Why 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

and decisions for efficient conservation.

1 **Why choose Random Forest to predict rare species distribution with few samples in large**

2 **undersampled areas? Three Asian crane species models provide supporting evidence**

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

4 [1]College of Nature Conservation, Beijing Forestry University, P.O. Box 159, Beijing 100083,

5 China

6 [2]Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic

7 Sciences and Natural Resources Research, University of Chinese Academy of Sciences, Beijing

8 100101, China

9 [3]EWHALE Lab, Department of Biology and Wildlife, Institute of Arctic Biology, University of

10 Alaska Fairbanks (UAF), 419 Irving I, P.O. Box 757000, AK 99775, USA

11 Corresponding author:

12 Yumin Guo

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

14 China

15 guoyumin@bjfu.edu.cn

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, and beyond. 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) In addition we developed an ensemble forecast by averaging predicted relative indeces of occurrence for the 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 most assessment methods, 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, in (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.

INTRODUCTION

40

41 Species distribution models (SDMs) are empirical ecological models that relate species

42 observations to environmental predictors (Guisan & Zimmermann, 2000; Drew et al., 2011).

43 SDMs have become an increasingly important and now essential tool in ecology, biogeography,

44 evolution and, more recently, in conservation biology (Guisan et al., 2013), management

45 (Cushman & Huettmann, 2010), impact assessments (Humphries & Huettmann, 2014) and

46 climate change research (Lei et al., 2011; Mi et al., 2016). To generalize and infer from a model,

47 or model transferability is defined as geographical or temporal cross-applicability of models

48 (Thomas & Bovee 1993; Kleyer 2002; Randin et al., 2006). It is one important feature in SDMs,

49 a base-requirement in several ecological and conservation biological applications (Heikkinen et

50 al., 2012). In this study, we used generality (transferability) as the concept of generalizing

51 distribution from sampled areas to unsampled areas (extrapolation beyond the data) in one study

52 area.

53 Detailed distribution data for rare species in large areas are rarely available in SDMs (Pearson

54 et al., 2007; Booms et al., 2010). However, they are among the most needed for their

55 conservation to be effective. Collecting and assembling distribution data for species, especially

56 for rare or endangered species in remote wilderness areas is often a very difficult task, requiring a

57 large amount of human, time and funding sources (Gwena et al., 2010; Ohse et al., 2009).

58 Recent studies have suggested that machine-learning (ML) methodology, may perform better

59 than the traditional regression-based algorithms (Elith et al., 2006). TreeNet (boosting; Friedman

60 2002), Random Forest (bagging; Breiman, 2001), CART (Breiman et al., 1984) and Maxent

61 (Phillips et al., 2004) are considered to be among the most powerful machine learning algorithms

62 and for common usages (Elith et al., 2006; Wisz et al., 2008; Williams et al., 2009; Lei et al.,

63   2011) and for obtaining powerful ensemble models (Araújo and New 2007; Hardy et al., 2011).

64   Although Heikkinen et al. (2012) compared the four SDMs techniques' transferability in their

65   study, they did not test with rare species and few samples in undersampled areas. It is important

66   to understand that the software platform of the former three algorithms (Boosted Regression

67   Trees, Random Forest and CART) applied by Heikkinen et al. (2012) from the R software

68   ("BIOMOD" framewok) comes without a GUI and lacks sophisticated optimization, sample

69   balancing and fine-tuning, but as they are commonly used though by numerous SDM modelers.

70   Instead, we here run these models in the Salford Predictive Modeler (SPM version 7) by Salford

71   Systems Ltd (https://www.salford-systems.com/). These algorithms in SPM are further optimized

72   and improved by one of the algorithm's original co-authors (especially for TreeNet and Random

73   Forest). It runs with a convenient GUI, and produces a number of descriptive results and graphics

74   which are virtually not available in the R version. While this is a commercial software, it is

75   usually available on a 30 days trial version (which suffices for most model runs we know. As

76   well, some of the features of the randomForest R package, most notably the ability to produce

77   partial dependence plots (Herrick 2013), are not directly implemented yet in SPM7 (but they can

78   essentially be obtained by running TreeNet in a Random Forest model).

79   Model generality (transferability) testing could offer particularly powerful for model

80   evaluation (Randin et al., 2006). Independent observations from a training data set has been

81   recommended as a more proper evaluations of models (Fielding & Bell 1997; Guisan and

82   Zimmermann 2000). So the use of an independent geographically (Fielding & Haworth, 1995) or

83   temporally (Boyce et al., 2002; Araujo et al., 2005b) testing data set is encouraged to assess the

84   generality of different SDMs techniques. Data from museum specimen, published literature

85   (Graham et al., 2004) as well as tracking are good source to assess model generality

86   (transferability) performance. In addition, how the distribution map links with reality data,

87   especially in undersampled areas where modelers want to make predictions should definitely be

88   employed as a metric to assess model performance and generalization. Arguably, if model

89   predictions perform very well there, great progress is provided and usually done cost-effective

90   too. Whereas, predictions on existing knowledge and data offers less progress. The model

91   prediction and conservation frontier obviously sits in the unknown and to provide progress there

92   (Breiman 2001, Drew et al. 2011).

93   In this study, we investigated models for the best-available data for three species in East Asia

94   as test cases: Hooded Cranes (*Grus monacha*, n=33), White-naped Cranes (*Grus vipio*, n=40) and

95   Black-necked Cranes (*Grus nigricollis*, n=75). Four machine-learning model algorithms

96   (TreeNet, Random Forest, CART and Maxent) were applied to map breeding distributions for

97   these three crane species which otherwise lack empirically derived distribution information. In

98   addition, two kinds of independent testing data sets (latest satellite tracking data, and compiled

99   literature data (Threatened Birds of Asia: Collar *et al.,* 2001) were obtained to test the

100 transferability of the four model algorithms. The purpose of this investigation is to explore

101 whether there is a SDM technique among the four algorithms that could generate reliable and

102 accurate distributions with high generality for rare species using few samples but in large

103 undersampled areas? Results from this research could be useful for the detection of rare species

104 and enhance fieldwork sampling in large undersampled areas which would save money and

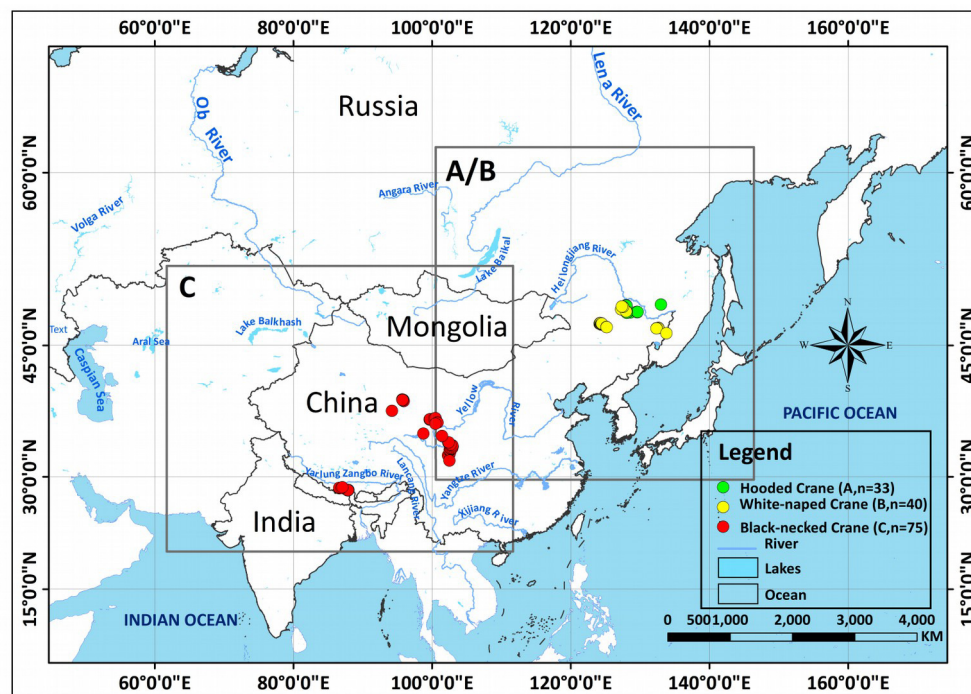105 effort, as well as advance the conservation management of those species.

106 ## MATERIALS AND METHODS

107 ## Species data

108 In our 13 combined years of field work, we have collected 33 Hooded Crane nests (2002-2014),

109 40 White-naped Crane nests (2009-2014) , and 75 Black-necked Crane nests (2014) (see Fig. 1),

110 during breeding seasons. We used these field samples (nests) to represent species presence points

111 referenced in time and space.

112 Put Fig. 1 here

113 Figure 1 Study areas for three species cranes.

Environmental variables

115   We used 21 environmental layers at a 30-s spatial resolution in a GIS format and that were known

116 to correlate with bird distribution and as proxies of habitats predictors. They included bio-

117 climatic factors (bio_1-7, bio_12-15), topographical factors (altitude, slope, and aspect), water

118 factors (distance to river, distance to lake, and distance to coastline), inference factors (distance to

119 road, distance to rail road, and distance to settlements), and land cover factors (for detailed

120 information, see Table 1). Most of these predictors were obtained from open access sources. Bio-

121 climate factors were obtained from the WorldClim database (http://www.worldclim.org), while

122 aspect and slope layer were derived from the altitude layer in ArcGIS, which was also initially

123 obtained from the WorldClim database. Road, railroad, river, lake and coastline and settlement

124 maps were obtained from the Natural Earth database (http://www.naturalearthdata.com). The land

125 cover map was obtained from the ESA database. We also made models with all 19 bio-climate

126   variables and 10 other environmental variables, and then reduced predictors by AIC, BIC,

127   varclust, PCA and FA analysis. When we compared the distribution maps overlaying with

128   independent data set generated by Random Forest model, we found the model based on 21

129   predictors have the best performance for Hooded Cranes, and the best level for White-naped

130   Crane and Black-necked Cranes (see Supplement S1). Therefore, we decided to constructed

131   models with 21 predictors for the all three cranes and four machine-learning techniques. All

132   spatial layers of these environmental variables were resampled in ArcGIS to a resolution of 30-s

133   to correspond to that of the bioclimatic variables and for a meaningful high-resolution

134   management scale.

135   Put Table 1 here

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

Environmental Layers	Description	Source	Website
Bio_1	Annual mean Temperature ($\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_2	Monthly mean (max temp - min temp) ($\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_3	Isothermality (BIO2/BIO7) (*100 $\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_4	Temperature seasonality (standard deviation *100 $\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_5	Max temperature of warmest month ($\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_6	Min temperature of Coldest month ($\mathrm{^\circ C}$)	WorldClim	http://www.worldclim.org/
Bio_7	Annual temperature range (BIO5-BIO6) ($\mathrm{^\circ C}$)	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	WorldClim	http://www.worldclim.org/

	month (mm)		
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 layer from Natural Earth	http://www.naturalearthdata.com/
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

138   We created TreeNet, Random Forest, CART, Maxent models and ensemble models (averaged

139   value of the former four model results) for Hooded Cranes, White-naped Cranes and Black-naped

140   Cranes. These four model algorithms are considered to be among the best performing machine

141   learning methods (more information about these four models can be seen in the references by

142   Breiman et al., 1984, Breiman 2001, Friedman 2002, Phillips et al., 2004, Hegel et al., 2010). The

143   first three machine learning models are binary (presence-pseudo absence) models and were

144   handled in Salford Predictive Modeler 7.0 (SPM). For more details on TreeNet, Random Forest

145   and CART in SPM and their performances, we refer readers to the user guide document online

146   (https://www.salford-systems.com/products/spm/userguide). Several implementations of these

147   algorithms exist. Approximately 10,000 'pseudo-absence' locations were selected by random

148   sampling across the study area for each species using the freely available Geospatial Modeling

149   Environment (GME; Hawth's Tools; Beyer 2013; see Booms et al., 2010 and Ohse et al., 2009 for

150   examples). We extracted the habitat information from the environmental layers for presence and

151   pseudo-absence points for each crane, and then constructed models in SPM with these data. In

152   addition, we used balanced class weights, and 1000 trees were built for all models to find an

153   optimum within, others used default settings.

154   For the predictions, we created a 'lattice' (equally spaced points across the study area;

155   approximately 5×5 km spacing for the study area). For the lattice, we extracted information from

156   the same environmental layers (Table 1) as described above for each point and then used the

157   model to predict ('score') bird presence for each of the regular lattice points. For visualization,

158   we imported the dataset of spatially referenced predictions ('score file') into GIS as a raster file

159   and interpolated for visual purposes between the regular points using inverse distance weighting

160   (IDW) to obtain a smoothed predictive map of all pixels for the breeding distributions of the three

161   cranes (as performed in Ohse et al., 2009 and Booms et al., 2010). The fourth algorithm we

162   employed, Maxent, is commonly referred to as a presence-only model; we used Maxent 3.3.3k (it

163   can be downloaded for free from http://www.cs.princeton.edu/~schapire/maxent/) to construct our

164   models. To run Maxent, we followed the 3.3.3e tutorial for ArcGIS 10 (Young et al., 2011) and
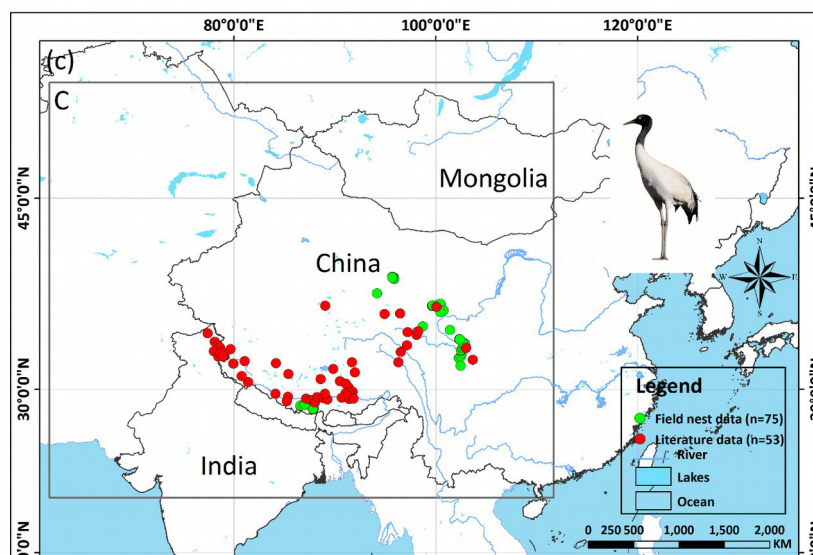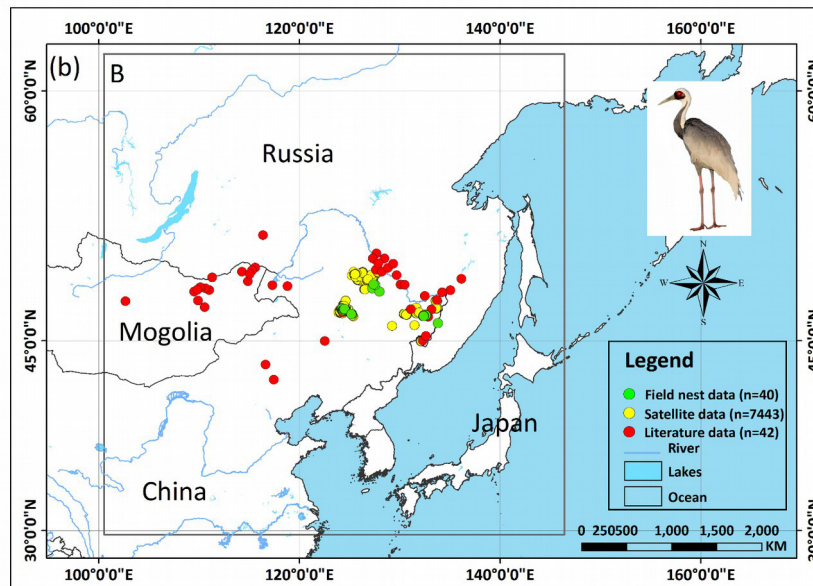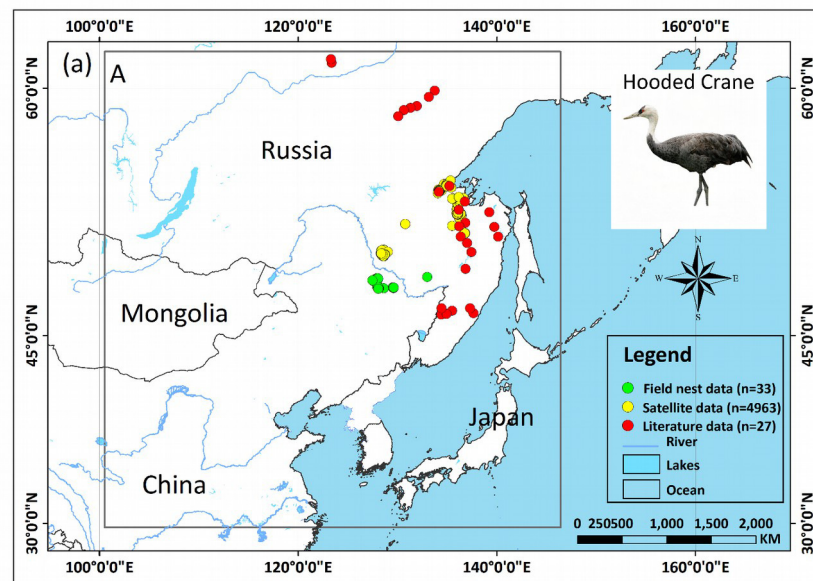
165   used default settings.

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). Here we digitized the only available 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 pseudo-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

190   distributional modeling (Manel *et al.,* 2001, McPherson *et al.,* 2004). TSS was also used to

191   evaluate model performance; we used TSS because it has been increasingly applied as a simple

192   but robust and intuitive measure of the performance of species distribution models (Allouche et

193   al., 2006).

194   Put Fig.2 here

195   Figure 2 Detailed study areas showing the presence of and testing data used for the three cranes.

196   2a) Hooded Cranes, 2b) White-naped Cranes, 2c) Black-necked Cranes.

197   To assess models transferability, we extracted the predictive value of the relative index of

198   occurrence (RIO) for testing data sets from the prediction maps using GME. We then constructed

199   resulting violin plots in R for these extracted RIOs to visualize their one-dimensional distribution.

200   This method allowed us to examine the degree of generalizability based on the local area with

201   samples to predict into undersampled areas that are otherwise unsampled in the model

202   development (=areas without training data). In addition, AUC is also commonly used to assess

203   model transferability in our study referring Randin et al. (2006).

RESULTS

Model performance

206   The results for AUC and TSS, two metrics commonly used to evaluate model accuracy, are

207   listed in Table 2. For the four SDMs technique, our results showed that the AUC values for

208   Random Forest were always highest (>0.625), ranking this model in first place, followed by

209   Maxent (>0.558), and then either CART or TreeNet (>=0.500). TSS showed us consistent results,

210   as was the case for AUC, and Random Forest performed the best (>0.250) followed by Maxent

211   (>0.137) for all three crane species, CART took the third place for Black-necked Cranes, and

212   TreeNet performed better than CART for White-naped Cranes. And the results showed there was

213   a trend that the value of these three metrics increased with an increase of nest site samples (33 to

214   75, Hooded Crane to Black-necked Crane, see Table. 2). Comparing the results of Random Forest

215   with ensemble models, we found their performance were close. Random Forest obtained better

216   models for Hooded Cranes and White-naped Cranes cases, the ensemble model performed better

217   for Black-necked Cranes.

218   Put Table 2 here

219   Table 2 AUC and TSS values for four machine learning models and their ensemble model with

220   three crane species based on literature testing data.

Accuracy metric	Species distribution model				
(samples)	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.
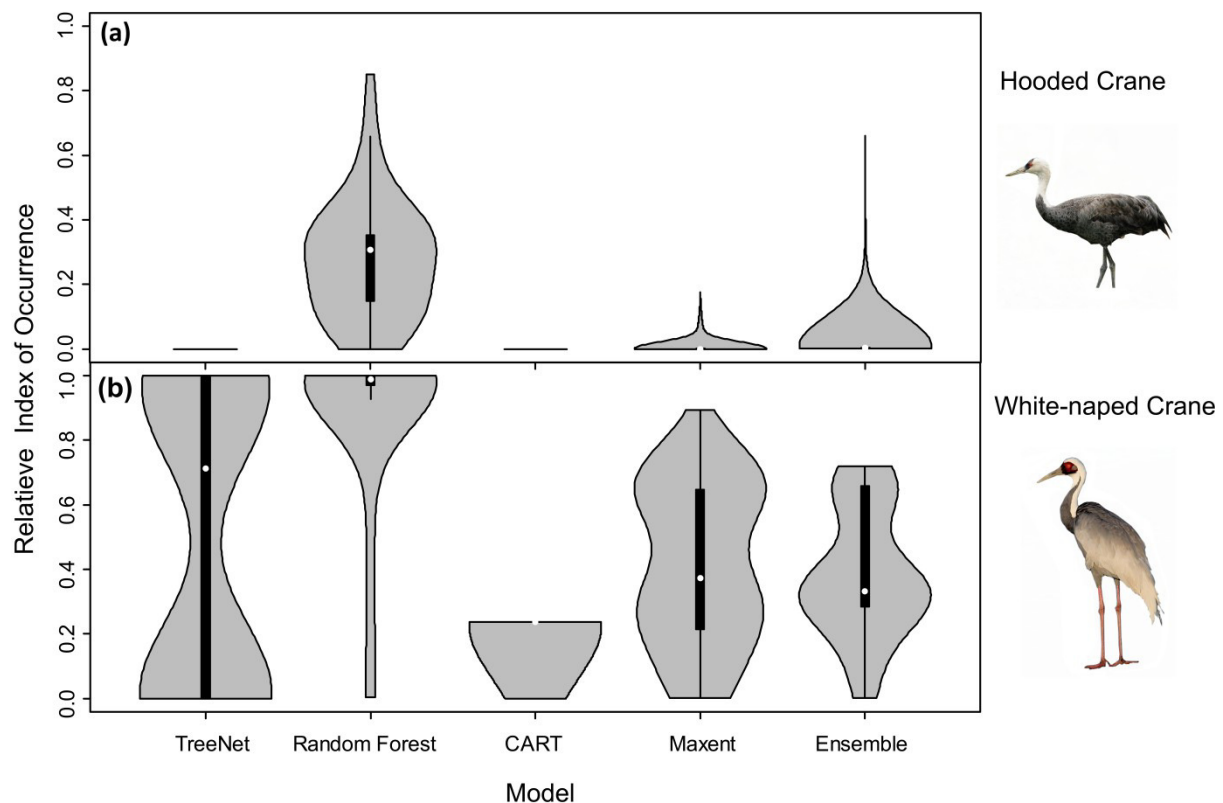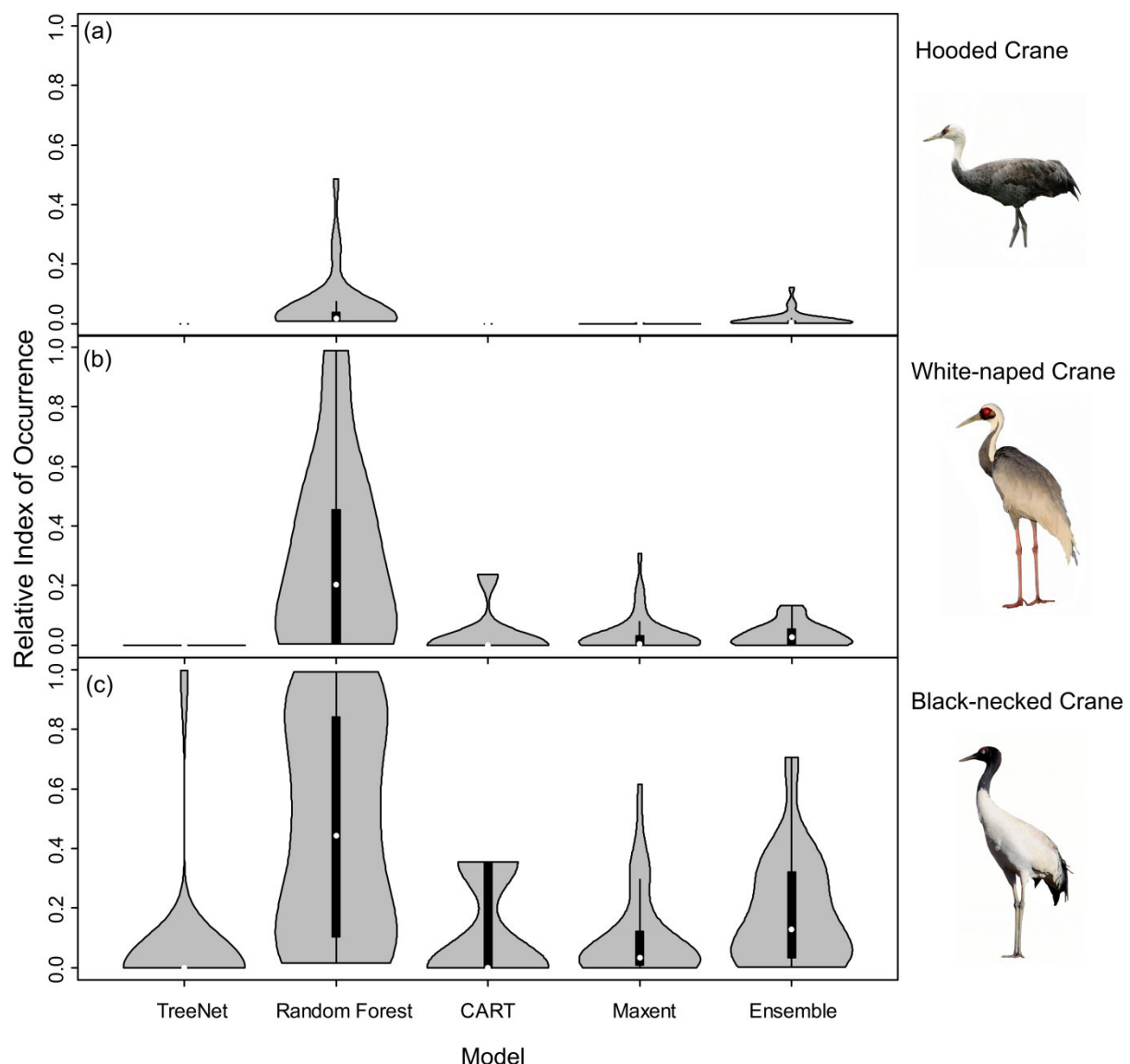
Put Fig. 3 here

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 the ensemble model and then Maxent. The best performer was still Random Forest, and its RIOs were distributed evenly to a certain extent with a median value

PeerJ

246 of 0.44. The results of AUC, as mentioned in "Model performance" part (Table 2), showed

247 consistent results with violin plots, Random Forest always get the highest value and has the best

248 generalization.

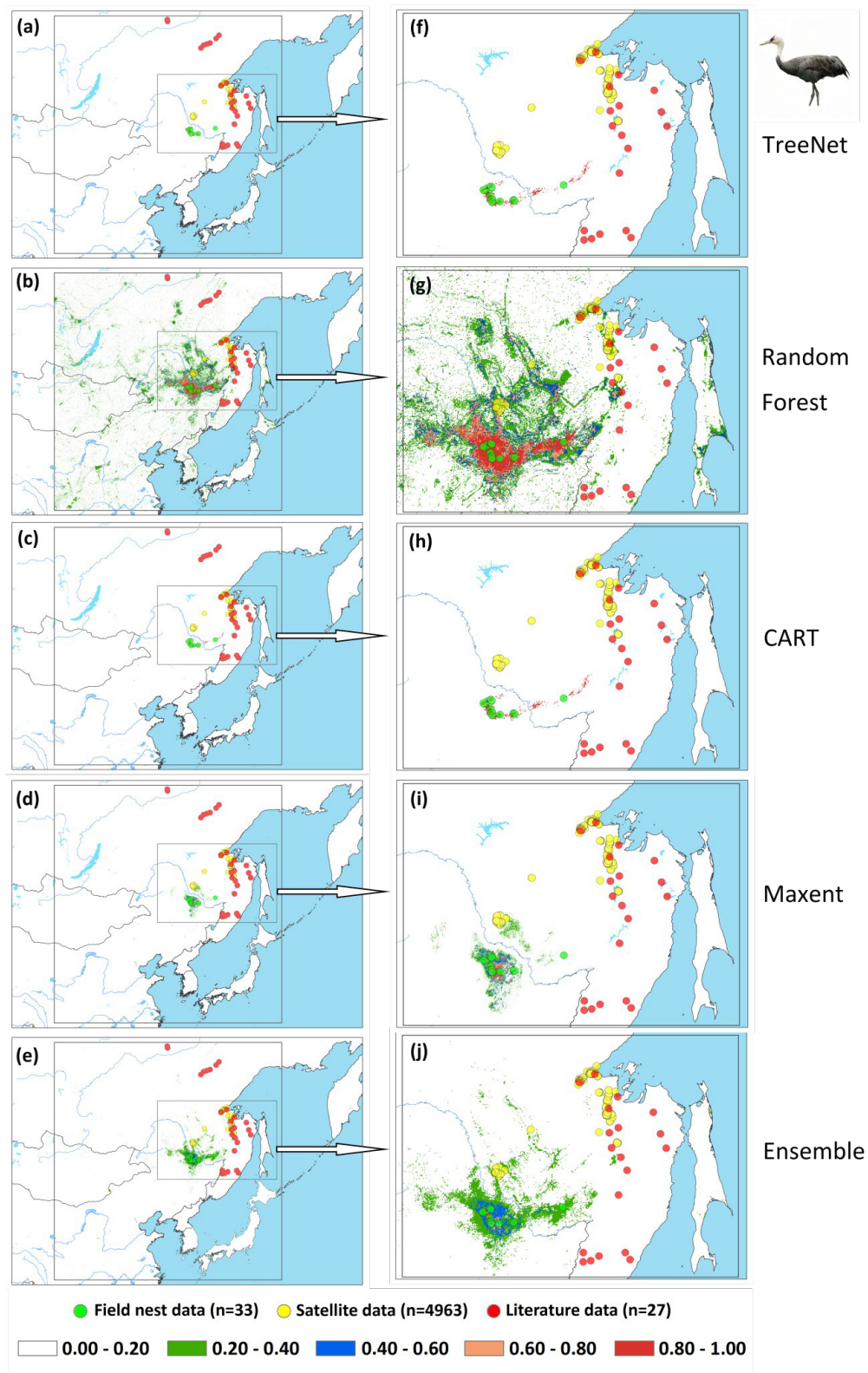249 Put figure 4 here

250 Figure 4 Violin plots of Relative Index of Occurrence (RIO) values for four SDMs and ensemble

251 model for three cranes based on calibration data from Threatened Birds of Asia. 4a) Violin plots

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

253 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, as expected and known, that a higher sample size makes models more robust and better to generalize from.

Put Fig. 5 Here

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

274   Cranes, 5f-5j) zoomed-in map for Hooded Cranes.

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

276   Random Forest, CART and Maxent) and ensemble model in detail. 6a-6e) prediction map for White-naped

277   Cranes, 6f-6j) zoomed-in map for White-naped Cranes.

278   Put Fig. 6 Here

Field nest data (n=75) Literature data (n=53)

| 0.00 - 0.20 | 0.20 - 0.40 | 0.40 - 0.60 | 0.60 - 0.80 | 0.80 - 1.00 |

279 Figure 7 Prediction maps for Black-necked Cranes and zoomed-in maps showing the four models

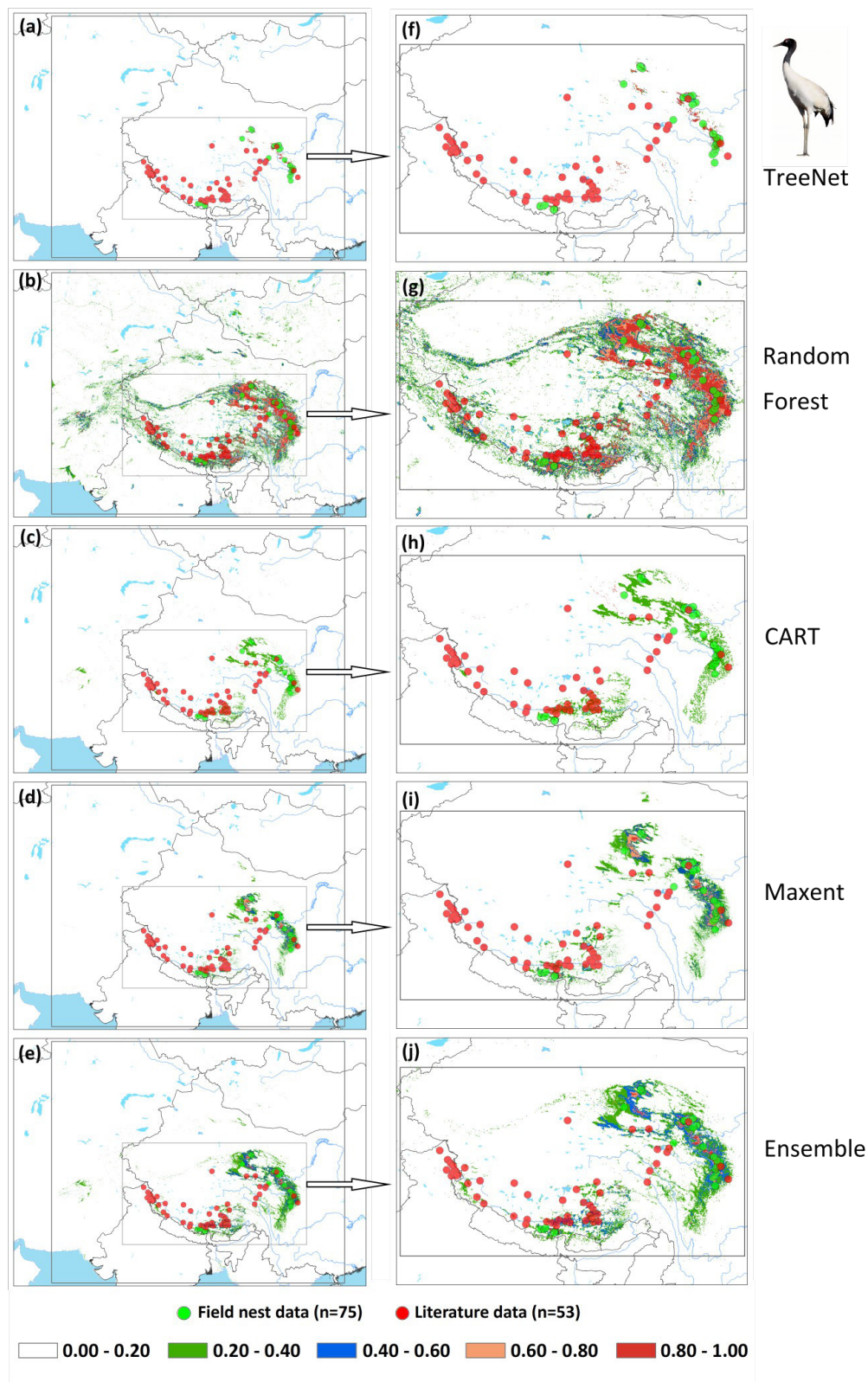280 (TreeNet, Random Forest, CART and Maxent) and ensemble model in detail. 7a-7e) prediction map for

281 Black-necked Cranes, 7f-7j) zoomed-in map for Black-necked Cranes.

282 DISCUSSION

283 Model generality (transferability)

284 Estimating species distributions in undersampled areas is a fundamental problem in ecology,

285 biogeography, biodiversity conservation and natural resource management (Drew et al., 2011).

286 That is specifically true for rare and difficult to be detected species and which are usually high on

287 the conservation priority. The use of SDMs and with machine learning has become the method

288 for deriving such estimates (Guisan & Thuiller, 2005; Drew et al., 2011; Guisan et al., 2013) and

289 could contribute to detect new and to confirm populations of rare species. However, the

290 application of a few samples to project a distribution area widely beyond the sample range is a

291 greater challenge and has rarely been attempted in the literature until recently. And only now

292 have conservationists realized its substantial value for pro-active decision making in conservation

293 management (see work by Ohse et al., 2010; Drew et al., 2011; Kandel et al., 2015 etc.). Our

294 results based on AUC, violin plots for RIOs and spatial assessment of testing data (satellite

295 tracking data and literature data) all suggest there are difference in the generalization

296 performance of different modeling techniques (TreeNet, Random Forest , CART and Maxent).

297 Moreover, among the acknowledged four rather powerful and commonly used machne-

298 learning techniques, Random Forest (bagging) in SPM usually had the best performance in each

299 case. Our results are in agreement with those of Prasad et al. (2006), Cutler et al. (2007) and

300 Syphard and Franklin (2009) indicating a superiority of Random Forest in such applications.

301 However, initially it appears to run counter to the conclusions off recent paper (Heikkinen et al.,

302 2012) with the poor transferability of Random Forest. But we propose this is due to the fact that

303 many Random Forest implementations exist (see the 100 classifier paper Fernández-Delgado et

304 al., 2014).

305   Here we applied Random Forest in SPM which has been optimized under one of the

306   algorithm's original co-authors, while Heikkinen et al. (2012) run a basic Random Forest with

307   BIOMOD framework in the R sofeware and which remains widely un-tuned and largely behind

308   the potential. The differences are known to be rather big (see Herrick 2013 for a comparison).

309   Furthermore, Maxent, a widely used SDM-method consequently greatly enjoyed by many

310   modelers (Phillips et al., 2006; Peterson et al. 2007; Phillips and Dudík 2008; Li et al., 2015,

311   etc.), didn't perform so good in regards to transferability in this study. This contrasts to those of

312   Elith et al. (2006) and Heikkinen et al. (2012), where Maxent and GBM perform well. We infer

313   this may be caused by sample size used as training data and due to the actual algorithms used.

314   When the sample size increased (33 to 75), the AUC and TSS value of all models rose (Table 2).

315   This indicates that higher sample sizes make models more robust and performing better. Sample

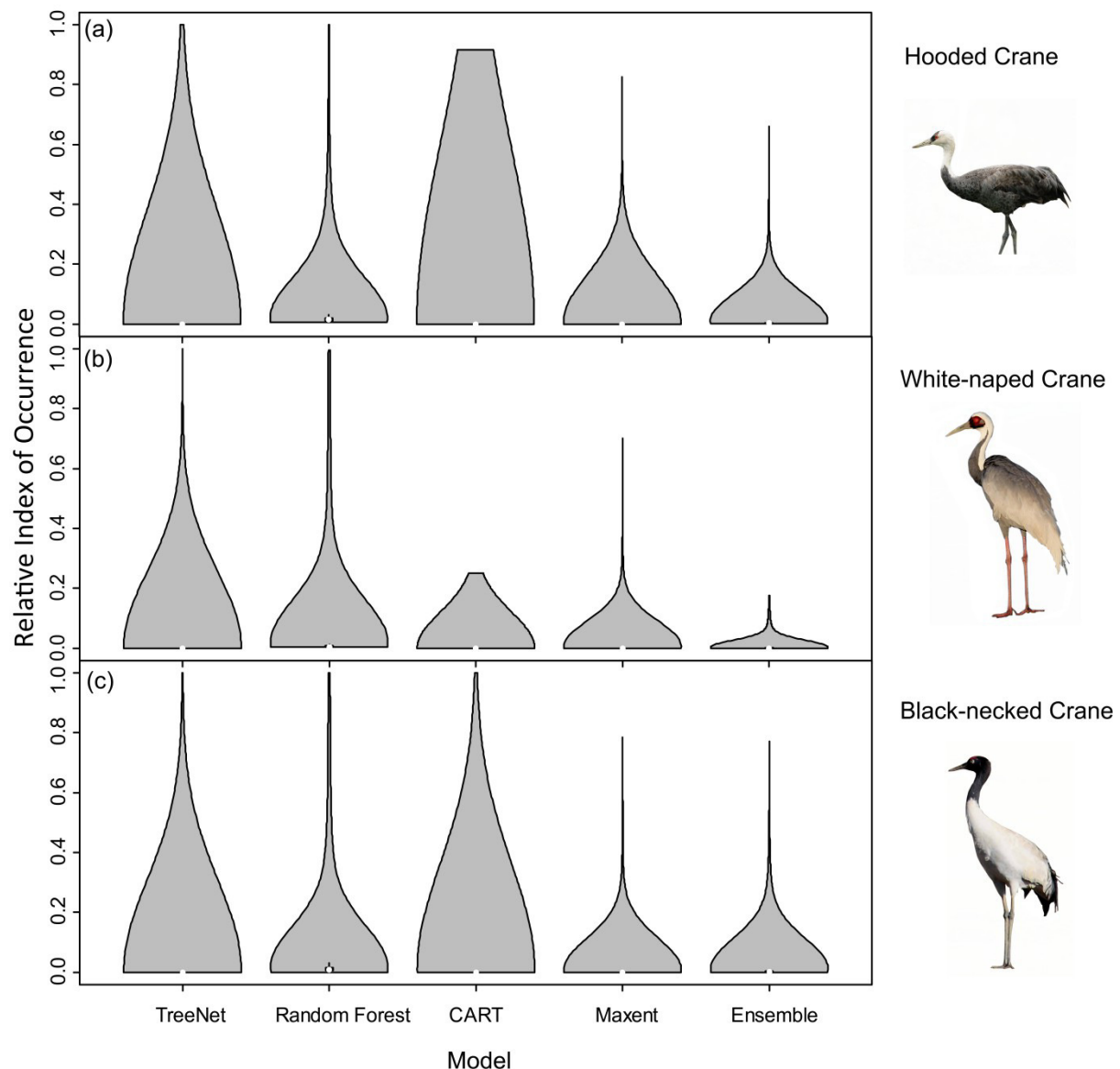316   sizes of 33 presence points still favor by Random Forest.

317   In Random Forest, random samples from rows and predictors are used to build hundreds of

318   trees. Each individual tree is constructed from a bootstrap sample and split at each node by the

319   best predictor from a very small, randomly chosen subset of the predictor variable pool (Herrick,

320   2013). These trees comprising the forest are each grown to maximal depth, and predictions are

321   made by averaged trees through 'voting' (Breiman et al., 2006). This algorithm avoids overfitting

322   by controlling the number of predictors randomly used at each split, using means of out-of-bag

323   (OOB) samples to calculate an unbiased error rate. And also, Random Forest in SPM utilizes

324   additional specific fine-tuning for best performance.

325   RIOs of random points

326   In order to explore whether Random Forest created higher RIOs for prediction maps in each grid,

327   which would result into higher RIOs of testing data, we generated 3,000 random points for

328   Hooded Cranes and White-naped Cranes, 5000 random points for Black-necked Cranes in their

329   related projected study areas. We made violin plots for RIOs of random points (Fig. 8), and we

330   found that more RIO values of random points for Maxent, Random Forest and ensemble models

331   were close to the lower value, and then followed by TreeNet. The distribution shapes of Random

332   Forest, Maxent and ensemble model are more similar to the real distribution of species in the real

333   world. The RIOs of White-naped Crane extracted from the CART model distributed in the range

334   of the low value. That means there were no points located in the high RIO areas of cranes, and

335   which is unrealistic. Consequently, we argued that Random Forest did not create higher RIOs for

336   prediction maps in each grid in our study.

337   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 typically have limited number of available occurrence records, which

345   poses challenges for the creation of accurate species distribution models when compared with

346   models developed with greater numbers of occurrences (Stockwell & Peterson, 2002; McPherson

347   et al., 2004; Hernandez et al., 2006). In this study, we used three crane species as case studies,

348   and their occurrence records (nests) totaled 33, 40, and 75, respectively (considering the small

349   numbers of samples and given that a low fraction of the area was sampled in the large projected

350   area). In our models, we found that model fit (AUC and TSS, see Table 2) of Random Forest that

351   had the highest index, while Maxent usually ranked second. In addition, we found that models

352   with few presence samples can also generate accurate species predictive distributions (Fig. 3 to 7)

353   with the Random Forest method. Of course, models constructed with few samples underlie the

354   threat of being biased more because few samples usually had not enough information including

355   all distribution gradients conditions of a species, especially for places far away from the location

356   of training presence points. However, the potential distribution area predicted by SDMs could

357   become the place where scholars could look for the birds (additional fieldwork sampling). And

358   also, these places could be used as diffusion or reintroduction areas! It's valuable and new

359   information either way.

360   Evaluation methods

361   In this study, we applied two widely-used assessment methods (AUC and TSS) in SDMs (Table

362   2). For an evaluation of these three values we used the approach recommended by Fielding &

363   Bell (1997), and Allouche et al. (2006), we found our model usually didn't obtain perfect

364   performance, and some of them were 'fair' in their performance. However, for macro-ecology

365   this is more than reasonable and ranks rather high. It's good conservation progress! We identified

366   Random Forest as always the highest performing. These results are consistent with the results of

367   violin plots of the Relative Index of Occurrence (RIO) using tracking as well as literature data

368   (Figs 3 and 4), and well as matching the spatial assessment results (Figs 5-7). And we

369   recommend when modelers assess model performance they should not only depend alone on

370   some metric (such as AUC and TSS), but also should base their assessments on the combined use

371   of visualization and expert knowledge. That means modelers should also assess how the species

372   distribution map actually looks and how it links with real data (see Huettmann & Gottschalk

373   2011). Spatial assessment metrics from alternative data should matter the most. Expert experience

374   and ecological common knowledge of the species of interest could sometimes also be highly

375   effective (Drew & Perera, 2011), albeit nonstandard, evaluation methods (see Kandel et al., 2015

376   for an example). Additionally, one alternative method for rapid assessment we find is to use a

377   reliable SDM, and thus Random Forest would be a good choice in the future given our consistent

378   results (Fig.s 3 to 7, Tables 3 to 5) in this study, which involved three species, a vast landscape to

379   conserve, and only limited data. Our work certainly helps to inform conservation decisions for

380   cranes in Northeast Asia.

381   Limitations and future work

382   Our study is not without limitations: 1) so far, only three species of cranes are used as a test case

383   in our study. That's because nest data for rare species in remote areas are usually sparse; 2) all our

384   species study areas are rather vast and confined to East-Asia. For future work, we would apply

385   Random Forest in more species and in more geographical conditions with differently distributed

386   features for a first rapid assessment and baseline to be mandatory for better conservation e,g, by

387   governments, IUCN and any impact and court decision. Then we would apply our prediction

388   results in specifically targeted fieldwork sampling campaigns and assess the model accuracy with

389   field survey results (ground-truthing) and with new satellite tracking and drone data, for instance.

390   This is to be fed directly into the conservation management process.

391   ACKNOWLEDGEMENTS

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.

422   Cohen J. 1960. A Coefficient of Agreement for Nominal Scales. Educational and Psychological

423   Measurement 20(1):37-46.

424   Collar NJ, Crosby R, Crosby M. 2001. Threatened birds of Asia: the BirdLife International red

425   data book. Volume 1.BirdLife International Cambridge, UK.

426   Cushman SA, Huettmann F. 2010. Spatial Complexity, Informatics, and Wildlife Conservation.

427   Springer, Springer Tokyo Berlin Heidelberg New York.

428   Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random

429   forests for classification in ecology. Ecology 88(11):2783-2792.

430   Drew CA, Perera AH. 2011. Expert knowledge as a basis for landscape ecological predictive

431   models. Pages 229-248 in Predictive Species and Habitat Modeling in Landscape Ecology.

432   Springer.

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

434   ecology. Springer.

435   Elith J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F,

436   Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura

437   M, Nakazawa Y, Overton JMM, Peterson AT, Phillips SJ, Richardson K, Scachetti-Pereira R,

438   Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE. 2006. Novel methods

439   improve prediction of species' distributions from occurrence data. Ecography 29(2):129-151.

440   Eskildsen, A., P. C. Roux, R. K. Heikkinen, T. T. Høye, W. D. Kissling, J. Pöyry, M. S. Wisz, and

441   M. Luoto. 2013. Testing species distribution models across space and time: high latitude

442   butterflies and recent warming. Global ecology and biogeography 22(12):1293-1303.

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

444   W. 2013. Comparing mechanistic and empirical model projections of crop suitability and

445   productivity: implications for ecological forecasting. Global ecology and biogeography

446   22(8):1007-1018.

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

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

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

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

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

460   Friedman JH. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis
461   38(4):367-378.

462   Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT. 2004. New developments in museum-
463   based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution
464   19(9):497-503.

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

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

470   Guisan A, Tingley R, Baumgartner JB, Naujokaitis□Lewis I, Sutcliffe PR, Tulloch AIT, Regan
471   TJ, Brotons L, Mcdonald□Madden E, Mantyka□Pringle C. 2013. Predicting species

472   distributions for conservation decisions. Ecology letters 16(12):1424-1435.

473   Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological

474   Modelling 135(2):147-186.

475   Gwena LL, Robin E, Erika F, Guisan A. 2010. Prospective sampling based on model ensembles

476   improves the detection of rare species. Ecography 33(6):1015-1027.

477   Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating

478   characteristic (ROC) curve. Radiology 143(1):29-36.

479   Hanley JA, McNeil BJ. 1983. A method of comparing the areas under receiver operating

480   characteristic curves derived from the same cases. Radiology 148(3):839-843.

481   Hardy SM, Lindgren M., Konakanchi H, Huettmann F. 2011. Predicting the distribution and

482   ecological niche of unexploited snow crab (Chionoecetes opilio) populations in Alaskan

483   waters: a first open-access ensemble model. Integrative and comparative biology 51(4):608-

484   622.

485   Hegel TM, SA Cushman, J Evans, Huettmann F. 2010. Current State of the Art for Statistical

486   Modelling of Species Distributions. Spatial Complexity, Informatics, and Wildlife

487   Conservation:273-311.

488   Heikkinen RK, Marmion M, Luoto M. 2012. Does the interpolation accuracy of species

489   distribution models come at the expense of transferability? Ecography 35(3):276-288.

490   Hernandez PA, Graham CH, Master LL, Albert DL. 2006. The effect of sample size and species

491   characteristics on performance of different species distribution modeling methods. Ecography

492   29(5):773-785.

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

494   Hijmans RJ, Graham CH. 2006. The ability of climate envelope models to predict the effect of

495   climate change on species distributions. Global Change Biology 12(12):2272-2281.

496   Huettmann F, Gottschalk T. 2011. Simplicity, Model Fit, Complexity and Uncertainty in Spatial

497   Prediction Models Applied Over Time: We Are Quite Sure, Aren't We? Pages 189-208 in

498   Predictive Species and Habitat Modeling in Landscape Ecology. Springer.

499   Humphries GRW, Huettmann F. 2014. Putting models to a good use: a rapid assessment of Arctic

500   seabird biodiversity indicates potential conflicts with shipping lanes and human activity.

501   Diversity and Distributions 20(4):478-490.

502   Jiguet F, Barbet-Massin M, Chevallier D. 2011. Predictive distribution models applied to satellite

503   tracks: modelling the western African winter range of European migrant Black Storks Ciconia

504   nigra. Journal of Ornithology 152(1):111-118.

505   Kandel K, Huettmann F, Suwal MK, Regmi GR, Nijman V, Nekaris K, Lama ST, Thapa A,

506   Sharma HP, Subedi TR. 2015. Rapid multi-nation distribution assessment of a charismatic

507   conservation species using open access ensemble model GIS predictions: Red panda (Ailurus

508   fulgens) in the Hindu-Kush Himalaya region. Biological Conservation 181:150-161.

509   Keith DA, Elith J, Simpson CC. 2014. Predicting distribution changes of a mire ecosystem under

510   future climates. Diversity and Distributions 20(4):440-454.

511   Kessler A, Batbayar N, Natsagdorj T, Batsuur D, Smith A. 2013. Satellite telemetry reveals

512   long□distance migration in the Asian great bustard Otis tarda dybowskii. Journal of Avian

513   Biology 44(4):311-320.

514   Kleyer M. 2002. Validation of plant functional types across two contrasting landscapes. Journal

515   of Vegetation Science 13(2):167-178.

516   Lei, Z., L. Shirong, S. Pengsen, and WangTongli. 2011. Comparative evaluation of multiple

517   models of the effects of climate change on the potential distribution of *Pinus massoniana*.

518   Chinese Journal of Plant Ecology 35(11):1091-1105.

519   Li, R., M. Xu, M. H. G. Wong, S. Qiu, X. Li, D. Ehrenfeld, and D. Li. 2015. Climate change

520   threatens giant panda protection in the 21st century. Biological Conservation 182:93-101.

521   Maggini R, Lehmann A, Zbinden N, Zimmermann NE, Bolliger J, Schröder B, Foppen R,

522   Schmid H, Beniston M, Jenni L. 2014. Assessing species vulnerability to climate and land use

523   change: the case of the Swiss breeding birds. Diversity and Distributions 20(6):708-719.

524   Manel S, Williams HC, Ormerod SJ. 2001. Evaluating presence–absence models in ecology: the

525   need to account for prevalence. Journal of Applied Ecology 38(5):921-931.

526   McPherson J, Jetz W, Rogers DJ. 2004. The effects of species' range sizes on the accuracy of

527   distribution models: ecological phenomenon or statistical artefact? Journal of Applied

528   Ecology 41(5):811-823.

529   Mi C, Huettmann. F, Guo Y. 2016. Climate envelope predictions indicate an enlarged suitable

530   wintering distribution for Great Bustards (Otis tarda dybowskii) in China for the 21st century.

531   PeerJ 4:e1630.

532   Mingchang C, Guangsheng Z, Ensheng W. 2005. Application and comparison of generalized

533   models and classification and regression tree in simulating trees species distribution. ACTA

534   ECOLOGICA SINICA 25(8):2031-2040.

535   Navarro□Cerrillo R, Hernández□Bermejo J, Hernández□Clemente R. 2011. Evaluating models

536   to assess the distribution of Buxus balearica in southern Spain. Applied Vegetation Science

537   14(2):256-267.

538   Ohse B, Huettmann F, Ickert-Bond SM, Juday GP. 2009. Modeling the distribution of white

539   spruce (Picea glauca) for Alaska with high accuracy: an open access role-model for predicting

540   tree species in last remaining wilderness areas. Polar biology 32(12):1717-1729.

541   Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT. 2007. Predicting species distributions

542   from small numbers of occurrence records: a test case using cryptic geckos in Madagascar.

543   Journal of Biogeography 34(1):102-117.

544   Peterson AT., Monica P, Muir E. 2007. Transferability and model evaluation in ecological niche

545   modeling: a comparison of GARP and Maxent. Ecography 30(4):550–560.

546   Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic

547   distributions. Ecological Modelling 190(3):231-259.

548   Phillips SJ, Dudík M. 2008. Modeling of species distributions with Maxent: new extensions and a

549   comprehensive evaluation. Ecography 31(2):161-175.

550   Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution

551   modeling. ACM, 2004.

552   Prasad AM, Iverson LR, Liaw A. 2006. Newer Classification and Regression Tree Techniques:

553   Bagging and Random Forests for Ecological Prediction. Ecosystems 9(2):181-199.

554   Randin CF, Dirnböck T, Dullinger S, Zimmermann NE, Zappa M, Guisan A. 2006. Are niche-

555   based species distribution models transferable in space? Journal of Biogeography

556   33(10):1689-1703.

557   Romo H, García-Barros E, Márquez AL, Moreno JC, Real R. 2014. Effects of climate change on

558   the distribution of ecologically interacting species: butterflies and their main food plants in

559   Spain. Ecography 37(11):1063-1072.

560   Stockwell DR, Peterson AT. 2002. Effects of sample size on accuracy of species distribution

561   models. Ecological Modelling 148(1):1-13.

562   Stokes KL, Broderick AC, Canbolat AF, Candan O, Fuller WJ, Glen F, Levy Y, Rees AF, Rilov G,

563   Snape RT, Stott I, Tchernov D, Godley BJ. 2015. Migratory corridors and foraging hotspots:

564   critical habitats identified for Mediterranean green turtles. Diversity and Distributions 21(6):

565   665-674.

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

567   Syphard DA, Franklin J. 2009. Differences in spatial predictions among species distribution

568   modeling methods vary with species traits and environmental predictors. Ecography

569   32(6):907-918.

570   Thomas JA, Bovee KD. 1993. Application and testing of a procedure to evaluate transferability of

571   habitat suitability criteria. Regulated rivers 8:285-285.

572 Thuiller W. 2003. BIOMOD–optimizing predictions of species distributions and projecting

573   potential future shifts under global change. Global Change Biology 9(10):1353-1362.

574 Zhai T, Li X. 2012. Climate change induced potential range shift of the crested ibis based on

575   ensemble models. ACTA ECOLOGICA SINICA 32(8):2361-2370 (in Chinese).

576 Williams JN, Seo C, Thorne J, Nelson JK, Erwin S, O'Brien JM, Schwartz MW. 2009. Using

577   species distribution models to predict new occurrences for rare plants. Diversity and

578   Distributions 15(4):565-576.

579 Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A. 2008. Effects of sample size on

580   the performance of species distribution models. Diversity and Distributions 14(5):763-773.

581 Yen P, Huettmann F, Cooke F. 2004. A large-scale model for the at-sea distribution and

582   abundance of Marbled Murrelets (Brachyramphus marmoratus) during the breeding season in

583   coastal British Columbia, Canada. Ecological Modelling 171(4):395-413.

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

585 Zhang M, Zhou Z, Chen W, Cannon CH, Raes N, Slik JWF. 2014. Major declines of woody plant

586 species ranges under climate change in Yunnan, China. Diversity and Distributions 20(4):405-

587 415.