanism
- Model reduction
- Often forced to use proxies for important variables

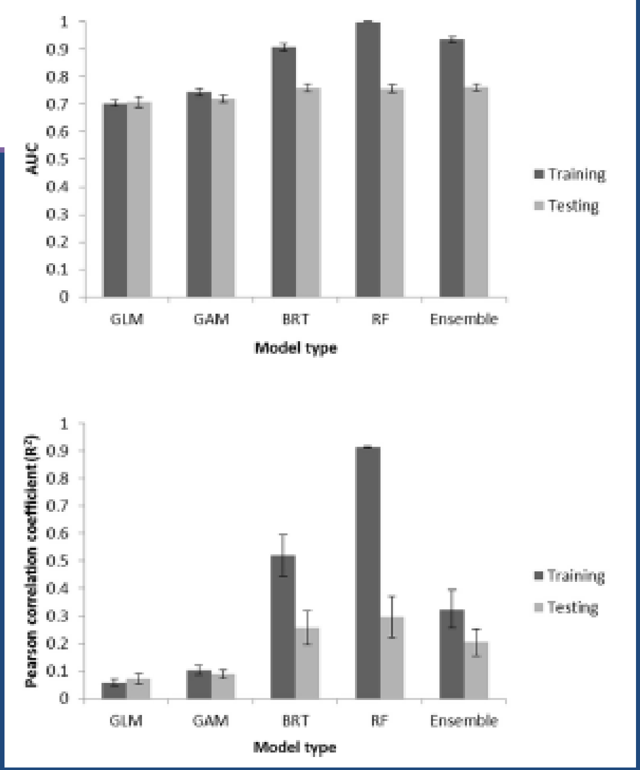


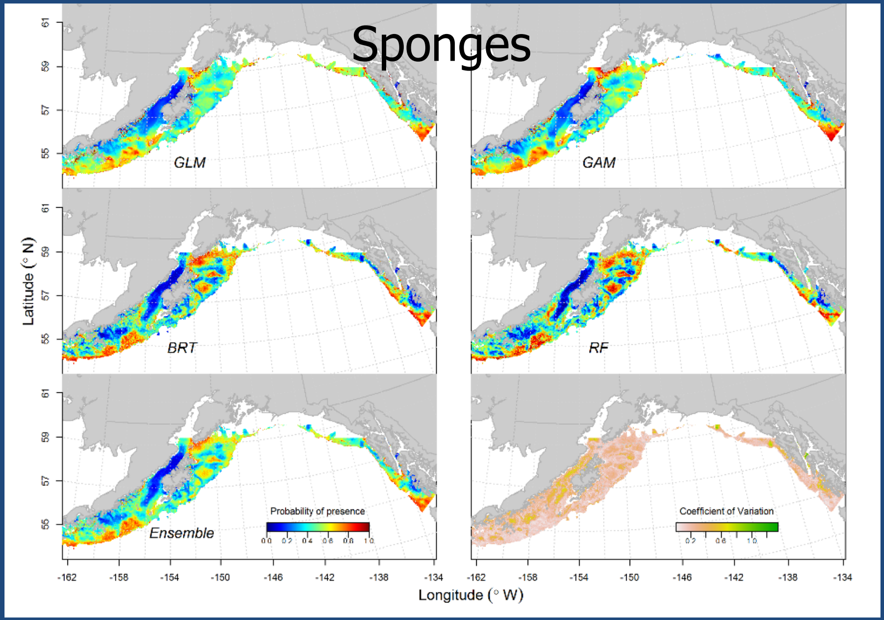


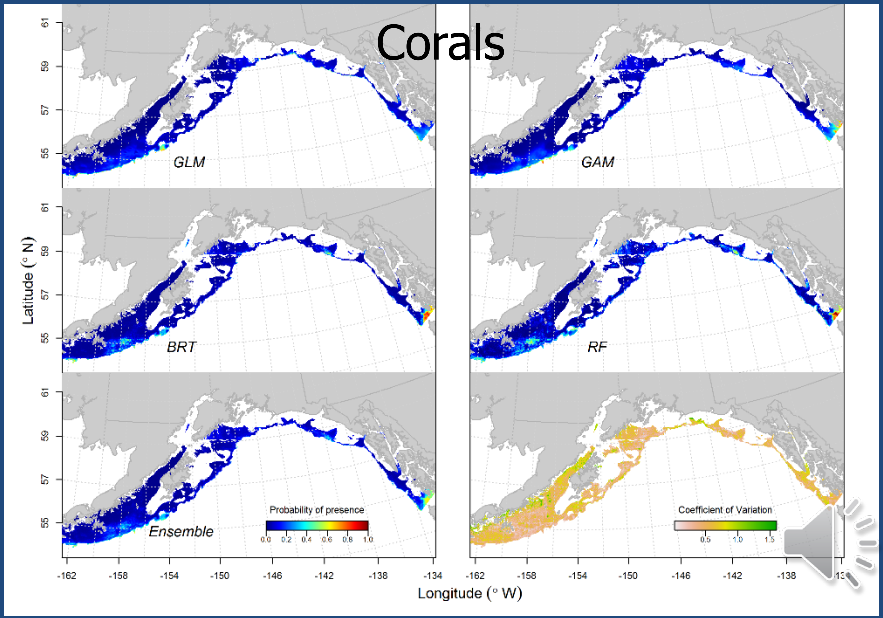


How to model (method)?

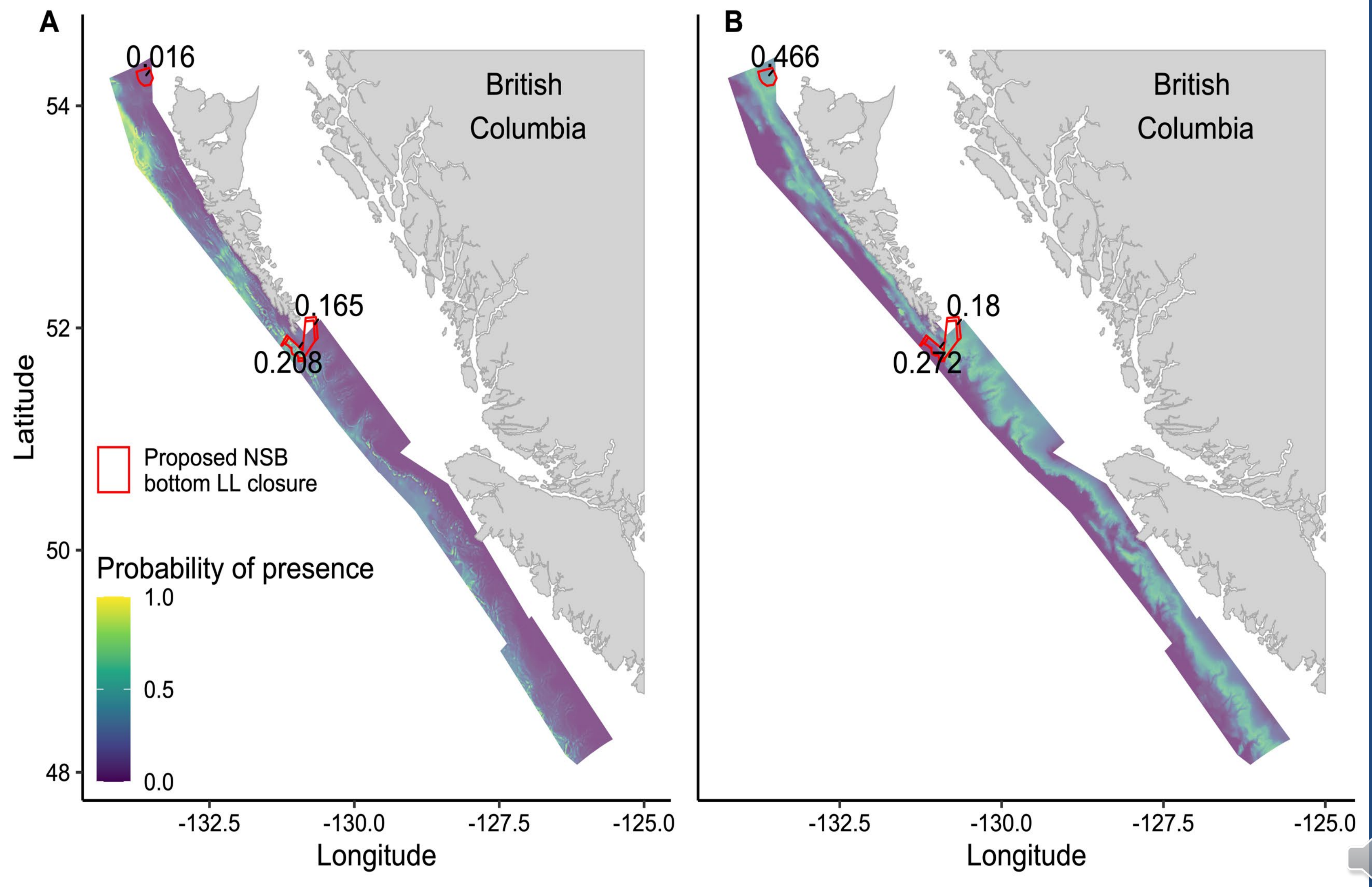
- Determined somewhat by data availability
- Maximum Entropy v. Statistical v. Machine Learning





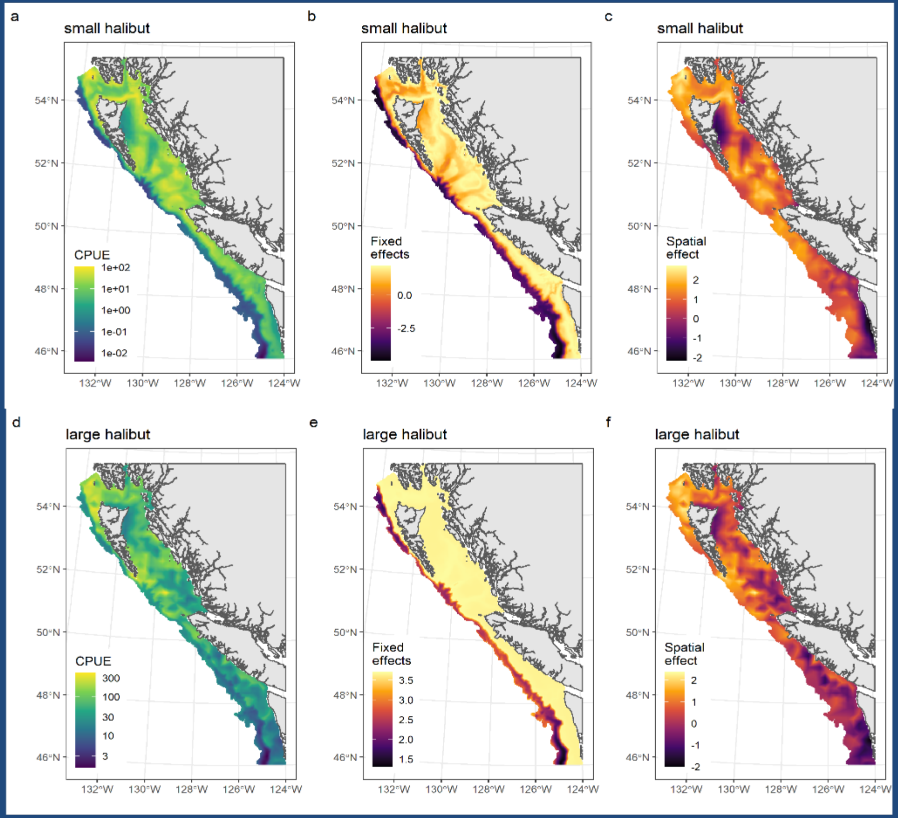


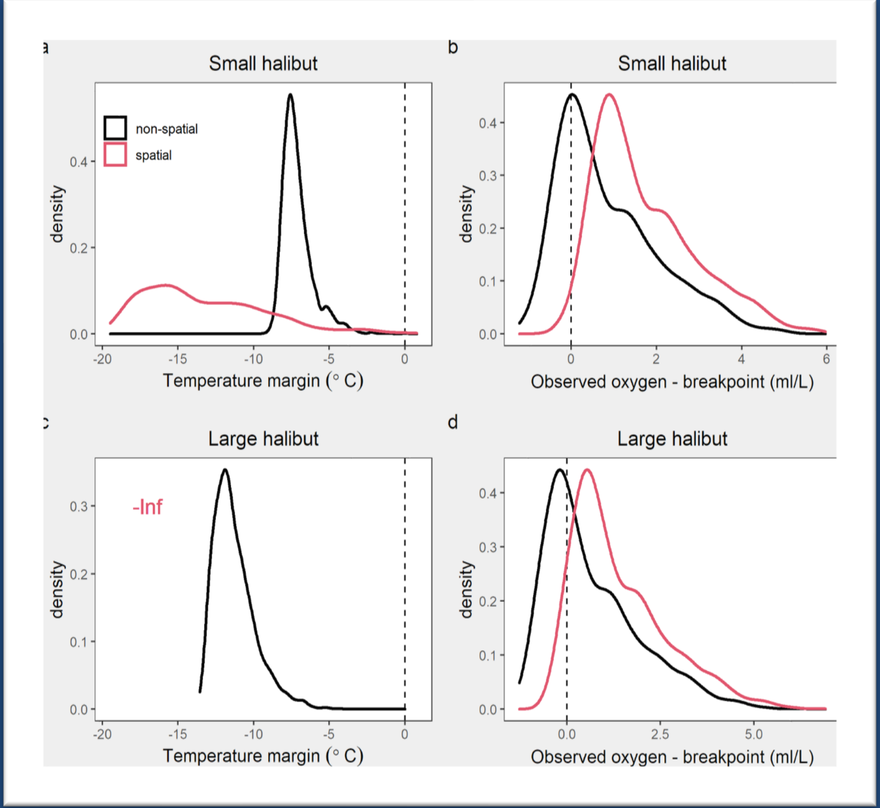
Problems with spatial patterns in the data



Accounting for unknown variables using spatial random fields (sdmTMB) – Pacific Halibut

Modeled group	R^2 (no spatial term)	R^2 (w/ spatial term)
Large halibut	0.07	0.53
Small halibut	0.13	0.33

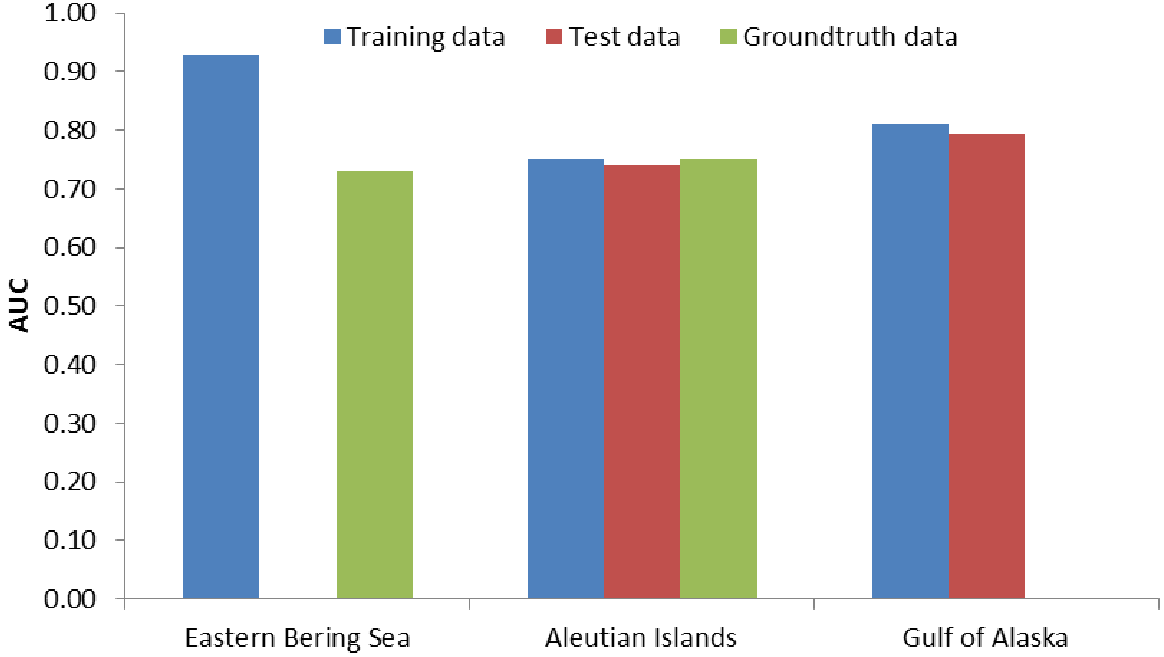




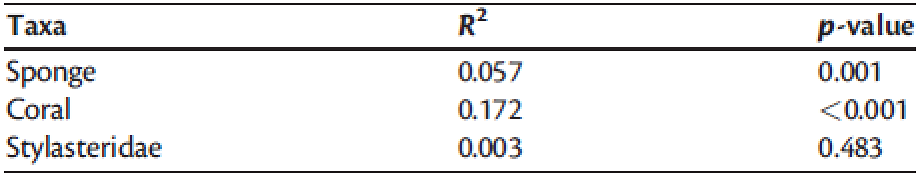
Thompson et al. in review

Model Fits to Independent Data

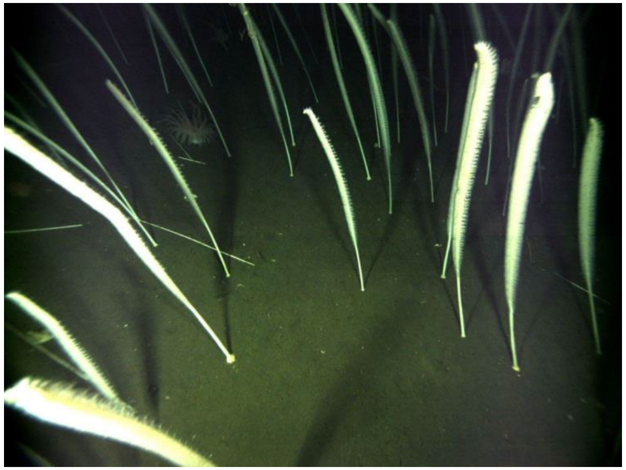
Presence/Absence models



Abundance models (AI only)



Taxa	R^2	*p*-value
Sponge	0.057	0.001
Coral	0.172	<0.001
Stylasteridae	0.003	0.483



Topics for discussion/lessons learned?

- The data is the only thing that matters
- Model predictions generally robust to method
- Validation is key to transmitting to management

Conclusions/Suggestions

- Most seamounts in the N Pacific have not been systematically surveyed
  - Mostly presence data from bycatch or targeted visual surveys
  - Shelf and slope relationships may not be applicable
- Both presence and absence data are needed from well designed surveys
- Substrate or proxies are the most important variables to know for coral and sponge SDM
- There are well thought out and reproducible guidelines for building SDM from the literature (beyond this ICES report)