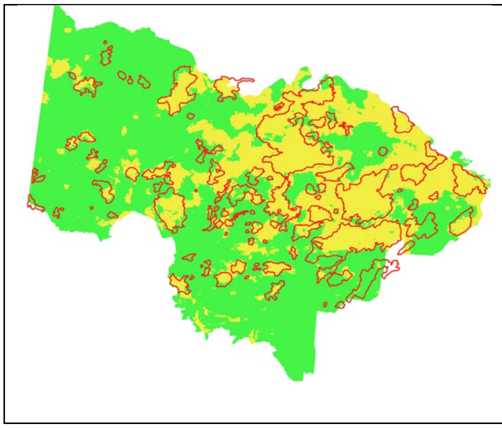
APPENDIX P

TERRESTRIAL RESOURCES TECHNICAL SUPPORT DOCUMENTS

Report on GHD Category 1 Predictive Mapping for Churchill, Berens, and Kinloch Ranges

In Support of Springpole Project Assessment (June 4th, 2024)



Rob Rempel, FERIT Environmental Consulting

https://www.ferit.ca

northernbio@gmail.com

Background: Understanding the impact of development on recruitment and population viability is related to the effect of development activities on calving and nursery areas. Although general range condition is mapped using predictive models of GHD categories 2 and 3 using the GHD Mapping tool, which include summer and spring probability of use, the effect of landscape configuration and development on Category 1 habitat is not modelled. Category 1 habitat is generally mapped using direct observations of cows and calves through aerial surveys in spring and summer, and then delineating areas using hand-drawn maps. This approach provides confirmed evidence of areas used by caribou for reproduction, but is not able to either identify areas that could serve as caribou calving and nursery areas, but currently are not used, or to predict the change in availability of calving areas as the landscape changes over time, either through the addition or removal of anthropogenic disturbances, or through the aging and succession of disturbed forest over time. A recent study by Walker et al. found that females displayed strong site fidelity in 15% of 52 females tracked, but 35% showed habitat fidelity, moving to new locations with similar habitat characteristics. The remainder showed a combination of site and habitat fidelity over the years monitored. The authors concluded that management for calving areas include a combination of site fidelity approaches, where mapped sites known to be used for calving be protected, but also habitat models should be developed to predict potential new locations for calving. For environmental impact assessments, it would be helpful to develop and apply a predictive model to assist evaluation of changes to calving and nursery areas over time.

One approach to developing such a model is to relate existing (unpublished), confirmed MECP delineations of calving and nursery (Category 1) areas to underlying environmental conditions, including GHD models of summer and spring probability of use. An overlay of Category 1 habitat on the existing map of GHD probability of high use in summer (Fig. 1) shows high spatial correlation between probability of high use (yellow) and mapped Category 1 areas (red outlines). This suggests that the variables used to map summer probability of use are likely to be also useful for predicting Category 1 habitat. Based on this initial assessment, we decided to move forward with developing a predictive model for calving and nursery areas that relates directly to the criteria MECP uses for identifying Category 1 habitat. The model would then be applied to proposed future conditions at project start and through time to estimate the relative change in availability of Category 1 habitat over time.

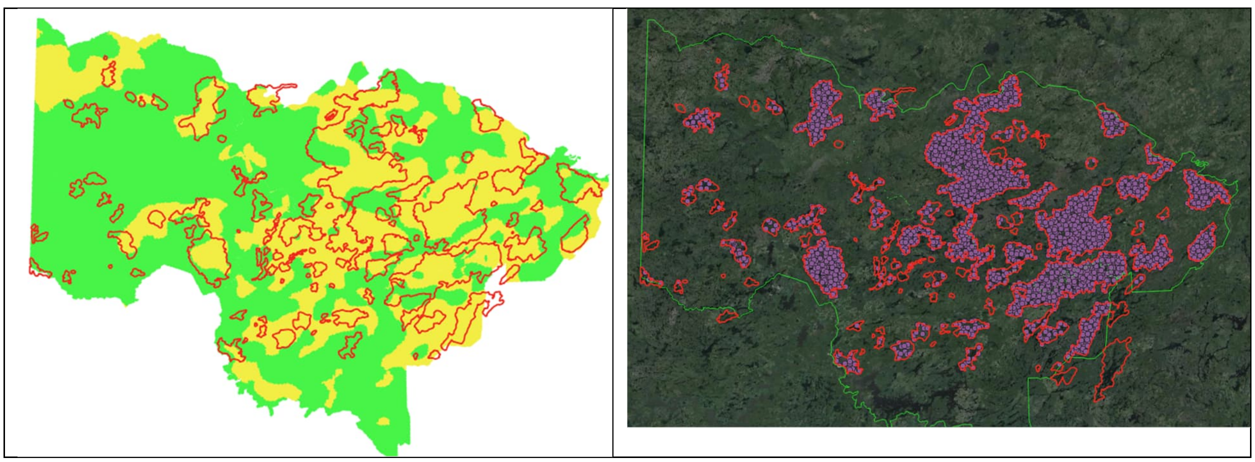


Figure 1. A: GHD Probability of High Use in Summer with overlay of MECP Category 1 (calving and nursery area) polygons; B. Overlay of sampled points within MECP Category 1.

Methods:

General Methods and Sampling:

The general approach to developing the predictive model for Category 1 habitat is to use a method based on a modelling technique called Resource Selection Function (RSF). This is related to binary logistic regression, but where data is assigned as either used (1) or available (0). In the context of this project, the "used" resource is defined by areas within the MECP Category 1 polygons, while the "available" resource is defined by a random set of points spread across the entire landscape. If conditions within the Category 1 polygons occur at a higher rate than average random conditions across the landscape, then it is inferred that these conditions are selected by caribou. This is somewhat different than a point-based GPS collar approach but is appropriate here because the nursery areas are based on a rather broad set of biophysical conditions, including site-level protective cover, predator detection capability, and predator escape opportunities (e.g., islands on lakes).

Sample data:

Points were created using the R package spsurvey (Dumelle et. al. 2023) to generate a generalized random tessellation stratified (GRTS) sample. GRTS is used to generate spatially dispersed random samples based on defined stratifications or proportions, minimum spatial separation, and other criteria. GRTS was used to generate 4021 sample points, with minimum spatial separation of 250 m. Sampling was restricted to polygons >= 1 ha, and the number of points per polygon set as a function of polygon size, with the restriction of no polygon having more than 100 points. This was to reduce undue leverage of large polygons on model estimates. The 4021 points were imported into the GIS, with points joined to the LSL shapefile that was produced by the GHD Mapping program. In QGIS 25000 random "available" points were created across the study area and joined to the LSL file. These files were then imported into R, where the used and available data were split into train and test data sets, with train data comprising 60% of the total sample. Splitting of "used" data was conducted to maintain proportional representation among size classes. Used and available data set splits were then joined to create the final train and test datasets.

Biophysical data:

The independent data used in the model included the same multi-scale data used in Category 2 and 3 habitat predictive maps (Table 1), but also included some additional variables, including distance to road and proportion of harvest or anthropogenic disturbances with a hexagon. Details on these variables can be found in the Category 2 and 3 predictive mapping report (ref xx). The 5,000 ha scale was used for the Category 2 RSF maps, so this scale was also used in developing this Category 1 predictive model. In addition to the biophysical variables, the predicted probability of spring use and summer use (representing high probability for calving and nursing areas), and a new calving RSF raster map based on GPS collar data (RSF 2024) was included in the candidate variable list.

Table 1. Candidate variables used in model development. For each of these variables, versions spatially averaged from the initial scale of 3.1 ha to the averaged scale of 5137 (S5) ha were included as candidates in the BRT Model.

Variable	Source	Calculation
Dense Deciduous	DEC_S5	Proportion within hexagon
Dense Mixed Forest	MIX_S5	Proportion within hexagon
Dense Conifer	CON_S5	Proportion within hexagon
Sparse Treed	ST_S5	Proportion within hexagon
Treed Peatland	LGTP_S5	Proportion within hexagon
Open Peatland	LGOP_S5	Proportion within hexagon
Water	LGW_S5	Proportion within hexagon
Natural Disturbance	DTN_S5	Proportion within hexagon

Eskers lines	ESL_S5	Density (m/ha)
Linear features – roads, railways, and transmission lines	TDENLF_S5	Density (m/ha)
Distance to nearest road	R_DIST	Distance (m)
Harvest disturbance	HARV	Proportion within hexagon
Anthropogenic disturbance	ANTHRO	Proportion within hexagon
Probability of Summer Use (high use category)	PSU_USE	Proportion of high use category within hexagon
Probability of Spring Use (high use category)	PSP_USE	Proportion of high use category within hexagon

Boosted Regression Tree:
The number of candidate variables was large, and interactions potentially complex, so a machine learning language approach was used to generate the predictive categories 1 classification map. Boosted regression tree (BRT) has proven to be one of the most powerful models for generating classifications and was selected as the modelling framework for the model. The R gbm package provides powerful tools for fine-tuning hyperparameters and to avoid overfitting the model.

Model Performance (Sensitivity, Specificity, and ROC):
The area under the curve (AUC) of the resource operating characteristic (ROC) plot reveals how well the model performs in terms of accurately predicting the location of Category 1 habitat in terms of both false positive and false negatives. In the context of testing a wildlife habitat model, sensitivity, specificity, and the ROC curve are essential metrics for evaluating how well the model predicts suitable versus unsuitable habitats. These concepts are applied to assess the model's accuracy in classifying a particular area as being a habitat or not, for binary classification tasks in statistical models and machine learning.

Sensitivity (True Positive Rate)

- Definition: Sensitivity is the proportion of actual positives (true habitats) correctly identified as such by the model. It indicates the model's ability to detect suitable habitats.
- Importance in Habitat Modeling: High sensitivity means the model is effective at identifying areas that are truly suitable habitats for wildlife species. This is crucial for conservation efforts, ensuring that critical habitats are not overlooked.

Specificity (True Negative Rate)

- Definition: Specificity measures the proportion of actual negatives (non-habitat areas) that are correctly identified by the model. It reflects the model's ability to recognize areas that are unsuitable for the species.
- Importance in Habitat Modeling: High specificity means the model accurately excludes areas that are not suitable as habitats. This helps in minimizing the allocation of conservation resources to non-essential areas and focusing on truly significant habitats.

ROC Curve (Receiver Operating Characteristic Curve)

- Definition: The ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings.
- Importance in Habitat Modeling:
  Threshold Evaluation: The ROC curve helps in choosing the optimal threshold for classifying an area as habitat or non-habitat, balancing the trade-offs between sensitivity and specificity.
- Model Comparison: It provides a tool to compare the performance of different habitat models. The area under the ROC curve (AUC) gives a single measure of overall accuracy that is not dependent on a specific classification threshold. A higher AUC indicates a model with better overall performance.

Results:

Cross-validation and over-fitting:

The BRT model will continue to improve the accuracy of its predictions as more variables are added into the model, as revealed by the black line in Fig. 2A, where the decrease in Bernoulli deviance indicates higher model performance. However, too many variables in the model can lead to over-fitting, where the model no longer performs well with a new, independent data set. To guard against this, cross-validation is used develop the model with primary data, and then test the model against a sub-set of data iteratively set aside from model development. I used 8-fold cross-validation, where 1/8th of the data was randomly selected and set aside, with the latest iteration of the model tested against the set-aside data. The graph reveals at what point higher model complexity no longer improves the model when tested against the set-aside data (green line).

An important additional symptom of over-fitting is when the model performs very well on the training data, but poorly on the test data. This can be revealed by comparing the sensitivity, specificity, and AUC statistics for models developed on the train and test data sets. The initial model performed very well on the train data set (AUC ~ 0.85), but less so on the test data (AUC ~ 0.59). Also, specificity was low for the test data model. Consequently, I tried to improve the model performance on the test data by adjusting model hyperparameters to reduce over fitting. I reduced model complexity to 1 level of interaction, increased model learning rate, increased the minimum number of observations for the final node, and reduced maximum number of trees. This resulted in a decrease in performance of the train model (AUC ~ 0.8), but increased performance of the test model. The specificity and sensitivity of the test model (i.e., false positive and false negative errors) were more balanced in the test model.

Initially 250 trees were specified as maximum model complexity, and at this level the cross-validation began to stabilize at 200 trees. To reduce over-fitting hyperparameters were adjusted. The maximum number of trees was reduced from 250 to 100, interaction depth reduced from 2 to 1, shrinkage rate increased from 0.1 to 0.7, and subsample (bag fraction) increased from 0.5 to 0.7. With these parameters, the solved model was simpler and stabilized more quickly at 88 trees versus 250 (Fig. 2B).

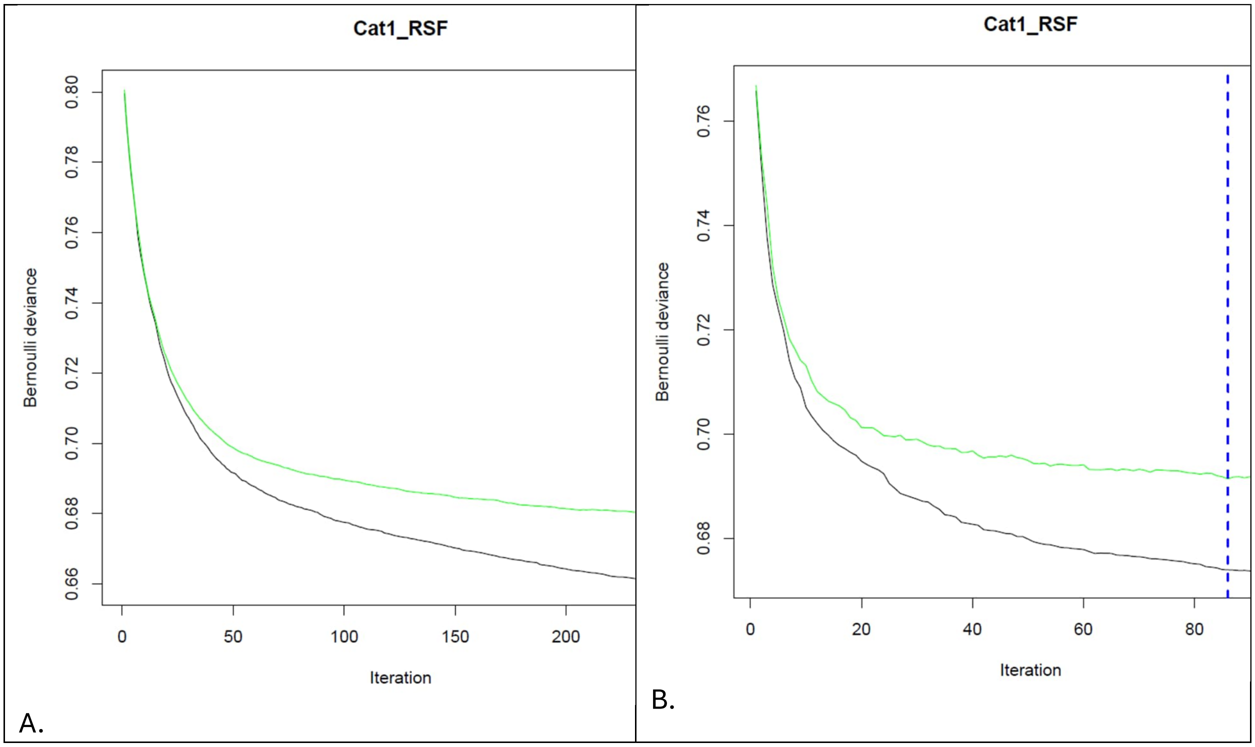


Figure 2. Model cross-validation for the initial model (A), and the revised model (B), with revised hyper-parameters to reduce risk of over-fitting.

Variable Influence:

The gbm program calculates variable influence to assess relative importance of each variable in the model. The most important variables included open water, probability of spring use, treed peatland, eskers, deciduous, conifer, and open peatland (Fig. 3). The revised model performed slightly better, and the order of relative variable importance changed (Fig. 3B). Prevalence of mature conifer (CON_S5), natural disturbance (DTN_S5), and linear feature density (TDENLF_S5) will all change as a result of forest aging and resource development activities, and thus the model variables will be useful for assessing simulated changes over time.

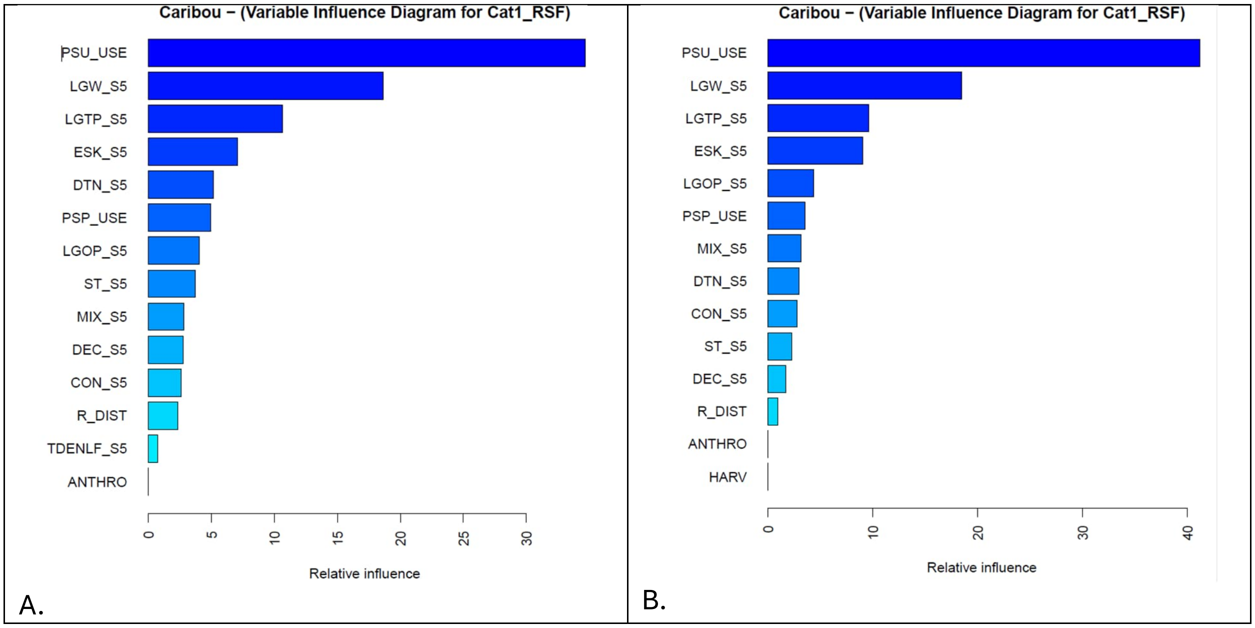


Figure 3. A: Model variable importance for the initial model; B. variable importance for the revised model.

Variable Response:

Although the variable influence diagram informs what variables are important in the model, it does not provide information on the nature of the response to the variable. To assist with this, variable response diagrams were created to inform the direction and magnitude of the response (Fig. 4). Open water (LGW_S5), treed peatland (LGTP_S5), probability of summer use and spring use (PSU_USE, PSP_USE), and conifer (ESK_S5) were all positive predictors of MECP Category 1 habitat. Open peatland (LGOP_S5) was positive up until the hexagon was about 15% open peatland, and thereafter had a negative association with Category 1. Natural disturbance (DTN_S5) had an immediate negative effect on Category 1 habitat.

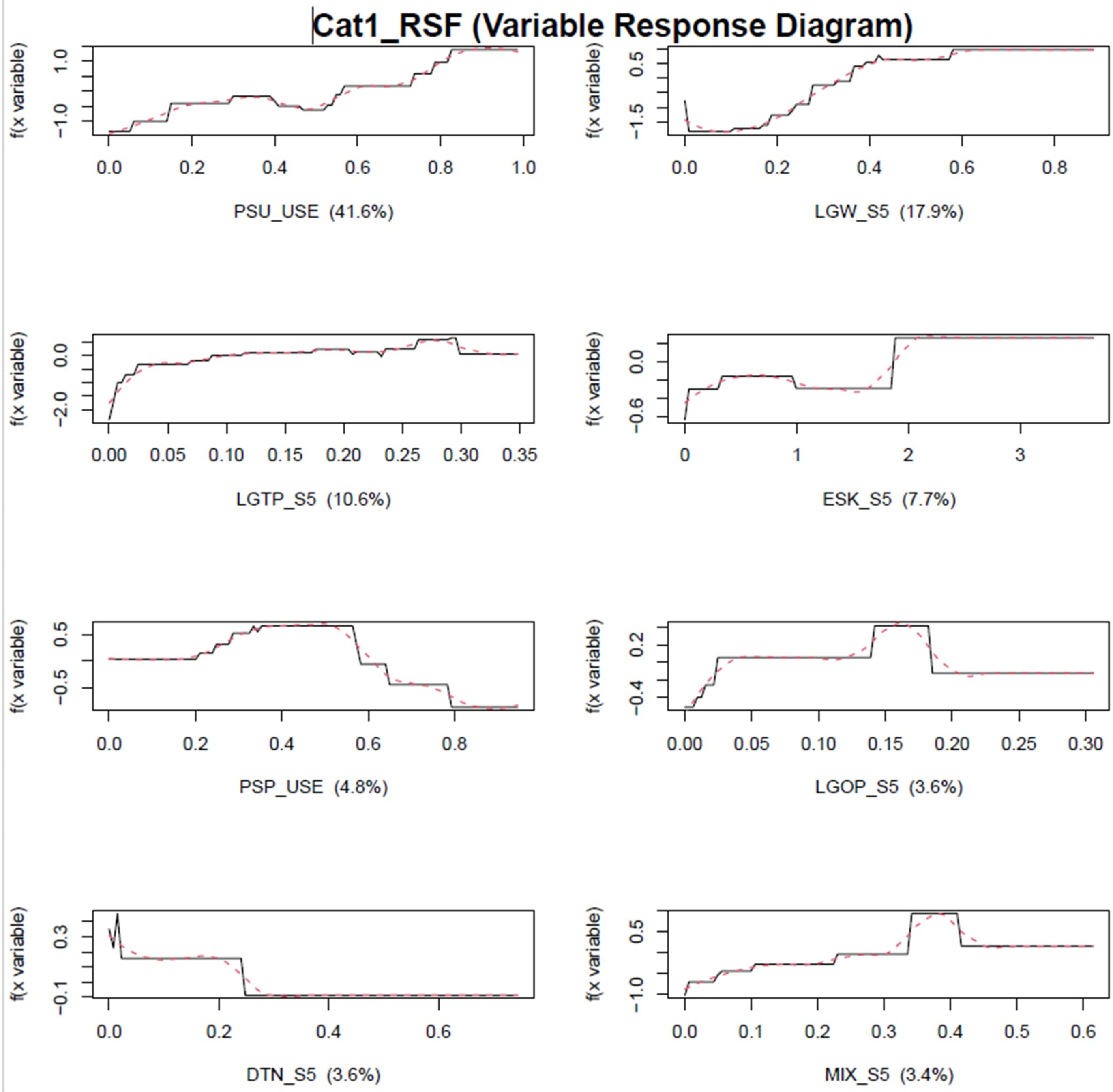


Figure 4. Variable response diagram.

Model Performance:

After adjusting the model for overfitting, the model performed very well for the train data (AUC = 0.792), and reasonably well for the test data (AUC = 0.786), although these values were slightly lower than the initial model (0.807 and 0.795 for train and test data, respectively) (Fig. 5). This was expected as the model was adjusted to reduce model over-fitting. Sensitivity and specificity were reasonably balanced for the test data (79.6 and 65.9, respectively).

Relative to the initial model, sensitivity decreased slightly, while specificity increased. This would result in the revised model predicting slightly less area as Category 1.

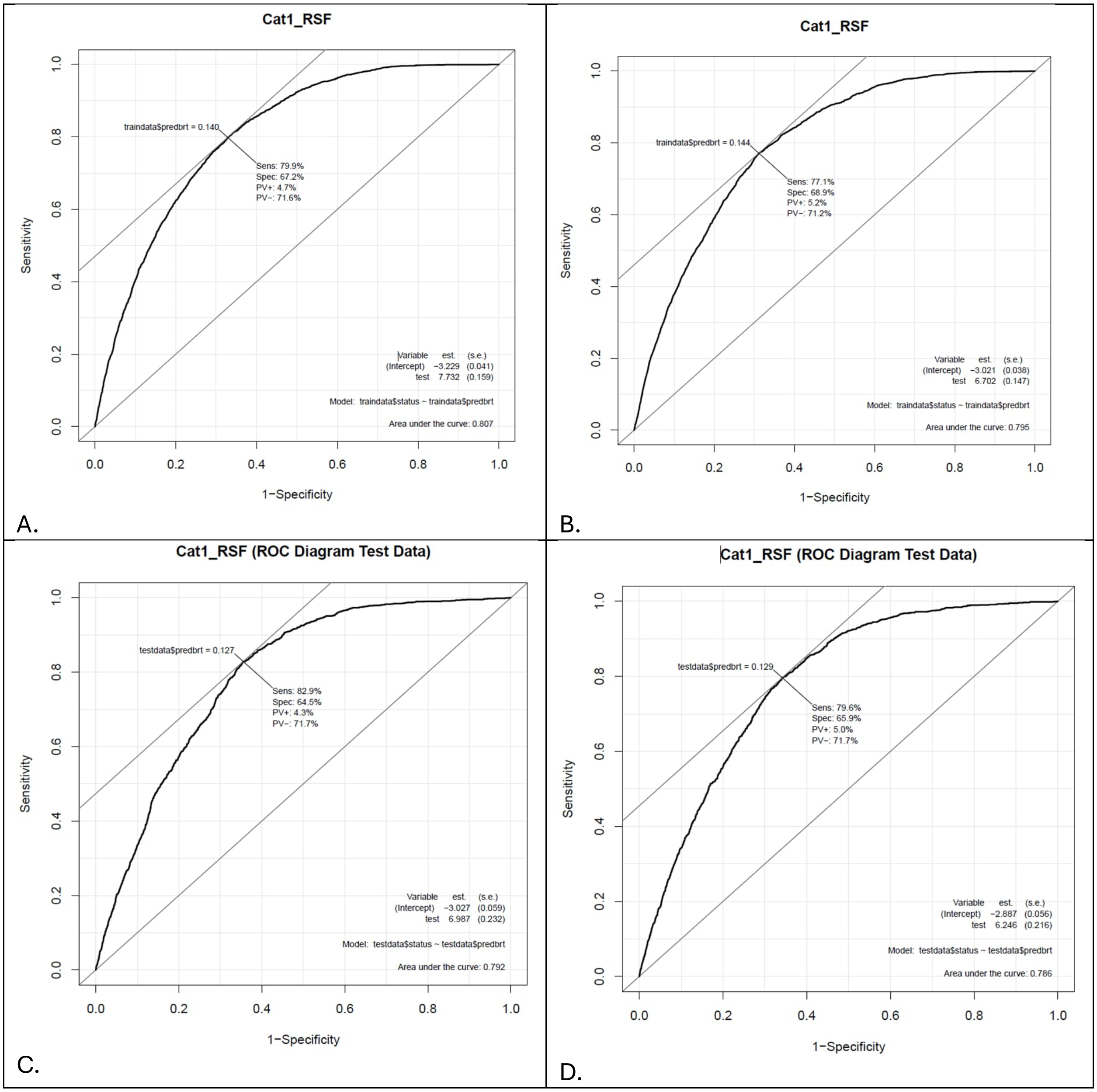


Figure 5. ROC plots for train and test models. A and C are for the initial model, B and D are for the revised model, with revised hyper-parameters to reduce risk of over-fitting.

Model Predictions:

The multi-variate RSF model made a marked improvement in the predictions of Category 1 habitat relative to using just a single variable, probability of summer use (high use category) (Fig. 6), with the new model increasing both sensitivity and specificity for Category 1 predictions. The spatial “fit” of the Category 1 RSF (yellow area in Fig. 6 B) is now more tightly aligned with MECP Category 1 map (red outlines) than if we had predicted Category 1 based on the summer Category 2 values alone (Fig. 6A). This strong correlation based on the predictive model now allows us to predict future changes and availability of Category 1 habitat based on simulated changes in anthropogenic and natural disturbance.

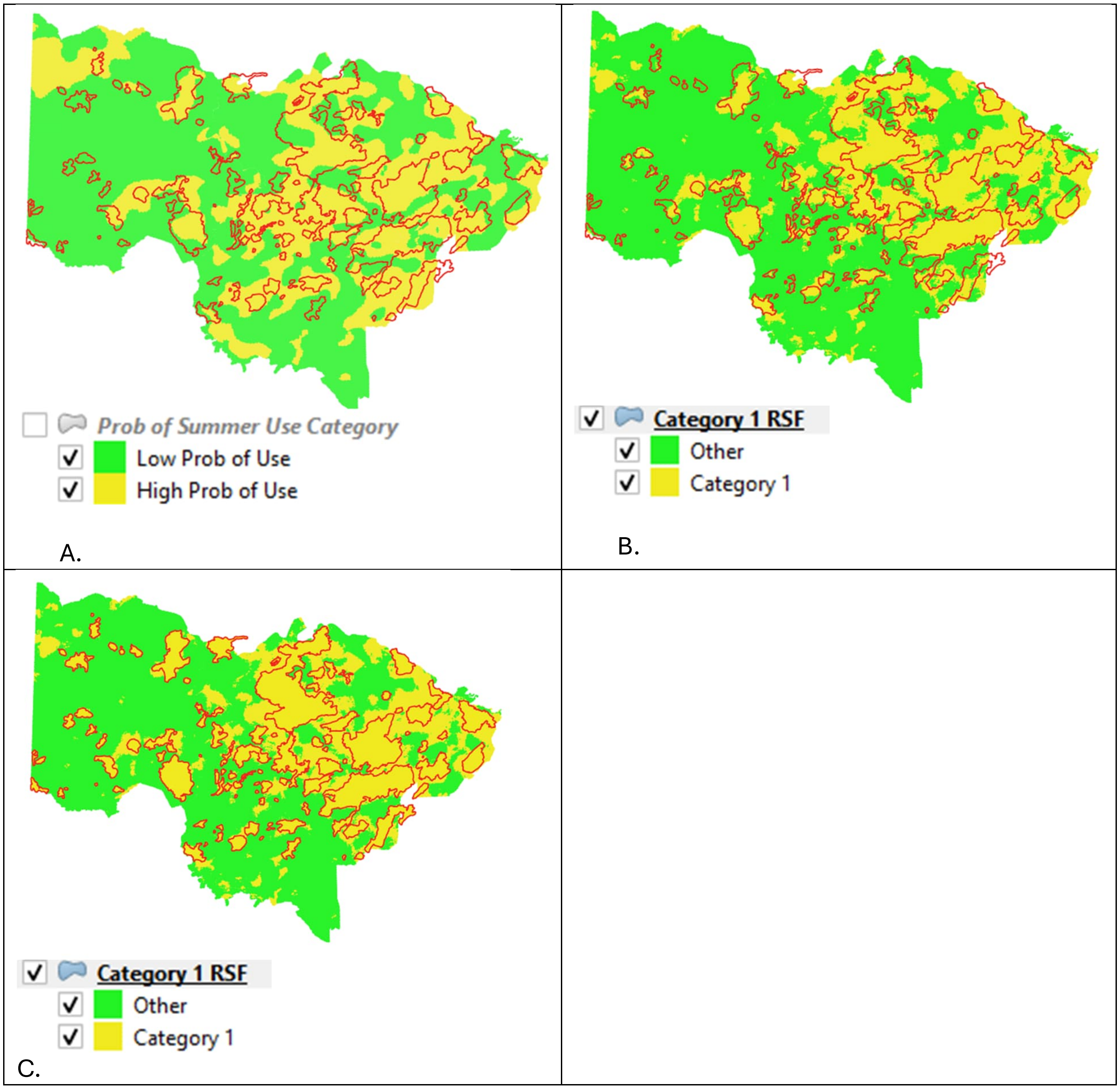


Figure 6. A: Summer High-Use Category, based on the Category 2 and 3 mapping models. Yellow is high, green is low. Red outlines are MECP mapped Category 1 areas; B. Model predictions (Time 0), where yellow is predicted Category 1; C. Combined predicted Category 1 and MECP mapped Category 1.

References:

Michael Dumelle, Tom Kincaid, Anthony R. Olsen, Marc Weber (2023). spsurvey: Spatial Sampling Design and Analysis in R. Journal of Statistical Software, 105(3), 1-29. doi:10.18637/jss.v105.i03