Why to choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence

Chunrong Mi [1,2] , **Falk Huettmann** [3] , **Yumin Guo** [Corresp., 1] , **Xuesong Han** [1] , **Lijia Wen** [1]

[1] College of Nature Conservation, Beijing Forestry University, Beijing, China

[2] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing, China

[3] EWHALE Lab, Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska Fairbanks (UAF), Fairbanks, Alaska, United States

Corresponding Author: Yumin Guo
Email address: guoyumin@bjfu.edu.cn

Species distribution models (SDMs) have become an essential tool in ecology, biogeography, evolution, and more recently, in conservation biology. How to generalize species distributions in large undersampled areas, especially with few samples, is a fundamental issue of SDMs. In order to explore this issue, we used the best available presence records for the Hooded Crane (*Grus monacha*, n=33), White-naped Crane (*Grus vipio*, n=40), and Black-necked Crane (*Grus nigricollis*, n=75) in China as three case studies, employing four powerful and commonly used machine learning algorithms to map the breeding distributions of the three species: TreeNet (Stochastic Gradient Boosting, Boosted Regression Tree Model), Random Forest, CART (Classification and Regression Tree) and Maxent (Maximum Entropy Models) Besides, we developed an ensemble forecast by averaging predicted probability of above four models results. Commonly-used model performance metrics (Area under ROC (AUC) and true skill statistic (TSS)) were employed to evaluate model accuracy. Latest satellite tracking data and compiled literature data were used as two independent testing datasets to confront model predictions. We found Random Forest demonstrated the best performance for the most assessment method, provided a better model fit to the testing data, and achieved better species range maps for each crane species in undersampled areas. Random Forest has been generally available for more than 20 years, and by now, has been known to perform extremely well in ecological predictions. However, while increasingly on the rise its potential is still widely underused in conservation, (spatial) ecological applications and for inference. Our results show that it informs ecological and biogeographical theories as well as being suitable for conservation applications, specifically when the study area is undersampled. This method helps to save model-selection time and effort, and it allows robust and rapid assessments

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

and decisions for efficient conservation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

Why to choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence

Chunrong Mi[1,2], Falk Huettmann[3], Yumin Guo[1], Xuesong Han[1] and Lijia Wen[1]

[1]College of Nature Conservation, Beijing Forestry University, P.O. Box 159, Beijing 100083, China

[2]Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing 100101, China

[3]EWHALE Lab, Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska Fairbanks (UAF), 419 Irving I, P.O. Box 757000, AK 99775, USA

Corresponding author:

Yumin Guo

College of Nature Conservation, Beijing Forestry University, P.O. Box 159, Beijing 100083, China

guoyumin@bjfu.edu.cn

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

ABSTRACT

Species distribution models (SDMs) have become an essential tool in ecology, biogeography, evolution, and more recently, in conservation biology. How to generalize species distributions in large undersampled areas, especially with few samples, is a fundamental issue of SDMs. In order to explore this issue, we used the best available presence records for the Hooded Crane (*Grus monacha*, n=33), White-naped Crane (*Grus vipio*, n=40), and Black-necked Crane (*Grus nigricollis*, n=75) in China as three case studies, employing four powerful and commonly used machine learning algorithms to map the breeding distributions of the three species: TreeNet (Stochastic Gradient Boosting, Boosted Regression Tree Model), Random Forest, CART (Classification and Regression Tree) and Maxent (Maximum Entropy Models) Besides, we developed an ensemble forecast by averaging predicted probability of above four models results. Commonly-used model performance metrics (Area under ROC (AUC) and true skill statistic (TSS)) were employed to evaluate model accuracy. Latest satellite tracking data and compiled literature data were used as two independent testing datasets to confront model predictions. We found Random Forest demonstrated the best performance for the most assessment method, provided a better model fit to the testing data, and achieved better species range maps for each crane species in undersampled areas. Random Forest has been generally available for more than 20 years, and by now, has been known to perform extremely well in ecological predictions. However, while increasingly on the rise its potential is still widely underused in conservation, (spatial) ecological applications and for inference. Our results show that it informs ecological and biogeographical theories as well as being suitable for conservation applications, specifically when the study area is undersampled. This method helps to save model-selection time and effort, and it

allows robust and rapid assessments and decisions for efficient conservation.

Keywords*:* Species distribution models (SDMs), Random Forest, Generality (transferability), Rare species, Undersampled areas, Hooded Crane (*Grus monacha*), White-naped Crane (*Grus vipio*), Black-necked Crane (*Grus nigricollis*)

INTRODUCTION

Species distribution models (SDMs) are empirical ecological models that relate species observations to environmental predictors (Guisan & Zimmermann, 2000, Drew et al., 2011). SDMs have become an increasingly important and now essential tool in ecology, biogeography, evolution and, more recently, in conservation biology (Guisan et al., 2013), management (Cushman & Huettmann, 2010), impact assessments (Humphries & Huettmann, 2014) and climate change research (Lei et al., 2011). To generalize and infer from a model, or model transferability is defined as geographical or temporal cross-applicability of models (Thomas & Bovee 1993; Kleyer 2002; Randin et al., 2006). It is one important feature in SDMs, a base-requirement in several ecological and conservation biological applications (Heikkinen et al., 2012). In this study, we used generality (transferability) as the concept of generalizing distribution from sampled areas to unsampled areas (extrapolation beyond the data) in one study area.

Detailed distribution data for rare species in large areas are rarely available in SDMs (Pearson et al., 2007; Booms et al., 2010). However, they are the most needed for their conservation to be effective. Collecting and assembling distribution data for species, especially for rare or endangered species in remote wilderness areas is often a very difficult task, requiring a large amount of human, time and funding source (Gwena et al., 2010; Ohse et al., 2009).

Recent studies have suggested that machine-learning (ML) methodology, may perform better than the traditional regression-based algorithms (Elith et al., 2006). TreeNet (boosting; Friedman

2002), Random Forest (bagging; Breiman, 2001), CART (Breiman et al., 1984) and Maxent (Phillips et al., 2004) are considered to be among the most powerful machine learning algorithms and for common usages (Elith et al., 2006; Wisz et al., 2008; Williams et al., 2009; Lei et al., 2011) and for obtaining powerful ensemble models (Araújo and New 2007; Hardy et al., 2011). Although Heikkinen et al. (2012) compared the four SDMs techniques' transferability in their study, they did not test with rare species and few samples in undersampled areas. It is important to understand that the software platform of the former three algorithms (Boosted Regression Trees, Random Forest and CARTs) applied by Heikkinen et al. (2012) from the R software ("BIOMOD" framewok) comes without a GUI and lacks sophisticated optimization and fine-tuning, but as they are commonly used though by numerous SDM modelers. Instead, we here run these models in the Salford Predictive Modeler (SPM) by Salford Systems Ltd. These algorithms in SPM are further optimized and improved by one of the algorithm's original co-authors (especially for TreeNet and Random Forest). It runs with a convenient GUI, and produces a number of descriptive results and graphics which are not available in the R version. While this is a commercial software, it is usually available on a 30 days trial version (which suffices for most model runs we know. As well, some of the features of the randomForest R package, most notably the ability to produce partial dependence plots (Herrick 2013), are not directly implemented yet in SPM7 (but they can essentially be obtained by running TreeNet in a Random Forest model).

Model generality (transferability) testing could offer particularly powerful for model evaluation (Randin et al., 2006). Independent observations from training data set has been recommended as a more proper evaluations of models (Fielding & Bell 1997; Guisan and Zimmermann 2000). So the use of an independent geographically (Fielding & Haworth, 1995) or temporally (Boyce et al., 2002; Araujo et al., 2005b) testing data set is encouraged to assess the generality of different SDMs techniques. Data from museum specimen, published literature (Graham et al., 2004) as well as tracking are good source to assess model generality (transferability) performance. In addition, how the distribution map links with reality data, especially in undersampled areas where modelers want to make predictions should definitely be as a metric to assess model performance and generalization. Arguably, if model predictions perform very well there, great progress is provided. Whereas, predictions on existing knowledge and data offers less progress. The model prediction and conservation frontier obviously sits in the unknown.

In this study, we modeled the best-available data for three species in East Asia as test cases: Hooded Cranes (*Grus monacha*, n=33), White-naped Cranes (*Grus vipio*, n=40) and Black-necked Cranes (*Grus nigricollis*, n=75). Four machine-learning models (TreeNet, Random Forest, CART and Maxent) were applied to map breeding distributions for these three crane species which otherwise lack empirically derived distribution information. In addition, two kinds of independent testing data sets (latest satellite tracking data, and compiled literature data (Threatened Birds of Asia: Collar *et al.,* 2001) were obtained to test the transferability of the four model algorithms. The purpose of this investigation is to explore whether there is a SDM technique among the four algorithms that could generate reliable and accurate distributions with high generality for rare species using few samples but in large undersampled areas? Results from this research could be useful for the detection of rare species and enhance fieldwork sampling in large undersampled areas which would save money and effort, as well as the conservation management of those species.

MATERIALS AND METHODS

Species data

In our 13 combined years of field work, we have collected 33 Hooded Crane nests (2002-2014), 40 White-naped Crane nests (2009-2014)， and 75 Black-necked Crane nests (2014) (see Fig. 1), during breeding seasons. We used these field samples (nests) to represent species presence points referenced in time and space.

Put Fig. 1 here

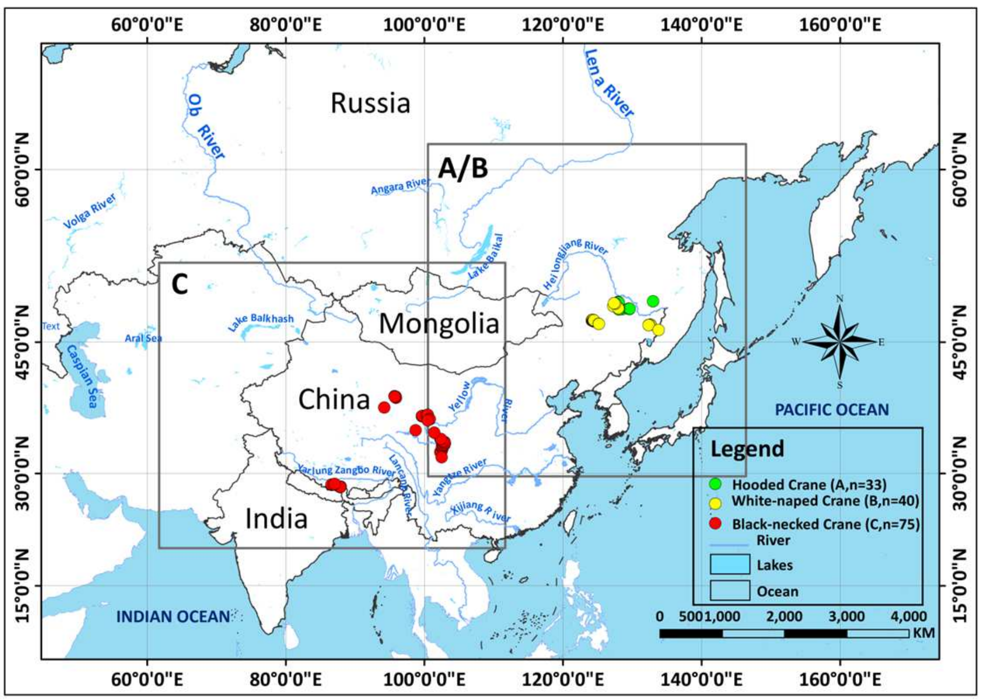


Figure 1 Study areas for three species cranes.

Environmental variables

We used 21 environmental layers at a 30-s resolution in GIS format and that were known to correlate with bird distribution and as proxies of habitats predictors. They included bio-climatic factors (bio_1-7, bio_12-15), topographical factors (altitude, slope, and aspect), water factors (distance to river, distance to lake, and distance to coastline), inference factors (distance to road, distance to rail road, and distance to settlements), and land cover factors (for detailed information, see Table 1). Most of these factors were obtained from open access sources. Bio-climate factors were obtained from the WorldClim database, while aspect and slope layer were derived from the altitude layer in ArcGIS, which was also initially obtained from the WorldClim database. Road, railroad, river, lake and coastline and settlement maps were obtained from the Natural Earth database. The land cover map was obtained from the ESA database. We also made models with all 19 bio-climate variables and 10 other environmental variables, and then reduced predictors by AIC, BIC, varclust, PCA and FA analysis. When we compared the distribution maps overlaying with independent data set generated by Random Forest model, we found the model based on 21 predictors have the best performance for Hooded Cranes, and the best level for White-naped Crane and Black-necked Cranes (see Supplement S1). Therefore, we decided to constructed models with 21 predictors for the all three cranes and four machine-learning techniques. All spatial layers of these environmental variables were resampled to a resolution of 30-s to correspond to that of the bioclimatic variables and for a meaningful high-resolution management scale.

Put Table 1 here

Table 1 Environmental GIS layers used to predict breeding distributions of three cranes.

Environmental Layers	Description	Source	Website
Bio_1	Annual mean Temperature (℃)	WorldClim	http://www.worldclim.org/
Bio_2	Monthly mean (max temp - min temp) (℃)	WorldClim	http://www.worldclim.org/

Bio_3	Isothermality (BIO2/BIO7) (*100℃)	WorldClim	http://www.worldclim.org/
Bio_4	Temperature seasonality (standard deviation *100℃)	WorldClim	http://www.worldclim.org/
Bio_5	Max temperature of warmest month (℃)	WorldClim	http://www.worldclim.org/
Bio_6	Min temperature of Coldest month (℃)	WorldClim	http://www.worldclim.org/
Bio_7	Annual temperature range (BIO5-BIO6) (℃)	WorldClim	http://www.worldclim.org/
Bio_12	Annual precipitation (mm)	WorldClim	http://www.worldclim.org/
Bio_13	Precipitation of wettest month (mm)	WorldClim	http://www.worldclim.org/
Bio_14	Precipitation of driest month (mm)	WorldClim	http://www.worldclim.org/
Bio_15	Precipitation seasonality (mm)	WorldClim	http://www.worldclim.org/
Altitude	Altitude (m)	WorldClim	http://www.worldclim.org/
Aspect	Aspect (°)	Derived from Altitude	http://www.worldclim.org/
Slope	Slope	Derived from Altitude	http://www.worldclim.org/
Landcover	Land cover	ESA	http://www.esa-landcover-cci.org/
Disroad	Distance to roads (m)	Road layer from Natural Earth	http://www.naturalearthdata.com/
Disrard	Distance to railways (m)	Railroad	http://www.naturalearthdata.com/

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

		layer from Natural Earth	
Disriver	Distance to rivers (m)	River layer from Natural Earth	http://www.naturalearthdata.com/
Dislake	Distance to lakes (m)	Lake layer from Natural Earth	http://www.naturalearthdata.com/
Discoastline	Distance to coastline (m)	Coastline layer from Natural Earth	http://www.naturalearthdata.com/
Dissettle	Distance to settlements (m)	Settle layer from Natural Earth	http://www.naturalearthdata.com/

Model development

We created TreeNet, Random Forest, CART, Maxent models and ensemble model (averaged value of the former four model results) for Hooded Cranes, White-naped Cranes and Black-naped Cranes. These four model algorithms are considered to be among the most accurate machine learning methods (more information about these four models can be seen in the references by Breiman et al., 1984, Breiman 2001, Friedman 2002, Phillips et al., 2004, Hegel et al., 2010). The first three machine learning models are binary (presence-pseudo absence) models and were handled in Salford Predictive Modeler 7.0 (SPM). For more details on TreeNet, Random Forest and CART in SPM, we refer readers to the user guide document online (https://www.salford-systems.com/products/spm/userguide). Several implementations of these algorithms exist. Approximately 10,000 'pseudo-absence' locations were selected by random sampling across the study area for each species using the freely available Geospatial Modeling Environment (GME;

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

Hawth's Tools; Beyer 2013; see Booms et al., 2010 and Ohse et al., 2009 for examples). We extracted the habitat information from the environmental layers for presence and pseudo-absence points for each crane, and then constructed models in SPM with these data. In addition, we used balanced class weights, and 1000 trees were built for all models to find an optimum within, others used default settings.

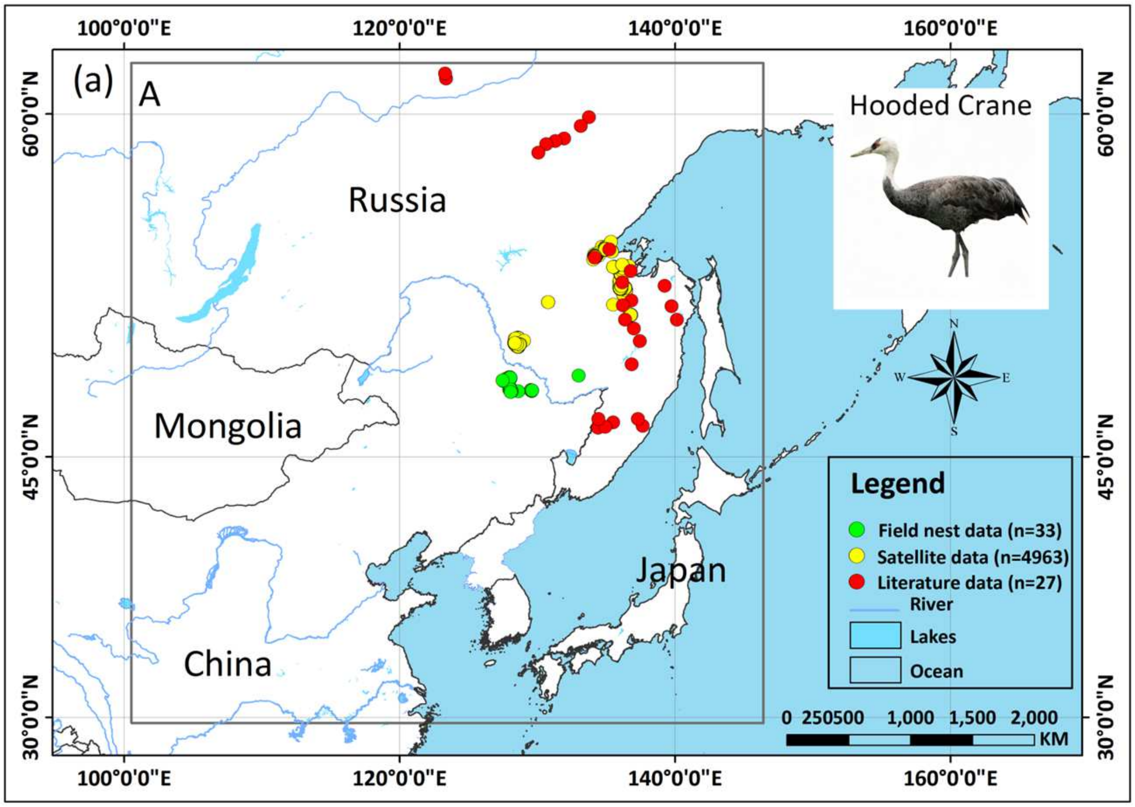
For the predictions, we created a 'lattice' (equally spaced points across the study area; approximately 5×5 km spacing for the study area). For the lattice, we extracted information from the same environmental layers (Table 1) as described above for each point and then used the model to predict ('score') bird presence for each of the regular lattice points. For visualization, we imported the dataset of spatially referenced predictions ('score file') into GIS as a raster file and interpolated for visual purposes between the regular points using inverse distance weighting (IDW) to obtain a smoothed predictive map of all pixels for the breeding distributions of the three cranes (as performed in Booms et al., 2010 and Ohse et al., 2009). The fourth algorithm we employed, Maxent, is commonly referred to as a presence-only model; we used Maxent 3.3.3k (it can be downloaded for free from http://www.cs.princeton.edu/~schapire/maxent/) to construct our models. To run Maxent, we followed the 3.3.3e tutorial for ArcGIS 10 (Young et al., 2011) and used default settings.

Testing data and model assessment

We applied two types of testing data in this study: one consisted of satellite tracking data, and the other was represented by data from the literature. Satellite tracking data were obtained from 4 individual Hooded Cranes and 8 White-naped Cranes that were tracked in the breeding regions at stopover sites (for more details regarding the information for tracked cranes, please see Supplement S2). The satellite tracking devices could provide 24 data points per day (Databases could be available upon request). Here, we chose points that had a speed of less than 5 km/h during the period from 1st May to 31th June for Hooded Cranes and 15th April to 15th June for White-naped Cranes as the locations of the breeding grounds for these two cranes. The total numbers of tracking data points were 4,963 and 7,712 (Hooded Cranes and White-naped Crane, respectively. We didn't track Black-necked Cranes, so there was no tracking testing data for this species). The literature data for this study were obtained by geo-referencing the location points of detections from 1980-2000 (ArcGIS 10.1) from Threatened Birds of Asia: the BirdLife International Red Data Book (Collar et al., 2001). From this hardcopy data source, we were able to obtain and digitize 27 breeding records for Hooded Cranes, 43 breeding records for White-naped Cranes, and 53 breeding records for Black-necked Cranes (see Fig. 2a, 2b, 2c). We digitized the only crane data for these three species in East-Asia into a database.

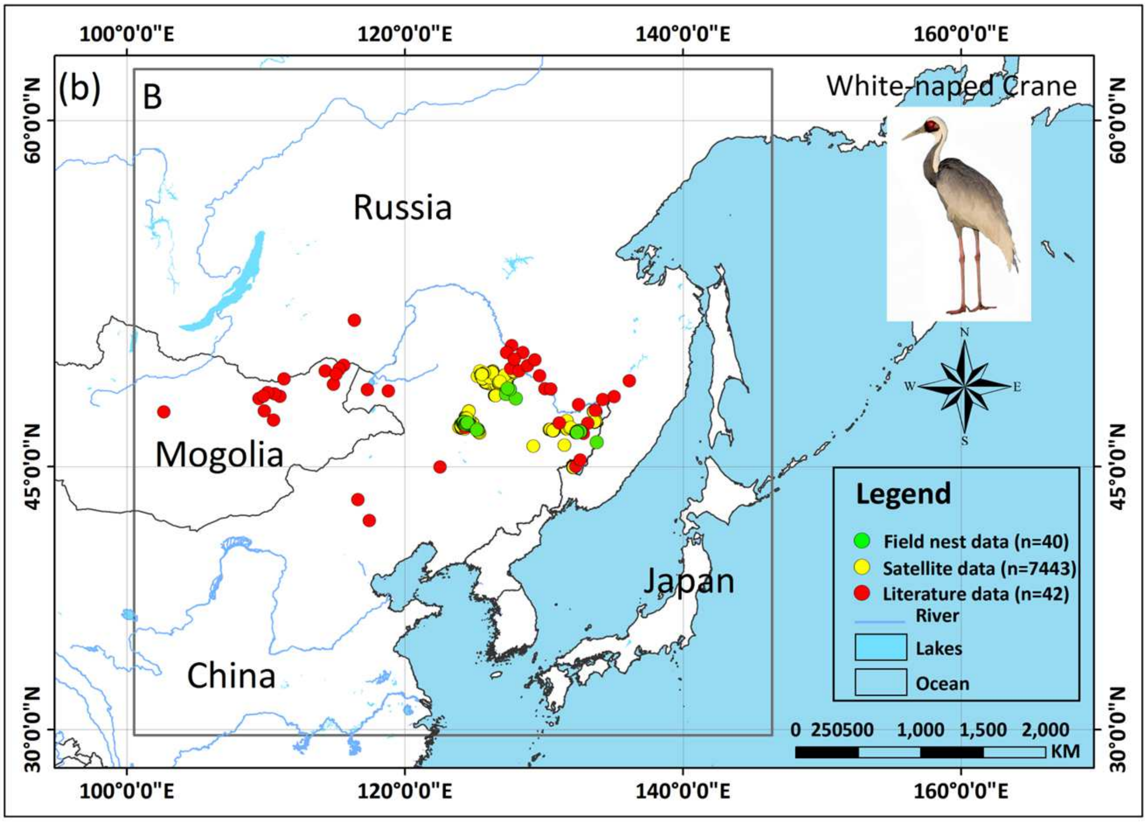
In addition, we generated 3,000 random points for Hooded Cranes and White-naped Cranes, and 5,000 random points for Black-necked Cranes as testing absence points in their respective study areas. And then, the literature locations (additional presence points for testing) and random points location (testing absence points) that contrasted with the associated predictive value of RIO extracted from the relative prediction map, which were used to calculate receiver operating characteristic (ROC) curves and the true skill statistic (TSS) (Hijmans and Graham, 2006). The area under the ROC curve (AUC) is commonly used to evaluate models in species distributional modeling (Manel *et al.,* 2001, McPherson *et al.,* 2004). TSS was also used to evaluate model performance; we used TSS because it has been increasingly applied as a simple but robust and intuitive measure of the performance of species distribution models (Allouche et al., 2006).

Put Fig.2 here



100°0'0"E
120°0'0"E
140°0'0"E
160°0'0"E
60°0'0"N
45°0'0"N
30°0'0"N
(a)
A
Russia
Mongolia
China
Japan
Hooded Crane
N
S
W
E
Legend
Field nest data (n=33)
Satellite data (n=4963)
Literature data (n=27)
River
Lakes
Ocean
0 250500 1,000 1,500 2,000
KM

225



(b)
B
White-naped Crane
Russia
Mogolia
China
Japan
Legend
Field nest data (n=40)
Satellite data (n=7443)
Literature data (n=42)
River
Lakes
Ocean
0 250500 1,000 1,500 2,000
KM
100°0'0"E
120°0'0"E
140°0'0"E
160°0'0"E
60°0'0"N
45°0'0"N
30°0'0"N

226

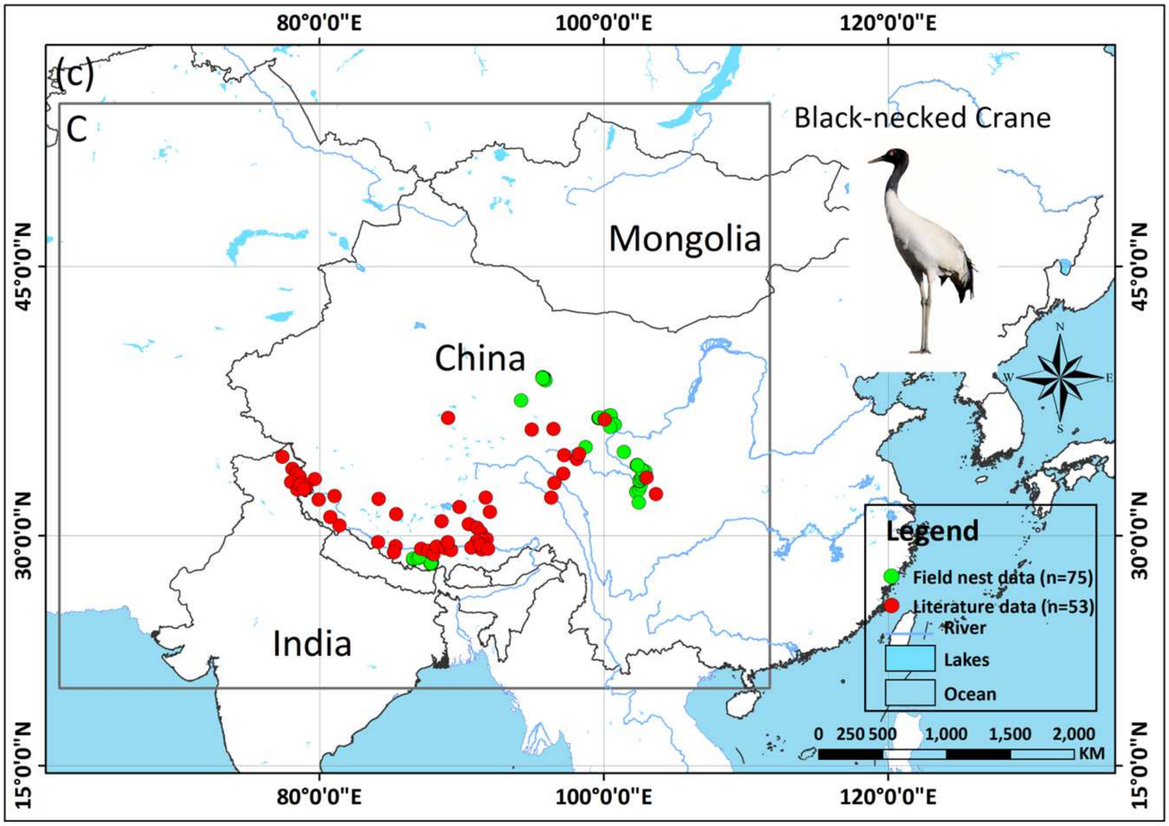


Figure 2 Detailed study areas showing the presence of and testing data used for the three cranes. 2a) Hooded Cranes, 2b) White-naped Cranes, 2c) Black-necked Cranes.

To assess models transferability, we extracted the predictive value of the relative index of occurrence (RIO) for testing data sets from the prediction maps using GME. We then constructed resulting violin plots for these extracted RIOs to visualize their one-dimensional distribution. This method allowed us to examine the degree of generalizability based on the local area with samples to predict into undersampled areas that are otherwise unsampled in the model development (=areas without training data). In addition, AUC is also commonly used to assess model transferability in our study referring Randin et al. (2006).

RESULTS

Model performance

The results for AUC and TSS, two metrics commonly used to evaluate model accuracy, are listed in Table 2. For the four SDMs technique, our results showed that the AUC values for Random Forest were always highest (>0.625), ranking this model in first place, followed by Maxent (>0.558), and then either CART or TreeNet (>=0.500). TSS showed us consistent results, as was the case for AUC, and Random Forest performed the best (>0.250) followed by Maxent (>0.137) for all three crane species, CART took the third place for Black-necked Cranes, and TreeNet performed better than CART for White-naped Cranes. And the results showed there was a trend that the value of these three metrics increased with an increase of nest site samples (33 to 75, Hooded Crane to Black-necked Crane, see Table. 2). Comparing the results of Random Forest with ensemble model, we found their performance were close. Random Forest obtained better model for Hooded Cranes and White-naped Cranes cases, ensemble model performed better for Black-necked Cranes.

Put Table 2 here

Table 2 AUC and TSS values for four machine learning models and their ensemble model with three crane species based on literature testing data.

Accuracy metric (samples)	Species distribution model				
	TreeNet	Random Forest	CART	Maxent	Ensemble
Hooded Crane (*Grus monacha,* n=33 sites)					
AUC	0.504	**0.625**	0.500	0.558	0.558

TSS	0.000	**0.250**	0.000	0.137	0.117
White-naped Crane (*Grus vipio*, n=40 sites)					
AUC	0.605	**0.754**	0.564	0.712	**0.754**
TSS	0.210	**0.509**	0.128	0.424	0.508
Black-necked Crane (*Grus nigricollis*, n=75 sites)					
AUC	0.528	0.830	0.672	0.805	**0.843**
TSS	0.055	0.660	0.345	0.611	**0.686**

Model generalization

Violin plots for RIOs with overlaid satellite tracking data (Fig. 3) showed that Random Forest for Hooded Cranes and White-naped Cranes performed better than the other three models. In the Hooded Crane models (Fig. 3a), the RIO for most satellite tracking data indicated that TreeNet, and CART predicted with a value around 0; Ensemble model demonstrated a slightly higher value than the other three models but was still much lower than Random Forest. Fig. 3b indicates the same situation than found in Fig. 3a: Random Forest still performed better than the other three models (median values in Random Forests were close to 1.00). TreeNet had a median RIO value of approximately 0.71, followed by Maxent (median was 0.37) and then ensemble and CART. While some tracking points had a low RIO value in TreeNet, the majority of RIO values for CART remained in the 0.20 range.

Put Fig. 3 here

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

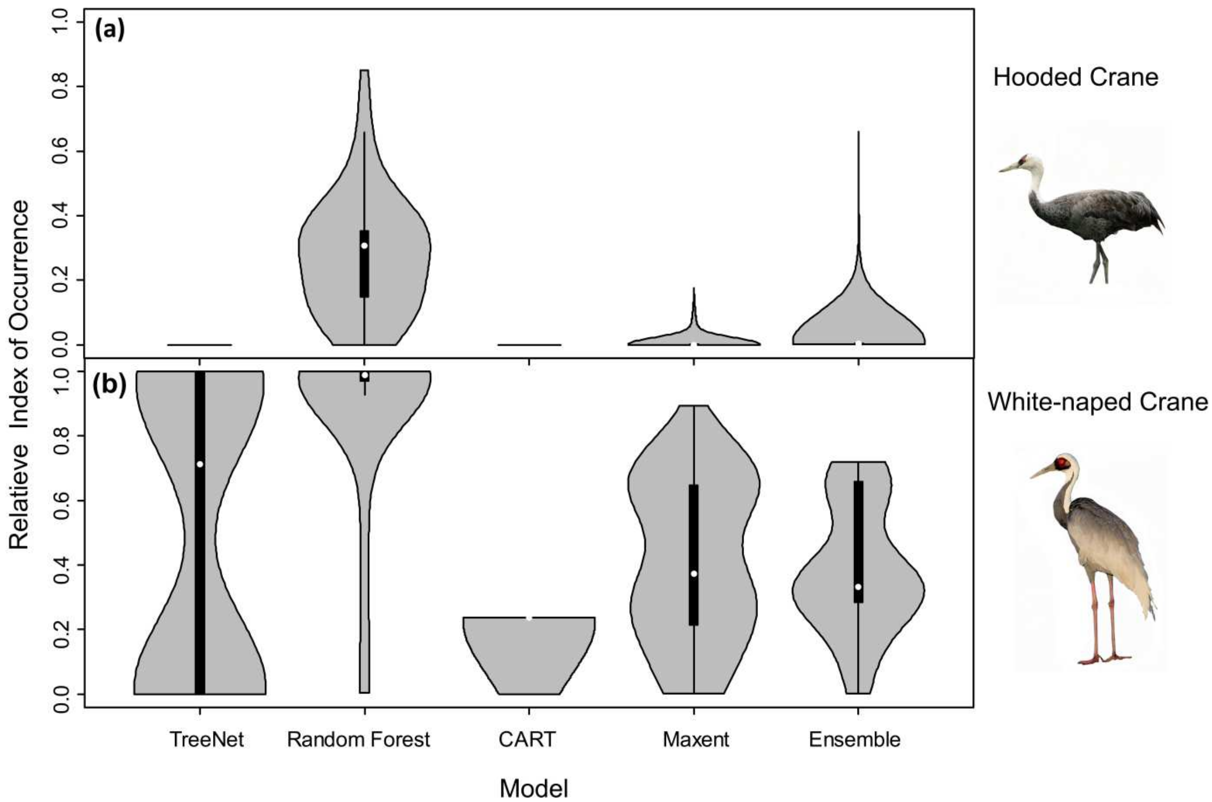


Figure 3 Violin plots of the Relative Index of Occurrence (RIO) for four SDMs and ensemble model for Hooded Cranes and White-naped Cranes based on satellite tracking data. 3a) violin plots of Hooded Cranes, 3b) violin plots of White-naped Cranes.

Violin plots of the RIOs values for the three cranes extracted for the literature data from the prediction maps (Fig. 4) demonstrated consistent trends (Fig. 3), indicating that Random Forest performed best across all models of the three species. In Fig. 4a, the RIO values for Random Forest ranged from 0 to 0.48, and most RIO values were below 0.1; the RIO values for the other three SDMs method were 0, the ensemble model performed a little bit better. As showed in Fig. 4b, most RIO values for Random Forest were below 0.7, and the median value was approximately 0.20, followed by Maxent and then CART. The violin plots for Black-necked Cranes (Fig. 4c) indicated that TreeNet performed the worst, although there were some pixels that had high RIO values, followed by ensemble and then Maxent. The best performer was still Random Forest, and its RIOs were distributed evenly to a certain extent with a median value of 0.44. The results of AUC, as mentioned in "Model performance" part (Table 2), showed consistent results with violin plots, Random Forest always get the highest value and has the best generalization.

Put figure 4 here

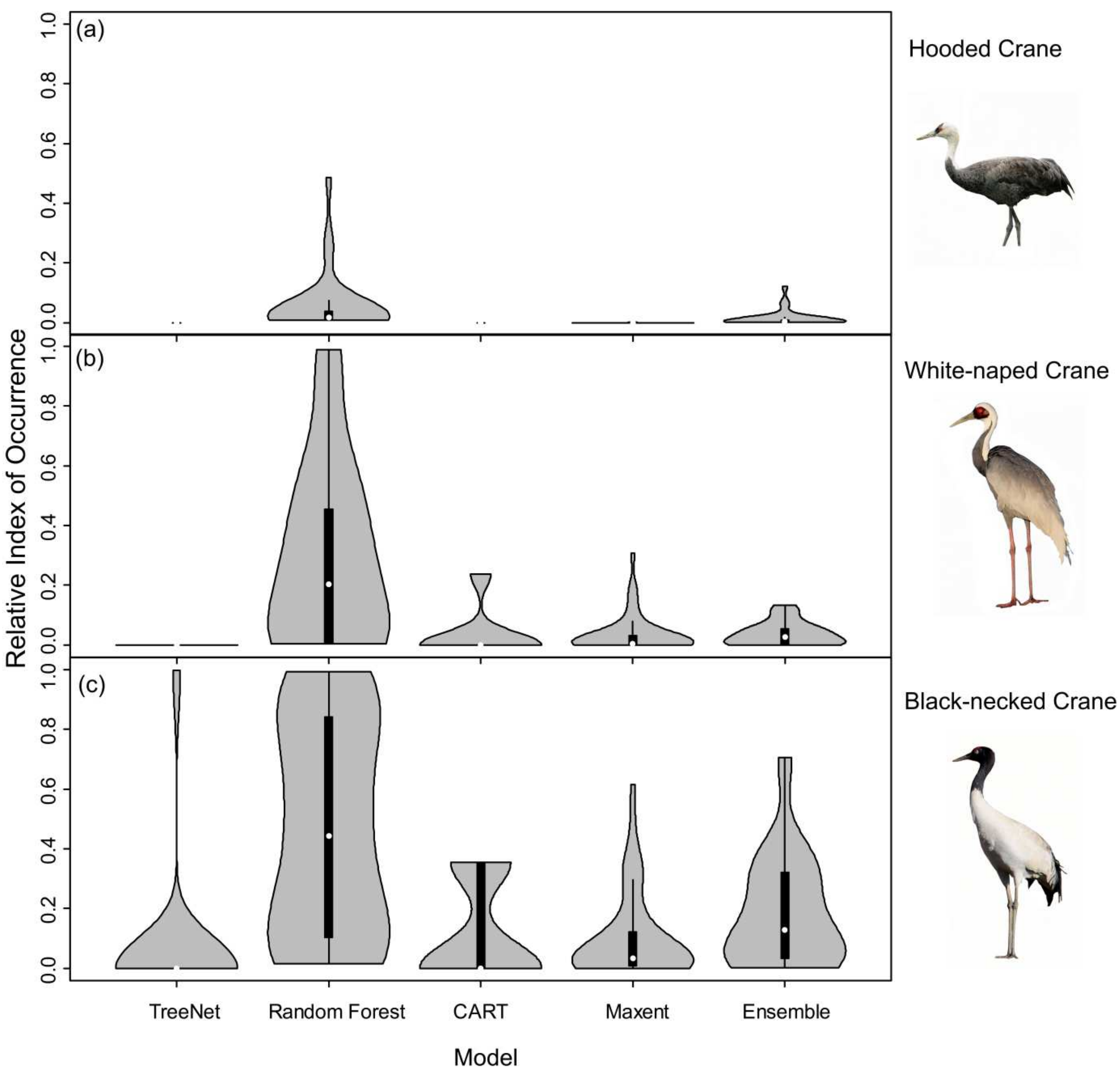


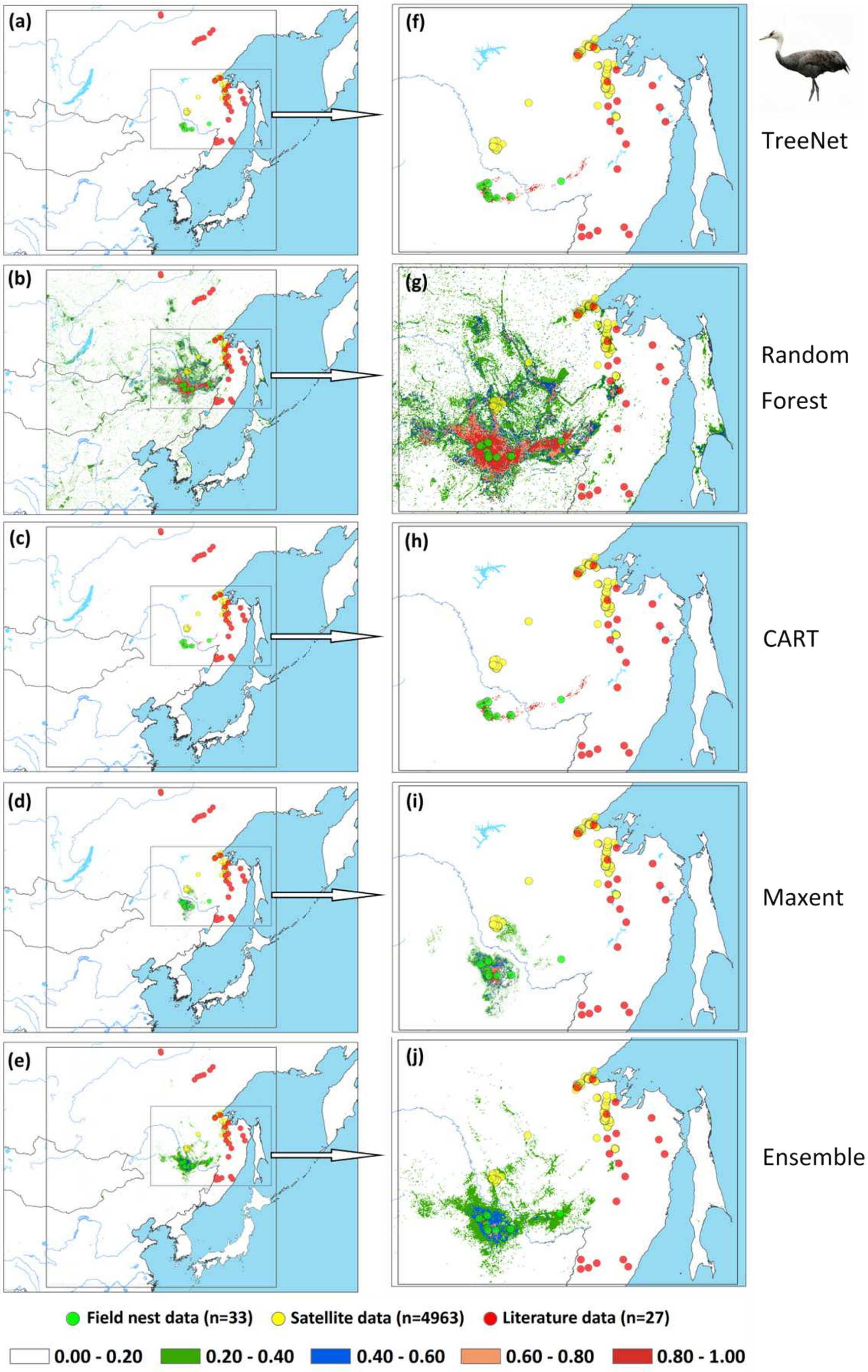
Figure 4 Violin plots of Relative Index of Occurrence (RIO) values for four SDMs and ensemble model for three cranes based on calibration data from Threatened Birds of Asia. 4a) Violin plots

for Hooded Cranes, 4b) violin plots for White-naped Cranes, 4c) violin plots for Black-necked Cranes.

Spatial assessment using a testing data overlay prediction map

An assessment of niche prediction beyond the local area where samples were located represents a real test of the generalizability of the model predictions in undersampled areas. This approach was used to evaluate whether testing data (satellite tracking data and literature data) locations matched predictions of the potential distribution area, as a spatial assessment of model performance. It's a spatial and visual method to show the transferability of SDMs from sampled to unsampled areas. From the results (Fig.s 5, 6 and 7. Digital version for each subgraph could be available request), we found that Random Forest demonstrated the strongest performance to handle generality (transferability), and a high fraction of testing data locations were predicted in the distribution areas of the three cranes (Fig.s 5b, 5g, 6b, 6g, 7b, 7g). The order of the generality of the remaining four models was: ensemble model followed by Maxent, CART and then TreeNet. Note, however, that the capacities of these models to predict well in undersampled areas were weaker than Random Forest, it holds particularly for areas that were further away from the sample areas (Fig.s 5, 6 and 7). In addition, we found that the generality increased with sample size (33 to 75, Hooded Crane to Black-necked Crane, see Fig.s 5, 6 and 7). This means a higher sample size make models more robust and better to generalize from.

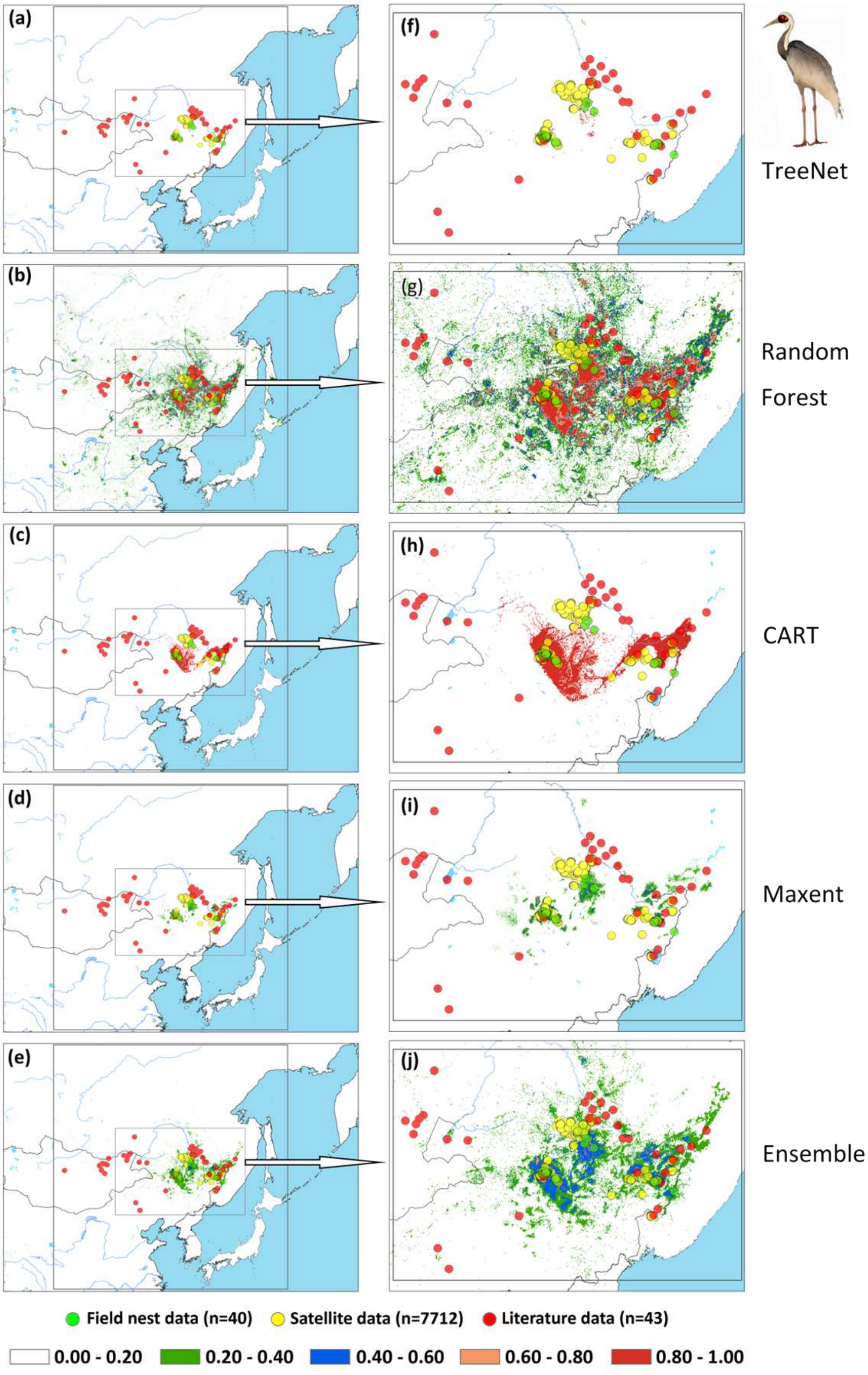
Put Fig. 5 Here



306

307 Figure 5 Prediction maps for Hooded Cranes and zoomed-in maps showing the four models (TreeNet,

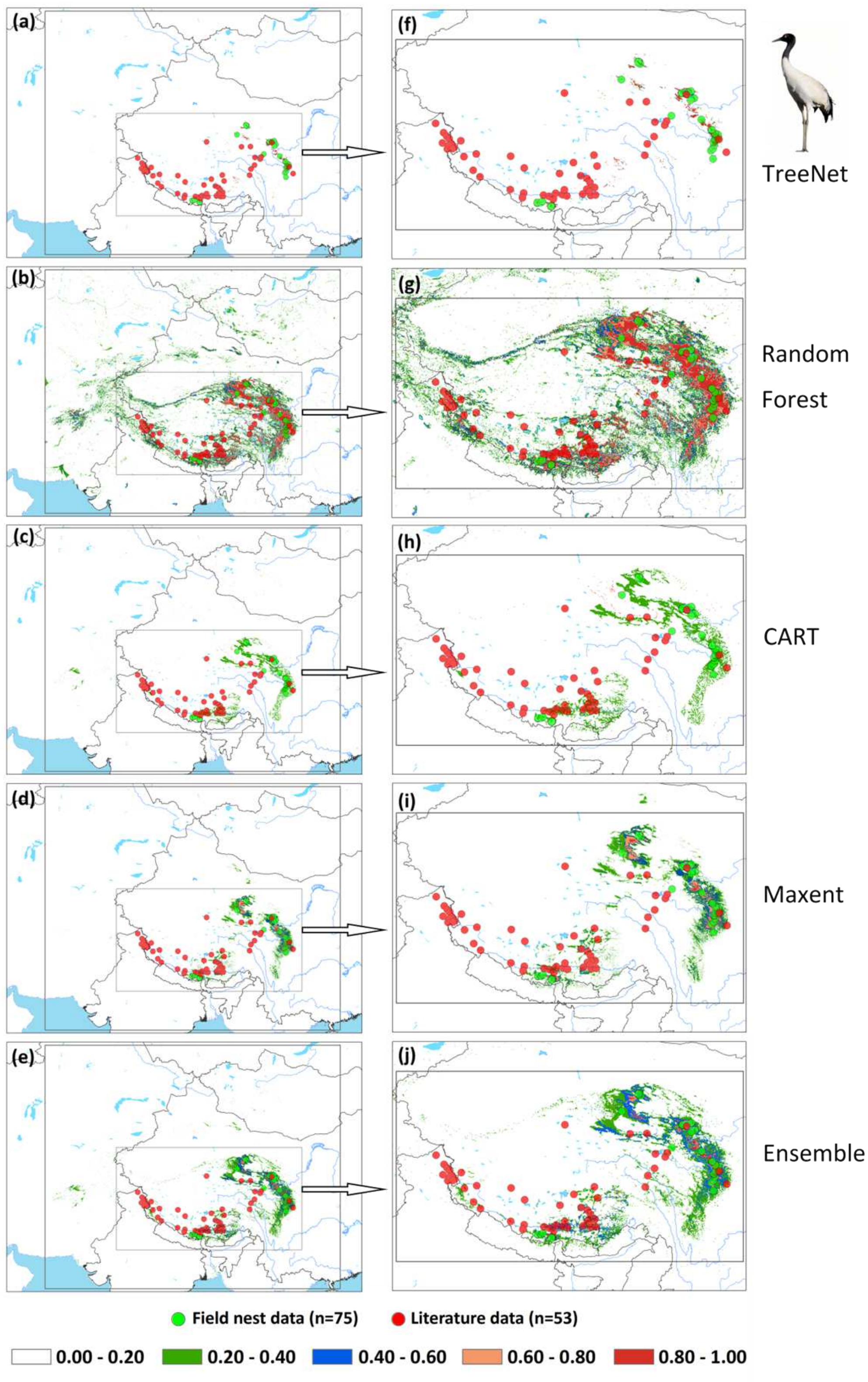
Random Forest, CART and Maxent) and ensemble model in detail. 5a-5e) prediction map for Hooded Cranes, 5f-5j) zoomed-in map for Hooded Cranes.



310

311 Figure 6 Prediction maps for White-naped Cranes and zoomed-in maps showing the four models (TreeNet,

Random Forest, CART and Maxent) and ensemble model in detail. 6a-6e) prediction map for White-naped Cranes, 6f-6j) zoomed-in map for White-naped Cranes. Put Fig. 6 Here



314

315 Figure 7 Prediction maps for Black-necked Cranes and zoomed-in maps showing the four models (TreeNet,

Random Forest, CART and Maxent) and ensemble model in detail. 7a-7e) prediction map for Black-necked Cranes, 7f-7j) zoomed-in map for Black-necked Cranes.

DISCUSSION

Model generality (transferability)

Estimating species distributions in undersampled areas is a fundamental problem in ecology, biogeography, biodiversity conservation and natural resource management (Drew et al., 2011). That is specifically true for rare and difficult to be detected species and which are usually high on the conservation priority. The use of SDMs has become the method for deriving such estimates (Guisan & Thuiller, 2005; Drew et al., 2011; Guisan et al., 2013) and could contribute to detect new populations of rare species. However, the application of a few samples to project a distribution area widely beyond the sample range is a greater challenge and has rarely been attempted in the literature. And only recently have conservationists realized its substantial value for pro-active decision making in conservation management (see work by Ohse et al., 2010; Drew et al., 2011; Kandel et al., 2015 etc.). Our results based on AUC, violin plots for RIOs and spatial assessment of testing data (satellite tracking data and literature data) all suggest there are difference in the generalization performance of different modeling techniques (TreeNet, Random Forest , CART and Maxent).

Moreover, among the acknowledged four rather powerful and commonly used machne-learning techniques, Random Forest (bagging) in SPM usually had the best performance in each case. Our results are in agreement with those of Prasad et al. (2006), Cutler et al. (2007) and Syphard and Franklin (2009) indicating a superiority of Random Forest in such applications. However, initially it appears to run counter to the conclusions off recent paper (Heikkinen et al., 2012) with the poor transferability of Random Forest. But we propose this is due to the fact that many Random Forest implementations exist (see the 100 classifier paper Fernández-Delgado et al., 2014).

Here we applied Random Forest in SPM which has been optimized under one of the algorithm's original co-authors, while Heikkinen et al. (2012) just run a basic Random Forest with BIOMOD framework in the R sofeware. The differences are known to be rather big (see Herrick 2013).

Furthermore, Maxent, a widely used SDM method enjoyed by many modelers (Phillips et al., 2006; Peterson et al. 2007; Phillips and Dudík 2008; Li et al., 2015, etc.), didn't perform so good in regards to transferability in this study. This contrasts to those of Elith et al. (2006) and Heikkinen et al. (2012), where Manxent and GBM perform well. We infer this may be caused by sample size used as training data. When the sample size increased (33 to 75), the AUC and TSS value of all models rose (Table 2). This indicates that higher sample sizes make models more robust and performing better. Sample sizes of 33 presence points still favor by Random Forest.

In Random Forest, random samples from rows and variables are used to build hundreds of trees. Each individual tree is constructed from a bootstrap sample and split at each node by the best predictor from a very small, randomly chosen subset of the predictor variable pool (Herrick, 2013). These trees comprising the forest are each grown to maximal depth, and predictions are made by averaged trees through 'voting' (Breiman et al., 2006). This algorithm avoiding overfitting by controlling the number of predictors randomly used at each split, using means of out-of-bag (OOB) samples to calculate an unbiased error rate. And also, Random Forest in SPM utilizes additional specific fine-tuning for best performance.

RIOs of random points

In order to explore whether Random Forest created higher RIOs for prediction maps in each grid, which would result higher RIOs of testing data, we generated 3,000 random points for Hooded Cranes and White-naped Cranes, 5000 random points for Black-necked Cranes in their related projected study areas. We made violin plots for RIOs of random points (Fig. 8), and found that more RIO values of random points for Maxent, Random Forest and ensemble models were close to the lower value, and then followed by TreeNet. The distribution shapes of Random Forest, Maxent and ensemble model are more similar to the real distribution of species in the real world. The RIOs of White-naped Crane extracted from the CART model distributed in the range of the low value. That means there were no points located in the high RIO areas of cranes, and which is unrealistic. Consequently, we argued that Random Forest did not create higher RIOs for prediction maps in each grid in our study.

Put Fig. 8 here

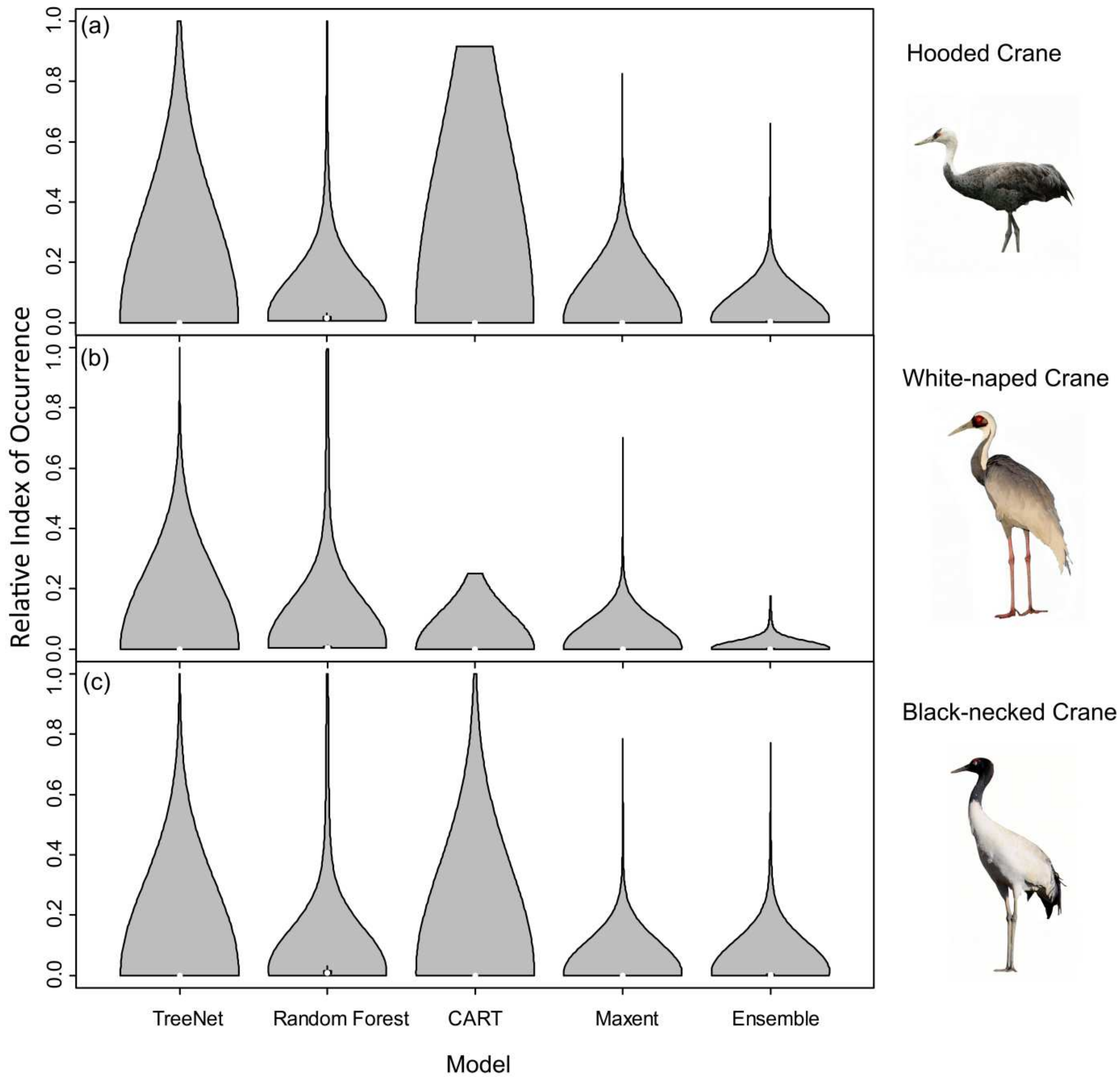


Figure 8 Violin plots of Relative Index of Occurrence (RIO) values for four SDMs and ensemble model for three cranes based on calibration data from Threatened Birds of Asia. 4a) Violin plots for Hooded Cranes, 4b) violin plots for White-naped Cranes, 4c) violin plots for Black-necked Cranes.

Models with small sample sizes

Conservation biologists are often interested in rare species and seek to improve their conservation.

These species usually have limited number of available occurrence records, which poses challenges for the creation of accurate species distribution models when compared with models developed with greater numbers of occurrences (Stockwell & Peterson, 2002; McPherson et al., 2004; Hernandez et al., 2006). In this study, we used three crane species as case studies, and their occurrence records (nests) totaled 33, 40, and 75, respectively (considering the small numbers of samples and given that a low fraction of the area was sampled in the large projected area). In our models, we found that model fit (AUC and TSS, see Table 2) of Random Forest that had the highest index, while Maxent usually ranked second. In addition, we found that models with few presence samples can also generate accurate species predictive distributions (Fig. 3 to 7) with the Random Forest method. Of course, models constructed with few samples underlie the threat of being biased more because few samples usually had not enough information including all distribution gradients conditions of a species, especially for places far away from the location of training presence points. However, the potential distribution area predicted by SDMs could become as the place where scholars could look for the birds (additional fieldwork sampling). And also, these places could be used as diffusion or reintroduction areas!

Evaluation methods

In this study, we applied two widely-used assessment methods (AUC and TSS) in SDMs (Table 2). For evaluation of these three values we used the approach recommended by Fielding & Bell (1997), and Allouche et al. (2006), we found our model usually didn't obtain perfect performance, and some of them were fair. However, for macro-ecology this more than reasonable and ranks rather high. It's a good conservation progress! We identified Random Forest as always the highest performing. These results are consistent with the results of violin plots of the Relative Index of Occurrence (RIO) using tracking as well as literature data (Fig.s 3, 4), and well as matching the spatial assessment results (Fig.s 5-7). And we recommend when modelers assess model performance they should not only depend alone on some metric (such as AUC and TSS), but also should base their assessments on the combined use of visualization and expert knowledge. That means modelers should also assess how the species distribution map actually looks and how it links with real data (see Huettmann & Gottschalk 2011). Spatial assessment metrics from alternative data should matter the most. Expert experience and ecological common knowledge of the species of interest could sometimes also be highly effective (Drew & Perera, 2011), albeit nonstandard, evaluation methods (see Kandel et al., 2015 for an example). Additionally, one alternative method for rapid assessment we find is to use a reliable SDM, and thus Random Forest may be a good choice in the future given our consistent results (Fig.s 3 to 7, Tables 3 to 5) in this study, which involved three species, a vast landscape to conserve, and only limited data. Our work helps to inform conservation decisions for cranes in Northeast Asia.

Limitations and future work

Our study is not without limitations: 1) so far, only three species of cranes are used as a test case in our study. That's because nest data for rare species in remote areas are usually sparse; 2) all our species study areas are rather vast and confined to East-Asia. For future, we would apply Random Forest in more species and in more geography conditions with different distributed feature for a first rapid assessment and baseline mandatory for better conservation. Then we would apply our prediction results in specifically targeted fieldwork sampling campaigns and assess the model accuracy with field survey results (ground-truthing) and more new satellite tracking data. This is to be fed directly into the conservation management process.

ACKNOWLEDGEMENTS

We thank Fengqin Yu, Yanchang Gu, Linxiang Hou, Jianguo Fu, Bin Wang, Jianzhi Li, Lama Tashi Sangpo, Baiyu Lamasery, and Nyainbo Yuze for their hard work in the field. Thanks to all data contributors to the book 'Threatened Birds of Asia'. Further we thank Salford Systems Ltd. for providing the free trial version of their data mining and machine learning software to the conservation research community.

REFERENCES

Allouche OA, Tsoar, Kadmon R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43(6):1223-1232.

Araújo MB, New M. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22(1):42-47.

Araújo MB, Whittaker R, Ladle R, Erhard M. 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology & Biogeography 14(6):529-538.

Ashtonw C, Perera AH. 2010. Expert Knowledge as a Basis for Landscape Ecological Predictive Models. Predictive Species & Habitat Modeling in Landscape Ecology:229-248.

Beyer H. 2013. Hawth's Analysis Tools for ArcGIS version 3.27 (software). in.

Booms TL, Huettmann F, Schempf PF. 2010. Gyrfalcon nest distribution in Alaska based on a predictive GIS model. Polar biology 33(3):347-358.

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FK. 2002. Evaluating resource selection functions. Ecological Modelling 157(2-3):281–300.

Braunisch V, Coppes J, Arlettaz R, Suchant R, Schmid H, Bollmann K. 2013. Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change. Ecography 36(9):971-983.

Breiman L. 2001. Random forests. Machine learning 45(1):5-32.

Breiman L., Friedman J, Stone CJ, Olshen RA. 1984. Classification and regression trees. CRC press.

Bucklin DN, Basille M, Benscoter AM, Brandt LA, Mazzotti FJ, Romanach SS, Speroterra C, Watling JI. 2015. Comparing species distribution models constructed with different subsets of environmental predictors. Diversity and Distributions 21(1):23-35.

Cohen J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20(1):37-46.

Collar NJ, Crosby R, Crosby M. 2001. Threatened birds of Asia: the BirdLife International red data book. Volume 1.BirdLife International Cambridge, UK.

Cushman SA, Huettmann F. 2010. Spatial Complexity, Informatics, and Wildlife Conservation. Springer, Springer Tokyo Berlin Heidelberg New York.

Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random forests for classification in ecology. Ecology 88(11):2783-2792.

Drew CA, Perera AH. 2011. Expert knowledge as a basis for landscape ecological predictive models. Pages 229-248 in Predictive Species and Habitat Modeling in Landscape Ecology. Springer.

Drew CA, Wiersma Y, Huettmann F. 2011. Predictive species and habitat modeling in landscape ecology. Springer.

Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JMM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129-151.

Eskildsen, A., P. C. Roux, R. K. Heikkinen, T. T. Høye, W. D. Kissling, J. Pöyry, M. S. Wisz, and M. Luoto. 2013. Testing species distribution models across space and time: high latitude butterflies and recent warming. Global ecology and biogeography 22(12):1293-1303.

Estes L, Bradley B, Beukes H, Hole D, Lau D, Oppenheimer M, Schulze R, Tadross M, Turner

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

W. 2013. Comparing mechanistic and empirical model projections of crop suitability and productivity: implications for ecological forecasting. Global ecology and biogeography 22(8):1007-1018.

Fernández-Delgado M, Cernadas E, Barro S, Amorim D. 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15(1):3133-3181.

Ferrier S, Watson G, Pearce J, Drielsma M. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodiversity & Conservation 11(12):2275-2307.

Fielding AH, Bell JF. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental conservation 24(1):38-49.

Fielding AH, Haworth PF. 1995. Testing the Generality of Bird‐Habitat Models. Conservation biology 9(6):1466-1481.

Ferrier S, Watson G, Pearce J, Drielsma M. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. Biodiversity & Conservation 11(12):2275-2307.

Friedman JH. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38(4):367-378.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19(9):497-503.

Guillera‐Arroita G, Lahoz‐Monfort JJ, Elith J, Gordon A, Kujala H, Lentini PE, McCarthy MA, Tingley R, Wintle BA. 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Global ecology and biogeography 24(3):276-292.

Guisan A, Thuiller W. 2005. Predicting species distribution: offering more than simple habitat models. Ecology letters 8(9):993-1009.

Guisan A, Tingley R, Baumgartner JB, Naujokaitis‐Lewis I, Sutcliffe PR, Tulloch AIT, Regan TJ, Brotons L, Mcdonald‐Madden E, Mantyka‐Pringle C. 2013. Predicting species distributions for conservation decisions. Ecology letters 16(12):1424-1435.

Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135(2):147-186.

Gwena LL, Robin E, Erika F, Guisan A. 2010. Prospective sampling based on model ensembles improves the detection of rare species. Ecography 33(6):1015-1027.

Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29-36.

Hanley JA, McNeil BJ. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148(3):839-843.

Hardy SM, Lindgren M., Konakanchi H, Huettmann F. 2011. Predicting the distribution and ecological niche of unexploited snow crab (Chionoecetes opilio) populations in Alaskan waters: a first open-access ensemble model. Integrative and comparative biology 51(4):608-622.

Hegel TM, SA Cushman, J Evans, Huettmann F. 2010. Current State of the Art for Statistical Modelling of Species Distributions. Spatial Complexity, Informatics, and Wildlife Conservation:273-311.

Heikkinen RK, Marmion M, Luoto M. 2012. Does the interpolation accuracy of species distribution models come at the expense of transferability? Ecography 35(3):276-288.

Hernandez PA, Graham CH, Master LL, Albert DL. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29(5):773-785.

Herrick K. 2013. Predictive Modeling of Avian Influenza in Wild Birds. Veterinary Research.

Hijmans RJ, Graham CH. 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12(12):2272-2281.

Huettmann F, Gottschalk T. 2011. Simplicity, Model Fit, Complexity and Uncertainty in Spatial Prediction Models Applied Over Time: We Are Quite Sure, Aren't We? Pages 189-208 in Predictive Species and Habitat Modeling in Landscape Ecology. Springer.

Humphries GRW, Huettmann F. 2014. Putting models to a good use: a rapid assessment of Arctic seabird biodiversity indicates potential conflicts with shipping lanes and human activity. Diversity and Distributions 20(4):478-490.

Jiguet F, Barbet-Massin M, Chevallier D. 2011. Predictive distribution models applied to satellite tracks: modelling the western African winter range of European migrant Black Storks Ciconia nigra. Journal of Ornithology 152(1):111-118.

Kandel K, Huettmann F, Suwal MK, Regmi GR, Nijman V, Nekaris K, Lama ST, Thapa A, Sharma HP, Subedi TR. 2015. Rapid multi-nation distribution assessment of a charismatic conservation species using open access ensemble model GIS predictions: Red panda (Ailurus fulgens) in the Hindu-Kush Himalaya region. Biological Conservation 181:150-161.

Keith DA, Elith J, Simpson CC. 2014. Predicting distribution changes of a mire ecosystem under future climates. Diversity and Distributions 20(4):440-454.

Kessler A, Batbayar N, Natsagdorj T, Batsuur D, Smith A. 2013. Satellite telemetry reveals long‐distance migration in the Asian great bustard Otis tarda dybowskii. Journal of Avian

Biology 44(4):311-320.
Kleyer M. 2002. Validation of plant functional types across two contrasting landscapes. Journal of Vegetation Science 13(2):167-178.
Lei, Z., L. Shirong, S. Pengsen, and WangTongli. 2011. Comparative evaluation of multiple models of the effects of climate change on the potential distribution of Pinus massoniana. Chinese Journal of Plant Ecology 35(11):1091-1105.
Li, R., M. Xu, M. H. G. Wong, S. Qiu, X. Li, D. Ehrenfeld, and D. Li. 2015. Climate change threatens giant panda protection in the 21st century. Biological Conservation 182:93-101.
Maggini R, Lehmann A, Zbinden N, Zimmermann NE, Bolliger J, Schröder B, Foppen R, Schmid H, Beniston M, Jenni L. 2014. Assessing species vulnerability to climate and land use change: the case of the Swiss breeding birds. Diversity and Distributions 20(6):708-719.
Manel S, Williams HC, Ormerod SJ. 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38(5):921-931.
McPherson J, Jetz W, Rogers DJ. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? Journal of Applied Ecology 41(5):811-823.
Mingchang C, Guangsheng Z, Ensheng W. 2005. Application and comparison of generalized models and classification and regression tree in simulating trees species distribution. ACTA ECOLOGICA SINICA 25(8):2031-2040.
Navarro ‐ Cerrillo R, Hernández ‐ Bermejo J, Hernández ‐ Clemente R. 2011. Evaluating models to assess the distribution of Buxus balearica in southern Spain. Applied Vegetation Science 14(2):256-267.
Ohse B, Huettmann F, Ickert-Bond SM, Juday GP. 2009. Modeling the distribution of white spruce

(Picea glauca) for Alaska with high accuracy: an open access role-model for predicting tree species in last remaining wilderness areas. Polar biology 32(12):1717-1729.

Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34(1):102-117.

Peterson AT., Monica P, Muir E. 2007. Transferability and model evaluation in ecological niche modeling: a comparison of GARP and Maxent. Ecography 30(4):550–560.

Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190(3):231-259.

Phillips SJ, Dudík M. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31(2):161-175.

Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution modeling. ACM, 2004.

Prasad AM, Iverson LR, Liaw A. 2006. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. Ecosystems 9(2):181-199.

Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A. 2006. Are niche-based species distribution models transferable in space? Journal of Biogeography 33(10):1689-1703.

Romo H, García-Barros E, Márquez AL, Moreno JC, Real R. 2014. Effects of climate change on the distribution of ecologically interacting species: butterflies and their main food plants in Spain. Ecography 37(11):1063-1072.

Stockwell DR, Peterson AT. 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148(1):1-13.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2517v1 | CC BY 4.0 Open Access | rec: 11 Oct 2016, publ: 11 Oct 2016

Stokes KL, Broderick AC, Canbolat AF, Candan O, Fuller WJ, Glen F, Levy Y, Rees AF, Rilov G, Snape RT, Stott I, Tchernov D, Godley BJ. 2015. Migratory corridors and foraging hotspots: critical habitats identified for Mediterranean green turtles. Diversity and Distributions 21(6): 665-674.

Swets JA 1988. Measuring the accuracy of diagnostic systems. Science 240(4857):1285-1293.

Syphard DA, Franklin J. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. Ecography 32(6):907-918.

Thomas JA, Bovee KD. 1993. Application and testing of a procedure to evaluate transferability of habitat suitability criteria. Regulated rivers 8:285-285.

Thuiller W. 2003. BIOMOD–optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9(10):1353-1362.

Zhai T, Li X. 2012. Climate change induced potential range shift of the crested ibis based on ensemble models. ACTA ECOLOGICA SINICA 32(8):2361-2370 (in Chinese).

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. Diversity and Distributions 15(4):565-576.

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A. 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14(5):763-773.

Yen P, Huettmann F, Cooke F. 2004. A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (Brachyramphus marmoratus) during the breeding season in coastal British Columbia, Canada. Ecological Modelling 171(4):395-413.

Young N, Carter L, Evangelista P. 2011. A MaxEnt Model v3.3.3e Tutorial.

Zhang M, Zhou Z, Chen W, Cannon CH, Raes N, Slik JWF. 2014. Major declines of woody plant species ranges under climate change in Yunnan, China. Diversity and Distributions 20(4):405-415.