Modelling the distribution of the tick *Ixodes ricinus* in England and Wales using passive surveillance data from citizen science reports

Authors: Mark Gideon Burdon* (UK Health Security Agency CDO Group), Maximilian Ayling (UK Health Security Agency CDO Group), Nyall Jamieson (University of Manchester), Julie Day (UK Health Security Agency CDO Group), Jolyon Medlock (UK Health Security Agency CSO Group), Kayleigh Hansford (UK Health Security Agency CSO Group), G. R. William Wint (University of Oxford), Thomas Ward (UK Health Security Agency CDO Group)

*Corresponding author: Mark Gideon Burdon (UK Health Security Agency CDO Group), email: mark.burdon@ukhsa.gov.uk

Abstract

Background: The tick *Ixodes ricinus* is the most common tick species in the UK and a significant vector of *Borrelia burgdorferi* s.l. (causative agent of Lyme borreliosis) and Tick-Borne Encephalitis virus (TBEv) to humans and *Anaplasma phagocutphilum, Babesia divergens* and louping ill virus to animals.

Methods: The Tick Surveillance Scheme (TSS) administered by the UK Health Security Agency (UKHSA) contains validated reports of tick encounters from the last twenty years sent in by human and animal health providers, as well as members of the public. We modelled the probability of tick presence across England and Wales using data sourced from the TSS and a combination of biotic and abiotic factors. TSS presence records between 2013 and 2023 are combined with background points generated through a combination of random sampling and target group sampling. An ensemble of statistical and machine learning models were then trained to classify points as presence or background.

Results: The ensemble model had an out-of-sample continuous Boyce index of 0.99 and area under the receiver-operator curve (ROC AUC) of 0.84 on 2024 testing data. The greatest contributors to ROC AUC were variables relating to roe deer (*Capreolus capreolus*) distribution and land cover type. Normalised Difference Vegetation Index and other climatic variables made little contribution to the model's performance. Most of southern England, as well as other areas with known tick populations such as the New Forest and the Lake District, are assigned some of the highest predicted probabilities of tick presence.

Interpretation: Unstructured citizen science data was suitable for creating a high-performing species distribution model for *I. ricinus* after addressing spatial and demographic biases. This model is now being used to inform local public health awareness showing the advantage of passive surveillance through to modelling and public health awareness.

Introduction

The early 21st century has seen a notable shift in the incidence and distribution of vector-borne diseases in Europe. A range of causes have been proposed to explain this change, including climate change, globalisation, changes in land use and ecological factors (Medlock & Leach, 2015; Semenza & Suk, 2017). Because of the differences in behaviour, life cycle and host ecology between tick and mosquito vectors, these are expected to respond differently to future climate trends, with tick species' range changing steadily as rainfall and temperature patterns shift (Ogden & Lindsay, 2016) as well as differences in host distributions, which are also impacted by climatic changes.

The tick *Ixodes ricinus* is the primary vector of *Borrelia burgdorferi s.l.*, the causative agent of Lyme disease. Infection rates of *Borrelia* in *I. ricinus* at a national scale have been assessed for England and Wales as ~4% (Cull et al., 2021) with some regional studies showing infection rates can vary from 4-13% over a five year period (Medlock et al., 2022). Since 2019, *I. ricinus* has been confirmed as a vector of Tick-borne encephalitis virus in the UK, with now a small number of probable and confirmed cases recorded having been acquired locally (Holding et al., 2020). From a veterinary perspective, *I. ricinus* is a known vector of louping ill virus (Gilbert, 2015), tick-borne fever (caused by *Anaplasma phagocytophilum*) (S. Gandy et al., 2022) and bovine babesiosis (caused by *Babesia divergens*) (S. Gandy et al., 2024).

Each year since 2015, there have been between 1,000 and 2,000 laboratory-confirmed cases of Lyme disease in England and Wales, and cases were rising before the COVID-19 pandemic (UKHSA, 2022). The true incidence of Lyme disease is estimated to be much higher when cases treated without laboratory testing are taken into account (Mavin et al., 2024). *Ixodes ricinus* ticks appear to have also spread into new areas of England since the early 2000s (S. L. Gandy et al., 2023; Medlock et al., 2013), coinciding with a steady increase in Lyme case numbers. Lyme incidence varies widely between regions (Office for Health Improvement and Disparities, 2025). A central component to understanding and mitigating the risk to human health for diseases like Lyme disease and tick-borne encephalitis is a reliable understanding of where *I. ricinus* populations are likely to be found and where humans might come into contact with ticks. Targeted field surveillance for ticks carried out by scientifically trained professionals can be resource intensive and therefore have limited spatial and temporal coverage (Hochachka et al., 2012; Ribeiro et al., 2019). Citizen science projects to record species observations are therefore highly valuable for use in passive surveillance of infectious disease vectors (see also Nieto et al. (2018)), which can be enhanced when combined with verification processes (through submission of samples for expert identification) to validate records. Such projects are widespread in Europe and elsewhere for ticks and mosquitoes, and are widely used to supplement formal surveillance programmes (e.g. Mosquito-Alert, Pragmatick, Citique). Ensuring this identification validation process is critical to prevent morphologically similar species (for example *Ixodes hexagonus*), with varying ecologies, being misclassified. The ability to estimate spatial vector distributions from unstructured citizen science data is important for public health and, at the time of writing, no detailed tick risk map for England and Wales was accessible online.

This study models the spatial distribution of *I. ricinus* across England and Wales using the Tick Surveillance Scheme (TSS) data collected by the Medical Entomology group in the UK Health Security Agency (UKHSA, previously in Public Health England and Health Protection Agency). We include a range of environmental and host datasets in an ensemble model comprised of a penalised generalised linear model (pGLM), a generalised additive model (GAM), an extreme gradient boosted trees (XGBoost) model and a Random Forest model.

Methodology

Data

Vector Presences

This study sourced *I. ricinus* presence records from the TSS administered by UKHSA. The TSS contains expert validated reports of tick encounters sent in by human and animal health professionals and the public (with tick submissions after a bite of a human or animal). Submissions must include a live or dead tick specimen, and include a form that provides details of how and where the tick was encountered. The form requests information on the date the tick was found; any recent travel of the host; the type of host (e.g. adult, child, animal); the geographic location where the tick was likely acquired; the part of the

body where the tick was found; and any other information on the tick encounter. UKHSA's Medical Entomology and Zoonoses Ecology team identify each specimens to species level using morphological keys. The team also interpret and compile the information and record geographic coordinates for the likely acquisition location, noting how these were derived. Multiple tick specimens can be reported in one single report, but as this study is focused on modelling presence, the number of specimens submitted is disregarded in this analysis.

We used 4,083 TSS records from England and Wales that were verified by UKHSA as *I. ricinus* from between 2013 and 2023 inclusive, with 2024's observations held back for use only in model testing to avoid spatial autocorrelation resulting in inflated performance statistics (Wadoux et al., 2021). Note that some 2024 records had not yet been processed at the time of data extraction, so this does not represent a full year of data. With most of the reports in England and Wales, data from Scotland and Northern Ireland was excluded from the study. Duplicate reports from the same recorder in the same location and year were also excluded. Figure 1 shows the geographical distribution of *I. ricinus* reports included in the model from each year, aggregated into 40km-wide hexagonal areas for easier comparison.

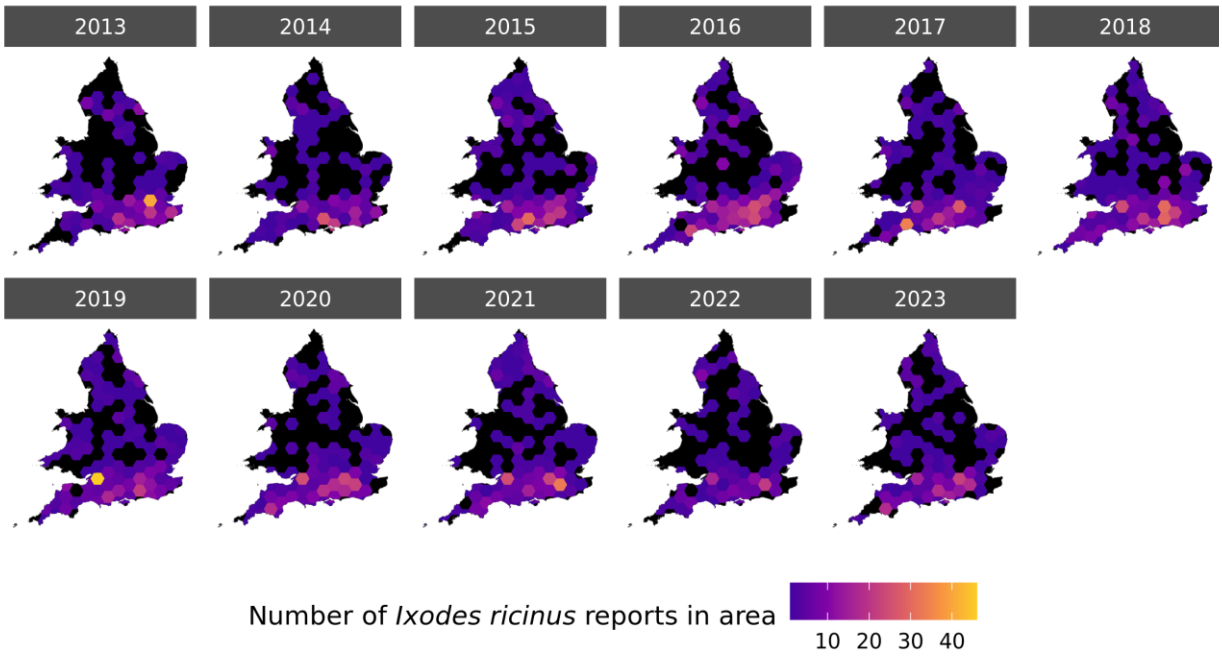

Figure 1: Gridded density map showing the number of I.ricinus reports in the TSS per year in each hexagonal area, after cleaning. Black represents areas with no reports in that year.

The TSS is based on passive surveillance data from citizen science reports, which introduces spatial biases. There is heterogeneous sampling effort across the study area due to differences in where individuals with a higher propensity to submit reports live or visit - such as urban areas or popular nature parks, as opposed to inaccessible private woodland or farmland (Dennis & Thomas, 2000; Kadmon et al., 2004). Because this pro-urban spatial heterogeneity was apparent in the raw data, we addressed it by a) varying importance weights for presences based on population density at the tick report location; and b) using target-group sampling to generate background points.

To reduce the spatial bias arising from higher observer presence in populous areas, we used importance weights that assigned more influence to ocurrences in sparsely populated areas. Weighting for sampling effort has previously been shown to increase accuracy in modelling bird presence (Johnston et al., 2020). In this study, presence points were assigned importance weights based on normalised log population density of the tick location's 2021 Middle layer Super Output Area (ONS 2021). The differences in these weights can be seen graphically in Figure 3.

Background Points

Previous modelling research has noted that obtaining true absence data for ticks is challenging because a species' niche changes over time (Estrada-Peña, 2008) and with weather conditions (Uusitalo et al., 2022) and seasonality, and sampling via dragging is resource-intensive (Ribeiro et al., 2019). Standard practice in species distribution modelling has been to generate either "pseudo-absence" points, often stratified by how confident the researcher is of absence, or a high number of "background" points that represent the overall environmental space, including areas where the species may be present. However, neither of these methods addresses the issue of spatial bias due to differential sampling effort; instead, we used target-group background sampling. This involves generating background points based on presence reports of other species in the same taxonomic group as the one targeted by the SDM. This technique has been shown to be effective at reducing spatial bias in a range of models (Barber et al., 2022; Phillips et al., 2009). We anticipate that this will also reduce biases arising from differential propensity to report ticks among different human demographics, as these would be constant across tick species. We did this by taking the density of TSS reports for other tick species as a proxy for tick sampling effort. The "other tick species" density raster is then used to generate background points in a 2:1 ratio with *I. ricinus* presences.

Iterative model development showed that relying entirely on target-group sampling resulted in a lack of background points in some areas (e.g. the north Pennines) with very few TSS reports for any species. To address these gaps, geographically random background points were generated within the boundaries of England and Wales, with a minimum distance of 5km from presences (to reduce overlap) and a ratio of 2:1 background to presence points. A version of the model without randomly generated background points, and a version with more background points, are demonstrated in the Sensitivity Testing section, as are alternative weighting schemas (e.g. placing more emphasis on target-group sampled points). The effect of this combination is that background points were distributed throughout the geographical and environmental space of the study area, but are more likely to be produced in areas where other ticks have been reported.

The background points were then re-weighted to have approximately equal total group weight to the presence points; this is in line with recommendations from Liu et al. (2019) to create background points as a small multiple of the total presences, but assign each group equal weight, as shown in Figure 2. Figure 3 shows the location of both presence and background points in the training set and their relative weighting.

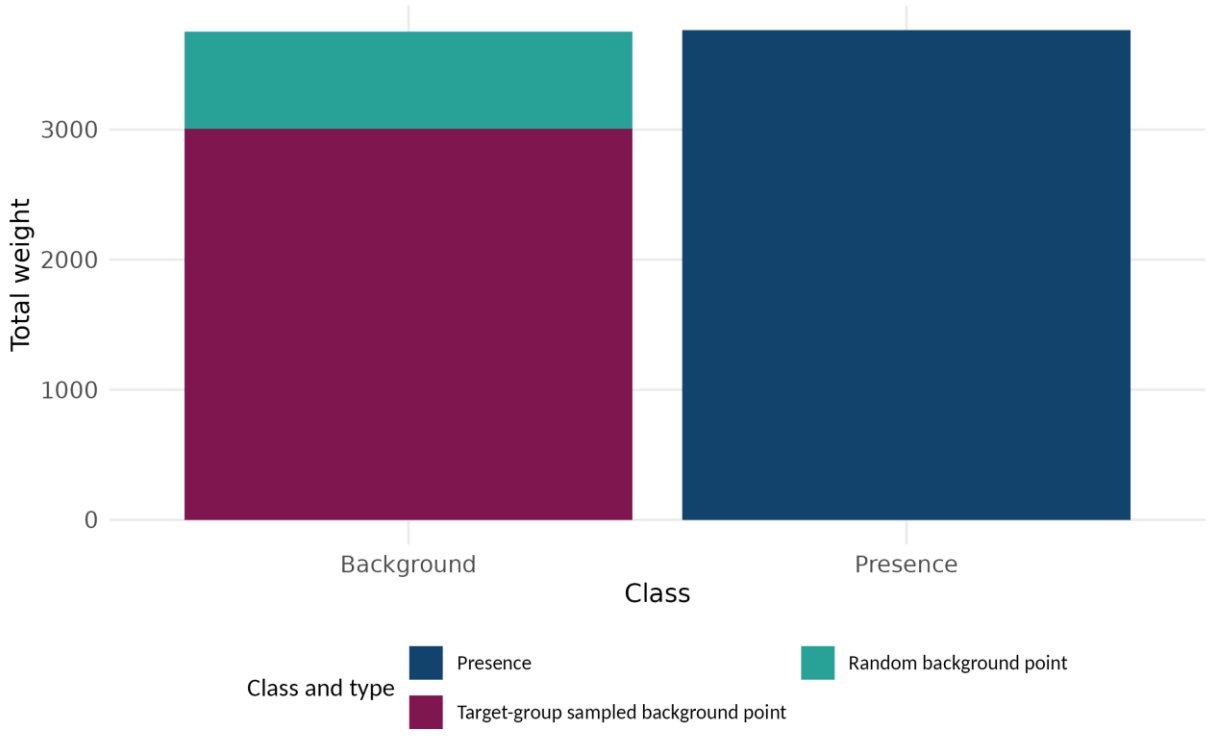

Figure 2: Bar plot showing the total weight assigned to presences and background points, with background points broken down into random and target-group sampled subclasses.

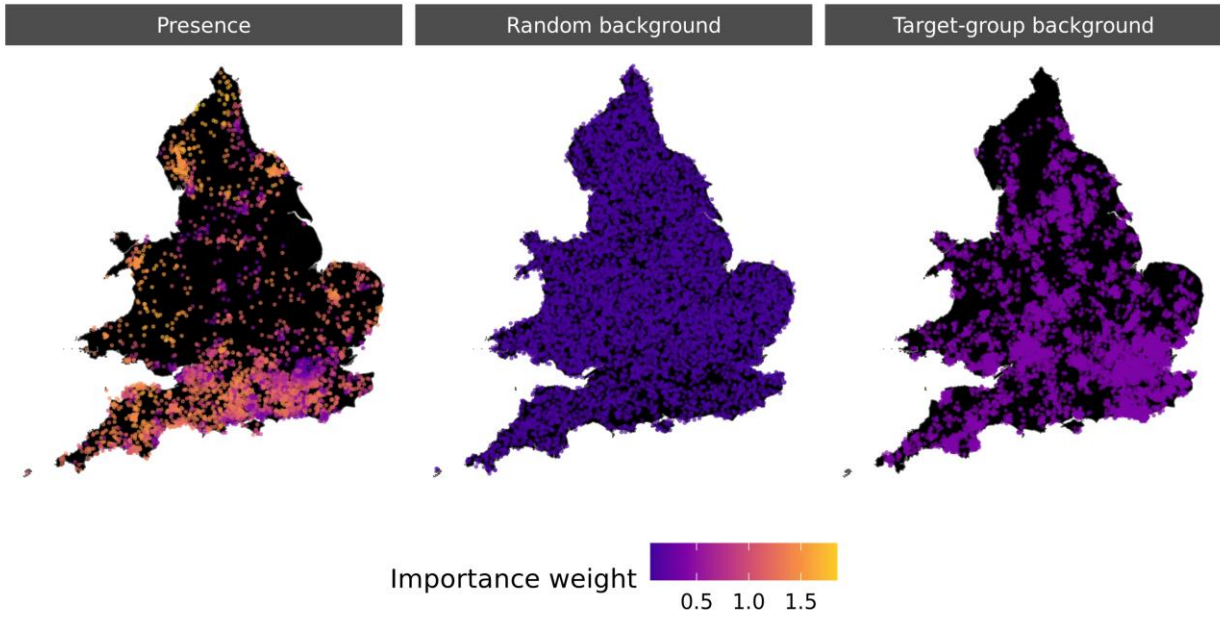

Figure 3: Maps showing the locations and relative weights given to presence points, randomly generated background points and target-group sampled background points. Presences are weighted by population sparsity. Randomly generated background points are spread throughout the extent of England and Wales but assigned a low weighting. Target-group sampled background points are given more weight than random points, but are clustered in some areas of the country. Areas with no points are filled in black. All points' locations are randomly jittered by up to 5km for pseudonymisation.

Environmental Data

A set of environmental predictor variables, which have previously been associated with the presence of *I. ricinus*, were identified and collated from publicly available data sources. These predictor variables fall into the following categories: climatic conditions, land cover, soil type, geology, host prevalence (cattle, pig, sheep and six species of deer), NDVI and elevation.

Although the *I. ricinus* tick can only crawl a few metres (Perret et al., 2003), it has a wide range of hosts including deer, livestock, rodents, birds and small mammals, all of which can help with tick dispersal and it can adapt to live in a range of habitats (e.g. [deciduous, coniferous, mixed] woodland, woodland edge, moorland, heathland, grazed grassland, urban parks) provided there is a suitable microclimate to support off-host survival (Milne, 1950). As a result, many different factors may contribute in complex ways to determine whether if the species is introduced to an area, and if it is able to thrive in that locality. The increase in deer numbers and their expanding range across the UK, along with as land-use changes and climactic variation, may have contributed to an enlarged geographic range for *I. ricinus* (Cunze et al., 2022; S. L. Gandy et al., 2023; Olsthoorn et al., 2025). We therefore aimed to include a diverse range of factors that have some *a priori* ecological justification as to their impact on *I. ricinus*, and are not confounded with spatial bias.

Existing modelling indicates that temperature has a non-linear effect on *I. ricinus* presence and abundance (Medlock et al., 2013). We therefore include data on temperature ranges, sunshine hours, ground frost and snow lying days from HadUK-Grid 1km v1.3.0 (Hollis et al., 2019). HadUK-Grid rainfall and humidity were also extracted as ticks are more prevalent in humid environments and risk desiccation in drier climates (Gray et al., 2021), though we recognise that there are concerns about the relevance of macroclimactic humidity to the microclimate experienced by ticks (Ostfeld & Brunner, 2015; Ribeiro et al., 2019). Soil type has been shown to be predictive of tick presence (Boulanger et al., 2024; Goldstein et al., 2018); we included WRBLV1 from the European Soil Data Centre's European Soil Database v2.0 (Panagos et al., 2006; "The European Soil Database Distribution Version 2.0," 2004). We used land cover type from the UKCEH Land Cover Map for 2021 (Marston et al., 2023) as we expect woodland and scrubby, grazed grassland areas to be high-risk areas for tick presence (Estrada-Peña, 2001; Hansford et al., 2023; Janzén et al., 2023; Medlock et al., 2022; Ribeiro et al., 2019); and to give a broader view of the likely ecology and biodiversity of the area, we also use superficial deposit type from the British Geological Survey ("BGS Geology 625K," 2024), as superficial geology and soil permeability has been linked to tick presence (Medlock et al., 2008). NDVI is often important in predicting *I. ricinus* presence and abundance (Bisanzio et al., 2014; Kjær et al., 2019; Ribeiro et al., 2019; Rochat et al., 2020; Signorini et al., 2019); the median NDVI for each pixel is calculated across April to August in each year after masking the "cloud" and "cloud shadow" layers. This time period was chosen as the peak in TSS reporting activity (Hansford et al., 2023), and in line with previous SDMs' exclusion of winter NDVI (Signorini et al., 2019). NDVI was calculated based on satellite images from USGS Landsat 8 Level 2, Collection 2, Tier 2, accessed via the *rgee* interface to Google Earth Engine (Aybar et al., 2020). Modelled estimates for presence of and environmental suitability for six deer species (fallow, roe, red, Chinese

water deer, Chinese muntjac and Japanese sika) from Croft et al. (2019) were included, as were estimated densities for cattle (Animal and Plant Health Agency, 2024a), pigs (Animal and Plant Health Agency, 2024b) and sheep (Animal and Plant Health Agency, 2024c). Finally, because the absence of hosts at high altitudes can affect tick distribution (Medlock et al., 2013), we also include elevation from the NASA Shuttle Radar Topography Mission (SRTM) digital elevation data (Developers, 2024).

Imprecision of geographical recording of presence points may arise if the human observer is unclear or mistaken about the location of the vector encounter, or later in data processing if the geographic location reported is not precise enough. To minimise this bias, we expressed categorical variables as grid-square proportions. This was done by segregating each category within land cover, soil type and geology type as their own binary raster layer, then calculating the mean of that layer for each square of a 1km-resolution grid, thus yielding the proportion of the grid square that was classified as that category. These variables therefore supply the model with an approximate picture of the composition of the area around the tick report (for example 56% built-up areas and gardens, 28% arable, etc.) and make the precise coordinates of the tick report less important, mitigating somewhat the problem of spatial inaccuracy in the vector reports.

All spatial data was re-projected to 1km x 1km resolution and cropped to the extent of England and Wales. To reduce inference problems caused by high collinearity between covariates, pairs of variables with the highest Pearson correlation above 0.7 were identified, and one variable was removed from each pair using the *step_corr* function from the *recipes* R package (Kuhn et al., 2024). For the machine learning models, predictors were scaled and centered to zero prior to model training for reasons of computational efficiency.

Table 1 shows summary statistics for the training data, split into presences (PR) and background (BG) points. All statistics are unweighted.

Table 1: Unweighted summary statistics for variables in the training set (2013-2023) of the Tick Surveillance Scheme. Minimum ("Min"), maximum ("Max") and mean are calculated within each class. "BG" denotes background points, while "PR" denotes presence points.

	Min		Max		Mean	
Variable	**BG**	**PR**	**BG**	**PR**	**BG**	**PR**
Ground frost days: minimum	0.00	0.00	2.25	2.56	0.07	0.08
Ground frost days: maximum *	0.96	2.27	31.00	30.77	18.22	18.12
Rainfall (mm): minimum	0.00	0.08	133.24	146.45	16.55	15.72
Rainfall (mm): maximum	58.45	57.56	1181.59	1192.61	154.04	166.34
Rainfall (mm): mean *	32.36	35.61	425.41	354.22	75.57	79.23
Highest average air temperature (°c): maximum	13.38	13.25	28.08	27.84	22.53	22.86
Humidity (%): minimum	57.41	58.46	82.80	82.85	72.29	71.93

Humidity (%): maximum	83.51	84.71	96.08	96.26	89.93	89.71
Sunshine hours: minimum *	9.69	10.87	84.14	79.37	41.57	42.27
Sunshine hours: maximum *	158.22	167.56	366.94	382.88	248.14	254.94
Sunshine hours: mean	88.51	91.76	182.08	189.55	133.28	136.19
Soil type: arenosol	0.00	0.00	1.00	1.00	0.01	0.01
Soil type: cambisol	0.00	0.00	1.00	1.00	0.30	0.24
Soil type: fluvisol	0.00	0.00	1.00	1.00	0.05	0.02
Soil type: gleysol	0.00	0.00	1.00	1.00	0.12	0.06
Soil type: histosol	0.00	0.00	1.00	1.00	0.02	0.01
Soil type: leptosol	0.00	0.00	1.00	1.00	0.05	0.09
Soil type: luvisol	0.00	0.00	1.00	1.00	0.34	0.35
Soil type: no information	0.00	0.00	1.00	1.00	0.01	0.01
Soil type: podzol	0.00	0.00	1.00	1.00	0.02	0.08
Soil type: regosol	0.00	0.00	1.00	1.00	0.00	0.00
Soil type: town	0.00	0.00	1.00	1.00	0.08	0.13
Soil type: water body	0.00	0.00	0.56	0.56	0.00	0.00
Geology: alluvium	0.00	0.00	1.00	1.00	0.10	0.07
Geology: blown sand	0.00	0.00	1.00	1.00	0.01	0.00
Geology: brickearth	0.00	0.00	1.00	1.00	0.01	0.01
Geology: clay with flints formation	0.00	0.00	1.00	1.00	0.01	0.03
Geology: crag group	0.00	0.00	1.00	1.00	0.00	0.00
Geology: drift geology not mapped	0.00	0.00	1.00	1.00	0.00	0.00
Geology: glacial sand and gravel	0.00	0.00	1.00	1.00	0.04	0.02
Geology: lacustrine deposits	0.00	0.00	1.00	1.00	0.01	0.00
Geology: landslide deposits	0.00	0.00	1.00	1.00	0.00	0.01
Geology: missing	0.00	0.00	1.00	1.00	0.51	0.64
Geology: peat	0.00	0.00	1.00	1.00	0.02	0.01

Geology: raised marine and coastal zone deposits	0.00	0.00	1.00	1.00	0.00	0.00
Geology: river terrace deposits	0.00	0.00	1.00	1.00	0.06	0.08
Geology: sand and gravel of uncertain age and origin	0.00	0.00	1.00	1.00	0.00	0.04
Geology: till	0.00	0.00	1.00	1.00	0.22	0.08
Land cover: arable *	0.00	0.00	1.00	1.00	0.32	0.14
Land cover: broadleaf woodland *	0.00	0.00	1.00	1.00	0.05	0.12
Land cover: built up areas and gardens *	0.00	0.00	1.00	1.00	0.17	0.33
Land cover: coastal	0.00	0.00	1.00	1.00	0.01	0.01
Land cover: coniferous woodland *	0.00	0.00	1.00	1.00	0.02	0.03
Land cover: freshwater	0.00	0.00	1.00	0.99	0.01	0.00
Land cover: improved grassland	0.00	0.00	1.00	1.00	0.31	0.29
Land cover: mountain, heath and bog *	0.00	0.00	1.00	1.00	0.03	0.03
Land cover: saltwater	0.00	0.00	1.00	0.87	0.01	0.00
Land cover: semi natural grassland	0.00	0.00	1.00	1.00	0.07	0.04
Chinese water deer current distribution *	0.00	0.00	1.00	1.00	0.15	0.09
Fallow deer current distribution	0.00	0.00	1.00	1.00	0.64	0.75
Muntjac deer current distribution	0.00	0.00	1.00	1.00	0.71	0.76
Red deer current distribution *	0.00	0.00	1.00	1.00	0.34	0.54
Roe deer current distribution *	0.00	0.00	1.00	1.00	0.64	0.84
Sika deer current distribution *	0.00	0.00	1.00	1.00	0.35	0.55
Chinese water deer suitability	0.00	0.00	0.93	0.93	0.15	0.14
Fallow deer suitability	0.00	0.00	0.93	0.93	0.18	0.20
Muntjac deer suitability	0.00	0.00	0.97	0.97	0.26	0.21
Red deer suitability	0.00	0.00	0.97	0.97	0.16	0.14
Sika deer suitability	0.00	0.00	0.92	0.89	0.06	0.07
Cattle population	0.04	0.05	181.57	161.36	34.28	30.90

Pig population (per km^2)	0.01	0.01	614.07	402.37	25.69	15.49
Sheep population *	0.10	0.14	442.58	399.90	62.96	46.92
Elevation: standard deviation	0.00	0.00	142.00	128.26	11.37	13.62
NDVI (Normalized Difference Vegetation Index)	-0.04	0.01	0.51	0.49	0.34	0.34
Weights	0.10	0.06	0.40	1.85	0.25	0.99
Single asterisk (*) : variable included in generalised additive model (GAM) formula.						

Models

A set of four SDMs (two statistical models and two machine learning models) were trained on the 2013-2023 training data to distinguish *I. ricinus* presences from background points. Cross-validation based on random sampling can result in overly optimistic performance metrics because of spatial correlation between nearby points (Ploton et al., 2020), so spatial block cross-validation from the *tidysdm* R package (Leonardi et al., 2024) was applied. The training data was divided into five spatial block folds and performance metrics averaged across each fold. For the GLM and the machine learning models, these folds were also used to tune hyperparameters by grid search with racing (Kuhn, 2024), which drops out hyperparameter combinations that analysis of variance (ANOVA) determined would not be the most performant. In each case, 12 possible combinations of hyperparameters were fitted and the model specification with the highest average continuous Boyce index on the training data was taken forward for ensembling. This process is shown in Figure 4.

The ensemble prediction was created using a simple arithmetic mean of the predicted probability of presence for each data point. Ensembling models often results in better model performance and reduces the influence of individual modelling parameters on the combined output (Yates et al., 2018). For species distribution modelling tasks in particular, diverse ensembles of bespoke models compare favourably to individual models or off-the shelf solutions (Valavi et al., 2022). Alternative methods for ensembling were also tested and are compared in the Supplementary Materials.

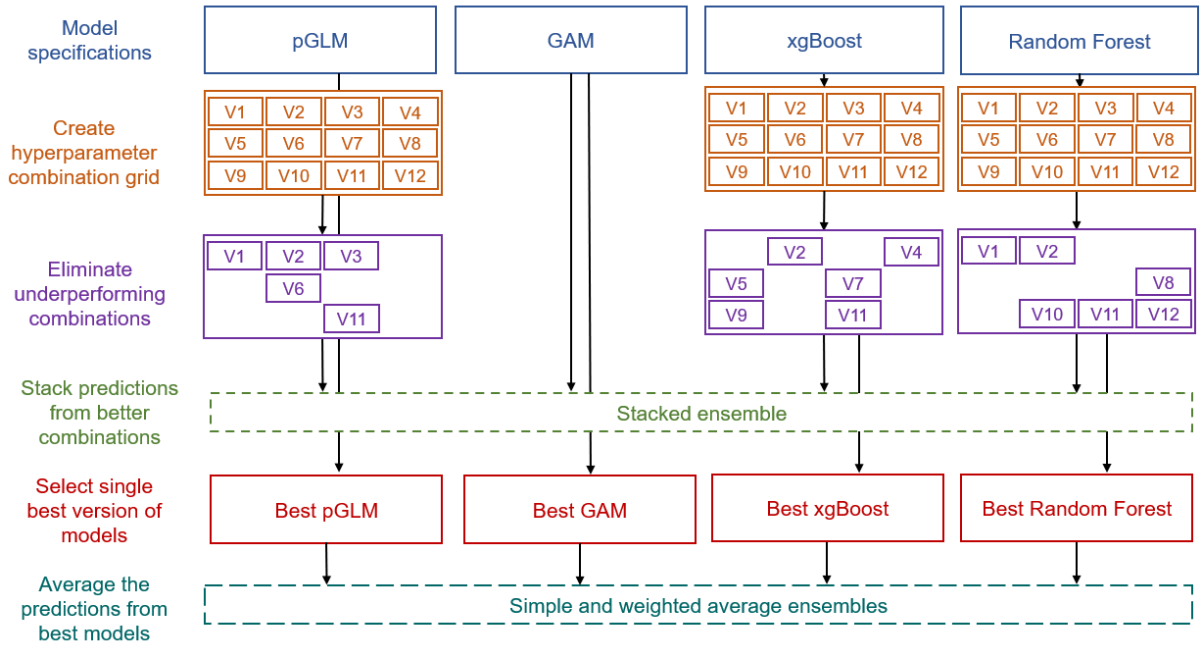

Figure 4: Model flow diagram showing the relationship between base model specifications, hyperparameter combinations and the model ensemble.

Statistical Models

A penalised generalised linear model (pGLM) and a generalised additive model (GAM) were estimated using the *parsnip* package as part of a broader *tidymodels*-based workflow.

The pGLM was specified as a penalised logistic regression model with alpha and lambda terms selected by cross-validation. After applying regularisation and penalisation, the model equation is that of a standard logistic regression model:

$$\text{logit}(p) = \log\left(\frac{p(\text{Presence})}{1-p(\text{Presence})}\right)$$
$$= \beta_0 + \beta_1 \text{groundfrost}_{\min} + \beta_2 \text{groundfrost}_{\max} + \beta_3 \text{rainfall}_{\min} + \cdots + \beta_p x_p.$$

The Generalised Additive Model (GAM) was specified as a binomial-family GAM and was fit using a cross-validation approach with five folds. In our model, the equation that defined our GAM was written as follows:

$$\text{logit}(p) = \log\left(\frac{p(Presence)}{1-p(Presence)}\right) = \beta_0 + s_1(\text{groundfrost max, sun min, sun max, rainfall mean})$$
$$+ s_2(\text{landcover type coniferous woodland, landcover type broadleaf woodland, landcover type built up areas and gardens, landcover type arable})$$
$$+ s_3(\text{landcover type mountain heath and bog})$$
$$+ s_4(\text{red suitability, roe current, red current, cwd current})$$
$$+ s_5(\text{sheep pop}),$$

where

$$i \in \{1,2,3,4,5\}$$

were thin-plate regression splines.

Key variables were combined thematically into splines (weather, land cover type, deer presence and suitability, and sheep) to allow the model to use combinations of conditions (e.g. areas with a mix of forest and grassland habitats). “Mountain, heath and bog” was removed from the weather spline and put in its own spline due to cross-validation issues related to its sparsity. The number of knots in each spline was based on an iterative approach of varying the number of knots in each spline independently and a preferable number of knots was based on maximising model accuracy. The number of knots in the splines was chosen as five for splines s_1, s_2, s_3, seven knots for s_4 and six knots for s_5.

Machine Learning Models

An extreme gradient boosted trees (XGBoost) model from the *xgboost* package (Chen & Guestrin, 2016) and a Random Forests model from the *ranger* R package (Wright & Ziegler, 2015) were estimated in classifier mode via *parsnip*. For the XGBoost model, the number of trees was set at 1000, with early stopping enabled to allow a smaller number of trees to be generated as required. The learning rate was set at 0.01. Other hyperparameters were chosen by cross-validation. Details of hyperparameters for the final estimated models are shown in the Supplementary Materials.

Results

Model Performance

Predictions on the testing set (unseen 2024 data) were compared with the true class to assess the ensemble model’s ability to distinguish risk of *I. ricinus* presence. As these analyses were carried out in 2024 and climate data for that year would not be not available until the following year, these variables were set as their mean values for the 2021-2023 period, in effect creating a naive forecast for 2024. Performance metrics were calculated on unweighted counts - unlike the model training, where points were weighted according to their class, population sparsity (for presences) and generation mechanism (for background points). Of the 309 presences in 2024, 195 (63%) were correctly classified as presences, leaving 114 (37%) incorrectly classified as background points. The model’s performance was almost exactly equal for background points: of the points, 1048 (86%) were correctly classified as background points. The ensemble model therefore classified the majority of points correctly, and there was no significant imbalance in terms of over-predicting points as either presence or background.

The Wilkinson dot plots in Figure 5 show the modelled probabilities that the ensemble model assigned to the 2024 testing data, both for true presence points and for the randomly generated background points.

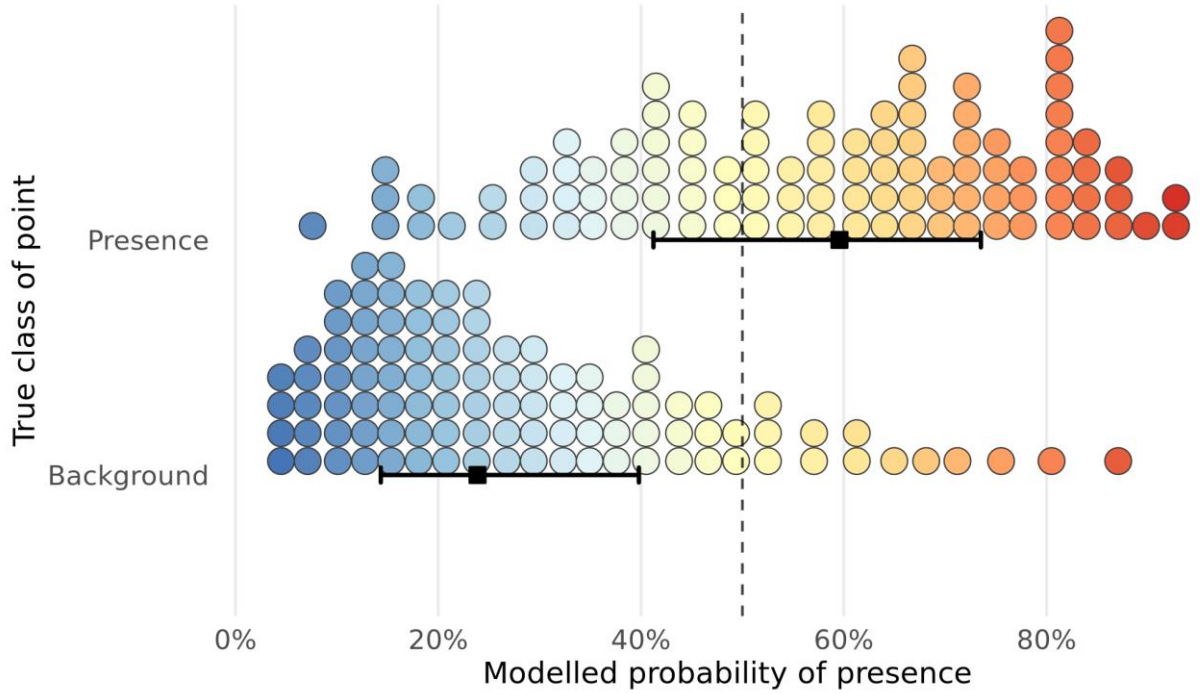

Figure 5: Wilkinson dot plots and intervals showing the distribution of the probability assigned to points in the 2024 testing data by the ensemble model, split by class (whether the point was a true tick presence point or a background point). Each bin represents 1% of the distribution, and each bin represents an equal number of observations (Kay 2023). The square point shows the median, and the thick black horizontal line shows the central two quartiles of the distribution. The dashed vertical line represents the 50% threshold.

As the intervals show, the majority of points were correctly classified. Predicted probabilities ranged from 3% to 95%. The median prediction for true presence points was 60%, and for background points was 24%. This indicates that the model was more skilled at classifying background points than presence points in the 2024 testing data. Figure 6 replicates this plot for each of the ensemble's base models.

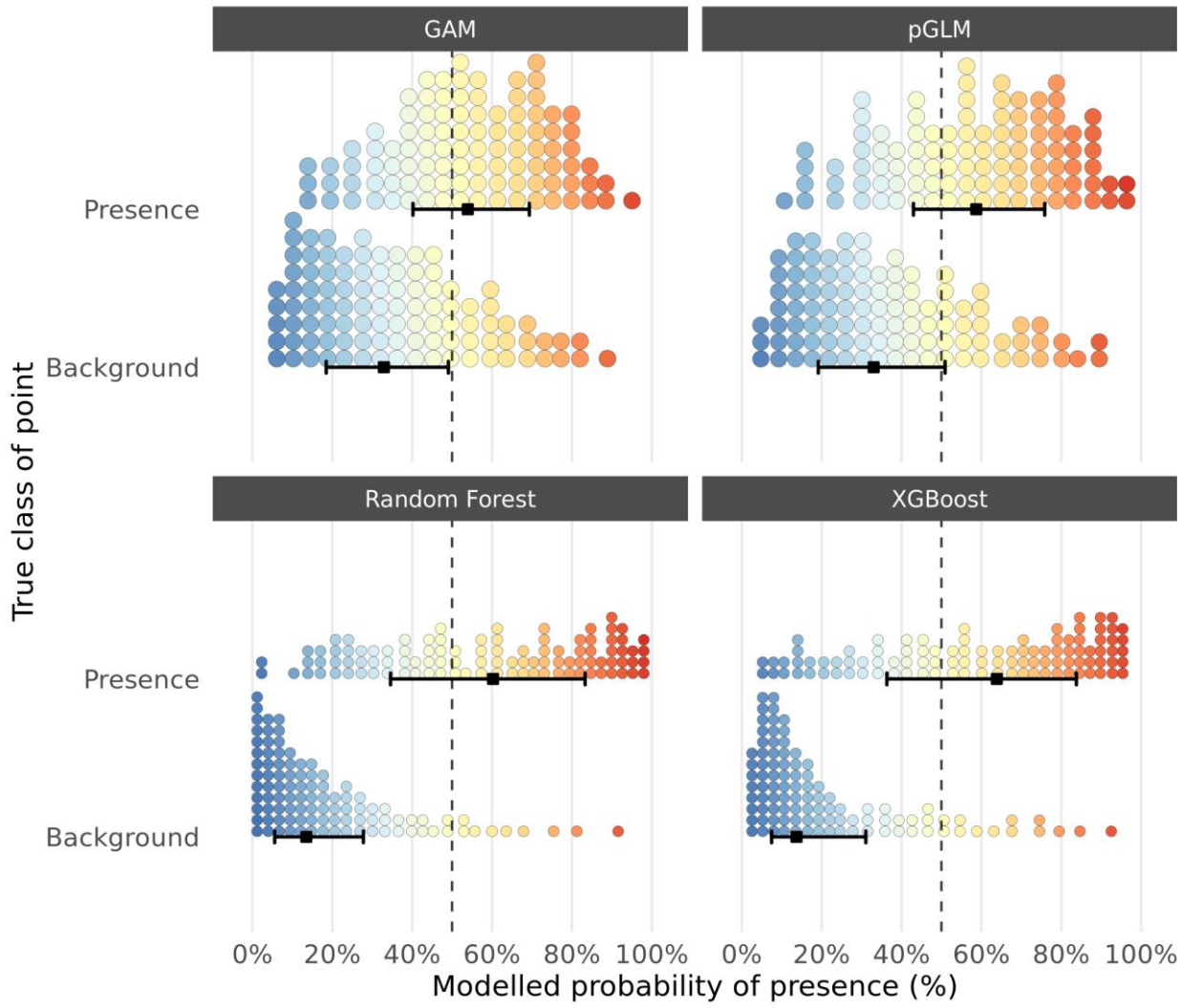

Figure 6: Wilkinson dot plots and intervals showing the distribution of the probability assigned to points in the 2024 testing data by each base model, split by class (whether the point was a true tick presence point or a background point). Each bin represents 1% of the distribution, and each bin represents an equal number of observations (Kay, 2023). The square point shows the median, and the thick black horizontal line shows the central two quartiles of the distribution. The dashed vertical line represents the 50% threshold. Model names shown above each subplot.

In the 2024 testing data, the machine learning models were more confident and accurate in assigning background points correctly: for example, the Random Forest model gave a median prediction of 14% to background points. This likely reflects the flexible non-linear nature of tree-based machine learning methods, which can effectively account for different relationships between variables in distinct habitats or regions. The median predictions for true presence points were similar across models, although there was more variation in the range of predictions among the machine learning models compared to the statistical models. Table 2 shows a range of model performance metrics for the simple ensemble model and the underlying base models. They are divided into metrics that evaluate the models based on their class predictions (such as accuracy) and those that evaluate the underlying predicted probabilities (such as Brier score).

Table 2: Out-of-sample predictive performance metrics for the simple ensemble and base models

	Class predictions					Probability predictions			
Model	**Accuracy**	**Kappa**	**MCC**	**Sensitivity**	**Specificity**	**Boyce Continuous**	**ROC AUC**	**Brier score**	**Max TSS**
Simple ensemble	0.82	0.47	0.47	0.63	0.86	0.99	0.84	0.14	0.53
pGLM	0.71	0.29	0.31	0.63	0.73	0.99	0.76	0.19	0.41
GAM	0.72	0.28	0.30	0.58	0.76	0.97	0.74	0.19	0.39
xgBoost	0.83	0.48	0.48	0.62	0.88	0.98	0.85	0.12	0.56
Random Forest	0.84	0.50	0.50	0.58	0.91	0.99	0.87	0.11	0.57

As the table shows, the machine learning-based models generally scored higher on measures of overall performance (accuracy, Cohen's kappa, continuous Boyce index, Brier score, max True Skill Score and ROC AUC). The statistical models' sensitivity (ability to correctly detect presence points) was very similar, but their specificity was lower (they incorrectly assigned more background points as presences), by comparison to the machine learning models. Despite assigning equal weight to the statistical models, the simple ensemble's scores on these metrics were closer to those of the machine learning models, and in some cases were equal to the XGBoost model.

Prediction Maps

As well as statistical model performance assessment, it is important to sense-check model predictions against expert opinion and other available evidence. The maps in Figure 7 show the modelled presence probabilities across England and Wales based on the average environmental data from 2021-2023 as a proxy for 2024's conditions, as well as the locations of actual tick reports for 2024.

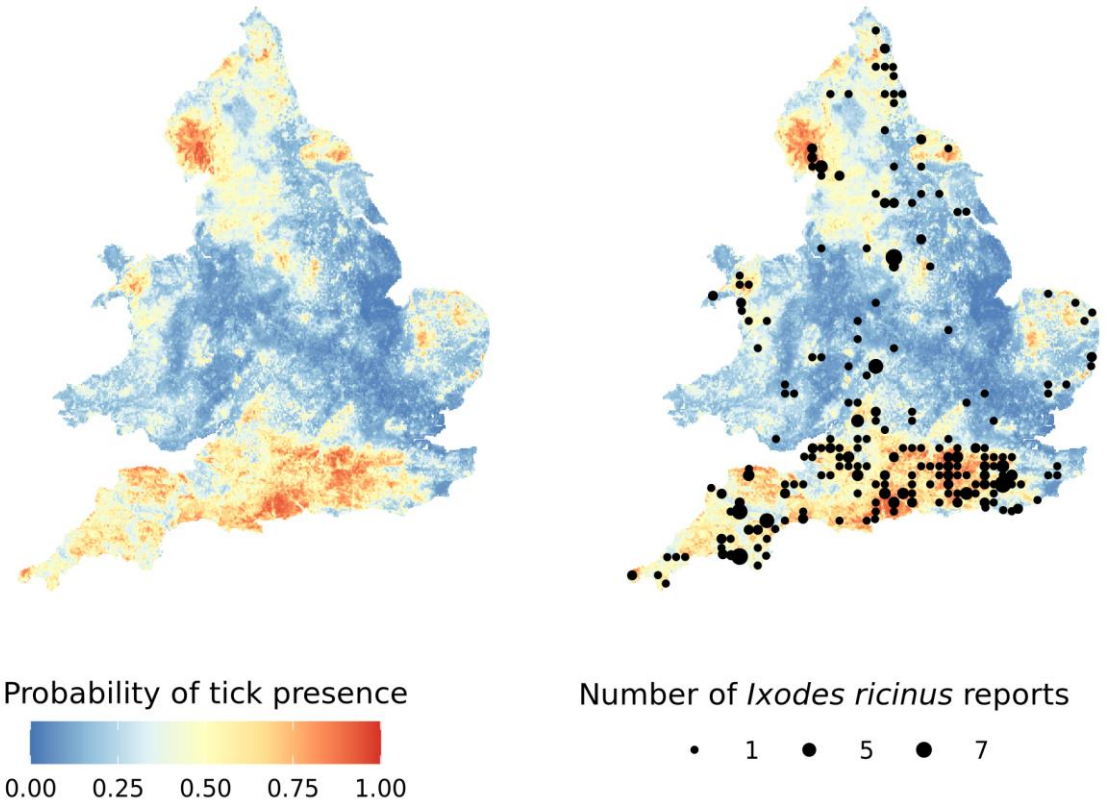

Figure 7: Modelled presence probabilities and actual tick reports from the 2024 testing data. To prevent reidentification of exact locations, reports were aggregated in a 10km x 10km grid and are shown at the centroid of the grid square. Larger dots represent multiple reports in the same grid square.

Areas with high modelled probability of *I. ricinus* presence include parts of northern England including notably Cumbria and the North Yorkshire Moors. In Wales, the highest modelled probabilities are in parts of Eryri in the north-west. In the Midlands and Anglia, there is currently lower suitability for *I. ricinus*, especially in Lincolnshire, largely on account of lower deer densities, although areas such as Thetford Forest and the Cotswolds are exceptions. The main predicted area for *I. ricinus* is in southern England, particularly Dorset, the New Forest, the South Downs, Exmoor and Dartmoor. The only exception is parts of east Kent; however, this may change as deer populations spread to occupy this area. If we translate this into public spaces with the highest probability of *I. ricinus* presence, the New Forest, Lake District and Exmoor are the National Parks with the highest predictions, while the Brecon Beacons, Pembrokeshire Coast and the Norfolk and Suffolk Broads had the lowest. For the National Landscapes, the Surrey Hills, Arnside & Silverdale and East Devon have high predictions, while the Lincolnshire Wolds, Wye Valley and Cannock Chase have with the lowest predictions.

The map was inspected by UKHSA medical entomology co-authors for expert review, checking the maps aligned with the understanding of I. ricinus distribution. In general, areas with very high modelled presence probabilities (75%+) were those where other sampling has found *I. ricinus*, whereas several areas in the 50%-75% range were noted as having similar environments to known tick habitats, without ticks yet being regularly detected. Two areas where the prediction maps did not align with the co-authors' understanding of the distribution were in the sheep farming areas of North Wales where ticks are expected to be more prevalent than the maps suggest; conversely, in the sheep-grazed uplands of the Lake District, there is anecdotally little evidence of ticks at high altitudes (as opposed to the forested valleys). The second of these potential issues may arise due to there being sharp differences in habitat over a small distance between the valleys and peaks in the Lake District that are effectively smoothed over by the model's grid system for summarising land type.

In addition, recognised tick and Lyme hotspots in the medical literature such as the New Forest, Exmoor, the South Downs, the Lake District and the North Yorkshire Moors (Dubrey et al., 2014) are all assigned much higher presence probabilities than average; see Figure 8 for a comparison of National Parks, and Figure 9 for a comparison of National Landscapes (formerly Areas of Outstanding Natural Beauty).

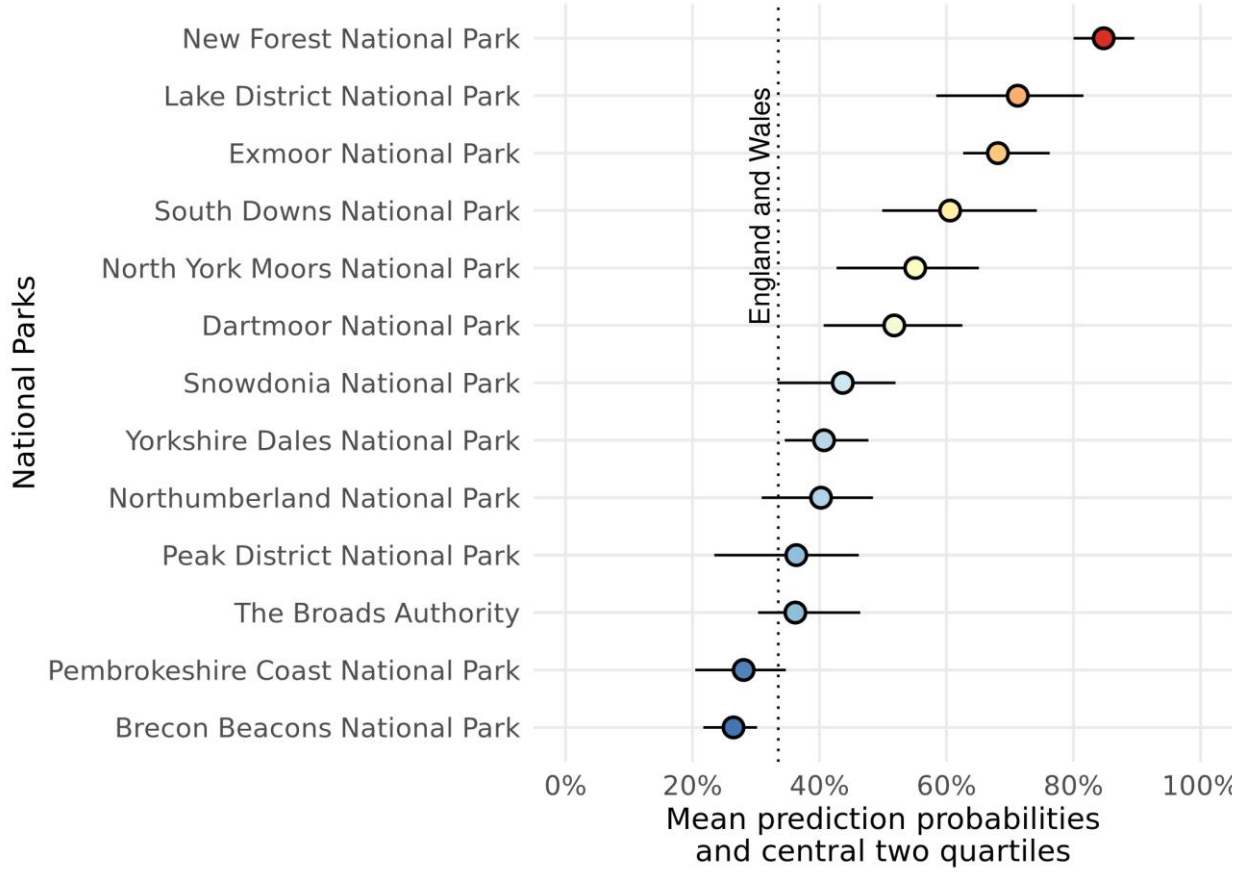

Figure 8: Mean, 25th percentile and 75th percentile I. ricinus presence probabilities for National Parks in England and Wales. The dotted line shows the mean presence probability predicted for England and Wales overall.

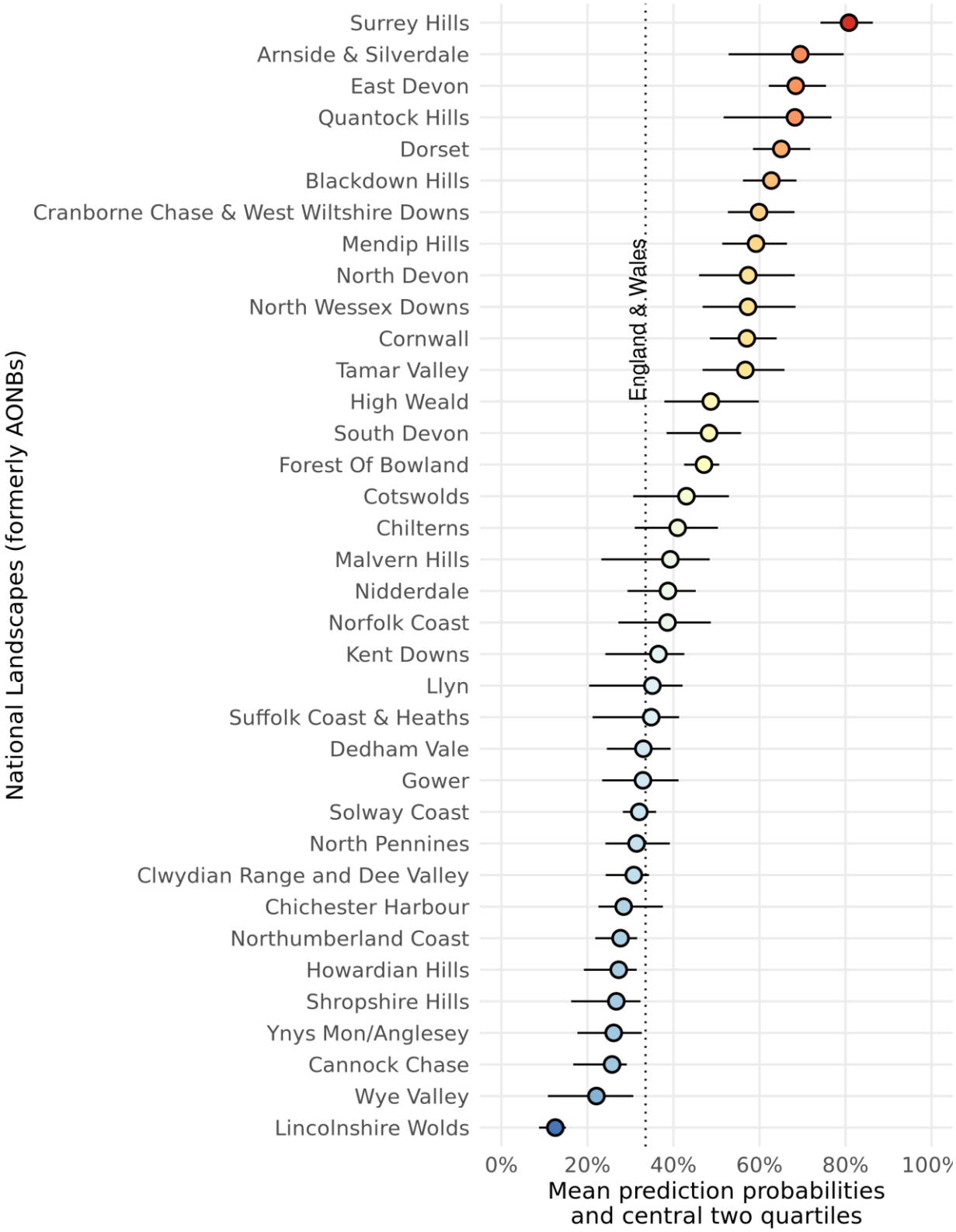
Surrey Hills
Arnside & Silverdale
East Devon
Quantock Hills
Dorset
Blackdown Hills
Cranborne Chase & West Wiltshire Downs
Mendip Hills
North Devon
North Wessex Downs
Cornwall
Tamar Valley
High Weald
South Devon
Forest Of Bowland
Cotswolds
Chilterns
Malvern Hills
Nidderdale
Norfolk Coast
Kent Downs
Llyn
Suffolk Coast & Heaths
Dedham Vale
Gower
Solway Coast
North Pennines
Clwydian Range and Dee Valley
Chichester Harbour
Northumberland Coast
Howardian Hills
Shropshire Hills
Ynys Mon/Anglesey
Cannock Chase
Wye Valley
Lincolnshire Wolds
England & Wales
National Landscapes (formerly AONBs)
0%
20%
40%
60%
80%
100%
Mean prediction probabilities
and central two quartiles

Figure 9: Mean, 25th percentile and 75th percentile I. ricinus presence probabilities for National Landscapes (formerly Areas of Outstanding Natural Beauty) in England and Wales. The dotted line shows the mean presence probability for England and Wales overall.

By contrast, heavily urban areas such as central London are not designated as hotspots, with the exception of extensive green spaces like Richmond Park where ticks and deer are known to be present (Dubrey et al., 2014; Hansford et al., 2021). The mean presence prediction in the Lower layer Super Output Areas (LSOAs) that cover most of Richmond Park (Richmond upon Thames 012B, 012C and 012 in the 2021 LSOA boundaries) is 63%. This compares to the median LSOA's presence probability of 29%. Lower layer Super Output Areas (2021) Boundaries EW BSC were taken from the ONS Open Geography Portal (Office for National Statistics, 2024).

Figure 10 gives an idea of uncertainty by visualising the standard deviation of modelled presence probabilities across the four base models (penalised GLM, GAM, XGBoost and Random Forest).

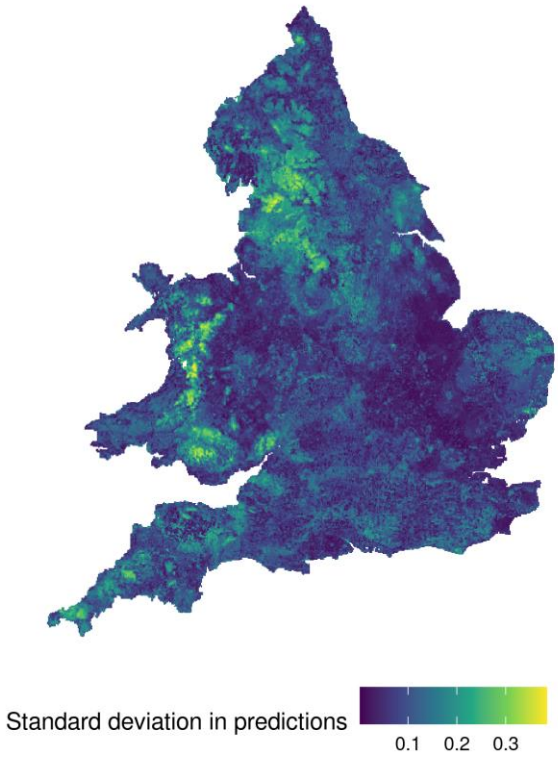

Figure 10: Map showing the standard deviation of modelled presence probabilities across the four base models. Areas with higher standard deviation can be viewed as more uncertain, as their predictions are more affected by model design.

The areas where predictions vary most between model are in the North West of England and in parts of Wales. These areas are likely to be relatively more suitable according to the simpler linear statistical regressions than the more complex patterns modelled by the machine learning models; as Figure 6 showed, XGBoost and Random Forest assigned low probabilities more often than the pGLM and GAM.

Variable Importance and Interpretability

Although we have used an ensemble that includes machine learning algorithms that allow for complex non-linear relationships between variables, post-hoc explainable artificial intelligence (XAI) methods exist that enable some interpretation of the predictions both at the global and local levels (Ryo et al., 2021). These give some indication of the general importance and average directionality of predictors (in the global case) and why a particular prediction was made (in the local case). As this model aims to predict rather than explain, these associations should not be viewed as causal. The global importance of individual variables can be ascertained using model-agnostic permutation methods that introduce small permutations in each variable and measure the effect on model predictive accuracy (Fisher et al., 2019; Oppel et al., 2009). We use the *DALEX* R package (Biecek, 2018) to determine variable importance and show the ten most important across 25 iterations in Figure 11.

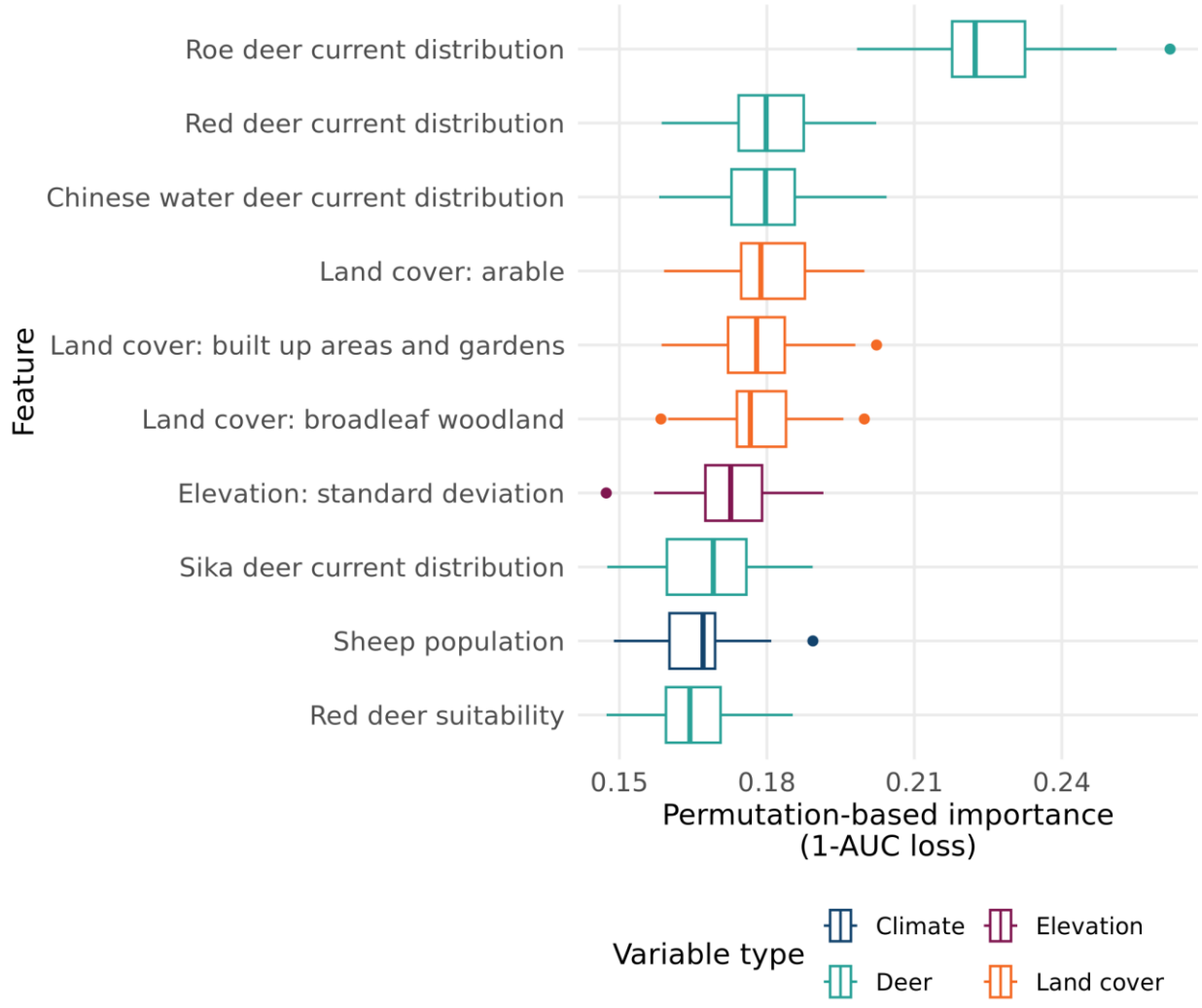

Figure 11: Permutation-based variable importance for simple ensemble predictions on the testing set. The plot shows the ten most important variables on average across 25 iterations, along with confidence intervals. Loss is measured as impact on 1-ROC AUC.

Partial dependence plots (PDP) show how average predictions vary across the range of a predictor when all other predictors are held at their mean. This gives a general idea of directionality, even in complex machine learning models. Figure 12 contains a PDP for each of the ten most important variables in the simple ensemble. Note that PDPs do not account for interactions and may be misleading when variables are highly correlated, in which case an accumulated local effects (ALE) plot is preferred (Apley & Zhu, 2020). In this case, both plots are very similar - see the Supplementary Materials for the ALE plot.

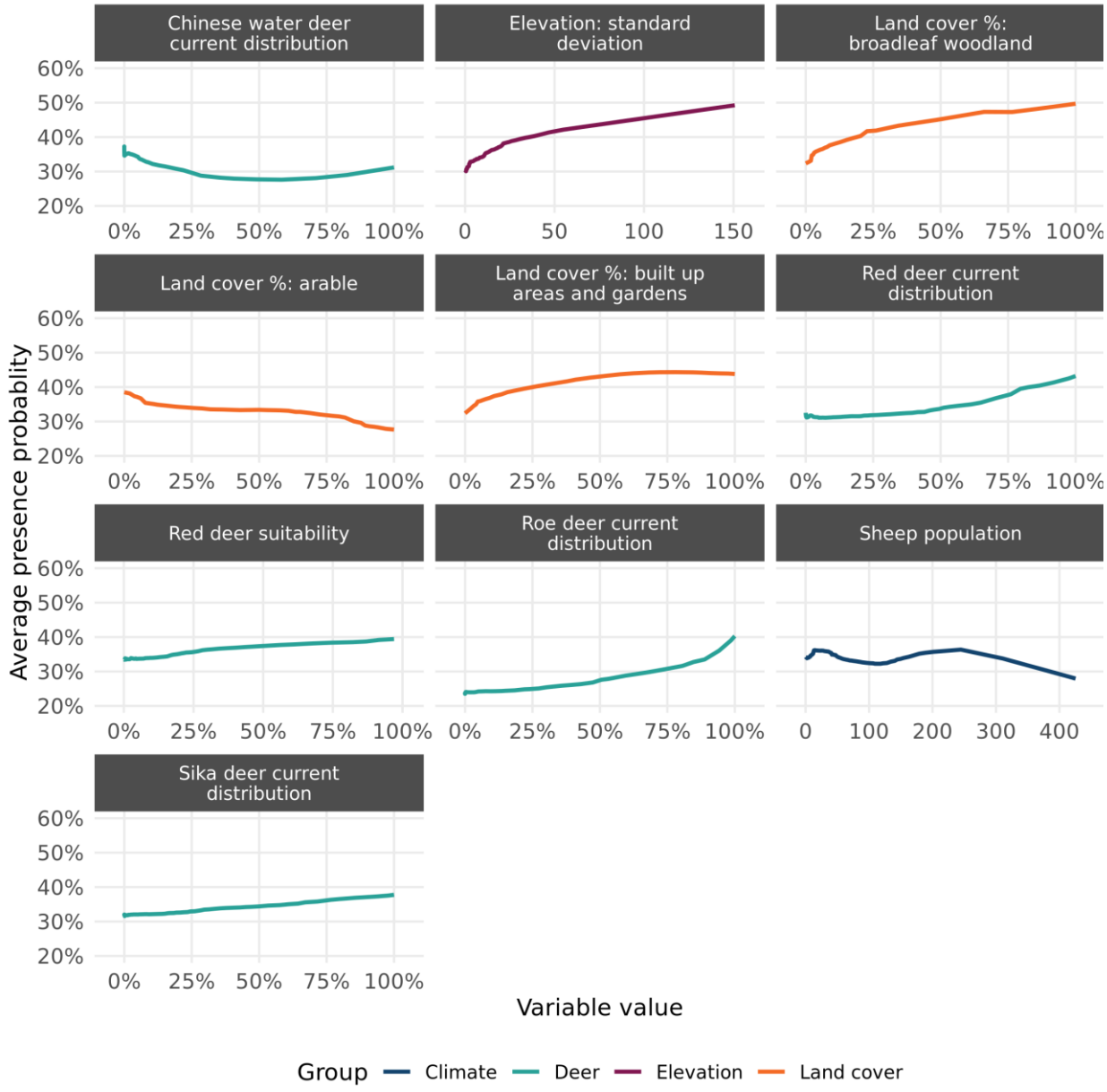

Figure 12: Partial dependence plots (PDPs) showing the average presence probability from the simple ensemble model, for the full range of the ten most important predictors in the testing data. All other predictors are held at their mean in each plot.

Sensitivity Testing

To check how sensitive the model's outputs are to different researcher choices, we carried out a range of sensitivity checks. In this section we show statistical summaries from a range of these alternative specifications ("scenarios"). For each scenario, results are shown for the simple ensemble and only one change is made from the baseline specification described in the main body of the paper. Three scenarios are tested that alter the presence points: in the first, presence points have spatial thinning applied with a minimum distance of 10km; in the second, no population weighting is applied; and in the third, only human records are used (both in the training data and in the generation of target-group sampled background points). Prediction maps from these scenarios are shown in Figure 13. Further sensitivity tests are described in the Supplementary Materials.

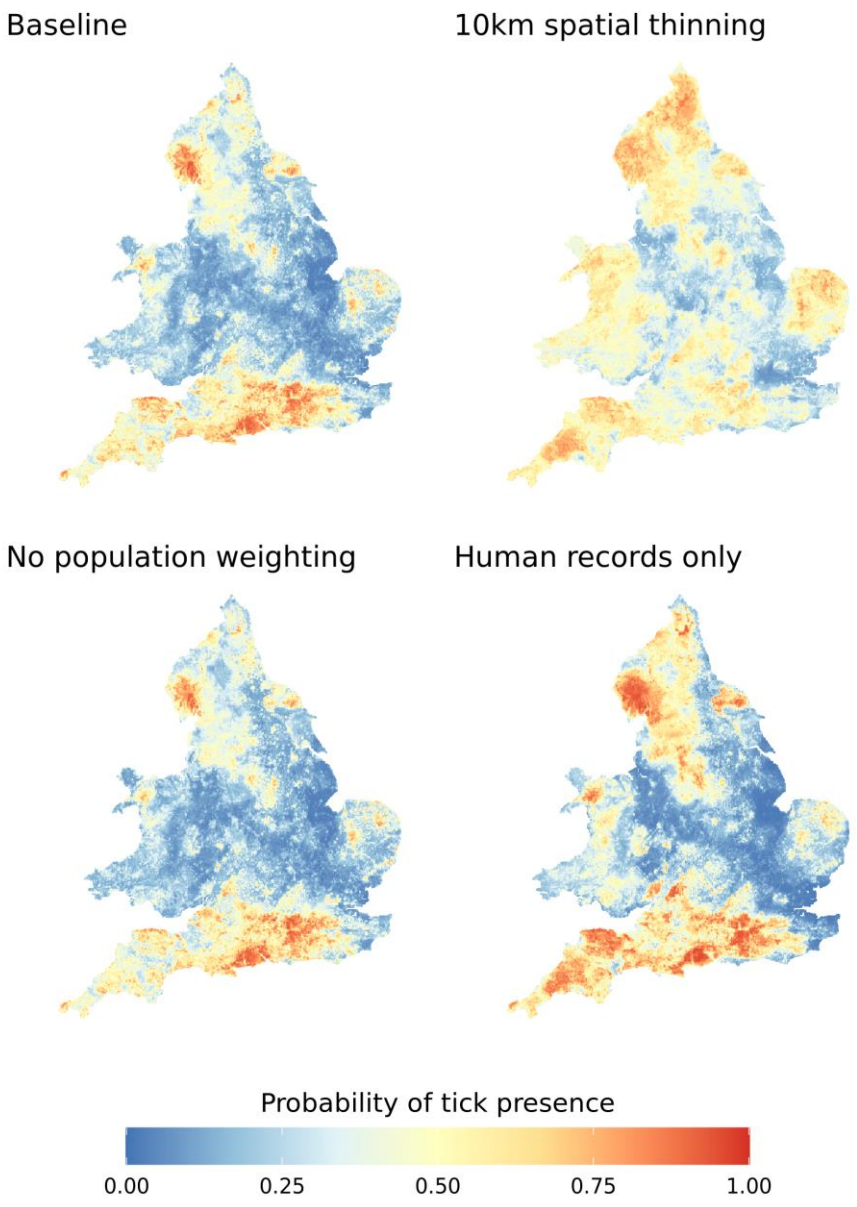

Figure 13: Prediction maps from the baseline scenario and three sensitivity testing scenarios that alter the presence points.

A further five scenarios alter how background points are generated. In the first, only random background points are used to train the model; in the second, random and target-group sampled points are equally weighted, and in the third only target-group sampled points are used to train the model. The fourth scenario increases the ratio of background to presence points from 4:1 (2:1 randomly generated and 2:1 target group sampled) to 10:1 (5:1 random and 5:1 target group). Finally, the fifth scenario removes the distance limit between presence points and background points (which is 5km in the baseline scenario). Prediction maps from these scenarios are shown in Figure 14.

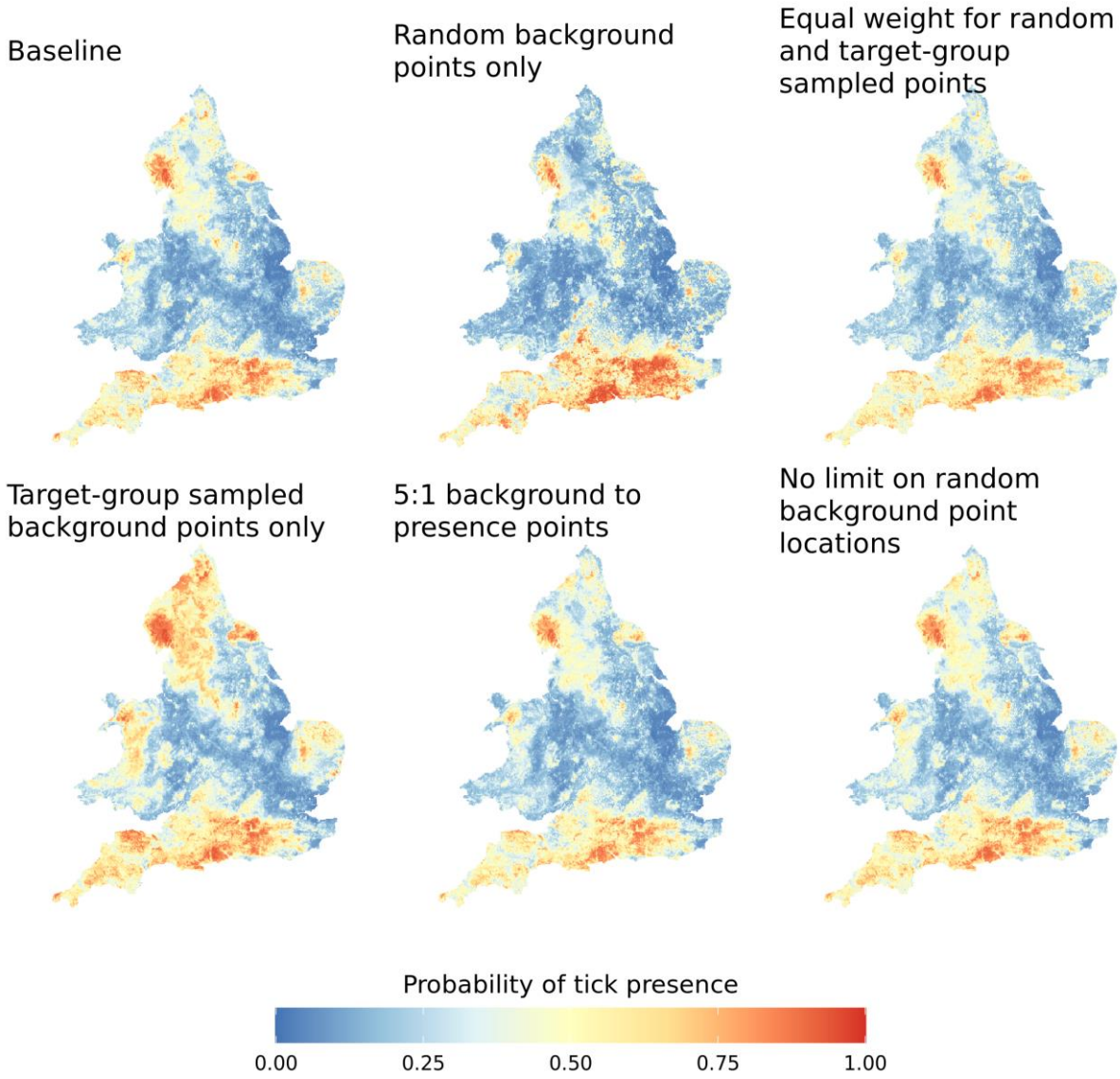

Figure 14: Prediction maps from the baseline scenario and five sensitivity testing scenarios that alter the background points.

To facilitate visual comparison between similar prediction maps, Figure 15 and Figure 16 show the difference in predictions between each scenario and the baseline, with differences divided into bins, for the changes affecting presences and the changes affecting background points, respectively.

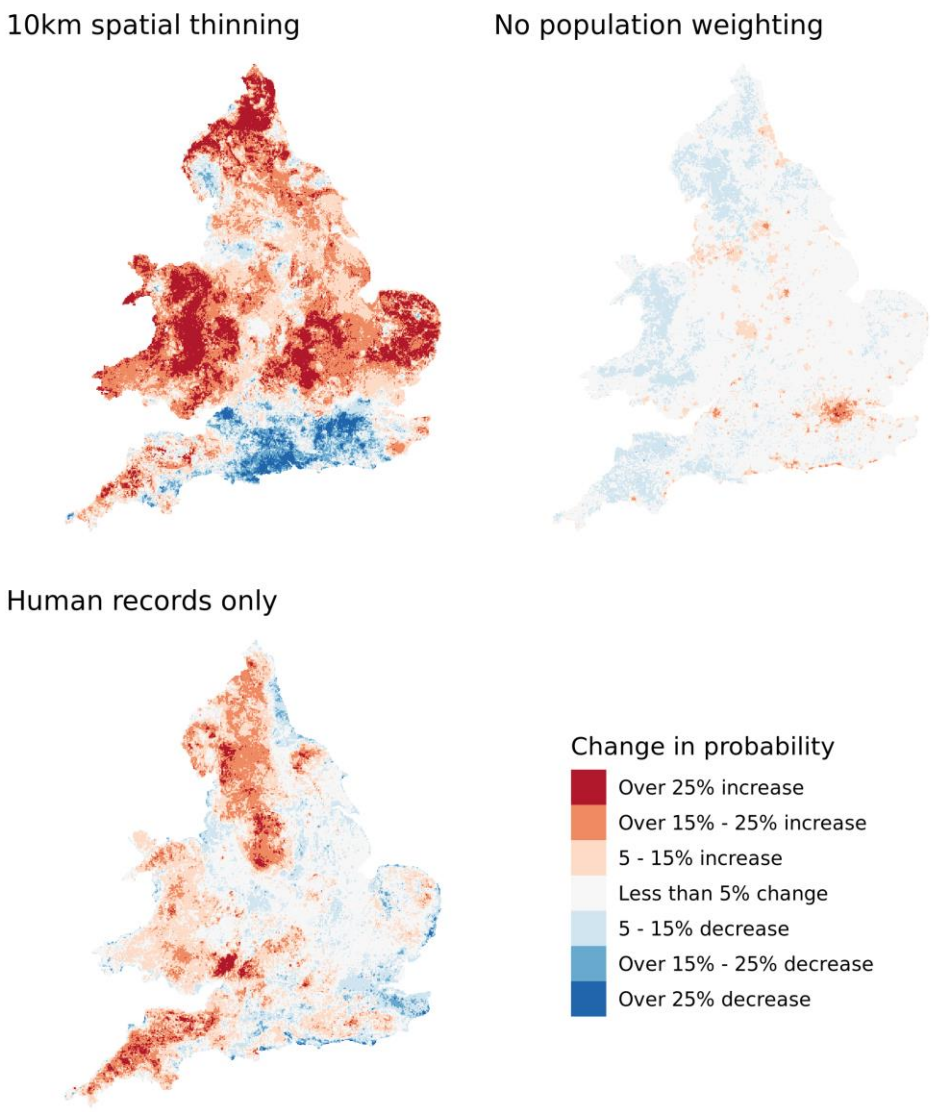

Figure 15: Maps showing differences in predictions between the sensitivity testing scenarios that alter presence points and the baseline scenario.

As Figure 15 demonstrates, introducing spatial thinning at 10km distance has considerable effects on the model predictions in both directions across most of the study area. Removing population weighting, as expected, increases the probability of presence in urban centres; filtering the data to human only increases the probability of presence in south-east Wales, Cornwall, and in much of the North West and Cumbria (note that some of this effect will be due to the reduction in overall numbers of presences).

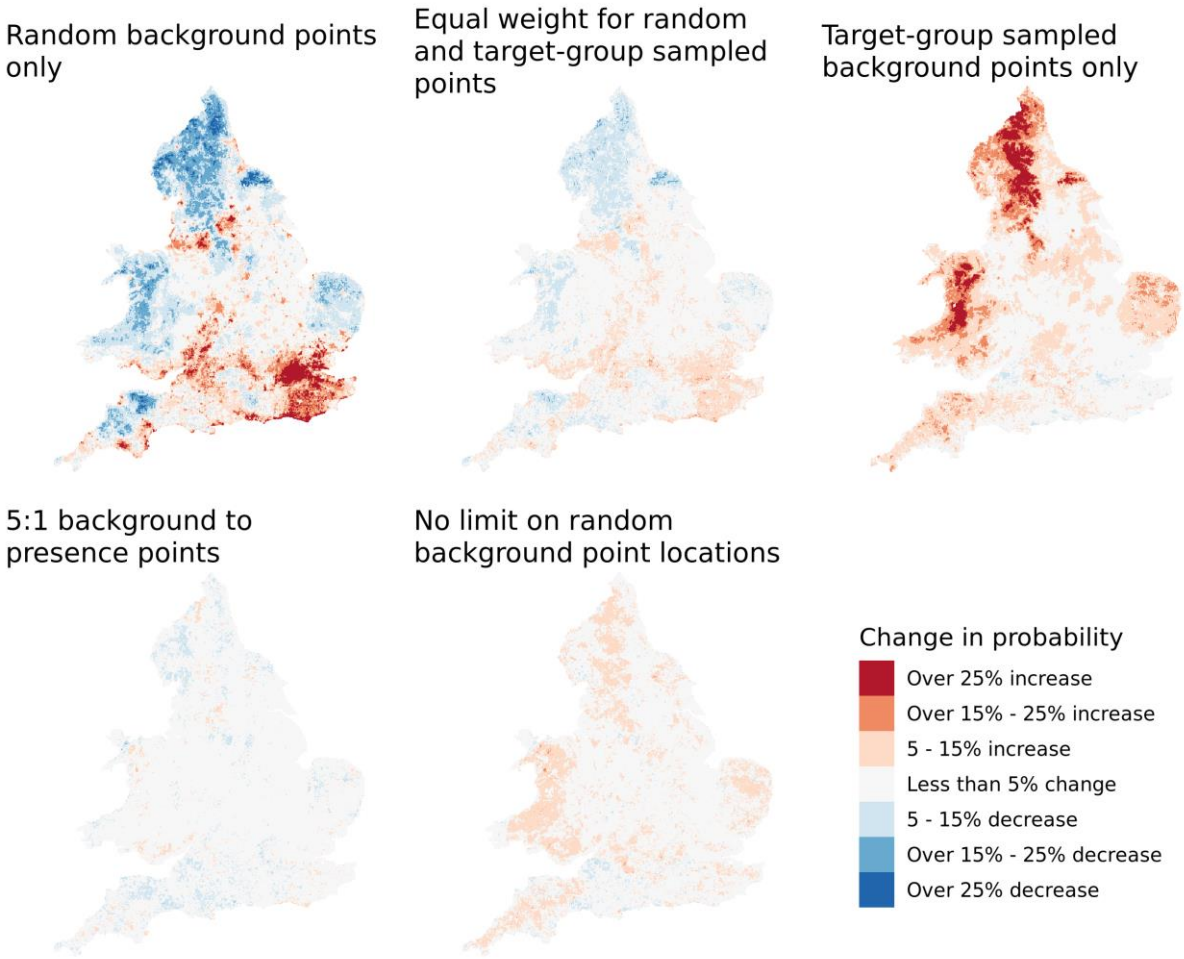

Figure 16: Maps showing differences in predictions between the sensitivity testing scenarios that alter background points and the baseline scenario.

Figure 16 shows that increasing numbers of background points or removing the limit on how close random background points can be to presence points makes only minor differences to the predictions. Changes in the proportion of target-group sampled and random background points have significant effects on Greater London and the South, as well as Wales, the North West and the North York Moors. This suggests that the target-group sampling strategy is effective in putting emphasis on rural environments and addressing potential pro-urban spatial bias.

Figure 17 shows the overall distribution of changes from the baseline scenario. To quantify differences between predictions, Table 3 shows the Spearman rank correlation and mean absolute difference between each scenario and the baseline. The spatial thinning scenario has the highest mean absolute difference and lowest Spearman rank correlation.

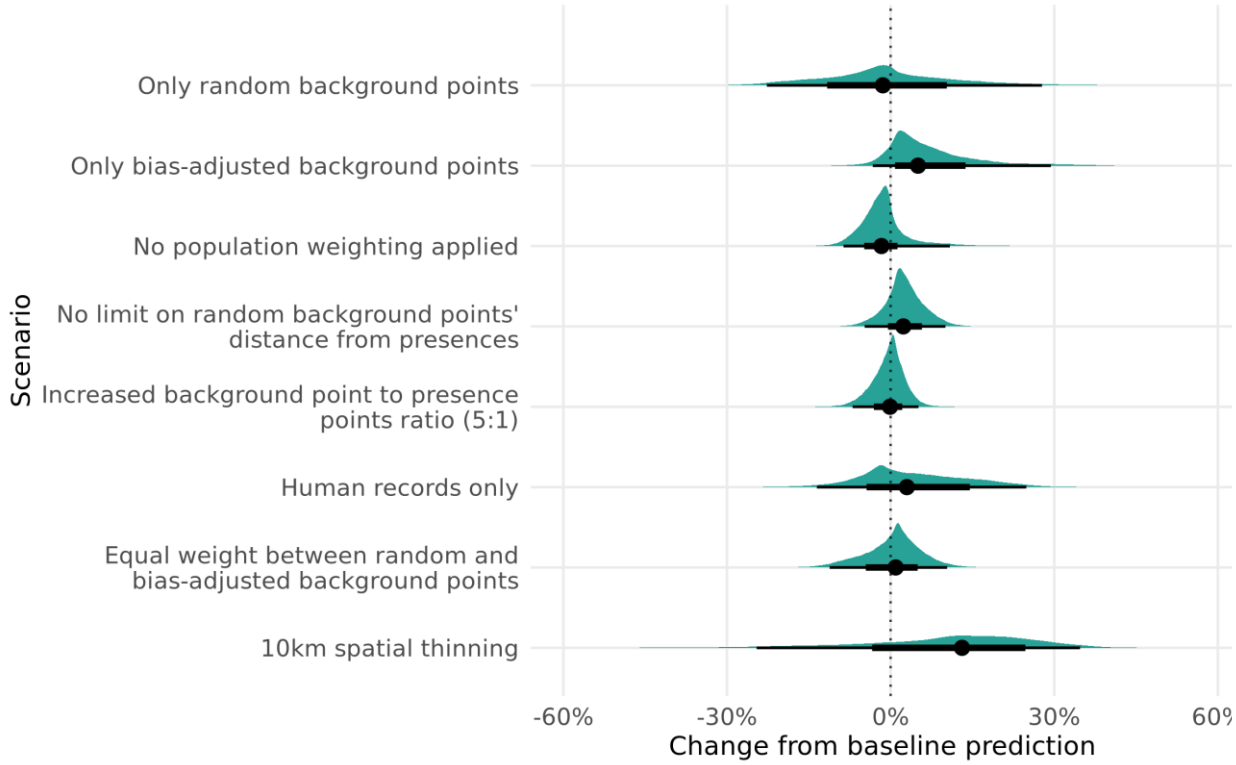

Figure 17: Density plot showing differences in predictions between all sensitivity testing scenarios and the baseline scenario.

Table 3: Spearman rank correlation and mean absolute difference between each scenario and the baseline.

Scenario	Spearman correlation	Mean absolute difference
Human records only	0.92	0.08
10km spatial thinning	0.71	0.16
Increased background point to presence points ratio (5:1)	0.99	0.02
No limit on random background points' distance from presences	0.99	0.03
Only random background points	0.83	0.09
Only bias-adjusted background points	0.95	0.08
Equal weight between random and bias-adjusted background points	0.96	0.04
No population weighting applied	0.97	0.03

Discussion

In this study we used unstructured, passive surveillance data from citizen science reports to model the distribution of the deer/sheep tick *I. ricinus*, vector of *Borrelia burgdorferi* s.l. (causative agent of Lyme borreliosis) and Tick-Borne Encephalitis virus (TBEv), across England and Wales. This paper produces, to our knowledge, the first published species distribution model for *I. ricinus* in England and Wales at the 1km resolution or similar. Previously published risk maps for *I. ricinus* in England have been modelled at a coarser resolution than this study (Estrada-Peña et al., 2006; Noll et al., 2023), making them unsuitable for local health interventions. Other UK studies investigating the risk of tick-borne diseases have focused on particular locations within the country, such as national parks (Cull et al., 2021) and urban green spaces (Hansford et al., 2021), both of which were focused on the variability of infection rates and the drivers for risk. We were interested in determining whether passive surveillance datasets, such as the UKHSA Tick Surveillance Scheme (Hansford et al., 2023), are suitable for species distribution modelling, and what steps public health agencies can take to improve the reliability of models based on these data collection schemes. We also want to embed reproducibility, business continuity and open science practices in our workflows as researchers in the field of health protection. To enable this, we aimed to use open-source software wherever possible, and in particular chose to use software packages that are compatible with the general purpose *tidymodels* modelling framework (Leonardi et al., 2024).

We found that applying less weight to presence data from densely populated areas and using target-group background sampling was effective in reducing presence predictions in towns and cities, and we expect this to also apply to other human biases introduced by the citizen science sampling process. Using an ensemble of statistical and machine learning models resulted in a 'risk map' showing the modelled probability of tick presence that was consistent with expert knowledge of I. ricinus ecology and distribution, as well as with previous small-scale survey-based models. Consistent with the literature, areas with broadleaf woodland and deer presence were modelled as higher risk for I. ricinus. Some areas with moderately high presence probabilities may be as yet unrealised parts of the species' environmental niche. Further work is needed to validate or challenge predictions in areas such as North Wales and the Lake District where some granular predictions did not match expectations based on the ecology.

However, we recognise that our modelling approach has limitations and makes simplifying assumptions. Probably the most significant assumption is that we cannot be sure that we have adequately addressed biases in the dataset without a ground truth for comparison (Matutini et al., 2021). Adding sampling study data could improve the model by providing these true absence points, particularly if under-sampled areas of geographical and environmental space were included. Further information on samplers could potentially be used to mitigate demographic differences in propensity to report vectors, and therefore further reduce bias in the model. We also recognise that population density is not necessarily indicative of human activity in popular rural destinations, and a more direct proxy for outdoor footfall would be more effective in reducing sampling bias.

Another significant limitation is that we have made many choices as researchers, a limited number of which have been demonstrated in the sensitivity tests, that have influenced the modelling process and therefore the distribution maps; but other reasonable choices could have yielded different results. We note in particular that the choice of background point generation methods and how these are combined has a significant impact on the model outputs, although the choices of number of points and buffering appear not to have been very impactful. We have aimed to blend data-first machine learning methods with some more directed methods (in particular, the GAM) to reduce our reliance on one model type, or on our priors, or on the dataset itself, but this prevented us quantifying uncertainty in as direct a

manner as a Bayesian approach. We opted to ensemble the base models using a simple average across the two statistical models and the two machine learning models, despite the overall higher performance of the machine learning models. This decision was taken for two reasons: first, that varying the weights given to each model based on performance would add another potential source of overfitting; and secondly, that we placed extra value on the ability to detect true presences (sensitivity). Without true absence data, we were concerned that the machine learning models' extra flexibility to non-linear patterns (as compared to statistical models) implied a higher risk of overfitting to the data without generalising to unrealised parts of the species' niche, or even to areas where the species is present but no presence records yet exist. Giving equal weight to simpler statistical models in the ensemble mitigates that risk somewhat. Researchers using models for other purposes may place greater emphasis on overall performance or specificity.

For the specific case of *I. ricinus*, this model also underlines the importance of deer as a host at a time when deer are present in increasing numbers across the UK. More broadly, we hope that as public awareness of vector-borne disease increases, this will lead to better passive reporting of ticks and mosquitoes. The bias mitigation strategies applied in this study may be applicable to other passive vector surveillance datasets, enabling public health researchers to better model and understand the risks posed by these species. Better availability of risk maps for vectors has a range of potential public health benefits - these include informing local partners' risk reduction strategies; informing high-risk groups and outdoor space users; highlighting areas with high predicted risk, but lower numbers of actual samples for enhanced surveillance and engagement; supporting serosurveillance studies to understand how presence translates to human health risk; and aiding health care practitioners to assess patient exposure, and appropriately prioritise a differential diagnosis of vector-borne disease.

Notes

Acknowledgements

We are grateful to others in the UKHSA who assisted in the collation and validation of the Tick Surveillance Scheme data and who have offered feedback on earlier drafts of the paper and methods, including Jacob Brolly, Susie Cant and Robert S. Paton. We also acknowledge the valuable technical support of Owen Jones and support from the rest of the Infectious Disease Modelling team.

Conflict of Interest

The authors have declared that no competing interests exist.

Data Availability Statement

UKHSA operates a robust governance process for applying to access protected data that considers:

- the benefits and risks of how the data will be used
- compliance with policy, regulatory and ethical obligations
- data minimisation
- how the confidentiality, integrity, and availability will be maintained
- retention, archival, and disposal requirements

- best practice for protecting data, including the application of 'privacy by design and by default', emerging privacy conserving technologies and contractual controls

Access to protected data is always strictly controlled using legally binding data sharing contracts.

UKHSA welcomes data applications from organisations looking to use protected data for public health purposes.

To request an application pack or discuss a request for UKHSA data you would like to submit, contact DataAccess@ukhsa.gov.uk.

References

Animal and Plant Health Agency. (2024a). Livestock demographic data group: Cattle population report. In *GOV.UK: Livestock population reports for Great Britain, using July 2023 data*. https://www.gov.uk/government/publications/cattle-population-in-great-britain-annual-reports

Animal and Plant Health Agency. (2024b). Livestock demographic data group: Pig population report. In *GOV.UK: Livestock population reports for Great Britain, using 2022 to 2023 data*. https://www.gov.uk/government/publications/pig-population-in-great-britain-annual-reports

Animal and Plant Health Agency. (2024c). Livestock demographic data group: Sheep population report. In *GOV.UK: Livestock population reports for Great Britain, using December 2022 / January 2023 data*. https://www.gov.uk/government/publications/sheep-and-goat-population-in-great-britain-annual-reports

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086.

Aybar, C., Wu, Q., Bautista, L., Yali, R., & Barja, A. (2020). Rgee: An r package for interacting with google earth engine. *Journal of Open Source Software*, *5*(51), 2272.

Barber, R. A., Ball, S. G., Morris, R. K., & Gilbert, F. (2022). Target-group backgrounds prove effective at correcting sampling bias in maxent models. *Diversity and Distributions*, *28*(1), 128–141.

BGS geology 625K: Superficial deposits. (2024). In *British Geological Survey*. https://www.bgs.ac.uk/datasets/bgs-geology-625k-digmapgb/

Biecek, P. (2018). DALEX: Explainers for complex predictive models in r. *Journal of Machine Learning Research*, *19*(84), 1–5.

Bisanzio, D., Amore, G., Ragagli, C., Tomassone, L., Bertolotti, L., & Mannelli, A. (2014). Temporal variations in the usefulness of normalized difference vegetation index as a predictor for ixodes ricinus (acari: Ixodidae) in a borrelia lusitaniae focus in tuscany, central italy. *Journal of Medical Entomology*, *45*(3), 547–555.

Boulanger, N., Aran, D., Maul, A., Camara, B. I., Barthel, C., Zaffino, M., Lett, M.-C., Schnitzler, A., & Bauda, P. (2024). Multiple factors affecting ixodes ricinus ticks and associated pathogens in european temperate ecosystems (northeastern france). *Scientific Reports*, *14*(1), 9391.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Croft, S., Ward, A. I., Aegerter, J. N., & Smith, G. C. (2019). Modeling current and potential distributions of mammal species using presence-only data: a case study on British deer. *Ecology and Evolution*, *9*(15), 8724–8735. https://doi.org/10.1002/ece3.5424

Cull, B., Hansford, K. M., McGinley, L., Gillingham, E. L., Vaux, A. G. C., Smith, R., & Medlock, J. M. (2021). A nationwide study on *Borrelia burgdorferi* s.l. infection rates in questing *Ixodes ricinus*: a six-year snapshot study in protected recreational areas in England and Wales. *Medical and Veterinary Entomology*, *35*(3), 352–360. https://doi.org/10.1111/mve.12503

Cunze, S., Glock, G., Kochmann, J., & Klimpel, S. (2022). Ticks on the move - climate change-induced range shifts of three tick species in Europe: current and future habitat suitability for Ixodes ricinus in comparison with Dermacentor reticulatus and Dermacentor marginatus. *Parasitology Research*, *121*(8), 2241–2252. https://doi.org/10.1007/s00436-022-07556-x

Dennis, R., & Thomas, C. (2000). Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. *Journal of Insect Conservation*, *4*, 73–77.

Developers, G. (2024). NASA SRTM Digital Elevation 30m Earth Engine Data Catalog. In *Google for Developers*. https://developers.google.com/earth-engine/datasets/catalog/USGS_SRTMGL1_003

Dubrey, S. W., Bhatia, A., Woodham, S., & Rakowicz, W. (2014). Lyme disease in the united kingdom. *Postgraduate Medical Journal*, *90*(1059), 33–42.

Estrada-Peña, A. (2001). Distribution, abundance, and habitat preferences of ixodes ricinus (acari: Ixodidae) in northern spain. *Journal of Medical Entomology*, *38*(3), 361–370. https://doi.org/10.1603/0022-2585-38.3.361

Estrada-Peña, A. (2008). Climate, niche, ticks, and models: What they are and how we should interpret them. *Parasitology Research*, *103*, 87–95.

Estrada-Peña, A., Venzal, J., & Sánchez Acedo, C. (2006). The tick ixodes ricinus: Distribution and climate preferences in the western palaearctic. *Medical and Veterinary Entomology*, *20*(2), 189–197.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Gandy, S. L., Hansford, K. M., & Medlock, J. M. (2023). Possible expansion of ixodes ricinus in the united kingdom identified through the tick surveillance scheme between 2013 and 2020. *Medical and Veterinary Entomology*, *37*(1), 96–104.

Gandy, S., Hansford, K., McGinley, L., Cull, B., Smith, R., Semper, A., Brooks, T., Fonville, M., Sprong, H., Phipps, P., Johnson, N., & Medlock, J. M. (2022). Prevalence of Anaplasma phagocytophilum in questing Ixodes ricinus nymphs across twenty recreational areas in England and Wales. *Ticks and Tick-Borne Diseases*, *13*(4), 101965. https://doi.org/10.1016/j.ttbdis.2022.101965

Gandy, S., Medlock, J., Cull, B., Smith, R., Gibney, Z., Sewgobind, S., Parekh, I., Harding, S., Johnson, N., & Hansford, K. (2024). Detection of Babesia species in questing Ixodes ricinus ticks in England and Wales. *Ticks and Tick-Borne Diseases*, *15*(1), 102291. https://doi.org/10.1016/j.ttbdis.2023.102291

Gilbert, L. (2015). Louping ill virus in the UK: a review of the hosts, transmission and ecological consequences of control. *Experimental and Applied Acarology*, *68*(3), 363–374. https://doi.org/10.1007/s10493-015-9952-x

Goldstein, V., Boulanger, N., Schwartz, D., George, J.-C., Ertlen, D., Zilliox, L., Schaeffer, M., & Jaulhac, B. (2018). Factors responsible for ixodes ricinus nymph abundance: Are soil features indicators of tick abundance in a french region where lyme borreliosis is endemic? *Ticks and Tick-Borne Diseases*, *9*(4), 938–944.

Gray, J., Kahl, O., & Zintl, A. (2021). What do we still need to know about Ixodes ricinus? *Ticks and Tick-Borne Diseases*, *12*(3), 101682. https://doi.org/10.1016/j.ttbdis.2021.101682

Hansford, K. M., Gandy, S. L., Gillingham, E. L., McGinley, L., Cull, B., Johnston, C., Catton, M., & Medlock, J. M. (2023). Mapping and monitoring tick (acari, ixodida) distribution, seasonality, and host associations in the united kingdom between 2017 and 2020. *Medical and Veterinary Entomology*, *37*(1), 152–163.

Hansford, K. M., McGinley, L., Wilkinson, S., Gillingham, E. L., Cull, B., Gandy, S., Carter, D. P., Vaux, A. G., Richards, S., Hayes, A., et al. (2021). Ixodes ricinus and borrelia burgdorferi sensu lato in the royal parks of london, UK. *Experimental and Applied Acarology*, *84*(3), 593–606.

Hochachka, W. M., Fink, D., Hutchinson, R. A., Sheldon, D., Wong, W.-K., & Kelling, S. (2012). Data-intensive science applied to broad-scale citizen science. *Trends in Ecology & Evolution*, *27*(2), 130–137. https://doi.org/10.1016/j.tree.2011.11.006

Holding, M., Dowall, S. D., Medlock, J. M., Carter, D. P., Pullan, S. T., Lewis, J., Vipond, R., Rocchi, M. S., Baylis, M., & Hewson, R. (2020). Tick-borne encephalitis virus, united kingdom. *Emerging Infectious Diseases*, *26*(1), 90–96. https://doi.org/10.3201/eid2601.191085

Hollis, D., McCarthy, M., Kendon, M., Legg, T., & Simpson, I. (2019). HadUK-grid—a new UK dataset of gridded climate observations. *Geoscience Data Journal*, *6*(2), 151–159.

Janzén, T., Hammer, M., Petersson, M., & Dinnétz, P. (2023). Factors responsible for Ixodes ricinus presence and abundance across a natural-urban gradient. *PLOS ONE*, *18*(5), e0285841. https://doi.org/10.1371/journal.pone.0285841

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, *14*(2), 401–413.

Kjær, L. J., Soleng, A., Edgar, K. S., Lindstedt, H. E. H., Paulsen, K. M., Andreassen, Å. K., Korslund, L., Kjelland, V., Slettan, A., Stuen, S., Kjellander, P., Christensson, M., Teräväinen, M., Baum, A., Klitgaard, K., & Bødker, R. (2019). Predicting and mapping human risk of exposure to Ixodes ricinus nymphs using climatic and environmental data, Denmark, Norway and Sweden, 2016. *Eurosurveillance*, *24*(9), 1800101. https://doi.org/10.2807/1560-7917.ES.2019.24.9.1800101

Kuhn, M. (2024). *Finetune: Additional functions for model tuning*. https://CRAN.R-project.org/package=finetune

Kuhn, M., Wickham, H., & Hvitfeldt, E. (2024). *Recipes: Preprocessing and feature engineering steps for modeling*. https://CRAN.R-project.org/package=recipes

Leonardi, M., Colucci, M., Pozzi, A. V., Scerri, E. M., & Manica, A. (2024). Tidysdm: Leveraging the flexibility of tidymodels for species distribution modelling in r. *Methods in Ecology and Evolution*.

Liu, C., Newell, G., & White, M. (2019). The effect of sample size on the accuracy of species distribution models: Considering both presences and pseudo-absences or background sites. *Ecography*, *42*(3), 535–548. https://doi.org/10.1111/ecog.03188

Marston, C. G., O'Neil, A. W., Morton, R. D., Wood, C. M., & Rowland, C. S. (2023). LCM2021–the UK land cover map 2021. *Earth System Science Data*, *15*(10), 4631–4649.

Matutini, F., Baudry, J., Pain, G., Sineau, M., & Pithon, J. (2021). How citizen science could improve species distribution models and their independent assessment. *Ecology and Evolution*, *11*(7), 3028–3039.

Mavin, S., Guntupalli, S., & Robb, M. (2024). Incidence and management of Lyme disease: a Scottish general practice retrospective study. *BJGP Open*, *8*(3), BJGPO.2023.0241. https://doi.org/10.3399/bjgpo.2023.0241

Medlock, J. M., Hansford, K. M., Bormane, A., Derdakova, M., Estrada-Peña, A., George, J.-C., Golovljova, I., Jaenson, T. G. T., Jensen, J.-K., Jensen, P. M., Kazimirova, M., Oteo, J. A., Papa, A., Pfister, K., Plantard, O., Randolph, S. E., Rizzoli, A., Santos-Silva, M. M., Sprong, H., … Van Bortel, W. (2013). Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasites & Vectors*, *6*(1), 1. https://doi.org/10.1186/1756-3305-6-1

Medlock, J. M., & Leach, S. A. (2015). Effect of climate change on vector-borne disease risk in the UK. *The Lancet Infectious Diseases*, *15*(6), 721–730. https://doi.org/10.1016/s1473-3099(15)70091-5

Medlock, J. M., Pietzsch, M. E., Rice, N. V. P., Jones, L., Kerrod, E., Avenell, D., Los, S., Ratcliffe, N., Leach, S., & Butt, T. (2008). Investigation of Ecological and Environmental Determinants for the Presence of Questing Ixodes ricinus (Acari: Ixodidae) on Gower, South Wales. *Journal of Medical Entomology*, *45*(2), 314–325. https://doi.org/10.1093/jmedent/45.2.314

Medlock, J. M., Vaux, A. G. C., Gandy, S., Cull, B., McGinley, L., Gillingham, E., Catton, M., Pullan, S. T., & Hansford, K. M. (2022). Spatial and temporal heterogeneity of the density of *Borrelia burgdorferi*-infected *Ixodes ricinus* ticks across a landscape: A 5-year study in southern England. *Medical and Veterinary Entomology*, *36*(3), 356–370. https://doi.org/10.1111/mve.12574

Milne, A. (1950). The ecology of the sheep tick, *Ixodes ricinus* L. spatial distribution. *Parasitology*, *40*(1-2), 35–45. https://doi.org/10.1017/s0031182000017832

Nieto, N. C., Porter, W. T., Wachara, J. C., Lowrey, T. J., Martin, L., Motyka, P. J., & Salkeld, D. J. (2018). Using citizen science to describe the prevalence and distribution of tick bite and exposure to tick-borne diseases in the United States. *PLOS ONE*, *13*(7), e0199644. https://doi.org/10.1371/journal.pone.0199644

Noll, M., Wall, R., Makepeace, B. L., Newbury, H., Adaszek, L., Bødker, R., Estrada-Peña, A., Guillot, J., Fonseca, I. P. da, Probst, J., Overgaauw, P., Strube, C., Zakham, F., Zanet, S., & Rose Vineer, H. (2023). Predicting the distribution of Ixodes ricinus and Dermacentor reticulatus in Europe: a comparison of climate niche modelling approaches. *Parasites & Vectors*, *16*(1). https://doi.org/10.1186/s13071-023-05959-y

Office for Health Improvement and Disparities. (2025). *Public health profiles*. https://fingertips.phe.org.uk

Office for National Statistics. (2024). *Lower layer super output areas (december 2021) boundaries EW BSC (V4)*. https://geoportal.statistics.gov.uk/datasets/ons::lower-layer-super-output-areas-december-2021-boundaries-ew-bsc-v4-2/about

Ogden, N. H., & Lindsay, L. R. (2016). Effects of Climate and Climate Change on Vectors and Vector-Borne Diseases: Ticks Are Different. *Trends in Parasitology*, *32*(8), 646–656. https://doi.org/10.1016/j.pt.2016.04.015

Olsthoorn, F., Gilbert, L., Fonville, M., Blache, N., May, L., Mondini, F., Rotbarth, R., Schlierenzauer, C., Gandy, S., Sprong, H., & Ghazoul, J. (2025). Woodland expansion and deer management shape tick abundance and Lyme disease hazard. *Ecological Solutions and Evidence*, *6*(1). https://doi.org/10.1002/2688-8319.12403

Oppel, S., Strobl, C., & Huettmann, F. (2009). Alternative methods to quantify variable importance in ecology. *Technical Report, University of Munich Department of Statistics*, *65*.

Ostfeld, R. S., & Brunner, J. L. (2015). Climate change and *Ixodes* tick-borne diseases of humans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1665), 20140051. https://doi.org/10.1098/rstb.2014.0051

Panagos, P. et al. (2006). The european soil database. *GEO: Connexion*, *5*(7), 32–33.

Perret, J.-L., Guerin, P. M., Diehl, P. A., Vlimant, M., & Gern, L. (2003). Darkness induces mobility, and saturation deficit limits questing duration, in the tick*Ixodes ricinus*. *Journal of Experimental Biology*, *206*(11), 1809–1815. https://doi.org/10.1242/jeb.00345

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

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540.

Ribeiro, R., Eze, J. I., Gilbert, L., Wint, G. R. W., Gunn, G., Macrae, A., Medlock, J. M., & Auty, H. (2019). Using imperfect data in predictive mapping of vectors: A regional example of Ixodes ricinus distribution. *Parasites & Vectors*, *12*(1), 536. https://doi.org/10.1186/s13071-019-3784-1

Rochat, E., Vuilleumier, S., Aeby, S., Greub, G., & Joost, S. (2020). Nested Species Distribution Models of *chlamydiales* in Ixodes ricinus (Tick) Hosts in Switzerland. *Applied and Environmental Microbiology*, *87*(1), e01237–20. https://doi.org/10.1128/AEM.01237-20

Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, *44*(2), 199–205.

Semenza, J. C., & Suk, J. E. (2017). Vector-borne diseases and climate change: a European perspective. *FEMS Microbiology Letters*, *365*(2). https://doi.org/10.1093/femsle/fnx244

Signorini, M., Stensgaard, A.-S., Drigo, M., Simonato, G., Marcer, F., Montarsi, F., Martini, M., & Cassini, R. (2019). Towards improved, cost-effective surveillance of ixodes ricinus ticks and associated pathogens using species distribution modelling. *Geospatial Health*, *14*(1). https://doi.org/10.4081/gh.2019.745

The european soil database distribution version 2.0. (2004). In *European Commission and the European Soil Bureau Network*. https://esdac.jrc.ec.europa.eu/content/european-soil-database-v20-vector-and-attribute-data

UKHSA. (2022). Lyme disease epidemiology and surveillance. In *GOV.UK*. https://www.gov.uk/government/publications/lyme-borreliosis-epidemiology/lyme-borreliosis-epidemiology-and-surveillance

Uusitalo, R., Siljander, M., Lindén, A., Sormunen, J. J., Aalto, J., Hendrickx, G., Kallio, E., Vajda, A., Gregow, H., Henttonen, H., Marsboom, C., Korhonen, E. M., Sironen, T., Pellikka, P., & Vapalahti, O. (2022). Predicting habitat suitability for Ixodes ricinus and Ixodes persulcatus ticks in Finland. *Parasites & Vectors*, *15*(1). https://doi.org/10.1186/s13071-022-05410-8

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., & Elith, J. (2022). Predictive performance of presence-only species distribution models: A benchmark study with reproducible code. *Ecological Monographs*, *92*(1), e01486.

Wadoux, A. M. J.-C., Heuvelink, G. B. M., Bruin, S. de, & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wright, M. N., & Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv Preprint arXiv:1508.04409*.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802.

Modelling the distribution of the tick *Ixodes ricinus* in England and Wales using passive surveillance data from citizen science reports: Supplementary Materials

Supplementary Materials

Model Hyperparameters

In the final selected penalized GLM model, alpha (the mixture argument) was set to 0.471 and lambda (the penalty argument) was set to 0. The pGLM was therefore an elastic net model (Zhou and Hastie 2005). The variables rainfall (mm): minimum, sunshine hours: maximum *, soil type: cambisol and geology: missing were removed by the penalization.

For the xgBoost model, the proportion of variables included in training, the maximum tree depth, the minimum loss reduction and the number of early stopping iterations are all set by tuning as described earlier. The single best hyperparameter combination in the main xgBoost specification used all variables, a maximum tree depth of 8, a minimum loss reduction of 4.2878068^{-4} and 17 early stopping rounds.

For the Random Forest model, hyperparameter tuning determined the number of trees, the number of predictors sampled for each split, and the minimum node size. In the best version of the Random Forests model, 198 trees are used, with 46 predictors randomly sampled for each split, and a minimum of 10 datapoints is required to split a node further.

Population Sparsity

Figure S1 shows the population sparsity weighting applied to each Middle layer Super Output Area (MSOA). The histogram shows how many MSOAs had sparsity weights within 1% bins.

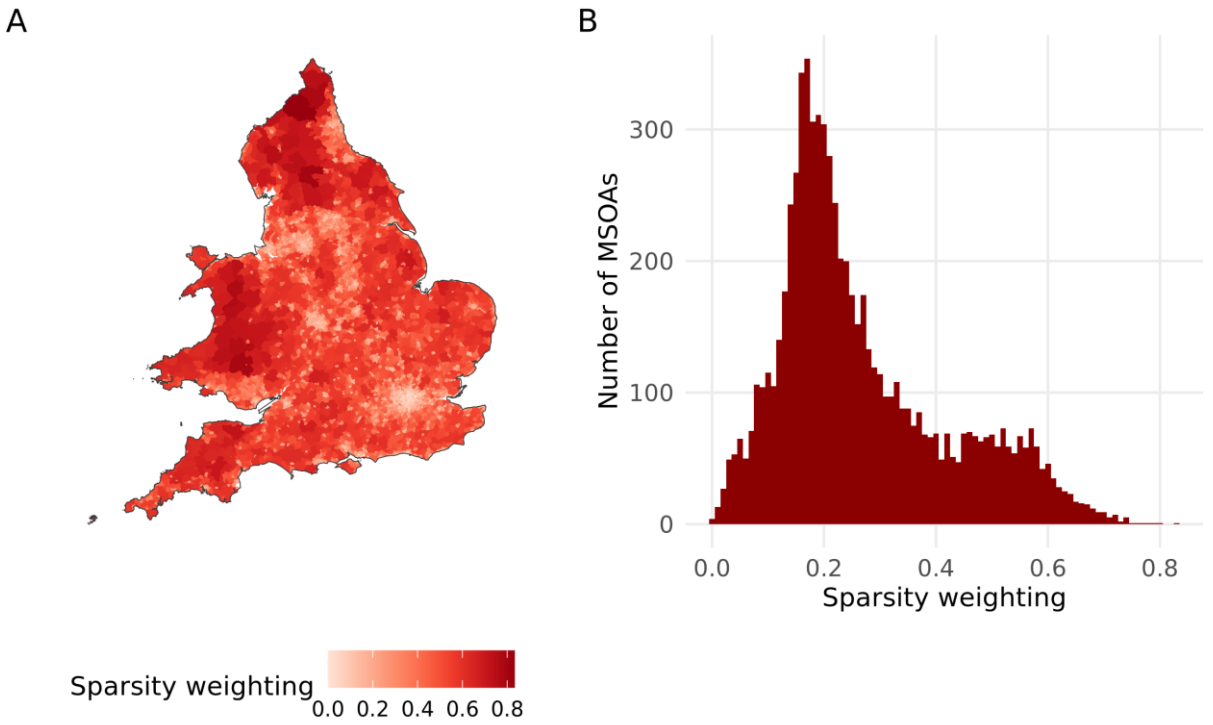

Figure S1: A: Map of England and Wales showing the sparsity weighting for each area. B: Histogram showing the number of MSOAs and their sparsity weightings.

Additional Sensitivity Testing

As a final check on how the scenarios described in the main text of the paper differ, Table S1 shows how the success metrics vary by scenario. Note that these are not fully comparable due to differences in background points. All scenarios except "Only random background points" are tested against both bias adjusted and random points, without weighting applied; that scenario is tested only against random background points, to make for a fairer comparison.

Table S1: Model performance metrics for the different sensitivity analysis scenarios.

	Class predictions					Probability predictions			
Scenario	Accuracy	Kappa	MCC	Sensitivity	Specificity	Boyce Continuous	ROC AUC	Brier score	Max TSS
Baseline	0.81	0.46	0.46	0.62	0.86	0.99	0.84	0.14	0.54
Human records only	0.82	0.51	0.52	0.74	0.84	0.97	0.86	0.14	0.61
10km spatial thinning	0.68	0.11	0.11	0.36	0.77	0.66	0.60	0.21	0.19
Increased background point to presence points ratio (5:1)	0.84	0.33	0.35	0.60	0.86	0.99	0.83	0.13	0.51
No limit on random background points' distance from presences	0.78	0.40	0.40	0.62	0.82	0.98	0.82	0.15	0.52
Only random background points	0.77	0.49	0.49	0.67	0.82	0.95	0.85	0.15	0.54
Only bias-adjusted	0.80	0.56	0.56	0.69	0.86	0.98	0.85	0.15	0.58

background points									
Equal weight between random and bias-adjusted background points	0.80	0.43	0.44	0.63	0.84	0.98	0.83	0.15	0.54
No population weighting applied	0.82	0.47	0.47	0.63	0.86	0.99	0.84	0.14	0.53

In addition to taking a simple average of predictions (the 'simple ensemble' strategy), we also tested putting more weight on some models than others. We took weighted averages of the predictions from the best iterations of the individual base models, both using metric targeting and manual choice. To target performance metrics, we took the sum of the ROC AUC and maximum True Skill Score metrics on the training data separately (NB: our preferred overall performance metric is the continuous Boyce index, but there was little variation between the models on this metric in the training set). We then weighted each model proportionally to its contribution to the summed metric across the four models; this aims to give more weight to models that give better performance. We also manually specified two weighting schemes to illustrate the impact of changing weights: a statistics-focused scheme that assigned 45% of the weight to each of the pGLM and GAM, and 5% to each of the xgBoost and Random Forest models, and a machine learning-focused scheme with the inverse weightings (45% for xgBoost and for Random Forest, 5% each for pGLM and GAM). Table S2 shows the performance on the testing set from the alternative model ensembling strategies, with the simple ensemble performance repeated for ease of comparison.

Table 2: Predictive performance metrics for the alternative model ensembles. These metrics are based on the testing set, which is reports from 2024. All metrics are based on unweighted data.

	Class predictions				Probability predictions			
Model	**Accuracy**	**Kappa**	**Sensitivity**	**Specificity**	**Boyce Continuous**	**ROC AUC**	**Brier score**	**Max TSS**
Simple ensemble	0.82	0.47	0.63	0.86	0.99	0.84	0.14	0.53
ML-focused	0.83	0.49	0.61	0.88	0.99	0.86	0.12	0.57
Stats-focused	0.74	0.33	0.61	0.78	0.99	0.79	0.17	0.45
Max AUC	0.82	0.47	0.62	0.87	0.99	0.85	0.13	0.55

Max TSS	0.83	0.49	0.62	0.88	0.99	0.85	0.13	0.56

The weighted average-based ensembles did not diverge dramatically from the simple average ensemble; placing more weight on machine learning models again increased specificity at the expense of sensitivity, whereas extra weight on the GAM and pGLM resulted in higher sensitivity than any other model strategy. Attempting to maximise specific performance metrics based on the training data was unsuccessful, indicating that the differences in model performance were not generalisable to new data.

Average Local Effect Plots

Figure S2 shows average local effect plots as an alternative to partial dependence plots. Comparison to the main text shows that the trends are very similar between the two plot types.

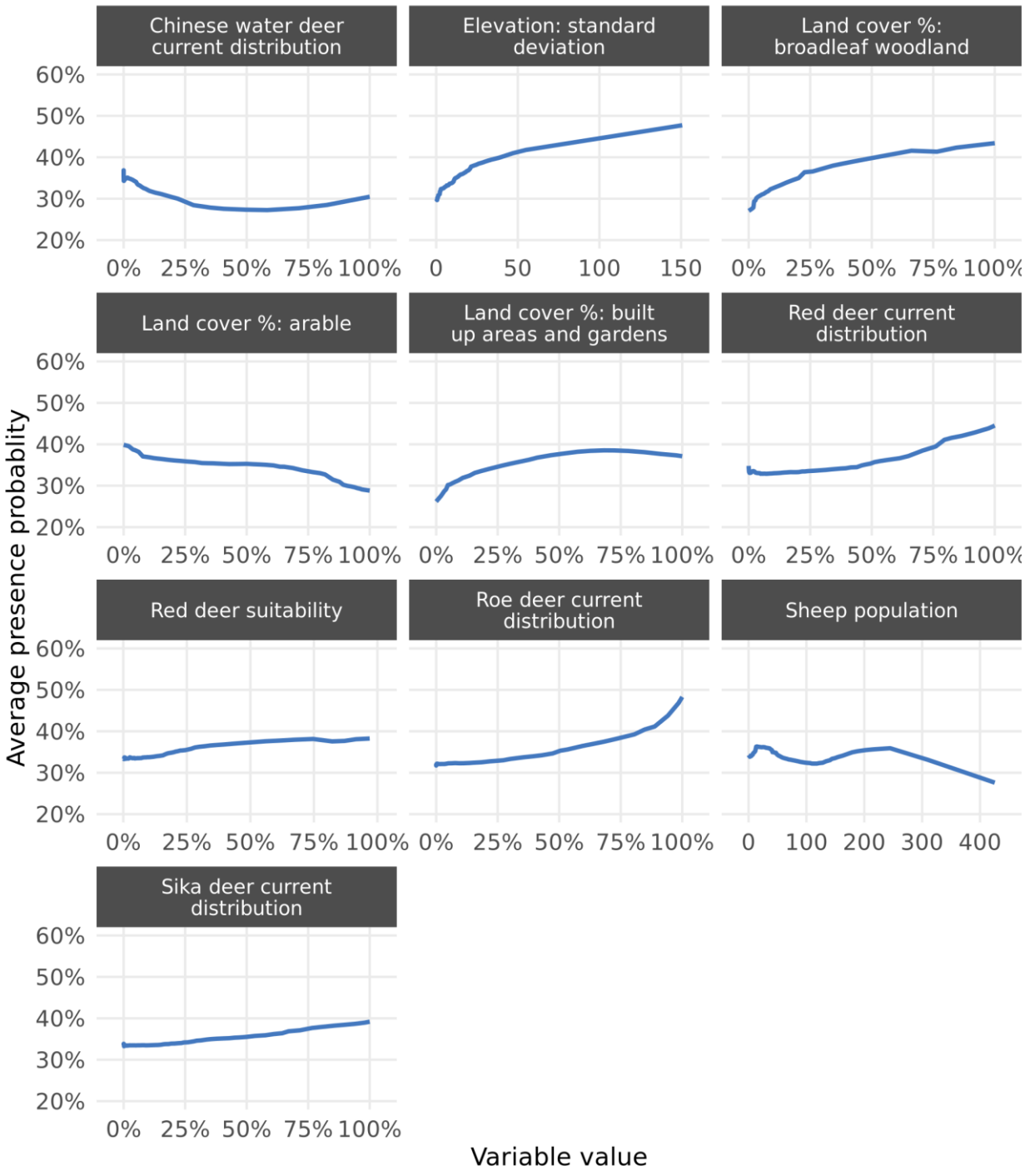

Figure S2: Average Local Effect plots (ALEs) showing local predictions from the simple ensemble model, for the full range of the ten most important predictors in the testing data. Other predictors are set to locally relevant values.