Final Report: Common swift nest site selection in the United Kingdom

1. Introduction

Currently, the global climate is undergoing a rapid warming process[1]. Impacted by climate change apparent from mid to late 20th century onwards, distribution ranges of many species have been dramatically altered[2,3]. For example, elevational and longitudinal range shifts, lowland biotic attrition are frequently observed in species of tropical latitudes in recent years[4,5]. The global-wide species redistribution can cause great disturbances to the ecosystems and threaten many species since adapting a new habitat can be especially difficult[6]. It is, therefore, of great importance for conservational biologists to study why species migrate, and where they go in respond to climate changes and thus help them adapt in the new environment while manage to maintain the healthy function of ecosystems.

In the last 20 years, researchers have developed a great number of species distribution models (SDMs) to predict species' potential range selection under future climate condition[7]. However, deficiency of fine-scale spatial-tempo data, the complexity of species-environment relationship and the limitations of models have significantly hampered our ability to make accurate predictions and thus feasible inferences[8,9]. Given that mathematical models can never capture the whole scope of real-world biological phenomena, researches have proposed the concept of Ensemble models[10], which make comprehensive comparisons between different models and incorporate their result to optimize the prediction precision and accuracy. Models used in ensemble analysis can be of all kinds, ranging from classical linear regression, classification trees to sophisticated machine learning algorithms such as artificial neural network, random forest, etc.

Common swifts (Apus apus) are medium size birds in the order of Apodiformes. They are well known for their dominant flying ability, being able to not touching the land for up to 10 months continuous flight, and outstanding adaptability to different environments[11]. The migratory species is a popular summer visitor from Africa to the United Kingdom, breeding throughout the entire country and live harmoniously with artificial environment alongside humans[12]. Sadly, the swift breeding population has been suffering from a rapid decline over the past 20 years (53% reduction between 1995 and 2016). It is believed that the losses of suitable nesting sites are to blame for the decline. Most of the swift nest sites surveyed in the UK are built high up in the rooftop of old buildings such as dilapidated houses, church, and barns. Unfortunately, the tendency of old building renovation and conversion to luxury residences has many swift nest sites knocked down or blocked. It is particularly disastrous for migratory species like common swift, as they tend to return to the same spot (building) each year during the summer breeding season. The severe population decline in the past decades has made swift an amber-listed species, meaning that human help is needed for the recovery of the swift population.

Currently, swift conservation in the UK mainly focuses on the nesting site protection in the buildings, but little work has been done to understand how environment plays a role in swifts' decision on choosing which building to nest on[13]. The aim of this mini-research project is, therefore, to identify the environmental factors that may influence swift nest site decisions in the UK using ensemble modeling, and subsequently to predict swift breeding hotspot around the entire country. This can potentially increase the efficiency of the swift nest survey, which now relies heavily on public participation in nest sighting and reporting. The result of this project may also avail pertinent management plans at the predicted swift breeding hotspots, and thus provide theoretical decision basis to long-term swift conservation under different future climate situations in the UK.

2. Method

2.1 Data

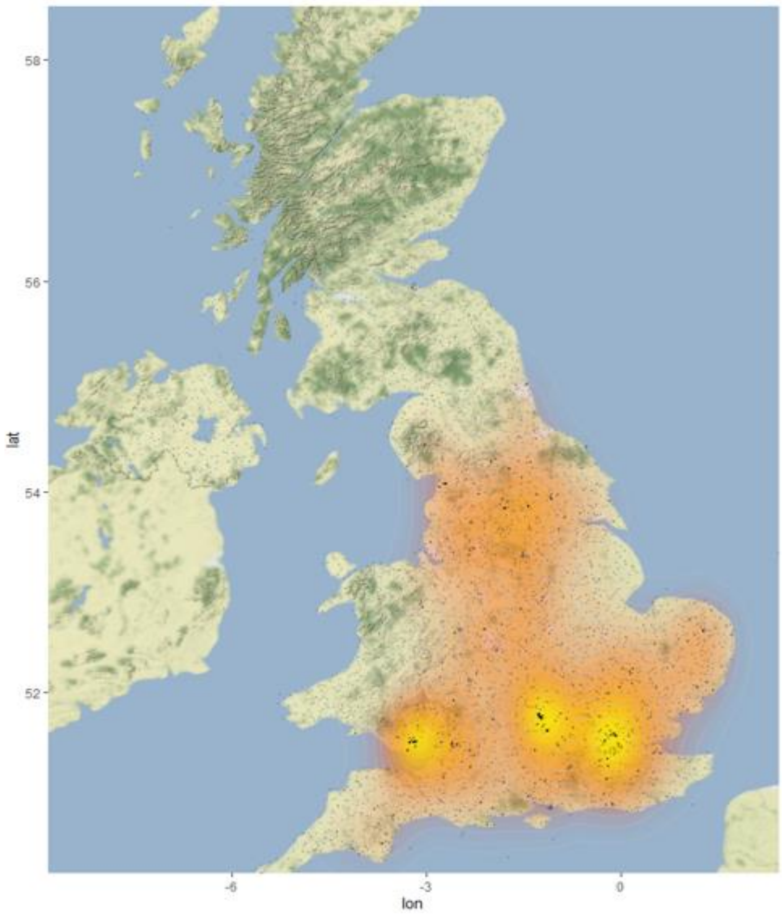

*Figure 1: **Locations of swift nests surveyed in 2019 by RSPB Swift Survey** (black) and **generated pseudo-absent data** (steel blue). 2D kernel density estimation of swift nest site is displayed with coloured contours (orange to yellow). Lighter colour represents higher density.*

Data used were directly download from RSPB Swift Survey[14], a public complied data portal which contains detailed information about when, where and how many swift nests are sighted inside British and Northern Irish national grid from 2016 to 2019. Breeding season is a critical period which influences swifts' population dynamic, and nest sites during the time can effectively reflect swift's habitat preferences and demands such as appropriate temperature and humidity, proper elevation, insect-abundant area and adequate wetland resources. Locations of surveyed swift nest sites are recorded in latitudinal and longitudinal scales, but the precisions of these spatial points are unknown. As records may duplicate across years, only data from 2019 with a total number of 2249 swift nest site occurrences are used for the analysis.

Our study area is restricted to the mainland UK territory (242,495 kilometre square). Inside the focused region, points are uniformly and randomly sampled within every 0.1 longitudinal and latitudinal intervals as 'pseudo-absent' sites in which we assume it has been surveyed and no swift nests are found. Generated spatial points that are a certain geographical distance away (about 7 kilometers) from any genuine existing records are retained for the model construction (3317 points in total) while the others are discarded (*Figure 1*). Binary class variable, ***is.nest***, which indicates the presence/absence of swift nests is then created and assigned to each data point as the response variable for the downstream analyses.

Considering that swifts usually visit and stay in UK during the summer, we only selected two spatially interpolated climate data: Mean Temperature of Warmest Quarter (variable name: BIO10) and Mean Precipitation of Warmest Quarter (variable name: BIO18), from WorldClim Bioclimatic database to represent climate information that might be relevant to swift nest site selection[15]. Future climate condition (Average BIO10, BIO18 values for year 2041-2060) is predicted by CCSM4 (Community Climate System Model) hosted at WorldClim[16]. Climate change are modelled under RCP8.5 Greenhouse gas concentration pathway, which is the highest global warming increase projection adopted by IPCC for its fifth Assessment Report (AR5). All climate data are retrieved and processed with 'raster' and 'rgdal' package in R[17,18].

Following 5 environmental and geographical factors that cannot be predicted or might not change in a short time period are also recruited as predictors in the analyses: Elevation (obtained from R package: 'elevatr'); Standard Deviation of Elevation (calculated on Elevation raster using terrain function in 'raster' package). Human Footprint (obtained from NASA Socioeconomic Data and Applications Center, SEDAC). Land and Water Area (obtained from SEDAC).

Parameters used are summarised as follows (Table 1):

Variable abbreviation	Variable Full Name	Data Type
is.nest	Common Swift nest occurrence	Binary Response
bio10	*Mean Temperature of Warmest Quarter*	Numerical predictor
bio18	*Mean Precipitation of Warmest Quarter*	Numerical predictor
elev	Elevation	Numerical predictor
elev.sd	Standard Deviation of Elevation	Numerical predictor
hfp	Global Human Footprint	Numerical predictor
land	Surface areas of land in square kilometers per 5km pixel	Numerical predictor
water	Surface areas of water in square kilometers per 5km pixel	Numerical predictor
f.bio10	Prediction of BIO10 in Year 2050	New data for prediction
f.bio18	Prediction of BIO18 in Year 2050	New data for prediction

*Table 1: **Summary of variables used in the study.** Two climate factors and five environmental variables are used as predictors in subsequent model*

2.2 Models

Generalized Linear Model (package: *'stats'* in R)

Generalized linear model (GLM) is the expansion to the ordinary least square regression, allowing for inferences on non-normally distributed data from exponential family using maximum-likelihood estimation. Here in this study, binary class variable (***is.nest***) is fitted with GLM using logistic regression which models swifts' nesting decisions as independent Bernoulli trails, whose successing probabilities (decide to nest at a certain location) are linked to the linear predictors (environment factors) via logistic functions.

Generalized Addictive Model (package: *'mgcv'* in R)[19]

Generalized addictive model (GAM) further expands GLM by smoothing the linear predictor with non-parametric smothers, permitting relax modelling of complex non-linear relationship between predictors and response variables. Each predictor provided to GAM are iteratively smoothed to minimize the residual sum square (also known as Backfitting process), and the resulting estimates are addictively combined for a complete model.

Multivariate Adaptive Regression Splines (package: 'earth' in R)[20]

Multivariate Adaptive Regression Splines (MARS) is a complex, non-parametric regression model that fits piecewise regression lines on subgroups of data which are recursively partitioned across ranges of predictor values. Regression lines within each subgroup are automatically defined and connected by basis functions to form a regression spline, in which gradients of regression lines can vary to flexibly adapt non-linear variable relationships, non-addictive variable contributions and multivariate interactions in high dimensional data.

Flexible Discriminant Analysis (package: 'mda' in R)[21]

Flexible discriminant analysis (FDA) is a popular multi-response classification technique that partitions the predictor space into subspaces of discriminant regions separated by non-linear decision boundaries. The decision boundaries are formed by adaptive non-parametric regression on dummy response variables transformed by optimal scoring to represent grouping information of the data. Here, we choose to use MARS as the regression method to be implemented in FDA in this study.

Naïve Bayes Classifier (package: 'e1071' in R)[22]

Naïve Bayes Classifier (NBC) is a simple but powerful probabilistic classifier that is based on the Bayesian theorem for conditional probability. It assumes that both the class variable (presence of swift nest) and the predictors (environmental factors) have their own probability distribution. The marginal distribution of class variable is referred to as prior in Bayesian inference, while the conditional distributions of predictors given the class variable are known as likelihoods. Prior distribution of class variable can be informed and upgraded to its posteriori form when empirical observations of predictors are provided to the model. Decision is made in NBC by assign the class label (nest is present/ not present) with the largest posteriori probability to the given new data point.

Classification tree analysis (package: 'tree' in R)[22]

Classification tree analysis (CTA) regroups the data according to series of decision criteria on the predictors, forming nodes, branches and leaves of a decision tree. At each node of the tree, a test is designed on one of the environmental variables to split the data into several subgroups such that the within group homogeneity of class variable is maximized. Prediction of nest occurrences is made in CTA by traversing the new environmental data through the trained decision tree until it hits a leaf node to which a class label is assigned.

Random Forest (package: 'randomForest' in R)[23]

Random Forest, as its name indicates, is a forest of random decision trees that are constructed as the same way that are done in CTA. However, each decision tree in RF is trained on random subsets of data with replacement using randomly selected predictors at each split of the tree. When an unsurveyed new data point is passed to RF, each tree in the forest can give their own prediction on whether a swift nest will found at the site or not. The class label (found / not found) that receives majority of the votes will be returned as the classification outcome by RF.

Artificial neural network (package: 'nnet in R)[24]

Artificial neural network (ANN) constructs models in a system that is structurally analogous to the actual biological neural network of animal brains in which neurons are interconnected to process information collaboratively.

Similarly, neurons (models) in ANN are connected by edges to achieve multiple model combination. An ANN is composed of three different layers, an input layer that intake the environmental data, a hidden layer consisting of neurons that process and aggregates the signals passed from the input layer, and an output layer that returns the prediction of nest occurrence. The most commonly used model in ANN is generalized linear model. By defining and adjusting the weights of edges during the learning process, ANN can attain the effect of any smoothing functions used in non-parametric regression analyses.

Generalized boosting model (package: 'xgboost' in R)[25]

Generalized boosting model (GBM) is an alternative decision tree based algorithm that is similar to the Random Forest, except that tree models are grown and boosted in a sequential manner in GBM. Model parameters and residuals from precedent trees inform the successive tree growth, and predictors are consistently reweighted to reduce the prediction error as the boosting proceeds. GBM is controlled by three important tuning parameters: Tree Depth, number of boosting round and shrinkage parameter eta. Specifically, eta determines individual trees' contribution to the growing model. Smaller the eta, the more conservative the boosting process is, and thus the less susceptible the model is to the overfitting. In R package xgboost, stochastic gradient boosting method is used to resample the training data and its associated environmental predictors for each round of tree growth.

2.3 Model Assessment.

Diagnostic ability of each individual model is evaluated by ROC (package: '*pROC*' in R) and Cohen's Kappa coefficient (package: '*psych*' in R)[26,27]. ROC, or Receiver operating characteristic, is a curve that plots false positive rate against true positive rate (sensitivity) of model predictions under different decision thresholds. Larger the area under the ROC curve (higher AUC), better the model performs in making binary decisions. Cohen's Kappa coefficient is a statistic that is used to measure the agreement between random guessing and model prediction, therefore checking the possibility that the correct model predictions are occurring by chance. 10 fold cross validation (data is partitioned by '*createFolds*' function in R package '*caret*')[27] is performed to increase the statistical significance of the model assessments[28].

2. Result

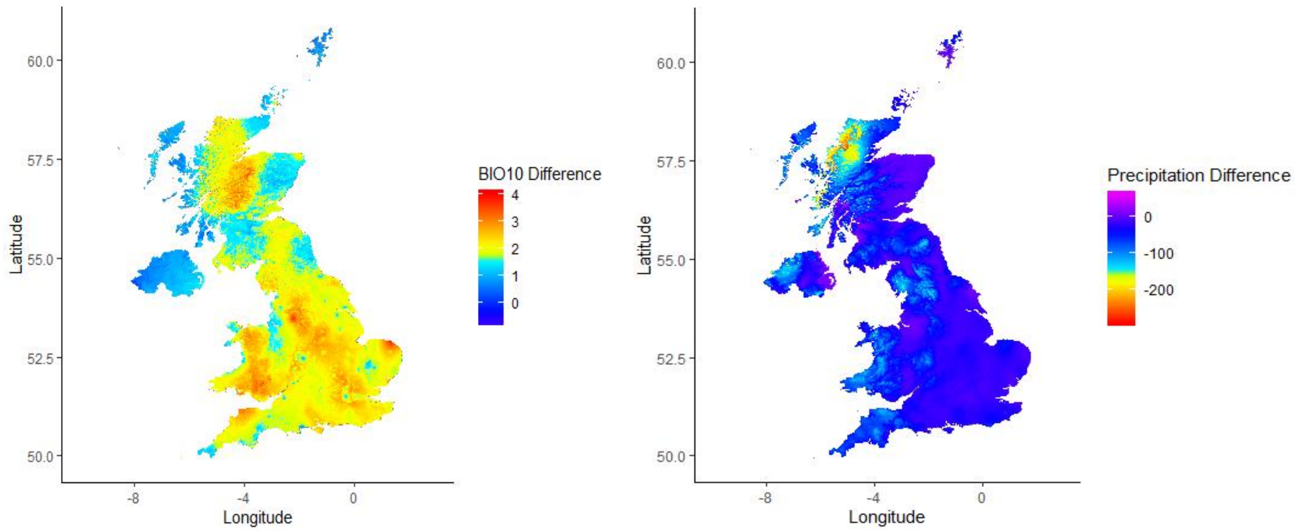

Figure 2: ***Differences in Mean Temperature and Precipitation in the Warmest Quarter in the UK between the present (1970-2000) and the future (2050-2059).*** *Comparison is made based on WorldClim Global Climate Database and CCSM4 climate prediction. Values are calculated by subtracting the current precipitation/temperature levels from the future's in each of the raster cell.*

According to *WorldClim CCSM4* future climate projection under *RCP8.5* Greenhouse gas emission pathway, mean temperature in the warmest quarter in the UK by 2050 will be 1.93 °C warmer than the 1970-2000 average. Temperature change in Northern Ireland and south east Scotland is relatively minor compared to the other regions. Aside from the rising temperature, the UK will also undergo a nationwide precipitation decrease by 45.02mm in 50 years, and the situation is especially severe in the Scottish Highlands.

Principle component analysis (PCA) with seven environmental predictors shows that swift nest site selections are primarily influenced by temperature, precipitation, elevation and human's footprint. Severe collinearity exists between variables on the first component, but is not addressed in this study for simplicity's sake. The separation between nest occurrence and non-occurrence samples on the first PC indicates that swifts generally prefer to nest in lowland plains with warm and dry atmosphere. Loading of human footprint also agrees with our knowledge of swift nesting

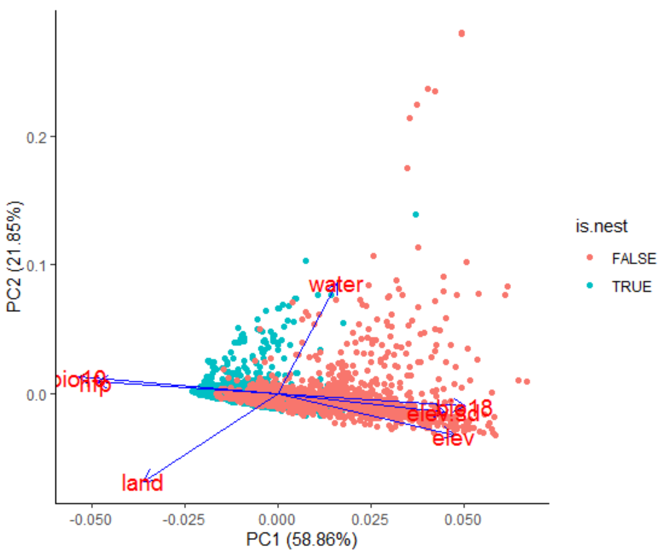

Figure 3 : ***Principle component analysis on 7 environmental and geographical predictors.*** *Blue dots: Swift nests surveyed in 2019 by Swift Survey. Red dots: Randomly generated pseudo-absent points.*

preference for old buildings or other artificial environments. First two PCs explained about 80% of the total variation.

Models (description)	Importance Weights of Environmental and Geographical Predictors (%)						
	Bio10	*Bio18*	*elev*	*elev.sd*	*land*	*water*	*Hfp*
GLM (first order, no interaction)	22.11	5.66	0.93	8.24	9.38	5.00	**48.69**
GAM (Default spline smooth)	13.53	2.43	1.42	4.62	1.92	1.15	**74.93**
MARS (Default parameters)	**59.78**	0.00	8.15	3.48	4.99	0.00	23.60
FDA (Default parameters)	24.04	0.32	7.35	3.65	1.57	0.00	**63.07**
NBC (Default parameters)	20.21	17.01	14.52	12.10	15.53	0.00	**20.63**
CTA (Split by gini)	24.13	16.87	14.17	0.15	18.63	0.18	**25.86**
RF (1000 trees)	26.30	12.39	10.18	6.45	17.08	1.17	**26.44**
ANN (4 nodes in hidden layer)	21.61	8.67	6.01	6.51	5.09	8.13	**43.98**
GBM (1000 boosting round)	34.93	5.29	5.48	6.13	7.69	0.66	**39.82**

*Table 2: **Variable Importance Weights for nine model assessed**. Note that variable importance for different models are calculated with different measurement. Hence these values are not inter-comparable between different models.*

Models	Model Assessment Indexes (Averaged score across each round of cross validation)					
	AUC (%)	*accuracy (%)*	*kappa*	*specificity*	*sensitivity*	*threshold*
GLM	95.1745	88.2602	0.7686	87.0697	90.1376	0.5889
GAM	95.1567	87.9851	0.7749	85.6673	91.6369	0.5849
MARS	95.2213	88.9532	0.7694	89.0681	88.7736	0.6546
FDA	95.1946	88.5248	0.7587	88.8443	88.0246	0.6939
NBC	93.3891	87.1256	0.7183	88.6215	84.7823	0.9697
CTA	89.5489	86.2549	0.7422	83.6063	90.4174	0.5146
RF	96.6453	90.8568	0.8099	90.8410	90.8858	0.6275
ANN	95.1699	88.0958	0.7708	90.7483	90.7483	0.5943
GBM (tree)	95.5085	90.2849	0.7796	92.0228	87.5527	0.8839

*Table 3: **Model Assessment on Each of the Nine Models.*** 10-fold cross-validation is performed to increase the viability of the assessment. Scores present here are averaged across each round of cross validation.

All nine models have made their predictions on swift potential habitat in the UK. Potential habitat is the model predicted suitable swift nesting range under future climate conditions, given the premise that current swift nest sites can effectively reflect swift's optimal range selections. Nine probability maps of swifts' current breeding hotspots are produced. Calculation of variable importance (function '*varImp'* in R package '*caret'*) shows that Human footprint index (*hfp*) has

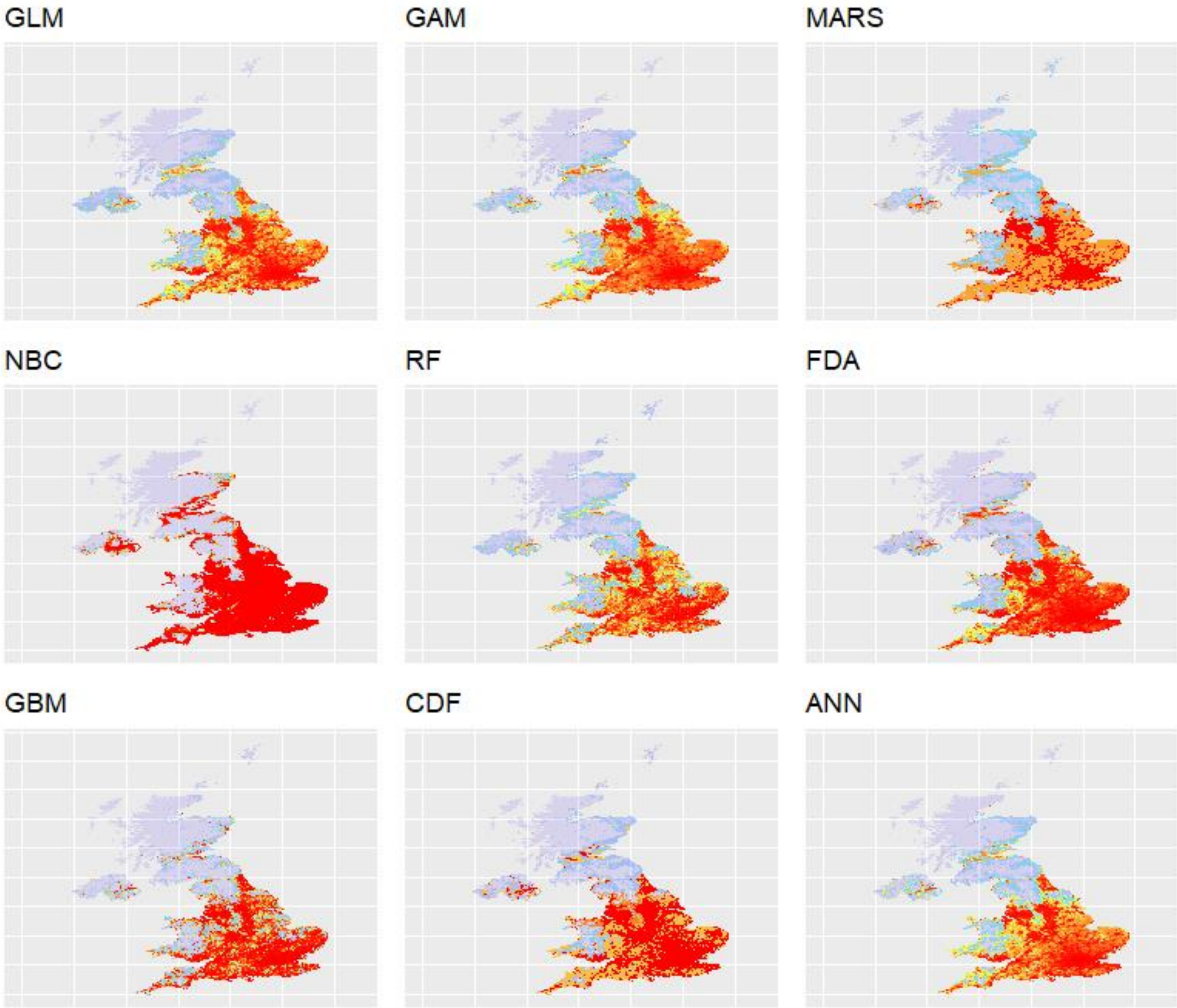

Figure 4: ***Model Estimated Probability Mapping on Swift Nest Occurrences across the UK****. The probability of finding a swift nest at a certain location, on the scale from 0 to 1, is present in the plot as colours ranging from blue to red. Bearing in mind that theses plots do not correspond to the true probability of finding a swift nest, but should only be used as density plots to demonstrate where swift nests are more likely to be found.*

the largest predictive power on swift nest occurrences among proposed seven predictors (*Table 2)*.

As reported by AUC and Cohen's Kappa coefficient with 10-fold cross validation, all nine models perform pretty well in predicting the presence of swift nests (*Table 3*). Random Forest, the model which obtained the highest performance scores, will be used as the representative model to demonstrate climate changes' potential impact on swift population and range selection in the UK.

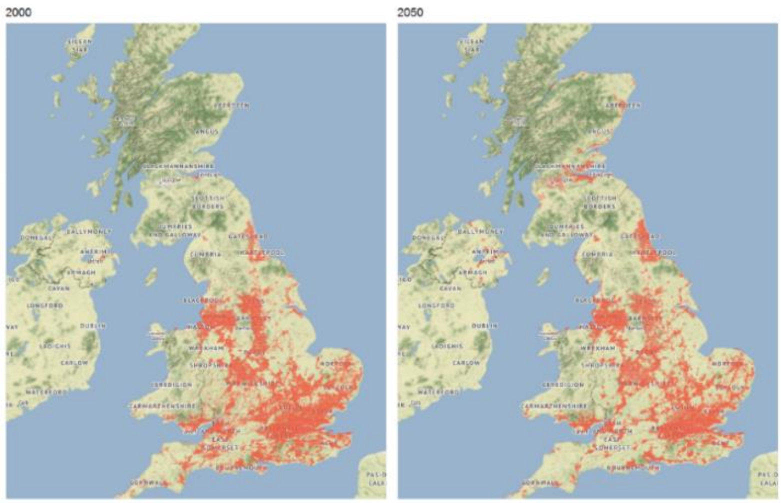

*Figure 5: **Predicted Swift Nest Site Occurrence in UK in 2019 and 2050**.* In each of the raster cells in the map (450 million in total), the presence of swift nest is modelled as an independent Bernoulli trial with probability predicted by Random Forest. The probability is scaled to make the total number of occurrences agrees with current British Swift population.

Random Forest forecasts that swift's habitat in south and central England will narrow sharply by year 2050, and a trend of northern range shift towards south east Scotland might occur. Swift's distribution centroid is predicted to move from today's 52.40' N to 53.05' N in 2050. By the time, cities and towns in Scottish Lowlands and Northern Ireland can all become swift's breeding hotspots during the summer. Swift population in Gateshead and Hartlepool may also increase in the future. Partial interaction plot shows that British swifts' ideal nesting temperature and precipitation level is about 15 degree Celsius and 220 mm respectively (figure 7).

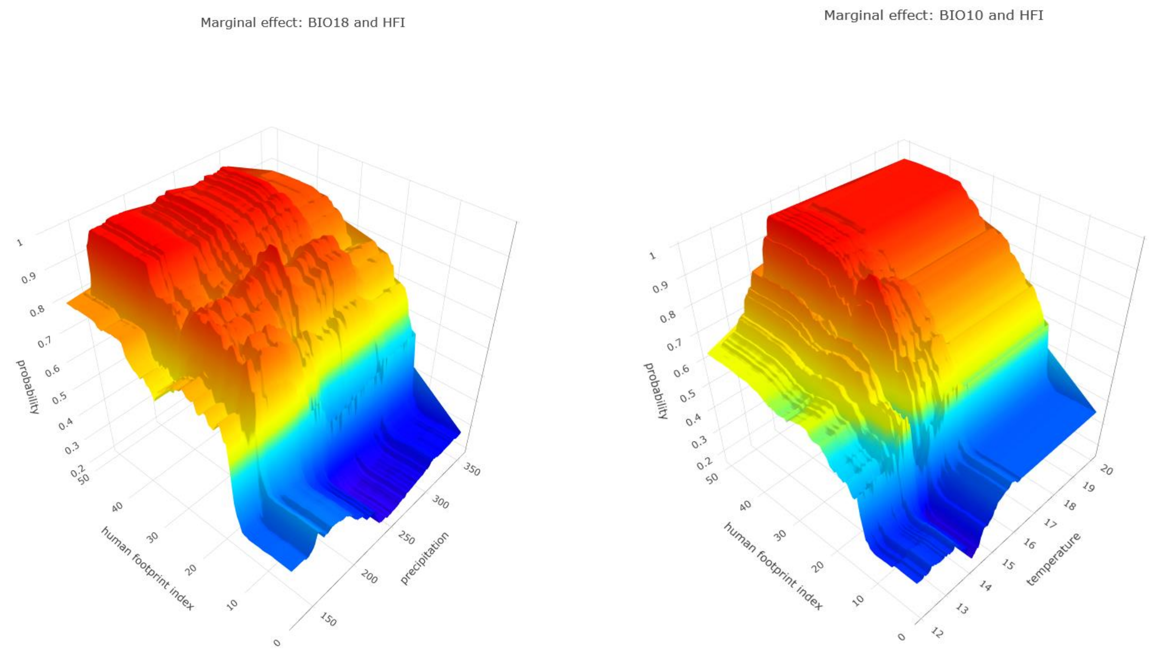

*Figure 6**: Interaction Plots**. On the left: Precipitation and Human Footprint Index interaction. On the Right: Temperature and Human Footprint Index interaction. When two variables are plotted against the predicted probability of nest site occurrences, remaining predictors are controlled at their mean values as constants.*

3.1 Discussion

Using the concept of ensemble modelling with nice different classification models, we (1) assessed the impact of seven environmental factors on swift nest site selection during the summer breeding season in the UK, (2) and forecasted British Swift's potential habitat in the future climate conditions. Each of the nine models have given their predictions on swift nesting hotspots in the UK with probability maps representing the likelihoods of finding swift nests at different parts of the country. Different models have different assessments on variable importance, but most of them, except MARS, agree that Human footprint index is the largest contributing factor to model predictions on swift nests occurrences. However, this result could be misleading due to the sampling bias. It is expected that the swift nest sites in this study would concentrate in the areas with strong human impacts since a large proportion of data in Swift Survey are reported by public sightings. In contrast, pseudo-absent points are more likely to be present in the wild given the way they are generated. Thus, the effect of hfp variable might not be purely biological, but rather artificial. If *hfp's* significance is not a result of sampling bais or incorrect pseudo-absent point generation, we might be able to conclude that swifts in the UK incline to build their nests in artificial environments like cities and towns whose biome is heavily influenced by humans. This can be also be seen from the consistent patterns between model predicted swift nests distribution map and the heatmap of *hfp* (*figure 5* and *figure 6*). Random Forest has the best performance in making correct predictions on swift nest occurrences among nine models we used. By 2050, Average temperature during summer in midland England, in which most of the current swift nests are found, is projected to rise from 15.34 °C to 17.53 °C, and the average precipitation will drop from 186.84 mm to 149.45 mm. Climate change is most significant around England major cities, probably resulting from large greenhouse gas emissions. Thus, according to RF's prediction, midland England may no longer be the optimum nesting choice for British Swifts in the near future. Instead, Northern British cities such as Edinburg, Glasgow, Newcastle and Belfast will become popular as the temperature/precipitation continue to increase/decrease. Nonetheless, our prediction did not take into consideration of (1) the energy expenditure of longer distance migration for swifts in trade of better habitats. (2) Strong correlations between variables as demonstrated by Principle component analysis: Prediction results should still be valid, but variable effects interpretation and importance evaluation should not be this straight forward. More importantly, swift range shift prediction in this study is only driven by the future climate changes while human activity's dominant impact on swift nest site selection has been ignored

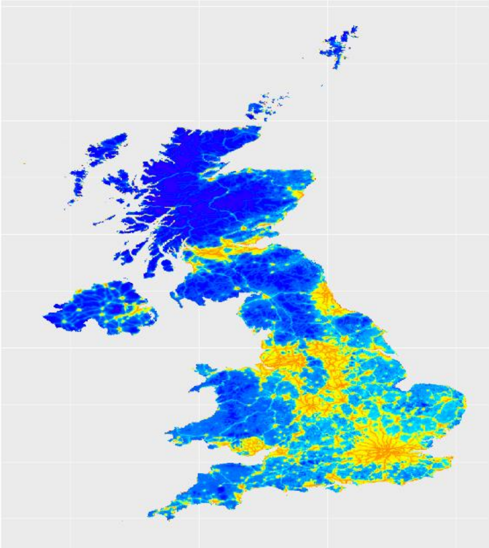

Figure 7: ***Human Foot Print Index Heatmap***

(assumed to be consistent across years). British swifts' future distribution range and population dynamic will still mainly depends on the future human activities and urban planning in the UK.

Reference

1. Araújo, M. & Rahbek, C. How Does Climate Change Affect Biodiversity? *Science (New York, N.Y.)* **313**, 1396–7 (2006).
2. Beaumont, L. J., Pitman, A. J., Poulsen, M. & Hughes, L. Where will species go? Incorporating new advances in climate modelling into projections of species distributions. *Global Change Biology* **13**, 1368–1385 (2007).
3. Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B. & Thomas, C. D. Rapid Range Shifts of Species Associated with High Levels of Climate Warming. *Science* **333**, 1024–1026 (2011).
4. Buermann, W. *et al.* Predicting species distributions across the Amazonian and Andean regions using remote sensing data. *Journal of Biogeography* **35**, 1160–1176 (2008).
5. Colwell, R. K., Brehm, G., Cardelús, C. L., Gilman, A. C. & Longino, J. T. Global Warming, Elevational Range Shifts, and Lowland Biotic Attrition in the Wet Tropics. *Science* **322**, 258–261 (2008).
6. Moore, T. T. CLIMATE CHANGE AND ANIMAL MIGRATION. *Environmental Law* **41**, 393–405 (2011).
7. Sinclair, S., White, M. & Newell, G. How Useful Are Species Distribution Models for Managing Biodiversity under Future Climates? *Ecology and Society* **15**, (2010).
8. Chen, X., Dimitrov, N. B. & Meyers, L. A. Uncertainty analysis of species distribution models. *PLOS ONE* **14**, e0214190 (2019).
9. Martínez-Minaya, J., Cameletti, M., Conesa, D. & Pennino, M. G. Species distribution modeling: a statistical review with focus in spatio-temporal issues. *Stoch Environ Res Risk Assess* **32**, 3227–3244 (2018).

10. Forester, B. R., DeChaine, E. G. & Bunn, A. G. Integrating ensemble species distribution modelling and statistical phylogeography to inform projections of climate change impacts on species distributions. *Diversity and Distributions* **19**, 1480–1495 (2013).
11. Hedenström, A. *et al.* Annual 10-Month Aerial Life Phase in the Common Swift Apus apus. *Current Biology* **26**, 3066–3070 (2016).
12. Åkesson, S., Klaassen, R., Holmgren, J., Fox, J. W. & Hedenström, A. Migration Routes and Strategies in a Highly Aerial Migrant, the Common Swift Apus apus, Revealed by Light-Level Geolocators. *PLOS ONE* **7**, e41195 (2012).
13. Swift Bird Conservation Project - The RSPB. https://www.rspb.org.uk/our-work/conservation/conservation-and-sustainability/safeguarding-species/help-us-help-swifts/.
14. Swift Survey. *The RSPB* https://www.swiftsurvey.org/Rspb/Home/Index.
15. Fick, S. E. & Hijmans, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315 (2017).
16. Gent, P. R. *et al.* The Community Climate System Model Version 4. *J. Climate* **24**, 4973–4991 (2011).
17. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. R package version 2.9-5. (2019).
18. Bivand, R., Keitt, T. & Rowlingson, B. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.4-4. (2019).
19. Wood, S. N. Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* **99**, 673–686 (2004).
20. Friedman, J. H. Multivariate Adaptive Regression Splines. *Ann. Statist.* **19**, 1–67 (1991).
21. Hastie, T., Tibshirani, R. & Buja, A. Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association* **89**, 1255–1270 (1994).
22. Ripley, B. tree: Classification and Regression Trees. R package version 1.0-40. (2019).
23. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
24. Venables, W. & Ripley, B. Modern Applied Statistics With S. in *Springer* (2002). doi:10.1007/b97626.
25. Chen, T. & He, T. xgboost: eXtreme Gradient Boosting. 4.
26. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

27. Revelle, W. R. psych: Procedures for Personality and Psychological Research. (2017).

28. Kuhn, M. caret: Classification and Regression Training. *Astrophysics Source Code Library* ascl:1505.003 (2015).