Supplementary Material S2: Catch Model Fitting and Performance

Described below are the seven leatherback turtle, swordfish, and blue shark correlative catch models used to define this study's eight operating models (Table 1). A model selection process was used for each model to determine a parsimonious set of covariates for each model. Static covariates (latitude, longitude, day-of-year, depth standard deviation) were deliberately included in only some models to explore structural differences, but all models included dynamic environmental covariates. These covariates included ocean variables from a data-assimilative configuration of the Regional Ocean Modelling System (ROMS), which were matched to each DGN set in the observer data based on each set's date and location. A broad set of ocean variables were examined for swordfish and blue sharks, prioritising sea-surface temperature (SST), isothermal layer depth (ILD), and finite-time Lyapunov exponent (FTLE), as they are known drivers of swordfish distribution (Brodie *et al.* 2018; Scales *et al.* 2018). Collinearity among included covariates was evaluated, and all covariates had variance inflation factors < 3 (Zuur *et al.* 2010). A reduced set of covariates was used for the leatherback turtle model, due to the small number of observed bycatch events. This covariate set was selected by minimizing the out-of-bag (OOB) error rate for models with 3-6 covariates from the available covariates. SST and SSH were prioritized, having been previously identified as influential for leatherback turtles (Eguchi *et al.* 2017).

Model performance for swordfish and blue shark models was measured using k-fold cross validation (k = 10). This process fitted each model to a majority of training data, and measured model predictive power by comparing observed values of a withheld minority of testing data against their predicted values from the trained model (using RMSE and Pearson correlation). This was repeated 'k' times using different subsets (or folds) of training and testing data, and the mean RMSE and correlation of these subsets represented model predictive performance. For the leatherback turtle random forest model, performance was measured using a cross-validated (i.e. out-of-bag, OOB) confusion matrix, returned by the 'print' function in the 'randomForest' R package (Liaw and Wiener 2002). AUC was not used, because it was unreliable given that most predicted values were of the 'no catch' class. Absolute model performance is less important in a closed-loop simulation, because the results are conditional on the operating model being true. However, it was nonetheless useful to verify that the models performed well, as this meant our MSE evaluated a plausible system.

LB Catch Models

Our MSE explored four possible distributions of leatherback turtles: LB1, LB2, LB1seas, and LB2scal (Table 1). Maps of these distributions are illustrated in Fig. S2.1.

The LB1 random forest catch model had the form:

$$CPUE = SST + SST_{sd} + SSH + Lon + Lat + H_{set}$$

The LB2 model had the form:

$$CPUE = SST + SST_{sd} + SSH + Curl + H_{set}$$

Where *CPUE* is the probability of catching one leatherback turtle per set. Additional covariates and their sources are detailed in Table S2.1. These models were fit using the 'randomForest' R package. We specified 700 trees, which was deemed sufficient by testing different tree numbers and examining OOB results. We used default values for the number of covariates randomly sampled as candidates at each split ('mtry'), and for the minimum size of terminal nodes ('nodesize'). Fitted responses for LB1 and LB2 are illustrated in Fig. S2.2, and performance results presented in Table S2.2.

Classification trees are sensitive to imbalance between classes (here, no catch = 0 and catch = 1), so we used a down-sampling procedure to address this imbalance. The majority class was down-sampled when training the random forest, such that each tree used a stratified bootstrap sample of the data with equal numbers of majority (0s) and minority (1s) class observations (Stock *et al.* 2019). Given the rarity of bycatch events, we set the sample size equal to the number of minority class observations. We found that another approach to address class imbalance, the synthetic minority over-sampling technique (Chawla *et al.* 2002; Stock *et al.* 2019) over-fit the data. Down-sampling increased model skill (43% error in LB1 identifying catches in down-sampled model, compared to 100% error with standard random forest), although at the expense of accuracy (down-sampled LB1 model also misclassified 27% of no catches as catches; Table S2.2).

Like other authors (Stock *et al.* 2018), we found that class imbalance-corrected random forests overpredicted bycatch rates, and these required rescaling. As stated in the main article, we also wanted to rescale catch rate to an inflated level suitable for our MSE simulation. We rescaled using Elkan's general updating approach for probability estimates for machine learning methods (Elkan 2001). Rescaling of values from a classification tree can be done when the base rate (i.e. occurrence of a class; in this case a turtle catch) used to fit the model differs from the population to be used for prediction. According to Equation 1 in Dankowski and Ziegler (2016), rescaling was done by:

$$P'(y) = \frac{b'(y - bP(y))}{b + b'P(y) - bP(y) - bb'}$$

Where $P(y)$ are the probabilities of a turtle catch ($y = 1$) in the California Current predicted using the down-sampled random forest (i.e. the *CPUE* specified above), b is the 'base rate' of turtle occurrence in the down-sampled data ($b = 0.5$), and $P'(y)$ are the updated probabilities given the desired base rate of turtle occurrence b'. We found the desired b' with an iterative process, by altering b' until the predicted number of presences for the observer data set achieved the desired catch rate ($b' = 0.1$). For both LB1 and LB2, this increased the actual catch rate from 23 turtle bycatch events to ~380 bycatch events (in the observed ~5700 sets).

The LB1seas turtle catch model was identical to LB1 in all aspects, but was multiplied by an additional factor to define each day's catch rate. The factor was a logistic function, defining the proportion of the LB1 catch rate remaining at a given day of the year (Fig. S2.3). The logistic function was parametrized to have a sharp decline in catch rate, with essentially zero catch rate by end of November. The LB2scal turtle catch model was fitted as LB2, but used a different form of rescaling than the general updating approach described above. Rescaling of

the catch rate was done by multiplying the down-sampled catch rate by the ratio of desired:actual bycatch events. The desired value was specified to give, as for the other models, a prediction of ~380 bycatch events (in the observed ~5700 sets). Because LB2 and LB2scal used the same catch model, the habitat preferences of turtles represented by these models were largely identical, with the only difference being that the LB2 model made an increased distinction between 'good' and 'bad' habitat (Fig. S2.1).

Table S2.1. Description and sources of covariates used in the seven catch models. All dynamic variables were resolved at a daily time-step. 'Range' is the minimum and maximum values in the 1990-2000 observer data used to fit the models.

Covariate	Units	Range	Description and Source
SST	°C	11.8, 23.9	Sea-surface temperature. Sourced from ROMS.
SST_{sd}	°C	0.04, 1.45	The spatial standard deviation of sea-surface temperature. Calculated over a 0.7° square. The variation of SST can indicate thermal fronts. Derived from ROMS
ILD	m	1.4, 109.4	Isothermal layer depth, calculated as the depth corresponding to a 0.5°C temperature difference relative to sea surface temperature. *ILD* is an index of water column structure, specifically the depth of surface mixing. Derived from ROMS
SSH	m	-0.04, 0.42	Sea-surface height. Sourced from ROMS
EKE	$m^2\ s^{-2}$	-13.9, -1.4	Eddy kinetic energy. This was calculated as the sum of eastward surface current velocity squared and northward surface current velocity squared, divided by two. *EKE* indicates the presence and intensity of eddies. Derived from ROMS
FTLE	d^{-1}	-0.28, 0	Finite-time Lyapunov exponent. *FTLE* is a Lagrangian coherent structure that measures the maximum separation of close-by particles of a time-dependent flow field after a fixed, finite particle advection time (Watson *et al.* 2018), and indicates fronts. Derived from ROMS
Curl	$N\ m^{-2}$	-1.58×10^{-6}, 2.85×10^{-6}	Wind stress curl. Sourced from ROMS.
Z_{sd}	m	10.3, 1334	The standard deviation of ocean bottom depth at each set (rugosity), calculated over a 0.3˚ x 0.3˚ square. Derived from ETOPO1 (interpolated to 0.1°), obtained from https://www.ngdc.noaa.gov/mgg/global/global.html
Lat, Lon	degrees	30.17, 46.43 -129.15, -117.25	The latitude and longitude of each set. Taken as the coordinates specified by the observer
Time	years	0, 10.4	Continuous time (decimal years)
DOY	day	1, 366	Day of calendar year
Distance	km	3.4, 453	Distance of each set from the coast. Calculated as the haversine distance using the location specified by the observer
H_{set}	h	1, 20	The duration of each DGN set. This is specified in the model prediction to estimate mean catch for a given set duration. This was evaluated as an offset term, but the response was non-linear so it was included as a covariate
Vessel		132 vessels	A vessel identifier. Used in the GAMM as a random effect to account for the dependency in catch rates by the same vessel

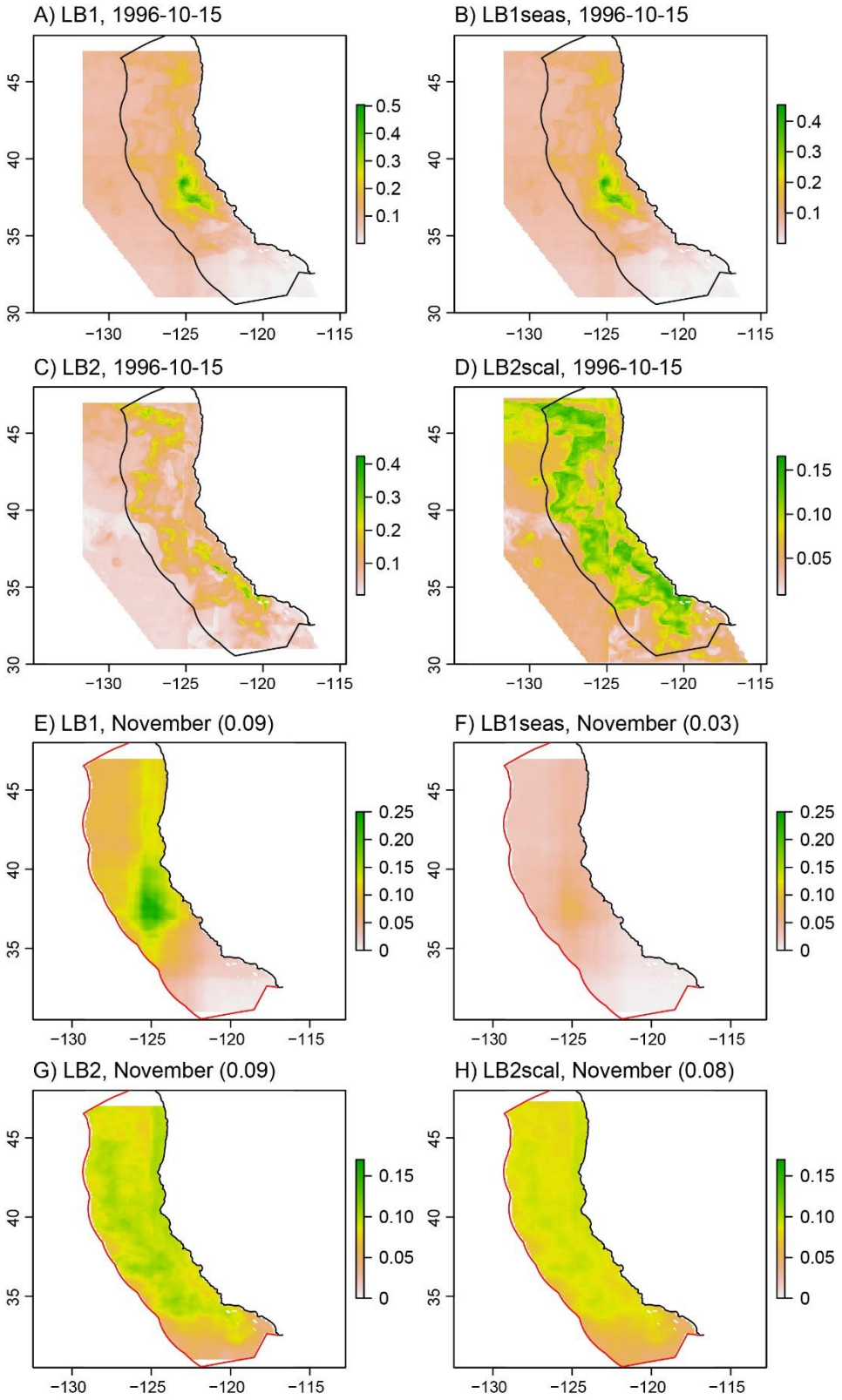

Fig. S2.1. Predicted catch by the four leatherback turtle catch models (**Table 1**), for an example date (A-D) and as the mean of all November days in 1991-2000 (E-H, mean in parentheses). Color is the mean probability of catching one turtle per 12h set. Note that LB1 and LB1seas have an identical spatial distribution, but the LB1seas catch rate declines due to a forced migration (Fig. S2.3). Note LB2 and LB2scal have near identical catch distributions, but with less difference between high and low catch rate areas for LB2scal.

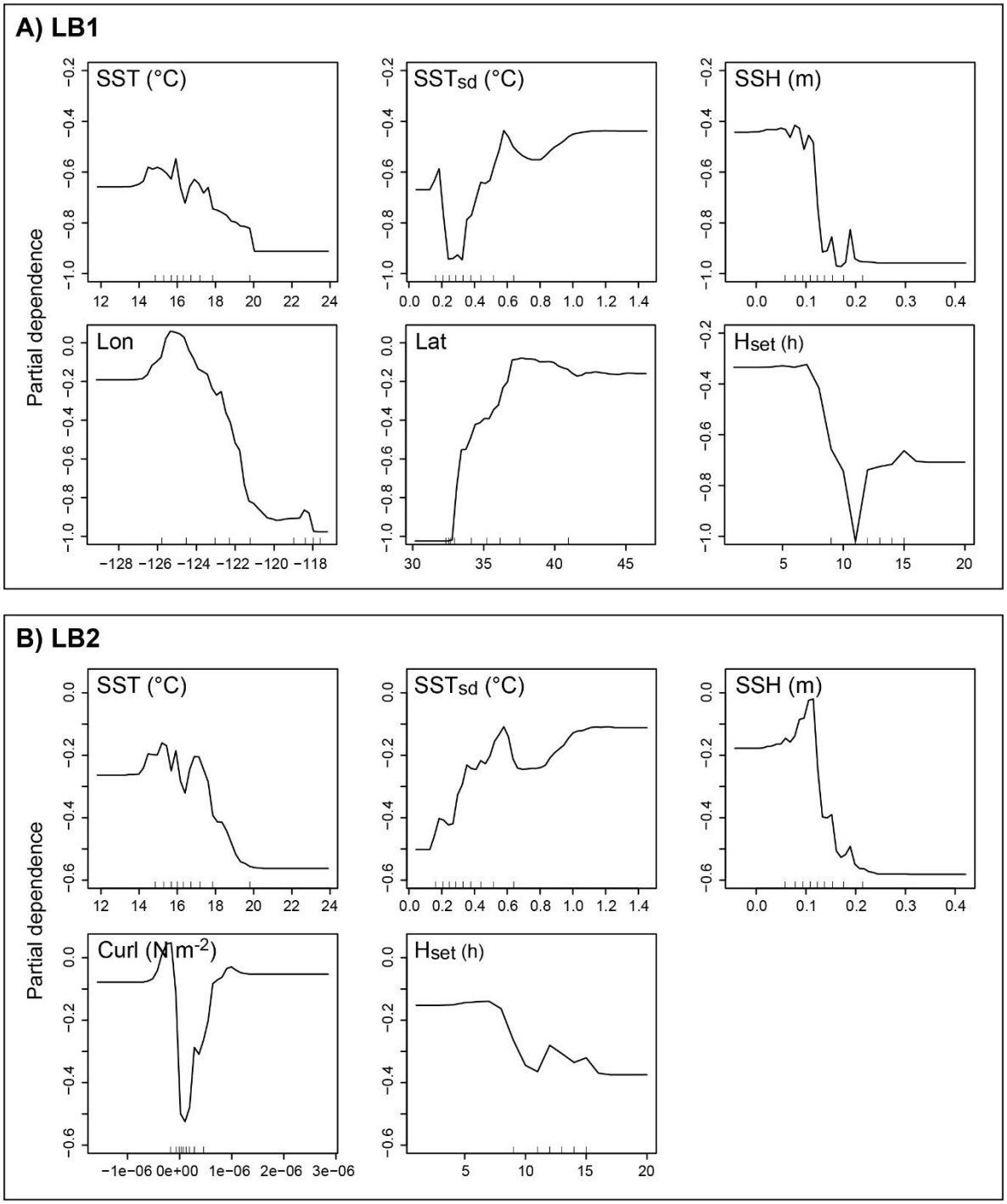

Fig. S2.2. Fitted response curves for the LB1 and LB2 catch models. Increasing values indicate an increased probability of a bycatch event.

Table S2.2. Cross-validated confusion matrix of the prediction (based on out-of-bag OOB data) for the down-sampled leatherback turtle catch random forest models (LB1 and LB2). The class-error represents prediction error; e.g. for LB1 there was 43% error predicting an OOB catch (10 of the 23 observed catch events were incorrectly predicted to be no catch). The OOB estimates of total model error rate were LB1 = 26.7% and LB2 = 24.5%.

LB1	Predicted No Catch	Predicted Catch	Class error
Observed No Catch	4221	1530	0.27
Observed Catch	10	13	0.43
LB2	Predicted No Catch	Predicted Catch	Class error
Observed No Catch	4351	1400	0.24
Observed Catch	14	9	0.61

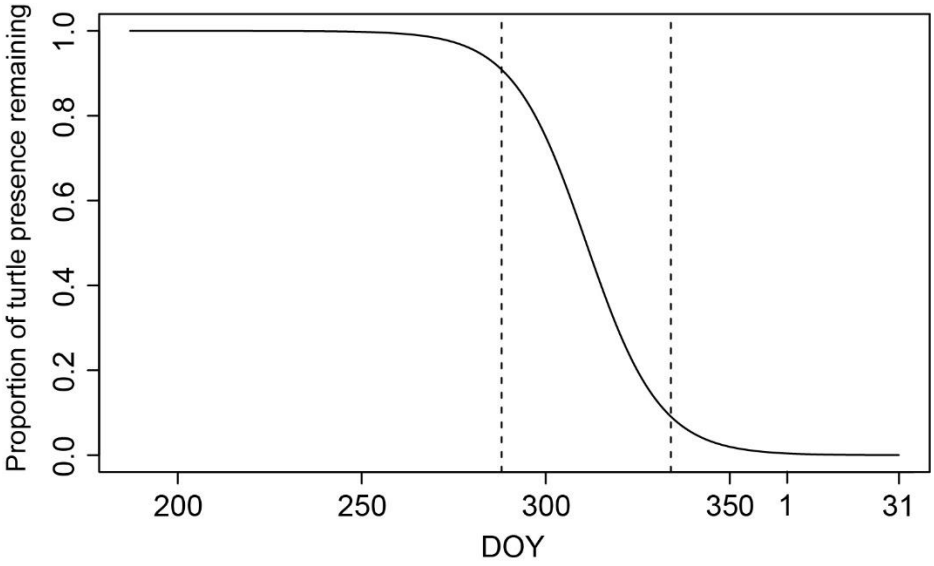

Fig. S2.3. The logistic 'migration' function used to reduce the LB1 catch rates, reducing the bycatch risk to ~0 by end of November. The vertical dotted lines indicate two dates: 15th Oct. and 31st Nov. DOY is day-of-year.

SF Catch Models

Our MSE explored two possible distributions of swordfish in our MSE, as operating models: SF1 and SF2 (**Table 1**). Example maps of these distributions are illustrated in Fig. S2.4.

The SF1 BRT swordfish catch model had the form:

$$CPUE = SST + SST_{sd} + ILD + SSH + EKE + Z_{sd} + FTLE + Dist + Lat + H_{set}$$

The SF2 GAMM catch model had the form:

$$CPUE = s(SST) + s(SST_{sd}) + s(ILD) + s(SSH) + s(EKE) + s(Z_{sd}) + s(FTLE) + te_{sw,cr}(Lon, Lat, Time) + s_{cc}(DOY) + s(H_{set}) + s_{re}(Vessel)$$

CPUE is the number of swordfish per set; additional covariates and their sources are detailed in Table S2.1. For the GAMM models, *s* indicates a thin plate regression spline, s_{re} indicates a smooth representing an i.i.d random effect, s_{cc} indicates a cyclic cubic regression spline (to model seasonality), and $te_{sw,cr}$ indicates a tensor product smooth with soap film smoother for the spatial covariates and a cubic regression spline for time.

As in Smith *et al*. (2020), the BRT was fitted to observed swordfish catches using a learning rate of 0.01, a tree complexity of 3, and a bag fraction of 0.6 (Elith *et al.* 2008), using the function 'gbm.step' in the 'dismo' R package (Hijmans *et al.* 2017). Random effects cannot be added to this form of BRT, but we examined the relationship between BRT residuals and *Vessel* to identify an effect on *CPUE* (Buston and Elith 2011). There was a residual effect, as measured with a GLRT test (section 3.5; Wood 2017), but we consider this minor as the interquartile range of the residuals for most vessels encompassed zero.

The GAMM was structurally quite different to the BRT, by modelling space and continuous time. The GAMM was fitted using the 'mgcv' R package (Wood 2011). We used soap film smoother to allow more control of catches at the boundary (Wood 2017), with swordfish catches shrinking to zero at the EEZ (using the 'sw' smoother). This was done to help reduce movement of vessels into areas far offshore that were predicted to be good fishing habitat, but were uncertain and potentially misidentified. We specified the western soap film boundary to be the EEZ, the northern and southern boundaries to be slightly beyond the Washington border (where fishing is not permitted) and slightly beyond the EEZ respectively, and the eastern boundary to be slightly inland (slightly inland to not unrealistically shrink catches at the coast). Knot locations for the soap-film smoother were selected manually using the 'locator' function (Wood 2017), and spread approximately evenly throughout the fishable domain (recommended for spatially unbalanced data; Thorson 2019). The boundary and knots used for the soap film smoother are shown in Fig. S2.5. We used 91 knots, which balanced flexible fit and computational feasibility.

Fitted responses for SF1 and SF2 are illustrated in Fig. S2.6 and S2.7, and model performance is reported in Table S2.3.

BS Catch Models

Our MSE explored one distribution of blue sharks in our MSE: BS1 (Table 1). This Poisson BRT had the form:

$$CPUE = SST + SST_{sd} + ILD + SSH + EKE + FTLE + Curl + H_{set}$$

CPUE is the number of blue sharks per set; additional covariates and their sources are detailed in Table S2.1. This BRT was fitted as per the swordfish BRT. An example map of this distribution is illustrated in Fig. S2.8, the fitted responses are illustrated in Fig. S2.9, and model performance is reported in Table S2.4.

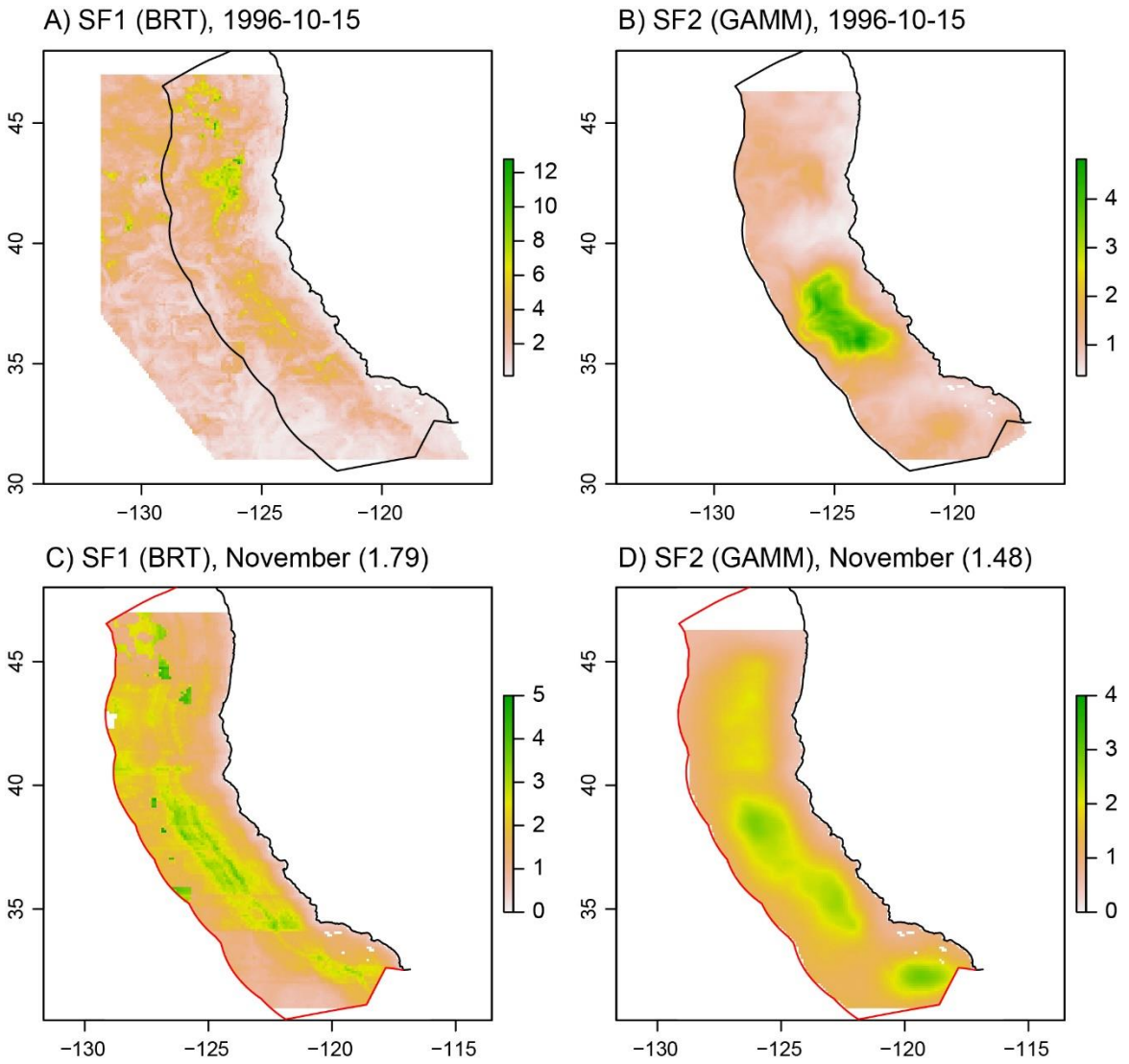

Fig. S2.4. Predicted catch by the two swordfish catch models (**Table 1**), for an example date (A-B) and as the mean of all November days in 1991-2000 (C-D, mean in parentheses). Color is the predicted mean number of swordfish per 12h set.

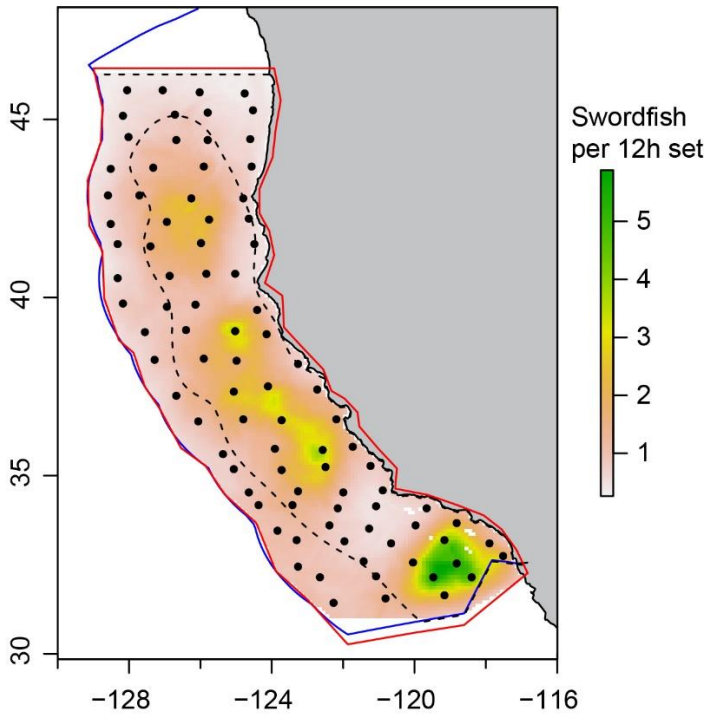

Fig. S2.5. The boundary (red line) and knots (dots) used for the soap film smoother. Also shown is an example prediction of swordfish catch from this model (color; 20th Dec. 1997), and a kernel density contour enclosing 99% of observed fishing effort (1990-2000; dashed line).

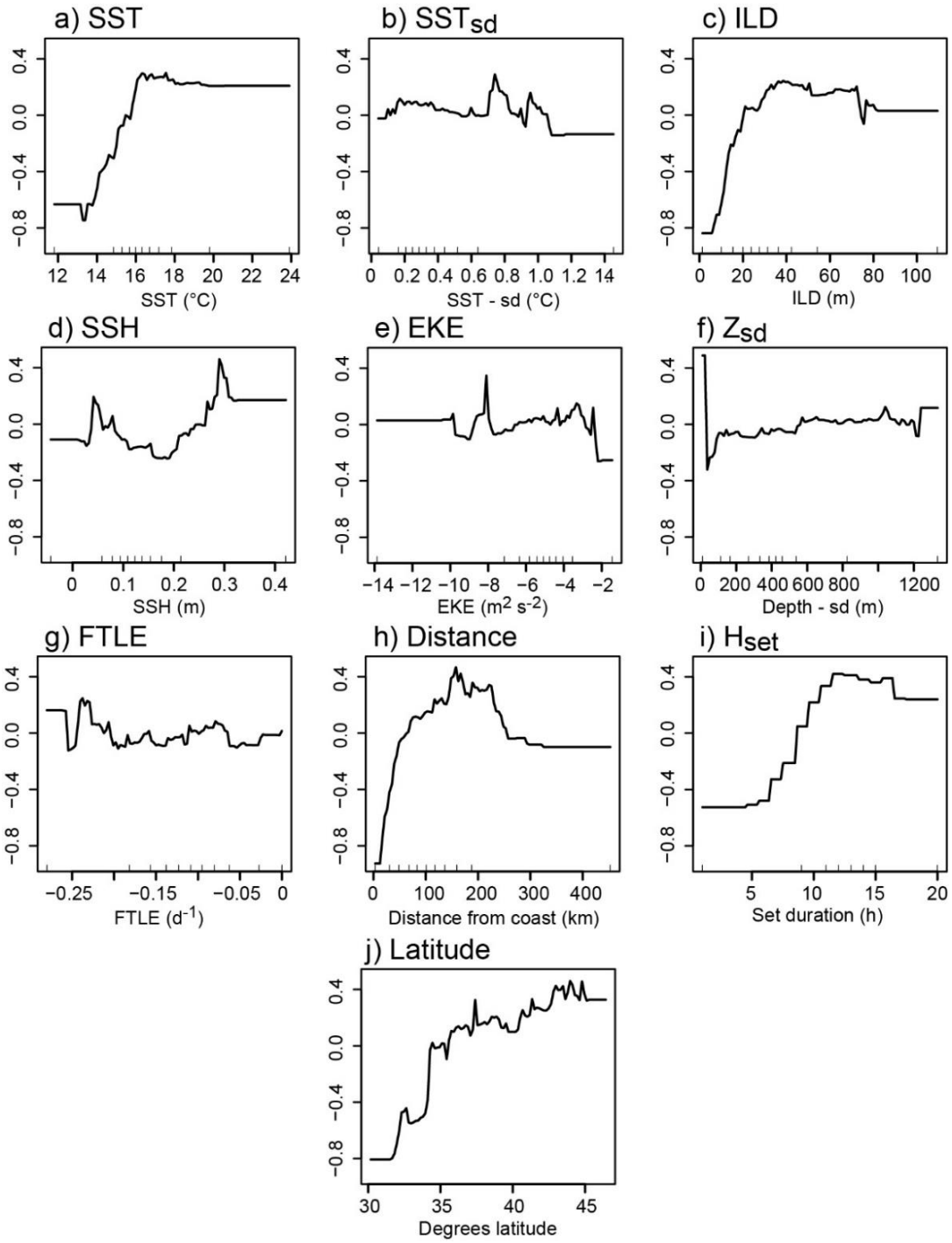

Fig. S2.6. Fitted responses in the swordfish BRT (SF1). See Table S2.3 for model performance.

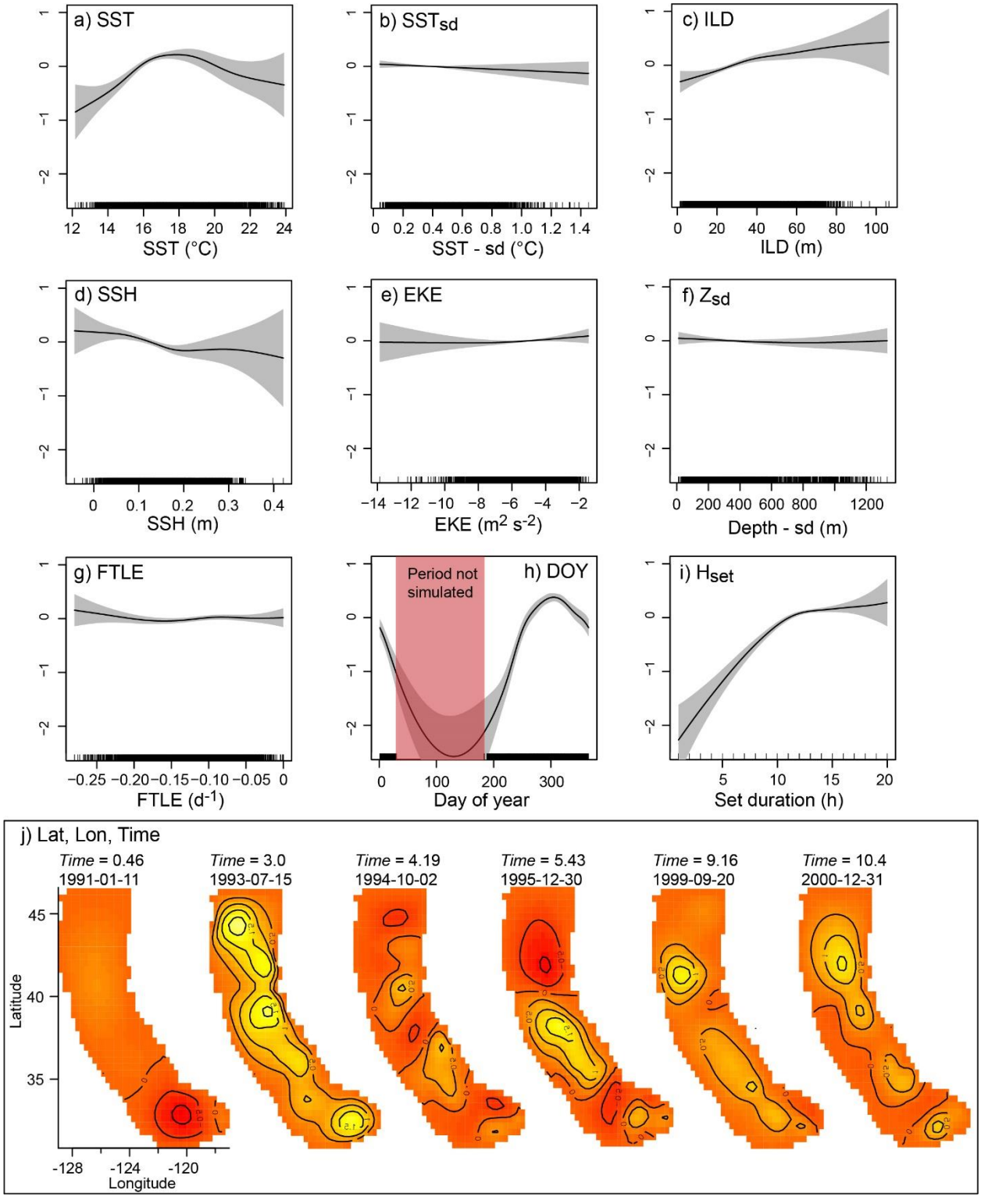

Fig. S2.7. Fitted smoothers for the swordfish spatial-temporal GAMM (SF2), with twice the standard error (grey area). The period shaded red for DOY was not part of the simulated fishing season. Panel j) illustrates the fitted catch for a subset of dates from the space-time smoother (yellow = high, red = low). See Table S2.3 for model performance.

Table S2.3. Results from k-folds cross-validation of the swordfish catch models. We fitted each model to a training majority of data ('train'), and tested the predictive power of a model using the held back minority of data ('test'). The mean root-mean-square-error (RMSE) and Pearson correlation (Cor) of the predicted and observed *test* data represents predictive power (better model highlighted grey). The RMSE of the *train* data, and the explained deviance of the model fitted to train data, represent model goodness-of-fit (best model in bold).

Metric	SF1 (BRT)	SF2 (GAMM)
RMSE test mean	2.496	2.449
RMSE test sd	0.156	0.173
Cor. test mean	0.483	0.513
Cor. test sd	0.051	0.049
RMSE train mean	**1.993**	2.210
RMSE train sd	0.035	0.024
Expl. Dev. % mean	**46.36**	42.43
Expl. Dev. % sd	1.15	0.84

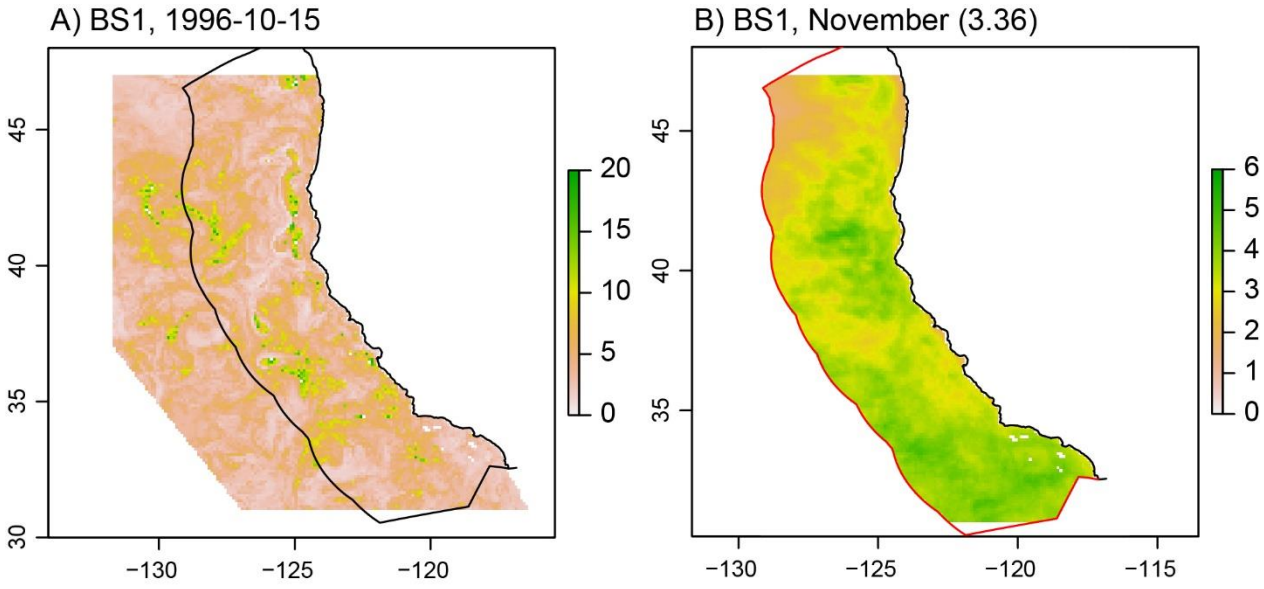

Fig. S2.8. Predicted catch by the blue shark catch model (**Table 1**), for an example date (A) and as the mean of all November days in 1991-2000 (B, mean in parentheses). Color is the predicted mean number of blue sharks per 12h set.

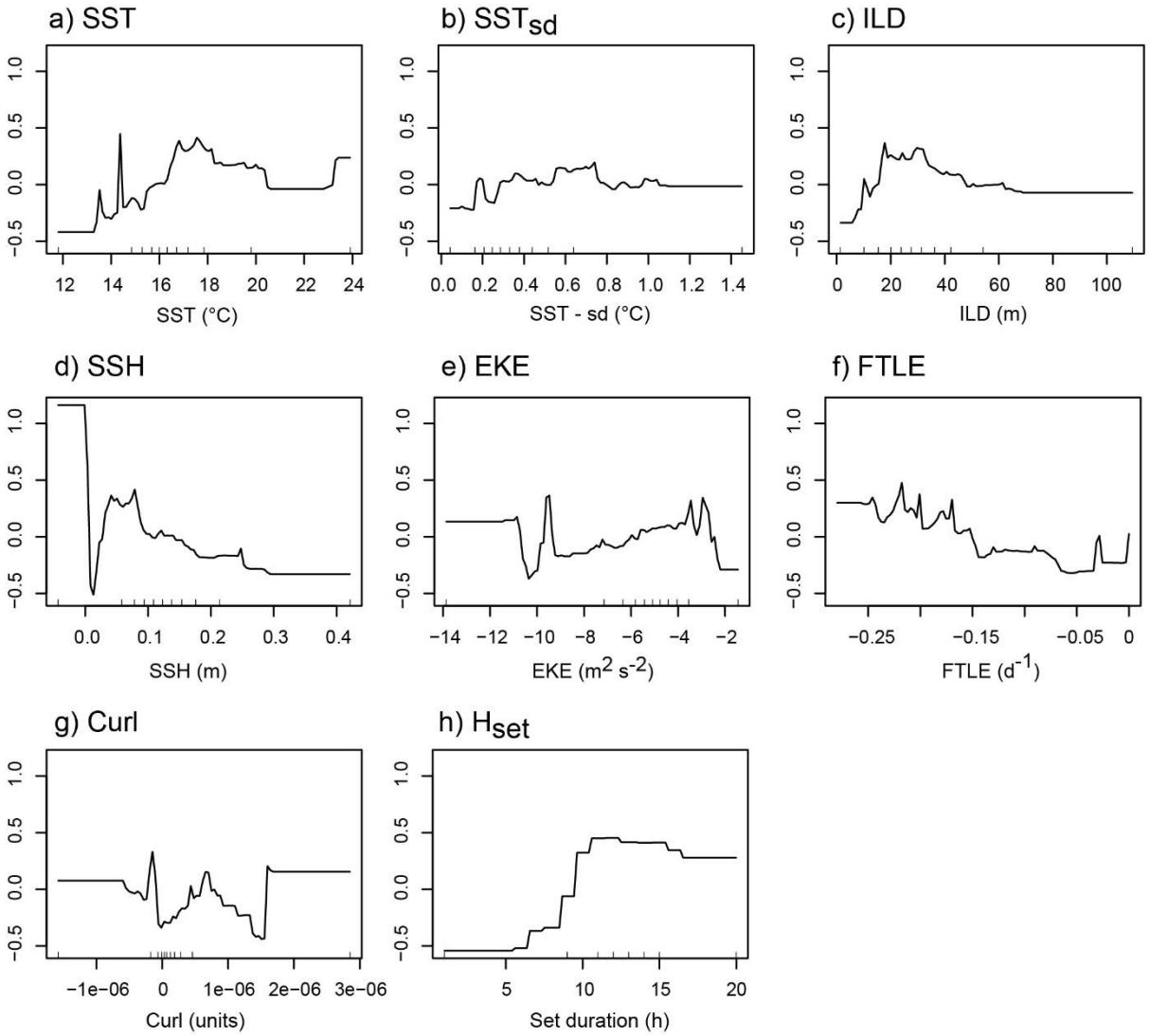

Fig. S2.9. Fitted responses in the blue shark BRT (BS1). See Table S2.4 for model performance.

Table S2.4. Results from k-folds cross-validation of the blue shark catch models. The mean root-mean-square-error (RMSE) and Pearson correlation (Cor) of the predicted and observed *test* data represents predictive power. The RMSE of the *train* data, and the explained deviance of the model fitted to train data, represent model goodness-of-fit.

Metric	**BS1 (BRT)**
RMSE test mean	5.924
RMSE test sd	0.689
Cor. test mean	0.313
Cor. test sd	0.040
RMSE train mean	4.740
RMSE train sd	0.104
Expl. Dev. % mean	40.21
Expl. Dev. % sd	2.18

References

Brodie, S., Jacox, M.G., Bograd, S.J., Welch, H., Dewar, H., Scales, K.L., Maxwell, S.M., Briscoe, D.M., Edwards, C.A., Crowder, L.B., Lewison, R.L., and Hazen, E.L. (2018) Integrating dynamic subsurface habitat metrics into species distribution models. *Frontiers in Marine Science* **5**(219). doi:10.3389/fmars.2018.00219

Buston, P.M., and Elith, J. (2011) Determinants of reproductive success in dominant pairs of clownfish: a boosted regression tree analysis. *Journal of Animal Ecology* **80**(3), 528-538. doi:10.1111/j.1365-2656.2011.01803.x

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002) SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321-357.

Dankowski, T., and Ziegler, A. (2016) Calibrating random forests for probability estimation. *Statistics in medicine* **35**(22), 3949-3960.

Eguchi, T., Benson, S.R., Foley, D.G., and Forney, K.A. (2017) Predicting overlap between drift gillnet fishing and leatherback turtle habitat in the California Current Ecosystem. *Fisheries Oceanography* **26**(1), 17-33. doi:10.1111/fog.12181

Elith, J., Leathwick, J.R., and Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology* **77**(4), 802-813. doi:10.1111/j.1365-2656.2008.01390.x

Elkan, C. The foundations of cost-sensitive learning. In 'International joint conference on artificial intelligence', 2001, pp. 973-978

Hijmans, R.J., Phillips, S., Leathwick, J., and Elith, J. (2017). dismo: Species Distribution Modeling. R package version 1.1-4. https://CRAN.R-project.org/package=dismo.

Liaw, A., and Wiener, M. (2002) Classification and regression by randomForest. *R news* **2**(3), 18-22.

Scales, K.L., Hazen, E.L., Jacox, M.G., Castruccio, F., Maxwell, S.M., Lewison, R.L., and Bograd, S.J. (2018) Fisheries bycatch risk to marine megafauna is intensified in Lagrangian coherent structures. *Proceedings of the National Academy of Sciences* **115**(28), 7362-7367. doi:10.1073/pnas.1801270115

Smith, J.A., Tommasi, D., Sweeney, J., Brodie, S., Welch, H., Hazen, E.L., Muhling, B., Stohs, S.M., and Jacox, M.G. (2020) Lost opportunity: Quantifying the dynamic economic impact of time-area fishery closures. *Journal of Applied Ecology* **57**(3), 502-513. doi:10.1111/1365-2664.13565

Stock, B.C., Ward, E.J., Eguchi, T., Jannot, J.E., Thorson, J.T., Feist, B.E., and Semmens, B.X. (2019) Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. *Canadian Journal of Fisheries and Aquatic Sciences* **77**(1), 146-163. doi:10.1139/cjfas-2018-0281

Stock, B.C., Ward, E.J., Thorson, J.T., Jannot, J.E., and Semmens, B.X. (2018) The utility of spatial model-based estimators of unobserved bycatch. *ICES Journal of Marine Science* **76**(1), 255-267. doi:10.1093/icesjms/fsy153

Thorson, J.T. (2019) Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research* **210**, 143-161. doi:10.1016/j.fishres.2018.10.013

Watson, J.R., Fuller, E.C., Castruccio, F.S., and Samhouri, J.F. (2018) Fishermen follow fine-scale physical ocean features for finance. *Frontiers in Marine Science* **5**(46). doi:10.3389/fmars.2018.00046

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(1), 3-36. doi:10.1111/j.1467-9868.2010.00749.x

Wood, S.N. (2017) 'Generalized Additive Models: An Introduction with R.' 2nd edn. (Chapman and Hall/CRC Press: Boca Raton, Florida)

Zuur, A.F., Ieno, E.N., and Elphick, C.S. (2010) A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* **1**(1), 3-14. doi:10.1111/j.2041-210X.2009.00001.x