Comparison of Three Modeling Techniques To predict the Spatial Distribution and Environmental Preferences of Red Kite *(Milvus, milvus)*.

ABSTRACT

Intro: As a recovering species it is important to understand the habitat preferences of Red Kites. Species distribution models have become powerful tools in understanding habitat distribution. However, the effect of using different models on predictions remains unclear especially when using spatially biased species presence data.

Methods: Maximum entropy (MaxEnt), boosted regression trees (BRT) and Random Forests (RF) predictions of Red Kite habitat were generated for Wales from spatially biased presence only data. 15 iterations were produced and averages were compared.

Results: All models obtained a AUC greater than 0.7. However, the emphasis on environmental variables differed, leading to dissimilar spatial predictions of habitat suitability.

Conclusions: Due to deviation in spatial predictions of Red Kite habitat, conservation efforts should employ an ensemble of species distribution models. Results should be interpreted with reference to any bias in the species distribution data.

Keywords:
Red Kite, Species Distribution Models, MaxEnt, Boosted Regression Trees, Random Forests, Bias.

INTRODUCTION

Global biodiversity loss is occurring at unprecedented rates (Sala *et al.*, 2000; Atley and Morad, 2009; Durant, 2014; McCallum, 2015). Therefore, it is important to understand the environmental changes which increase biodiversity (Yusoff, 2011; Seippel *et al*., 2012). In the UK in 1946 only seven breeding pairs of Red Kite (*Milvus milvus)* existed in mid-Wales (Davis, 1993; Newton *et al.*, 1996; Evans *et al*., 1999), However there are now 1600 (Evans *et al.*, 2008; RSPB, 2012). It is critical to understand the nature of Red Kite habitat in order to protect its current extent and to facilitate biodiversity increase, through reintroduction of species with similar characteristics (Evans *et al*., 2008).

However, the factors limiting the global and national scale dispersal of Red Kites remain largely unknown (Heuck *et al*., 2013). Species observation data is typically limited to subsets due to a lack of resources and logistical constraints (Elith *et al.*, 2006; Oppel *et al*., 2012). However, the advent of inexpensive computing power and remote sensing products, has led to increases in the use of statistical models, to predict species distributions across large areas (Elith and Leathwick, 2009; Tremblay *et al*., 2009; Hoffmann *et al.*, 2014). For example, a study by Heuck *et al.* (2013) used land cover products and generalised additive models (GAM) to investigate the effect of land cover on Red Kite distribution in Germany. The study concluded that the interactions between anthropogenic and climatic variables were important in determining habitat.

Numerous different species distribution models (SDM) have been used to map the extent of terrestrial bird habitat (Gutzweiller and Barrow, 2001; Schulte *et al.*, 2005; Fuller *et al.*, 2007; Vallecillo *et al.*, 2009). Although model performance has been compared several times (Segurado and Araújo, 2004; Elith *et al*., 2006; Elith and Graham, 2009; Marmion *et al*., 2009a) the exact effect different SDM methods have on predictions remains unclear (Elith and Graham, 2009). Especially when handling biased, low-density species sighting data (Luck, 2007; Hernandez *et al.*, 2006; Opell *et al*., 2011). However, the majority of species information is randomly sampled presence-only data (Evans and Hammond, 2004; Elith *et al.*, 2006; Opell *et al*., 2011; Buk and Knight, 2012; Zhu *et al.*, 2015). As SDM outputs are a function of both data quality and handling it is important to understand how these factors influence each other (Hernandez *et al*., 2006). Especially as there is growth in the use of SDM to set conservation priorities with the aim of increasing biodiversity (Araujo, 2005; Opell *et al*., 2011).

This study aims to compare the deviation in three SDM predictions of Red Kite habitat across Wales. Results will inform conservation practices about the variation in modelling techniques. In addition, the study aims to explore the effect of unrepresentative sighting data on SDM predictions.

SPECIES DISTRIBUTION MODELS

Universally SDM require species location data and environmental variables (Elith *et al*., 2006; Marmion *et al.*, 2009a). SDM assume sighting locations exhibit the environmental characteristics of good habitat (Elith *et al.*, 2006). Therefore, it is important the samples are representative and unbiased to produce the most accurate results (Elith and Leathwick, 2009; Marmion *et al*., 2009b). A study by Hernandez *et al.* (2006) found some SDM, like MaxEnt, can produce meaningful results with 5-10 sightings. However, a higher number is preferable as it allows the model to refine the relationship between predictor and variables (Elith *et al.*, 2006).

Machine learning SDM have gained popularity due to their ability to handle large iteration numbers and accurately fit non-parametric data (Hastie *et al*., 2009; Elith and Leathwick, 2009). Three commonly used SDM are maximum entropy (MaxEnt), boosted regression trees (BRT) and random forests (RF) (Elith *et al.*, 2006; Elith and Leathwick, 2009).

MaxEnt contrasts presence locations and background values, derived from environmental parameters, to calculate species suitability (Phillips *et al*., 2006; Guillera-Arroita *et al.*, 2012). The theory operates on the principle that, of the numerous possible species distributions, the one which fits environmental constraints and is associated with maximum information entropy is correct (Phillips *et al*., 2006). MaxEnt also allows for consideration of sampling bias by weighting data (Phillips *et al.,* 2006).

Unlike MaxEnt, BRT and RF require presence and absence locations (Hastie *et al*., 2009). Where solely presence data is available pseudo-absence can be generated from the un-sampled study area (Phillips *et al.*, 2009). However this introduces error, as absence locations may contain un-sampled presence (Phillips *et al.*, 2009). BRT produce a regression model, from a stage wise progression of classification trees (Elith *et al.*, 2008). Model parameters, like the complexity and learn rate of each tree, can be varied in order to best fit the data (Elith *et al.*, 2008).

RF also constructs a regression model, but bootstraps the data to create many small inaccurate classification trees (Breiman, 2001; Marmion *et al.*, 2009a). Each tree captures different regularities between species and environmental data, which are combined into a predictive model (Breiman, 2001).

THE RED KITE IN WALES, UK

Welsh forests are typically a mixture of conifer and broadleaf deciduous trees (Natural Resources Wales, 2016). These typically have dense canopies which are not favoured by ‘swooping’ Red Kite hunting techniques (Davis and Davies, 1973; RSPB, 2012). However, the agricultural land cleared for crops and livestock may be more suitable for hunting. As the small mammals Red Kites consume are visible and carrion is readily available (Davis and Davis, 1981; Natural Resources Wales, 2016).

In Welsh lowland areas the mean temperature is 9.5-10.5ºC (Met Office, 2013). Temperature decreases by ~0.5ºC for every 100m gain in altitude (Met Office, 2013) and is subject to seasonal variation of ~20ºC (Met Office, 2015). In Europe Red Kites are seasonal visitors. However, Welsh Red Kite are residents, as such they may favour areas with lower seasonal variability (RSPB, 2012). The proximity of Wales to the North Atlantic Oscillation (Hurrell, 1995) coupled with variable topography creates high precipitation levels (Met Office, 2013). Although precipitation does not appear to effect Red Kite habitat, the aspect of the hilly terrain may affect the locations of breeding sites (Newton *et al.*, 1996).

METHODS

RED KITE SIGHTING DATA

In order to obtain enough data to accurately model Red Kite distribution, presence only data from several sources was combined, see Table 1. Data was cleaned to remove duplicates and sightings were georeferenced to the origin of Ordinance survey (OS) grid squares. As previously mentioned, inaccurate sighting data can introduce error, but climatic characteristics do not significantly vary within OS girds (New *et al.*, 2000).

Table 1 | Sources and number of records of Red Kite sighting data.

Source	Sightings
BIS Casual Records 2013	40
Breconshire Birds 2013	20
Miscellaneous Unitary Authorities	24
iRecord	8
MapMate	99
Mixed Taxa Records	10
Casual Records	18
Unknown	25
Total	**244**

ACCOUNTING FOR BIAS

As Figure 1 illustrates sightings were often near roads, conservation areas and population centres. This introduces bias as higher populations often cause increased sighting records of the same number of Red Kites (Oppell *et al.*, 2012; Howard and Davis, 2015). As previously mentioned SDM assume sighting locations to be typical for the species (Elith *et al.*, 2006; Syfert *et al.*, 2013). In MaxEnt bias was accounted by the inclusion of a bias grid, which gave clusters of sightings and those near roads smaller weighting in analysis (Phillips *et al.*, 2006; Syfert *et al.*,2013). In

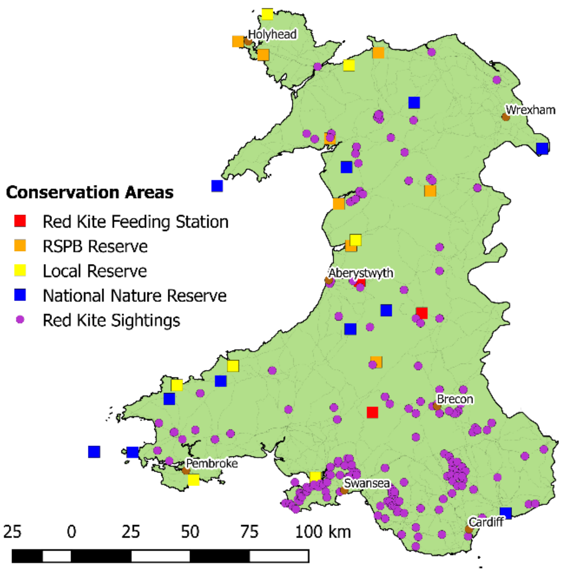

Figure 1 | Spatial bias of sighting data near conservation areas, population centres and main roads.

order to account for bias in the BRT and RF models, sightings within 2500m of another were removed. Furthermore, absence points were only generated within 500m of a primary road in order to counteract the bias in the presence data (Phillips *et al.*, 2009; Lobo and Tognelli, 2011; Barbat-Massin *et al.*, 2012).

ENVIRONMENTAL DATA

Relevant environmental variables were obtained from a variety of sources, see Table 2. Variable collinearity was tested (Frank, 2016). Where collinearity was calculated >0.7 one variable was removed.

MODEL PARAMETERS

See Table 3 for list of modelled environmental variables.

MAXENT

MaxEnt version 3.3.3K was used to model Red Kite distribution. MaxEnt was selected due to its increasing popularity as a SDM and ability to model with small numbers of sighting data sets (Phillips *et al.*, 2006; Hernandez *et al.*, 2006). The convergence percentage was set at 0.00001. In order to generate precision data, the model was run fifteen times under a bootstrap replication with 20% of the sighting data used for testing (Elith *et al.*, 2006).

Table 3 | Environmental and anthropogenic variables used in the final models.

Environmental	Anthropogenic
Aspect	Area Farmed
Distance to Stream	Distance to Conservation Area
GPP	Number of Crops
Max Precipitation	Habitat Land Cover
Min Temperature	
Mean Temperature	
Slope	
Temperature Seasonality	
TWI	

BOOSTED REGRESSION TREES

The GBM and Dismo packages (Ridgeway, 2015; Robert *et al.*, 2016) were used to model Red Kite distribution. BRT was selected due to its popularity as a SDM (Elith *et al*., 2006). Model parameters were determined using the GBM.Step function (see, Elith *et al.*, 2008) and systematically varying learning rate and bag fraction (Elith and Leathwick, 2016). A tree complexity of 3 provides the lowest levels of predictive deviance for data of under 250 sightings (see, Elith *et al.*, 2008). The final prediction had a learning rate of 0.0001, a bag fraction of 0.5 and produced over 1000 trees. The model was run fifteen times to test precision and 20% of the sighting data was used for testing.

Table 2 | Environmental variables, source and validation.

Environmental Variable	Source	Pixel Size (m)	Validation
Elevation	OS Contours	30	Counters produce a hydrologically correct DEM. Elevation causes variance in macro-climate and slope (MetOffice, 2013).
Slope	Elevation data	30	Slope can affect the 'swooping' of Red Kite hunting (RSPB, 2012).
Aspect	Elevation data	30	Aspect can affect the nesting habits of Red Kites (Newton *et al*., 1996)
Topographic Wetness Index (TWI)	Elevation data	30	Increased soil moisture can lead to increased numbers of mammals, Red Kite prey (Davies and Davis, 1973; RSPB, 2012)
Euclidean Distance to Stream Network	Elevation data and Natural Resources Wales	30	Streams provide water for Red Kites, potential prey and livestock.
Euclidean Distance to Conservation Area	Natural Resources Wales	30	Conservation areas are designed as habitat and may introduce bias in sighting numbers
Landscape Scale Habitat Mapping	Natural Resources Wales	30	Land cover has been found to be important to Red Kite, this land cover classification puts emphasis on habitat type.
Land Cover Classification	Landsat 5 Imagery	30	Nine class land cover classification of Wales.
Gross Primary Productivity (GPP)	MODIS	1000	Vegetation productivity can relate to crop and forest areas, both potential Red Kite habitat (Davies and Davis, 1973).
Area Farmed	2013 Agricultural Census	30	Kriging interpolation of area farmed from data points at 2km intervals. Agricultural areas are considered to be good Red Kite habitat (RSPB, 2012).
Number of Livestock	2013 Agricultural Census	30	Kriging interpolation of livestock numbers from data points at 2km intervals. Livestock keep grass levels short and stop succession which creates a good environment for Red Kites to hunt. In addition, they can provide carrion (RSPB, 2012).
Number of Crops	2013 Agricultural Census	30	Kriging interpolation of crop numbers from data points at 2km intervals. Ploughing and harvesting practices can provide good hunting conditions for Red Kite (RSPB, 2012).
Mean Temperature	BIOCLIM	1000	Climatic Variable
Temperature Seasonality	BIOCLIM	1000	Seasonal variation may have an effect on resident Red Kites (RSPB, 2012).
Maximum Temperature	BIOCLIM	1000	Some areas may be undergo too much seasonal variation.
Minimum Temperature	BIOCLIM	1000	Some areas may be too cold in Winter.
Temperature Range	BIOCLIM	1000	Seasonal variation may have an effect on resident Red Kite habitat (RSPB, 2012).
Maximum Precipitation	BIOCLIM	1000	Climatic Variable, may correlate with areas with more farming and GPP.
Minimum Precipitation	BIOCLIM	1000	Climatic Variable, may correlate with areas of less farming and GPP.

RANDOM FORESTS

The randomforest and rfUtilities packages in R (Liaw and Weiner, 2002; Evans and Murphey, 2014) were used to model Red Kite distribution. As variation of model parameters had little to no impact on model variance RF was run using appropriate defaults (see, Evans and Murphey, 2015 and Breiman *et al*., 2015). The model was fitted using 1001 trees and tested with a training sample of 20%. Fifteen iterations were run to test model precision.

PROCESSING OF OUTPUTS AND ACCURACY ASSESMENT

Outputs were manipulated in R (v.3.2.4), ArcMap (v.10.3) and Microsoft Excel to display comparable data. Area Under Curve (AUC) values, a standard measure of SDM accuracy (Elith *et al.*, 2006; Elith and Leathwick, 2009), was produced for training and test data for each run of every SDM.

RESULTS

MODEL ACCURACY

All iterations of every model were above 0.7 AUC and therefore, are considered to be accurate (Phillips *et al*., 2006). BRT produced the lowest mean AUC scores from both training and testing data, but also had the lowest standard deviation (0.005 and 0.024). RF models training AUC was comparatively high and precise, but the test data had the most standard deviation (0.05), suggesting that some predictions could be considerably imprecise, Figure 2. MaxEnt produced the median training and testing AUC values with similar standard deviation to BRT.

INFLUENCE OF ENVIRONMENTAL VARIABLES

Although the SDM produced similar AUC values, the level of importance assigned to the different environmental variables changed drastically, Figure 3. For example, seasonality was one of the two most influential variables for both BRT and RF, but had the least influence on MaxEnt predictions. In addition, similar rankings of importance of environmental variables were not produced by the same combination of models. In some cases, BRT and RF agree (GPP) and others MaxEnt and RF (maximum precipitation). This suggests that there is no significant relationship between different SDM and level of importance assigned to environmental variables.

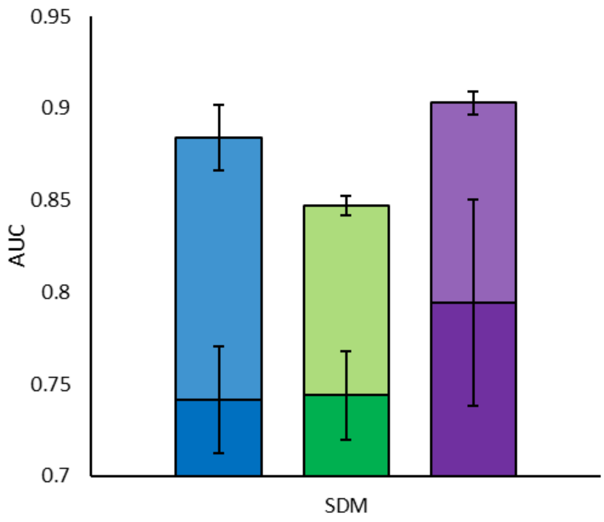

Figure 2 | AUC values for MaxEnt (blue), BRT (green), RF (purple). Test data AUC is denoted by the darker colour. Standard deviation is shown.

RED KITE HABITAT DISTRIBUTION PREDICTIONS

The different levels of influence assigned to environmental variables caused differing predictions of spatial distributions of good habitat, see Figures 3, 5 and 6. Despite the removal of some bias from the clustered sightings in the south east, both BRT and RF mapped highly suitable habitat in this area. MaxEnt combated the data bias most effectively. Figure 5 illustrates that all three models predicted areas of habitat in mid and West Wales, but the exact location of habitat often varied, Figure 6.

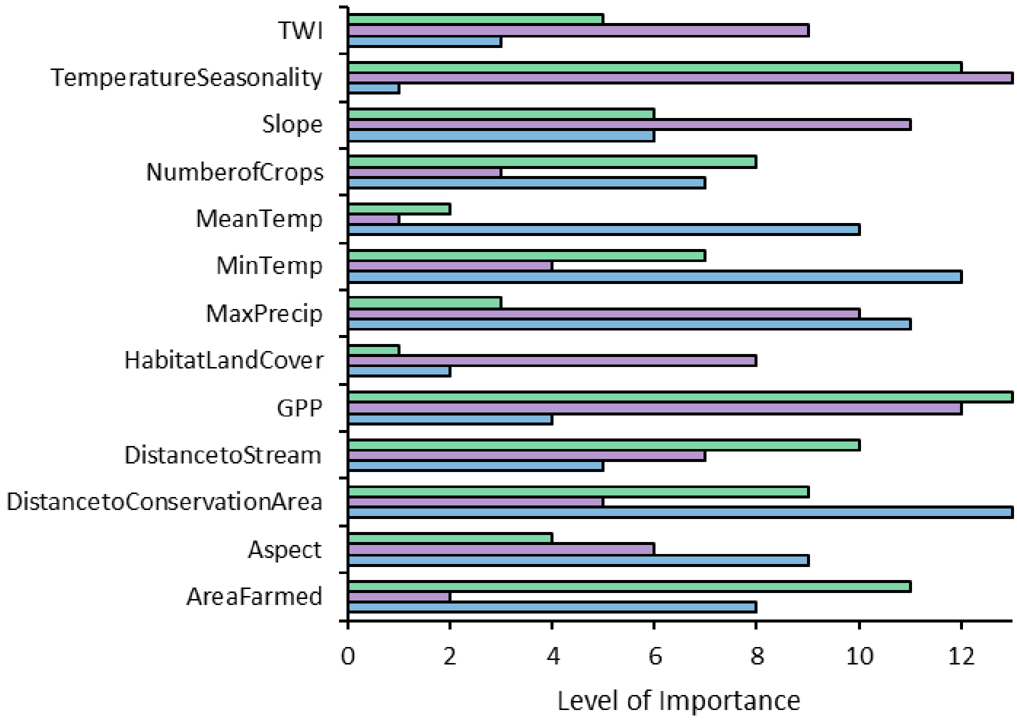

Figure 3 | In order to compare the relative influence across different models environmental variables were standardised into ranks of importance, where thirteen is the most influential on model output. MaxEnt = blue, BRT = green, RF = purple.

DISCUSSION

ENVIRONMENTAL FACTORS FOR RED KITE CONSERVATION IN WALES

Unlike the results in Figure 3, previous distribution modeling of Red Kites found land cover to be an important factor in determining good habitat (Heuck *et al*., 2013). However, Heuck *et al.* (2013) may over represent land cover, as research cautions against the use of numerous land cover products in SDM (Vallecillo *et al*., 2009; Princé *et al.,* 2013; Bryan *et al*., 2015). Land cover is a function of a variety of different processes and it is unclear how effective SDM are at modeling the relationships between these processes, environmental variables and species locations (Smith, 2003; Vallecillo *et al*., 2009).

The differences in environmental variables likely caused the disparity between the studies. For example, agricultural land, an established habitat type of Red Kites (Davies and Davis, 1973; RSPB, 2012), was better represented by other environmental layers like GPP and Area farmed, than land cover. These factors were important across the three SDM results but were unrepresented by Heuck *et al*. (2013) (Figures 3 and 4).

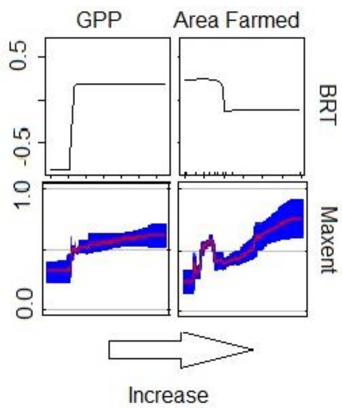

Figure 4 | Fitted function and logistic output of GPP and area farmed variables for BRT and MaxEnt

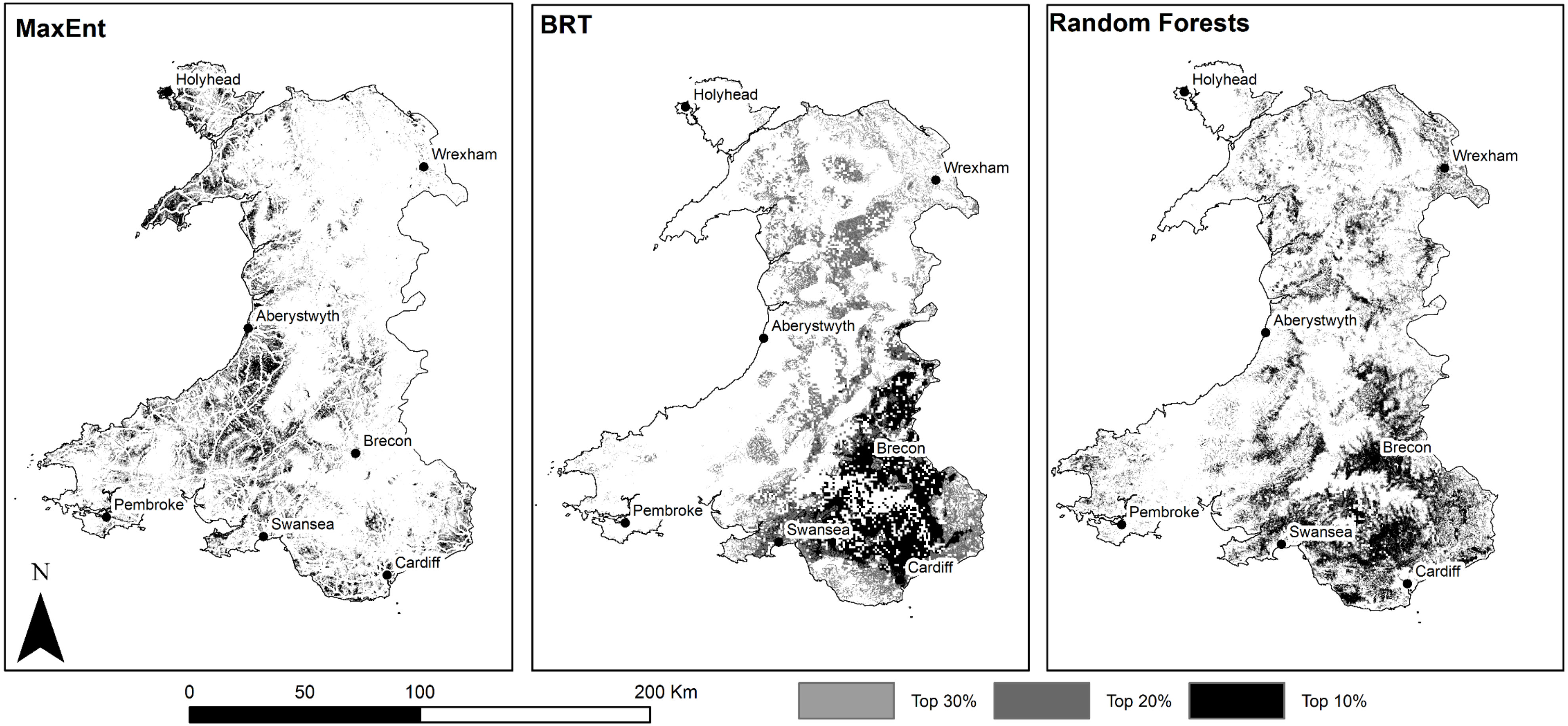

Figure 5 | Locations of habitat suitability displayed in percentiles. Top thirty, twenty, and ten percent for MaxEnt, boosted regression trees and random forests.

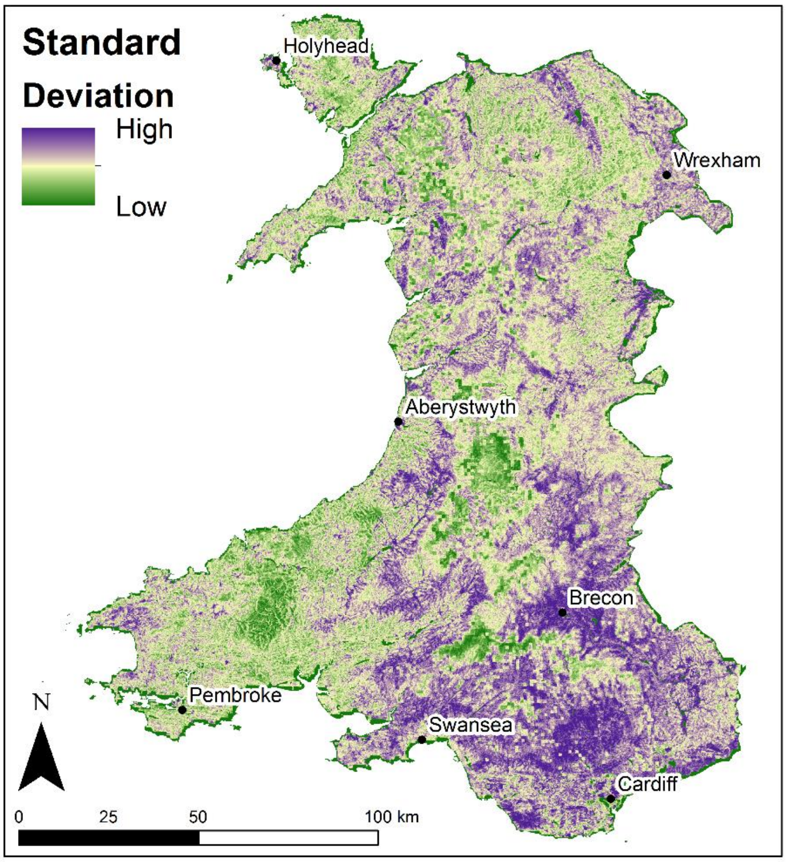

Figure 6 | Standard deviation of each pixel between the MaxEnt, boosted regression trees and random forest models. Areas with high standard deviation are shown in purple and low deviation in green. Areas of greatest standard deviation were generally locations of bias.

Often, the same interactions between Red Kite and the environment can be inferred from the three SDM despite the differing importance assigned to environmental variables (Elith *et al*., 2006; Hernandez *et al*., 2006). The models suggest that temperature seasonality could affect Red Kite habitat. Seasonal temperature variation was a highly influential predictor in both BRT and RF and MaxEnt found both min and mean temperature to be of importance, which could be related to seasonality, see Figure 3 and 7 (Li *et al*., 2015).

MODELS ABILITY TO HANDLE UNREPRESENTATIVE SIGHTING DATA

In order to utilise the majority of current Red Kite sighting statistics, it is important for

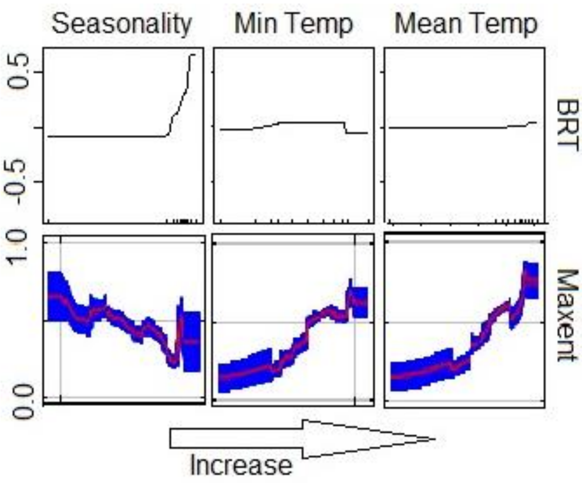

Figure 7 | Fitted function and logistic output of Temperature Seasonality, minimum temperature and mean temperature variables for BRT and MaxEnt models to produce results accurate to ground truth from poor quality sighting data (Elith *et al.*, 2006; Elith and Leathwick, 2009). In this regard MaxEnt removed bias most effectively, likely due to the unique manner in which it handles environmental and background data (Phillips *et al.*, 2006). This finding agrees with other studies conclusions; that MaxEnt is able to produce accurate predictions from small, biased data sets (Hernandez *et al*., 2006; Elith *et al*., 2006; Syfert *et* al.,2013; Hefley and Hooten, 2015; Marshall *et al*., 2015).

However, the lack of deviation in BRT AUC, Figure 2, suggests BRT may define the relationships between environmental variables and Red Kite more precisely. Therefore, accurate sighting data could allow BRT to outperform MaxEnt (Elith *et al*., 2006; Elith and Leathwock, 2009; Couce *et al*., 2012; Hertzog *et al*., 2014). RF most likely requires more species data to produce a model less subject to overfitting with precise test AUC values, Figure 2 (Hernandez *et al.*, 2008; Feeley and Silman, 2011). However, all the models provide useful data regarding the environmental interactions of Red Kites, if the nature of the sighting data is properly understood and accounted for in interpretation of results.

However, this is not always possible. Figures 1 and 8 illustrate that locations close to conservation areas exhibit increased Red Kite presence. This could be due to the use of conservation areas for leisure, creating a bias in sightings (Elith *et al.*, 2006). However, as conservation areas are actively managed to increase biodiversity, the area is also likely to exhibit the characteristics of good Red Kite habitat. Therefore, the effects of bias and habitat are impossible to separate.

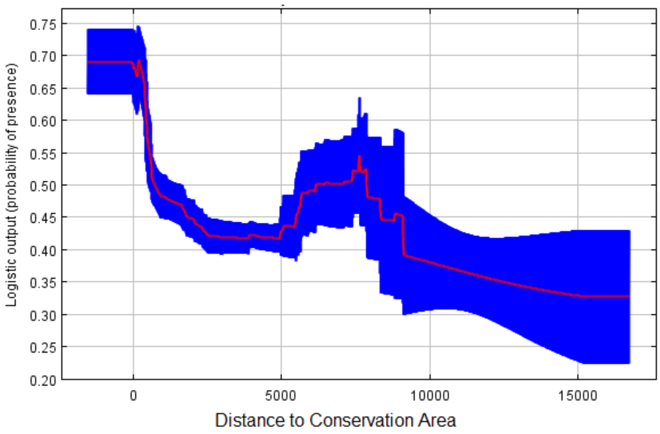

Figure 8 | Effect of increased distance between sighting location and conservation area on computed probability of presence in MaxEnt.

UTILITY OF MODELS FOR DETERMINING CONSERVATION AREAS

Figure 6 illustrates that modelling methods have the potential to impact the selection of geographic regions for conservation. Despite similar accuracy measurements (Figure 2) the nature of predictions varies drastically due to different emphasis and handling of environmental variables (Elith *et al*., 2006; Phillips *et al*., 2006; Elith *et al*., 2008). The differences in fitting and function between techniques are a likely explanation for this phenomenon (Oppel *et al*., 2012; Elith and Graham, 2009). For example, boosted regression and random forest classification trees rely on fitting constants to regions which have homogenous responses to predictors (Elith *et al.*, 2008; Breiman, 2001). If the majority of trees have the same environmental variable in the first branch, this variable will have a large effect on the model prediction (Elith *et al.*, 2006; Elith *et al.*, 2009).

Whereas, MaxEnt determines importance through multiple environmental variables simultaneously (Phillips *et al*., 2006). Therefore, the environmental variables are handled differently, redistributing emphasis and habitat predictions (Hernandez *et al.*, 2006; Oppel *et al*., 2012). This results in similar AUC scores but different selection of habitat, and priorities for conservation, Figure 5.

A study by Oppel *et al*. (2012) compared five SDM, including MaxEnt, BRT and RF to determine seabird conservation areas. They concluded that similar accuracy and precision does not equate to the same emphasis of environmental variables and distribution predictions, a finding that is consistent with other studies (Elith and Graham, 2009; Ready *et al*., 2010). However as previously discussed, all predictions provide information regarding the environmental parameters of Red Kite habitat, which can be used in conjunction to improve overall conservation efforts (Araújo and New, 2007; Coetzee *et al*.,2009; Farrand *et al*., 2011; Marmion *et al*., 2009; Oppel *et al*., 2012).

CONCLUSION

All three models produced predictions with AUC values greater than 0.7. MaxEnt removed the bias of the sighting data most effectively. However, both BRT and RF yielded suitable predictions, if the nature of the sighting data was accounted for during interpretation. The SDM suggests that Red Kite habitat is dictated by agricultural activity, low seasonal temperature variation and the presence of nearby conservation areas.

Although environmental preferences of Red Kites were approximately determined, the various statistical methods placed importance on differing environmental variables. This produced contrasting spatial habitat predictions and illustrates why a single species distribution model is insufficient to determine Red Kite conservation areas. However, an ensemble of SDM can produce useful and accurate predictions, even from unrepresentative species data.

Future conservation efforts could incorporate a variety of modeling techniques, environmental variables and available species presence data. Research is needed to understand the causes of varying levels of environmental importance and the effect of using different variables.

REFERENCES

Araujo, M., Thuiller, W., Williams, P. and Reginster, I. (2005). Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecology and Biogeography*, 14(1), pp.17-30.

Araújo, M.B., New,M. (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* 22, 42–47.

Atley, T. and Morad, M. (2009). Mind the gap: the UK's new planning system, landscape conservation and biodiversity. *International Journal of Environmental Studies*, 66(6), pp.785-790.

Barbet-Massin, M., Jiguet, F., Albert, C. and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many?. *Methods in Ecology and Evolution*, 3(2), pp.327-338.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, pp.5-32.

Bryan L. Nuse, Robert J. Cooper, Elizabeth A. Hunter. (2015) Prospects for predicting changes to coastal wetland bird populations due to accelerated sea level rise, *Ecosphere*, , **6**, 12, art286

Buk, K. and Knight, M. (2012). Habitat Suitability Model for Black Rhinoceros in Augrabies Falls National Park, South Africa. *South African Journal of Wildlife Research*, 42(2), pp.82-93.

Coetzee, B.W.T., Robertson, M.P., Erasmus, B.F.N., van Rensburg, B.J., Thuiller, W.,2009. Ensemble models predict Important Bird Areas in southern Africa will become less effective for conserving endemic birds under climate change. *Global Ecology and Biogeography* 18, 701–710.

Couce, E., Ridgwell, A. and Hendy, E. (2012). Environmental controls on the global distribution of shallow-water coral reefs. *Journal of Biogeography*, 39(8), pp.1508-1523.

Davis, P. (1993). The Red Kite in Wales: setting the record straight. *British Birds*, 86(7), pp.295-298.

Davies, P. and Davis, P. (1973). The ecology and conservation of the Red Kite in Wales. *British Birds*, pp.183-224.

Davis, P. and Davis, J. (1981). The food of the Red Kite in Wales. *Bird Study*, 28(1), pp.33-40.

Durant, S. (2014). Non-native species: UK bill could prompt biodiversity loss. *Nature*, 512(7514), pp.253-253.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J.,Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A.,Manion,G.,Moritz,C.,Nakamura, M.,N akazawa, Y.,Overton,J.M.,Peterson,A.T.,P hillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J.,Williams, S., Wisz, M.S. , Zimmermann, N.E., (2006). Novel metho ds improveprediction of species' distributions from occurrence data. *Ecography* 29, 129–151.

Elith,J. and Graham, C.H.,(2009).Dothey?Howdothey? WHYdotheydiffer?Onfindingreasons for differing performances of species distribution models. *Ecography* 32, 66–77.

Elith, J., Leathwick, J. and Hastie, T. (2008). A working guide to boosted regression trees.*Journal of Animal Ecology*, 77(4), pp.802-813.

Elith, J., Leathwick, J.R., (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Reviewof Ecology, Evolution, andSystematics* 40, 677–697

Elith, J. and Leathwick, J. (2016). Boosted Regression Trees for ecological modeling. *CRAN*.

Evans, I., Cordero, P. and Parkin, D. (2008). Successful breeding at one year of age by Red Kites Milvus milvus in southern England. *Ibis*, 140(1), pp.53-57.

Evans, J. and Murphy, M. (2015). Package 'rfUtilities'. *CRAN*.

Evans JS and Murphy MA (2014). 'rfUtilities'. R package version 1.0-0, <URL: http://CRAN.R-project.org/package=rfUtilities>.

Evans, I. and Hammond, P. (2004) Monitoring cetaceans in European waters. *Mammal Review*, 34(1-2), pp.131-156.

Evans, I., Summers, R., O'toole, L., Orr-Erwing, D., Evans, R., Snell, N. and Smith, J. (1999) Evaluating the success of translocating Red Kites Milvus milvus to the UK. *Bird Study.* **46**(2) pp. 129-144.

Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2016). Hmisc: Harrell Miscellaneous. R package version 3.17-2. http://CRAN.R-project.org/package=Hmisc

Feeley, K. and Silman, M. (2011). Keep collecting: accurate species distribution modelling requires more collections than previously thought. *Diversity and Distributions*, 17(6), pp.1132-1140.

Fuller, R., Devereux, B., Gillings, S., Hill, R. and Amable, G. (2007). Bird distributions relative to remotely sensed habitats in Great Britain: Towards a framework for national modelling. *Journal of Environmental Management*, 84(4), pp.586-605.

Guillera-Arroita, G., Lahoz-Monfort, J. and Elith, J. (2014). MaxEnt is not a presence-absence method: a comment on Thibaud et al . *Methods Ecol Evol*, 5(11), pp.1192-1197.

Gutzwiller, K. and Barrow, W. (2001). Bird-Landscape Relations in the Chihuahuan Desert: Coping with Uncertainties about

Predictive Models. *Ecological Applications*, 11(5), p.1517.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Hegel,T.M.,Cushman,S.A.,Evans, J.,Huettmann,F.,(2010). Current state of the art for statistical modelling of species distributions. In: Cushman, S., Huettmann, F.(Eds.), Spatial Complexity, Informatics, and Wildlife Conservation. Springer,Tokyo, pp. 273–311

Hefley, T. and Hooten, M. (2015). On the existence of maximum likelihood estimates for presence-only data. *Methods Ecol Evol*, 6(6), pp.648-655.

Hernandez, P., Franke, I., Herzog, S., Pacheco, V., Paniagua, L., Quintana, H., Soto, A., Swenson, J., Tovar, C., Valqui, T., Vargas, J. and Young, B. (2008). Predicting species distributions in poorly-studied landscapes. *Biodiversity and Conservation*, 17(6), pp.1353-1366.

Hernandez, P., Graham, C., Master, L. and Albert, D. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5), pp.773-785.

Hertzog, L., Besnard, A. and Jay-Robert, P. (2014). Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. *Diversity and Distributions*, 20(12), pp.1403-1413.

Heuck, C., Brandl, R., Albrecht, J. and Gottschalk, T. (2013). The potential distribution of the Red Kite in Germany. *J Ornithol*, 154(4), pp.911-921.

Hoffmann, A., Penner, J., Vohland, K., Cramer, W., Doubleday, R., Henle, K., Kõljalg, U., Kühn, I., Kunin, W., Negro, J., Penev, L., Rodríguez, C., Saarenmaa, H., Schmeller, D., Stoev, P., Sutherland, W., Ó Tuama, É., Wetzel, F. and Häuser, C. (2014). The need for an integrated biodiversity policy support process – Building the European contribution to a global Biodiversity Observation Network (EU BON). *Nature Conservation*, 6, pp.49-65.

Howard, E. and Davis, A. (2015). Investigating Long-Term Changes in the Spring Migration of Monarch Butterflies (Lepidoptera: Nymphalidae) Using 18 Years of Data From Journey North, a Citizen Science Program. *Ann Entomol Soc Am*, 108(5), pp.664-669.

Hurrell, J. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, 269(5224), pp.676-679.

Li, Y., Zeng, Z., Zhao, L. and Piao, S. (2015). Spatial patterns of climatological temperature lapse rate in mainland China: A multi-time scale investigation. *Journal of Geophysical Research: Atmospheres*, 120(7), pp.2661-2675.

Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Liaw, A. and Weiner, M. (2015). Package ‘randomForest’. *CRAN*.

Lobo, J. and Tognelli, M. (2011). Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, 19(1), pp.1-7.

Luck, G. (2007). A review of the relationships between human population density and biodiversity. *Biological Reviews*, 82(4), pp.607-645.

Marmion, M., Luoto, M., Heikkinen, R. and Thuiller, W. (2009b). The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecological Modelling*, 220(24), pp.3512-3520.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. and Thuiller, W. (2009a). Evaluation of consensus methods in predictive species distribution modelling. *Diversity and Distributions*, 15(1), pp.59-69.

Marshall, L., Carvalheiro, L., Aguirre-Gutiérrez, J., Bos, M., de Groot, G., Kleijn, D., Potts, S., Reemer, M., Roberts, S., Scheper, J. and Biesmeijer, J. (2015). Testing projected wild bee distributions in agricultural habitats: predictive power depends on species traits and habitat type. *Ecol Evol*, 5(19), pp.4426-4436.

Met Office (2013). Wales: climate. [online] Available at: http://www.metoffice.gov.uk/climate/uk/regional-climates/wl.

Met Office (2015). UK Mean Temperatures [online] Available at: http://www.metoffice.gov.uk/climate/uk/summaries

McCallum, M. (2015). Vertebrate biodiversity losses point to a sixth mass extinction.*Biodiversity and Conservation*, 24(10), pp.2497-2519.

Murphy MA, Evans JS and Storfer A (2010). "Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics." _Ecology_, pp. 252-261.

NaturalResourcesWales. (2016). *Natural Resources Wales / Forestry*. [online] Available at: http://naturalresources.wales/forestry/?lang=en [Accessed 17 Mar. 2016].

New, M., Hulme, M. and Jones, P. (2000). Representing Twentieth-Century Space–Time Climate Variability. Part II: Development of 1901–96 Monthly Grids of Terrestrial Surface Climate. *Journal of Climate*, 13(13), pp.2217-2238.

Newton, I., Davis, P. and Moss, D. (1996). Distribution and Breeding of Red Kites Milvus milvus in Relation to Afforestation and Other Land-Use in Wales. *The Journal of Applied Ecology*, 33(2), p.210.

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

Phillips, S., Anderson, R. and Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4), pp.231-259.

Phillips, S., Dudík, M., Elith, J., Graham, C., Lehmann, A., Leathwick, J. and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1), pp.181-197.

Princé, K., Lorrillière, M., Barbet-Massin, F. and Jiguet F. (2013) Predicting the fate of French bird communities under agriculture and climate change scenarios, *Environmental Science & Policy*, **33**, 120

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ready, J., Kaschner, K., South, A.B., Eastwood, P.D., Rees, T., Rius, J., Agbayani, E.,Kullander, S., Froese, R., 2010. Predicting the distributions of marine organisms at the global scale. Ecological Modelling 221, 467–478

Greg Ridgeway with contributions from others (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. https://CRAN.R-project.org/package=gbm

Robert J. Hijmans (2015). raster: Geographic Data Analysis and Modeling. R package version 2.5-2. https://CRAN.R-project.org/package=raster

Robert J. Hijmans, Steven Phillips, John Leathwick and Jane Elith (2016). dismo: Species Distribution Modeling. R package version 1.0-15. https://CRAN.R-project.org/package=dismo

Roger Bivand, Tim Keitt and Barry Rowlingson (2015). rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.1-3. https://CRAN.R-project.org/package=rgdal

RSPB, (2012). *Red kite*. [online] The RSPB. Available at: http://www.rspb.org.uk/discoverandenjoyn ature/discoverandlearn/birdguide/name/r/redkite/ [Accessed 17 Mar. 2016].

Sala, O., Ill, S., Armesto, J. *et al.* (2000). Global Biodiversity Scenarios for the Year 2100 . *Science*, 287(5459), pp.1770-1774.

Schulte, L., Pidgeon, A. and Mladenoff, D. (2005). One Hundred Fifty Years of Change in Forest Bird Breeding Habitat: Estimates of Species Distributions. *Conservation Biology*, 19(6), pp.1944-1956.

Seippel, Ø., De Marchi, B., Garnåsjordet, P. and Aslaksen, I. (2012). Public opinions on biological diversity in Norway: Politics, science, or culture?. *Norsk Geografisk Tidsskrift - Norwegian Journal of Geography*, 66(5), pp.290-299.

Segurado, P., Araújo, M.B., (2004). An evaluation of methods for modelling speciesdistributions. *Journal of Biogeography* 31, 1555–1568.

Smith, J. (2003). Land-Cover Assessment of Conservation and Buffer Zones in the BOSAWAS Natural Resource Reserve of Nicaragua. *Environmental Management*, 31(2), pp.252-262.

Syfert, M., Smith, M. and Coomes, D. (2013). The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. *PLoS ONE*, 8(2), p.e55158.

Tremblay, Y., Bertrand, S., Henry, R.W., Kappes, M.A., Costa, D.P., Shaffer, S.A., (2009).
Analytical approaches to investigating sea bird-environment interactions: areview.
Marine *Ecology Progress Series* 391, 153–163.

Vallecillo, S., Brotons, L. and Thuiller, W. (2009). Dangers of predicting bird species distributions in response to land-cover changes. *Ecological Applications*, 19(2), pp.538-549.

Yusoff, K. (2011). Aesthetics of loss: biodiversity, banal violence and biotic subjects.*Transactions of the Institute of British Geographers*, 37(4), pp.578-592.

Zhu, A., Zhang, G., Wang, W., Xiao, W., Huang, Z., Dunzhu, G., Ren, G., Qin, C., Yang, L., Pei, T. and Yang, S. (2015). A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. *International Journal of Geographical Information Science*, 29(10), pp.1864-1886.