Vol. 43: 183–197, 2020
https://doi.org/10.3354/esr01060

ENDANGERED SPECIES RESEARCH
Endang Species Res

Published October 8

Ensemble Random Forests as a tool for modeling rare occurrences

Zachary A. Siders[1,*], Nicholas D. Ducharme-Barth[2], Felipe Carvalho[3], Donald Kobayashi[3], Summer Martin[3], Jennifer Raynor[4], T. Todd Jones[3], Robert N. M. Ahrens[3]

[1]UF/IFAS SFRC Fisheries and Aquatic Sciences Program, University of Florida, Gainesville, FL 32611, USA
[2]Oceanic Fisheries Programme, Pacific Community, Nouméa 98800, New Caledonia
[3]NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI 96818, USA
[4]Department of Economics, Wesleyan University, Middletown, CT 06457, USA

ABSTRACT: Relative to target species, priority conservation species occur rarely in fishery interactions, resulting in imbalanced, overdispersed data. We present Ensemble Random Forests (ERFs) as an intuitive extension of the Random Forest algorithm to handle rare event bias. Each Random Forest receives individual stratified randomly sampled training/test sets, then down-samples the majority class for each decision tree. Results are averaged across Random Forests to generate an ensemble prediction. Through simulation, we show that ERFs outperform Random Forest with and without down-sampling, as well as with the synthetic minority over-sampling technique, for highly class imbalanced to balanced datasets. Spatial covariance greatly impacts ERFs' perceived performance, as shown through simulation and case studies. In case studies from the Hawaii deep-set longline fishery, giant manta ray *Mobula birostris* syn. *Manta birostris* and scalloped hammerhead *Sphyrna lewini* presence had high spatial covariance and high model test performance, while false killer whale *Pseudorca crassidens* had low spatial covariance and low model test performance. Overall, we find ERFs have 4 advantages: (1) reduced successive partitioning effects; (2) prediction uncertainty propagation; (3) better accounting for interacting covariates through balancing; and (4) minimization of false positives, as the majority of Random Forests within the ensemble vote correctly. As ERFs can readily mitigate rare event bias without requiring large presence sample sizes or imparting considerable balancing bias, they are likely to be a valuable tool in bycatch and species distribution modeling, as well as spatial conservation planning, especially for protected species where presence can be rare.

KEY WORDS: Rare event bias · Species distribution modeling · Protected species · Bycatch · Machine learning · Random Forest

1. INTRODUCTION

Machine learning algorithms have proven to be a ubiquitous tool for modeling species distributions. A subset of these approaches are nonparametric and have the potential to model complex, non-linear environmental responses that drive some species distributions (Breiman 2001b, Olden et al. 2008, Stock et al. 2020). For marine species, many environmental covariates are proxies for a given organism's environmental cues, increasing the importance of modeling non-linear responses (Pielke et al. 2003, Robinson et al. 2011). Unfortunately, many machine learning algorithms are designed for 'big data',

*Corresponding author: zsiders@ufl.edu

Publisher: Inter-Research · www.int-res.com

which is lacking for many species (Martin et al. 2015), especially those that are priority conservation concerns, where observations are often a rare occurrence. Modeling these rare events requires special consideration, but the resulting species distribution is critical for determining key environmental covariates of the distribution, establishing conservation areas, and identifying potential conservation threats (Millar et al. 2007, Elith & Leathwick 2009, Evans et al. 2011).

Rare events induce an intrinsic imbalanced data problem where the presences of a species are underrepresented, skewing the distribution of classes (e.g. the presence and absence of a given species) (He & Garcia 2009, Kuhn & Johnson 2013). This has been broadly referred to as 'rare event bias', where the data are typically overdispersed, exhibiting greater variability than expected from standard probability distributions (Martin et al. 2015). A variety of methods have been utilized in ecology to deal with overdispersion (e.g. zero-inflated, hurdle, or delta–generalized linear models), most of which implement a mixture model in some fashion (Zuur et al. 2009, Campbell 2015, Stock et al. 2020). While these models can be well suited to modeling rare event data, large sample sizes are often needed to overcome the low proportion of presences and estimate model parameters with adequate precision (He & Garcia 2009, Rodriguez-Galiano et al. 2012). Further, many of these models are conditioned on linear responses that can be inadequate for modeling species distributions (Breiman 2001b, Austin 2007, Olden et al. 2008, Merow et al. 2014).

Decision tree approaches, in particular Random Forest (RF) (Breiman 2001a), are well suited to modeling rare event data (Strobl et al. 2007). Random Forest is a supervised machine-learning algorithm based on generating a 'forest' of decision trees grown in parallel, referred to as bagging in machine learning. For each decision tree in the forest, classified data are passed through the tree and covariates are used to separate the data into classified groups (typically binomial data but can be categorical or continuous) (Breiman 2001a). Categorical and continuous covariates are then used (with cutoffs applied to continuous datasets) to make decisions at every node in individual decision trees to attempt to best separate the classes. A random subset of data is provided for each tree (the in-the-bag dataset) and a random subset of covariates are tried at each node in each tree. Each classification tree votes on a datum's class and the proportion of trees that vote correctly equal the probability of that class occurring for each observation.

Relative to other machine learning approaches, RF tends to balance overfitting and prediction well through the generation of many 'weak learners' (the individual decision trees) and the creation of an ensemble of their predictions (the leaves) to generate a 'strong learner' (the forest) (Breiman 2001b, Evans et al. 2011). Additionally, the performance of RF has been generally high in building species distribution models with rare species (Williams et al. 2009, Rodriguez-Galiano et al. 2012). In such models, randomness is imparted in 2 ways: (1) a random subset of the data is passed through a given tree in the 'forest'; and (2) random covariates are drawn at each node split when growing the tree, resulting in high diversity within the 'forest' as well as accounting for covariate interactions. Nonetheless, rare event bias and the imbalanced data problem result in low performing RFs due to 2 phenomena: (1) successive partitioning of the data when growing the decision trees causes them to 'see' fewer and fewer of the rare events, thus fitting more and more to the majority class (the absences); and (2) interactions between covariates can go unlearned by the decision trees due to the sparseness of the data induced by partitioning (Fig. 1A) (He & Garcia 2009).

A variety of sampling methods have been proposed to deal with the class imbalance problem in RF (Kuhn & Johnson 2013). Down-sampling methods select data points from the minority and majority classes such that both classes end up with approximately the same sample size. Internally down-sampled approaches accomplish this balancing via a stratified random sample of the full dataset to generate the subsets passed to each tree (Kuhn & Johnson 2013). However, down-sampling imparts a bias against the majority class, resulting in weak learning RF and poor model performance (Fig. 1B) (He & Garcia 2009). Up-sampling methods generate pseudo-replicates of the minority class to correct the class imbalance. Chawla et al. (2002) describe the synthetic minority oversampling technique (SMOTE) that randomly selects a minority class datum, calculates the K nearest neighbors, and randomly blends data from these nearest neighbors into pseudo-replicates of the minority class. Additionally, the SMOTE approach can down-sample the majority class to further help balance the classes. As Kuhn & Johnson (2013) discuss, these resampling methods can resolve class imbalances, but there is little consensus on the best approach.

In this study, we present Ensemble Random Forests (ERF) as a method to mitigate the bias against the majority class that arises from balancing imbalanced

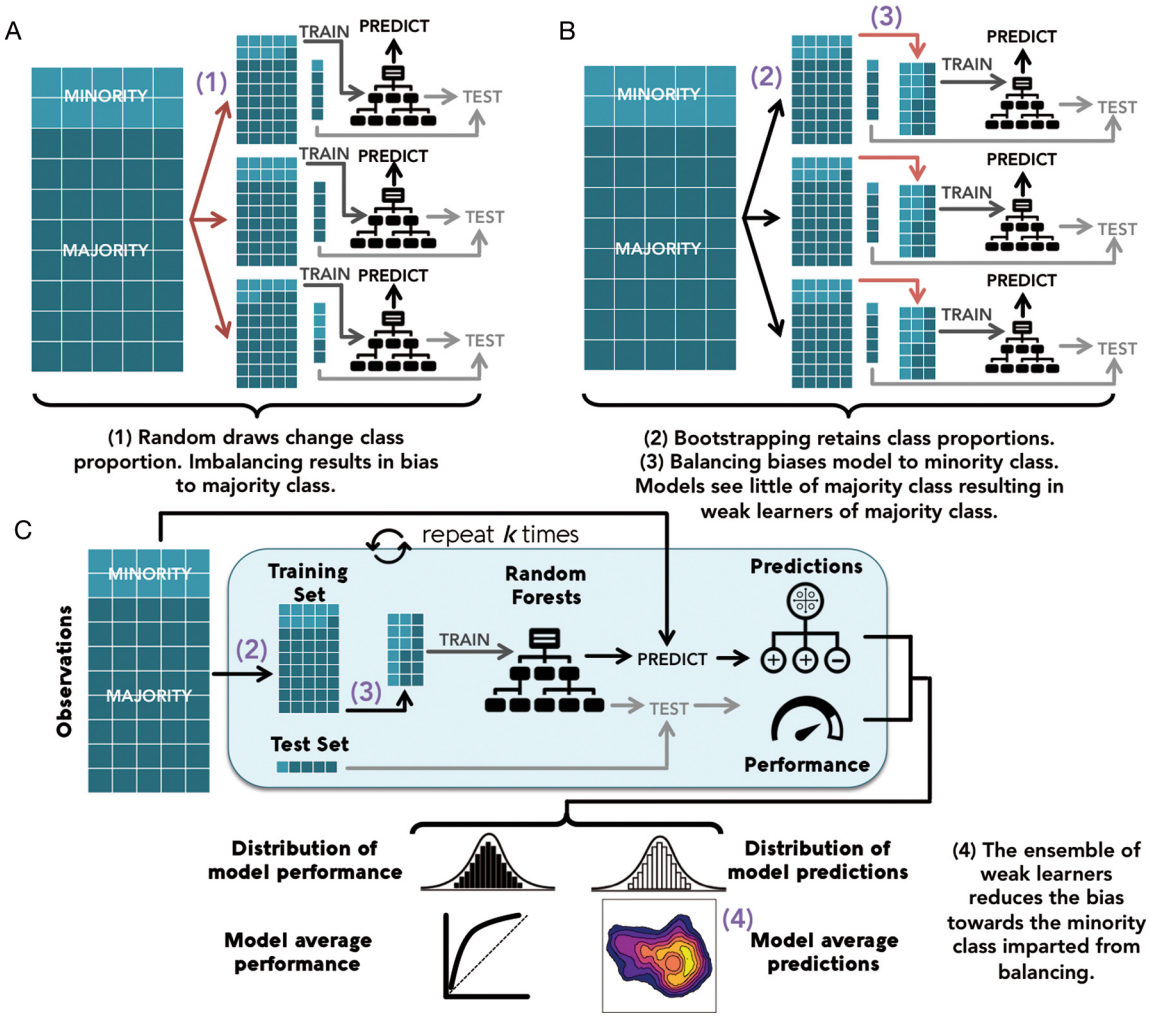

Fig. 1. A schematic of a classical Random Forest implementation for a balanced dataset (A) using random draws (1) to generate the training/test sets. (B) Bootstrapping (2) and balancing (3) are techniques used to modify the classical implementation to deal with imbalanced class datasets. (C) The Ensemble Random Forests modeling framework proposed in this study iterates these modifications to generate an ensemble of Random Forest models. Observations (presence–absence data) are split into training and test datasets using bootstrapping (2), the former is balanced (3) and used to train a Random Forest model, while the latter is used to assess model performance, and the combined datasets are used to make model predictions. This process can be repeated *k* times to generate *k* models and distributions of model performance metrics and predictions. The mean and uncertainty of model performance metrics and predictions (4) can then be extracted to generate the results of the Ensemble Random Forests models

datasets and, as a result, reduce rare event bias. The overall concept is simple: train an ensemble of RFs with down-sampling and average over the predictions of each RF in the ensemble. Using simulations, we compare ERF to other RF-based rare event approaches and assess the performance of ERF across a range of spatial covariance in the simulated data. Specifically, this study (1) compares, using simulation, the performance of the standard, down-sampling, SMOTE, and ensemble implementations of RFs as a function of class imbalance for 3 performance metrics; and (2) evaluates the importance of spatial contrast in ERF performance through simulation. To demonstrate the functionality of ERF to model real rare events, we modeled the distribution of pelagic species bycatch in the central north Pacific using fisheries-dependent data from the Hawaii-based deep-set longline fishery. Modeling the distributions of pelagic species is particularly difficult: few scientific surveys are directed at capturing species presence–absence, resulting in a reliance on bycatch in fisheries. However, these bycatch events are often rare (between 100:1 and 1000:1), or ultra-rare (between 1000:1 and 10 000:1) (Martin et al. 2015). We

used 2 case studies to (1) evaluate the performance and prediction uncertainty of ERF as a function of class imbalance using a commonly caught, secondary target species in the Hawaiian deep-set longline fishery, wahoo *Acanthocybium solandri*; and (2) compare spatial contrast and ERF predictive ability for 3 Endangered Species Act listed species: giant manta ray *Mobula birostris* syn. *Manta birostris* (White et al. 2018) (Threatened), scalloped hammerhead *Sphyrna lewini* (Endangered), and false killer whale *Pseudorca crassidens* (Endangered). The purpose of these simulation and case study objectives was to demonstrate the effects of sample size and sample covariation on model performance and the utility of ERFs as a bycatch and species distribution modeling tool.

2. MATERIALS AND METHODS

2.1. Ensemble Random Forests

2.1.1. Background

We refer to ERF as an intuitive extension of RF for 2 reasons. The first is because the ERF method builds upon the existing algorithmic architecture of RF that bags the data into subsets for each decision tree within the forest (Fig. 1A). ERFs extend this bagging procedure to double bagging: bagging first within each RF to grow decision trees and then by generating train/test data subsets for each RF within the ensemble (Fig. 1). The second reason is that ensemble modeling is recommended as common practice for species distribution models (Araújo & New 2007) and, in practice, each individual RF is already an ensemble of decision trees. The procedure is roughly equivalent to a *k*-fold cross-validation procedure where *k* paired training and test partitions are drawn from the full dataset. ERFs differ from *k*-fold cross-validation in that each model's predictions are used to form an ensemble prediction.

Overall, ERFs are straightforward to implement, employing many of the same 'best practices' recommended for RF when using imbalanced class datasets (Cutler et al. 2007, Rogan et al. 2008, Evans et al. 2011, Kuhn & Johnson 2013). Each RF receives mutually exclusive random subsets of the data: one to grow the decision trees, the training set, and a second to conduct internal validation, the test set. These datasets are created by stratified random draws with replacement from the total dataset, in order to generate training and test datasets with class proportions close to those of the total dataset. Each RF grows decision trees on subsets of their individual training set, called the in-the-bag dataset, and holds out some data, the out-of-bag dataset, used to internally measure the prediction error of the RF. When growing the decision trees, each RF uses a balanced weighting scheme (i.e. equal proportions of each class), also called down-sampling (Kuhn & Johnson 2013), to reduce the bias imparted by class imbalance (Fig. 1B). Each RF then predicts on the test dataset, to generate individual RF performance metrics, and on the full dataset. The predictions of each RF on the full dataset are then averaged to form the ERF predictions (Fig. 1C) (Marmion et al. 2009). Performance metrics are also calculated for the ensemble predictions.

2.1.2. Implementation

The base RF procedure, implemented using the randomForest package (Liaw & Wiener 2002) in R 3.6.1 (R Core Team 2018), was extended to create the ERF routine. Each RF received a training set containing 90% of the data and a test set of 10%, where the train and test subsets received a representative sample (i.e. the proportion of classes, presence and absence, was equal to their proportion in the whole dataset). The individual decision trees were trained on balanced in-bag datasets (i.e. an equal number of presences and absences was provided to each tree). Each RF model grew 500 decision trees and randomly drew 3 covariates at each node in each tree. The number of decision trees to fit and the number of covariates to try at each node was optimized using a single RF model for each case study. The number of RFs in the ensemble was optimized by determining the minimum number of RFs that resulted asymptotic behavior in the variance of model performance across a suite of case studies.

2.1.3. Performance

The test dataset for each RF in the ensemble was used for internal validation and to generate metrics of individual model performance. We chose to focus on 3 performance metrics: (1) the area under the curve (AUC), (2) the root mean squared error (RMSE), and (3) the true skill statistic (TSS). The receiver operator characteristic curve (ROC), which plots 1-specificity against the sensitivity for various thresholds (binary cutoff values) between 0 and 1 was used

to determine AUC. ROC metrics were calculated using the ROCR package (Sing et al. 2005) in R. AUC values range between 0 and 1; models with values less than 0.5 perform worse than random, models with values of 0.5 equal random, and models with values more than 0.5 perform better than random. Typically, models with AUCs > 0.7 are deemed useful (Phillips 2005), and models with AUCs > 0.95 are deemed to perform exceptionally well. RMSE was determined using:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{x}_i - x_i\right)^2}{N}} \quad (1)$$

where $\hat{x}_i$ is the prediction of a given model, x_i are the observations, and N is the total number of observations. TSS was calculated as the maximum of the sum of the true positive rate (TPR) and true negative rate (TNR) minus 1:

$$\mathrm{TSS} = \max\left(\mathrm{TPR} + \mathrm{TNR} - 1\right) \quad (2)$$

Code to generate the simulations and the ERF routine is provided at https://osf.io/q9wfn/.

2.2. Simulation

2.2.1. Simulated covariates

Ten covariates were simulated using the RandomFields package (Schlather et al. 2020) in R over a 100 × 100 cell grid. Each simulated covariate, X_n, was assumed to be a multivariate normal process (MVN) (Eq. 3) with stationary isotropic Matérn covariance, Σ_{X_n}, based on the distance r between 2 points (Eq. 4) (following Thorson et al. 2017):

$$X_n \sim \mathrm{MVN}\left(0, \Sigma_{X_n}\right) \quad (3)$$

$$\Sigma_{X_n} = \Sigma(r) = \frac{\sigma_\Sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\kappa_{X_n}|r|\right)^\nu K_\nu\left(\kappa_{X_n}|r|\right) \quad (4)$$

$$\kappa_{X_n} = \frac{\sqrt{2\nu}}{l} \quad (5)$$

$$l \sim U(10, 30) \quad (6)$$

where σ_Σ^2 controls the pointwise variance, κ_{X_n} controls the geostatistical range of correlations (Eq. 5), ν controls the smoothness of the covariance matrix (we assumed $\nu = 1$), and K_ν is the Bessel function. Variance parameters, σ_Σ^2, were fixed at 0.0001 and scale parameters, l, were drawn from a uniform distribution (Eq. 6). Each scale parameter draw was used to simulate the Matérn covariance function over the 100 × 100 grid. We assumed all simulated covariates were moderately correlated and the correlations between covariates, ρ_{X_n,X_n}, were drawn from a Beta distribution (Eq. 7). The standard deviation of each covariate, σ_{X_n}, was determined from the variation of the simulated spatial field, X_n and used to convert correlation between covariates, ρ_{X_n,X_n}, to a covariance matrix (Eq. 8). The means of each covariate, μ_{X_n}, were drawn from a standard normal distribution (Eq. 9) and along with the covariance matrix, $\Sigma_{X,X}$, were used to draw coefficients for the interactions between environmental covariates from a multivariate normal distribution (Eq. 10). A probability of presence field, p_1, was calculated following a linear combination of environmental covariates, X_n (Eq. 11).

$$\rho_{X_n,X_n} \sim \mathrm{Beta}(2,2) \times 2 - 1 \quad (7)$$

$$\Sigma_{X,X} = \begin{bmatrix} \sigma^2_{X_n} & \cdots & \rho_{X_n,X_n}\sigma_{X_n}\sigma_{X_N} \\ \vdots & \cdots & \vdots \\ \rho_{X_N,X_n}\sigma_{X_N}\sigma_{X_n} & \cdots & \sigma^2_{X_N} \end{bmatrix} \quad (8)$$

$$\mu_{X_n} \sim N(0, 1) \quad (9)$$

$$\beta_{X_n,X_n} \sim \mathrm{MVN}\left(\mu_{X_n}, \Sigma_{X,X}\right) \quad (10)$$

$$\begin{aligned}\mathrm{logit}\left(p_1\right) = {} & \beta_n X_n + \ldots + \beta_N X_N + \beta_{X_n,X_n} X_n X_n \\ & + \ldots + \beta_{X_n,X_N} X_n X_N\end{aligned} \quad (11)$$

To generate point patterns from the probability of presence maps, a random cell was drawn and subsequent detections in that cell were assumed to have process and observation components (Royle & Kéry 2007) (Eqs. 12 & 13, respectively).

$$\hat{y}_n \sim \mathrm{Bernoulli}\left(p\right) \quad (12)$$

$$y_n \mid \hat{y}_n \sim \mathrm{Bernoulli}\left(\delta \times p \times \hat{y}_n\right) \quad (13)$$

where $\hat{y}_n$ is the true presence, y_n is the observed presence, δ is the background detection probability, and $\delta \times p$ is the effective detection probability, always $\leq \delta$.

2.2.2. Titration simulation

Point patterns with 50 000 samples were generated from p_1 (Eqs. 12 & 13) across a suite of detection probabilities ranging from ultra-rare events to an equal number of presences and absences (0.001,

0.0025, 0.005, 0.01, 0.025, 0.05, 0.1, 0.25, 0.5). For each detection probability, 25 point patterns were generated (resulting in 225 point patterns) and 10% of the full dataset was held out from all models for external validation. The standard RF, an RF with down-sampling (RF-DS), an RF with the synthetic minority over-sampling technique (RF-SMOTE), and an ERF were fit to each point pattern. The standard RF was implemented with a stratified random sample (90% training, 10% test) from the full dataset (minus the external validation set) and no internal resampling. The RF-DS was implemented with the same stratified random sampling from the full dataset and internal down-sampling (i.e. balancing the subset passed to each decision tree). The RF-SMOTE was implemented in the same manner as the RF-DS but after generating pseudo-replicates of the minority class using a single nearest neighbor following Stock et al. (2020). The number of pseudo-replicates generated was equal to the minority class size. The ERF model was implemented with 100 forests in the ensemble and each RF was implemented in the same manner as the RF-DS approach. For each model and detection probability combination, model performance metrics (i.e. AUC, RMSE, and TSS) were calculated using the external holdout dataset as well as the whole dataset to evaluate changes in model performance as a function of the degree of class imbalance. For the ERF model, performance metrics were assessed on the ensemble predictions (i.e. the mean across RF predictions within the ensemble). Uncertainty in performance metrics for each model and detection probability combination was assessed across the 25 point pattern samples by calculating the 95% confidence intervals..

2.2.3. Spatial contrast simulation

Two probability of presence fields, p_2 and p_3, were calculated from p_1 by rescaling to shrink probabilities towards 0.5 with ranges 0.25–0.75 and 0.45–0.55, respectively (see Fig. S1 in the Supplement at www.int-res.com/articles/suppl/n043p183_supp.pdf). Rescaling and shrinking the range of the probability of presence decreased the degree of clustering (Fig. S1, right panel) while preserving the relative autocorrelation in the environment covariates (Fig. S1, left panel). We assumed the detection probability for the point pattern process, δ, was 2% or a 50:1 chance of detection (i.e. a semi-rare event). For each probability of presence map, a point pattern was generated using this detection probability (Eqs. 12 & 13), 10% of the presences and absences were held out completely from any model fitting, and an ERF was fit to the remaining 90% of the point pattern data from which 90% train/test sets (90/10%) were generated for each RF in the ensemble. Each RF in the ensemble made predictions on the individual test sets, the overall test set, the total point pattern, and the gridded covariates used to generate the probability of presence maps, p_x. Model performance metrics (i.e. AUC, RMSE, and TSS) were calculated to evaluate changes in model performance as a function of the degree of spatial clustering of detections for the test sets generated for each RF, the held out test set that no RFs in the ensemble trained on, and the overall point pattern. Gridded residuals for each individual RF in the ensemble were calculated as the difference between individual RF predictions over the grid and the ensemble predictions (mean across RFs). The gridded coefficient of variation was calculated by determining the standard deviation of the individual RF predictions divided by the ensemble prediction in a grid cell.

2.3. Case studies

2.3.1. Fisheries-dependent data

Fisheries-dependent data from the Hawaii-based deep-set longline fishery targeting bigeye tuna *Thunnus obesus* were used to demonstrate the functionality of ERF. Data for 2005–2017 were provided by the National Oceanographic and Atmospheric Administration (NOAA) Fisheries' National Observer Program (NOP). The NOP conducts independent, at-sea data collection for commercial fishing and processing vessels (Allen & Gough 2007). During fishing trips, observers record information about catch by species, the location of fishing effort, and the rigging of fishing gear. Observers currently accompany a target of 20% of deep-set trips, which are selected quasi-randomly. Of the 2 pelagic longline fisheries in Hawaii, the deep-set fishery comprises approximately 96–99% of the total trips (NMFS 2017). Typically, the deep-set fishery sets around ~230 m deep and fishes an area roughly 10–35° N and 180–135° W (Bigelow et al. 2006). In 2017, a 10 yr peak was reached in effort, with 145 active vessels making 1539 trips and setting 19 647 longlines, roughly 13 sets per trip (WPRFMC 2018).

From the NOP-provided data, the beginning and end coordinates where the longline was set as well as the beginning and end coordinates where the longline was hauled were extracted. These 4 co-

ordinates were converted into a polygon for each set using the sp package (Pebesma & Bivand 2005) in R and the average across the polygon was extracted for each covariate. All NOP sets from 2005–2017 were used and datasets were generated for wahoo, giant manta ray, scalloped hammerhead, and false killer whale by dummy coding all sets without a given species as absences and sets with a given species as presences. Multiple individuals in a set were ignored, such that any set with at least one individual of a given species was treated as a presence record.

2.3.2. Covariates

To estimate bycatch distribution models from the NOP-provided data, we extracted environmental covariates at multiple spatial and temporal scales (Wiens 1989) (Table S1). Additionally, the RF algorithm can provide metrics of variable importance and we were interested in determining the variability of the variable importance rank across the ensemble of RFs. The environmental covariates we extracted can be largely split into 4 groups: (1) static (e.g. bathymetry, distance to shore, and distance to seamounts); (2) time dynamic (i.e. lunar phase and El Niño Southern Oscillation); (3) surficial spatiotemporally dynamic (e.g. sea surface temperature [SST], chlorophyll *a* [chl *a*], sea level anomaly, and blended sea winds); and (4) subsurface spatiotemporally dynamic (e.g. mixed layer depth, temperature at the mixing layer, and depth to the dissolved oxygen [DO] minimum). Gear covariates (e.g. set length, float length, number of floats, and bait type) were not included as model covariates, nor were set locations. These covariates were omitted to focus on the environmental drivers of bycatch interactions. Most covariates were used directly in the model; however, a subset of covariates required additional processing or derivation from existing covariates. Both SST and chl *a* were extracted as level-3 data products, meaning that cloud interference resulted in occasional data gaps. Fixed rank kriging implemented using the FRK package (Zammit-Mangion 2018) in R was used to interpolate the data layer and create a cloud-free product. Frontal structures and the gradient across identified structures were calculated from cloud-free products using the grec package (Lau-Medrano 2018) in R. Geostrophic current and wind divergence and vorticity were calculated based on the N–S and E–W current speed of adjacent cells based on the equations shown in Fig. S2.

2.3.3. Titration case study

As a commonly caught species, wahoo were an ideal case study to titrate rare event bias in the dataset and monitor the effects of rare events on model performance and uncertainty. Using the wahoo dataset from the fisheries-dependent data and environmental covariates, we derived rare event sub-datasets. For each sub-dataset, a given number of presences were drawn randomly from the original dataset and combined with the full set of absences. Presence sample sizes were 15, 25, 35, 50, 75, 100, 150, 250, 500, 100, 2500, 5000, and 10 000. Each rare event dataset was partitioned into 90/10% training/testing subsets and used to model the distribution of wahoo bycatch using ERF. All models were implemented in the same manner as outlined above (see Section 2.1.2), with 20 RFs comprising each ensemble to reduce the computational overhead. Model performance metrics (i.e. AUC, RMSE, and TSS) were assessed on the full dataset while the standard deviation in AUC (a measure of inter-model variability) and the standard deviation in variable importance were assessed for individual RFs in the ensemble. For each possible presence sample size, 10 random draws were conducted and the results were averaged to reduce the effect of noise from a single random draw.

2.3.4. Spatial contrast case studies

From the NOP-provided dataset, datasets were constructed for giant manta ray, scalloped hammerhead, and false killer whale presence and absence. The RF, RF-DS, RF-SMOTE, and ERF models were fit to each dataset. Each ERF model had 100 RFs implemented as outlined in Section 2.1.2. As so few presences occurred, a new stratified random draw was made from the full dataset for each species and model performance metrics (i.e. AUC, RMSE, and TSS) were evaluated for this new dataset as well as for the internal test datasets for all models. Probability of presence maps were generated for each species by making predictions on the full species' dataset and then averaging predictions over a hexagonal grid with cells roughly 100 km^2 in area. To protect the confidentiality of fishing locations, hexagonal grid cells with fewer than 3 vessels fishing the cell (279 of 3018 cells) were removed. Similarly, the presence points of each species were visualized by highlighting grid cells that contained presences rather than individual set locations. Rip-

ley's K functions, a measure of spatial clustering, were calculated from the probability of presence maps using the spatstat package (Baddeley et al. 2015) in R.

3. RESULTS

3.1. Simulation

3.1.1. Titration simulation

For each of the point patterns resulting from different detection probabilities, the spatial autocorrelation in the point pattern presences was similar (Fig. S3), mitigating the spatial autocorrelation effect on model performance. The ERF approach did as well or slightly better than RF-DS and RF-SMOTE approaches for AUC, RMSE, and TSS performance metrics across the range of detection probabilities for the external holdout dataset (Fig. 2A–C) and the full dataset (Fig. 2D–F). The RF approach outperformed all other approaches for the RMSE performance metric for all detection probabilities that resulted in class imbalance ($\delta \geq 0.25$) (Fig. 2B,E). This is a result of the RF approach learning the absences better than the other tested models as a function of the declining number of presences, as exemplified by the co-occurring decline in the TSS statistic for the RF approach. By random chance, the RF-DS and RF-SMOTE approaches perform on par with ERF for ultra-rare events ($\delta = 0.001$ or 1000:1 absences: presence), but on average ERF perform better according to AUC and TSS performance metrics (Fig. 2A,C,D,F). In general, uncertainty generated from different point pattern samples declined for the RF-DS, RF-SMOTE, and ERF approaches as class imbalance decreased. The same held for the RF approach, with the excep-

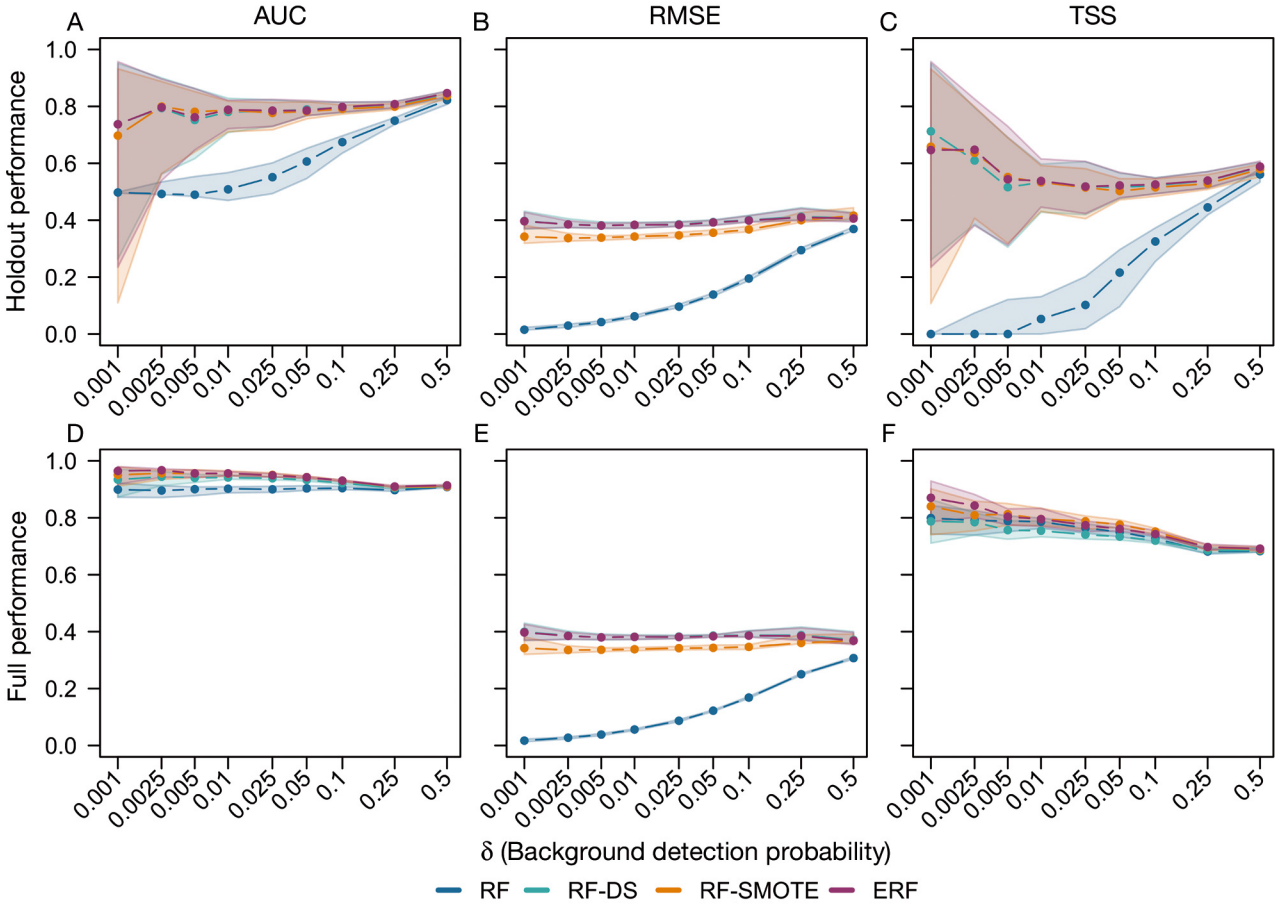

Fig. 2. Performance metrics for (A–C) the external holdout dataset and (D–F) the full dataset. (A,D) area under the curve (AUC), (B,E) root mean square error (RMSE), and (C,F) true skill statistic (TSS) for the standard Random Forest (RF), the Random Forest with down-sampling (RF-DS), the Random Forest with synthetic minority over-sampling technique (RF-SMOTE), and Ensemble Random Forests (ERF) across a range of detection probabilities, δ. The shaded region for each model indicates the 95% confidence interval calculated from 25 iterations of the simulation

tion of detection probabilities <0.01, where uncertainty decreased as the RF approach ignored the few presences in the datasets during training.

3.1.2. Spatial contrast simulation

The random fields simulation of spatial clustering (Fig. 3) yielded 10 simulated environmental covariates (Fig. 3A–J) and 3 probability of presence maps (Fig. 3K–M) with similar spatial autocorrelation (Moran's I between 0.26 and 0.27; Fig. S1), declining spatial clustering from p_1 to p_3 (Fig. S1, right panel), and declining contrast in the probability of presence (Figs. 3K–M & S1, left panel). Residuals between individual RFs (Fig. 3N–P) and the true probability of presence maps (Fig. 3K–M) highlight the variability in learning the majority and minority classes between individual RFs. By treating p_1 as the true probability of presence pattern and comparing p_1 to the ERF predictions (Fig. 3Q–S), the decline in model performance can be visually assessed. The ERF predictions (Fig. 3Q–S) resulted in the test AUC declining from p_1 to p_3 and as a function of declining spatial clustering (test AUC ± SD: 0.74 ± 0.03, 0.64 ± 0.04, and 0.52 ± 0.04, respectively). Performance on the held out test set was similar for AUC (overall test AUC: p_1 = 0.72, p_2 = 0.70, and p_3 = 0.51), RMSE (overall test RMSE: p_1 = 0.38, p_2 = 0.41, and p_3 = 0.42), and TSS (overall test TSS: p_1 = 0.37, p_2 = 0.25, and p_3 = 0.02). However,

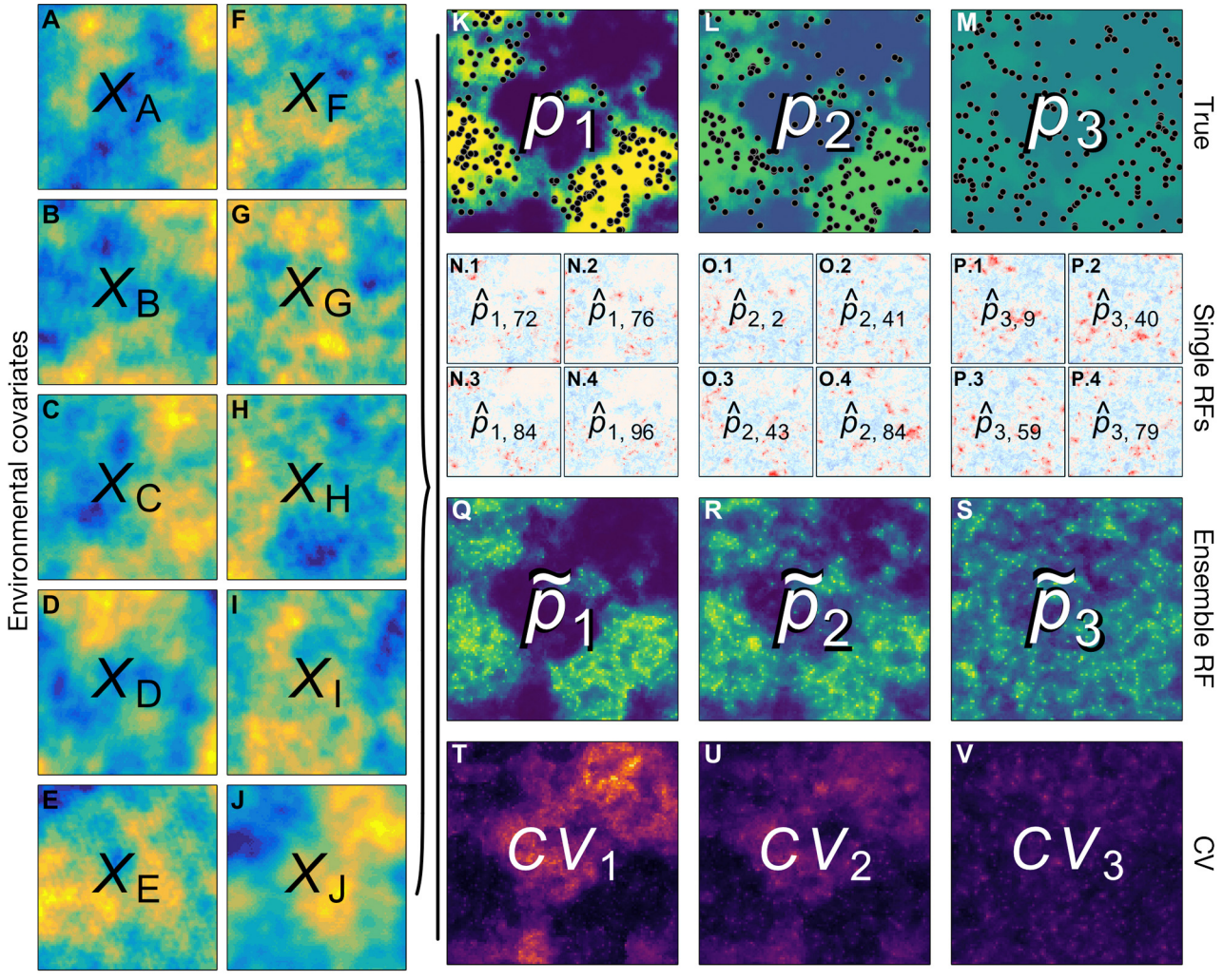

Fig. 3. (A–J) Simulated environmental covariates, X_n, and (K–M) the corresponding 3 simulated probability of presence maps, $p_{1...3}$, using a linear model of the covariates and fixed parameters. The probabilities from p_1 were shrunk to reduce contrast and produce the probabilities for p_2 and p_3 (range 0.25–0.75 and 0.45–0.55, respectively). (N–P) Top 4 deviating Random Forests within the ensemble ($\hat{p}_{n,x}$), colored by positive (red) or negative deviations (blue), (Q–S) predictions of the ensemble ($\tilde{p}_n$), and (T–V) the coefficient of variation of the ensemble predictions (CV_n) are shown. Warmer colors indicate higher values

the ensemble AUC increased as a function of declining spatial clustering (ensemble AUC: $p_1 = 0.965$, $p_2 = 0.971$, $p_3 = 0.981$). The TSS (ensemble TSS: 0.817, 0.838, 0.907, respectively) and RMSE (ensemble RMSE: 0.387, 0.403, 0.413, respectively) also increased from p_1 to p_3, though the latter indicates poorer model performance as spatial clustering declines. Comparing the ERF predictions (Fig. 3Q–S) to the true probability of presence pattern (Fig. 3K–M), the ability of the ensemble to smooth across the individual RF predictions (Fig. 3N–P) to generate a central tendency can be observed. As spatial clustering declines, the ERF predictions gradually fill in areas of low probability of presence. Despite this, the ERF still delineates areas of high and low probability of presence and inflates the difference between these areas (as shown by comparing Fig. 3L & Fig. 3R). Differences in the test AUC do not equate to one probability of presence map being more 'true' than any other. Rather, declining AUC corresponds to the model's ability to discriminate between the true/false positives and negatives. As spatial clustering declines, the ability of an individual RF to identify patterns within the data that match those in the environmental covariates can be expected to decline.

3.2. Case studies

3.2.1. Titration case study

Titrating rare event bias (decreasing number of presences) into the NOP-provided deep-set longline data for wahoo (21 279 presences: 26 076 absences) resulted in declining model performance in terms of raw metrics and increased metric uncertainty across RFs in the ensemble (Fig. 4). The median AUC derived from ROCs decreased as a function of increasing rare event bias (Fig. 4A). All presence sample sizes resulted in models with AUC values greater than 0.5 (better than random). RMSE was similar from 15 to 2500 presences and decreased as presences increased over 5000 (Fig. 4B). The median TSS decreased as a function of increasing rare event bias (Fig. 4C). The top 5 covariates (in order) were day of year, distance to current front, distance to chl *a* front, distance to SST front, and current divergence. The standard deviation in the variable importance metric increased exponentially with increasing rare event bias (Fig. 4D). All covariates had similar uncertainty in variable importance.

3.2.2. Spatial contrast case studies

Available presences for giant manta ray ($n = 24$) and scalloped hammerhead ($n = 23$) were ultra-rare (~1:2000 presence:absence) while false killer whale presences ($n = 62$) were considered rare (~1:750 presence:absence) in the 47 355 observed longline sets. The presence–absence point patterns of these 3 case study species (represented in Fig. 5A–C) resulted in contrasting ERF predictions and, as expected, different areas of high probability of presence. Giant manta ray and scalloped hammerhead had a high probability of presence concentrated south of the main Hawaiian Islands (Fig. 5A,B). The probability of presence map for giant manta ray also had an area of high probability southwest of the main Hawaiian islands near 12° N, 165° W (Fig. 5A). False killer whale probability of presence was dispersed but an area north–northeast of the main Hawaiian Islands had higher probability than the surrounding areas (Fig. 5C). AUC declined from giant manta ray (0.78 ± 0.09, $\mu \pm \sigma$), to scalloped hammerhead (0.75 ± 0.11), to false killer whale (0.55 ± 0.09) (Fig. 5D–F). External model performance test metrics (Table 1) and internal test metrics (Table S2) showed ERFs performed better than the other models, with RF-SMOTE a close second, followed by RF-DS, and then RF. This follows the simulation spatial clustering results, as test AUC for each RF in the ensemble declined as a function of spatial clustering (Fig. 5G). However, the ensemble AUC was high across all 3 species, with false killer whale having the highest (0.999), followed by scalloped hammerhead (0.918), and then giant manta ray (0.916). Spatial clustering of giant manta ray and scalloped hammerhead was similar and greater than spatial clustering of false killer whale, while all 3 species had greater clustering than expected by chance (Fig. 5G). RMSE was highest for false killer whale (0.401), then giant manta ray (0.384), and lowest for scalloped hammerhead (0.374). Similarly to AUC, TSS was high for all species but highest for false killer whale (0.999), then scalloped hammerhead (0.996), and then giant manta ray (0.987).

4. DISCUSSION

Ensemble Random Forests (ERFs) is an intuitive method for reducing rare event bias by generating stratified randomly sampled training/test sets and training multiple RFs (Breiman 2001a,b, Cutler et al. 2007) with down-sampling (Kuhn & Johnson 2013) to generate an ensemble of 'strong learners'. Here, the

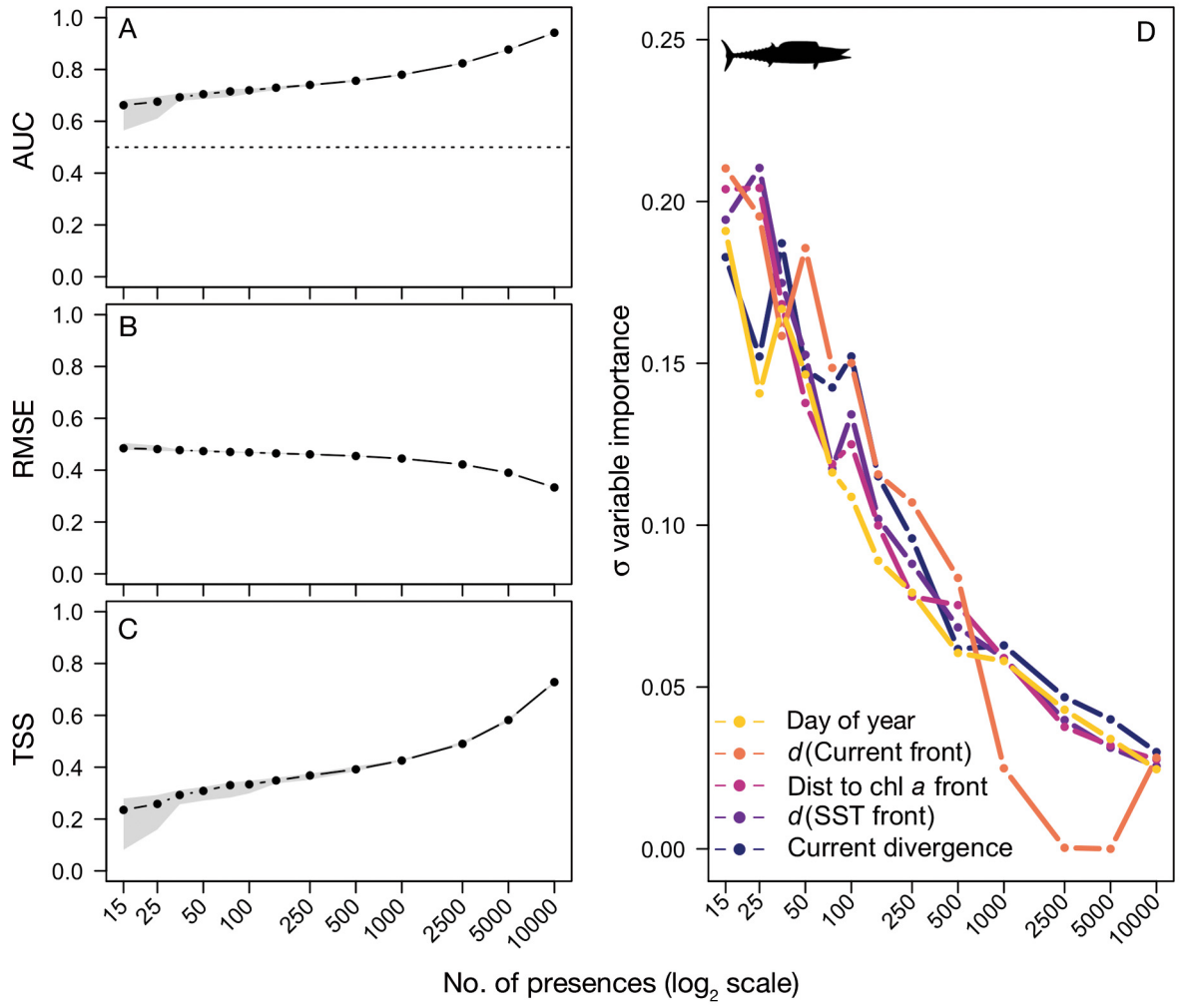

Fig. 4. For each of the wahoo case study titration sub-models (see Section 2.3.3), the median (the line) and 95% confidence interval (the gray shading) of (A) the area under the curve (AUC) model performance metric calculated using individual receiver operator characteristic curves, (B) the root mean square error (RMSE), and (C) the true skill statistic (TSS) as a function of the number of presences. (D) Standard deviation of the variable importance metric (mean decrease in accuracy) for the 5 most important covariates (determined across all titration submodels), with warmer colors indicating higher average importance

performance of ERF as a function of rare event bias is demonstrated through simulation and case studies. In simulations, ERF performed better on average than 3 alternative RF approaches for rare events across highly class imbalanced to balanced datasets. However, the performance of individual RFs in the ensemble is related to the strength of spatial clustering in a given dataset. In the case studies, class imbalance impacted the overall performance of the ensemble, while contrast in the presence–absence point patterns (Evans et al. 2011) negatively impacted individual RF models but did not greatly impact the ensemble performance, similar to the simulations. Nonetheless, the ERF routine performs admirably even for very low sample sizes of presences given enough absences.

In the titration simulation, ERF performed, on average, better in regards to full performance than the RF with a stratified training/test set, RF with down-sampling, and RF with synthetic minority oversampling technique for AUC and TSS. On average, ERF, RF-DS, and RF-SMOTE converged as class imbalance increased for all holdout performance metrics. Not surprisingly, the base RF approach performed the worst with respect to AUC and TSS and the best for RMSE. This performance reflects the influence of successive partitioning on the growth of decision trees, where fewer presences are re-

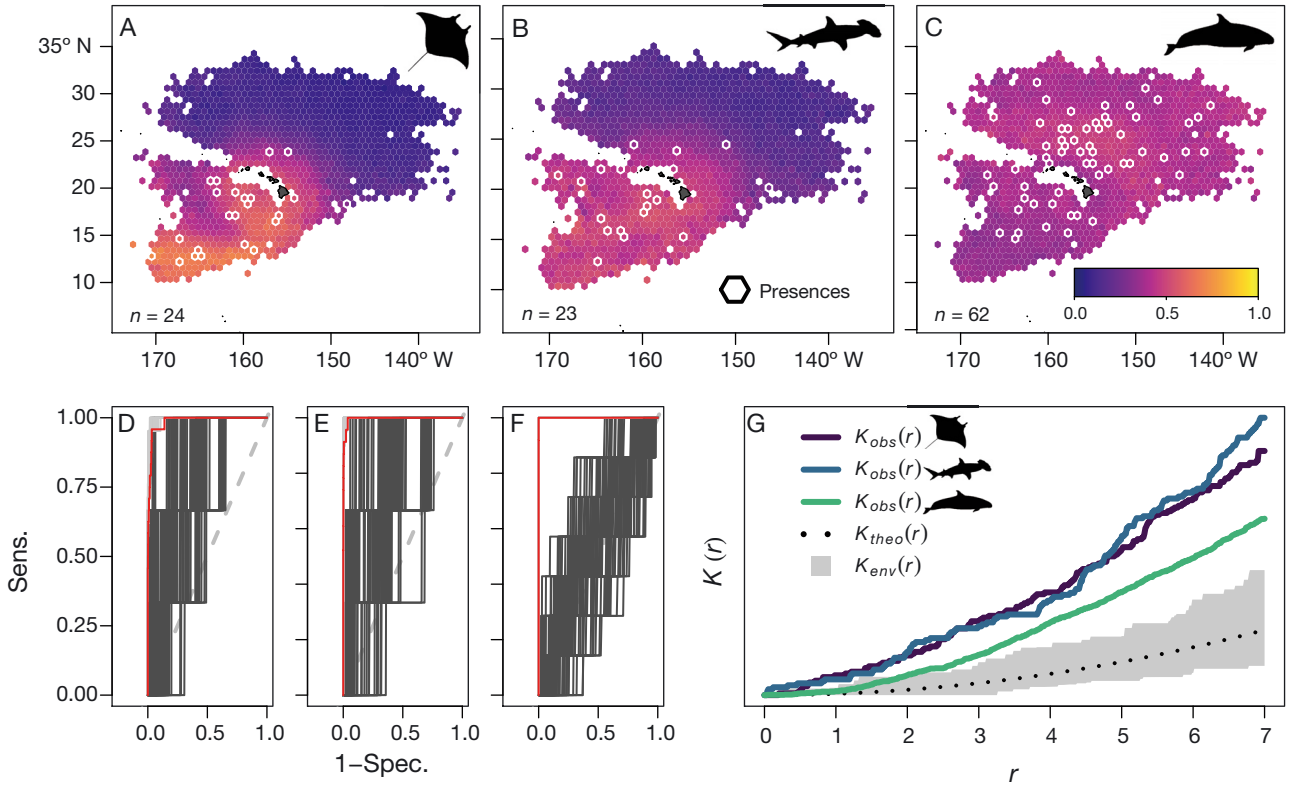

Fig. 5. Ensemble Random Forests spatial predictions in the central north Pacific for (A) giant manta ray, (B) scalloped hammerhead, and (C) false killer whale, with warmer colors indicating a higher probability of presence. The receiver operator characteristic curves for the training set (light gray lines), the test set (dark gray lines), and the ensemble predictions (red line) against the random expectation (dashed 1:1 line) for (D) giant manta ray, (E) scalloped hammerhead, and (F) false killer whale. (G) Ripley's K functions for the theoretical expected clustering (K_{theo}, dotted line) and theoretical envelope (K_{env}, gray polygon), and for the case studies (K_{obs}): giant manta ray (purple line), scalloped hammerhead (blue line), and false killer whale (teal line)

Table 1. Model performance metrics, area under the curve (AUC), root mean squared error (RMSE), and true skill statistic (TSS), measured on the stratified random draw dataset for the spatial covariation case study species using Random Forest (RF), RF with downsampling (RF-DS), RF with synthetic minority over-sampling technique (RF-SMOTE), and Ensemble Random Forests (ERF)

Performance metric	AUC	RMSE	TSS
Giant manta ray			
RF	0.61	0.03	0.28
RF-DS	0.91	0.39	0.81
RF-SMOTE	0.95	0.33	0.91
ERF	1.00	0.38	0.99
Scalloped hammerhead			
RF	0.63	0.03	0.33
RF-DS	0.96	0.38	0.91
RF-SMOTE	0.94	0.31	0.81
ERF	1.00	0.37	1.00
False killer whale			
RF	0.55	0.04	0.14
RF-DS	0.61	0.41	0.31
RF-SMOTE	0.96	0.33	0.77
ERF	1.00	0.40	1.00

tained as trees grow, resulting in the model training predominantly on absences (He & Garcia 2009). Effects from successive partitioning are most noticeable when $\delta \leq 0.005$, where the effects become so severe that multiple RFs trained on independent point pattern samples all perform similarly. Interestingly, the RF-DS and RF-SMOTE methods can perform on par or, in the case of RF-SMOTE, better than ERF by random chance. This is similar to the conclusion of Kuhn & Johnson (2013) that the performance of different approaches to mitigating rare event bias can differ between datasets and there is little consensus on a 'best' approach. However, ERF was more consistent in the AUC and TSS performance metrics than the RF-DS and RF-SMOTE approaches across independent point pattern samples over the whole range of class imbalances. Thus, ERF may be an option to mitigate rare event bias regardless of the dataset.

Across a range of spatial clustering in presences, ERF overall performance actually increased for AUC and TSS as spatial clustering declined. However, the

test performance of individual RFs within the ensemble and the overall test performance of the ERF declined. This indicates that ERF tends to overfit as a function of declining spatial clustering (the difference between the overall performance metric and the test performance metric increases). Poor test performance is not unanticipated, as there is little to distinguish an absence from a presence in regards to the true probability of presence in the p_3 probability of presence maps and it would be surprising if models performed well in these circumstances. Thus, caution should be exercised in interpreting performance metrics generated for the ensemble predictions, as they tend to inflate the ability of model to predict to test sets.

In the titration of rare event bias into the wahoo dataset, models performed better than random according to AUC for all presence sample sizes. Uncertainty in performance metrics increased as sample sizes fell below 100 presences (250:1 absences:presences). As expected, as presence sample sizes declined, the uncertainty in variable importance increased. This should advise users of ERF, and realistically any model with rare event data, to interpret variable importance metrics with caution, as the ranking changes considerably between individual RFs in the ensemble. ERFs for giant manta ray and scalloped hammerhead had greater test model performance than ERFs for false killer whale, even with only 24 and 23 presences, respectively. This highlights the tradeoff between spatial covariation and sample size: fewer samples are needed for individual RFs to discern patterns as covariation increases. However, as demonstrated by the spatial contrast simulation results, the high performance of the ensemble predictions can be misleading and mean test performance metrics should be used in interpreting the performance of ERF in the case studies.

There are 4 distinct advantages of ERF for rare events: (1) no individual RF trains on the whole dataset but the ERF does 'see' the whole dataset, reducing the effects of successive partitioning (e.g. there is minimal data loss by setting test sets) (He & Garcia 2009); (2) uncertainty in the RF algorithm can be propagated forward; (3) a balanced weighting scheme allows for a better accounting of interacting covariates when growing the decision trees (Kuhn & Johnson 2013); and (4) the propagation of Type II error (absences predicted as presences) from rebalancing the dataset to a balanced weighting scheme is minimized in the ensemble as the majority of RFs will vote correctly on absences (He & Garcia 2009). These advantages allow further gains to be made from base RF implementation for small sample sizes. In the case of our applied example, the Hawaii-based longline deep-set fishery, 78% of species caught by the gear had at least a 10:1 absence to presence ratio, 62% had at least 100:1, and 44% had at least 1000:1. These bycatch occurrences are fairly typical in a specialized fishery with specific target species. Fisheries of this nature occur globally, drastically increasing the utility of ERF in directly analogous systems. As sample size increases, the performance from ERF converges towards the base RF method. Similarly, the uncertainty between RFs declines as sample size increases, reducing the necessity for incorporating inter-model variability. Overall, ERF is likely to be a useful tool for generating probability of presence maps (relative to the sampling regime) for any species, as it propagates uncertainty in model fit and includes interactions between covariates. However, the highest model gains are likely to occur with rarely occurring species.

There are disadvantages to ERF relative to other statistical applications. Linear models offer an explicit ease-of-interpretation that is lacking in 'black box' machine-learning approaches (Breiman 2001b, Olden et al. 2008). Relative to RF or the ERF implementation presented here, the commonly used maximum entropy species distribution models do have an advantage: total sample size (presences and absences) can be considerably smaller, as these models only rely on presence data points and a background sample (from the environmental covariates) (Phillips et al. 2006, Elith et al. 2011). Despite high model performance with ultra-rare presences, each of our case studies had the advantage of a plethora of absence records (presences + absences = 47 355 records). This difference does currently limit the application of ERF to datasets with known absences.

In the marine environment where industry-based and related sampling regimes are common, many conservation priority species have both presences and absences. In terrestrial systems, ERF is likely to be most applicable to modeling rare species with extremely few occurrences within a larger sampling scheme (e.g. quadrat-based sampling), due to its reliance on informative absences. RFs have been used to classify presence-only data and performed better or on par with MaxEnt (Williams et al. 2009). It stands to reason that ERF may perform similarly on presence-only datasets; however, testing is necessary to demonstrate such capabilities. Broadly, there is a pressing need to evaluate the performance of species distribution models using simulation and tools like

RandomFields, which can greatly ease the computational overhead in generating semi-realistic covariates. Future avenues for ERF include extension to presence-only datasets as well as management strategy evaluations on the inclusion of species distribution models in conservation planning.

Rare event bias is a difficult statistical property of many datasets, and is challenging to model appropriately, often relying on high sample sizes of rare events or conditioning on linear responses (He & Garcia 2009). For many species, rare occurrences prohibit the former and the use of proxy or indirect covariates often necessitates interacting covariates or non-linear responses. ERF is a useful tool to reduce biases resulting from dealing with the imbalanced data problem in machine-learning algorithms. Additionally, ERF reduces the reliance on a high number of positives in the data needed in GLMs, GAMs, or other mixture models, and incorporates interacting and non-linear covariate responses (even though we used linear effects for simplicity in our simulation). The application of ERF to rare event presence–absence data is a straightforward bagging of a down-sampled RF routine and mostly involves capturing the individual RF predictions and collating them to generate inter-model uncertainty and model averages. For the rarely occurring, ESA-listed species in the Hawaii-based deep-set longline dataset, application of ERF generated datum-specific and spatially averaged predictions of interactions with the fishery gear. These maps have direct utility in evaluating threats to species across their range and in the rebuilding plans mandated by the Endangered Species Act. Within the datasets for similar fisheries and other large surveys around the world, there are likely many species where the application of ERF could be informative in assessing their spatial distribution and determining the influence of the environment in driving ecological patterns.

Data availability. Due to the confidential nature of the NOAA NOP Hawaii-based deep-set longline fishery observer data, data for the case studies are solely available upon request from the NOAA NOP. All code for the simulation has been provided at https://osf.io/q9wfn/

Acknowledgements. We thank the Western Pacific Regional Fishery Management Council along with NOAA Fisheries Pacific Islands Fisheries Science Center and NOAA Pacific Island Regional Office for funding this work. We also thank Asuka Ishizaki for facilitating this collaboration between WPRFMC, NOAA PIFSC, and NOAA PIRO. This work was funded by NA15NMF4410008 and NA15NMF4410066.

LITERATURE CITED

Allen SD, Gough A (2007) Hawaii longline fishermen's experiences with the Observer Program. NOAA technical memorandum NMFS-PIFSC-8

Araújo MB, New M (2007) Ensemble forecasting of species distributions. Trends Ecol Evol 22:42–47

Austin M (2007) Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol Model 200:1–19

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. Chapman and Hall/CRC, London

Bigelow K, Musyl MK, Poisson F, Kleiber P (2006) Pelagic longline gear depth and shoaling. Fish Res 77:173–183

Breiman L (2001a) Random forests. Mach Learn 45:5–32

Breiman L (2001b) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16:199–231

Campbell RA (2015) Constructing stock abundance indices from catch and effort data: Some nuts and bolts. Fish Res 161:109–130

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random Forests for classification in ecology. Ecology 88:2783–2792

Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40:677–697

Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ (2011) A statistical explanation of MaxEnt for ecologists. Divers Distrib 17:43–57

Evans JS, Murphy MA, Holden ZA, Cushman SA (2011) Modeling species distribution and change using random forest. In: Drew CA, Wiersma YF, Huettmann F (eds) Predictive species and habitat modeling in landscape ecology: concepts and applications. Springer, New York, NY, p 139–159

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21:1263–1284

Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York, NY

Lau-Medrano W (2018) Grec: GRadient-Based RECognition of Spatial Patterns in Environmental Data. R package version 1.4.1. https://CRAN.R-project.org/package=grec

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2:18–22

Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2009) Evaluation of consensus methods in predictive species distribution modelling. Divers Distrib 15:59–69

Martin SL, Stohs SM, Moore JE (2015) Bayesian inference and assessment for rare-event bycatch in marine fisheries: a drift gillnet fishery case study. Ecol Appl 25: 416–429

Merow C, Smith MJ, Edwards TC, Guisan A and others (2014) What do we gain from simplicity versus complexity in species distribution models? Ecography 37:1267–1281

Millar CI, Stephenson NL, Stephens SL (2007) Climate change and forests of the future: managing in the face of uncertainty. Ecol Appl 17:2145–2151

NMFS (2017) The Hawaii limited access longline logbook summary report, January to December 2016. PIFSC Data

Report DR-17-009. NOAA Fisheries doi:10.7289/V5/DR-PIFSC-17-009

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83:171–193

Pebesma EJ, Bivand RS (2005) Classes and methods for spatial data in R. R News 5(2), https://cran.r-project.org/doc/Rnews/

Phillips SJ (2005) A brief tutorial on Maxent. AT&T Res 190:231–59

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Modell 190:231–259

Pielke RA, Schellnhuber HJ, Sahagian D (2003) Nonlinearities in the Earth system. Global Change Newsletter 55:11–15

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Robinson LM, Elith J, Hobday AJ, Pearson RG, Kendall BE, Possingham HP, Richardson AJ (2011) Pushing the limits in marine species distribution modelling: lessons from the land present challenges and opportunities. Glob Ecol Biogeogr 20:789–802

Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP (2012) An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J Photogramm Remote Sens 67: 93–104

Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D (2008) Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. Remote Sens Environ 112:2272–2283

Royle JA, Kéry M (2007) A Bayesian state-space formulation of dynamic occupancy models. Ecology 88:1813–1823

Schlather M, Malinowski A, Oesting M, Boecker D and others (2020) RandomFields: simulation and analysis of Random Fields. R package version 3.3.8, https://cran.r-project.org/package=RandomFields

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21:7881

Stock BC, Ward EJ, Eguchi T, Jannot JE, Thorson JT, Feist BE, Semmens BX (2020) Comparing predictions of fisheries bycatch using multiple spatiotemporal species distribution model frameworks. Can J Fish Aquat Sci 77:146–163

Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 8:25

Thorson JT, Jannot J, Somers K (2017) Using spatio-temporal models of population growth and movement to monitor overlap between human impacts and fish populations. J Appl Ecol 54:577–587

White WT, Corrigan S, Yang L, Henderson AC, Bazinet AL, Swofford DL, Naylor GJP (2018) Phylogeny of the manta and devilrays (Chondrichthyes: mobulidae), with an updated taxonomic arrangement for the family. Zool J Linn Soc 182:50–75

Wiens JA (1989) Spatial scaling in ecology. Funct Ecol 3: 385–397

Williams JN, Seo C, Thorne J, Nelson JK, Erwin S, O'Brien JM, Schwartz MW (2009) Using species distribution models to predict new occurrences for rare plants. Divers Distrib 15:565–576

WPRFMC (2018) Annual stock assessment and fishery evaluation report for U.S. Pacific island pelagic fishery ecosystem plan 2017. Western Pacific Regional Fishery Management Council, Honolulu, HI

Zammit-Mangion A (2018) FRK: Fixed Rank Kriging. R package version 0.2.2.1. https://CRAN.R-project.org/package=FRK

Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM (2009) Mixed effects models and extensions in ecology with R. Springer-Verlag, New York, NY

Editorial responsibility: Brendan Godley, University of Exeter, Cornwall Campus, UK

Submitted: October 9, 2019; Accepted: July 13, 2020
Proofs received from author(s): September 24, 2020