Assessment of a large number of empirical plant Species Niche Models by elicitation of knowledge from two national experts

Simon M Smart[1,#], Susan Jarvis[1],Toshie Mizunuma[2], Cristina Herrero-Jáuregui[3], Zhou Fang[4], Adam Butler[4], Jamie Alison[5], Mike Wilson[1], Robert H Marrs[6]

1 NERC Centre for Ecology & Hydrology, Lancaster University campus, Library Avenue, Bailrigg, Lancaster, United Kingdom

2 Department of Botany, National Museum of Nature and Science, Amakubo 4-1-1, Tsukuba, Ibaraki 305-0005, Japan

3 Department of Ecology, Complutense University of Madrid, C/ José Antonio Novais 12 28040 Madrid, España

4 Biomathematics & Statistics Scotland, JCMB, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, United Kingdom

5 NERC Centre for Ecology & Hydrology, Deiniol Road, Bangor, Gwynedd, United Kingdom

6 School of Environmental Sciences, Nicholson Building, University of Liverpool, United Kingdom

 # Corresponding author: ssma@ceh.ac.uk, (SMS) Orcid ID: https://orcid.org/0000-0003-2750-7832

SS was funded in part by an Urgency Grant (NE/P003044/1) awarded by the Natural Environment Research Council The expert assessments were funded by a grant from the Botanical Society of Britain and Ireland.

Keywords: vascular plants, bryophytes, global change, forecasting, biodiversity, species distribution model, statistical model

Abstract

Quantitative models play an increasing role in exploring the impact of global change on biodiversity. To win credibility and trust they need validating. We show how expert knowledge can be used to assess a large number of empirical species niche models constructed for the British vascular plant and bryophyte flora. Key outcomes were; a) scored assessments of each modelled species and niche axis combination, b) guidance on models needing further development, c) exploration of the trade-off between presenting more complex model summaries, which could lead to more thorough validation, versus the longer time these take to evaluate, d) quantification of the internal consistency of expert opinion based on comparison of assessment scores made on a random subset of models evaluated by both experts. Overall, the experts assessed 39% of species and niche axis combinations to be 'poor' and 61% to show a degree of reliability split between 'moderate' (30%), 'good' (25%) and 'excellent' (6%). The two experts agreed in only 43% of cases, reaching greater consensus about poorer models and disagreeing most about models rated as better by either expert. This low agreement rate suggests that a greater number of experts is required to produce reliable assessments and to more fully understand the reasons underlying these differences of opinion. While AUC statistics showed generally very good ability of the models to predict random hold-out samples of the data there was no correspondence between these and the scores given by the experts and no apparent correlation between AUC and species prevalence. Crowd-sourcing further assessments by allowing web-based

access to model fits is an obvious next step. To this end we developed an on-line application for inspecting and evaluating the fit of each niche surface to its training data.

Introduction

Quantitative biodiversity models have become an important tool in our attempts to understand past ecological change and to predict what may lie ahead as humans increasingly dominate the Earth system (Ellis 2015). The development and application of ecological models is a burgeoning field yet producing models that are credible when applied in predictive mode and easy to use is a major challenge (Evans et al. 2013, Houlahan et al. 2017). Independent validation of the performance of models is critical if they are to win credibility and be deployed to address real problems. Recent decades have seen a rapid increase in the development and application of statistical Species Distribution or Species Niche Models (hereafter SNM) that reproduce the distributions of species based on correlative matching of presence/absence or presence-only datasets to environmental covariates (Elith & Leathwick 2009; Guillera-Arroita et al 2015). The advantage of such models is that they are easy to develop and apply. However, they have been criticised on a number of grounds: These include reliance on the assumption of niche conservatism as conditions change (Pearman et al 2007), inappropriate extrapolation to future potentially novel configurations of environmental conditions (Yates et al 2018;), omission of demographic processes and biotic interactions (Merow et al 2014; Zurrell et al 2009), omission of parameters linked to adaptive capacity such as phenotypic and genotypic variation and rate of likely evolution (Cartullo et al 2015). Building models that address these criticisms is essential but remains heavily data constrained given the number of

species of interest. Moreover, there is no guarantee of an improvement in accuracy even if models are trained on demographic data that ought to confer realistic dynamism (Crone et al 2011 but see Chapman et al 2014; Merow et al 2014). Therefore, empirical SNM are still likely to see continued development and use but in parallel with the move to accumulate and build more sophisticated hybrid models. Wise application of SNM is also fostered by the guidance emerging from a growing number of large scale tests of model transferability in space and time (Norberg et al 2019; Yates et al 2018; Dobrowski et al 2011; Pearman et al 2008).

The urgency of the problems typically addressed by SNM has also meant an increase in the formal inclusion of expert knowledge in model building (Low Choy et al 2009; Shirk et al 2010: Addison et al 2013) and testing (Drew & Perera 2012; van Zonneveld et al 2014). Confidence in the use of SNM increases if there is a degree of consensus between model predictions and independent expert judgement. Using statistical models of the realised niche of vascular plants and bryophytes in Britain, we investigated how expert opinion can be used to rapidly evaluate a large number of SNM that have been developed for a significant fraction of the British flora, covering all community dominants and numerous rare and subordinate species. The models are freely available within an R package called MultiMOVE (Henrys et al. 2015). It is more likely that these models will be used and gain credibility if they can be shown to reproduce the response of each plant species to major ecological gradients reliably. This can be done quantitatively, by testing the ability of each model to reproduce random samples of the training data, but also by seeking the view of experts not involved in model construction but who possess comprehensive knowledge of the British flora. In this paper we apply and compare the results of both approaches.

Each SNM in the MultiMOVE package is a statistical representation of the realised niche of each species across British ecosystems. That is, each niche is a modelled probability space defined by the main effects and interactions between climate, vegetation height, indicators of substrate pH, fertility and substrate wetness across the time interval in which the model-building data were collected. A large database of species presence-absence data from quadrat locations across Britain was used to build models for 1188 vascular plants and bryophytes (Fig 1). The availability of fine resolution co-located soil measurements lends the models potentially greater accuracy in defining each realised niche (Coudun et al 2006; Wamelink et al 2014) while also allowing models to be used to explore scenarios of environmental change that drive change in soil variables (De Vries et al 2010; Smart et al 2010b). Species presence/absence data used to build the models were available at relatively fine resolution (maximum 200m^2 (14.14 x 14.14m) to minimum 4m^2). This lessens the chance of poor model fit resulting from the averaging of environmental heterogeneity (Huston 1999). SNM were derived by fitting species presence and absence to the explanatory variables using five different statistical modelling techniques (Henrys et al. 2015). While the model development process is rigorous and scientific, in as much as it is clearly documented and therefore repeatable, it is not a given that each model represents he true realised niche of each species. For example, a model may be missing important predictors, there may be insufficient occurrences to parameterise the model, or the data may not fit the assumptions of the model. To address these issues, an ensemble of modelling techniques was used recognising that there is no single best statistical approach to specties niche modelling (Araújo and New 2006; Smart et al 2010b; Norberg et al 2019). Moreover, the notion that it is possible to define the 'true' realised niche as a spatially and

123 temporally invariant pattern is problematic even though the concept of the niche remains
124 extremely useful (Pulliam 2000, Chase and Liebold 2003, Araújo and Guisan 2006). We
125 assume pragmatically that the shape of each species' niche is stable enough to be usefully
126 approximated by popular niche modelling methods and, as we explore here, embodied in
127 the experiential knowledge that can be elicited from experts (Drew and Perera
128 2012;O'Hagan et al. 2006). Many of the species that we modelled have ranges that extend
129 into the European mainland. Restrictions on data availability resulted in models that only
130 included presence/absence for Britain thereby constraining the environmental range of
131 some of the models to a subset of their occupied area (c.f. Thuiller et al. 2004, McCune
132 2016, Yates et al. 2018). A useful consequence is that we did not require experts to
133 demonstrate knowledge of the ecological preferences of species outside Britain.
134
135 We report the results of a model assessment exercise carried out by two independent
136 expert botanists covering all niche axes of all species in the MultiMOVE R package (Fig 1).
137 Both experts were deemed sufficiently familiar with the habitat preferences of the British
138 flora to be able to judge the quality of each species' model as a representation of its realised
139 niche. Our aim was ultimately to generate species-specific guidance for users, alerting them
140 to potentially good and bad representations of the realised niche of each species and to
141 help identify models in need of improvement. Clearly, the experiential impression of each
142 niche can differ between experts depending upon the geographic and ecological scope of
143 their familiarity with British vegetation. In this respect, two experts are better than one but
144 not as good as an even greater number. We return to this issue in the discussion in light of
145 an analysis of the consistency between the two experts in their assessment results for a
146 random 5% sub-sample of the vascular plant species models.

147

148 The assessment made by the expert is also likely to be influenced by the methods used to
149 summarise model fit. Each species model can be thought of as comprising three
150 components each of which could be subjected to a separate assessment question: 1) Do the
151 response curves resulting from each of the five modelling techniques reproduce the
152 expected niche response of the species according to the experience of the expert? 2) Since
153 each model is fitted to a dataset of presences and absences does each model accurately
154 predict the observations that were used to build the model? 3) Does the observed
155 presence/absence data adequately represent the ecological range of the species in Britain?
156 A poor representation of the niche could for example arise from biased or unrepresentative
157 model-building data despite the model being a good fit to these data. Since a total of 1188
158 species models needed to be assessed we asked each expert to inspect the modelled
159 response to each abiotic niche axis averaged across model types rather than evaluating each
160 of the model types along each niche axis. Thus our principal objective was to address
161 question 1 via an inspection of the ability of each of the ensemble models to represent the
162 realised niche averaged across the five modelling techniques (Fig 1). We then address
163 question 2 by generating AUC statistics describing the fit of each model to random hold-out
164 samples of the training data. The correspondence between the experts' evaluations and the
165 model fit statistics were then compared with the expectation that better fitting models
166 should coincide with higher expert scores for the species and niche axis combinations
167 making up each model (Fig 1). In light of these results we discuss the trade-off between the
168 time required to evaluate more complex graphical representations of model fit versus the
169 possibility that more information-rich visualisations could yield more accurate and
170 comprehensive validation.

In summary, we sought to answer the following questions:

1. How did the two experts rate the ability of the models to capture the niche of each species?
2. To what extent did the experts agree with each other based on joint validation of a random sub-sample of the vascular plant models?
3. Did modelled species and niche axis combinations judged to be better representations of the species' niche coincide with higher quantitative model fit statistics for each species model?

Methods

Selection of experts

We circulated a request for experts to colleagues within the vegetation surveying community in Britain. Two experts were selected both of whom were prepared to commit themselves to the validation task. While we can assume that a greater number of experts should lead to more robust consensus (Drew & Perera 2012), our investigation was limited by the funding available to pay each expert for the large number of assessments required. A previous expert-based assessment of the habitat affinities of British plant species successfully employed three experts, hence we had no prior reason to expect that just two experts with comprehensive knowledge of the British flora would be insufficient (McInnes et al. 2017). However, In order to further identify the strengths and weaknesses of this approach we carried out a literature review of papers documenting the use of expert knowledge in validating statistical species distribution or niche models (Supplementary file

194 S1). We were especially interested in the range of variation in the ratio of experts to
195 numbers of species, and in conclusions as to the usefulness of expert assessment and the
196 levels of agreement found between experts and between experts and models.
197
198 The two expert botanists were recommended to us by colleagues. Both satisfied the six
199 criteria for selection of experts in elicitation studies listed by O'Hagan et al. (2006), a)
200 Tangible evidence of expertise, b) Reputation, c) Availability and willingness to participate,
201 d) Understanding of the problem area, e) Impartiality, f) Lack of an economic or personal
202 stake in the findings. Neither of the experts were previously acquainted with the authors
203 either in a personal or professional capacity. Both agreed to take part in the assessment
204 exercise and in doing so felt that their levels of botanical experience were sufficient to
205 tackle the national scope of the assessment. Their expertise and experience of the British
206 flora is summarised below:
207
208 Expert 1: This expert trained as a botanist and vegetation ecologist gaining a master degree
209 in ecology and then further plant identification qualifications from the British Natural
210 History Museum. The expert has 15 years' experience practicing as a professional botanist
211 and, in the last 8 years as a professional bryologist. The expert has been a vice-county
212 recorder for the Botanical Society of Britain and Ireland (BSBI) for the past 12 years and a
213 regional recorder for the British Bryological Society for 8 years.
214
215 Expert 2: This expert is a vegetation ecologist, bryologist and botanist with over 20 years'
216 experience in the nature conservation sector. The expert specialises in detailed vegetation
217 surveys especially the UK National Vegetation Classification, designing & implementing

vegetation monitoring programmes, training in identification and survey skills, bryophyte surveys and statistical analysis of ecological data.

In this instance, the two experts are not considered to be human research subjects in the sense of the Declaration of Helsinki and so it was not deemed necessary to seek approval and review by an Institutional Ethics Committee.

Assessment methodology

The modelled responses of each species along each of the seven niche axes were made available to each expert as a 'shiny' application (Chang et al. 2016) allowing each species to be selected by the expert for inspection and scoring via a user-friendly interface (see Fig S2.1 – Supplementary Material). The modelled response curve for each niche axis was plotted as the average of the predictions generated from the GLM, GAM, MARS and Neural Network models for the species. The Random Forest models were excluded because of the frequent occurrence of abrupt spikes in the modelled curves that were uninterpretable and probably reflected local over-fitting (Wenger and Olden 2012). The resource constraints of the project meant that only one average curve was plotted per niche axis rather than separate curves for each method with uncertainty intervals on each. Had we done so this would have increased the number of required assessments four-fold from 8316 to 33264 (1188 species * 7 niche axes * 4 model methods) and confronted the expert with a more complex representation of each niche that would have needed longer to evaluate. We return to this issue in the discussion. The modelled response curves were derived by solving each model for values of the respective predictor. The range of the predictor variable on each x-axis was defined by the maximum and minimum values in the complete training

242 dataset used to build the models and was therefore the same for every species assessed
243 (Henrys et al. 2015). Since each niche model included terms to be solved for other
244 predictors these also needed to contribute to the solution of each model along each
245 ecological gradient. This was done by setting the value of all other predictors to their
246 median value in the training data ; the default option in MultiMOVE. Hence, when
247 inspecting a species response along a single gradient, model predictions were generated by
248 varying the input values for this gradient only and fixing the input value for all other
249 covariates at the median of each covariate across the training data. An alternative approach
250 is to set the values of the background predictors to their observed values in each of the
251 sampled locations in the training data. We explore this option later in the paper. Raw
252 probabilities from each species' model were rescaled to account for varying prevalence in
253 the model-building data with the result that all values ranged between 0 and 1 (Real et al.
254 2006).
255
256 The experts were introduced to the use and installation of the software and the assessment
257 methodology via email and telephone. A guidance note on carrying out the assessment was
258 also circulated (see Supplementary Material). Bryophyte species (n=307) were assigned to
259 one of the experts who had particular experience of the British bryophyte flora. The vascular
260 plants (n=881) were split between the two experts at random. From this pool, 45 vascular
261 plants (5% of the total) were selected at random to be assessed by both experts. These were
262 included among the larger list given to each expert so that neither expert knew the identity
263 of the species that would also be inspected by the other. Experts were asked to assess the
264 accuracy of each niche axis using four categories; poor, moderate, good, excellent (
265 Supplementary file S1). No attempt was made to define this scale hence assessment was left

entirely to the judgement of the expert. The exact quote from the guidance note issued to each expert is as follows *"[The niche of each species is described in terms of seven environmental axes that are all shown together on each species page;] …..[You should evaluate each of these separately by comparing what the response curve implies about the species' preference with your experience of the species in British habitats. If unsure because you cannot understand the response or you suspect you do not have enough experience of the species' preferences throughout its range then don't hesitate to select 'Cannot evaluate']".*

Analysis

The results of the validation exercise are presented showing the frequency of species assigned to each class. The results for niche axes and species combinations that were assessed independently by both experts are presented as a confusion matrix showing the number of times the experts agreed and the frequency of disagreements by pairs of score; for example, by indicating how often expert 1 gave an assessment of 'good' when expert 2 gave an assessment of 'poor'. From these data % agreement was calculated as follows;

%agreement = (total number of identical assessments/ total number of assessments) * 100

By restricting the two sums above to just pairs containing one of the assessment categories, agreement values can also be readily calculated for each, showing for example whether experts were more likely to disagree when applying the 'excellent' score or the 'poor' score.

Comparison with quantitative model fit statistics

Area under the Receiver-Operator Curve (AUC) statistics for each species and each model type in the MultiMOVE ensemble were computed as follows: The presence absence data for each modelled species were split randomly into a 75% training and 25% test set. For each species and modelling method we train on the training set and predict the probability of presence on the test set. From this we calculated AUC values on the test set using the 'evaluate' function in the R package dismo (Hijmans et al. 2011). For each species and modelling method we repeated this process 10 times and extracted the average of the AUC values. Scatter plots and a loess smoother were used to explore whether the assessment category awarded by each expert to each species x niche axis combination varied systematically with the mean AUC of the respective species model. We would for example, expect models that best predicted a hold-out sample of their observations to be a better description of their niche and to attract a better assessment. This assumes that the observations used to build the model are representative of the species ecological range as perceived by each expert. Prevalence was plotted against mean AUC because the high true negative rates associated with species that rarely occur in the data would be expected to result in higher AUC values (Peterson et al. 2008, Lobo et al. 2007). The Area Under Curve (AUC) statistic is simply the area beneath the ROC curve, and provides a single value that is used to summarize overall performance (e.g. McCune 2016, Boria and Blois 2018, Yates et al. 2018).

Results and Discussion

Expert assessment results

312 Overall, the experts assessed 39% of niche axes to be ‘poor’ and 61% to show a degree of
313 reliability split between ‘moderate’ (30%), ‘good’ (25%) and ‘excellent’ (6%) (Fig 2A). The
314 two experts exhibited differing tendencies in their approach to model assessment. Expert 1
315 assigned a greater proportion of models to categories associated with stronger model
316 performance (Fig 2B). Expert 2 showed the reverse tendency, in particular assigning a much
317 greater proportion of modelled niche axes to the ‘poor’ category (Fig 2C). Since species
318 were allocated randomly these differences cannot be attributed to any prior ecological bias
319 in the species assessed. Expert 1 was the only expert to assess the bryophyte models. The
320 distribution of scores was similar to results for vascular plants; 36% of model axes being
321 considered ‘poor’, 28% ‘moderate’, 29% ‘good’ and 7% ‘excellent’ (Fig 2D).
322
323 Joint assessment of a 5% random sub-set of vascular plant models yielded 43% agreement
324 between experts. They were more likely to agree on the assessment of poor niche axes with
325 increasingly less consensus about niche axes considered to be better by at least one of the
326 experts (Table 1). These levels of disagreement are interesting; in 14 cases expert 2 assigned
327 ‘poor’ where expert 1 assigned ‘good’ and in 5 cases expert 1 assigned ‘poor’ where expert
328 2 gave ‘good’ consistent with the tendency for expert 2 to judge more harshly than expert 1.
329 In nine cases, disagreements centred on climate axes, in seven cases on the
330 succession/disturbance axis conveyed by vegetation height and in the remaining 3 cases on
331 abiotic substrate conditions. Species-specific examples of model fits are discussed below.
332 Model assessment scores for all species and niche axes are available in Supplementary
333 Material (S4).
334

335 ## Quantitative assessment of model fit

Mean AUC statistics for the species models were invariably greater than 0.8 with most species having scores greater than 0.9 suggesting good and excellent ability to predict the test data, respectively (Fig 3) (Swets 1988). A large number of absences tends to decrease the false positive rate thereby increasing AUC. Interestingly, while this effect cannot be ruled out, mean AUC was in fact lowest at the very lowest levels of prevalence. Regardless of the relationship between AUC and prevalence, there was no obvious difference in AUC between assessment categories for either expert (Fig 3). There was a weak indication that species models with higher AUC were more likely to be assigned as 'excellent' by expert 2. However, the smoothed lines did not differ by any meaningful amount (Fig 3b).

Assessment results in light of the literature review

We located 25 published papers that reported an independent assessment of statistical species distribution models using expert opinion (Supplementary file S1). Compared to these papers, our assessment involved by far the lowest ratio of experts to study organisms (1 to 307 for bryophytes and 1 to 881 for vascular plants with 45 species evaluated by both experts). It would however, be wrong to assume that these low ratios are an accurate measure of the fraction of knowledge that could be applied by each expert to each species in the assessment. The experts were chosen based on their experience and expertise in surveying British plant communities. As such, this experience should enable assessment of the habitat preferences of each of the species embedded within the mixed species assemblages widely encountered by the experts. Familiarity with the UK National Vegetation Classification by both experts also brings with it an awareness of the way many individual species respond to changing abiotic conditions within the context of the plant

community. We also encouraged the experts to select the 'cannot evaluate' category if they felt unable to evaluate a model through lack of experience. Even so, the levels of disagreement between the experts suggests that various unquantified biases may have influenced their judgement. For example, a species whose abiotic niche varies geographically will be wrongly evaluated if the expert's home-range did not include the full range of the species (Drew & Perera 2012; Murray et al. 2009; Supplementary file S1). In addition to these expert-centred sources of variation, we suspect that the simplicity of the univariate model summaries may have also mitigated against more accurate (nearer to the truth) and more precise (less uncertainty surrounding estimates of the truth) assessments.

Trade-offs between simple versus complex model summaries

At least three factors come into play when evaluating each model; i) the effectiveness of the way model fit was summarised for the expert, ii) the extent to which each model reproduces the observations used to build the model, iii) the extent to which the observational data adequately represents the ecological preferences of the species. The AUC statistics address the second issue. Across the prevalence range, mean AUC values indicated generally very good fits between the model predictions and hold-out samples of the training data. We might therefore have expected fewer 'poor' and 'moderate' expert assessment scores. The two experts were able to validate the fit of each species model to each abiotic axis based on a plot of the simple model average for the five model types across each separate niche axis. Raw predicted probabilities were also standardised to range between 0 and 1 thereby allowing species to be compared on an equal basis (Fig S1.1, S2 Supplementary file). This simple presentation was designed to make the assessment as

quick as possible. More realistic yet complex presentations are however possible, including graphing outputs from all available model types with attached confidence intervals rather than presenting just the average prediction. Expert assessors may have responded differently to such treatments but their complexity may well have meant prohibitively greater time spent on each assessment and additional training to help interpret more complex graphs. For example *Coeloglossum viride*, an orchid of shortly grazed calcareous grassland with an expected optimum at high pH and short vegetation height, was assessed by both experts. Plotting the predictions from each type of model shows how the model average can arise by combining models that are consistent with expectation versus models that completely fail to reproduce the expected ecological response (Fig 4). The inspection of the full range of models on the same graph would have allowed assessment and scoring of each model type as well as each axis however this will have meant a longer assessment process requiring significantly greater resourcing and training.

Further insight into the way each species model represents the realised niche can be gained from examining observed data and modelled occurrence simultaneously along more than one niche axis. Such plots are better able to reveal peaks in the probability of occurrence that are not visible when predictions are averaged for all other possible axes. For example the modelled maximum probability of occurrence for *C. viride* increases when the joint response to substrate pH and vegetation height is plotted (Fig 5A). T The result is a more accurate depiction of the modelled response for *C.viride* because its optimum is approximated more clearly by two rather than one niche axis (Fig 5A). The 2D plot highlights the dependence of the species on both pH and vegetation height, responses that are averaged out by examining only one dimension. However, had we presented these plots to

406 the experts for every pair of axes this would have increased the volume of assessment
407 material from seven graphs to 21 graphs per species.
408

409 The critical importance of the background variables

410 Another important difference in the way model responses can be summarised centres on
411 the choice of values for background variables; that is those explanatory variables other than
412 the ones that define the particular abiotic gradient being assessed. The default setting in
413 MultiMOVE is to set the background variables to the median for the input data. This
414 effectively holds all other variables constant allowing predictions to vary only in response to
415 the gradient of interest. However, the assessment results show that this can lead to
416 predictions being made for unrealistic combinations of explanatory variables while at the
417 same time missing those conditions that are optimal with respect to the observed
418 occurrences of the species. Turning again to *C.viride*, when all explanatory variables other
419 than pH and vegetation height are set to the median values for the training data
420 unrealistically high predictions are generated outside of the observed range of the species
421 and coinciding with vegetation that would appear too tall to be suitable (Fig 5B). Predicting
422 across the same two gradients but solving the model based on observed values at each
423 sampled location for all other explanatory variables results in the region of highest
424 prediction coinciding much more closely with the observed range of the species (Fig 5A).
425 This is a clearer test of the ability of the model to reproduce the abiotic responses in the
426 observations used to build the model. As such we must be clear that this is not a test of the
427 transferability of the model to predict new, independent observations (Wenger and Olden
428 2012; Yates et al 2018). Rather it is a validation of the fit of the model to the observations

upon which the model was based. The greatest difference between the two methods for introducing background variables is to be expected where a species exhibits multiple optima so that the median values of explanatory variables for the training data are not representative of any of the individual realised peaks in occurrence. *Schoenus nigricans*, a tussock-forming rush that has distinct ecological loci in base-rich soligenous mires in the low rainfall south east of Britain and in the lower pH, higher rainfall north west, is an example (Fig 6). Interestingly the model predicts lower values away from the high and low rainfall extremes despite a large number of observations being found in this range (Fig 6A). The model therefore appears to be a poor fit to the observations even though the observations are a reasonable representation of the ecological range of the species in these two dimensions. However, when based on median values for background explanatory variables the pattern is substantially worse (Fig 6B). The highest probabilities all occur outside of the observed ecological range of the species. Solving the models based on median background variables in the training data is therefore likely to have resulted in an assessment of poorer model fit to either axis than if model predictions were based on observed values at each sample point.

These considerations suggest that there are a number of ways of achieving improved model presentation for assessment . More complex yet information-rich summaries of the modelled niche are possible to produce but they are likely to take longer to evaluate. Surface plots showing observed presences overlaid with model predictions more clearly show the extent to which the small ensemble of model types has reproduced the observed data. Solving the models using observed values of explanatory variables for each location rather than median values across all locations also avoids applying unrealised and unrealistic

453 combinations of input variables that do not do justice to the fit of the model to
454 observations.
455

456 The value of expert elicitation

457 Human judgement is affected by a range of known biases (Tversky and Kahneman 1974,
458 McCarthy et al. 2004) and experts are no exception yet their opinions carry greater weight
459 than the non-expert and therefore have the potential for great benefit if correct (Ellenberg
460 2014) or grave disbenefit if false (Hill 2004). Having two experts assess our niche axes was
461 better than having one. Yet just as the power of the ensemble approach to modelling relies
462 on a consensus among models that reduces the eccentric influence of any one model
463 (Araújo and New 2006, Smart et al. 2010b) it would be desirable to have more experts carry
464 out the model assessment. The size of the task is large however, given the many species and
465 niche axis combinations. A way forward would be to expose the MultiMOVE models to
466 crowd-sourced expertise. We have implemented this step by presenting bivariate modelled
467 niche surfaces and associated training data in a publicly available online application
468 (https://shiny-apps.ceh.ac.uk/find_your_niche/). Here assessments can now be captured
469 along with a self-reported indicator of level of expertise. Such an approach allows for more
470 complex yet informative model summaries to be presented since volunteer assessors can
471 take as much or as little time as required for each species of interest. The disadvantage is
472 that no prior control can be exercised over the expertise of the assessor.
473

474 Our results show that statistical and expert assessments of models can be very different for
475 a number of reasons: models can be a poor representation of the phenomena of interest
476 but fit their training data well indicating that the shortcoming is with the observations

rather than the modelling method. In addition, simple model summarises, designed to be readily evaluated by the ecologist but non-expert in statistics and modelling, can be over-simplifications. Moreover, experts may have too much faith in the transferability of their own expertise. Our results also confirm the variation that can occur among experts when asked the same question despite their expertise ostensibly covering the same knowledge domain; in this instance the habitat preferences of the British vascular plant flora (e.g. Gastón et al 2014; Murray et al 2009; Supplementary file S1). Having more experts assess the models becomes an obvious requirement when a small number fail to reach consensus. The key lessons from our investigation are a) that a robust consensus among experts should be based on as large a number of experts as possible, b) that excessively simple model summaries should be avoided even though this will necessitate additional time for assessment and additional training of experts to interpret more complex model summaries.

Data Availability

- The MultiMOVE R package is freely available via the Centre for Ecology & Hydrology data catalogue at https://doi.org/10.5285/94ae1a5a-2a28-4315-8d4b-35ae964fc3b9
- An on-line shiny application for submitting assessments of the modelled niche surfaces for British plant species is available at https://shiny-apps.ceh.ac.uk/find_your_niche/). This is best viewed in Chrome.

Acknowledgements

499 We thank the two anonymous experts who assessed the models and commented on a draft
500 of this paper; they can freely identify themselves. We do not do so in keeping with
501 maintaining their independence from this summary. We also thank the Botanical Society of
502 Britain and Ireland Research Fund for a grant to carry out the assessment and to Clive Lovatt
503 and Alex Lockton for handling the administration of the grant. We thank David Elston, two
504 anonymous referees and the journal editors for comments that much improved an earlier
505 version of the manuscript.

506

507 Supporting information

508 **S1 File.** Literature review of published papers involving expert assessment of species niche
509 models.
510 **S2 File.** Guidance provided to the two experts on the model validation process.
511 **S3 File.** Average model response curves for each of the species and niche axes discussed in
512 the text. These curves represent the information that was provided to each expert for
513 assessment.
514 **S4 File.** Excel file containing model validation results for all species and niche axis
515 combinations.

516

517

518 **References**

519 Addison, P.F.E., Rumpff, L., Sana Bau, S., Carey, J.M., En Chee, Y., Jarrad, F.C., McBride, M. &
520 Burgman, M.A. Practical solutions for making models indispensable in conservation
521 decision-making. Diversity and Distributions, 19, 490-502.
522
523 Araújo, M.B. & Guisan, A. 2006. Five (or so) challenges for species distribution modelling. – J.
524 Biogeogr. 33: 1677-1688.
525
526 Araújo, M.B. & New, M. 2006. Ensemble forecasting of species distributions. - Trends Ecol
527 Evol. 22: 42-47.
528
529 Boria, R.A. and Blois, J.L. 2018. The effect of large sample sizes on ecological niche models:
530 Analysis using a North American rodent, *Peromyscus maniculatus*. - Ecol. Mod. 386: 83-88.
531
532 Chang, W. et al. 2016. Shiny: Web Application Framework for R. R package version 0.14.1.
533 https://CRAN.R-project.org/package=shiny.
534
535 Chapman, D.S., Haynes, T., Beal, S., Essl, F. Bullock, J.M. 2014. Phenology predicts the native
536 and invasive range limits of common ragweed. *Glob. Change.Biol.* 20, 192-202.
537
538 Chase, J.M. and Liebold, M.A. 2003. Ecological Niches: linking classical and contemporary
539 approaches. Chicago: University of Chicago Press.
540
541

542 Coudun, C., Gegout, J-C., Piedallu, C., Rameau, J-C. 2006. Soil nutritional factors improve
543 models of plant species distribution: an illustration with *Acer campestre* (L.) in France.
544 Journal of Biogeography: 33, 1750-1763.

545

546 Crone, E.E. et al 2011. How do plant ecologists use matrix population models? Ecol.Letts. 14,
547 1-8.

548

549 de Vries, W, *et al* (2010). Use of dynamic soil-vegetation models to assess impacts of
550 nitrogen deposition on plant species composition and to estimate critical loads: an
551 overview. *Ecol. Applications*. **20**, 60-79.

552

553 Dobrowski, S.Z. et al. 2011. Modelling plant ranges over 75 years of climate change in
554 California, USA: temporal transferability and species traits. - Ecol. Monographs 81: 241-257.

555

556 Drew, C.A., Perera, A.H. 2012. Expert knowledge as a basis for landscape ecological
557 predictive models. In: A.H.Perera et al (eds.), *Expert Knowledge and its Application in*
558 *Landscape Ecology.* Springer, New York. Pgs 229-248.

559

560 Ellenberg, J. 2014. How not to be wrong: the hidden maths of everyday life. Chapter 1:
561 Abraham Wald and the missing bullet holes. London: Penguin Random House.

562

563 Ellis, EC. 2015. Ecology in an anthropogenic biosphere. - Ecol. Mon. 85: 287-331.

564

565 Elith, J. and Leathwick, J. R. (2009) 'Species distribution models: ecological explanation and
566 prediction across space and time', *Annual review of ecology, evolution, and systematics*.
567 Annual Reviews, 40, pp. 677–697.
568
569 Evans, M.R. et al. 2013. Predictive systems ecology. - Proc. R. Soc. Lond. B. 280: 20131452.
570
571 Gastón, A., García-Viñas, J.I., Bravo-Fernández, A.J., López-Leiva, C., Oliet, J.A., Roig, S.,
572 Serrada, R. 2014. Species distribution models applied to plant species selection in forest
573 restoration: are model predictions comparable to expert opinion? New Forests 45, 641-
574 653.
575
576 Guillera-Arroita, G. et al. 2015. Is my species distribution model fit for purpose? Matching
577 data and models to applications. Glob Ecol. Biogeogr. 24: 276-292.
578
579 Henrys, P.A. et al. 2015. Niche models for British plants and lichens obtained using an
580 ensemble approach. - New J. Bot. 5: 89-100.
581
582 Hijmans, R.J. et al. 2011. Package 'dismo'. 2011. Available online at: http://cran.r-
583 project.org/web/packages/dismo/index.html.
584
585 Hill, R. 2004. Multiple sudden infant deaths – coincidence or beyond coincidence? - Paed.
586 Peri. Epid. 18: 320-326.
587

588 Houlahan, J.E. et al. 2017. The priority of prediction in ecological understanding. – Oikos
589 126: 1-7.
590
591 Huston, M.A. 1999. Local processes and regional patterns: appropriate scales for
592 understanding variation in the diversity of plants and animals. *Oikos* **86**, 393-401.
593
594
595 Lobo, J.M. et al. 2007. AUC: a misleading measure of the performance of predictive
596 distribution models. - Glob. Ecol. Biogeo. 17: 145-151.
597
598 Low Choy, S., O'Leary, R., Mengersen, K. 2009. Elicitation by design in ecology: using expert
599 opinion to inform priors for Bayesian statistical models. Ecology 90, 265-277.
600
601 Martin, T.G. et al. 2012. Eliciting expert knowledge in conservation science. - Cons Biol. 26:
602 29–38.
603
604 McCarthy, M.A. et al. 2004. Comparing predictions of extinction risk using models and
605 subjective judgement. - Acta Oecol. 26: 67–74.
606
607 McCune, J.L. 2016. Species distribution models predict rare species occurrences despite
608 significant effects of landscape context. - J. Appl. Ecol. 53: 1871-1879.
609
610 McInnes, R.N. et al. 2017. Mapping allergenic pollen vegetation in UK to study
611 environmental exposure and human health. - Sci. Tot. Env. 599: 483-499.

612

613 Merow, C., Latimer, A.M., Wilson, A.M., McMahon, S.M., Rebelo, A.G., Silander Jr, J.A. 2014.
614 On using integral projection models to generate demographically driven predictions of
615 species' distributions: development and validation using sparse data. *Ecography* 37, 1167-
616 1183.

617

618 Murray, J.V., Goldizen, A.W., O'Leary, R.A., McAlpine, C.A., Possingham, H.A., Low Choy, S.
619 2009. How useful is expert opinion for predicting the distribution of a species within and
620 beyond the region of expertise? A case study using brush-tailed rock-wallabies *Petrogale*
621 *penicillata.* J.Appl.Ecol. 46: 842-851. 2009.

622

623 Norberg, A. et al. 2019. A comprehensive evaluation of predictive performance of 33 species
624 distribution models at species and community levels. Ecol. Monographs 89: e01370

625

626 O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley,
627 J.E., Rakow, T. 2006. Uncertain judgements: Eliciting Experts' Probabilities. John Wiley &
628 Sons Ltd.

629

630 Pearman, P.B. et al. 2008. Prediction of plant species distributions across six millennia. -
631 Ecol.Letts. 11: 357-369.

632

633 Pearman, P.B., Guisan, A., Broennimann, O., Randin, C.F. 2007. Niche dynamics in space and
634 time. Trends.Ecol.Evol. 23, 149-158.

635

636 Peterson, A.T. et al. 2008. Rethinking receiver operating characteristic analysis applications
637 in ecological niche modelling. – Ecol. Mod. 213, 63-72.
638
639 Pulliam, H.R. 2000. On the relationship between niche and distribution. - Ecol. Letts. 3: 349-
640 361.
641
642 Real, R. et al. 2006. Obtaining environmental favourability functions from logistic regression.
643 - Environ. Ecol. Stat. 13: 237-245.
644
645 Shirk, A.J., Wallin, D.O., Cushman, S.A., Rice, C.G., Warheit, K.I. 2010. Inferring landscape
646 effects on gene flow: a new model selection framework. Molecular Ecology 19: 3603-3619.
647
648 Smart, S.M. et al. 2010a. Empirical realized niche models for British higher and lower plants
649 – development and preliminary testing. - J. Veg. Sci. 21: 643-656.
650
651 Smart, S.M. et al. 2010b. Impacts of pollution and climate change on ombrotrophic
652 *Sphagnum* species in the UK: analysis of uncertainties in two empirical niche models. - Clim
653 Res. 45: 163-177.
654
655 Swets, J. 1988. Measuring the accuracy of diagnostic systems. - Science 240: 1285–1293.
656
657 Thuiller, W. et al 2004. Effects of restricting environmental range of data toproject current
658 and future species distributions. – Ecography 27: 165-172.
659

660 Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. -
661 Science 185: 1124-1131.
662
663 van Zonneveld, M., Castañeda, N., Scheldeman, X., van Etten, J., Van Damme, P. 2014.
664 Application of consensus theory to formalize expert evaluations of plant species distribution
665 models. Appl.Veg.Sci. 17, 528-542. 2014.
666
667 Wamelink, G.W.W., Goedhart, P.W. & Frissel, J.Y. 2014. Why some plant species are rare.
668 *PLoS ONE* 9(7): e102674.
669
670 Wenger, S.J. and Olden, J.D. 2012. Assessing transferability of ecological models: an under-
671 appreciated aspect of statistical validation. - Methods Ecol. Evol. 3: 260–267.
672
673 Yates, K.L. et al. 2018. Outstanding challenges in the transferability of ecological models. -
674 Trends Ecol. Evol. 33: 790-802.
675
676 Zurell, D., Jeltsch, F., Dormann, C.F., Schröder, B. 2009. Static species distribution models in
677 dynamically changing systems: how good can predictions really be? Ecography 32, 733–744
678

679 Table 1. Confusion matrix of results for species assessed by both experts. Numbers refer to
680 the count of niche axes and species combinations that were assessed. Thus the diagonal
681 gives the number of assessments where both experts agreed. The figure in brackets is the %
682 agreement for each category of score.

683

Expert 1 / Expert 2	excellent	good	moderate	poor	Expert 2 totals
excellent	2 (8)	2	1	1	6
good	9	16 (17)	7	5	37
moderate	9	39	44 (25)	14	106
poor	1	14	62	64 (40)	141
Expert 1 totals	21	71	114	84	126 (43)

684

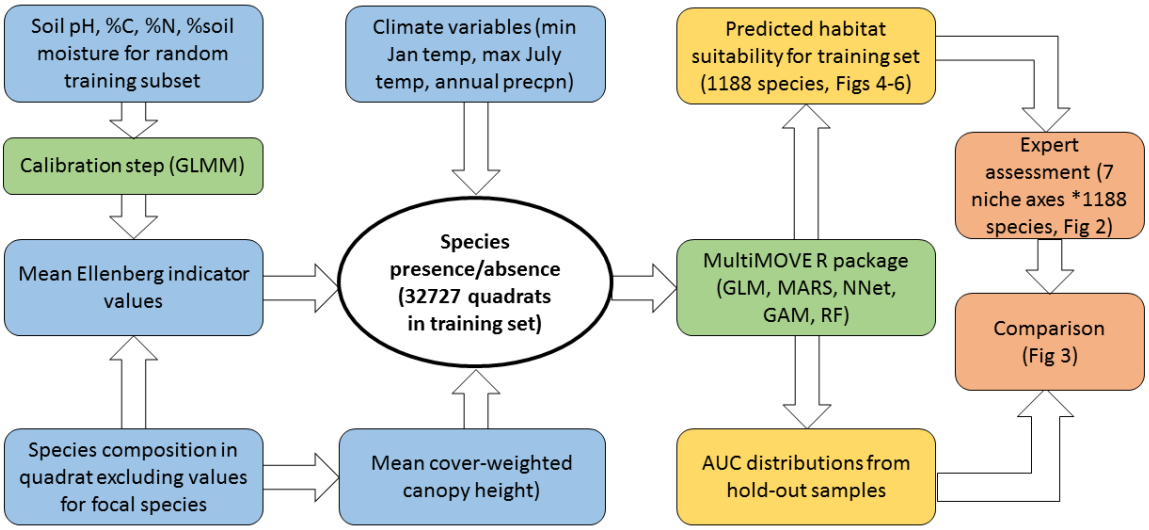
Figure legends:

Fig 1. Steps involved in building and assessment of the MultiMOVE species niche models based on expert judgement and comparison with AUC. Colour codes are as follows: Blue = model inputs. Green = quantitative modelling steps. Orange = Model outputs. Light red = model assessment steps. See Henrys et al (2015) and Smart et al (2010a) for detailed accounts of the construction of the species niche models including descriptions of the input data.

Fig 2. Results from assessments of the MulitMOVE models by two independent experts: A. both experts combined, B. Expert 1, vascular plants only, C. Expert 2, vascular plants only, D. Expert 1, bryophytes only.

Fig 3. Comparison of expert assessments – A. expert 1, B. expert 2 - for each species-niche axis combination versus AUC statistics for the associated model and the prevalence of each species in the training data used to build each model. Loess smoothers are fitted to each species*niche axis combination grouped by the assessment category awarded by the expert. Thus each point is a species * niche axis combination whose position is defined by its prevalence on the X axis and the mean AUC for the species model on the Y axis. Note that prevalence (the proportion of presences / total number of quadrats) was square-root transformed to spread the data more evenly across the X axis.

706 Fig 4. Modelled response of *Coeloglossum viride* to an indirect indicator of substrate pH.
707 The modelled response was assessed by both experts as moderate (expert 1) and poor
708 (expert 2). Their assessment would have been based solely on inspection of the unweighted
709 model average (brown line). Raw probabilities have been rescaled to between 0 and 1. Grey
710 ribbons indicate the 95% confidence region for the relevant modelled response.
711
712 Fig 5. Modelled response of *Coeloglossum viride* to vegetation height (1, <10cm, 8 >=15m),
713 (assessed as poor by both experts) and an indirect indicator of substrate pH (assessed as
714 moderate and poor by the two experts). Colours indicate the weighted average model
715 prediction for all training plots in the MultiMOVE database. The red line encloses all
716 observed occurrences of the species (black dots) in the training data. The grey polygon
717 encloses the ecological space defined by the training data; A. model predictions based on
718 observed values of background explanatory variables in each training plot, B. background
719 explanatory variables set to their median values in the training data.
720 Fig 6. Modelled response of *Schoenus nigricans* to precipitation (assessed as good) and an
721 indirect indicator of substrate pH (assessed as moderate). Colours indicate the weighted
722 average model prediction for all training plots in the MultiMOVE database. The red line
723 encloses all observed occurrences of the species (black dots) in the training data. The grey
724 polygon encloses the ecological space defined by the training data; A. predictions based on
725 observed values of background explanatory variables in each training plot, B. background
726 explanatory variables set to their median values in the training data.
727
728



729

730 FIG 1.

731

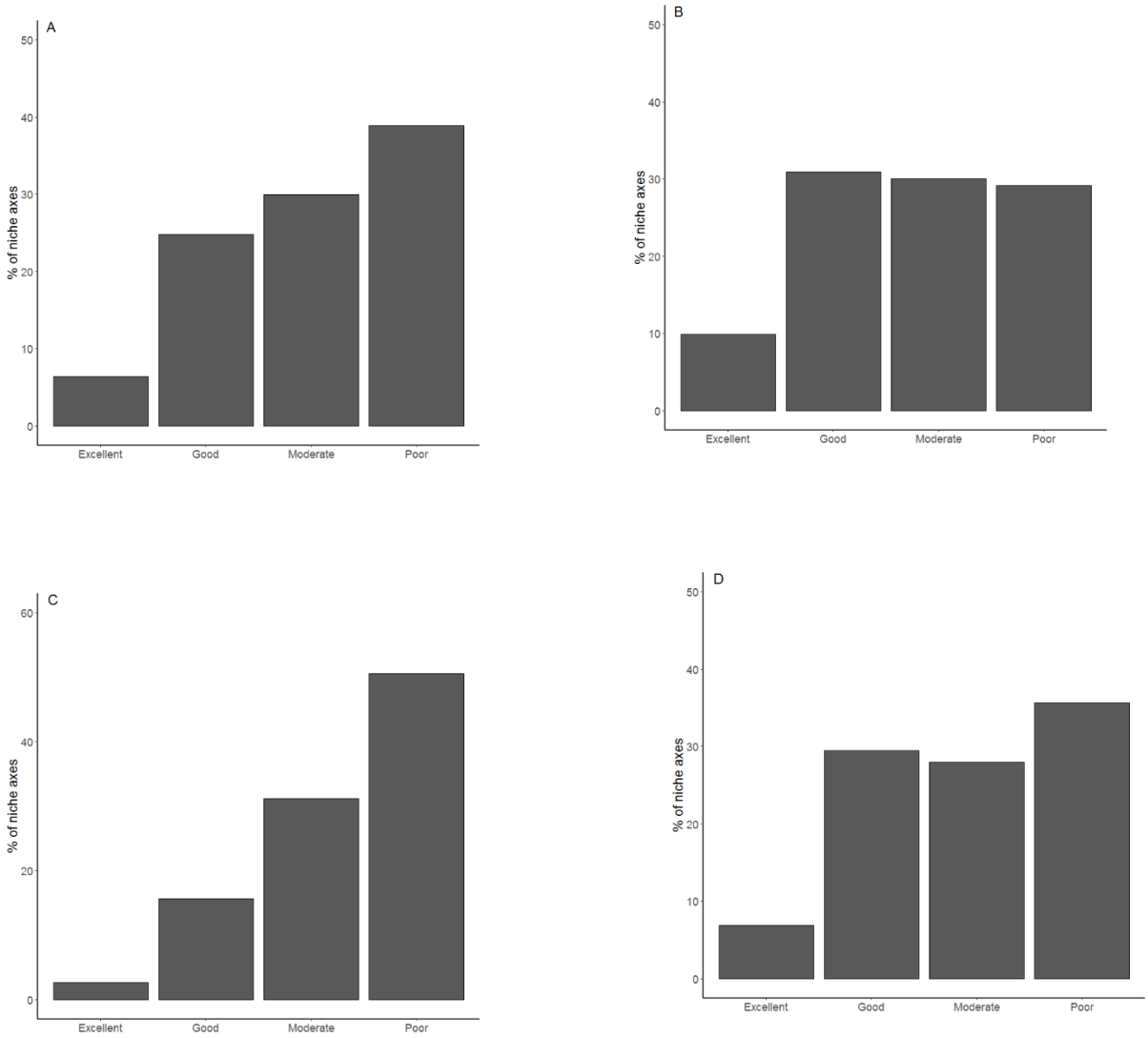
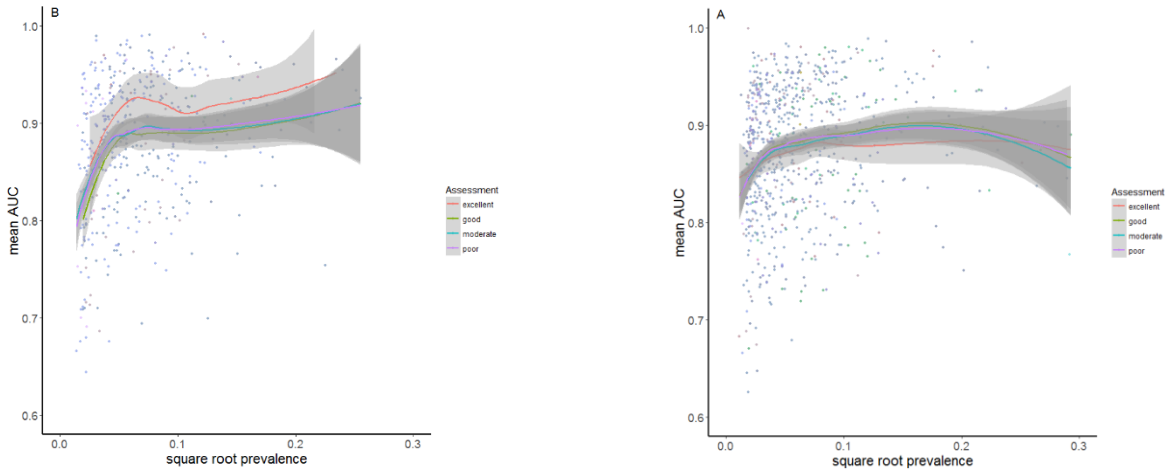

A
% of niche axes
0
10
20
30
40
50
Excellent
Good
Moderate
Poor
B
% of niche axes
0
10
20
30
40
50
Excellent
Good
Moderate
Poor
C
% of niche axes
0
20
40
60
Excellent
Good
Moderate
Poor
D
% of niche axes
0
10
20
30
40
50
Excellent
Good
Moderate
Poor

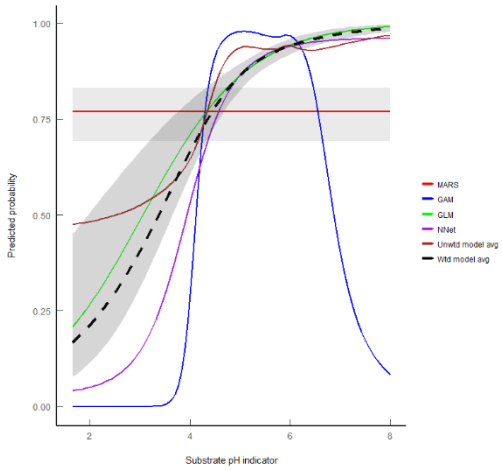
FIG 2 A-D.



746

747 FIG 3 A-B

748



749

750 FIG 4

751

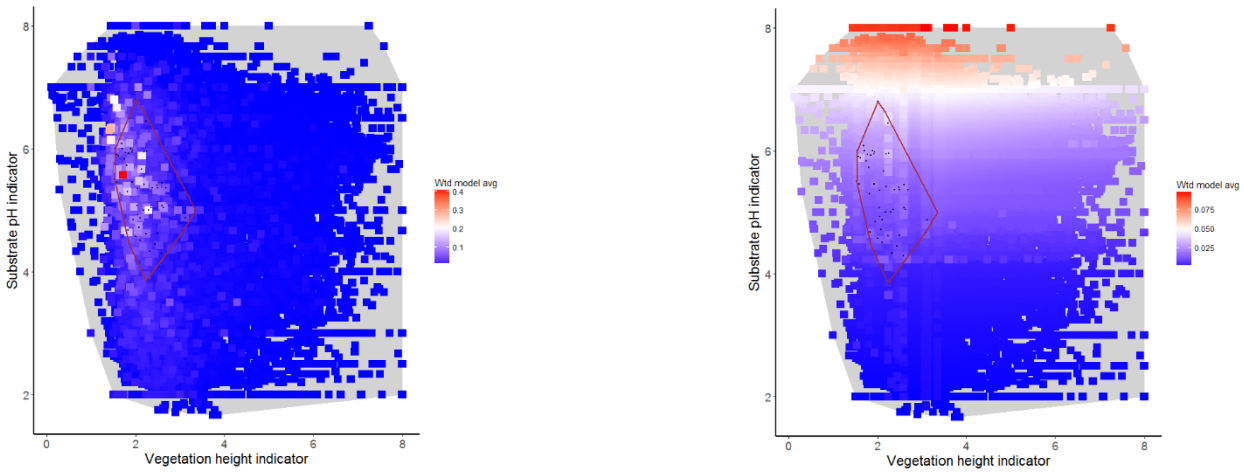

Substrate pH indicator
Vegetation height indicator
Wtd model avg
0.4
0.3
0.2
0.1
0.075
0.050
0.025

FIG 5

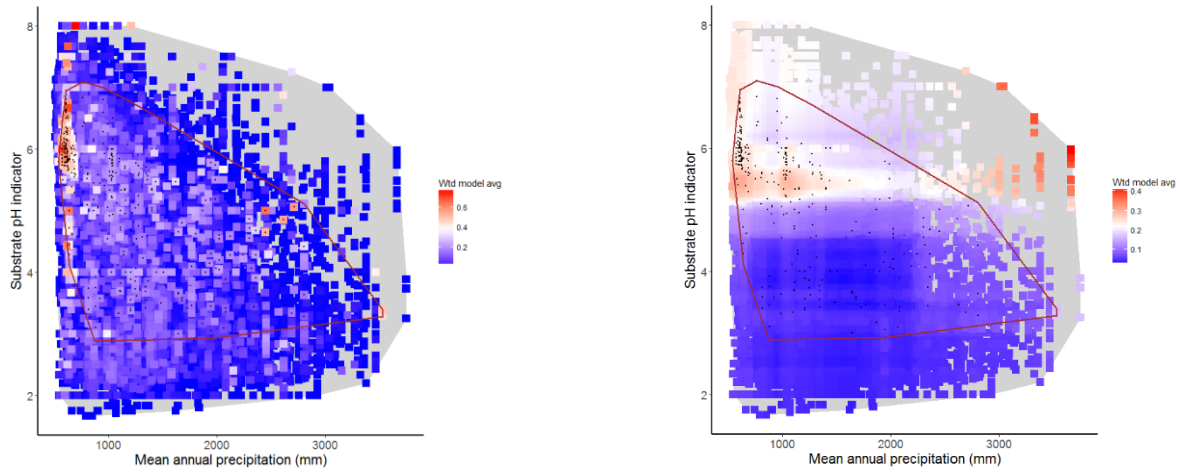

Substrate pH indicator
Mean annual precipitation (mm)
Wtd model avg
0.6
0.4
0.2

FIG 6