The following supplements accompany the article

Environmental correlates of nearshore habitat distribution by the critically endangered Māui dolphin

Solène Derville*, Rochelle Constantine, C. Scott Baker, Marc Oremus, Leigh G. Torres

*Corresponding author: solene.derville@ird.fr

Marine Ecology Progress Series 551: 261–275 (2016)

Supplement 1. Environmental variables

A. Turbidity index

To assess the potential for remotely sensed chlorophyll-*a* concentration to be used as an index of water turbidity we compared in situ Secchi disk measurements collected in 2015 along the central west coast North Island and MODIS chlorophyll-*a* concentrations for the same positions and dates. The log-transformed chlorophyll-*a* concentrations and Secchi disk measurements showed a linear relationship (Fig. S1). Yet, Secchi depths were distributed with a greater variance in waters with a log-transformed chlorophyll-*a* concentrations below -0.5. Therefore, we considered the linear relationship to be robust for waters with log-transformed chlorophyll-*a* concentration > -0.5 and limited predictions to these conditions. As a result, predictions were only applied to the relatively turbid waters found in nearshore waters within a few kilometres from the coast.

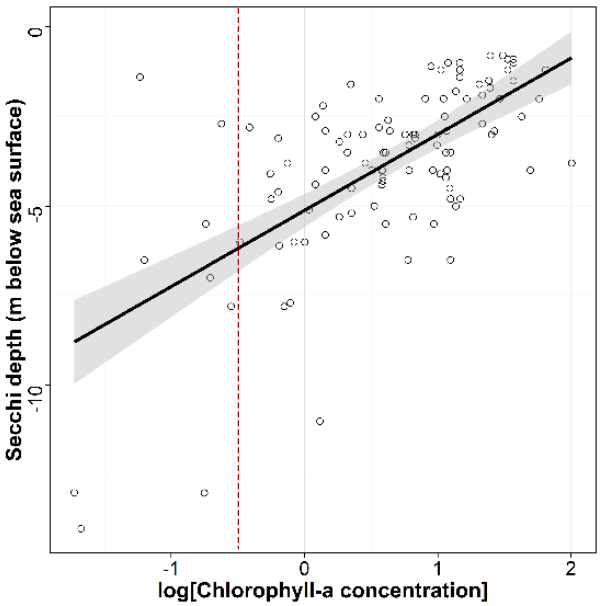

Fig. S1. Secchi disk measurements collected in 2015 with relation to log-transformed remotely sensed chlorophyll-a (NASA, Aqua Modis 8-days composites). Adjusted R^2 = 0.43, df = 107, regression coefficient = 2.12 ± SE 0.24; Pearson's coefficient = 0.65, paired sample two-sided t-test: t = -8.906, df = 107, p = $1.52e^{-14}$. Grey areas indicate 95% confidence intervals.

B. Watersheds

Watersheds were classified into major watersheds (major river mouths + harbour mouths) and minor watersheds (minor river mouths; Fig. S2). Maps of minor rivers were obtained from NZ River Centrelines Topo50 Maps series at 1:50,000 scale, provided by Land Information New Zealand under a Creative Commons Attribution 3.0 New Zealand Licence (https://koordinates.com/layer/20-nz-river-centrelines-topo-150k/, last updated 08 Jan 2015, accessed 11 March 2015). Maps of Major rivers were provided by NZ Forest Service at 1:250,000 scale, under a Public Domain Licence (https://koordinates.com/layer/306-nz-major-rivers/ , last updated 14 October 2011, accessed 05 May 2015). Harbour mouths were geolocated manually using GoogleEarth (version 7.1.2.2041). All minor rivers emptying inside harbours were pulled out of the data. We checked that presence/absence positions could be linked to their closest minor/major watershed by straight segments without

crossing over land. Then we calculated Euclidean distances between points and closest watersheds after projecting in a UTM coordinate system. The same procedure was applied during the prediction stage to calculate distances between major watersheds/minor watersheds and the centre of all grid cells in the map of North Island coastal waters. In the Ahipara Bay, a local correction on distance was applied so that the watersheds located on the eastern side of the strip of land were not considered as the closest watersheds to grid cells located in Ahipara Bay.

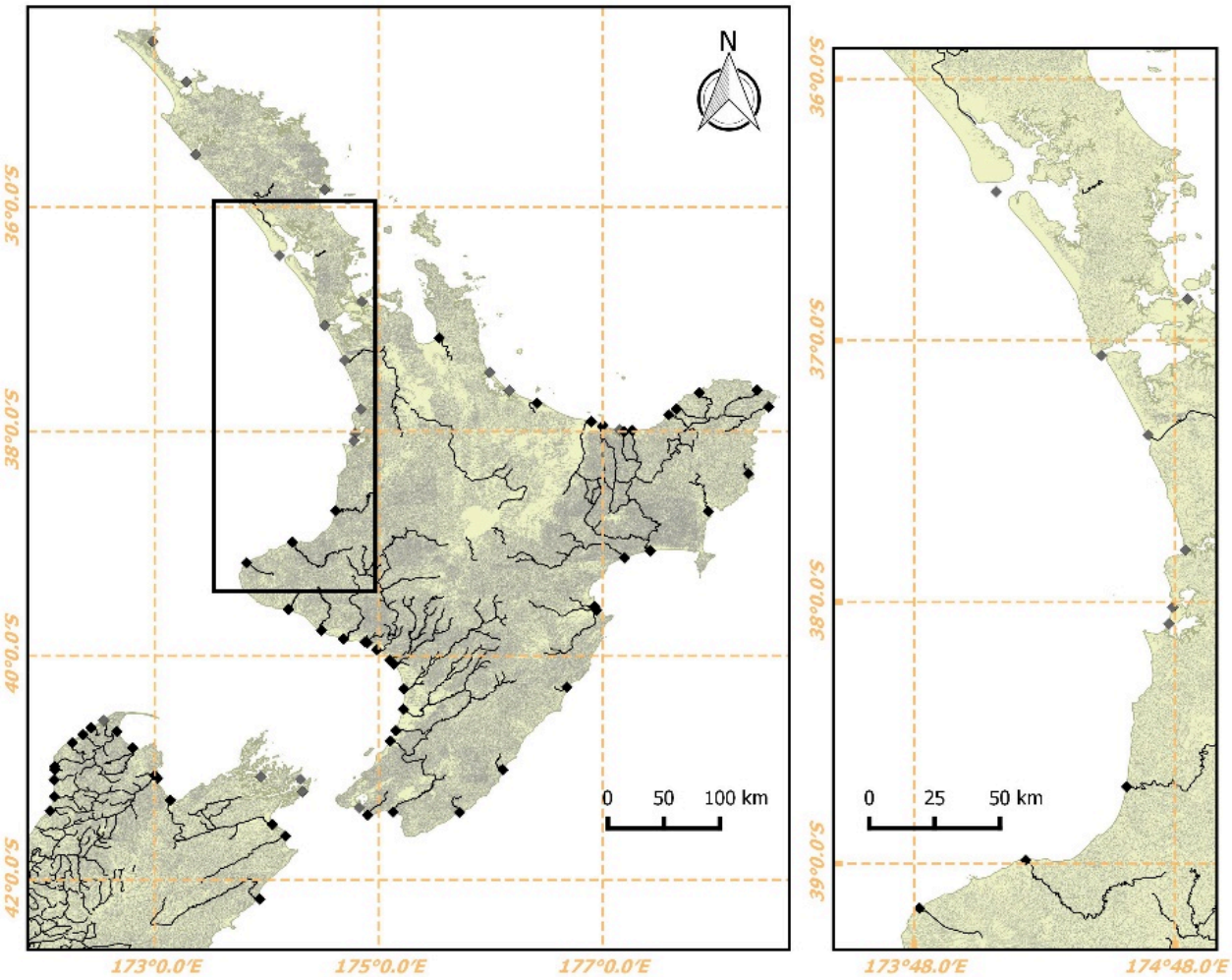

Fig. S2. Watersheds included in model calibration stage (right panel) and prediction stage (left panel). Grey diamonds indicate harbour entrances (i.e. major watersheds) and black diamonds indicate major river mouths (i.e. major watersheds). Major rivers are shown in black. Minor rivers (i.e. minor watersheds) are shown in grey but their mouth is not reported here.

C. Depth raster

Bathymetry was acquired from a raster with 250m resolution provided by NIWA (https://niwa.co.nz/). Due to the abrupt terrain in certain parts of the coastline, the depth raster was occasionally affected by slight imprecisions in the areas closest to the shore. As a result, high positive values of depth were assigned to a few presence/absence positions located in nearshore waters along west coast North Island. Corrected depth values were assigned to these positions after shifting them by one or two grid cells westward.

Supplement 2. Boosted Regression Trees

A. Collinearity between environmental variables

Correlation between environmental variables was assessed using the training dataset composed of 1626 positions (Fig. S3). We also investigated the eventual correlation between these environmental variables included in the BRT model and latitude. DIST_COAST was significantly correlated to DEPTH (Spearman's test: S = 1286, p-value < 2.2e-16) and DIST_MINWATERSHEDS (Spearman's test: S = 3407, p-value < 2.2e-16). Even if BRT modelling generally copes well with correlated or interacting explanatory variables, carefully selecting predictors is desirable to prevent overfitting when using small datasets. Based on this correlation matrix, we conducted a predictor selection step before releasing our final model. We compared the full model (6 predictors) to simplified models, excluding DIST_MINWATERSHEDS and either DIST_COAST or DEPTH.

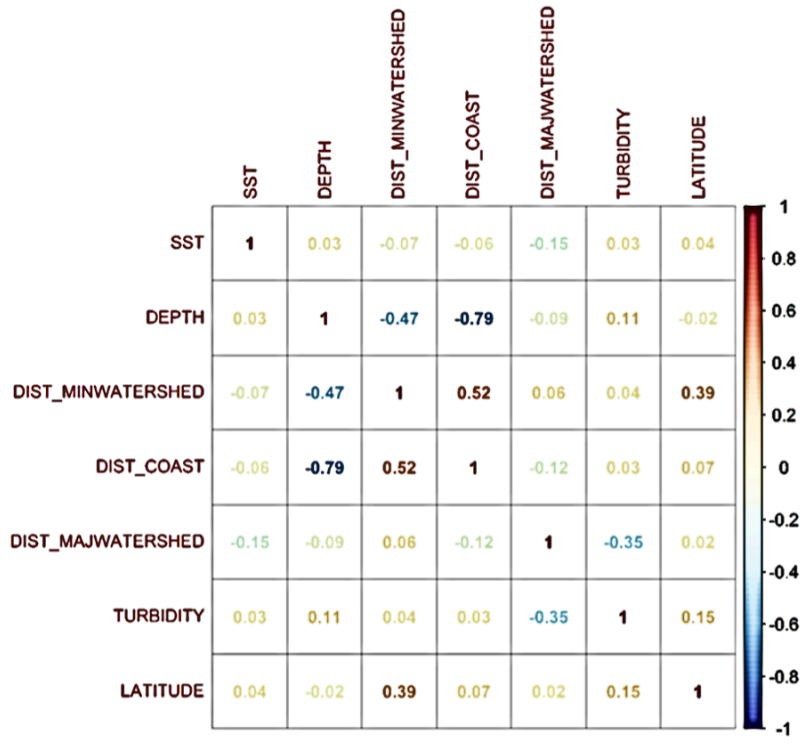

Fig. S3. Correlation matrix between all environmental variables included in the BRT model 1. Numbers correspond to Spearman correlation coefficient for each pair of environmental variables.

B. BRT model selection metrics

BRT model selection was performed using two different metrics: the area under the receiver operating curve (*AUC*) and the deviance explained by the model (*dev*).
Deviance explained corresponds to the percentage of deviance for the null model explained by the fitted model (*dev* = 100% for a perfect model).

$$\boldsymbol{Deviance\ explained} = \frac{\boldsymbol{Null\ deviance - Residual\ deviance}}{\boldsymbol{Null\ deviance}} * \mathbf{100}$$

The function *calc.deviance()* in dismo package calculates the deviance between 2 vectors v_1 and v_2 and uses a different method depending on the type of data (Bernoulli, Poisson, Laplace or Gaussian). For vectors of type "Bernoulli":

$$\boldsymbol{deviance} = -\mathbf{2} * \sum_{\boldsymbol{i=0}}^{\boldsymbol{N}} \boldsymbol{v_{1i}} * \mathbf{log}\,(\boldsymbol{v_{2i}}) + (\mathbf{1} - \boldsymbol{v_{1i}}) * \mathbf{log}\,(\mathbf{1} - \boldsymbol{v_{2i}})$$

With *N* the sample size. Using this function, residual deviance was calculated between observed presence/absence data (v_1) and predicted probability of presence (v_2), and null deviance was calculated between observed presence/absence data (v_1) and mean observed probability of presence (v_2). When calculating deviance explained by the model over the training dataset (int.dev), null and residual deviances were directly provided by the *gbm.step()* function in dismo package. In contrast, when calculating deviance explained in the evaluation

dataset (portion of data withheld during model building; ext.dev), null and residual deviances must be manually estimated using *predict.gbm()* and *calc.deviance()* functions (see code beneath).
AUC (area under the curve) of the ROC (receiver operating characteristic) evaluates the capacity of a model to predict the right probability of presence over locations where observations were present or absent. The ROC plots sensitivity ([true positive rate]: portion of presence points classified as presence) versus 1- specificity ([1-false positive rate]: portion of absence points classified as absences) at various increasing thresholds above which prediction is considered positive. AUC ranges between 0 and 1; an AUC of 0.5 indicates that the model predicts occurrence no better than random while an AUC of 1 implies a perfect prediction. In *gbm.step()* function, AUC is automatically calculated during Cross-Validation and averaged over all folds (cv.AUC). Here, AUC was also calculated manually over the evaluation data set (ext.AUC) using the *roc()* function from pROC package (see code beneath).

```
#production of a BRT model based on a training dataset "training.df".
model = my.gbm.step(training.df,
                    gbm.x = …, # columns of training.df containing predictors
                    gbm.y =…, # column of training.df containing presence/absence binomial data
                    tree.complexity = 4,
                    learning.rate = 5e-4,
                    family = "bernoulli",
                    n.trees = 25,
                    bag.fraction=0.1,
                    plot.main = TRUE,
                    fold.vector = training.df$fold,
                    n.folds = 4,
                    keep.fold.fit = T,
                    keep.fold.vector = T)
#Model internal metrics
cv.AUC = model$cv.statistics$discrimination.mean
cv.dev = model$cv.statistics$deviance.mean # average residual deviance over all folds during cross-validation
int.null.deviance = model$self.statistics$mean.null
int.residual.deviance = model$cv.statistics$deviance.mean
int.dev = (int.null.deviance – int.residual.deviance)/int.null.deviance
#Model external metrics , calculated over an evaluation dataset "evaluation.df"
pred = predict.gbm(model, evaluation.df, n.trees = model$gbm.call$best.trees, type = "response")
ext.AUC = roc(evaluation.df$presence, pred)$auc
ext.residual.deviance = calc.deviance(evaluation.df$presence, pred, calc.mean=T)
ext.null.deviance =
calc.deviance(evaluation.df$presence,rep(mean(evaluation.df$presence),nrow(evaluation.df)), calc.mean=T)
ext.dev=(ext.null.deviance - ext.residual.deviance)/ext.null.deviance
```

C. BRT parameters calibration

Boosted regression trees were fitted using the gbm and dismo libraries in R. In order to estimate the optimal number of trees (*nt*) using Cross-Validation, this method requires the specification of 4 main parameters: bag fraction (*bf*), learning rate (*lr*) and tree complexity (*tc*).
During model calibration step, we tested different combinations of *bf* (0.1, 0.5, 0.75), *lr* (0.001, 0.0005, 0.0001, 0.00005, 0.00001) and *tc* (1, 2, 3, 4). The number of trees to add at each cycle of the boosting algorithm, referred to as "step size" may also be specified. We tested all combinations of step size (25 or 50) with the other parameters. Overall, these different combinations of parameters resulted in 120 models from which 12 converged (Table S1). The optimal parameters setting was selected based on explained deviance and AUC. Two models stood out: they shared the same *bf*, *lr* and *tc* (*tc* = 4, *lr* = 0.00005, *bf* = 0.1) but had different step size. By default, we selected the model with smallest step size (25; in bold in Table S1) as this setting allows a slower convergence of the boosting algorithm, thus more precision in the selection of the optimal number of trees.

Table S1. Summary of parameters and evaluation metrics for 12 BRT models. BRT boosting algorithm parameters: tc = tree complexity, lr = learning rate, bf = bag fraction, step.size = number of trees to add at each cycle during boosting, nt=number of trees allowing best predictive performance**.** Selected model is shown in bold.

tc	*lr*	*bf*	step.size	*nt*	cv.AUC	ext.AUC	int.dev	ext.dev
2	5,00E-04	0.1	50	1200	0.648	0.839	0.080	0.063
2	1,00E-04	0.1	50	2200	0.636	0.815	0.036	0.033
3	5,00E-04	0.1	25	1175	0.673	0.845	0.103	0.076
3	5,00E-04	0.1	50	1150	0.669	0.849	0.098	0.094
3	1,00E-04	0.1	25	1350	0.667	0.830	0.031	0.024
3	1,00E-04	0.1	50	3400	0.669	0.838	0.066	0.055
3	5,00E-05	0.1	50	1950	0.663	0.835	0.023	0.022
4	**5,00E-04**	**0.1**	**25**	**1025**	**0.693**	**0.853**	**0.104**	**0.087**
4	5,00E-04	0.1	50	1050	0.681	0.852	0.106	0.098
4	1,00E-04	0.1	25	1675	0.682	0.846	0.043	0.035
4	1,00E-04	0.1	50	3800	0.669	0.850	0.085	0.069
4	5,00E-05	0.1	50	3300	0.676	0.845	0.042	0.037

D. BRT parameters calibration

During the predictors selection step, we tested 5 different models (Table S2) using the best parameters setting (*tc* = 4, *lr* = 0.00005, *bf* = 0.1, step.size = 25). DIST_MINWATERSHED, DEPTH and DIST_COAST were successively excluded based on their relatively low contribution to model 1 containing all 6 predictors. Each model was run over 1000 bootstrap samples of the training dataset. Each bootstrap sample was selected randomly but with a constant presence to pseudo-absence ratio. Even if little variation was observed on average in the performance metrics calculated over the 5 models, model 1 maximized 3 metrics out of 4 (ext.AUC, int.dev, ext.dev) and was therefore conserved as our best BRT model.

Table S2. Summary of evaluation metrics for 5 BRT models built with different predictors. For each metric, we report the mean value and the coefficient of variation obtained from bootstrap re-sampling. Selected model is shown in bold.

model	Predictors		cv.AUC	ext.AUC	int.dev	ext.dev
1	SST + TURBIDITY + DIST_MAJWATERSHEDS + DEPTH + DIST_MINWATERSHEDS + DIST_COAST	mean	0.803	**0.870**	**0.313**	**0.203**
		coeff_var	1.872	0.930	7.684	6.441
2	SST + TURBIDITY + DIST_MAJWATERSHEDS + DEPTH+ DIST_COAST	mean	0.795	0.860	0.300	0.192
		coeff_var	1.735	1.038	7.229	6.616
3	SST + TURBIDITY + DIST_MAJWATERSHEDS + DIST_COAST	mean	0.775	0.865	0.272	0.197
		coeff_var	1.859	1.043	8.062	6.348
4	SST + TURBIDITY + DIST_MAJWATERSHEDS + DEPTH	mean	**0.808**	0.859	0.308	0.187
		coeff_var	1.611	0.992	6.626	6.236
5	SST + TURBIDITY + DIST_MAJWATERSHEDS	mean	0.762	0.841	0.230	0.165
		coeff_var	1.768	0.863	8.192	5.740