# How to assess species distribution model accuracy: using internal-aspatial or external-spatial methods?

**Chunrong Mi** [1, 2, 3] , **Falk Huettmann** [4] , **Yumin Guo** [Corresp. 3]

[1] Chinese Academy of Science, Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Beijing, China
[2] University of Chinese Academy of Science, Beijing, Beijing, China
[3] Beijing Forestry University, College of Nature Conservation, Beijing, China
[4] University of Alaska Fairbanks (UAF), EWHALE Lab- Department of Biology and Wildlife, Institute of Arctic Biology, Fairbanks, Alaska, USA

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

Species distribution models (SDMs) have become an increasingly important tool in ecology, biogeography, evolution and, more recently, in conservation management, landscape planning and climate change research. The assessment of their predictive accuracy is one fundamental issue in the development and application of SDMs. Accuracy assessments for models should have a close connection to the intended use of the model. However, we found that the common evaluation method (we named internal-aspatial) usually ignored how the spatial prediction map actually looks like, and achieves for the real-world species distribution and for application. Therefore, in this research we proposed a spatial method to evaluate model performance by assessing how the prediction maps look like (we named external-spatial). We took Hooded Crane (*Grus monacha*) as a case, in this research, to compare these two methods (internal-aspatial and external-spatial) performance. Both of the two methods were expressed with three commonly used SDM evaluation criteria (AUC, Kappa and TSS). In addition, model accuracy was also assessed via evaluating the prediction maps with knowledge of the study species and alternative occurrence data assistance. We used two popular data mining algorithms (Random Forest and TreeNet) and ran 8 experiments using 1, 3, 5, 8, 11, 21, 29 and 78 predictors, allowing to develop overall 16 models for this assessment. Results indicated that AUC had a significant linear relationship with Kappa and TSS. Both of interal-aspatial and external-spatial methods could get higher AUC values and they were close. This indicated that internal-aspatial model assessments can serve as powerful assessment-aspatiual metrics without the need of secondary data even! However, internal-aspatial, external-spatial, prediction map evaluation and alternative occurrence data could not distinguish well models with different sets of predictors. This is the first time the concept of spatial assessment criteria is expressed and assessed. Overall, we hope to see more study on meaningful spatial criteria and proposed more and better methods to evaluate SDMs and

distribution map in the future.

# How to assess species distribution model accuracy: using internal-aspatial or external-spatial methods?

Chunrong Mi[1,2,3], Falk Huettmann[4], Yumin Guo[3]

[1]Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[2]University of Chinese Academy of Sciences, Beijing, China

[3]College of Nature Conservation, Beijing Forestry University, Beijing 100083, China

[4]EWHALE Lab- Department of Biology and Wildlife, Institute of Arctic Biology, University of Alaska Fairbanks (UAF), 419 Irving I, Fairbanks Alaska 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

# ABSTRACT

Species distribution models (SDMs) have become an increasingly important tool in ecology, biogeography, evolution and, more recently, in conservation management, landscape planning and climate change research. The assessment of their predictive accuracy is one fundamental issue in the development and application of SDMs. Accuracy assessments for models should have a close connection to the intended use of the model. However, we found that the common evaluation method (we named internal-aspatial) usually ignored how the spatial prediction map actually looks like, and achieves for the real-world species distribution and for application. Therefore, in this research we proposed a spatial method to evaluate model performance by assessing how the prediction maps look like (we named external-spatial). We took Hooded Crane (*Grus monacha*) as a case, in this research, to compare these two methods (internal-aspatial and external-spatial) performance. Both of the two methods were expressed with three commonly used SDM evaluation criteria (AUC, Kappa and TSS). In addition, model accuracy was also assessed via evaluating the prediction maps with knowledge of the study species and alternative occurrence data assistance. We used two popular data mining algorithms (Random Forest and TreeNet) and ran 8 experiments using 1, 3, 5, 8, 11, 21, 29 and 78 predictors, allowing to develop overall 16 models for this assessment. Results indicated that AUC had a significant linear relationship with Kappa and TSS. Both of interal-aspatial and external-spatial methods could get higher AUC values and they were close. This indicated that internal-aspatial model assessments can serve as powerful assessment-aspatiual metrics without the need of secondary data even! However, internal-aspatial, external-spatial, prediction map evaluation and alternative occurrence data could not distinguish well models with different sets of predictors. This is the first time the concept of spatial assessment criteria is expressed and assessed. Overall, we hope to see more study on meaningful spatial criteria and proposed more and better methods to evaluate SDMs and distribution map in the future.

***Keywords:*** Species distribution models (SDMs); internal-aspatial; external-apatial; AUC; Kappa;

TSS; Random Forest; TreeNet (Boosting); Hooded Crane (*Grus monacha*)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27257v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

# INTRODUCTION

Species distribution models (SDMs) are well-established numerical spatial tools that combine field observations of species occurrence or abundance with environmental predictor variables (Guisan and Zimmermann, 2000; Elith and Leathwick, 2009; Drew et al., 2011). In recent years, predictive model of species distribution has become an increasingly important science-based management tool for policy (e.g. Drew et al., 2011; Mi et al., 2016). The trend of SDMs goes towards data mining of data heavy applications to address many wider and more holistic and interdisciplinary issues in ecology, biogeography, evolution and, more recently, in conservation management and climate change research, and usually gets done now on a global level but with a high resolution (Austin et al., 1990; Franklin, 1995; Peterson et al., 2002; Guisan and Thuiller, 2005; Wei et al., 2010). A wide variety of statistical and machine learning methods have been introduced, often in conjunction with geographic information systems (GIS), remote-sensing (Aspinall and Veitch, 1993; Franklin et al., 2000; Hegel et al., 2010).

Only slowly, a number of 'data mining' approaches for modeling data that contain non-linear and other complex and interacting dependencies have appeared now in the wildlife literature, too (Derrig and Francis, 2006; Cushman and Huettmann 2010; Mi et al 2014). Relatively new methods are based on data hungry networks, or ensembles and that can handle complex and even marginal data situations. Over 100 machine learning algorithms exist (Fernandez-Delgado et al. 2014). Some of them use statistical trees and include algorithms such as Random Forest (bagging; Breiman, 2001a), TreeNet (stochastic gradient boosting (Friedman, 2002)), and other methods (Araujo and New 2007; see Biomod2 package in R). Random Forest and TreeNet were used in this research, because model construction process was fast and convenient (Mi et al., 2014 and 2016), and also model prediction performed very well in similar investigations (Oppel et al., 2012, Mi et al., 2014), and they were also non-parametric. These two algorithms usually do not really require or make relevant *a priori* assumptions about the relationship between the response and predictor variables. It does not limit the number of predictor variables, and it is capable of uncovering the underlying structure of data that are non-additive, interacting or hierarchical in nature (Prasad et al., 2006; Hasti et al., 2009; Cushman and Huettmann, 2010; Drew et al., 2011).

While somewhat overlooked, the deeper and ecological assessment of predictive accuracy is one fundamental issue in machine learning and specifically in the development of species distribution models (Fielding and Bell, 1997; Pearce and Ferrier, 2000; Guisan and Thuiller, 2005; Allouche et al., 2006). A quantitative assessment of model performance assists in determining the suitability of the model for specific applications (Vaughan and Ormerod, 2005; Barry and Elith, 2006; Guisan et al., 2006). Model performance assessment can also provide a basis for comparing alternative modelling techniques (Loiselle et al., 2003; Segurado and Araujo, 2004; Pearson et al., 2006) and it enables the user to investigate how different properties of the data and/or the species affect the accuracy of predictive maps generated by the model (Kadmon et al., 2003; Segurado and Araujo, 2004; Reese et al., 2005; Seoane et al., 2005).

To assess model performance, we found that most scholars used evaluation criteria (e.g. the Area Under the ROC Curve (AUC), the Kappa Statistic (Kappa) and the True Skill Statistic(TSS)) were created by the software itself (e.g. Anderson and Gonzalez Jr, 2011; Elith et al., 2011). Typically, such tests are based on hold-out data, such as from boot strapping or jackknifing applied to the (large) training data. This creates unmapped (aspatial) models and then calculates metrics (Kappa, TSS). We named this: internal-aspatial method. However, this method ignored how the spatial prediction map actually looks like, and achieves for the real-world species distribution and for application. Actually, accuracy assessments for models should have a close connection to the intended use of the model (Fielding, 2002). One major role of species distribution models is to model complex ecology and to support an efficient conservation management, such as conservation planning, design reserve networks that maintain biodiversity (Guisan and Thuiller, 2005). Therefore, in this research we proposed a spatial method to evaluate model performance by assessing how the prediction maps look like. We named this: external-spatial method. This spatial metrics mean it mapped predictions first, and then used presence-absence points to overlay the created prediction map and get the relative index of occurrence (RIO), and then calculated accuracy metrics to obtain the estimate of accuracy, such as AUC, Kappa and TSS. The aim of this research was to explore which evaluation method was better, internal-aspatial or external-spatial? In addition, we also assess model accuracy using prediction map with experts' knowledge of target species' distribution, and alternative occurrence data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27257v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

# MATERIALS AND METHODS

## Study species put to a test

The Hooded Crane is listed as a vulnerable (VU) species in the IUCN Red List. This species breeds in Eastern Russia and Northeastern China (Guo, 2005; Simonov and Dahmer, 2008; Mi et al., 2018). Its global population is estimated to be 11,160 individuals (Birdlife international, 2014) and the population size is declining (IUCN, 2012). In recent years, more than 10,500 (~ 94%) Hooded Cranes winter in Izumi, Japan (Birdlife international, 2014). This presents a risk and therefore, it is badly needed to find suitable places and methods to disperse the Hooded Crane from Izumi in order to diversify and reduce the population density there and to minimize local risks. Otherwise it can for instance lead to epidemic diseases of birds and their population crashes, such as Avian Influenza (Mi et al., 2018). Thus, here we tried to construct a winter distribution model for Hooded Cranes and to see where they would stay and for making a conservation plan and obtain more management methods.

## Species occurrence data

The Hooded Crane winter occurrence data was collected from our own fieldwork, also using previously published literature in East Asia (Fig. 1). In general, the data were initially provided by location name. To ensure the exact position for a valid geo-referencing, we then searched the location using a map with coordinates, Google Earth and also consulted experts for confirmation. Overall, we obtained 112 data points that were observed for this species during 1980-2013. This compiled data represents the best available geo-referenced data set for Hooded Crane wintering in China we know (Supplement S1). Initially, we considered that the data points maybe overly dense in some locations (oversampled or cluster sampling); thus, we created a concentric buffer around a data point with a 2-km radius in ArcGIS 10.1 (Toolboxes/System Toolboxes/Analysis Tools/Proximity/Buffer). However, in our data, we did not find any overlap of the 2-km scale. Therefore, we continued to use the all the data points as intended.

## Environmental layers

The environmental predictor variables we used to develop models in this study describe climate, topography, terrain and human factors. We chose a set of 21 predictors (see Supplement S2) to develop models as the ordinary baseline model, which was often used in our previous research (e.g. Mi et al., 2017). Based on this step, we then added another 8 bio-climate layers (Supplement S2) to construct models with overall 29 predictors (Mi et al., 2016; Han et al., 2018). In addition, we tried to make models with the entire predictor set (78 predictors). The Salford Predictive Modeler (SPM) suite we applied can generate a variable importance ranking table from the obtained trees; here we chose the top 1, top 3, top 5, top 8 and top 11 predictor variables. For that initial approach, we refer readers to Harrell et al. (1996), who promote for multiple regressions that the variable (predictor) quantity should not exceed n/10 (n means sample size, in our case n=112) for multivariable regression models). In all, we created 8 Random Forest models and 8 TreeNet models. All data layers were publicly available and had a global-wide coverage (Supplement S2). We re-projected layers into WGS-1984 Mercator (in meters) and merged them for a study area coverage in ArcGIS. Slope and aspect layers were derived in ArcGIS from the DEM. We also calculated the Euclidean distance to road, railroad, river, lake, coastline, settlement using the Euclidean distance tool in ArcGIS 10.1. Layers and raw data can be obtained from the College of Nature Conservation, Beijing Forestry University upon request to the autuors.

Put Figure 1 Here

## Selection of model algorithms

In SPM, we choose Random Forest (hereafter RF) and TreeNet (hereafter TN) as our species distribution models. We used them as a set of representative algorithms for the wider machine learning (ML) family of methods. RF and TN are specific stand-alone software products from Salford Systems Ltd that can outcompete R implementations (Herrick 2013), and each performs one specific technique. Here we used them as representative ML methods because when using these algorithms, model construction is fast and convenient, they offer a very high degree of fault tolerance for messy and incomplete data (Friedman, 2001; Craig and Huettmann 2008, Mi et al., 2014; Jiao et al. 2016). For more details on Random Forest and TreeNet, we refer readers to the user guide (https://www.salford-systems.com/products/spm/userguide).

## Model development

We created 10,000 random points across the study area using the freely available Geospatial Modeling Environment (GME; Beyer, 2013; Booms et al., 2010) and compared these points to the 112 bird locations. The ratio of 112 presence points versus 10,000 pseudo-absence points is commonly used in the machine learning modeling literature (Booms et al., 2010, Mi et al., 2016) and even more so when it comes to data mining (Hastie et al. 2009, Jiao et al. 2016) We extracted information from environmental layers at bird location sites and random points using GME.

We generally used the powerful default settings in SPM (e.g. Mi et al., 2014). Our distribution models were constructed in SPM by using 'classification' and the balanced class weights option to account for unequal sample sizes of presence and availability (pseudo-absence). For the predictions, we created equally-spaced point lattice grids of 1,047,746 regularly spaced points across our study area (approximately a 5×5 km spacing for the study area). We extracted information from the environmental layers (Supplement S2) described above for each point, and then used the model to predict (='score') birds occurrence as a relative index of occurrence at each lattice point based on the extracted environmental data. For visualization, we imported the dataset of spatially referenced predictions 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, as it is commonly done (e.g. Kandel et al. 2015, Regmi et al. 2018). We used that resulting prediction surface for our spatial assessment and comparison.

## Evaluation criteria

In this study, we obtained three metrics: AUC (the area under the ROC Receiver Operator Curve), the Kappa Statistic (Kappa) and the True Skill Statistic (TSS). They were among the most popular measures that are commonly used to assess the accuracy of prediction models (Fielding and Bell, 1997; Pearce and Ferrier, 2000; Manel et al., 2001; Pearson et al., 2004; Huettmann and Gottschalk, 2011; Liu et al., 2013). They are based on the confusion matrix (Fielding and Bell, 1997; Pearce and Ferrier, 2000), and AUC is more generally applied whereas Kappa and TSS offer specific advantages in terms of prevalence (Manel et al., 2001). None of those metrics take the spatial distribution, arrangement or autocorrelation of the points into account though (Betts et al. 2009). We obtained these three criteria using two methods (internal-aspatial and external-spatial). When SPM creates the model, it offers internal-aspatial AUC value, and a threshold table which could be transformed as a subsequent confusion matrix table, then we calculated internal–aspatial Kappa and TSS in R 3.0.3 according to the formula in Supplement S3. For the spatial metric, we extracted the relative index of occurrence (RIO) for all of the presence and pseudo-absence points from each model prediction map with GME software. Using the above RIO and the "SDMTools" package in R 3.0.3, we could obtain external-spatial AUC, Kappa, and TSS. Kappa and TSS values can be shown with different thresholds (0-1, interval 0.01), in this research, we used the maximum value (max-Kappa, max-TSS) as the final metric.

## Prediction map assessment

Model prediction maps were evaluated by the 'true' distribution of Hooded Cranes in winter, as we know them from our own field experience. We ranked the prediction map based on following reasons:

(1) Closeness between the predicted distribution and the distribution we knew;

(2) Whether the predicted distribution reflects ecological and biology realities of Hooded Crane (such as food, water availability etc);

(3) Whether some places were not predicted as part of the distribution area, and some places should be not considered as the distribution area but they were predicted (such as a settlement for instance)

## Alternative occurrence data assessment

Alternative data from other sources, such as other research, citizen observtions, speciemen, or new field investiantion data were very important testing data for us to assess model accuracy (Magness et al., 200; Huettmann and Gottschalk, 2011). In this study, we used the occurrence data of Hooded Cranes from Global Biodiversity Information Facility (GBIF, http://www.gbif.org/) as alternative data to assess our model performance. We applied all of the 90 locations, which were observed by people and recorded GPS location in winter time (October to next February) from 1994 to 2013. Then we extracted the relative index of occurrence (RIO) for each point from 8 Random Forest and 8 TreeNet spatial distribution map.

# RESULTS

## Static metric evaluation

AUCs and TSSs from both, internal-aspatial and external-spatial metrics were close among models with different number of variables in Random Forest, while the Kappa value diversified more between different models. The internal-aspatial AUCs were slightly greater than the related external-spatial metric; however, TSSs had a contrasting trend. For TreeNet models, internal-aspatial AUCs and TSSs were always larger than related external-spatial AUCs and TSSs (Fig. 2a and 2c). For the Kappa statics of Random Forest, it showed a somewhat contrasting result from RF3 to RF78 for internal and external metrics. The trend of internal metrics was increasing first and then decreasing; while external-spatial metrics kept increasing, except for RF21 and RF29 they were smaller than RF11 (Fig. 2b).

From the linear regression analysis (Table 1), we found that AUC had a significant positive relationship with TSS and Kappa ($P \leqslant 0.001$, $R^2 > 0.510$), both for internal-aspatial and external-spatial metrics, except for the internal-aspatial Kappa metric of the Random Forest model. Therefore, in the remaining analysis, we just used AUC to evaluate model accuracy.

Put Figure 2 Here

Put Table 1 Here

We used three-way ANOVA analysis between AUC and model algorithm (RF or TN), evaluation method (internal-aspatial or external-spatial), number of predictors, and their interaction factors. The result showed that AUC was only effected by evaluation methods ($P$ = 0.001) and model algorithm × evaluation methods among these factors. In addition, the results of interaction plots (Fig. 2) showed that internal-aspatial AUC were usually larger than external-spatial AUC across all models with same variables (Fig. 3a), and for RF and TN models (Fig. 2b). However, we found only TN model had significant difference between internal-aspatial and external-spatial AUC with Paired t-test ($P$ = 0.000, t=10.727, df=7), but not for RF model ($P$ = 0.221, t=1.344, df=7).

Put Figure 3 Here

## Prediction map assessment

According to the distribution of Hooded Crane known to us and also when compare with the source of "The BirdLife International Red Data Book" (Collar et al., 2001), for Random Forest, first, we listed RF1, RF3, RF5, RF8 as the worst predictor set of models (ranked as the fourth place; see Fig. 3 and Table 2). It shows that models with the least predictors actually perform worst. This is due to the distribution map not reflecting well the true distribution situation of Hooded Cranes, and it goes against the ecological amplitude of hooded crane distribution and biology, especially in the Far Eastern part (Fig. 4a). Second, it should be fewer areas predicted as the winter distribution area in Sakhalin Island (Russia), and whether Lake Biwa (Japan) can be predicted. Sakhalin Island is just recorded as rare breeding area of Hooded Crane in history (Collar et al., 2001). RF78, RF29 predicted slightly better than RF21 and RF11 judged on the Sakhalin Island prediction that too many areas were predicted in the middle and upper part and along this island in the RF21 and RF29 model. Third, there were largely areas of RIO ranging from 0.41 to 0.60 in the prediction map of RF78. Therefore, we regarded RF29 as the best and rank it as the first place in Random Forest.

For TreeNet: first, TN1, TN3 and TN5 are ranked as worst in our set for the same reason with Random Forest. Next, ranked 2, came TN8, TN21 and TN29, because it should be fewer areas predicted as the winter distribution area in Sakhalin Island (Russia) and Shanghai (China). Third, Lake Biwa (Japan) should be predicted, Poyang Lake and Dongting Lake (China) should predict more area and, meanwhile fewer areas should be included in the east coast of Vietnam. Thus, TN78 is ranked higher than TN11. We think that the model evaluation through a prediction map assessment may still carry bias in some extent (e.g. in our study, prediction maps from models with predictor number from 11 to 78 were very close), but it is less than going purely by internal metrics.

Put Figure 4 Here

Put Table 2 Here

## Alternative data assessment

Alternative presence data from GBIF were also used to evaluate model accuracy. From the Fig. 5, we found that most Random Forest model performed good, especiealy of RF11, RF21, RF29 and RF78. For TreeNet, TN3 performed significantly good, and TN21, TN29, TN78 looks similar. Comparing with distribution maps (Fig. 3), these results were acceptable, becauce the general distribution were close and it looks like good prediction models. Therefore, it was difficult to distingusih which one was better than another one. Comparing the Ralative index of Occurrence (RIO), we found more record points had higher RIOs in Random Forest than in TreeNet.

Regression analysis was used to compare the AUC value with median and mean RIO of each model in Fig. 5. We found there was a significant linear regression relationship among internal-aspatial and external-spatial AUC with two kinds of RIOs in Random Forest model ($P <$ 0.03), but the $R^2$ of spatial method (mean = 0.942) were larger than internal-aspatial mehthod (mean = 0.706); while in TreeNet model, the linear relationship were not so obvious ($P > 0.189$).

Put Figure 5 Here

Figure 5 (a) violin plot of Random Forest with different predictors model; (b) violin plot of TreeNet with different predictors model. The thick black bar in the centre represents the interquartile range, the thin black line extended from it represents the 95% confidence intervals, and the white dot is the median.

# DISCUSSION

In this study, we have proposed two methods to obtain three evaluation criteria (AUC, the Kappa statics (Kappa) and the true skill statics (TSS)), we refer to them as internal-aspatial and external-spatial approaches (Fig. 2). Further, we used a prediction map based on experience knowledge (Fig. 4 and Table 2). Overall, regardless of the evaluation criteria (AUC and TSS) the internal-aspatial or external-spatialmetric, the AUC and TSS in these models with different predictors were close to each other. In comparison, Kappa performed slightly more distinct, especially in the Random Forest model (Fig. 2). In addition, we found there were obviously linear relationships among three evaluation criteria, no matter of what approach was used (Table 1).

When put to a test, the results of the three-way ANOVA showed that model accuracy based on AUC was only influenced by the evaluation approach (internal-aspatial or external-spatial) and the interaction of the evaluated metric and model algorithms (Random Forest or TreeNet). Though the internal values were larger than the external-spatial value in same models in most cases (Fig. 2a and Fig. 3), we found only TN model had a significant difference between the internal and the spatial AUC, but not for the RF model when using a Paired t-test. It means that the effect of the internal-aspatial and external-spatial metric to evaluate model accuracy was close in Random Forest; but would somewhat mislead in TreeNet. Based on a rough classifying system, AUC can be interpreted as follows: ≥ 0.9 are excellent, 0.80 ~ 0.90 "good", 0.70 ~ 0.80 "fair", 0.60 ~ 0.70 "poor"and 0.50 ~ 0.60 "fail" (Allouche et al., 2006). Therefore, one same TreeNet model would be listed as different accurate classes models when referring to internal-aspatial and external-spatial metric, which was also seen in Kappa and TSS (Fig. 2b and 2c).

From both of the internal-aspatial and external-spatial metrics of AUC and TSS values, we found it was difficult to tell which model was better, when the number of predictors ranged from 3 to 78. But spatial Kappa for Random Forest showed distinctly different among models, internal-aspatial Kappa showed an inconsistent result with external-spatial Kappa and the other two statistic criteria. Combining the rank of prediction maps assessed through our field knowledge (Fig. 4 and Table 2), we found that the Random Forest result was consistent with the external-spatial Kappa metric. We thought taking external-spatial Kappa as the criteria was the best choice among above for Random Forest model in our case, but maybe do not perform well for all models and species. This needs more applications and study. Also, several studies have criticized the kappa statistic for being inherently dependent on prevalence and they claimed that this dependency introduces bias and statistical artefacts to estimates of accuracy (Byrt et al., 1993; Lantz and Nebenzahl, 1996; Manel et al., 2001; McPherson et al., 2004).

The high AUC values (> 0.85) and the slightly difference among all 16 Hooded Crane models (Fig. 2a, Supplement S4) show that all models were accurate and performed similar, as values above 0.75 generally indicate an adequate model performance for most applications (Pearce and Ferrier, 2000). However, we would list RF1, RF3, TN1, TN3 and TN5 as 'bad models' because of the poor spatial prediction map (Fig. 4). It means that both of the internal-aspatial and external-spatial AUC did not perform so well to distinguish model predictions, but the prediction maps did. Therefore, we argue that model accuracy evaluation should not only be based on a static number, but also should care more about models' spatial prediction as assessed with external-spatial data!

In this research, we evaluated how the prediction map compares in the light of the experts' knowledge on species and its real winter distribution, to determine which model is more accurate and reliable. We clearly agree with the thought of Fielding (2002): accuracy assessments for models should have a close connection to the intended use of the model. One major gain of species distribution models is to model and quantify complex ecology and to support an efficient conservation management, such as conservation planning (Drew et al., 2011). It could be used for instance to design reserve networks that maintain biodiversity (Guisan and Thuiller, 2005; Han et al., 2018). Thus, in order to obtain a more accurate and reliable species distribution model in less time and with less money, an external-spatial approach will be more meaningful than just to get a numerically perfect statistical model with unproven output and assumptions. In times of limited and competing, science budgets, such things really matter, and when large scales are to be handled well; valid inference remains 'key'. On such scales even small decimal improvements can be of major value. Wilson et al. (2005) already concluded that efforts should be directed towards producing the most reliable predictions for use in conservation planning, and for instance to find the reserve network that is most robust to the uncertainty in the predictions.

In addition, alternative occurrence data from other sources were also used to assess model accuracy. The results were similar with prediction maps and statistic metrics for Random Forest models, but not really for TreeNet. This also was proved through the regression analysis result between internal-aspatial, and external-spatial AUC and RIOs. The alternative data used, in this research, was only 'presence' data. In the future true absence data (=species are not occurring) should also be collected, though collecting such absence data remains difficult. However, in all the ways we used here and also studies elsewhere, most studies people used point data (like presence, pseudo-absence points) to assess distribution area accuracy (Baltensperger et al., 2013; Mi et al., 2017). Whether points can stand for the area (polygon) and how much representation they own should be a discussion in future study. It's a question of detection distances as assessed through Distance Sampling for instance!

In this study, we found the accuracy of highly non-parsimonious models RF78 and TN78 performed very well, and those were close to models with a set of predictors reaching from 11 to 29 across most of evaluation criteria. That means high-dimension variables models can also predict species distribution very well. In contrast we found that, so far, high-dimension variables models are widely avoided by ecological researchers (few study use more than 30 variables, we referred 30 papers published, see Supplement S5). The published advice was that high dimensionality is unwanted, dangerous (Meisel, 1972), poorly fitted or overfitted (Harrell et al., 1996). But Breiman (2001b) stated for long time differently, and recent work has shown that dimensionality can be a blessing (Drew et al. 2011) and any proxy predictor can often improve the prediction, specifically when dealing with large scales and when decimals mean a lot beyond just template thresholds. This presents a massive paradigm shift for the sciences. Though using complex predictors may be unpleasant perhaps, and requires skill and some time, the soundest path for valid inference – the goal of science (generalization; as per textbook) - is to go for predictive accuracy first, then try to understand why and to infer (Breiman, 2001b; Hilborn and Mangel 1997, Drew et al. 2011).

# Conclusion

In this study, we used two methods to assess model accuracy: internal-aspatial and external-spatial. We found internal-aspatial and external-spatial metrics can get higher model performance (AUC > 0.85), but both of them can't distinguish models with different predictors well, while the prediction maps did a little better than them. Therefore, we argued that model accuracy evaluation also should care more about models' spatial prediction and has a close connection to the intended use of the model! Certainly, all above conclusion is limited to Random Forest and TreeNet from lots of SDM options available, and only one species. Whether other algorithm implementations and species have the same results should be tested further. As we know, other than Breiman (2001b) and related papers by the authors, this maybe the first time the concept of external-spatial assessment criteria for model accuracy is formerly promoted, and with a quest to assess model accuracy through prediction maps for inference and applications. Overall, we hope to see more study on proposing better methods and data to assess species distribution models (SDMs) and prediction distribution map for valid inference, and sustainable conservation management worldwide.

# ACKNOWLEDGEMENTS

This research was possible because of the large investment of field effort, money, time, personal interest, and dedication by researchers and people who believe in sustainable science-based conservation management. Thanks to all data contributors to the book 'Threatened Birds of Asia', to the support of State Forestry Administration and Whitley Fund for Nature (WFN). We heartily thank all those who contributed to the Hooded Crane protection and species distribution models, and Salford Systems Ltd. This research was funded by The National Forestry Bureau of China.

# REFERENCES


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

Anderson, R.P., Gonzalez Jr, I., 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: an implementation with Maxent. *Ecological Modelling* 222, 2796-2811.

Aspinall, R., Veitch, N., 1993. Habitat mapping from satellite imagery and wildlife survey data using a Bayesian modeling procedure in a GIS. *Photogrammetric Engineering and Remote Sensing* 59, 537-543.

Austin, M., Nicholls, A., Margules, C.R., 1990. Measurement of the realized qualitative niche: environmental niches of five Eucalyptus species. *Ecological Monographs* 60, 161-177.

Baltensperger, A., Mullet, T., Schmid, M., Humphries, G., Kövér, L., Huettmann, F., 2013. Seasonal observations and machine-learning-based spatial model predictions for the common raven (Corvus corax) in the urban, sub-arctic environment of Fairbanks, Alaska. *Polar Biology* 36, 1587-1599.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. *Journal of Applied Ecology* 43, 413-423.

Betts, M.G., Ganio, L., Huso, M., Som, N., Huettmann, F., Bowman, J., Wintle, B.A., 2009. Comment on “Methods to account for spatial autocorrelation in the analysis of species distributional data: a review” *Ecography* 32: 374-378.

Beyer, H., 2013. Hawth’s Analysis Tools for ArcGIS version 3.27 (software).

Birdlife international, 2014. IUCN Red list for birds.

Booms, T.L., Huettmann, F., Schempf, P.F., 2010. Gyrfalcon nest distribution in Alaska based on a predictive GIS model. *Polar Biology* 33, 347-358.

Breiman, L., 2001a. Random forests. *Machine Learning* 45, 5-32.

Breiman, L., 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16, 199-231.

Byrt, T., Bishop, J., Carlin, J.B., 1993. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46, 423-429.

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

Craig, E., Huettmann F., 2008. Using "blackbox" algorithms such as TreeNet and Random Forests for data-mining and for finding meaningful patterns, relationships and outliers in complex ecological data: an overview, an example using golden eagle satellite data and an outlook for a promising future. Chapter IV in Intelligent Data Analysis: Developing New Methodologies through Pattern Discovery and Recovery (Hsiao-fan Wang, Ed.). IGI Global, Hershey, PA, USA. pp 65 -83.

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

Derrig, R., Francis, L., 2006. Distinguishing the forest from the TREES: A comparison of tree based data mining methods, Casualty Actuarial Society Forum. Citeseer, pp. 1-49.

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

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40, 677.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17, 43-57.

Fielding, A.H., 2002. What are the appropriate characteristics of an accuracy measure. *Predicting species occurrences: issues of accuracy and scale*, 271-280.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation* 24, 38-49.

Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. *Progress in Physical Geography* 19, 474-499.

Franklin, J., McCullough, P., Gray, C., 2000. Terrain variables used for predictive mapping of vegetation communities in Southern California. Terrain analysis: principles and applications/edited by John P. Wilson, John C. Gallant.

Fernandez-Delgado, M., Cernadas, E, Barro, S., Dinan, A., 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning* 15: 3133-3181.

Friedman, J., 2001. TreeNet™.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis* 38, 367-378.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., OVERTON, J., Aspinall, R., Hastie, T., 2006. Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology* 43, 386-392.

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

Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. *Ecological modelling* 135, 147-186.

Guo, Y., 2005. Breeding behaviror ecology of Hooded Crane (Grus monacha) in Lesser Xingan Mountains. Northeast Forest University.

Han, X., Huettmann, F., Guo, Y., Mi, C., Wen, L., 2018. Conservation prioritization with machine learning predictions for the black-necked crane Grus nigricollis, a flagship species on the Tibetan Plateau for 2070. *Regional Environmental Change* 1-10.

Harrell, F., Lee, K.L., Mark, D.B., 1996. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* 15, 361-387.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition.Springer New York.

Hegel, T.M., Cushman, S.A., Evans, J., Huettmann, F., 2010. Current state of the art for statistical modelling of species distributions, Spatial complexity, informatics, and wildlife conservation. Springer, pp. 273-311.

Herrick, K.A., Huettmann, F., Lindgren, M.A., 2013. A global model of avian influenza prediction in wild birds: the importance of northern regions. *Vet Res* 44, 42.

Hilborn, R., Mangel, M.,1997. The Ecological Detective: Confronting Models with Data. 1997. Princeton New Jersey.

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?, Predictive Species and Habitat Modeling in Landscape Ecology. Springer, pp. 189-208.

IUCN, 2012. IUCN red list of threatened species. Version 2012.2. International Union for the

Conservation of Nature Gland, Switzerland.

Jiao, S., Huettmann, F., Guo, Y., Li, X., Ouyang, Y., 2016. Advanced long-term bird banding and climate data mining in spring confirm passerine population declines for the Northeast Chinese-Russian flyway. *Global and Planetary Change* 144: 17-33

Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications* 13, 853-867.

Kandel K. Huettmann, F., Suwal, M. K., Regmi, G.R., Nijman, V ., Nekaris, K.A.I., Lama, S.T., Thapa, A., Sharma, H.P., Subedi, T.R., 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.

Lantz, C.A., Nebenzahl, E., 1996. Behavior and interpretation of the statistic: Resolution of the two paradoxes. *Journal of clinical epidemiology* 49, 431-434.

Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H., 2003. Avoiding pitfalls of using species distribution models in conservation planning. *Conservation biology* 17, 1591-1600.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence–absence models in ecology: the need to account for prevalence. *Journal of applied Ecology* 38, 921-931.

McPherson, J., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of applied ecology* 41, 811-823.

Mi, C., Huettmann, F., Guo, Y., 2014. Obtaining the best possible predictions of habitat selection for wintering Great Bustards in Cangzhou, Hebei Province with rapid machine learning ßanalysis. *Chinese Science Bulletin*, 1-9.

Mi, C., Huettmann, F., Guo, Y., 2016. Climate envelope predictions indicate an enlarged suitable wintering distribution for Great Bustards (Otis tarda dybowskii) in China for the 21st century. *PeerJ* 4:e1630.

Mi, C., Huettman, F., Guo, Y., Han, X., Wen, L., 2017. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. *PeerJ* 5:e2849.

Mi, C., Møller, A.P., Guo, Y., (2018) Annual spatio-temporal migration patterns of Hooded

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27257v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

Cranes wintering in Izumi based on satellite tracking and their implications for conservation. *Avian Research* 9, 23.

Oppel, S., Meirinho, A., Ramírez, I., Gardner, B., O'Connell, A.F., Miller, P.I., Louzao, M., 2012. Comparison of five modelling techniques to predict the spatial distribution and abundance of seabirds. *Biological Conservation* 156, 94-104.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling* 133, 225-245.

Pearson, R.G., Dawson, T.P., Liu, C., 2004. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. *Ecography* 27, 285-298.

Pearson, R.G., Thuiller, W., Araújo, M.B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T.P., Lees, D.C., 2006. Model-based uncertainty in species range prediction. *Journal of Biogeography* 33, 1704-1711.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R.H., Stockwell, D.R., 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature* 416, 626-629.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9, 181-199.

Reese, G.C., Wilson, K.R., Hoeting, J.A., Flather, C.H., 2005. Factors affecting species distribution predictions: a simulation modeling experiment. *Ecological Applications* 15, 554-564.

Regmi G.R., Huettmann, F., Suwal, M. K., Nijman, V., Nekaris, K.A.I., Kandel, K., Sharma, N. and Coudrat, C., 2018. First Open Access Ensemble Climate Envelope Predictions of Assamese Macaque Macaca Assamensis in South and South-East Asia: A new role model and assessment of endangered species. *Endangered Species Research* 36: 149–160.

Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. *Journal of Biogeography* 31, 1555-1568.

Seoane, J., Carrascal, L.M., Alonso, C.L., Palomino, D., 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. *Ecological Modelling* 185, 299-308.

Simonov, E.A., Dahmer, T.D., 2008. Amur-Heilong river basin reader. Ecosystems.

Vaughan, I., Ormerod, S., 2005. The continuing challenges of testing species distribution models. *Journal of Applied Ecology* 42, 720-730.

Wei, C.L., Rowe, G.T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M.J., Soliman, Y., Huettmann, F., Qu, F., Yu, Z., 2010. Global patterns and predictions of seafloor biomass using random forests. *PLoS One* 5, e15323.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. *Biological Conservation* 122, 99-112.

Figure Legend

**Figure 1** Map of study area and study species locations.

**Figure 2** (a) barplot of AUC from internal-aspatial and external-spatial metrics of RF and TN; (b) barplot of Kappa from internal-aspatial and external-spatial source of RF and TN; (c) barplot of TSS from internal-aspatial and external-spatial source of RF and TN.

**Figure 3** Interaction plot of AUC value between predictor number, model algorithms and evaluation methods. (a) Interaction plot of AUC value between predictor number (1, 3, 5, 8, 11, 21, 29 and 78) and two evaluation methods (internal-aspatial and external-spatial), (b) Interaction plot of AUC between model algorithms (Random Forest) and two evaluation methods (internal-aspatial and external-spatial). Dots with same line represent the AUC value (internal-aspatial and external-spatial) of same model with certain predictor number (1, 3, 5, 8, 11, 21, 29 and 78).

**Figure 4** Prediction map of Random Forest and TreeNet with 8 different predictor numbers. (a) Prediction map of Random Forest; (b) Prediction map of TreeNet.

**Figure 5** (a) violin plot of Random Forest with different predictors model; (b) violin plot of TreeNet with different predictors model. The thick black bar in the centre represents the interquartile range, the thin black line extended from it represents the 95% confidence intervals, and the white dot is the median.

Table Legend

**Table 1** Linear Regression analysis of AUC, Kappa and TSS.

**Table 2** Rank of model by prediction map assessment.

# Table 1(on next page)

Table

Table

Table 1 Linear Regression analysis of AUC, Kappa and TSS

| | Slope | $R^2$ | *P* |
|---|---|---|---|
| AUC~Kappa | 0.165 | 0.510 | 0.000 |
| AUC~TSS | 0.481 | 0.839 | 0.000 |
| AUC_RF~Kappa_RF | 0.159 | 0.556 | 0.001 |
| AUC_RF~TSS_RF | 0.455 | 0.773 | 0.000 |
| AUC_TN~Kappa_TN | 0.203 | 0.582 | 0.000 |
| AUC_TN~TSS_TN | 0.533 | 0.937 | 0.000 |
| AUC_RF_spatial~Kappa_RF_spa | 0.203 | 0.847 | 0.001 |
| AUC_RF_spa~TSS_RF_spa | 0.507 | 0.999 | 0.000 |
| AUC_RF_aspa~Kappa_RF_aspa | 0.120 | 0.348 | 0.123 |
| AUC_RF_aspa~TSS_RF_aspa | 0.598 | 0.970 | 0.000 |
| AUC_TN_spa~Kappa_TN_spa | 0.234 | 0.968 | 0.000 |
| AUC_TN_spa~TSS_TN_spa | 0.503 | 1.000 | 0.000 |
| AUC_TN_aspa~Kappa_TN_aspa | 0.074 | 0.968 | 0.000 |
| AUC_TN_aspa~TSS_TN_aspa | 0.263 | 0.943 | 0.000 |

Note: aspa represents internal-aspatial metric, spa represents external-spatial metric.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27257v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

Table 2 Rank of model by prediction map assessment.

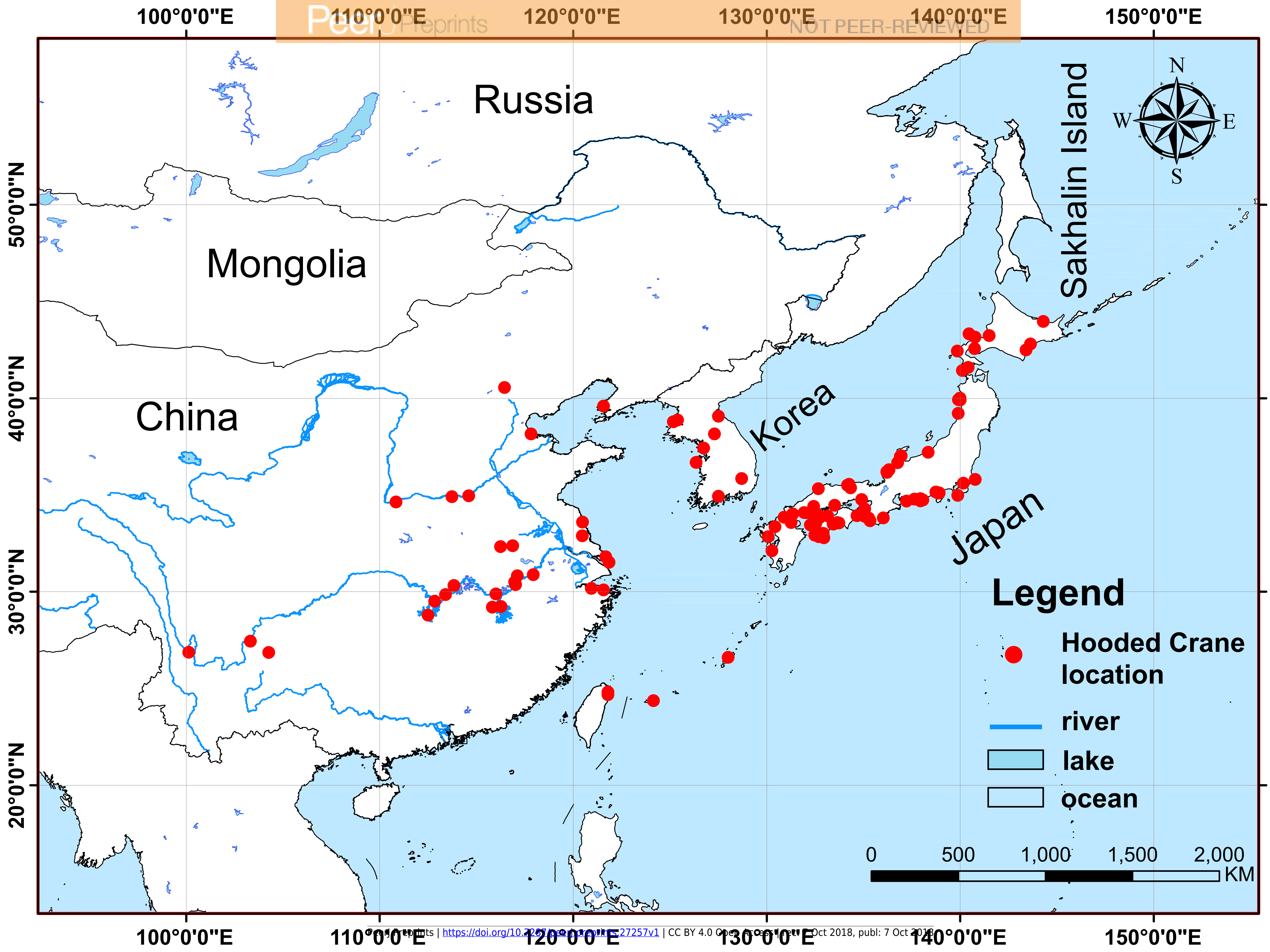
| Rank | Random Forest | TreeNet |
| --- | --- | --- |
| 1 | RF29 | TN78 |
| 2 | RF78 | TN11, |
| 3 | RF11, RF21 | TN29, TN21, TN8 |
| 4 | RF1, RF3, RF5, RF8 | TN1, TN3, TN5 |

Note: 1 means the best; 2 means better; 3 means good; 4 means less good.

## Figure 1(on next page)

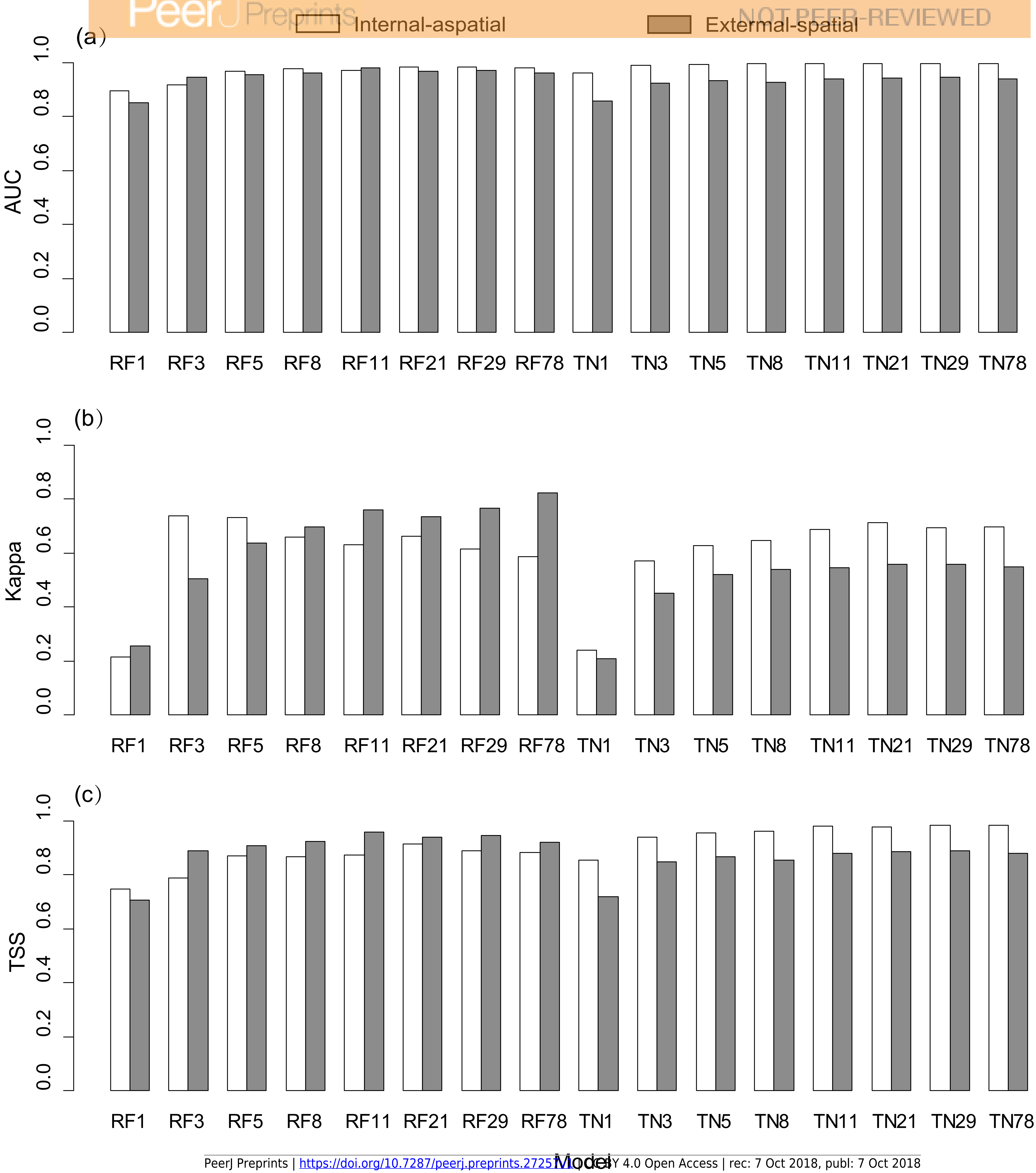
Figure 1 Study area

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27257v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018


100°0'0"E
110°0'0"E
120°0'0"E
130°0'0"E
140°0'0"E
150°0'0"E
50°0'0"N
40°0'0"N
30°0'0"N
20°0'0"N
Russia
Mongolia
China
Korea
Japan
Sakhalin Island
N
S
W
E
Legend
Hooded Crane location
river
lake
ocean
0
500
1,000
1,500
2,000
KM

## Figure 2(on next page)

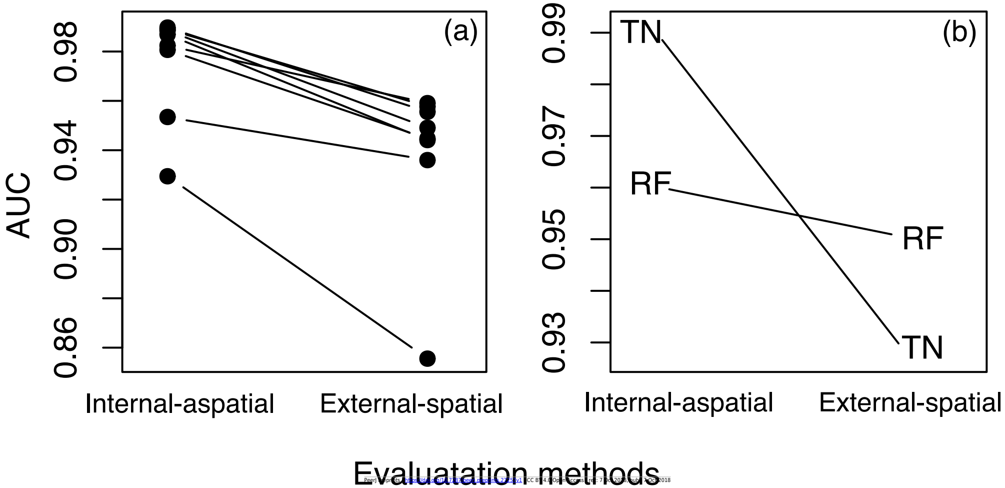
Figure 2




PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27251v1 | CC BY 4.0 Open Access | rec: 7 Oct 2018, publ: 7 Oct 2018

## Figure 3(on next page)

Figure 3


(a)
AUC
0.98
0.94
0.90
0.86
Internal-aspatial
External-spatial
(b)
0.99
0.97
0.95
0.93
TN
RF
RF
TN
Internal-aspatial
External-spatial
Evaluatation methods

## Figure 4(on next page)

Figure 4a


(a)
1 predictor
3 predictors
5 predictors
8 predictors
11 predictors
21 predictors
29 predictors
78 predictors
Legend
Random Forest
with different predictors
Predicted Occurrence
0.00 - 0.20
0.21 - 0.40
0.41 - 0.60
0.61 - 0.80
0.81 - 1.00

## Figure 5(on next page)

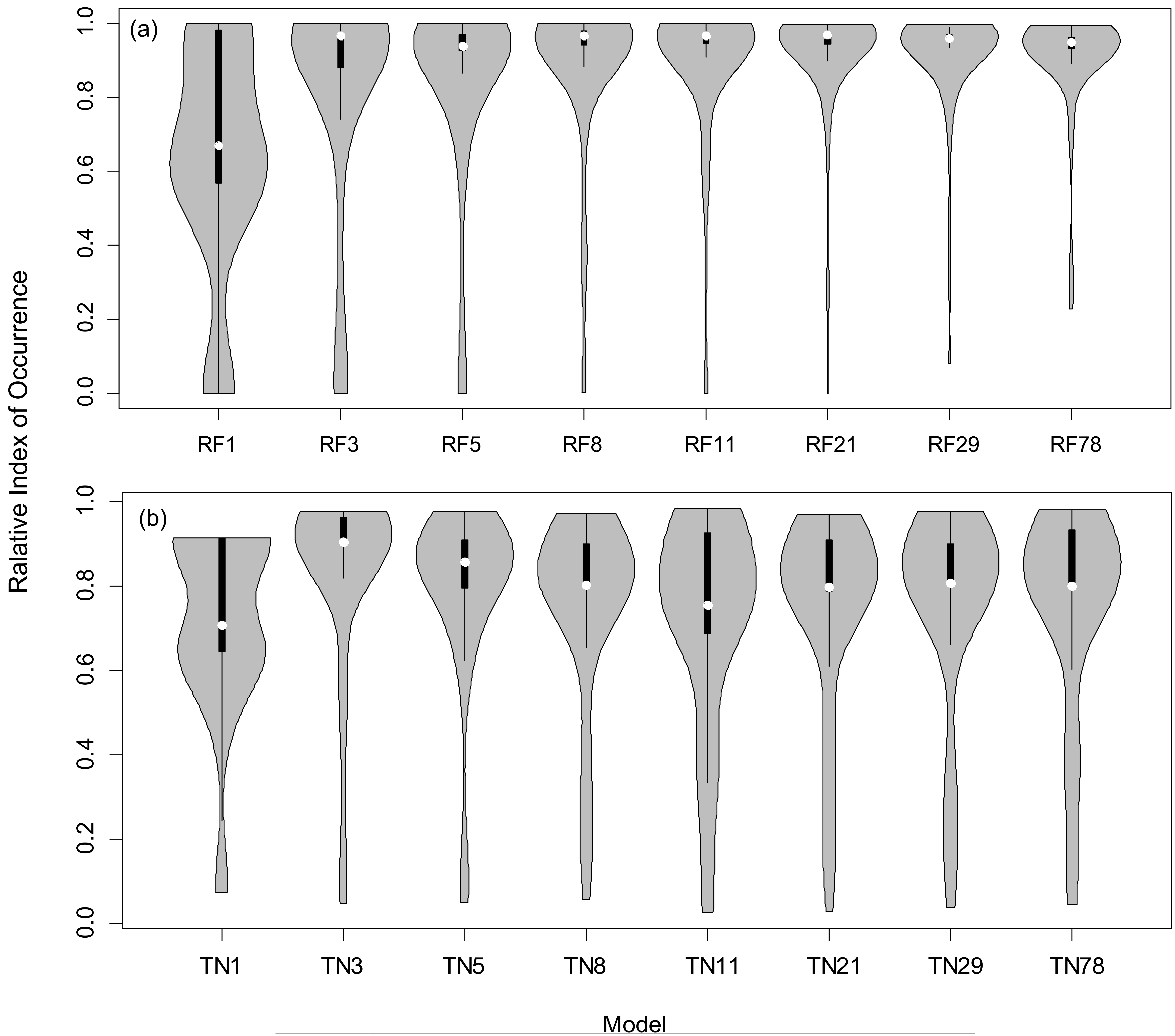
Figure 4b

(b)


1 predictor
3 predictors
5 predictors
8 predictors
11 predictors
21 predictors
29 predictors
78 predictors
Legend
TreeNet with different predictors
Predicted Occurrence
0.00 - 0.20
0.21 - 0.40
0.41 - 0.60
0.61 - 0.80
0.81 - 1.00

## Figure 6(on next page)

Figure 5


(a)
(b)
Ralative Index of Occurrence
Model
0.0
0.2
0.4
0.6
0.8
1.0
RF1
RF3
RF5
RF8
RF11
RF21
RF29
RF78
TN1
TN3
TN5
TN8
TN11
TN21
TN29
TN78